METHODS OF MICROARRAY DATA ANALYSIS II

edited by
Simon M. Lin
Kimberly F. Johnson

KLUWER ACADEMIC PUBLISHERS

METHODS OF MICROARRAY DATA ANALYSIS II

Papers from CAMDA '01

This page intentionally left blank

METHODS OF MICROARRAY DATA ANALYSIS II

Papers from CAMDA '01

edited by

Simon M. Lin

and

Kimberly F. Johnson

Duke University Medical Center

KLUWER ACADEMIC PUBLISHERS
NEW YORK, BOSTON, DORDRECHT, LONDON, MOSCOW

eBook ISBN: 0-306-47598-7
Print ISBN: 1-4020-7111-6

Visit Kluwer Online at: http://kluweronline.com
and Kluwer's eBookstore at: http://ebooks.kluweronline.com

Contents

Contributors

Baggerly, Keith, Department of Biostatistics, UT M.D. Anderson Cancer Center, Houston TX
Bidaut, Ghislain, Biomedical Informatics, Fox Chase Cancer Center, Philadelphia PA and Structural and Genetic Information Laboratory, CNRS-AVENTIS, Marseille, France
Bolognie, Serge, Department of Computer Science and Engineering, University of California at San Diego, San Diego CA
Bobshev, Georgiy, Statistics Research Division, RTI, Research Triangle Park NC
Carin, Lawrence, Department of Electrical Engineering and Computer Engineering, Duke University, Durham NC
Chang, Jeong-Ho, Biointelligence Laboratory, School of Computer Science and Engineering, Seoul National University, Seoul Korea
Coombes, Kevin, Department of Biostatistics, UT M.D. Anderson Cancer Center, Houston TX
Das, A, Statistics Research Division, RTI, Research Triangle Park NC
Das, S, Statistics Research Division, RTI, Research Triangle Park NC
Dasgupta, Nilanjan, Department of Electrical Engineering, Duke University, Durham NC
Dopazo, Joaquin, Bioinformatics Unit, Spanish National Cancer Center, Madrid Spain
Fowlkes, Charless, Department of Computer Science, University of California at Berkeley, Berkeley CA
Gold, David, Department of Biostatistics, UT M.D. Anderson Cancer Center, Houston TX
Goldschmidt, Pascal, Department of Cardiology, Duke University Medical Center, Durham NC
Grant, Jeffrey, Biomedical Informatics, Fox Chase Cancer Center, Philadelphia PA
Herrero, Javier, Bioinformatics Unit, Spanish National Cancer Center, Madrid Spain
Hwang, Kyu-Baek, Biointelligence Laboratory, School of Computer Science and Engineering, Seoul National University, Seoul Korea
Johnson, Kimberly, Duke Bioinformatics Shared Resource, Duke University Medical Center, Durham NC
Lee, Sang-Joon, Department of Biostatistics, UT M.D. Anderson Cancer Center, Houston TX
Li, Yi-Ju, Center for Human Genetics, Duke University Medical Center, Durham NC
Lin, Simon, Duke Bioinformatics Shared Resource, Duke University Medical Center, Durham NC

Liao, Xuejun, Department of Electrical Engineering and Computer Engineering, Duke University, Durham NC
Lockhart, David, Ambit Biosciences, San Diego, CA and Laboratory of Genetics, The Salk Institute for Biological Studies, La Jolla CA
Malik, Jitendra, Departments of Computer Science and Molecular Cell Biology, University of California at Berkeley
Manion, Frank, Biomedical Informatics, Fox Chase Cancer Center, Philadelphia PA
Martin, Eden, Center for Human Genetics, Duke University Medical Center, Durham NC
Mateos, Alvaro, Bioinformatics Unit, Spanish National Cancer Center, Madrid Spain
McConnell, Patrick, Duke Bioinformatics Shared Resource, Duke University Medical Center, Durham NC
Moloshok, Thomas, Biomedical Informatics, Fox Chase Cancer Center, Philadelphia PA
Ochs, Michael, Biomedical Informatics, Fox Chase Cancer Center, Philadelphia PA
Page, Grier, Department of Biometry and Epidemiology, Medical University of SC, Charleston SC
Shan, Qun, Department of Molecular Cell Biology, University of California at Berkeley, Berkeley CA
Sluka, James, Inpharmix, Inc., Greenwood IN
Speer, Marcy, Center for Human Genetics, Duke University Medical Center, Durham NC
Stivers, David, Department of Biostatistics, UT M.D. Anderson Cancer Center, Houston TX
Sung, Hsi-Guang, Department of Biostatistics, UT M.D. Anderson Cancer Center, Houston TX
Tamames, Javier, ALMA Bioinformatics SL, Madrid Spain
Vata, Korkut, Department of Cardiology, Duke University Medical Center, Durham NC
Wang, Jing, Department of Biostatistics, UT M.D. Anderson Cancer Center, Houston TX
Zhang, Byoung-Tak, Biointelligence Laboratory, School of Computer Science and Engineering, Seoul National University, Seoul Korea
Zhang, Hong, Department of Computer Science, Armstrong Atlantic State University, Savannah GA
Zhang, Zhen, Center for Biomarker Discovery, Department of Pathology, Johns Hopkins Medical Institutions, Baltimore MD, and 3Z Informatics, Mt. Pleasant SC
Zhang, Ling, Center for Human Genetics, Duke University Medical Center, Durham NC, and Bioinformatics Group, Statistics Department, North Carolina State University, Raleigh NC

Acknowledgements

The editors would like to thank the contributing authors for their fine work, as well as Anna Menzies, Patrick McConnell, and Emily Allred for their assistance in preparing this volume. We especially thank our supporters at Duke University; John Harer, Vice Provost for Academic Affairs and Jim Siedow, Vice Provost for Research and Interim Director of the Center for Bioimformatics and Computational Biology. CAMDA would not be possible without the contributions of the scientific committee and other reviewers who contribute to the scientific review process. Our thanks for the time and effort they commit to CAMDA. We particularly appreciate the continued encouragement of John Weinstein at the NCI. Finally, we would like to acknowledge the many years of support and mentorship provided by Michael Colvin of the Duke Comprehensive Cancer Center. His vision for a Bioinformatics Shared Resource allows CAMDA to flourish.

Reviewers

David Adams (Duke)
Cindy Afshari (NIEHS)
Bruce Aranow (U Cincinatti)
Burns Blaxall (Duke)
George Bobashev (RTI)
Philippe Broet (INSERM)
Chris Gorton (CIIT)
Joaquin Dopazo (CNIO)
J. Gormley (MBI)
Greg Grant (U Penn)
Patrick Hurban (Paradigm Genetics)
Stuart Hwang (COR Therapeutics)
Bret Jessee (AnVil Informatics)
Warren Jones (UAB)
D. P. Kreil (EBI)
Elisabetta Manduchi (U Penn)
Robert Nadon (Imaging Research)
Jean Roayaei (UNC)
John Rockett (EPA)
Raymond Samaha (Applied Biosystems)
Jennifer Shoemaker (Duke)
Dawn Wilkins (U Mississippi)
Thomas Wu (Genentech)
Fei Zou (UNC)

This page intentionally left blank

Preface

Advances in microarray technology continue to increase the amount of data available to researchers, and the need for new analytical tools has never been greater. The search for new methods continued with the second CAMDA conference held in October of 2001. The second volume of Methods of Microarray Data Analysis highlights ten papers presented at the conference and presents three review papers to provide readers with a broad overview of microarrays, experimental design, and analytical methods. As editors, we have not comprehensively edited these papers, but have provided comments to the authors to encourage clarity and expansion of ideas. Each paper was peer-reviewed and returned to the author for further revisions.

Again, we do not propose these methods as the *de facto* standard for microarray analysis. However, the CAMDA conference continues to provide a forum for the scientific community to work toward a standard protocol. If you have insights into new analytical methods for microarray data, please join us at the 2002 CAMDA conference.

Kimberly Johnson

Simon Lin

This page intentionally left blank

INTRODUCTION

The year 2001 marked the release of the working draft of the human genome. This monumental achievement has fueled continuous improvement of DNA microarray technology. In parallel, we have seen an accelerated emergence of novel proteomics and metabolomics technologies with the resulting data in a format analogous to DNA microarrays [Oliver *et al, Metab Eng.* 2002, 4(1):98-106; Albala, *Expert Rev Mol Diagn.* 2001, l(2):145-52.]. The challenge of analyzing this tremendous amount of bio-array data has caught the attention of many statisticians and computer scientists. To provide a forum for the comparative assessment of new analytical methods, the second Critical Assessment of Microarray Data Analysis (CAMDA) conference was held in October, 2001 with 150 researchers from nine countries in attendance. The scientific committee selected twelve papers for oral presentation, with ten highlighted here. The presentations were complemented by opening remarks and a keynote address by Dr. Roland Stoughton of Rosetta Inpharmatics. The second keynote address was presented by Dr. David Lockhart of The Salk Institute and Ambit Biosciences. Closing remarks were provided by John Weinstein from the NCI. At the end of the conference, attendees voted on the "Best Presentation" with the Scientific Committee providing weighted votes. The CAMDA'0l Best Presentation went to:

Kevin R. Coombes, Keith A. Baggerly, David N. Stivers, Jing Wang, David Gold, Hsi-Guang Sung, and Sang-Joon Lee from M.D. Anderson Cancer Center for their paper "Biology-driven Clustering of Microarray Data, Applications to the NCI 60 Data Set."

In this introduction, we describe the CAMDA '01 data sets and then briefly mention each paper in this volume, organized by specific topics. While we have tried to assign each paper to a topic, it is often difficult to accomplish this because many papers cross categories. We compare and contrast the methods presented and point out the relevant research issues associated with each method. Finally, we highlight the web companion to this book.

CAMDA 2001 Data Sets

The CAMDA 2000 papers analyzed a spotted cDNA array of yeast cell cycle data, originated by Spellman *et al. [Mol Biol Cell* 1998, 9:3273-3279], and an Affymetrix leukemia data set by Golub *et al.* [*Science* 1999, 286:531-537]. The majority of participants easily discriminated ALL from AML in the leukemia data set, whereas the yeast data presented a bigger challenge. Fewer participants attempted to analyze the yeast data. While the 2000 datasets were representative of the different types of arrays, this year's data sets were selected to represent the complexity of biological systems.

The scientific committee chose the Rosetta Compendium [*Cell.* 2000; 102(1): 109-26], from a study of 300 expression profiles of yeast mutants and chemical treatments, and the NCI60 Cancer Cell Lines with Drug Treatments [*Nature Genetics.* 2000;24(3):236-44], a pharmacogenomic database. The Rosetta Compendium represents a model organism where the entire genome is known and documented. The challenge was to extract useful biological information from this overabundance of array data. In contrast, the challenge of the NCI-60 data set was to model the relationships between gene expression levels and drug treatment response. These relationships represent critical questions in pharmacogenomics as well as the promise of clinically relevant uses for microarrays in patient care. Both data sets provide the opportunity for researchers to explore a variety of new methods.

Feature Selection and Extraction

Feature selection and extraction play an important role in genome analysis. From a pattern recognition point of view, we can think of biological samples as objects, and genes as features to describe each object. In a typical microarray data set, the number of objects is small (usually <50), but the number of features measured is often greater than ten thousand, with many of the features being either correlated or irrelevant. To circumvent this "curse of dimensionality," feature selection or extraction is necessary prior to applying pattern analysis algorithms [Jain *et al., IEEE Transactions on Pattern Analysis and Machine Intelligence. 22*,(l):4-37]. Feature selection relies on methods that determine the best features to study. Feature extraction, on the other hand, combines information from individual features into components, and describes each object by these new components instead of individual features. Identifying genes participating in the signal

transduction pathway and then analyzing them as a grouped component rather than as individual genes reduces the dimensionality while addressing the issue of biological relevance.

Four papers focus on feature selection and extraction. Zhang utilizes a supervised component analysis approach to extract features in Chapter 9, and then uses them to 'fish out' interesting profiles from the database. In Chapter 10, Dasgupta uses partial least squares (PLS) modeling as another supervised feature extraction method and combines it with predictive modeling. In contrast to the supervised strategies, the Bayesian deposition proposed by Bidault *et al.*can be viewed as an unsupervised strategy to extract biological pathways and then analyze their patterns. This approach is discussed in Chapter 7. Lin *et al.* (Chapter 8) also decomposes the data, but uses independent component analysis (ICA) to extract features. In addition, the Lin paper discusses feature selection by utilizing knowledge from the Gene Ontology annotation of the yeast genome to select relevant features based on 'expert opinion'. Both Zhang (figure 2 and 3, Chapter 9) and Lin (figure 1, Chapter 8) demonstrate that clustering results based on selected or extracted features are more comprehensible in terms of their biology than results which do not use this step.

Clustering Strategies

Clustering is a classical unsupervised learning methodology that has been applied to microarray data since 1998 (Eisen, 1998). It is still actively investigated in both clustering in algorithm development and in its application in microarrays. Fowlkes *et al.* (Chapter 5) present a new globally divisive algorithm called GENECUT as opposed to the locally agglomerative algorithm used in Hughes *et al.* GENECUT performs top-down k-way partitioning instead of commonly used binary splits. In Chapter 11, Chang *et al.* use a soft topographic vector quantization (SQVT) algorithm which defines the microarray clustering problem as an energy minimization problem. Similar to self-organizing maps (SOM), SQVT is able to provide a visual clue for the topology of the resulting clusters. Mateos (Chapter 6) presents a new algorithm to take advantage of both the hierarchical presentation of clusters and the robustness of neural networks. This self-organizing tree algorithm (SOTA) is now available as a user-friendly Java application (http://www.almabioinfo.com). A comparison of average linkage, SOM, and SOTA can be seen in table 3 of Chapter 3.

Clustering algorithms do not provide an endpoint to the analytical process. Biology-driven clustering, SQVT, and SOTA all indicate the close resemblance of the putative breast cancer cell lines (MDA-MB-435 and MDA-N) to melanomas. These methods further support speculation by Ross

et al. in the original analysis of the NCI data set concerning a breast-melanoma connection and indicate that the biology of these connections deserve further investigation.

Additional applications for clustering are also presented. The Fowlkes paper suggests that further study of the conserved transcription factor binding sites of the co-regulated genes is needed. The Chang paper suggests the use of clustering to reduce the amount of information. Without clustering first, it is impossible to run some Bayesian Network learning algorithms. Similarly, Mateos argues that, by using the mean profile of each cluster, noise could be reduced. These papers show that clustering can be an important preliminary step in modeling complex biological systems.

Modeling Complex Systems

Modeling complex biological systems has been a challenge for decades. Several papers in this volume show both the potential benefits and difficulties of modeling biological pathways. Two papers discuss the decomposition of gene expression data into multiple biological pathways. In Chaper 7, Bidaut *et al.* propose a Bayesian decomposition approach, whereas in Chapter 8, Lin *et al.* propose an independent component analysis (ICA) approach. A clear advantage of these models is that a gene is no longer constrained to a group as defined by cluster analysis. Instead, the behavior of a gene can be explained by its involvement in one or more signaling pathways.

In addition to modeling pathways, another challenge is the prediction of the behavioral response of a complex system. Chang *et al.* utilize Bayesian networks to model dependencies among gene expression, drug activity and cancer type. The visual representation of the dependency relationship in figure 4 of Chapter 11 could direct biologists to investigate further. Interestingly, the PDQ_ MED software (Chapter 13) provides a similar topology of the relationship among the concepts when used to check the co-occurrence of the items in figure 4 of Chapter 11 (see figure 1 below). The PDQ_MED software seems to further support the Chang model of the relationship between these variables although this needs additional investigation.

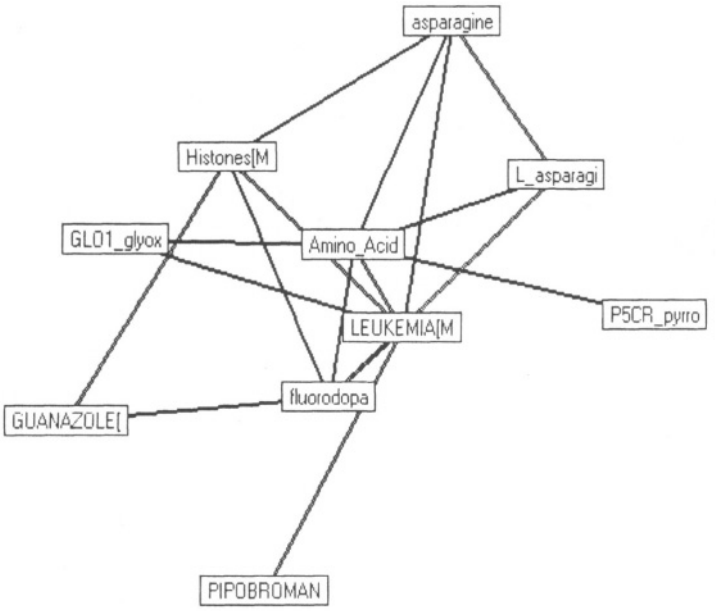


Figure 1. The co-occurrence topology in the literature of the items obtained from Bayesian learning using the Chang method. The text data analysis is done with PDQ_MED. This topology is closely related to the topology of Figure 4 of Chapter 11.

In recognition of the high dimensionality of the data and the possible correlation among features, as well as the non-linear nature of the biological system, Dasgupta (Chapter 10) explored kernel-PLS modeling to predict the pharmacogenomic response of the cell lines. The predictive capability of kernel-PLS on Taxol response (shown in figure 8 of Chapter 10) is striking, but more experimental data are necessary to validate this model.

When beginning to model complex biological systems, biologists often need to determine the differentially expressed genes and the number of replicates needed to detect these genes. Li *et al.* (Chapter12) investigates this issue and suggests a pooled variance strategy to discover the differentially expressed genes. They also discuss the minimum number of replicates required for detecting differential expression.

Ontologies, Semantic Understanding, and Functional Genomics

An ancient school of Chinese philosophers believed that the problem of making complex connections in the physical world originates in the

difficulty of naming things. In functional genomics, with extensive use of computers for reasoning, this problem becomes more severe. Usually, a gene name does not tell the whole story of its function, but functional inference is at the center of functional genomics.

For example, a gene named 'Galectin-1', is also called "GBP", "Lectin, galactose-binding, soluble, 1", "LGALS1", and "Hs.227751" in the literature and databases. A knowledgeable biologist might infer an association with galactoside-binding and lectin, but it requires some reading and a literature search to associate this gene with apoptosis [Pace *et al., J Immunol* 2000, 165(5):2331-4]. Computers can not make this inference without an explicit functional annotation of each gene. The Gene Ontology was designed by a consortium of genome biologists and bioinformaticians to solve this problem. For example, the Gene Ontology annotation of galectin-1 is: 1) galactose binding lectin, and 2) apoptosis. By using the Gene Ontology, we are one step closer to a semantic understanding of gene functions by computers and databases.

The CAMDA conference has seen an evolution of emphasis on this issue. Many of the papers in this volume include the Gene Ontology in the analytical process. However, encoding the existing biological knowledge using the Gene Ontology requires enormous effort. The paper by Sluka (Chapter 13) brought an alternative solution in the form of text data mining using a commercial software tool. PDQ_MED taps into the wealth of information in the medical literature. While the CAMDA conference would not normally include a commercial presentation of this nature, the scientific committee felt that the Sluka presentation had special merit in its focus on text data mining. Many attendees at CAMDA'01 were surprised by the ability of PDQ_MED to discover hidden links in the list of genes available from high-throughput studies. This presentation was of such high quality that attendees voted it second on the Best Presentation ballots.

A standard protocol?

The two data sets at CAMDA'00 and those of CAMDA '01 continue to be used as standard data sets in many areas such as workshops, graduate courses and publications. Without an agreed upon synthetic data set as a benchmark, the CAMDA data sets have become *de facto* standards when testing new algorithms due to the number of papers available for reference.

With continuing technological improvements, microarray analysis may become routine in many labs. The CAMDA conference has served to catalyze the interdisciplinary research of bioinformatics and genomics. In the year to come, we expect to see more new data analysis strategies emerge. These new methods bring us closer to a consensus on data analysis methods.

Web Companion

Additional information for many of these chapters can be found at the CAMDA website, where links to algorithms, color versions of several figures, and conference presentation slides can be found. Information about future CAMDA conferences is available at this site as well. Please check the website regularly for the call for papers and announcements about the next conference.

www.camda.duke.edu

This page intentionally left blank

1

AN INTRODUCTION TO DNA MICROARRAYS

Patrick McConnell[1], Kimberly Johnson[1], David J. Lockhart[2]
[1] *Duke Bioinformatics Shared Resource*
[2] *Ambit Biosciences & The Salk Institute for Biological Studies, Laboratory of Genetics*

Abstract: Oligonucleotide and DNA arrays, or microarrays, have proven to be useful tools to investigate biological function, and are the focal point of an increasing number of studies. However, most papers shed little light on the underlying basis and best use of microarray technology, often leaving a number of important questions unanswered. Under what conditions are microarrays helpful? How should microarray data be analyzed? What data analysis methods should be avoided? Which biological questions can microarrays address? Which biological questions are not best answered by microarrays? Here, we examine the technology itself, the data produced, proper experimental design, data analysis techniques, and experimental validation. These are issues important to all users of DNA arrays, from the mathematician who may have only limited knowledge of the biology behind the technology, to the biologist who is concerned with experimental design and the details of data analysis. Finally, we stress the importance of great experimental care, sample and data triage, well-characterized and rigorous analysis, and the need for appropriate follow-up and verification, especially when using animal or human tissue.

Key words: microarrays, functional genomics, data analysis, validation, experimental design

1. INTRODUCTION TO FUNCTIONAL GENOMICS

Biological and biomedical research is in the midst of a significant transition that is being driven by two primary factors: the massive increase in the amount of DNA sequence information and the development of technologies to exploit its use. Consequently, we find ourselves at a time when new types of experiments are possible, and observations, analyses and discoveries are being made on an unprecedented scale. Over the past few

years, more than 60 organisms have had their genomes completely sequenced, with another 170 or so in progress (see www.tigr.org or genomes@ncbi.nlm.nih.gov for a list). The sequence of the human genome has been deciphered, in both public and private efforts, and the complete sequence of the mouse and other animal and plant genomes are close behind. Unfortunately, the billions of bases of DNA sequence do not tell us what all of the genes do, how cells work, how cells form organisms, what goes wrong in disease, how we age or how to develop a drug. Thus, functional genomics has become an increasingly important scientific discipline.

The purpose of functional genomics is to understand biology, not simply to identify the component parts, and new experimental and computational methods take advantage of as much sequence information as possible. Unlike the genome sequencing efforts, functional genomics is less a specific project or program than it is a mindset and general approach to problems. The goal is not simply to provide a catalogue of all the genes and information about their functions, but to understand how the components work together to comprise functioning cells and organisms.

2. MICROARRAY TECHNOLOGY

To take advantage of the large and rapidly increasing body of sequence information, new technologies are required. Among the most powerful and versatile tools for genomics are high-density arrays of oligonucleotides (short strands of nucleic acids) or complementary DNAs (see Figure 1 for an overview) [Lockhart *et al.*, 1996; Schena *et al.*, 1995]. DNA arrays work by hybridization (non-covalent chemical bonding) of fluorescently labeled RNA or DNA in solution to DNA molecules (probes) that are attached to specific locations on the chip surface. The hybridization reactions take place in parallel across the entire array at the same time. Thus, the hybridization of a sample to an array is, in effect, a highly parallel search by each molecule for a matching partner on an 'affinity matrix,' with the eventual binding of labeled molecules to the surface-bound probe determined by the rules of molecular recognition. The process is straightforward, highly parallel (all sequences are counted simultaneously), and, if done correctly, quantitative.

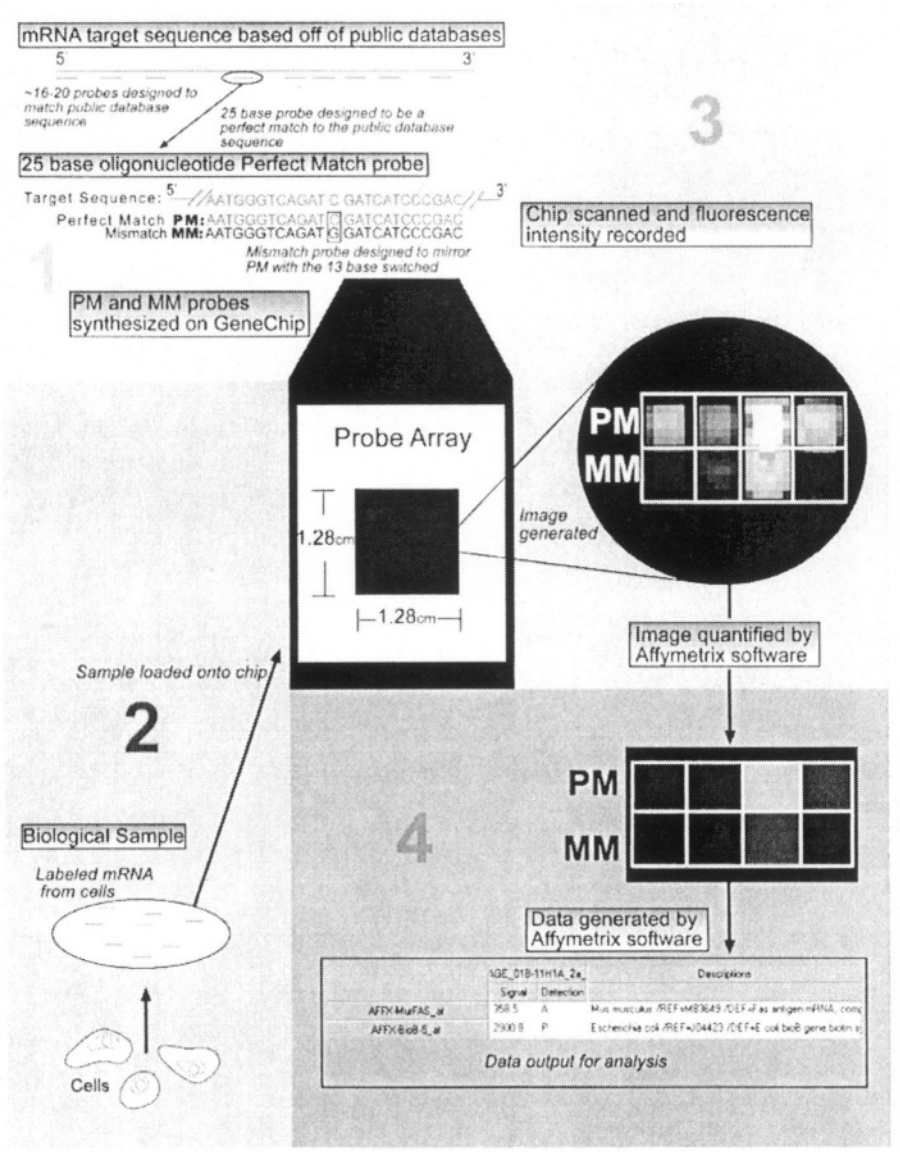


Figure 1. Overview of microarray expression analysis on the Affymetrix platform - cDNA arrays follow a slightly different procedure, but the overall process is similar. Step 1: identify a set of genes to be probed and create an array (performed by chip manufacturers). Step 2: extract mRNA from cells. Step 3: scan the array, generating a quantitative image for perfect match (PM) and mismatch (MM) probes. Step 4: convert the hybridization intensity values to quantitative expression levels.

[Figure taken from http://www.ohsu.edu/gmsr/amc/AMC_Technology.shtml, copyright 2001 Edwin Quick, CI BBSR/GMSR BBC at OHSU]

There are two dominant types of DNA arrays (often called 'microarrays') that have been used for most global gene expression measurements. The first are high-density oligonucleotide arrays that are synthesized *in situ* on a glass surface using light-directed combinatorial synthesis (commercially available from Affymetrix) [Fodor *et al.*, 1991]. These oligonucleotide arrays can contain more than 500,000 probes, typically 25-mers (25 bases long), in approximately 20 x 20 micron features in a total area smaller than one half-inch square. The arrays are designed and synthesized on the basis of sequence information alone, and it is possible to cover tens of thousands of genes and ESTs on a single array [Lipshutz *et al.*, 1999]. The other main array type is made by spotting cDNAs (or, alternatively, pre-synthesized oligonucleotides) at specific locations on a glass slide. The cDNAs, usually polymerase chain reaction (PCR) products that are 500 to 1,000 bases in length, are spaced about 100 to 300 microns apart, allowing for more than 10,000 spots to be placed on a standard glass microscope slide.

Before using a microarray, samples are often amplified and then labeled with fluorescent dyes. The samples are then hybridized to the microarray, and they bind to complimentary probes affixed to the microarray surface. The arrays are then scanned, producing a fluorescent image. The fluorescent intensity at any particular probe location indicates the relative concentration of the complimentary DNA sequence in the sample.

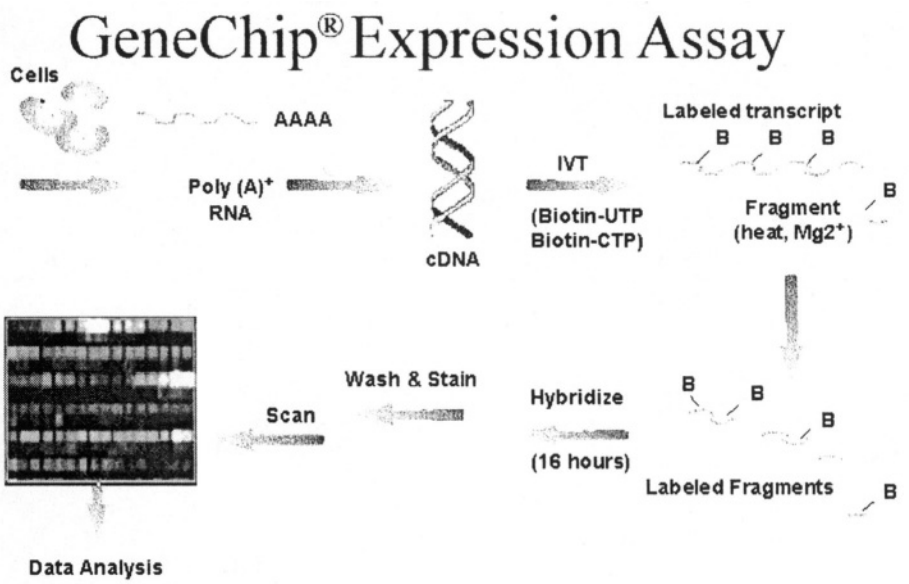
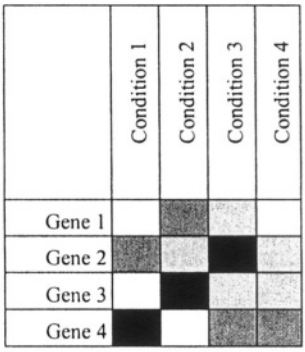


Figure 2. Overview of gene expression measurements with the Affymetrix platform. The process begins with mRNA samples from cells which are labeled with a fluorescent dye. Messenger RNA expression levels are determined using the quantitative fluorescent images. [Figure is Copyrighted by and provided courtesy of Affymetrix Inc.]

3. MICROARRAY DATA

Microarray data analysis begins with the final scanned image of fluorescent intensities as shown in Figure 2. Each feature or spot (probe) on the array is identified by a feature or spot detection and grid alignment algorithm. Following background subtraction, intensities are translated into numerical values and normalized (scaled) so that microarray experiments from different samples and different arrays can be directly compared [Brazma *et al.*, 2001]. The result is a data matrix of gene expression levels for each condition (see Figure 3). The resulting data can then be processed in a number of different ways depending on the purpose of the experiment (see section 6).



	Condition 1	Condition 2	Condition 3	Condition 4
Gene 1	1	10	5	1
Gene 2	10	5	20	5
Gene 3	1	20	5	5
Gene 4	20	1	10	10

Figure 3. Example of the conversion of hybridization intensities to relative gene expression levels. Darker squares represent higher gene expression levels, as seen in the numerical translations ranging, in this example, from 0-20.

4. MICROARRAY EXPERIMENT GOALS

Genomics and gene expression experiments can be used to identify new genes involved in a pathway, potential drug targets or expression markers that can then be used in a predictive or diagnostic fashion. Because the arrays can be designed and made on the basis of only partial sequence information, it is possible to include genes on an array that are completely uncharacterised. In many ways, the spirit of this approach is akin to that of classical genetics in which mutations are made broadly and at random (not only in specific genes), and screens or selections are set up to discover mutants with an interesting phenotype.

Such broad discovery experiments are probably best described as 'question-driven' rather than hypothesis-driven in the conventional sense. But that is not to diminish their value for understanding biological processes and even for understanding and treating human disease. For example, by analyzing multiple samples obtained from individuals with and without acute leukemia or diffuse large B-cell lymphoma, gene expression (mRNA) markers were discovered that could be used in the classification of these cancers [Golub *et al*., 1999; Alizadeh *et al*., 2000; Caldas and Aparico, 2002; Van'T Veer, 2002]. The importance of monitoring a large number of genes was clearly illustrated in these studies. Golub *et al.* found that reliable predictions could not be made based on any single gene, but that predictions based on the expression levels of 50 genes (selected from the more than 6,000 monitored on the arrays) were highly accurate [Golub *et al*., 1999]. The results of the Golub and Alizadeh studies indicated that measurements with more individuals and more genes will be needed to identify robust

expression markers that are predictive of clinical outcome. But even with the limited initial data, it was possible to help clarify an unusual case (classic leukemia presentation but atypical morphology) and to use this information to guide the patient's clinical care. The Golub and Alizadeh studies led to the van't Veer study, which successfully predicted cancer disease outcome with higher confidence than traditional procedures by using microarray analysis. More discussion of various data analysis strategies applied to the Golub *et al.* data set can be found in the CAMDA'00 proceedings [Lin and Johnson, 2000].

The use of genomics tools such as arrays does not, of course, preclude hypothesis-driven research. For fully sequenced organisms, arrays containing probes for every annotated gene in the genome have been produced [DeRisi *et al.*, 1997; Wodicka *et al.*, 1997]. With these one can ask, for example, whether a transcription factor has a global role in transcription (affecting all genes) or a specific role (affecting only some). Holstege *et al.* used this type of application in genome-wide expression analysis in yeast to functionally dissect the machinery of transcription initiation [Holstege *et al.*, 1998]. Similarly, genes located near the ends of chromosomes in yeast (as well as genes at the mating-type locus) are known to be transcriptionally 'silent'. Full-genome arrays allowed the chromosomal landscape of silencing to be mapped, and make it possible to test whether what is true for a handful of well-studied genes near the telomeres is true for all telomeric genes, and whether any centromere-proximal genes are also transcriptionally silenced [Wyrick *et al.*, 1999].

An often overlooked aspect of global measurement of gene expression is that the sequence or even the origin of the arrayed probe does not need to be known to make interesting observations – the complex profiles, consisting of thousands of individual observations, can serve as transcriptional 'fingerprints'. The fingerprints can be used for classification purposes or as tests for relatedness, in a similar manner to the way in which DNA fingerprints are used in paternity testing. In one example, transcriptional fingerprints have been used to determine the target of a drug [Marton *et al.*, 1998]. The basic idea is that if a drug interacts with and inactivates a specific cellular protein, the phenotype of the drug-treated cell should be very similar to the phenotype of a cell in which the gene encoding the protein has been genetically inactivated, usually through mutation. Thus, by comparing the expression profile of a drug-treated cell to the profiles of cells in which single genes have been individually inactivated, specific mutants can be matched to specific drugs, and therefore, targets to drugs. For instance, fingerprints were used to identify a drug 'mechanism' by utilizing the Rosetta data set [Hughes *et al.*, 2000]. Similarly, profiles have been used in the classification of cancers and the classification schemes did not depend

on any specific information about the genes involved [Golub *et al*., 1999; Alizadeh *et al*., 2000], although that information can be used to draw further biological and mechanistic conclusions. Finally, expression profiles can be used to classify drugs and their mode of action [Ulrich and Friend, 2002]. For example the functional similarity and specificity of different purine analogues have been determined by comparing the genome-wide effects on treated yeast, murine and human cells [Gray *et al*., 1998; Rosania *et al*., 2000]. Ulrich and Friend [2002] also discuss the benefits and challenges of using "toxicogenomics" as a tool to help identify drug effects earlier in the drug-discovery process.

5. MICROARRAY EXPERIMENTAL DESIGN

One of the final steps when conducting an experiment is validation of experimental results. The most basic method of validation for microarray results is independent replication of experiments. Thus, one of the first steps in experimental design is to determine the number of replicates needed to obtain meaningful results. More measurements make it possible to detect patterns and relationships that would not have been obvious or have sufficient statistical significance with smaller data sets. But, while it is preferable to have as many replicates as possible, the high cost of chips makes it more practical to use a smaller number of replicates. To determine the number of replicates, biologists must balance confidence, cost and efficiency against the desire to explore more experimental conditions. Setting limits on the number of different experimental conditions to be observed then becomes part of the experimental design process.

On the basis of our experience and that of others, we recommend in almost all cases, that experiments be conducted at least in duplicate. Using as many as four microarray replicates for each experiment is preferable, especially when using spotted cDNA arrays, although experimental conditions can dictate otherwise [Schulze and Downward, 2001]. This is consistent with the statistical estimation of the minimum number of replicates required by Li *et al.* [Chapter 12 in this volume].

Regardless of the replication strategy, sample-to-sample variability cannot, in general, be completely minimized. Thus, it is important to consider the best ways to maximize consistency between samples. In many types of studies, it is not possible to control completely all variables, and there may be considerable variability due to experimental difficulties (for example, tissue inhomogeneity or variations in sample procedures) or individual genetic variation (for example, different patients or different tumours). But such factors do not preclude the discovery of some genes that

are consistently different and that clearly 'cluster' or differentiate between the sample sets. For example, meaningful results can be extracted from the analysis of human tissue collected at different hospitals, by different surgeons and at different times. An essential requirement in these types of studies is that a sufficient number of experiments be performed across multiple individuals and multiple tissue or tumour samples to account for individual variation and possible tissue inhomogeneity.

Preparation of samples for replicates should be done as independently as possible (for example, different mice or independent dissections of a region, independent sample preparations and independent hybridizations to physically different arrays). It is not sufficient to merely remake samples from the same extracted RNA from the same mouse or tissue sample, or to simply re-hybridize samples to other arrays, as has been done in some studies. Furthermore, if genetically identical inbred animals are not used, then it is necessary to do more experiments or to pool samples from multiple animals to effectively average out differences due to genetic inhomogeneity. The same considerations apply when using human tissue or samples from any genetically inhomogeneous source.

6. MICROARRAY DATA ANALYSIS

Microarray data analysis is central to successful experimentation, but it is a large and complex topic. There are many ways to analyze, categorize or divide the data [Quackenbush, 2001], including several widely used clustering analysis techniques. Such methods are useful for finding genes that are activated together (or that interact with each other) in particular pathways. Other algorithms, such as Principle Component Analysis (PCA), have been applied to reduce the dimensionality of the data in order to discover genes or conditions that contribute to variability. Freely distributed software, such as Eisen's Cluster and TreeView, can be used to process and visualize the data. Commercially available products such as Genespring and Partek offer additional tools.

Microarray data analysis is an area of ongoing research, and the primary subject of the CAMDA conference. The rest of this volume and the previous volume in the series address a variety of data analysis issues. We briefly summarize a few main points in the following sections, but for a more in-depth review of data analysis methods, readers should refer to Jagota [2001] and Quackenbush [2001].

7. RESULT VALIDATION

On the basis of our experience and that of others, we cannot stress strongly enough the importance of great experimental care, well-characterized and rigorous analysis, and the need for appropriate follow-up and verification when performing highly parallel expression experiments, especially when using animal or human tissue. As previously discussed, replicate experiments should be performed for validation purposes. Additional methods to validate results include both statistical and biological procedures.

7.1 Sample and Data Triage

There are several initial sample processing and quality control steps that can maximize the validity of microarray results. Quality control emphasizes care and consistency in handling animals, tissue, and cells. Sample RNA should be handled appropriately to minimize degradation, and samples should be tested for suitable quality. Sample quality checks include tests for RNA or cDNA size distributions, and measurements of the quality and amount of labeled RNA. For a more complete summary of sample triage processes, see Lockhart and Barlow, [200la].

Before performing further analysis of data produced from microarray experiments, a data "triage" step should be completed to determine if the data is of sufficient consistency and quality. Background, noise, overall signal strength, and the percentage of genes scored as "present" should be measured. These measures [Lockhart and Barlow, 200la] should be appropriately assessed and reviewed before data analysis begins in order to help ensure that false positive rates are low and quantitative values are reliable.

7.2 Statistical Validation

In some cases, data analysis algorithms make assumptions about the structure of the data that can produce variations and uncertainties in results. Experimental noise also contributes to variation and uncertainty. Fortunately, many of these effects, which are not always obvious, can be addressed statistically. For example, statistical resampling [Levine and Domany, 2001; Zhang and Zhou, 2000] has been used to assess the reliability of clustering in unsupervised learning. Here, subsets of the original data are randomly selected, and the data-analysis algorithm is applied to each subset individually. Supervised learning models make use of cross-validation techniques [Dubitzky *et al.*, 2000] and have been used

extensively as well. A large portion of the data is used as a 'training set' from which a model is built via a data-analysis algorithm. This model is then tested with the remaining data. If the model fits the remaining data, then it is considered a valid model. A review of these and other techniques, including cluster analyses, neural network, and Bayesian network analyses can be found in Chapter 2.

7.3 Biological Validation

In additional to statistical validation, at least some fraction of the genes observed to be differentially expressed should be confirmed with independent methods on independent samples (not the same RNA that was used for the array experiments), especially if subtle expression differences are to be interpreted. For example, northern blots or quantitative RT-PCR experiments are used to check particularly interesting findings, and to confirm a result that might be the basis for follow-up experiments, such as the creation of a knockout mouse. The use of western blots to measure corresponding protein levels, and immunohistochemistry and *in situ* hybridization to measure cell or region specificity of proteins and mRNAs is also highly recommended.

Although array-based expression measurements can be made quantitative and reproducible, specific genes that are found to be differentially expressed on arrays should be viewed as high probability candidates but not as completely confirmed. Global expression measurements should be considered a starting point for the understanding of a biological problem, and as a valuable tool for obtaining information concerning a large number of genes. They should be used in the context of other types of measurements, knowledge and information, and it should be understood that findings will need to be followed up with further experiments of various, more conventional types.

8. CONCLUSION

In summary, the goal of genomics is to understand biology, and DNA microarrays provide a versatile and powerful way to monitor the functional expression of tens of thousands of genes at a time. But which data analysis methods are most appropriate and most useful is still open to question. Despite their impressive and rapidly growing resume, microarray technologies are still in their infancy, with plenty of room for technical improvements, further development, and more widespread acceptance and accessibility. New experimental methods, along with sequence information,

computational tools, and integrated knowledge databases are now coupled with traditional basic science approaches to the study of biology and medicine. The combination of new methods with more traditional techniques will help us understand the function and regulation of all genes and proteins, decipher the underlying workings of the cell and hopefully lead to new ways to intervene with or prevent aberrant cellular processes in order to improve human health and well-being.

Note: some text from this article is reprinted with permission from *Nature* [Lockhart and Winzeler, 2000] and *Nature Reviews Neuroscience* [Lockhart and Barlow, 200lb].

REFERENCES

Alizadeh, AA, *et al.* Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature* 403 (2000): 503-510.

Brazma, A, Parkinson, H, Schlitt, T, Shojatalab, M. A quick introduction to elements of biology - cells, molecules, genes, functional genomics, microarrays. EMBL.

Caldas, C, Aparico, SAJ Cancer: The Molecular Outlook. *Nature* 415 (2002): 484-485.

DeRisi, JL, Iyer, VR, Brown, PO. Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science* 278 (1997): 680-686.

Dubitzky, W, Granzow, M, Berrar, D. Data Mining and Machine Learning Methods for Microarray Analysis. In: Lin SM, Johnson, KJ., eds. *Methods of Microarray Data Analysis: Papers from CAMDA 2000.* Norwell, MA: Kluwer Academic Publishers, (2001): 5-22.

Fodor, SPA, *et al.* Light directed, spatially addressable parallel chemical synthesis. *Science* 251 (1991): 767-773.

Golub, TR, *et al.* Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science* 286 (1999): 531-537.

Gray, NS, *et al.* Exploiting chemical libraries, structure, and genomics in the search for kinase inhibitors. *Science* 281 (1998): 533-538.

Holstege, FC *et al.* Dissecting the regulatory circuitry of a eukaryotic genome. *Cell* 95 (1998): 717-728.

Hughes, TR *et al.* Functional discovery via a compendium of expression profiles. *Cell* 102 (2000): 109-126.

Jagota, A. *Microarray Data Analysis and Visualization.* Bioinformatics by the Bay Press. Haywood, CA: 2001.

Levine, E, Domany, E. Resampling method for unsupervised estimation of cluster validity. *Neural Comput.* 13 (2001): 2573-2593.

Lin, SM, Johnson, KF (eds). *Methods of Microarray Data Analysis: Papers from CAMDA 2000.* Norwell, MA: Kluwer Academic Publishers, 2001.

Lipshutz, RJ, Fodor, SP, Gingeras, TR, Lockhart, DJ. High density synthetic oligonucleotide arrays. *Nature Genet.* 21 (1999): 20-24.

http://industry.ebi.ac.uk/%7Ebrazma/Biointro/biology.html (2001).

Lockhart, DJ, Barlow, C. *DNA Arrays and Gene Expression Analysis in the Brain.* Edited by H Chin and SO Moldin. Methods in Genomic Neuroscience. CRC Press, 2001.

Lockhart, DJ, Barlow, C. Expressing what's on your mind: DNA arrays and the brain. *Nature Reviews Neuroscience* 2 (2001): 63-68.

Lockhart, DJ, Winzeler, EA. Genomics, gene expression and DNA array. *Nature* 405 (2000): 827-836.

Lockhart, DJ, *et al.* Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nature Biotechnol.* 14 (1996): 1675-1680.

Marton, MJ, *et al.* Drug target validation and identification of secondary drug target effects using DNA microarrays. *Nature Med.* 4 (1998): 1293-1301.

Quackenbush, J. Computational Analysis of Microarray Data. *Nature Reviews Genetics* 2 (2001): 418-427.

Rosania, GR, *et al.* Myoseverin: a microtubule binding molecule with novel cellular effects. *Nature Biotechnol.* 18 (2000): 304-308.

Schena, M, Shalon, D, Davis, RW, Brown, PO. Quantitative monitoring of gene expression patterns with a complimentary DNA microarray. *Science* 270 (1995): 467-470.

Schulze, A, Downward, J. Navigating gene expression using microarrays - a technology review. *Nat. Cell Bio.* 3 (2001).

Ulrich, R, Friend, SH. Toxicogenomics and drug discovery: will new technologies help us produce better drugs? *Nature Reviews Drug Discovery* 1 (2002): 84-88.

van't Veer, LJ, *et al.* Gene Expression Profiling Predicts Clinical Outcome of Breast Cancer. *Nature* 415 (2002): 530-536.

Wodicka, L, Dong, H, Mittmann, M, Ho, M-H, Lockhart, DJ. Genome-wide expression monitoring in *Saccharomyces cerevisiae. Nature Biotechnol.* 15 (1997): 1359-1367.

Wyrick, JJ, *et al.* Chromosomal landscape of nucleosome-dependent gene expression and siliencing in yeast. *Nature* 402 (1999): 418-421.

Zhang, K, Zhou H. Assessing reliability of gene clusters from gene expression data. *Fund. Integr. Genomics* 1 (2000): 156-173.

This page intentionally left blank

2

EXPERIMENTAL DESIGN FOR GENE MICROARRAY EXPERIMENTS AND DIFFERENTIAL EXPRESSION ANALYSIS

G.V. Bobashev, S. Das, A. Das
RTI International

Abstract: Advances in microarray technology open new challenges in data collection, analysis and interpretation. In this review paper, we focus on issues related to obtaining trustworthy normalized data and results of differential expression analysis. In particular, we briefly summarize discussions on sources of biological and technological variation, experimental design, design of arrays, normalization and error models, differential expression and multiple comparison issues. These issues remain major bottlenecks to developing standards and obtaining useful and applicable results of future analysis such as cluster analysis, network modeling, etc.

Key words: genomics, microarray, experimental design, differential expression, normalization, data quality.

1. INTRODUCTION

In this review paper on microarray data, we focus on issues related to experimental design, normalization and differential expression analysis, (i.e., finding genes that are expressed differently between groups). These emerging issues are widely discussed in the microarray scientific community. Because of recent advances in computational and statistical techniques, many scientists are focusing on the analysis of microarray data and developing models from these data. At the same time, issues of data collection, quality, and standards remain major bottlenecks to obtaining useful and applicable results. This paper briefly highlights some of these issues. The paper is structured as follows: Section 2 discusses experimental

design with sources of variation and design principles. A brief mention of microarray design follows in Section 3, and an overview of technical components of normalization models is provided in Section 4. We discuss error models and multiple comparison issues in Section 5, and final remarks are in Section 6. Statistical methodology for normalized data is beyond the scope of this discussion and will be addressed in future work.

2. DESIGN OF MICROARRAY EXPERIMENTS

The goal of a microarray experiment is to measure and compare the relative expression levels of thousands of genes simultaneously among samples. Typically, these samples compare different stages of the cell cycle, cell types, healthy and diseased cells, or different treatments. It is important to understand and account for many sources of noise and variation. These sources of variation are largely responsible for problems such as measurement error, confounding, elevated false positive and false negative rates, and biased association. By understanding these sources of variation, an experiment can be designed in a manner that takes them into account.

Moreover, the analysis models and methods must be compatible with the experimental design to avoid artificial effects. For decades, statisticians have dealt with experimental design and analysis issues to separate the effects of interest from confounding, bias, measurement and random error [Fisher, 1951]. Replication is needed to provide statistical inference. However, due to the expense of microarray experiments, decisions about what, how and when to replicate must balance the additional cost against the level of better inference that is likely to be gained from replication. The answer to these questions greatly depends on sources of variation in the experiment.

With respect to microarray experiments it is important to differentiate between “technological” and “biological” variation. Technological variation could occur because of imprecise measurements of mRNA content, variation among microarray batches, imperfection of laboratory protocols and equipment, etc. Theoretically, with the development of technology and automation the technological variation could be virtually eliminated. Biological variation i.e., variation in gene expression among genes, cells, cell lines, animals, etc., would remain present even if the amount of mRNA could be measured precisely. These two components – technological and biological - contribute to the total variance in the experiment, and therefore researchers must account for both of them at the design and analysis stages.

In the remainder of this section we focus on the identification of different sources of variation.

2.1 Biological variation

2.1.1 Variations among subjects

Each microarray experiment begins with the selection of a sample. The sample should be representative of its population, so that valid inferences about this population can be drawn. For example, there are many different strains of lab mice, which may yield different results under the same sets of experimental conditions. Experimental results based on one strain of mice may not be generalizable across strains. Among-animals variation may also be confounded by within-gene variation. Because some of the genes may have a relatively short time course of activity, variation in the timing of sample preparation among several animals could alter gene expression, resulting in two otherwise identical animals producing different expression profiles. The authors could find little published literature on this issue, but this is clearly a substantial source of variation.

2.1.2 Variations among genes

A substantial body of literature concerning gene-specific modeling supports the concept that each gene has its own expression level in a tissue (in addition to variation among individuals and tissues), as well as its own expression time course [Kerr *et al*., 2001 a-d; Yang *et al*., 2001b; Hughes *et al.* 2000, Wolfinger *et al*., 2001]. More research is needed to understand fully whether gene expression from many genes on the array should be pooled together for a higher statistical power, or whether each individual gene should be treated separately [Dudoit *et al*., 2000].

What do we know about the consistency of same gene hybridization rates? How much variance is likely for one gene among multiple hybridizations? Do different genes have different hybridization rates? The answers to these questions are platform-dependent and little appears to have been published on these topics, perhaps because the underlying microarray experiments are expensive to perform.

Bartosiewicz *et al.* [2000] estimated variability among spots, slides and animals for mouse liver tissues. The largest variability occurred among animals with a coefficient of variation (standard deviation divided by the mean) ranging from 0.18 to 0.6, while among-slides and among-spots variability ranged from 0.08 to 0.18 [Wu, 2000]. In recent presentations, Elashoff [2001] suggests that each gene has its own variation which might be independent of the mean. This argues against using mean expression as a

direct predictor of the gene-specific variance. We discuss this in greater detail in the Differential Expression section.

2.2 Technological variations

The major steps in performing a microarray experiment are (1) creating or selecting an array, (2) extracting, amplifying, and reverse transcribing mRNA into cDNA, (3) fluorescent (or radioactive) labeling of cDNAs, (4) hybridization to a DNA microarray, (5) scanning the hybridized array, and (6) interpreting the scanned image. Below we discuss the potential technological pitfalls at each step that can lead to variation.

Step 1-Creating or selecting an array: Construction of an array involves identifying DNA fragments corresponding to various partial or whole genes (DNAs, gene fragments, alternate splicings, or oligonucleotides). Solutions containing the same concentrations of fragments are prepared. These solutions are used to spot the fragments onto an array. Variation is possible throughout this step. For example, the spotting process is sensitive to printer type, pin type, length of the printing cycle, and quality of maintenance of the instruments. Arrays are commonly printed in batches, so similar "production defects" may be shared by all arrays in the batch, in addition to slide specific factors. For printing purposes, slides are commonly divided into quadrants, so the same pin prints all spots in a quadrant. This may introduce quadrant effects (i.e., one pin behaving differently from the others) and spatial effects within the quadrant (i.e., the efficiency of a pin at a specific position) [Craig *et al.,* 2001]. However, array printing is generally performed at specialized sites, and efforts are made to minimize all sources of variability. Steps 2 through 6 usually contribute more to variability than step 1, especially when performed by inexperienced technicians.

Step 2-Extracting, amplifying, and reverse transcribing mRNA into cDNA: The transcription level of a gene is equal to the amount of its corresponding mRNA present in the cytoplasm. When isolating mRNA, it is important to use identical extraction methods, to use the minimum number of processing steps, to measure the amount of mRNA and to standardize concentration. One potential source of variability is the purity of mRNA among prepared tissue samples. Occasionally, the quantity of cells available does not yield sufficient mRNA to conduct the experiment and a pre-amplification process must be conducted, which may skew the relative abundance of different mRNA species (typically, some low-abundance mRNAs may not be amplified). Information concerning the details of RNA extraction must be provided so that its impact on array sensitivity can be estimated.

Step 3-Fluorescent (or radioactive) labeling of cDNAs: The success of this step can be quite variable. cDNAs that hybridize to the microarray probes must be labeled in order to be detected. The number of fluorescent dye molecules labeling each cDNA depends on its length and possibly its sequence composition, both of which are often unknown. Because of this, fluorescence intensities for different cDNAs cannot be quantitatively compared. However, identical cDNAs from two samples are comparable as long as the same number of label molecules have been incorporated into the target sequence in each sample. To equalize the total concentrations of the two cDNA samples before applying them to an array, the solutions are diluted to have the same overall fluorescence intensity. This procedure assumes the total amount of mRNA is identical in each source population. This assumption is difficult to check [Bier *et al*., 2001]. For two-dye arrays, it is important to ensure the same calibration is consistently used.

Step 4-Hybridization to a DNA microarray: In any hybridization experiment, the time required for complete hybridization is proportional to the concentration of the applied mRNA sample. Highly abundant transcripts will rapidly hybridize to completion, so their signal should be approximately constant regardless of how long hybridization is performed. For moderately abundant transcripts, it takes longer for complete hybridization, so the amount of transcript for such mRNA will be underestimated unless a sufficiently long hybridization time is used. Finally, for rare abundance transcripts, the hybridization will still be in the linear phase of the curve after a relatively long time, and thus the concentration of these rare species is likely to be greatly underestimated. Other potential causes of variation at this stage include the temperature at which hybridization is performed, and the effect of the different buffers used to prevent nonspecific binding [Bier *et al*., 2001].

Step 5-Scanning the hybridized array: The hybridized array is scanned to determine how much of each cDNA sample is bound to each spot on the array. Within a certain intensity range, the amount of signal detected is linearly proportional to the time of exposure. Usually, signal is accumulated by a reading device (e.g., CCD camera, Phosphor imager). Data are saved as a TIFF image, where the intensity of a given pixel is proportional to the amount of signal coming from the spot on the filter or slide. For highly abundant transcripts, beyond a certain amount of signal there may be little increase in intensity per unit time, and the spot will be saturated in the image. Moderately expressed genes may yield signals within the linear component of the scanner's detection range. For rare transcripts, it may not be possible to expose the slide long enough to get a detectable signal. One of the major sources of variation for spotted arrays is baseline intensity difference between the dyes. One channel is often consistently more

intensive than the other channel for the same concentration of the label [Kerr *et al.,* 2001a-d]. Successive scans of the same slide can lead to inconsistent images. For some technologies such as those based on a radioactive label, successive scans are not possible.

Step 6-Interpreting the scanned image: Interpreting data from a microarray experiment can be challenging and subjective. Quantification of the intensities of each spot is subject to noise for irregular spots, the position of the spot on the array, dust on the slide, and nonspecific hybridization. It can be difficult to determine the intensity threshold between spots and background, especially when the spots fade gradually around their edges. Spot intensities can be contaminated by neighboring spots bleeding across spot-boundaries. Detection efficiency may not be uniform across the slide, leading to excessive red intensity on one side of the array and excessive green on the other. Even after overcoming detection and calibration problems, the measured intensities for each spot only represent the ratio of cDNAs in each cell population. Low levels of cDNA due to reverse transcription bias, sample loss, or an inherently rare mRNA can cause large uncertainties in these ratios. Additionally, different software analysis packages produce different background and signal measurements and different quality checks (e.g., depending on the segmentation technique used to define the spots).

2.3 Microarray quality checklist

The statistical analysis of data assumes established protocols were used to obtain data of a consistent quality, without apparent technological errors or failures. In practice, many scientists prefer to inspect image quality before analysis and return to particular spots on arrays to confirm certain statistical findings. This is not an option for high-throughput screening, which requires a higher quality microarray system. However, for smaller capacity research laboratories this is an important step. The quality control recommendations for microarray chip production can be found in manuals such as the Clontech manual [Clontech manual, 2001] or on the web. For example, the Galbraith laboratory web page recommends a set of visual checks to detect problems such as high background before hybridization, irregular spot morphology, comet tails, streaks, high background around the margins of the cover slip, high background following hybridization, and low signal. See http://www.stressgenomics.org/stress.fls/expression/arraytech/troubleshooting/troubles_index.htm.

3. EXPERIMENTAL DESIGNS THAT INCORPORATE BIOLOGICAL AND TECHNOLOGICAL VARIATION

The previous section discussed different biological and technological sources of variation that can be present in gene microarray experiments and differential expression analysis (animal, gene type, experimental protocol, dyes, etc.). These sources of variation are not unique to microarray experiments; in fact, they are present in virtually every scientific experiment. In this section we briefly discuss some standard statistical techniques for designing experiments that can help increase the precision of experimental data. Below we will be using experimental design terminology such as "variety" which here represents factors of interest, e.g., time point, tissue type, type of treatment drug.

3.1 Block designs

The effects of variability can be reduced in the way the experiment is set up and the experimental material (i.e., the microarray) is handled. This can be done by grouping the microarrays so that gene expressions of one variety are closely comparable with those of another, especially with respect to all extraneous sources of variation. Some known sources of variation such as experimental protocol, dyes, etc., are usually not of primary interest, but need to be controlled. Grouping of experimental units into internally homogeneous batches is known as blocking, and experimental designs that use this strategy to minimize the effects of variability are known as block designs [Cochran and Cox, 1992].

Because block designs control sources of variation by keeping extraneous experimental conditions uniform within a block, the number of experimental units in each block should normally be the same as (or, a multiple of) the number of varieties being examined (complete block designs). However, this may become untenable for a large number of varieties, and incomplete block designs can handle such situations, albeit with more statistical complexity. In this context, with the two-dye system, microarray experiments can be essentially thought of as incomplete block designs with blocks of size two [Kerr *et al.*, 2001a-d].

A relatively simple example of blocking would be an experiment to compare gene expression between two cell lines using a one-color platform in which the units will need to be taken over two days. In order to account for the effects, if any, from taking units on different days, a block design would require that half of replicates from each cell line be done on the first day and the remaining half from each line on the second day. In this way,

any effects resulting from the different dates involved are balanced between the cell lines. If one cell line treatment had been conducted entirely on one day and the other on another day, any effects from the difference in days would have biased the outcome.

3.2 Randomization

Blocking is an effective mechanism for reducing variability from known sources by apportioning it equally among the different varieties of scientific interest. However, in real-life experiments it is often difficult to account for all factors that can contribute to experimental variability. Randomly assigning experimental units to one of two or more treatments, prevents such unknown or unmeasured sources of variation from introducing bias by ensuring that such factors are not differentially distributed among the different varieties of interest [Fisher, 1951]. Randomization also provides a sound objective basis for the appropriateness of certain statistical procedures and control of type 1 errors in hypothesis tests [Piantadosi, 1997].

Randomization and blocking are both necessary attributes of a well-designed experiment. When randomization is performed within blocks, the blocks are homogeneous with respect to extraneous sources of variation (both biological and technological), thus guaranteeing exactly the same distribution of blocking factors across treatments/varieties, while retaining the advantages of randomization. Thus, randomized block designs help achieve balance on both known and unknown sources of biological and technological variation.

3.3 Loop designs

The principles of randomization and blocking can substantially improve the accuracy and efficiency of microarray experiments. Figure 1 illustrates a situation where comparisons are made between more than two varieties. Instead of the standard practice of selecting one reference sample and including it in all hybridizations (Figure la), Kerr and Churchill [2001a] have proposed a loop design, where each sample is directly compared to at least two others (Figure 1b).

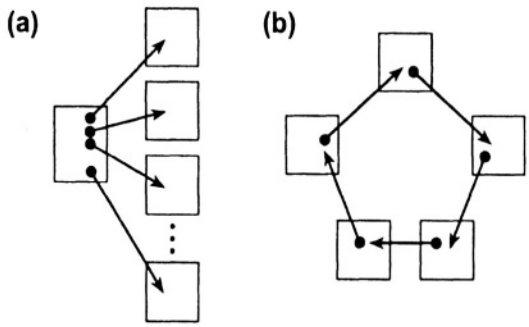


Figure 1. Experimental designs to study more than two varieties of interest (adopted from Kerr and Churchill, 2001a). Nodes represent varieties and edges represent arrays. Direction of the arrows (head or tail) represents the two dyes. Figure (a) represents the standard reference-based design, while figure (b) represents a balanced loop design. (Reproduced with the permission of the authors)

In contrast to the standard design that measures the expression level of the reference sample multiple times and the expression levels of the other study samples only once, such a balanced design obtains multiple measurements of all expression levels. Thus, the loop design obviates the need for a reference sample and doubles the amount of data on the genes of interest, without requiring additional resources. It has been shown that such a balanced design allows one to compare the varieties of interest with much greater precision than the standard design [Kerr *et al,* 2001 a-d]. Moreover, since this design is balanced with respect to dyes (because each variety is labeled once with each of the red and green dyes), any dye effects are not confounded with variety effects, so variety X gene effects are not confounded with dye X gene effects. In contrast to the standard design, any anomalous gene expression with respect to dyes will not bias the estimates of the effect of interest.

3.4 Split plot designs

In split plot designs, the experimental material (e.g., an array) can be naturally subdivided or classified into several subgroups (e.g., spots from different genes), which permits the inclusion of an extra factor in the experiment. In microarray experiments, a split plot design would be appropriate if there are two or more non-nested comparisons, (e.g., two treatments and two mouse strains). Depending on which comparison is more important, the researcher might want to split that comparison within less

important comparisons. For example, if treatment effects are more important than strain effects, each strain would be randomized for two treatments. Instead of treating array effects the same way for each fixed effect, split plot designs maximize the statistical power for one effect of interest. If that effect is the smallest one, it places the reduced significance of the larger effects in perspective (though it is usually not possible to predict effect sizes before conducting an experiment). For experiments contrasting the significance levels of a series of treatments directly, Jin *et al.* [2001] recommend complete randomization to assign each labeling reaction to specific arrays. To maximize the power for a particular contrast, they recommend keeping the contrast constant across all arrays, as long as the design assigns some replicates of each treatment to one dye and others to the second dye (called a dye swap).

3.5 Optimal designs

The search for an optimal design is a challenge in any complex experiment. This is especially true for microarray experiments, where there are strong resource constraints. More studies need to be done in this area. The works of Kerr, Churchill, Craig, and Speed, among others, have suggested some future directions. A comparison of efficiencies of several designs is done in Kerr and Churchill [2001 a,b], where the authors compare several designs and show the efficiency of the loop design as opposed to a conventional design where each array had a reference sample. The authors mention however, that even for an experiment with a relatively small (10 or more) number of arrays, the number of possible designs could be quite large and it could be computationally difficult to evaluate all of them. On the other hand, the families of efficient designs are evaluated and some of them could be found at appropriate links at Churchill's web-site http://www.jax.org/research/churchill. It is important to mention that optimality of the design strongly depends on the final objective of the study. Designs that are optimal for one goal may not be useful for another. For example, for a differential expression study, it is generally better to have more replicates for each treatment. For a time-course experiment, where gene expression patterns are the prime interest, it is more advantageous to increase the number of time points than it is to increase replicates. So far, most publications refer to designs using two-dye arrays. For Affymetrix or radioactively labeled arrays, design issues are generally similar, although effects such as dye effects are not present [Kerr *et al.,* 2001 a-d].

Having two dyes has an advantage of doubling the capacity of the array because each spot carries two pieces of information. In addition, it allows both treatment and control to be blocked on the same array. On the other

hand, the two dyes do not have exactly the same correspondence between mRNA concentration and the fluorescent intensity. This creates a need to account for such "dye effect" by incorporating a "dye swap".

4. DESIGN OF MICROARRAYS

There is little published literature addressing the issue of array design, primarily because most of these processes are proprietary to the array-producing companies. In a recent paper by Craig *et al*., [2001] normalization models were expanded to include effects that cause spot-to-spot variation. Using these expanded models, one could determine how gene replicates should be placed on the slide to avoid confounding these effects with the treatment-by-gene effect. Three factors related to slide construction were considered: 1) the effect of using different pins (on a multiple pin printing tool) to print the spots on the slide, 2) the effect of varying amounts of genetic material being retrieved by the pins on subsequent visits to the printing tool template (dip effect), and 3) the effect of washing the excess genetic material from the slide. A pilot study was conducted to assess the amount of systematic variability caused by these three factors. This study used one dye (Cy3), one treatment, and one gene at all 256 spots on each of 3 microarrays.

The results of the experiment revealed a nonrandom distribution of dye intensities among spots on each slide after hybridization and washing. One source of this variability appeared to be the slide orientation in the centrifuge, suggesting that labeled genetic material was unevenly washed off. Within the slides oriented in the same direction, a large percentage of the variability among spots was explained by slide and pin effects. The dip effect appeared to be small.

To separate printing and washing effects from the treatment-by-gene effects, gene replicates are needed, and must be placed on the slide in such a way as to avoid confounding. Craig *et al.* [2001] discouraged the use of one type of replication within the array when the same set of pins is dipped several times into the same set of wells. Because the same pin is associated with each gene, this would place replicates in the same region of the slide and potentially confound the pin effect with the washing and printing effects. Instead, replicates should be included in the template, and arranged so that replicate spots are spread throughout the slide. While theoretically this could be done on a well-by-well basis, this is not practical from an experimental viewpoint, since a multi-tip pipette is typically used to fill several template wells at once.

5. NORMALIZATION MODELS

Normalization is another term for extracting information about gene expression and accounting for confounding, bias, measurement and random error. If differential expression is the final goal then normalization implicitly becomes part of the analysis model. The majority of normalization models use Analysis of Variance (ANOVA) models also accounting for experimental design, estimating the amount of differential expression as well as conducting statistical testing. In other cases, the objective of the experiment is to conduct further analysis of gene expression such as cluster analysis. In this case, the output of the normalization models becomes normalized data, i.e., data that is supposed to be clean of confounding and on the same scale.

5.1 Data transformation and background removal

Prior to analysis, intensity data are usually transformed, most commonly by logarithmic transformation. There is some evidence that the logarithm of measured intensity is linearly (or sigmoidaly), related to probe concentration [Samartzidou, 2001; Kalnin, 2001]. Thus, the logarithmic transformation might be justified not only by mathematical convenience, but also by the physical properties of scanned intensities.

Background removal is often viewed as a statistical task, and there are a variety of approaches, ranging from removal of the global background [Clontech manual, 2001] to spatial modeling of local background values. Some scientists advocate background be removed on the logarithmic scale, but we did not find any published evidence supporting that contention. On the other hand, it appears that the problem of background removal may itself be removed when scientists master microarray technology. Usually, with increased experience in using array technology, the level of background noise drops to levels small enough that its removal would not substantially influence subsequent analyses.

Note that in this discussion we have not separately examined the situation where measurements are in ratios. We have done this deliberately, since ratios can be converted back to a linear scale by using the logarithmic transformation. Suppose, for i=1,2,...,n observations, we have ratio measures in the form of $R_i \approx X1_i/X2_i$. Then, we have $\log R_i = \log X1_i - \log X2_i$, whereby we are back to a linear relationship, albeit in the log scale. Thus the same principles and methods discussed previously for ordinary (i.e., continuous) measures are also applicable to ratio measurements.

5.2 Linear vs. non-linear effects

Some of the most widespread approaches use models for the mean expression such as ANOVA, where various factors accounting for disparities in gene expression levels are introduced as either linear or non-linear terms. Usually, the effects of spot, block on the array, array replicate, treatment, etc, are introduced as linear terms [Kerr *et al.,* 2000], while possibly non-linear relationships between mRNA concentration and intensity are modeled as a locally linear relationship [Kepler *et al,* 2000; Yang *et al.,* 2001b].

5.3 Random vs. fixed effects

Various effects in the linear models can be treated as either fixed or random with the difference in implication usually determined by the scientific goal. If the estimate for a particular effect is of interest and the experiment to test this effect is repeatable, it is usually considered fixed. For example, if differential expression analysis of a specific gene under a set of defined experimental circumstances is of interest, treatment is usually considered a fixed effect, because it was not selected at random from a population of choices of treatments. If the effect is not of specific interest such as a spot on an array, it should be considered random [Wolfinger *et al.,* 2001]. Introducing random effects into the model usually increases the efficiency of the estimates.

5.4 Ordinary least squares vs. orthogonal regression

Although linear regression analysis is generally used, orthogonal regression has been proposed as a better approach. The rationale behind orthogonal regression is that both parts of the equation are treated equally, thus estimating the relationship between them; while ordinary regression explains the variation of a dependent variable as a function of the independent variables [Sapir and Churchill, 2000]. If there is no apparent reason for treating one array as dependent and another as independent, the orthogonal approach is preferable. However, if we are relating gene expression to a standard reference category, ordinary regression is more appropriate.

5.5 Means vs. medians

Some models use analysis of median variation rather than mean. This approach may be more appropriate when typical normal distribution

assumptions are violated but it is not yet widely used. Quantile regression methodology has been developed by Gould and Rogers [1994]. Amaratunga *et al.* [2001] have developed normalization methods using a smooth monotonic function based on a set of sequential adjustments on the 10th, 20th, and up to the 90th percentiles. Although the use of medians rather than means seems to be attractive because the distribution of gene expression (even log-transformed) is often skewed, the application of such models is very limited because of a lack of theoretical statistical background and software support.

5.6 Self-consistency

One way to address self-consistency is to identify a set of genes for which expression is not expected to change during the experiment. This set is usually not known *a priori* but could be estimated using an iterative procedure that identifies a set of genes with the lowest ratio of between-treatments to within-treatment variance. [Kepler *et al.,* 2000]. This step is especially useful when a large number of genes are affected by the treatment as often happens in small-capacity, custom arrays.

5.7 Flagging outliers

Statistical modeling implies data are assumed to follow a certain distribution, (e.g., the normal distribution). Outliers could present significant problems for the analysis and thus bias the results. Detection and removal of outliers is often considered part of the normalization process. Yang *et al.* [2001a] showed that flagging weak spots improves normalization and ratio estimates in microarrays. Outliers are often defined as points that fall beyond three standard deviations from the mean, and are usually visible on a 2D plot [Houts, 2001; Samartzidou, 2001].

6. DIFFERENTIAL EXPRESSION

Analysis of differential gene expression follows directly from the normalization model. In fact, following the ANOVA framework, differential expression represents a contrast between gene-specific treatment effects. These contrasts may be tested using some version of statistical tests. There are two major problems with the statistical testing of differential gene expression. First, the use of gene-specific variance with few replicates leads

to imprecise variance estimates and low power. Second, multiple testing leads to increased rates of false positives and false negatives.

6.1 Error models

To increase the power of differential expression analysis, more precise estimates of gene-specific variance are needed. One approach is to employ error models that use information on variation in other genes within or across experiments. Older model formulations [Kerr *et al.,* 2000; Churchill and Oliver, 2001] assume a common gene variance and use a pooled estimate of variance to conduct a t-test. Others use a linear model that relates average gene expression to its variance [Kepler *et al.,* 2000]. Finally, a local regression approach can identify gene-specific variances assuming a smooth, but non-linear relationship between mean gene expression and gene variance [Kepler *et al.,* 2000; Yang *et al,* 2001b]. All these approaches lead to more refined estimates of gene-specific variance than the approaches based on a common pooled gene variance. Models that use some form of pooled variance often use a z-test rather than a t-test, reasoning that the variance pooled over such a large number of genes well approximates the "true" variance. The power of the z-test is much higher than that of the t-test, especially if the number of replicates is smaller than six [Casella and Burger, 1990].

Wolfinger *et al.* [2001] and Yang *et al.* [2001b] have adopted a more biologically focused approach, where the model is split into two components. In the first component, an array-level model uses pooled information about all genes; in the second component, the residuals from the first model were used to fit a gene-specific model that assumes variance at the individual gene level.

Thomas *et al.* [2001] have used more robust Wilcoxon rank tests . These tests have an advantage in that they do not assume any *a priori* distributions.

A principally different approach has been taken by a group from GeneLogic [Elashoff, 2001] where a database of "normal" gene-specific expression levels and variances has been developed. This database of normalized expression values is used as a reference for gene-specific variances in their normalization model. This approach has the advantage of being able to directly estimate the "true" variance, but it requires enormous resources to develop such a database.

6.2 Bayesian approach

A Bayesian approach can provide a combination of error models with individual gene models. Baldi and Long [2001] used a normalized

distribution of pooled variances as a prior and calculated posterior distributions for individual genes. Through a simulation study they have shown that the Bayesian test is much more powerful compared to the frequentist t-test, especially when the number of replicates is quite small.

6.3 Adjustment for multiple comparisons and power considerations

There is no uniform approach on how to appropriately deal with the large number of differential expression tests conducted in one microarray experiment. Depending on the purpose of the analysis, some scientists suggest a Bonferroni/Sidak-type adjustment [Thomas *et al.,* 2001; Kerr *et al.,* 2001 a-c], while Westfall and Young [1993] recommend a resampling algorithm.

Another less conservative approach is taken for example by Dudoit *et al.* [2000], and Storey *et al.* [2001 where False Discovery Rates (FDR) were applied to the problems of multiple comparison in microarrays. First introduced by Benjamini and Hochberg in 1995, the false discovery rate approach allows control of desired specificity based on the distribution of the p-values. The number of arrays needed for an experiment depends on statistical power calculations. When conducting power analysis, one needs to set up the false positive and false negative rates (or sensitivity and specificity), consider an approach for multiple testing, and obtain information on within-gene variation as well as other sources of variation. Such discussions and simulation results could be found for example in Wolfinger *et al.* [2001] and Zien *et al.* [2001].

7. FINAL REMARKS

We have presented approaches to experimental design, normalization, and differential expression analysis for microarray experiments. Different microarray platforms have their own specific features, and hopefully the cost of arrays will decrease and firm technological standards will be developed. These standards are needed for comparisons across platforms and experiments.

One crucial facet of the design and analysis of microarray experiments is the knowledge of the magnitudes of the sources of variation throughout the entire experimental process, from array production to data analysis. Database development should incorporate information about experimental design so that the data can be merged across experiments for meta-analysis.

In upcoming work, we discuss the issues that arise with further analysis of normalized microarray data, pattern/structure discovery, discriminant analysis and prediction, and modeling genetic mechanisms. The statistical challenges facing these analyses relate to the fact that microarray data usually involve relatively few analysis units, but a large number of measurements per unit.

ACKNOWLEDGEMENTS

The authors thank Dr. Amy Licata for valuable discussions.

REFERENCES

Amaratunga, D, Cabrera, J. A Resistant Walk through the Microarray Data Minefield. Presentation at Microarray Data Analysis Using Statistics and Standards to Navigate the Microarray Minefield, *http://www.healthtech.com/200I/mda/*(2001).

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

Bartosiewicz, M, Trounstine, M, Barker, D, Johnston, R, Buckpitt, A. Development of a toxicological gene array and quantitative assessment of this technology. *Arch Biochem Biophys* 376 (2000): 66–73.

Bassett, DE Jr., Eisen, MB, Boguski, MS. Gene Expression Informatics--It's All in Your Mine. *Nature Genetics* 21 (supplement) (1999): 51-55.

Benjamini, Y, Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society* Series B. 57 (1995): 289-300.

Bier, FF, Kleinjung, F, Fresenius, J. Feature-size limitations of microarray technology-a critical review. *Anal Chem* 371(2) (Sep 2001): 151-6.

Brazma, A, Vilo, J. Gene expression data analysis. *Microbes Infect* 3(10) (Aug 2001): 823-9.

Brown, CS, Goodwin, PC, Sorger, PK. Image metrics in the statistical analysis of DNA microarray data. *Proc Natl Acad Sci* U S A 98(16) (Jul 31 2001): 8944-9.

Casella, G, Berger, RL. *Statistical inference.* Belmont, CA: Wadsworth Publishing Company, 1990.

Churchill, GA, Oliver, B. Sex, flies and microarrays. *Nature Genetics* 29(4) (Dec 2001): 355-6.

Clontech manuals. *http://www.clontech.com/techinfo/manuals/index.shtml* (2001).

Cochran, WG, Cox, GM. *Experimental Designs.* New York: Wiley, 1992.

Craig, BA, Vitek, O, Black, MA, Tanurdzik, M, Doerge, RW. *Proceedings of the 2001 Kansas State University Conference on Applied Statistics in Agriculture.* 2001.

Dudoit, S, Yang, YH, Callow, MJ, Speed, TP. Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments. *http://www.stat.berkeley.edu/users/terry/zarray/TechReport/578.pdf* (2000).

Efron, B, Tibshirani, R, Storey, JD, Tusher, V. Empirical Bayes analysis of a microarray experiment. *Journal of the American Statistical Association* 96 (2001): 1151-1160.

Elashoff, M. Ensuring Good Microarray Data. Presentation at Microarray Data Analysis Using Statistics and Standards to Navigate the Microarray minefield. *http://www.healthtech.com/2001/mda/* (2001).

Fisher, RA. The Design of Experiments, 6th edition. London: Oliver and Boyd, 1951.

Galbraith laboratory web page: *http://www.stressgenomics.org/stress.fls/expression/array_tech/trouble_shooting/troubles_index.htm*

Gould, W, Rogers, WH. Quantile regression as an alternative to robust regression. *Proceedings of the Statistical Computing Section.* Alexandria, VA: American Statistical Association, 1994.

Hess, KR, Zhang, W, Baggerly, KA, Stivers, DN, Coombes, KR, Zhang, W. Microarrays: handling the deluge of data and extracting reliable information. *Trends Biotechnol* 19(11) (Nov 2001): 463-8.

Houts, T. Towards the quantitative microarray analysis pitfalls and Progress. Presentation at Microarray Data Analysis Using Statistics and Standards to Navigate the Microarray Minefield. *http://www.healthtech.com/2001/mda/*(2001).

Hughes,, TR, Marton, MJ, Jones, AR, Roberts, CJ, Stoughton, R, Armour, CD, Bennett, HA, Coffey, E, Dai, H, He, YD, Kidd, MJ, King, AM, Meyer, MR, Slade, D, Lum, PY, Stepaniants, SB, Shoemaker, DD, Gachotte, D, Chakraburtty, K, Simon, J, Bard, M, Friend, SH. Functional Discovery via a Compendium of Expression Profiles. *Cell* 102 (2000), 109-126.

Jin, W, Riley, RM, Wolfinger, RD, White, KP, Passador-Gurgel, G, Gibson, G. The contributions of sex, genotype and age to transcriptional variance in Drosophila melanogaster *Nature Genetics* 29(4) (Dec 2001): 389-95.

Kalnin, N. *Personal communication.* Clontech, 2001.

Kepler,T, Crosby, L, Morgan, KT. Normalization and analysis of DNA microarray data by self-consistency and local regression. *Nucleic Acids Research* (Submitted 2000): Santa Fe Institute preprint 00-09-055.

Kerr, MK, Churchill, GA. Experimental Design for Gene Expression Microarrays. *Biostatistics* 2(2) (2001), 183-201.

Kerr, MK, Churchill, GA. Statistical Design and the Analysis of Gene Expression Microarray Data. *Genetical Research* 77 (2001): 123-128.

Kerr, MK, Leiter, EH, Picard, L, Churchill, GA. Analysis of a designed microarray experiment. *Proceedings of the IEEE-Eurasip Nonlinear Signal and Image Processing Workshop* (June 3-6 2001).

Kerr, MK, Afshari, CA, Bennett, L, Bushel, P, Martinez, J, Walker, NJ, Churchill, GA. Statistical analysis of a gene expression microarray experiment with replication. *Statistica Sinica* (to appear 2001).

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

Koenker, R, Bassett, G. Regression Quantiles. *Econometrica* 46 (1978): 33-50.

Lee, ML, Kuo, FC, Whitmore, GA, Sklar, J. 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) (2000): 9834-9.

Long, AD, Mangalam, HJ, Chan, BY, Tolleri, L, Hatfield, GW, Baldi, P. Improved statistical inference from DNA microarray data using analysis of variance and a Bayesian statistical

framework. Analysis of global gene expression in *Escherichia coli* K12. *J Biol Chem* 276(23) (Jun 2001): 19937-44.

Mills, JC, Gordon, JI. A new approach for filtering noise from high-density oligonucleotide microarray datasets. *Nucleic Acids Res* 29(15) (Aug 2001): E72-2.

Piantadosi, S. *Clinical Trials: A Methodological Perspective,* New York: John Wiley, 1997.

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

Sapir, M, Churchill, GA. Estimating the posteriorprobability of differential gene expression from microarray data. Poster: *http://www.jax.org/research/churchill/*(2000).

Samartzidou, H. Validating Microarray Results: Using Control Reagents and Software Tools to Analyse, Standardize, and Compare Microarray Data. Presentation at Microarray Data Analysis Using Statistics and Standards to Navigate the Microarray Minefield *http://www.healthtech.com/2001/mda/* (2001).

Sen, Churchill, G. A Statistical framework for quantitative trait mapping, *Genetics* 159 (2001): 371-387.

Storey, JD, Tibshirani, R. Estimating false discovery rates under dependence, with applications to DNA microarrays. Submitted *to Journal of the American Statistical Society.* Technical Report 2001-28, Department of Statistics, Stanford University http://www-stat.stanford.edu/~jstorey/papers/dep.pdf (2001)

Thomas, JG, Olson, JM, Tapscott, SJ, Zhao, LP. An efficient and robust statistical modeling approach to discover differentially expressed genes using genomic expression profiles. *Genome Res* 11(7) (Jul 2001): 1227-36.

Tseng, GC, Oh, MK, Rohlin, L, Liao, JC, Wong, WH. Issues in cDNA microarray analysis: quality filtering, channel normalization, models of variations and assessment of gene effects. *Nucleic Acids Res* 29(12) (2001): 2549-57.

Wang, X, Ghosh, S, Guo, SW. Quantitative quality control in microarray image processing and data acquisition. *Nucleic Acids Res* 29(15) (2001): E75-5.

Westfal, P, Young, S. *Resampling-based multiple testing.* Whiley, 1993.

Wolfinger, RD, Gibson, G, Wolfinger, ED, Bennett, L, Hamadeh, H, Bushel, P, Afshari, C, Paules, RS. Assessing gene significance from cDNA microarray data via mixed models. *Journal of Computational Biology* 8(6) (2001): 625-637, http://brooks.statgen.ncsu.edu/ggibson/Pubs.htm

Wu, TD. Analysing gene expression data from DNA microarrays to identify candidate genes. *Journal of Pathology* 195(1) (Sep 2001): 53-65.

Yang, MC, Ruan, QG, Yang, JJ, Eckenrode, S, Wu, S, Mclndoe, RA, She, JX. A statistical method for flagging weak spots improves normalization and ratio estimates in microarrays. *Physiol Genomics* 7(1) (Oct 2001): 45-53.

Yang, YH, Dudoit, S, Luu, P, Speed, TP1. *Normalization for cDNA Microarray Data.* San Jose, California: SPIE BiOS, 2001.

Zien, A, Fluck, J, Zimmer, R, Lengauer, T. Microarrays: How Many Do You Need? *Proceedings, RECOMB'02,* to appear: http://cartan.gmd.de/~zien/paper/recomb02.pdf (2002).

This page intentionally left blank

3

MICROARRAY DATA PROCESSING AND ANALYSIS

Joaquín Dopazo
Bioinformatics Unit, Spanish National Cancer Center (CNIO), Melchor Fernández Almagro 3, 28029, Madrid, Spain

Abstract: DNA arrays technologies make possible the monitoring of changes in the expression patterns of thousands of genes. The analysis of such data has become a computationally-intensive task that requires technological developments at various stages, from the design of the array, to image analysis, database storage, data processing and clustering and information extraction. Here a review of the current trends in each of the various areas is provided.

Key words: gene expression, normalisation, databases, distances, clustering, gene networks, data mining.

1. INTRODUCTION

In recent years, a number of technologies in the field of genomics have developed to a level that has increased the volume of biological data available in orders of magnitude. The large amount of information coming from diverse sources including genome sequence projects and high throughput functional data is attracting the interest of biologists to the study of global mechanisms, leaving behind the old paradigm "one postdoc, one gene." In this new setting for biological research, DNA array technologies that allow for the simultaneous recording of thousands of gene expression levels in a single experiment have acquired a special role. This technology has opened new ways of looking at organisms in a genome-wide manner. Now it is possible to study complete genome patterns of gene expression in prokaryotes [Arfin *et al*., 2000] or in simple eukaryotes like yeast [Eisen *et al*., 1998] or *C. elegans* [Hill *et al*., 2000], while in higher organisms, like

humans, tens of thousands of genes related to a given system can be monitored [Zhang *et al.,* 1997].

There are two extensively used formats: cDNA microarrays [Schena *et al.,* 1995] and Affymetrix gene chips or oligonucleotide arrays [Lockhart *et al.,* 1996]. In the first format, each cDNA array consists of solid support (usually nylon or glass) where cDNA or oligonucleotides are arrayed in a fixed pattern. Fluorescent DNA derived from mRNA coming from the control and test samples is competitively hybridised to the complementary DNA probes on the array. The radioactive or fluorescence emissions of specifically bound probes are detected using an appropriate scanner, giving a quantitative estimate of each gene expression. These intensity values are supposed to be proportional to the amounts of RNA originally present in the cell. Oligonucleotide arrays, on the other hand, employ a different system to label the complex probe. These are directly synthesized on glass wafers using a photolithographic process.

A complete system for expression arrays requires the implementation and development of different experimental protocols but, even more important, is the development of database and bioinformatics tools for data collection and analysis. Computational tools for microarray data analysis are in rapid and continuous evolution and there is no clear consensus on which methods are best to cope with the complexities of such analysis [Brazma and Vilo, 2000; Quackenbush, 2001]. This review will go over the different steps necessary in the analysis of the DNA array data, and comment on the advantages and limitations of the different, most used data analysis methods.

2. DESIGN OF THE ARRAY

One of the first problems in any microarray assay is the selection of an appropriate set of genes to be used in the array. In some special cases, like prokaryotes or small eukaryotes such as yeast, an array can be prepared including the complete genome. Unfortunately, for many eukaryotic genomes the total number of genes is unknown and their intron/exon features are not well defined. In any case, the total number of genes is too large to be fitted in a single array. Therefore, a selection of a representative set of genes to be present in the array must be done. In general this selection work involves finding genes with known features: genes of known function, or having sequence or functional relationships with genes of interest, or sequence variants of genes of interest, or biological controls, etc. There are different databases and public data repositories containing information on the sequences that can be used for this selection process, such as UniGene (http://www.ncbi.nlm.nih.gov/UniGene), TIGR gene indices (TIG, in

http://www.tigr.org/tdb/tgi.shtml), DoTS (http://www.allgenes.org), etc. These resources provide high-quality annotation for the cDNA contained in them but the number of sequences is extremely high (more than 13 million sequences in the last GenBank release of year 2001). Due to the large numbers of sequences, integrated software that allows user-friendly querying, array project management, and clone tracking can be very useful [Tamames *et al*., 2002].

When using oligonucleotides as probes, the probe length and the melting point of the expected heteroduplex is of crucial importance [Hughes *et al.,* 2001]. Other factors, such as base composition and order, are important too. Several computer programs have been developed to optimise the choice of oligonucleotides [Drummond and Stamper, 1999], but in the case of cDNA arrays, problems often arise from a lack of reliability. Even in cleaned up, sequence verified, cDNA collections, error rates can be as high as 30% [Knight, 2001] due, mainly, to three different types of errors. (1) the corresponding sequence in the database is different from the one found in the cDNA clone; (2) the sequence is correct, but the annotation is wrong; and (3) the predicted orientation is wrong. Another common problem, derived from using cDNA from ESTs collections is redundancy. A widely used method for controlling redundancy is clustering of EST data [Miller *et al*., 1999]. Recently, alternative splicing has been identified as a common mechanism of the generation of variability at the level of gene products. It has been shown that at least 50% of the human genes are subjected to alternate splicing [International Genome Sequencing Consortium, 2001; Venter 2001] and this fraction is likely to be similar in other animals, including invertebrates [Brett, 2002]. This raises an obvious problem regarding the election of the probe representative of a particular gene. The use of tools such as GeneNest [Coward, 2002] to explore gene structure, including alternative splicing, based on a mapping of the EST consensus sequences to the complete human genome, can help in solving this problem.

Once genes are identified, it is important to determine the specific conditions to be studied, for example, the different time points of a biological process, or the distinct types of tissue, or drug treatments, etc., which can take different values, or classes. Therefore, considering all the variables involved in the experiment, there are four basic experimental factors to be taken into account: values (or classes), genes, dyes, and arrays. With these four factors there are $2^4 = 16$ possible experimental effects. The first step in choosing a good design is to identify which effects might possibly contribute to variation in the data. An ANOVA model can thus be considered as having global and gene-specific components [Kerr and Churchill, 2001a; Kerr *et al*., 2000]. Obviously, intra- and inter-array gene replications are necessary to have estimation on the different types of experimental errors.

The first steps of deciding on the set of genes to be placed in the array and determining the conditions to be studied are examples of the inherent complexities of the management of data in genomic methodologies. With just these factors to consider, the size and complexity of the microarray data set becomes apparent.

3. DATA ACQUISITION AND IMAGE ANALYSIS

After competitive hybridisation with the control and the query labelled DNAs, the array is scanned. Two images corresponding to the two fluorescent dyes are obtained. For cDNA arrays, these images must be analyzed to identify and quantify the spots corresponding to the probes. Usually, commercial microarray scanner manufacturers provide their own solutions for image processing. In addition there are several public domain software packages for the analysis of the images produced in DNA array experiments (see Table 1).

The processing of scanned images usually involves three tasks. (1) gridding, which is the process of assigning coordinates to each of the spots, (2) segmentation, which allows the classification of the pixels either as foreground or as background; and (3) intensity extraction, which implies calculating, for each spot on the array, red and green foreground fluorescence intensities, background intensities and, in some cases, quality measures [Yang *et al.*, 2001]. Background adjustment is necessary because measured intensities include a contribution due to the non-specific hybridization of the target to other elements in the slides (chemicals, etc). The most commonly used procedure to remove this background effect is subtracting the fluorescence intensity measured around the spots. Nevertheless, the use of means or medians around the spot tend to produce noisy measures [Yang *et al.*, 2001]. Some software packages use morphological opening, which is a way of calculating an average of the background along windows, which can be subtracted from the signal. This method performs better than the subtraction of a constant average, which can cause negative intensity values.

Some programs are very efficient in the precise localisation of the spots (e.g. QuantArray) while others provide convenient data analysis tools (e.g. GenePix and ArrayPro). ArrayVision is well-known as a powerful program for macroarray images. The quantification of macroarray (radioactive) images is often more complicated than for microarrays (fluorescent), as in the first case the spots tend to be tightly arranged, with no space between consecutive spots.

Table 1. Some available programs for image analysis. Upper part: commercial programs. Lower part: public domain programs.

Program and distributor	**Web page**
AIDA Array, Raytest GmbH	http://www.raytest.de
ArrayPro, Media Cybernetics	http://www.mediacy.com/
ArrayVision, Imaging Research Inc.	http://www.imagingresearch.com/
GenePix, Axon Instruments, Inc.	http://www.axon.com
ImaGene, BioDiscovery Inc.	http://www.biodiscovery.com/
Iconoclust, Clondiag	http://www.clondiag.com/
Iplab, Scanalytics Inc.	http://www.scanalytics.com/
Lucidea Automated Spotfinder, Amersham Bioscience	http://www.amershambiosciences.com/application /microarray/default.htm
Phoretix Array^2, Phoreticx	http://www.phoretix.com/index.htm
QuantArray, Packard BioScience	http://www.packardbiochip.com/
Spot, CSIRO	http://www.cmis.csiro.au/iap/spot.htm
Free software and author	**Web page**
ArrayViewer, NHGRI	http://www.nhgri.nih.gov/DIR/LCG/15K/HTML/ images.html
F-Scan/P-Scan, Center for Information Technology, NIH	http://abs.cit.nih.gov/fscan/ http://abs.cit.nih.gov/pscan/
Scanalyze from Michael Eisen, Lawrence Berkeley National Lab, Berkeley, California	http://rana.lbl.gov/EisenSoftware.htm
TIGR Spotfinder, TIGR	http://www.tigr.org/softlab/

4. NORMALISATION AND FILTERING

Image processing is the first step in the data analysis process. The distinct efficiencies in the labelling process and in the detection of the fluorescence in both channels, as well as differences in the initial amount of mRNA in the samples, not to mention problems derived from the manipulation of the samples, cause systematic biases in the measures. Then, the intensity values for red and green channels must be rescaled before proceeding with the analysis [Hill *et al.*, 2001]. This process is called normalisation and is based on the existence of some reference point. These references are either external, if some controls are "spiked" into the RNA before labelling, or internal, under the assumption that most of the genes (or at least a large subset of them) do not change their expression values in both cases. These controls are used to balance the biases in both channels. A common approach is to assume that the total amount of mRNA is the same

for both samples. Under this assumption, the total computed intensity in red and green channels (mean or median value) should be the same and can be used as a normalisation factor. This approach has been used both for Oligonucleotide and cDNA arrays. Nevertheless, there are more sophisticated approaches that involve specific assumptions. In many experiments it is possible to assume that a significant amount of the genes are expressing at the same level. A scatter plot of red versus green would cluster these genes along a straight line of slope 1 (for similar labelling efficiencies). Intensities are adjusted to get a slope of 1. Another approach consists of the use of "housekeeping" genes, whose expression level is supposed to be the same in closely related samples. An approximate probability density can be calculated for all the red and green ratio intensities for the housekeeping genes, and used for estimating confidence intervals to identify differentially expressed genes [Der *et al*., 1998]. So far there are few comparisons of the relative merits of the different normalisation procedures [Hill *et al.,* 2001; Li and Wong, 2001; Goryachev *et al.,* 2001]. In addition, normalised data, which are expression ratios, are usually log-transformed. The advantage of using this transformation is that the resulting data reflect the up-regulation and down-regulation values in a symmetrical scale. Ratios of down-regulated genes have their values between 1 and 0, while up-regulated ratios can have large values. For example, after log2-transformation, a change in the regulation by a factor of 2 (2 for up-regulation and 1/2=0.5 for down-regulation) would take the values of 1 and -1 for up- and down-regulation, respectively.

Whichever method is used, and whatever analysis is performed, it is judicious to filter out genes that do not change their expression level during the course of the experiment [Harrington *et al*., 2000]. The selection of informative genes is the first step in reducing the complexity of the data and so improving the signal to noise ratio.

5. DATA STORAGE

Data storage is a critical and often underestimated step. The ability to make comparisons between different experiments (usually from different laboratories) is highly dependent on the existence of a common, well-defined storage structure. In addition, the vast amount of information produced by the DNA array techniques needs an efficient relational database system for storage (see Table 2).

Table 2. *Most popular human DNA repository databases.*

Database	Web page
ArrayExpress, European Bioinformatics Institute	http://www.ebi.ac.uk/arrayexpress
ArrayDB, National Human Genome Research Institute	http://genome.nhgri.nih.gov/arraydb/
ChipDB, Whitehead, MIT	http://web.wi.mit.edu/young/chipdb
DRAGON, Kennedy Krieger Institute	http://207.123.190.10/dragon.htm
ExpressDB, Harvard	http://twod.med.harvard.edu/ExpressDB
Gene Expression Omnibus (GEO), NCBI	http://www.ncbi.nlm.nih.gov/geo
GeneX , NCGR	http://www.ncgr.org/genex
RIKEN cDNA Expression Array Database (READ), RIKEN	http://genome.gsc.riken.go.jp/READ/
RNA Abundance Database (RAD) University of Pennsylvania	http://www.cbil.upenn.edu/RAD2/query.html
Stanford Microarray Database (SMD), Stanford University	http://genome-www4.stanford.edu/MicroArray/SMD

There is currently a significant international effort to standardise the way this information is stored and managed, and the MGED (Microarray Gene Expression Database) working group (http://www.mged.org/) is perhaps, the most important example. MGED has four main objectives. First is providing a standard definition of a minimum set of data known as the MIAME (Minimum Information About a Microarray Experiment) initiative [Brazma, 2001]. Second is the development of a relational database schema for storing the data. The ArrayExpress relational database schema is available at the EBI ArrayExpress website (http://www.ebi.ac.uk/arrayexpress/), including tools for submitting and retrieving the data. Third is the establishment of a data exchange format (MAGE-ML, a specialised XML format) for microarray experiments. And fourth is the development of ontologies for microarray experiment descriptions and biological material (biomaterial) annotation.

The advantage of many systems of information storage is that they provide an interface to allow recovery of the data stored under a user-friendly system of queries. Users can easily filter the data, for example removing genes whose expression patterns do not change along the studied conditions, apply different transformations, and send data directly to analysis tools. There are different commercial PC-based packages that provide these capabilities to the same extent (depending on the price), that can fulfil the storage and analysis necessities of small groups. Nevertheless, when a large number of experiments are performed, the approach must be different. There are some public domain packages that can easily be implemented and allow the setup of a centralised resource for data storage and management. One of

these packages is the popular maxdSQL (Manchester University; http://bioinf.man.ac.uk/microarray/resources.html).

6. ADDRESSING BIOLOGICAL QUESTIONS

The possibility of determining in a simple experiment the expression level of thousands of genes opens up the possibility of obtaining answers for biological questions that, just two or three years before, we could not even dream. From the point of view of the pure methodology, we can distinguish between questions involving the comparison of two conditions (typically the condition of interest versus a reference) and the questions involving the study of many conditions (e.g. time courses, dosage series, series of patients, tissues, etc.). Comparison of two conditions involves the determination of the genes whose expression levels change significantly with respect to the reference condition. Multi-condition experiments usually provide the answer for more complexes questions, and involve more sophisticated methods for their analysis. The next two sections will describe the most commonly used methodologies for analyzing both situations.

Typically, the multi-condition experiments are represented in a matrix of gene expression values, with genes in rows and conditions in columns. In other words, each column represents a single microarray experiment. Depending on the experiment, the values of gene expression can be used to classify conditions (columns) or gene expression profiles (rows). Both cases involve a first step of clustering either to obtain sets of conditions with similar gene expression values or to obtain sets of genes with similar expression profiles along the studied conditions. Classification of different types of cancers is a typical example of the first type of experiments. The molecular signature of the different tumoral tissues has been demonstrated to be a valuable diagnostic tool (see for example [Alizadeh *et al*., 2000; Alon *et al.,* 1999; Perou *et al*., 1999; Ross *et al.,* 2000; Scherf *et al*., 2000; etc.]). The second type of experiment usually involves the study of time series or dosage series to detect which genes display highly correlated expression patterns. These genes most likely play similar roles in the cell (see, for example [Brown *et al.,* 2000; Eisen *et al*., 1998; Wen *et al.,* 1998; etc.]). In addition, it is possible to design experiments that infer networks of interactions between genes [D'haeseleer *et al.,* 2000]. It is possible, in theory, to extract information about these networks from the study of expression profile correlation and, in some cases, complement the results with external information.

7. DATA ANALYSIS

7.1 Two conditions comparison

The comparison of two independent samples (i.e. diseased versus normal tissue, etc.) is the simplest experimental situation, as previously mentioned. Although a number of statistical tests are available to assess the significance of the observed differences, most of the papers published to date use simple filtering rules that eliminate genes with less than two- or three-fold expression changes or, in general arbitrary assigned fold differences [Schena et al., 1996; Webb *et al.,* 2000]. Application of such thresholds completely fail to spot biologically important genes that have a small fold change, but which are highly significant statistically because they can be measured with high precision as a result of replication. On the other hand, many genes that have a large fold change in one array may also exhibit high variability across multiple arrays and thus possess little to no statistical significance. Thus, the appropriate determination of significance is a key issue that helps to appropriately distinguish between important biological changes and chance variation [Wittes and Friedman, 1999]. A recent work [Tanaka *et al.* 2000] illustrates the danger of false positives and false negatives when looking strictly at fold change.

Classical statistical techniques, like a standard t-test, are available to check the significance of the observed differences if two independent samples are compared. This test produces a P value that represents the probability that the difference is observed because of random chance. A very small P value will indicate that the tested gene is likely to be differentially expressed. The genes in the array can be ranked according to increasing p values and an appropriate threshold can be selected depending on the percentage of false positives chosen.

So far there are few examples where these tests have been applied [Rogge *et al.,* 2000; Glynne *et al.,* 2000], but in the majority of the cases, the strategy of choice seems to be based on ad hoc thresholding procedures, often based on arbitrary assigned fold differences [Schena *et al.,* 1996; Webb *et al.,* 2000].

7.2 Multiple conditions comparison.

7.2.1 Types of comparisons: genes or conditions

As already mentioned, the values of gene expression can be used to classify conditions (columns) or gene expression profiles (rows). In both cases a first clustering step is necessary. However, even though the goal of

clustering is the same, the characteristics of the input data are quite different. In the case of classifying conditions, a few (usually less that one hundred) column vectors with many components (all the genes in the array, that is, several thousands) need to be clustered. On the other hand, when gene expression patterns are to be classified, several thousand rows with few components need to be clustered. Individual components of the vectors have a non negligible noise component, and missing values are not infrequent. The different properties of the distinct clustering methods make them more suitable for one, both, or none of the types of comparisons, as will be discussed in the next sections.

7.2.2 Distances

The identification of genes with correlated expression across the conditions studied or, alternatively, the identification of conditions with similar expression values for all the genes is achieved through the comparison of the row or column vectors, respectively, by means of a distance function. The choice of a given metric depends very much on what properties the researcher wants to measure. There are two types of metrics extensively used in the comparison of expression profiles: Euclidean distance and Pearson's correlation coefficient.

Given two vectors, that can represent either genes (rows) or conditions (columns), with their corresponding expression patterns: $\mathbf{v_1}$ $(e_{11}, e_{12}, .. e_{1n})$ and $\mathbf{v_2}$ $(e_{21}, e_{22}, .. e_{2n})$, values for both distances are obtained as follows. *Euclidean* distance is obtained as the square root of the summation of the squares of the differences between all pairs of corresponding values:

$$d_{1,2} = \sqrt{\sum_i (e_{1i} - e_{2i})^2}$$

This metric measures the absolute distance between two points in an n-dimensional space, where n is the size of the vector. This distance considers two similar vectors whose components display similar magnitude of expression. Although this property may be useful in some cases, in the case of gene expression profiles it is biologically more interesting to search for vectors whose components may have different absolute values, but similar overall profile. Euclidean distance can be used for these purposes if the data are properly transformed (normalised in a statistical sense, that is, subtracting the mean and dividing by the variance). Nevertheless, the ideal metric for identifying profiles with similar shapes is the Pearson's correlation coefficient (r) that does not need any specific transformation of the data. This metric gives values between -1 (negative correlation) and 1

(positive correlation). The more the two profiles have the same trend; the closer to 1 is the r value, irrespective of their absolute values of expression. A very interesting property of the correlation coefficient is that it can be used to detect negatively correlated genes.

The distances described above are affected to some extent by the fact that there exists a degree of correlation between genes as well as between experimental conditions. This correlation tends to produce elliptical clusters, which can cause problems for clustering methods whose optimal performance occurs with compact, round clusters, such as k-means. There are, on the other hand, distances that can deal with datasets containing large numbers of measures with a high degree of internal correlation. Distances that take into account covariance between experiments, like Mahalanobis distance [Mahalanobis, 1936], may be useful for datasets with high internal correlation. The problems derived from the complex joint distribution of gene expression values, particularly their correlational structure and non-normality has been addressed by other authors [Hunter, 2001]. They argue that simple similarity metrics such as Euclidean distance or correlational similarity scores are suboptimal for use in this application and propose the use of Bayesian approaches. Nevertheless, they conclude that the use of more sophisticated approaches did not produce significantly better results than the euclidean or the correlation metrics.

Most of the metrics found in the literature are derived from the Euclidean distance or from the correlation coefficient.

7.2.3 Unsupervised clustering

Unsupervised clustering comprises a number of techniques that produce arrangements of the data based on a distance function. These methods do not use any external information for constructing groups of similar profiles of conditions or genes.

Despite the arsenal of methods used, the optimal way of classifying gene expression data is still open to debate. Here we will discuss the virtues and pitfalls of the most frequently used methods. The table below shows a list of the most current methods used for clustering, arranged on the basis of their properties and underlying algorithms.

Table 3. *Some of the most used clustering methods and their main properties.*

Method	Topology	Properties
PCA, SVD,	Non hierarchical	Exploratory algorithms
k-means, quality cluster	Non hierarchical	Runtimes > n^2
Average linkage	Hierarchical	Runtimes > n^2
SOM	Non hierarchical	Robust, runtimes ~n
SOTA	Hierarchical	Robust, runtimes ~n

Clustering techniques can be used in combination with other exploratory techniques, like Principal Component Analysis (PCA) [Everitt and Dunn, 1992], that help to visualise the complexity of the data in a two or three-dimensional space, and groups of genes can be visualised. Other related techniques, like Singular Value Decomposition (SVD) [Alter *et al.,* 2000] or correspondence analysis [Fellenberg *et al.,* 2000] have also been applied to cluster gene expression patterns. Nevertheless, some authors have pointed out that these techniques can produce misleading results when applied to gene expression data [Yeung and Ruzzo, 2001]

Depending on the way in which the data are clustered we can distinguish between hierarchical and non-hierarchical clustering. Hierarchical clustering allows detecting higher order relationships between clusters of profiles whereas the majority of non-hierarchical classification techniques work by allocating expression profiles to a predefined number of clusters, without any assumption on the inter-cluster relationships. Many authors prefer hierarchical clustering to the non-hierarchical alternatives due to the possibility of exploring different levels of the hierarchy. Aggregative hierarchical clustering in its different variants (average-linkage, single-linkage, complete-linkage, etc.) [Sneath & Sokal, 1973] is still one of the preferred choices for the analysis of patterns of gene expression, in part due to the availability of software either in standard statistical packages or specifically designed for gene expression data [Eisen *et al.,* 1998]. It produces a representation of the data with the shape of a binary tree, in which the most similar patterns are clustered in a hierarchy of nested subsets [Dopazo *et al.,* 2001]. Standard hierarchical clustering has been used to analyze several systems, including yeast [Eisen *et al.,* 1998; Chu *et al.,* 1998; Spellman *et al.,* 1998] and human cells [Wen *et al.,* 1998; Perou *et al.,* 1999; Voehringer *et al.,* 2000; Scherf *et al.,* 2000; Ross *et al.,* 2000; Roberts *et al.,* 2000].

As an alternative to hierarchical clustering, other non-hierarchical methods, like quality cluster [Heyer *et al.,* 1999] or k-means [Hartigan, 1975], have been used [Tavazoie *et al.,* 1999]. These algorithms start with a pre-defined number of clusters and, by iterative reallocation of cluster

members, minimise the overall within-cluster dispersion. Common criticisms of these types of algorithms focus on the fact that the number of clusters has to be fixed from the beginning of the procedure. Other authors [Ben-Dor *et al.,* 1999; Herwig *et al.,* 1999] proposed different versions of a progressive k-means procedure that find the number of different clusters from the data itself and is independent of an *a priori* specified number of clusters.

Standard hierarchical clustering works very well for clustering conditions (few items) but several authors [Tamayo *et al.,* 1999] have noted that standard clustering methods are not very robust when applied to clustering thousands of gene expression profiles. In addition, typical runtimes of standard methods based on distance matrices can range from N^2 to N^4 [Hartigan, 1975], which makes them very slow when thousands of items are to be analyzed. In an attempt to overcome these problems, some authors have proposed the use of neural networks as an alternative [Tamayo *et al.,* 1999; Törönen *et al.,* 1999; Herrero *et al.,* 2001]. A comparison of runtimes [Mateos *et al.,* 2002b (Chapter 6, this volume)] shows how neural network-based methods are clearly faster than their standard counterparts. A recent comparative study [Cummings, 2001], shows SOTA as one of the fastest methods. Unsupervised neural networks, such as Self-Organising Maps (SOM) [Kohonen, 1997] or the Self-Organising Tree Algorithm (SOTA) [Dopazo and Carazo, 1997], provide a more robust framework, appropriate for clustering large amounts of noisy data. Because of their properties, neural networks are suitable for the analysis of gene expression patterns. They can deal with real-world data sets containing noisy, ill-defined items with irrelevant variables and outliers, and whose statistical distributions do not need to be parametric.

Nevertheless, the SOM has some inherent problems. Firstly, it is a topology-preserving neural network. In other words: the number of clusters is arbitrarily fixed from the beginning, as in k-means. In addition, the training of the network (and, consequently, the definition of clusters) depends on the number of items in each cluster. Thus the clustering obtained is not proportional. If irrelevant data (e.g. invariant, "flat" profiles) or some particular type of profile is over represented, SOM will produce an output in which this type of data will populate the vast majority of clusters. As a “side effect”, the most interesting profiles tend to map in a few clusters and resolution can be poorer for them. Contrary to this, clustering obtained with SOTA is proportional to the heterogeneity of the data, instead to the number of items in each cluster. Thus, regardless of whether a given type of profile is abundant, all the similar items will remain grouped together in a single cluster and they will not directly affect to the rest of the clustering. This is because SOTA is distribution preserving while SOM is topology preserving [Dopazo and Carazo, 1997; Frizke, 1994].

To date there are few comparisons about the relative performance of different clustering methods. Recently, using the silhouette statistic [Hand, 1981] it was found that SOTA performed slightly better than average linkage, whereas SOM displayed the worst performance in terms of accuracy of classification in the proper cluster [Mateos *et al.,* 2002b (Chapter 6, this volume)].

Since the comparison operations are performed amongst the data and the average profiles in the nodes, the absence of some points (missing values) in a vector corresponding to a particular gene expression profile will have a negligible effect on the whole process of the network training. This makes unnecessary the use of methods for estimating missing values [Troyanoskaya *et al.,* 2001] required if average linkage or similar methods are used.

7.2.4 Definition of clusters: confidence intervals and statistical approaches

Statistical tests for assessing the reliability of the clusters found by the different methods have been scarcely used. Since the development of statistical models that account for clustering in the context of gene expression is still in a very preliminary phase [Baldi and Long, 2001], simulation [Kruglyak and Tang, 2001] or resampling techniques, like the bootstrap [Efron and Tibshirani, 1991], have been used [Kerr and Churchill, 2001b] as a practical alternative.

Other authors [Herrero *et al.,* 2001] used a permutation test, consisting of producing random permutations in the values of the rows (genes) to obtain an approximation of the random distribution of distances that one could expect in a data set with the same distribution of values that the original data set studied. Then, a simple one-tail test can be used to set a threshold, at the desired confidence level, for a distance value that cannot happen by chance. This test can be coupled to some clustering methods [Herrero *et al.,* 2001].

7.2.5 Supervised clustering

For most biological problems, there is some information available beforehand that can be exploited to produce clusters and to further classify new data. There are a number of methods for supervised clustering able to "learn" from this information the features defining each cluster, usually from a training set, and later use this knowledge to classify additional data. A detailed description of these methods is beyond the scope of this revision but it is worth mentioning that they have successfully been applied to cluster both genes and conditions.

Support Vector Machines (SVM) was the first example of the application of machine learning methods to classifying gene expression profiles. SVM are able to use prior information on the classes studied, in this case MIPS (http://www.mips.biochem.mpg.de/proj/yeast) functional classes [Brown *et al.,* 2000]. Later, perceptrons, a feed forward neural network, were used for the same task [Mateos *et al*., 2002a], showing some advantages over SVM, such as the possibility of classifying several classes at the same time (SVM can only separate two classes at a time).

Recent proposals suggest that supervised methods such as SVM [Furey *et al.,* 2000] or supervised neural networks [Khan *et al*., 2001] can be used to classify conditions too. Neural networks, in contrast to SVM, are able to discriminate amongst many different classes, and this is preferable for multi-class problems. Principal component analysis (PCA) was used to reduce the number of items to analyze [Khan *et al*., 2001]. One reason for this reduction is the consequent decrease in the number of parameters that the perceptron has to infer from the data, which depends on the size of the input layer (in this particular case the number of genes). Generally speaking, fewer parameters mean more generalisation power in the network. Nevertheless, the use of PCA results for clustering may not provide the best results [Yeung and Ruzzo, 2001]. Other authors have first clustered the gene profiles with SOTA, and used them for clustering conditions after defining the optimal level of information [Mateos *et al.,* 2002b (Chapter 6, this volume)].

Again there are few comparisons between supervised and unsupervised clustering methods, but in a recent example [Mateos *et al*., 2002b], supervised clustering seems to perform better for classifying conditions.

7.3 Gene networks

One of the subjects attracting more attention is the study of gene networks. The simplest approach is clustering the data and searching for regulatory control elements (e.g., promoters) in all co-expressing genes [Brazma *et al.,* 1998; Tavazoie *et al.,* 1999]. But the information provided by these approaches is limited to genes that are co-regulated, not to which gene is regulating which other gene. In network inference, the aim is to construct a model of the interactions between genes. This requires inference of the causal relationships among genes or, in other words, the reverse engineering of the network architecture from the gene expression profiles. Boolean networks [Somogyi *et al*., 1997; Liang *et al*., 1998] are utilised for reverse engineering gene networks, for example in *S. cerevisiae* [Friedman *et al*., 2000]. Depending on the connectivity (number of possible gene interactions), Boolean networks need more or less experimental points. Fully connected networks would need 2^N experimental measures for N genes,

which is a completely unrealistic number. But introducing restrictions derived from previous knowledge of the system can drastically decrease the number of points required [D'haeseleer *et al.,* 2000]. Also, systems of differential equations [Chen *et al.,* 1999] or other approaches inspired in electronic circuit theory [Arkin *et al.,* 1997] can be used for modelling simple gene regulation systems. A major challenge is to couple modelling with systematic experimental design. This is the key for the discovery of novel gene function and gene network connections trough expression profiling and computational inference.

8. CONCLUSIONS AND FUTURE PROSPECTS

The methods revised here, despite being a non-exhaustive collection, comprise the most widely used ones. Nevertheless, these methods constitute only the tip of the iceberg in terms of the magnitude of information that DNA array technologies, with the appropriate data analysis can offer.

The real purpose of genomic methodologies like DNA arrays is to convert the wealth of data produced into information and the information into knowledge [Basset *et al.,* 1999]. The methods described in the review can be considered the first generation of analysis tools, oriented to arrange the vast amounts of data in a comprehensive manner. The next step in the analysis consists of extracting the information and biological characteristics common to groups of genes of interest. Most of this information is in the biomedical literature and because of this, a considerable effort in developing automatic procedures for extracting it has been made in the last years. Different problems such as the detection of protein names [Fukuda, 1998], information related to groups of proteins [Andrade and Valencia, 1998], protein-protein interactions [Blaschke *et al.,* 1999], or building knowledge bases derived from the scientific literature [Ohta *et al.,* 1997] as well as general utilities for text retrieval [Wilbur and Coffee, 1994] have been addressed recently. Also, text mining techniques have been applied to the analysis of gene expression data [Tanabe *et al.,* 1999; Oliveros *et al.,* 2000; Jenssen *et al.,* 2001].

However, text mining (and data mining in general) is only one of the many topics in progress. The DNA array data analysis field is rapidly evolving and many new methods now under development will come to light soon. It is even possible that, in only a few years, DNA arrays will be substituted by other, more powerful, functional genomic screening techniques. Nevertheless, many of the ideas for the analysis of the data, perhaps implemented in different, more efficient algorithms, will still hold.

REFERENCES

Alizadeh, AA, Eisen, MB, Davis, RE, Ma, C, Lossos, IS, Rosenwald, A, Boldrick, JC, Sabet, H, Tran, T, Yu, X, Powell, JI, Yang, L, Marti, GE, Moore, T, Hudson, J Jr, Lu, L, Lewis, DB, Tibshirani, R, Sherlock, G, Chan, WC, Greiner, TC, Weisenburge,r DD, Armitage, JO, Warnke, R, Levy R, Wilson, W, Grever, MR, Byrd, JC, Botstein, D, Brown, PO, Staudt, LM. Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature* 403 (2000):503-511.

Alter, O, Brown, PO, Botsein, D. Singular value decomposition for genome-wide expression data processing and modelling. *Proc Natl Acad Sci USA* 97 (2000): 10101-10106

Alon, U, Barkai, N, Notterman, DA, Gish, K., Ybarra, S, Mack, D, Levine, AJ. Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed with oligonucleotide arrays. *Proc Natl Acad Sci USA* 96 (1999): 6745-6750.

Andrade, MA, Valencia, A. Automatic extraction of keywords from a scientific text: application to the knowledge domain of protein families. *Bioinformatics* 14 (1998): 600-607.

Arfin, SM, Long, AD, Ito, ET, Tolleri L, Riehle, MM, Paeglei, ES, Hatfield, GW. Global Gene Expression Profiling in *Escherichia coli* K12. *J. Biol. Chem.* 38 (2000): 29672-29682.

Arkin, A, Shen, P, Ross, J. A test case of correlation metric construction of a reaction pathway from measurements. *Science* 277 (1997): 1275-1279.

Baldi, P, Long, AD. A Bayesian framework for the analysis of microarray expression data: regularized t-test and statistical inferences of gene changes. *Bioinformatics* 17 (2001): 509-519.

Basset, DE, Eisen, MB, Boguski, MS. Gene expression informatics – It's all in your mine. *Nat Genet* 21 (1999): 51-55.

Ben-Dor, A, Shamir, R, Yakhini, Z. Clustering gene expression patterns. *J Comput Biol* 6 (1999): 281-297.

Blaschke, C, Andrade, AM, Ouzounis, C, Valencia, A. Automatic extraction of biological information from scientific text: protein-protein interactions. *Proc ISMB '99* (1999): 60-67.

Brazma, A, Hingamp, P, Quackenbush, J, Sherlock, G, Spellman, P, Stoeckert, C, Aach, J; Ansorge, W, Ball, CA, Causton, HC, Gaasterland, T,Glenisson, P, Holstege FC, Kim, IF, Markowitz, V, Matese, JC, Parkinson, H, Robinson, A, Sarkans, U, Schulze-Kremer, S, Stewart, J, Taylor, R, Vilo, J, Vingron, M. Minimum information about a microarray experiment (MIAME)-toward standards for microarray data. *Nat Genet.* 29 (2001): 365-371.

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

Brazma, A, Jonassen, I, Vilo, J, Ukkonen, E. Predicting gene regulatory elements in silico on a genomic scale. *Genome Res.* 8 (1998): 1202-1215.

Brett, D, Pospisil, H, Valcarcel, J, Reich, J, Bork, P. Alternative splicing and genome complexity. *Nat Genet* 30 (2001): 29-30.

Brown, MPS, Grundy, WN, Lin, D, Cristianini, N, Sugnet, CW, Furey, TS, Are,s M, Haussler, D. Knowledge-based analysis of microarray gene expression data using support vector machines. *Proc natl Acad Sci USA* 97 (2000): 262-267.

Chen, T, He, HL, Church, GM. Modeling gene expression with differential equations. *Pacific Symposium on Biocomputing* 4 (1999): 29-40. see http://www.smi.stanford.edu/projects/helix/psb99/Chen.pdf.

Chu, S, DeRisi, J, Eisen, M, Mulholland, J, Botsein, D, Brown, PO, Herskowitz, I. The transcriptional program sporulation in budding yeast. *Science* 282 (1998): 699-705.

Coward, E, Haas, SA, Vingron, M. SpliceNest: visualization of gene structure and alternative splicing based on EST clusters. *Trends Genet.* 18 (2002): 53-55.

Cummings, CA. Application of SOTA, a growing neural network algorithm, to gene expression profile clustering. *Briefings on Bioinformatics* 2 (2001): 402-404.

D'haeseleer, P, Liang, S, Somogyi, R. Genetic network inference: from co-expression clustering to reverse engineering. *Bioinformatics* 16 (2000): 707-726.

Der, SD, Zhou, A, Williams, BRG, Silverman, RH. Identification of genes differentially regulated by interferon a, b, or g using oligonucleotide arrays. *Proc. Natl. Acad. Sci. USA* 95 (1998): 15623–15628.

Dopazo, J, Carazo, JM. Phylogenetic reconstruction using a growing neural network that adopts the topology of a phylogenetic tree. *J. Mol. Evol* 44 (1997): 226-233.

Dopazo, J, Zanders, E, Dragoni, I, Amphlett, G, Falciani, F. Methods and approaches in the analysis of gene expression data. *J. Immunol Meth* 250 (2001): 93-112.

Drummond, M, Stamper, J. DNAPROBE, a computer program which generates oligonucleotide probes from protein alignments. *Nucl Acids Res* 27 (1999): 3493.

Efron, B, Tibsirani, R. Statistical data analysis in the computer age. *Science* 253 (1991): 390-395.

Eisen, M, Spellman, P L, Brown, PO, Botsein, D. Cluster analysis and display of genome-wide expression patterns. *Proc. Natl Acad. Sci. USA* 95 (1998): 14863-14868.

Everitt, BS, Dunn, G. *Applied multivariate data analysis.* New York: Oxford University Press, 1992.

Fellenberg, K, Hauser, NC, Brors, B, Neutzner, A, Hoheisel, JD, Vingron, M. Correspondence analysis applied to microarray data. *Proc Natl Acad Sci USA* 98 (2000): 10781-10786.

Friedman, N, Linial, M, Nachman, I, Pe'er, D. Using Bayesian networks to analyse expression data. *J Comput Biol* 7 (2000): 601-620.

Fritzke, B. Growing cell structures - a self-organizing network for unsupervised and supervised learning. *Neural networks* 7 (1994): 1141-1160.

Fukuda, K, Tsonuda, T, Tamura, A, Takagi, T. Information extraction: identifying protein names from biological papers. *Proc Pacific Symposium Biocomputing* (1998): 707-718.

Furey, TS, Cristianini, N, Duffy, N, Bednarski, DW, Schummer, M, Haussler, D. Support vector machine classification and validation of cancer tissue samples using microarray expression data. *Bioinformatics* 16 (2000): 906-914.

Glynne, R, Akkaraju, S, Healy, JI, Rayner, J, Goodnow, CC, Mack, DH. How self tolerance and the immunosuppressive drug FK506 prevent B-cell mitogenesis. *Nature* 403 (2000): 672-676.

Goryachev, AB, Macgregor, PF, Edwards, AM. Unfolding of microarray data. *J. Comput. Biol* 8 (2001): 443-461.

Hand, DJ. *Discrimination and classification.* NY: Wiley, 1981.

Harrington, CA, Rosenow, C, Retief, J. Monitoring gene expression using DNA microarrays. *Curr. Opin. Microbiol.* 3 (2000): 285-291.

Hartigan, JA. *Clustering algorithms.* New York: Wiley, 1975.

Herrero, J, Valencia, A, Dopazo, J. A hierarchical unsupervised growing neural network for clustering gene expression patterns. *Bioinformatics* 17 (2001): 126-136.

Herwig, R, Poustka, AJ, Müller, C, Bull, C, Lehrach, H, O'Brien, J. Large-scale clustering of cDNA-fingerprinting data. *Genome research* 9 (1999): 1093-1105.

Heyer, LJ, Kruglyak, S, Yooseph, S. Exploring expression data: identification and analysis of coexpressed genes. *Genome Res.* 9 (1999): 1106-1115.

Hill, AA, Brown, EL, Whitley, MZ, Tucker-Kellog, G, Hunter, CP, Slonim, DK. Evaluation of normalization procedures for oligonucleotide array data based on spiked cDNA controls. *Genome Biology* 2 (2001): resarch0055.1-0055.13.

Hill, AA, Hunter, CP, Tsung, BT, Tucker-Kellog, G, Broiwn, EL. Genomic analysis of gene expression in *C. elegans. Science* 290 (2000): 809-812.

Hughes, TR, Mao, M, Jones, AR, Burchard, J, Marton, MJ, Shannon, KW, Lefkowitz, SM, Ziman, M, Schelter, JM, Meyer, MR, Kobayashi, S, Davis, C, Dai, H, He, YD, Stephaniants, SB, Cavet, G, Walker, WL, West, A, Coffey, E, Shoemaker, DD, Stoughton, R, Blanchard, AP, Friend, SH, Linsley, PS. Expression profiling using microarrays fabricated by an ink-jet oligonucleotide synthesizer. *Nat. Biotechnol* 19 (2001): 342-347.

Hunter, L, Taylor, RC, Leach, SM, Simon, R. GEST: a gene expression search tool based on a novel Bayesian similarity metric. *Bioinformatics* 17 (2001): S115-S122.

International Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. *Nature* 409 (2001): 860–921.

Jenssen, T-K, Laegreid, A, Komorowski, J, Hovig, E. A literature network of human genes for high-throughput analysis of gene expression. *Nature Genetics* 28 (2000): 21-28.

Kerr, MK, Churchill, GA. Experimental design for gene expression microarrays. *Biostatistics* 2 (2001): 183-201.

Kerr, MK, Churchill, GA. Bootstrapping cluster analysis: assessing the reliability of conclusions from microarray experiments. *Proc Natl Acad Sci USA* 98 (2001): 8961-8965.

Kerr, MK, Martin, M, Churchill, GA. Analysis of variance for gene expression microarray data. *Journal Comput. Biol.* 7 (2000): 819-837.

Khan, J, Wei, JS, Ringnér, M, Saal, LH, Ladanyi, M, Westermann, F, Berthold, F, Schwab, M, Antonescu, CR, Peterson, C, Meltzer, PS. Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. *Nature Med.* 7 (2001): 673-579.

Knight, J. When the chips are down. *Nature* 410 (2001): 860-861.

Kohonen, T. *Self-organizing maps.* Berlin: Springer-Verlag, 1997.

Kruglyak, S, Tang, H. A new estimator of significance of correlation in time series data. *J. Comput Biol* 8 (2001): 463-470.

Li, C, Wong, WH. Model-based analysis of oligonucleotide arrays: expression index computation and outlier detection. *Proc Natl Acad Sci USA* 98 (2001): 31-36.

Lockhart, DJ, Dong, H, Byrne, MC, Follettie, MT, Gallo, MV, Chee, MS, Mittmann, M, Wang, C, Kobayashi, M, Horton, H, Brown, EL. Expression monitoring by hybridisation to high-density oligonucleotide arrays. *Nat. Biotechnol.* 14 (1996): 1675-1680.

Mateos, A, Dopazo, J, Jansen, R, Tu, Y, Gerstein, M, Stolovitzky, G. Systematic Learning of gene functional classes from DNA array expression data by using multi-layer perceptrons. *In press* (2002).

Mateos, A, Herrero, J, Tamames, J, Dopazo, J. *Supervised neural networks for clustering conditions in DNA array data after reducing noise by clustering gene expression profiles.* Edited by S Lin and K Johnson. Methods of Microarray Data Analysis II. Kluwer, 2002 (*in press*).

Mahalanobis, PC. On the generalized distance in statistics. *Proc. Natl. Inst. Sci. India* 12 (1936): 49-55.

Miller, RT, Christoffels, AG, Gopalakrishnan, C, Burke, J, Ptitsyn, AA, Broveak, TR, Hide, WA. A Comprehensive Approach to Clustering of Expressed Human Gene Sequence: The

Sequence Tag Alignment and Consensus Knowledge Base. *Genome Res.* 9 (1999): 1143-1155.

Oliveros, JC, Blaschke, C, Herrero, J, Dopazo, J, Valencia, A. Expression profiles and biological function. *Genome Informatics* 10 (2000): 106-117.

Ohta, Y, Yamamoto, Y, Okazaki, T, Uchiyama, I, Takagi, T. Automatic construction of knowledge base from biological papers. *Proc ISMB '97* (1997): 218-225.

Perou, M, Jeffrey, SS, van de Rijn, M, Ree, C, Eisen, MB, Ross, DT, Pergamenschikov, A, Williams, CF, Zhu, SX, Lee, JCF, Lashkari, D, Shalon, D, Brown, PO, Botstein, D. Distinctive gene expression patterns in human mammary epithelial cells and breast cancers. *Proc. Natl. Acad. Sci. USA* 96 (1999): 9112-9217.

Quackenbush, J. Computational analysis of microarray data. *Nature Rev Genet* 2 (2001): 418-427.

Roberts, CJ, Nelson, B, Marton, MJ, Stoughton, R, Meyer, MR, Bennett, HA, He, YD, Dal, H, Walker, WL, Hughes, TR, Tyers, M, Boone, C, Friend, SH. Signaling and circuitry of multiple MAPK pathways revealed by a matrix of global gene expression profiles. *Science* 287 (2000): 873-880.

Rogge, L, Bianchi, E, Biffi, M, Bono, E, Chang, SY, Alexander, H, Santini, C, Ferrari, G, Sinigaglia, L, Seiler, M, Neeb, M, Mous, J, Sinigaglia, F, Certa, U. Transcript imaging of the development of human T helper cells using oligonucleotide arrays. *Nat Genet.* 25 (2000): 96-101.

Ross, DT, Scherf, U, Eisen, MB, Perou, CM, Rees, C, Spellman, P, Iyer, V, Jeffrey, SS, van de Rijn, M, Waltham, M, Pergamenschikov, A, Lee, JC, Lashkari, D, Shalon, D, Myers, TG, Weinstein, JN, Botstein, D, Brown, PO. Systematic variation in gene expression patterns in human cancer cell lines. *Nat Genet.* 24 (2000): 227-35.

Schena, M, Shalon, D, Heller, R, Chai, A, Brown, PO, Davis, RW. Parallel human genome analysis: microarray-based expression monitoring of 1000 genes. *Proc. Natl. Acad. Sci. USA.* 93 (1996), 10614-10619.

Scherf, U, Ross, DT, Waltham, M, Smith, LH, Lee, JK, Tanabe, L, Kohn, KW, Reinhold, WC, Myers, TG, Andrews, DT, Scudiero, DA, Eisen, MB, Sausville, EA, Pommier, Y, Botstein, D, Brown, PO, Weinstein, JN. A gene expression database for the molecular pharmacology of cancer. *Nat Genet.* 24 (2000): 236-44.

Sneath, PHA, Sokal, RR . Numerical Taxonomy. W. H. Freeman: San Francisco, 1973.

Spellman, PT, Sherlock, G, Zhang, MQ, Iyer, VR, Anders, K, Eisen, MB, Brown, PO, Botstein, D, Futcher, B. Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization. *Mol. Bio. Cell* 9 (1998): 3273-3297.

Tamames, J, Clark, D, Herrero, J, Dopazo, J, Blaschke, C, Fernández, JM, Oliveros, JC, Valencia, A. Bioinformatics methods for the análisis of expresión arrays: data clustering and information extraction. *J. Biotechnol.* (2002, in press).

Tamayo, P, Slonim, D, Mesirov, J, Zhu Q, Kitareewan, S, Dmitrovsky, E, Lander, ES, Golub TR. Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation. *Proc. Natl. Acad. Sci. USA 96* (1999): 2907-2912.

Tanabe, L, Smith, LH, Lee, JK, Scherf, U, Hunter, L, Weinstein, JN. MedMiner: An internet tool for filtering and organizing bio-medical information, with application to gene expression profiling. *BioTechniques* 27 (1999): 1210-1217.

Tanaka, TS, Jaradat, SA, Lim, MK, Kargul, GJ, Wang, X, Grahovac, MJ, Pantano, S, Sano, Y, Piao, Y, Nagaraja, R, Doi, H, Wood III, WH, Becker, KG, Ko, MSH. Genome-wide

expression profiling of mid-gestation placenta and embryo using a 15,000 mouse developmental cDNA microarray. *Proc. Natl. Acad. Sci.* 97 (2000): 9127–9132.

Tavazoie, S, Hughes, JD, Campbell, MJ, Cho, RJ, Church, GM. Systematic determination of genetic network architecture. *Nature genetics* 22 (1999): 281-285.

Törönen, P, Kolehmainen, M, Wong, G, Castrén, E. Analysis of gene expression data using self-organizing maps. *FEBS letters* 451 (1999): 142-146.

Troyanskaya, O, Cantor, M, Sherlock, G, Brown, P, Hastie, T, Tibshirani, R, Botstein, D, Altman, RB. Missing value estimation methods for DNA microarrays. *Bioinformatics* 17 (2001): 520-525.

Venter, JC *et al.* The sequence of the human genome. *Science* 292 (2001): 1304-1351.

Voehringer, DW, Hirschberg, DL, Xiao, J, Lu, Q, Roederer, M, Lock, CB, Herzenberg, LA, Steinman, L, Herzenberg, LA. Gene microarray identification of redox and mitochondrial elements that control resistance or sensitivity to apoptosis. *Proc. Natl. Acad. Sci. USA* 97 (2000): 2680-2685.

Webb, GC, Akbar, MS, Zhao, C, Steiner, DF. Expression profiling of pancreatic b cells: Glucose regulation of secretory and metabolic pathway genes. *Proc Natl. Acad. Sci. USA* 97(2000): 5773-5778.

Wen, X, Fuhrman, S, Michaels, GS, Carr, DB, Smith, S, Barker, JL, Somogyi, R. Large-scale temporal gene expression mapping of central nervous system development. *Proc. Natl. Acad. Sci. USA* 95 (1998): 334-339.

Wilbur, WJ, Coffee, L. The effectiveness of document neighbouring in search enhancement. *Inf Process Manag* 30 (1994): 253-266.

Wittes, J, Friedman, HP. Searching for evidence of altered gene expression: A comment on statistical analysis of microarray data. *J. Natl. Cancer Inst.* 91 (1999): 400–401.

Yang, YH, Buckley, MJ, Speed, TP. Analysis of cDNA microarray images. *Briefings on Bioinformatics* 2 (2001): 341-349.

Yeung, KY, Ruzzo, WL. Principal component analysis for clustering gene expression data. *Bioinformatics* 17 (2001): 763-774.

Zhang, L, Zhou, W, Velculescu, VE, Kern, SE, Hruban, RH, Hamilton, SR, Volgestein, B, Kinzler, KW. Gene expression profiles in normal and cancer cells. *Science* 276 (1997): 1268-1272.

This page intentionally left blank

4

BIOLOGY-DRIVEN CLUSTERING OF MICROARRAY DATA

Applications to the NCI60 Data Set

Kevin R. Coombes, Keith A. Baggerly, David N. Stivers, Jing Wang, David Gold, Hsi-Guang Sung, and Sang-Joon Lee
Department of Biostatistics, UT M.D. Anderson Cancer Center, Houston TX 77030 USA

Abstract: In this note, we describe an approach to the analysis of microarray data that uses significant biological information early in the process. By combining information about the biological function and chromosomal location of genes with microarray expression data, we are able to get a more comprehensive picture of the heterogeneity of different kinds of cancer. We also get information about the importance of different chromosomes and biological processes for distinguishing cancers. In general, methods that use existing biological knowledge are likely to provide more meaningful and more interpretable results than completely unsupervised methods.

Key words: DNA microarrays, gene ontology.

1. INTRODUCTION

At present, the prototypical statistical analysis of a set of microarray data relies on methods that apply equally well to any data set that can be structured as a single large matrix. Very few methods attempt to use existing biological knowledge early in the analysis. In the study of cancer, this approach seems inadequate. We already know about many genes that are important in the study of cancer because of their involvement in specific biological processes; see, for instance, the curated cancer gene lists maintained by the National Cancer Institute [NCI, 2001]. We also know that cancers frequently exhibit systematic chromosomal abnormalities, and that these abnormalities can, in some cases, be detected by gene expression profiling with microarrays [Virtaneva *et al.*, 2001].

In this paper, we describe a way to use existing knowledge about the location of genes on chromosomes and existing functional annotations to analyze the NCI60 data set [Ross *et al.,* 2000; Scherf *et al.,* 2000]. We began by carefully updating the gene annotations, including chromosome mappings. Using the structured vocabulary produced by the Gene Ontology Consortium [GOC, 2000], we selected several of the functional categories of genes that are represented on the array for detailed analysis. We carried out separate hierarchical cluster analysis and principal component analysis using the genes from each chromosome and the genes assigned to each functional category. We found that this method is at least as good at clustering cell lines from related types of cancer as methods that use all genes or methods that select genes based on a variation filter. Moreover, this method allows us to interpret the genes involved without resorting to *ex post facto* arguments. By combining genes (or principal components of genes) from known functional categories that are identified by this analysis as useful for distinguishing different types of cancer, we should also be able to cluster the remaining expressed sequence tags (ESTs) on the microarray and make predictions about their functional roles.

2. THE ANNOTATION PROBLEM

2.1 Reannotating the Spots

The critical step in applying existing functional and chromosomal annotations from the public databases to the analysis of microarray data is to collect up-to-date information about the genes spotted on the array. Most of the spots on the microarrays used to acquire the NCI60 data were already annotated with GenBank accession numbers. Although the spot annotations included gene symbols and descriptions, this information must be considered suspect. Symbols and descriptions are properly associated to UniGene clusters or LocusLink identifiers rather than to GenBank accession numbers; every time UniGene is rebuilt, some accession numbers become associated with different clusters, or with no cluster at all.

To address this problem, we downloaded the latest build of UniGene (ftp://ncbi.nlm.nih.gov/repository/UniGene/Hs.data.Z, build 137 as of July 2001) and reannotated all 10,000 spots on the microarrays using a combination of perl scripts and SQL queries against an evolving Microsoft Access database. In the existing descriptions of the microarrays, most spots were annotated with two GenBank accession numbers, one from the 3' end and the other from the 5' end of the gene. Of the 10,000 spots on the array, 8,192 spots were annotated with both 3' and 5' accession numbers, 1,514

had only one accession number, and 294 had no accession number. The latter spots consisted primarily of empty spots and controls. The results of our updated analysis of the annotations are shown in Table 1. Only 7,478 of the spots on the microarray could be mapped to a unique cluster in the current build of UniGene. (These categories of spots are indicated with an asterisk in Table 1.) The 7,478 spots represented 6,614 distinct genes.

Table 1. Current UniGene status of the accession numbers of the spots on the NCI60 microarrays. Asterisks indicate categories of genes with usable UniGene annotations.

Number of Spots	Existing GenBank annotations	Current UniGene Status
294	No available accession number	None
128	Only 3' available	Unknown to UniGene
* 1379	Only 3' available	Known to UniGene
1	Only 5' available	Unknown to UniGene
* 6	Only 5' available	Known to UniGene
399	Both 3' and 5' available	Both unknown
763	Both 3' and 5' available	3' known, 5' unknown
291	Both 3' and 5' available	3' unknown, 5' known
646	Both 3' and 5' available	Both known, but disagree
* 6093	Both 3' and 5' available	Both known, and agree

Along with the current UniGene cluster numbers, we downloaded current gene symbols, names, chromosome mappings, cytogenetic positions, and LocusLink identifiers. Of the 6,614 distinct genes, 6,059 had been mapped to specific chromosomes in the current UniGene database. When compared with the total number of genes per chromosome listed in the MapViewer at the National Center for Biotechnology Information (NCBI) as of July 2001, most individual chromosomes are represented on the array in roughly the expected numbers of genes (data not shown). The only exception is the Y chromosome, which is represented by a single gene on the microarray.

2.2 Finding Functional Categories

2.2.1 From UniGene to LocusLink

As mentioned above, the UniGene annotations include a mapping from clusters to LocusLink identifiers. LocusLink provides a gateway to additional functional information about the genes. Of the 6,614 distinct genes on these microarrays, 5,074 were associated with LocusLink identifiers. The remaining 1,540 genes consisted primarily of expressed

sequence tags (ESTs) and "hypothetical proteins" with no further functional information.

2.2.2 From LocusLink to Gene Ontology

LocusLink includes functional annotations for many genes using the structured vocabulary being developed by the Gene Ontology Consortium. To access this information, we downloaded the latest release of LocusLink (ftp://ncbi.nlm.nih.gov/refseq/LocusLink/LL_tmpl) from the NCBI web site. Using a combination of perl scripts and SQL queries against our Access database, we located all the Gene Ontology numbers associated with each of the spots on the array that were already mapped to a LocusLink identifier.

The functional annotations in Gene Ontology form a directed acyclic graph, with a single root node and three top-level nodes representing biological processes, molecular functions, and cellular components. Of the 5,074 LocusLink identifiers represented on the microarray, we found that 2,989 had a functional annotation:

- 2,484 genes had at least one molecular function;
- 2,399 genes were involved in a biological process;
- 1,883 genes were localized to one cellular component.

For genes that had at least one Gene Ontology annotation, we found a total of 11,277 annotations, broken down into:

- 3,758 annotations of molecular function;
- 4,762 annotations of biological process;
- 2,757 annotations of cellular component.

We decided to restrict our attention to annotations of biological processes for further analysis of the gene expression data. We believe that biological processes (which explain why something is being done) provide a more interpretable conceptual level for grouping genes than molecular function (what is being done) or cellular component (where it is being done).

2.2.3 Disentangling Gene Ontology

The Gene Ontology structured vocabulary is a directed acyclic graph with thousands of nodes. The existing annotations are frequently quite detailed, mapping individual genes far down in the hierarchy. Our goal was to select a few nodes at reasonably high levels to group the genes into sets each containing approximately 100 to 500 genes. This goal required us to

determine how many genes were mapped into subnodes of any given node in the graph.

Table 2. Functional categories of genes.

Function	Annotations	Spots
Oncogenesis	140	180
Apoptosis	128	138
Physiological processes	180	210
Perception of external stimuli	238	150
Ectoderm development	129	152
Mesoderm development	92	102
Cell adhesion	111	140
Cell-cell signaling	137	166
Cell surface receptor linked signal transduction	222	228
Intracellular signaling cascade	110	110
Cell motility	120	153
Cell organization and biosynthesis	98	118
Cell shape and size control	78	101
Intracellular protein traffic	157	188
Transport	146	136
Cell proliferation	197	249
Stress response	599	372
Radiation response	147	136
Cell cycle	494	283
Nucleic acid metabolism	695	595
Protein metabolism	471	567
Lipid metabolism	146	156
Carbohydrate metabolism	103	97
Energy pathways	88	98

To accomplish this task, we downloaded the complete Gene Ontology graph in XML format from the Gene Ontology web site. Using a perl script, we mapped the frequency of annotation from the genes into each node, and then percolated those frequencies up the tree. We then inspected the graph and selected 24 categories. The categories are shown in Table 2, along with the number of annotations into subnodes in the graph and the number of spots on

the array placed into each functional category. The number of annotations can exceed the number of spots because of multiple annotations per gene or because of multiple pathways in the graph. The number of spots can exceed the number of annotations because some genes are represented multiple times on the microarray.

3. PRELIMINARY ANALYSIS

3.1 Data Preprocessing

We used the normalization procedure followed by the original authors [Ross *et al.,* 2000]. Thus, for each microarray, local background is subtracted from the estimate of signal intensity at each spot. The background-corrected values in the second channel are rescaled to set the median log ratio of the background-corrected values between the channels to equal one.

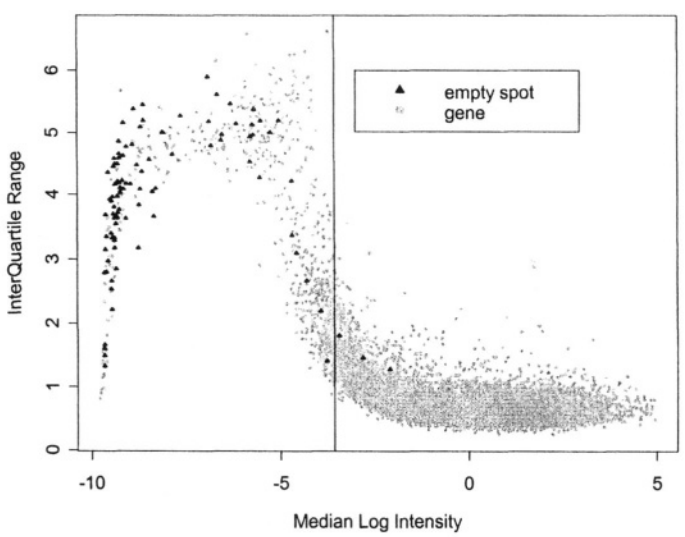


Figure 1. Variability of reference channel measurements as a function of log intensity.

We considered studentizing the log ratios; that is, taking each log ratio and dividing it by its standard deviation. The standard deviation can be estimated from the repeated measurements in the reference channel. To investigate this possibility, we normalized each channel independently,

rescaling the median intensity in the channel to one. We computed the interquartile range of each gene as the difference between the 75^{th} and 25^{th} percentile values of its log intensity. We plotted the interquartile range as a function of its median log intensity across experiments (Fig. 1). The vertical line in the graph marks the 97^{th} percentile of the empty spots on the arrays. Although studentizing would reduce the weight given to the spots to the left of this line, it is simpler to filter them out completely, since they are essentially unexpressed by any of the cell lines in this study. After this filtering step, 9,249 spots remain.

We next considered centering each row in the data matrix of log ratios; that is, subtracting the mean of each row from all entries in the row. Because the reference material is a mixture of twelve of the cell lines studied individually in these experiments, one might expect the average expression in the experimental channels to be roughly equal to the average expression in the reference channel; that is, the average log ratios across all experiments should be zero. To investigate this possibility, we plotted the median log ratio (over all experiments) of each gene as a function of its median log intensity in the reference channel (Fig. 2). We found that a substantial number of genes were consistently expressed at higher levels in the reference channel than in the experimental channel: 1,027 genes have median log ratios greater than 0.5. This result makes sense if these genes differ sharply in their expression between cell lines. We also performed cluster analyses using only the genes whose average log ratio was small (less than 0.5 in absolute value). These genes were still able to recover a great deal of information about how the cell lines were related (data not shown). In order to give both groups of genes comparable weight in our further analyses, we centered each row of the log ratio matrix.

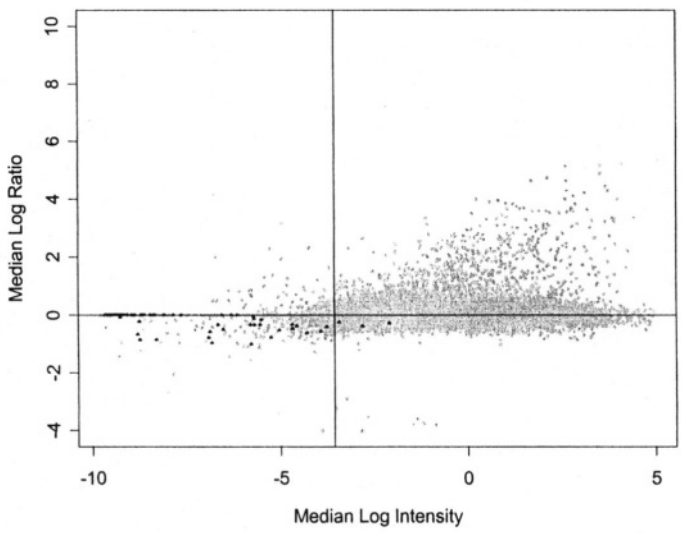


Figure 2. Median log ratio between channels as a function of the median log intensity in the reference channel.

3.2 Updating Cell Line Classifications

Preliminary cluster analysis suggested that the origins of some cell lines in this study were incorrectly annotated (data not shown). Cell line SNB-75 is described by the original authors as derived from a renal call carcinoma, but it tended to cluster with cell lines derived from tumors of the central nervous system (CNS) [Ross *et al.,* 2000]. At the NCI web site (http://dtpws4.ncifcrf.gov/DOCS/misc/common_files/cell_list.html), this cell line is listed as CNS-derived. A search of the medical literature found a number of references confirming this classification [Shi *et al.,* 1995]. Next, the cell line ADR-RES is described by Ross *et al.* as being of unknown origin. The NCI web site describes ADR-RES as a breast cancer cell line; published articles describe it as a multidrug resistant cell line developed from the breast cancer cell line MCF-7 [Nieves-Neira and Pommier, 1999]. We updated the annotations on both cell lines for further analysis. Finally, after confirming that replicate experiments from the same cell line (K562 or MCF7) clustered together repeatedly, we decided to use only one member of each replicate pair in order to avoid giving undue weight to the replicated cell lines.

3.3 Choosing a Distance Metric

Before performing analyses using sets of genes filtered by chromosomal location or by functional category, we had to choose a distance metric to compare both genes and samples. We considered four different metrics: Euclidean distance, the Pearson correlation coefficient, the Spearman rank correlation, and the Euclidean distance between rows of the Pearson correlation matrix. Each correlation-based measure was converted into a distance metric by taking *distance = (1 – correlation)/2*. We performed analyses on the data set consisting of all 9,249 nonempty spots using both multidimensional scaling and average linkage hierarchical clustering for each distance metric. All the correlation-based metrics gave comparable results (data not shown). Analyses using Euclidean distance seemed to be driven strongly by a few outlying samples (which were in turn driven by a small number of outlying genes). After reviewing these preliminary results, we decided to perform all further analysis using the distance metric based on the Pearson correlation coefficient. The cluster analysis using all 9,249 genes and this metric displays coherent patterns for 7 of the 9 kinds of cancer; only breast cancer and prostate cancer fail to cluster together strongly (Fig. 3). Analysis of the data using multidimensional scaling or principal components gave results that were consistent with the hierarchical clustering (data not shown).

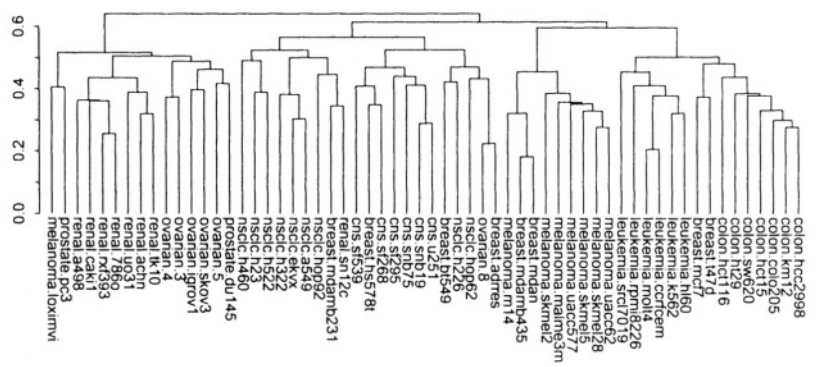


Figure 3. Hierarchical clustering using 9,249 genes, with distance determined by the Pearson correlation coefficient. Seven of the nine kinds of cancer cell lines cluster strongly together.

4. CHROMOSOMAL CLUSTERING

For each chromosome (except for the Y chromosome, for reasons described above), we formed a data matrix whose columns were the 60 experimental samples and whose rows were the genes on the microarray from that chromosome. Genes unexpressed in the reference channel were omitted. The matrix contained normalized log ratios between the experimental samples and the reference channel. In addition, the values in each row were centered to have mean zero. We performed principal components analysis and hierarchical clustering using the distance metric derived from the Pearson correlation coefficient. We scored each dendrogram for its ability to cluster samples into classes that correctly reflected the kind of cancer involved, using the following ordinal scale:

- A = a cluster contains all and only samples of that kind.
- B = a cluster contains all samples of that kind of cancer, but includes one or two extraneous samples.
- C = a cluster contains all but one sample of that kind.
- D = a cluster contains all but one sample of that kind, but includes one or two extraneous samples.
- E = a cluster contains all but two samples of that kind.
- F = samples of that kind are weakly clustered.

Using this scale, the scores for each cancer from the dendrogram of Figure 3 are shown in Table 3 (abbreviations: B = breast, C = colon, L = leukemia; M =melanoma; N = non small cell lung cancer; O = ovarian; P = prostate; R = renal; S = central nervous system).

Table 3. Scores for the dendrogram shown in Figure 3.

Cancer	B	C	L	M	N	O	P	R	S
Score		A	A	D	F	D		C	E

Using an ordinal scale allows us to rank the quality of a clustering based on how well it identifies individual cancer types, but it does not allow us to assign a single quantitative score for the overall quality of the cluster. Replacing A–F by numerical values 1–6 would allow us to compute numerical averages, but raises two potential difficulties. First, we do not know *a priori* how to weight the relative importance of clustering specific types of cancer. Second, we do not have a sound statistical basis for interpreting the significance of the scores that would result. For these reasons, we restricted ourselves to the ordinal scale.

Table 4. Scores showing ability of genes on specific chromosomes to cluster cancer types.

Chromosome	B	C	L	M	N	O	P	R	S
1		B	A	D	F			D	B
2		E	C	D		D	E	D	E
3		C	E	D				E	F
4			E	E			E	E	
5		A	A	D	F		E		
6		C	A	D			E	E	D
7		E	A	D		E		C	E
8		E		C				D	
9		B	C	D	F	E			D
10		C	C	E		E	E		
11		E		C				C	D
12		B	C	C		E	E	E	
13				D	E				
14		A	A		F				
15		C	B	C	F			C	
16									
17		A	A	D	F	E			E
18		E	D						
19				D		D			
20		E				C			
21									
22		A		E					E
X		B	A	D				E	D

The scores for dendrograms using one chromosome at a time are shown in Table 4. We can draw a number of conclusions from this table. First, genes on chromosomes 16 and 21 are unable to distinguish any single kind of cancer. At the opposite extreme, chromosomes 1, 2, 6, 7, 9, 12, and 17 can be used to distinguish the largest numbers of cancer types (see Fig. 4 for examples). Also, breast cancer samples do not cluster together regardless of the chromosome being used. Conversely, leukemias and colon cancers are both easily distinguished from other kinds of cancer. Note also that the prevalence of D scores in the melanoma column results from one melanoma sample (LOXI MVI) that consistently clusters apart from the other samples, and from two putative breast cancer samples (MDA-MB-435 and MDA-N, both of which came from the same patient) that consistently cluster as nearest neighbors within the majority of the melanoma samples.

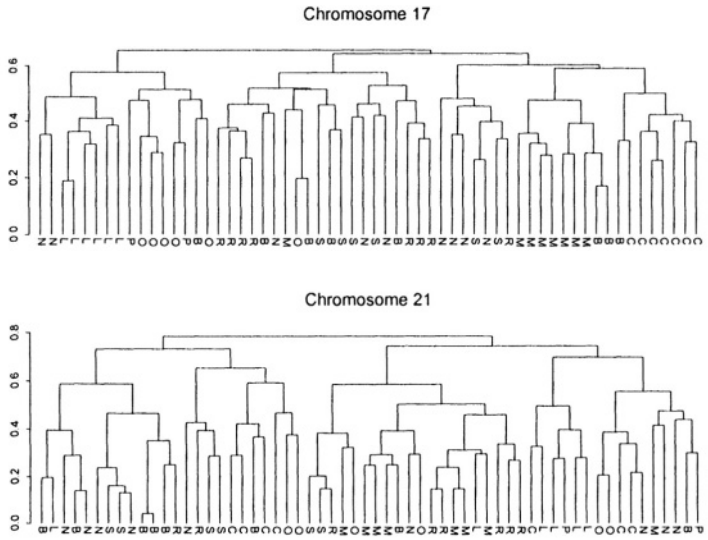


Figure 4. Dendrograms using genes restricted to a single chromosome. Chromosome 17 gives a highly structured result; chromosome 21 is essentially random.

5. FUNCTIONAL CLUSTERING

For each functional category (or biological process) of genes described above, we formed a data matrix whose columns were the experimental samples and whose rows were the spots on the microarray corresponding to genes involved in that biological process. Genes that were unexpressed in the reference channel were omitted from the data set. Each data set was analyzed as described for the analogous chromosome-specific data sets. The dendrograms were scored as described above; the result is shown in Table 5. Again, the colon cancers, leukemias, melanomas, and renal cancers are most easily distinguished, with the breast and lung cancers least likely to cluster together to any degree. Interestingly, genes involved in apoptosis do a poor job of distinguishing most kinds of cancers. The categories that do the best job of classifying melanomas are "radiation response" and "perception of external stimuli", both of which include a substantial number of genes involved in the perception of light. The categories that work best to distinguish many kinds of cancer are cell cycle, cell proliferation, and protein metabolism, suggesting that cancers use a wide range of different strategies to overcome the built-in controls on growth.

Table 5. Scores showing ability of genes in different functional categories to cluster cancer types.

Function	B	C	L	M	N	O	P	R	S
Oncogenesis		E	A	D				F	D
Apoptosis		A	D						
Physiological processes			E	F			F	D	
Perc. of external stimuli	F			C					
Ectoderm development				E	F				
Mesoderm development		D		D				E	D
Cell adhesion		D	D	B				D	
Cell-cell signaling	F		D	E					
Cell surf. rec. signal trans.		E	D	D	F	F		C	F
Intracell. signaling cascade		B	E	D		F			D
Cell motility		A	C	E					
Cell org. and biosyn.		C	C					F	
Cell shape and size ctrl.			B	F				D	F
Intracell. protein traffic		B						E	
Transport		B		D			B	F	E
Cell proliferation		D	E	D	F	C		C	F
Stress response		B	A	D		D		D	F
Radiation response		D	E	C					
Cell cycle		B	A	D		F	F	D	B
Nucleic acid metabolism		D	A	D		D		C	
Protein metabolism		A	C	D	F	D		C	E
Lipid metabolism	F		E	D			F	D	
Carbohydrate metabolism		C	C	D			B		
Energy pathways		E					B		

6. CONCLUSIONS

We can draw some general conclusions by looking at the dendrograms and the multidimensional scaling plots for all the chromosomes and all the functional categories. A complete set of these figures can be found in the supplementary data at our web site:

http://www.mdanderson.org/depts/cancergenomics/camda.html.

First, the colon cancer and leukemia cell lines appear to be the most homogeneous kinds of cancer included in this study, followed by the melanomas and the renal cancer cell lines. At the opposite extreme, the breast cancer cell lines seem to be the most heterogeneous, followed by the non-small cell lung cancers. The ovarian and central nervous system cancers fall somewhere in the middle of this spectrum. In addition, the colon cancers and leukemias share some significant common features, often clustering near one another. Similarly, the two prostate cancer cell lines frequently cluster near the ovarian cancer cell lines.

The tissue of origin of several cell lines in this study is in question. The original authors [Ross *et al.,* 2000; Scherf *et al.,* 2000] concluded that the putative breast cancer cell lines MDA-MB-435 and MDA-N were actually melanomas; our results support this conclusion. Another cell line, SNB-75, is listed as a renal cell carcinoma by Ross *et al.,* but was described previously as a CNS cancer [Shi *et al.,* 1995]. This cell line consistently clusters with other cancers of the central nervous system, and so one suspects that the earlier description of its origin is more likely to be correct. Finally, the ADR-RES cell lines is listed as being of "unknown" type by Ross *et al.,* but is described elsewhere as derived from a breast cancer cell line [Nieves-Neira and Pommier, 1999]. Across a wide variety of functional categories and chromosomes, this sample is most similar to one of the ovarian cancer cell lines, and so it is difficult to draw definitive conclusions about its origin.

We found substantial differences in the ability of collections of genes (on different chromosomes or in different functional categories) to distinguish between types of cancer. Some individual chromosomes were nearly as good at distinguishing cancer types as the full data set; others yielded essentially random permutations of the data. Similarly, some functional categories of genes were extremely good at distinguishing cancer types, and others were nearly useless.

Because this study did not include normal counterparts of the tissues of origin of the cancer cell lines, one must draw conclusions carefully. Categories of genes that distinguish cancer types may actually be reflecting differences in gene expression in the underlying tissues of origin. Categories of genes that fail to distinguish cancer types may do so either because they are unimportant for the study of cancer or because they are important – in similar ways – across a wide variety of cancers.

Apoptosis genes, for example, did not distinguish between different cancer types in this study. Without question, apoptosis genes play a critical role in the development and progression of cancer. In order to proliferate, every cancer must find a way to avoid the apoptotic pathways that lead to cell death. The inability of apoptosis genes to distinguish cancer types has

two possible interpretations: either most cancers avoid cell death by developing similar disturbances in the apoptotic pathways, or else each cancer finds an idiosyncratic way (having nothing to do with the tissue of origin) to avoid cell death. In either event, this finding suggests that cancer therapies targeted to specific genes in the apoptotic pathways are more likely to be useful across a wide range of types of cancer, but less likely to be useful for all patients with breast cancer, for instance.

ACKNOWLEDGEMENTS

This work was partially supported by the Tobacco Settlement Funds as appropriated by the Texas State Legislature, and by a generous donation from the Michael and Betty Kadoorie Foundation.

REFERENCES

Gene Ontology Consortium. Gene Ontology: tool for the unification of biology. *Nature Genetics* 25 (2000): 25-29. (See http://www.geneontology.org.)

National Cancer Institute. Cancer Genome Anatomy Project: Curated Cancer Gene Lists. *http://cgap.nci.nih.gov/ Genes/CuratedGeneLists* (2001).

Nieves-Neira, W, Pommier Y. Apoptotic response to camptothecin and 7-hydroxystaurosporine (UCN-01) in the 8 human breast cancer cell lines of the NCI Anticancer Drug Screen: multifactorial relationships with topoisomerase I, protein kinase C, Bcl-2, p53, MDM-2 and caspase pathways. *Int J Cancer* 8 (1999): 396–404.

Ross, DT, Scherf, U, Eisen, MB, Perou, CM, Rees, C, Spellman, P, Iyer, V, Jeffrey, SS, Van de Rijn, M, Waltham, M, Pergamenschikov, A, Lee, JC, Lashkari, D, Shalon, D, Myers, TG, Weinstein, JN, Botstein, D, Brown, PO. Systematic variation in gene expression patterns in human cancer cell lines. *Nature Genetics* 24 (2000): 227–235.

Scherf, U, Ross, DT, Waltham, M, Smith, LH, Lee, JK, Tanab,e L, Kohn, KW, Reinhold, WC, Myers, TG, Andrews, DT, Scudiero, DA, Eisen, MB, Sausville, EA, Pommier, Y, Botstein, D, Brown, PO, Weinstein, JN. A gene expression database for the molecular pharmacology of cancer. *Nature Genetics* 24 (2000): 236–244.

Shi, Q, Chen, K, Li, L, Chang, JJ, Autry, C, Kozuka, M, Konoshima, T, Este,s JR, Lin, CM, Hamel, E. Antitumor agents, 154 - Cytotoxic and antimitotic flavonols from Polanisia dodecandra. *J Nat Prod* 58 (1995): 475–482.

Virtaneva, K, Wright, FA, Tanner, SM, Yuan, B, Lemon, WJ, Caligiuri, MA, Bloomfield, CD, de La Chapelle, A, Krahe, R. Expression profiling reveals fundamental biological differences in acute myeloid leukemia with isolated trisomy 8 and normal cytogenetics. *Proc Natl Acad Sci USA* 98 (2001): 1124–1129.

This page intentionally left blank

5

EXTRACTING GLOBAL STRUCTURE FROM GENE EXPRESSION PROFILES

Charless Fowlkes[1], Qun Shan[2], Serge Belongie[3], and Jitendra Malik[1]
Departments of Computer Science[1] and Molecular Cell Biology[2], University of California at Berkeley; Department of Computer Science and Engineering, University of California at San Diego[3]

Abstract: We have developed a program, GENECUT, for analyzing datasets from gene expression profiling. GENECUT is based on a pairwise clustering method known as *Normalized Cut* [Shi and Malik, 1997]. GENECUT extracts global structures by progressively partitioning datasets into well-balanced groups, performing an intuitive k-way partitioning at each stage in contrast to commonly used 2-way partitioning schemes. By making use of the *Nyström* approximation, it is possible to perform clustering on very large genomic datasets.

Key words: gene expression profiles, clustering analysis, spectral partitioning

1. INTRODUCTION

DNA microarray technology empowers biologists to analyze thousands of mRNA transcripts in parallel, providing insights about the cellular states of tumor cells, the effect of mutations and knockouts, progression of the cell cycle, and reaction to environmental stresses or drug treatments. Gene expression profiles also provide the necessary raw data to interrogate cellular transcription regulation networks. Efforts have been made in identifying cis acting elements based on the assumption that co-regulated genes have a higher probability of sharing transcription factor binding sites.

There is a well-recognized need for tools that allow biologists to explore public domain microarray datasets and integrate insights gained into their

own research. One important approach for structuring the exploration of gene expression data is to find coherent clusters of both genes and experimental conditions. The association of unknown genes with functionally well-characterized genes will guide the formation of hypotheses and suggest experiments to uncover the function of these unknown genes. Similarly, experimental conditions that cluster together may affect the same regulatory pathway.

Unsupervised clustering is a classical data analysis problem that is still an active area of intensive research in the computer science and statistics communities [Ripley, 1996]. Broadly speaking, the goal of clustering is to partition a set of feature vectors into k groups such that the partition is "good" according to some cost function. In the case of genes, the feature vector is usually the degree of induction or suppression over some set of experimental conditions. As of yet, there is no clear consensus as to which algorithms are most suitable for gene expression data.

Clustering methods generally fall into one of two categories: *central* or *pairwise* [Buhmann, 1995]. Central clustering is based on the idea of prototypes, wherein one finds a small number of prototypical feature vectors to serve as "cluster centers". Feature vectors are then assigned to the most similar cluster center. Pairwise methods are based directly on the distances between all pairs of feature vectors in the data set. Pairwise methods don't require one to solve for prototypes, which provides certain advantages over central methods. For example, when the shape of the clusters are not simple, compact clouds in feature space, central methods are ill-suited while pairwise methods perform well since similarity is allowed to propagate in a transitive fashion from neighbor to neighbor. A family of genes related by a series of small mutations might well exhibit this sort of structure, particularly when features are based on sequence data.

Clustering algorithms can also often be characterized as greedy or global in nature. The agglomerative clustering method used by Eisen *et al.* [1998] to order microarray data is an example of a greedy pairwise method: it starts with a full matrix of pairwise distances, locates the smallest value, merges the corresponding pair, and repeats until the whole dataset has been merged into a single cluster. Because this type of process only considers the closest pair of data points at each step, global structure present in the data may not be handled properly.

Another unsupervised clustering approach that has been applied to gene expression analysis is the self –organizing map [Tamayo *et al*., 1999]. While this technique is useful for structuring data sets in some applications, the lack of an explicit "energy function" has made it difficult to analyze.

Our approach to clustering gene expression data is based on the *Normalized Cuts* (NCut) method introduced by Shi and Malik [1997; 2000]. Normalized Cuts is a pairwise clustering that finds a partitioning of the data set into well-balanced groups. The resulting clustering minimizes a well-defined, global cost function. Experience in the field of computer vision, VLSI layout and parallel computing suggests that spectral graph methods [Chung, 1997] such as Normalized Cuts provide excellent results on a wide range of practical problems. In Section 2, we outline the NCut method for clustering and in Section 3, demonstrate the application of NCut to the Rosetta yeast gene expression dataset [Hughes *et al.*, 2000].

2. CLUSTERING WITH NORMALIZED CUT

In this section we describe the NCut cost function, which provides a measure of cluster quality that takes into account both the within-group similarity and the between-group dissimilarity. We also outline the algorithm used for finding a clustering of the data that has low cost. The reader is referred to Shi and Malik [2000] and the references therein for additional detail.

2.1 The NCut Criterion

We use the Pearson correlation between vectors of expression data to capture the degree of similarity between two genes or two experiments. We will apply the same clustering algorithm to both the problem of clustering genes and that of clustering experiments so in this section we refer generically to the items being clustered. Let $\mathbf{W_{ij}}$ be the Pearson correlation between the *ith* and *jth* data points. First consider the case of partitioning the dataset into two groups (bi-partitioning). Let **V** denote the complete set of data which is broken into subsets **A** and **B**. The NCut cost function is defined as

$$NCut(A,B) = \frac{cut(A,B)}{assoc(A,V)} + \frac{cut(A,B)}{assoc(B,V)}$$

where the *cut* and *association*, defined as

$$cut(A,B) = \sum_{i \in A, j \in B} wij \qquad assoc(A,V) = \sum_{i \in A, j \in B} wij$$

are graph-theoretic terms that quantify the cost of this partition (the cut) and the total connection of the subset to the whole set (the association). Normalizing by the association term makes NCut different from graph theoretic techniques based on min-cut (applied to genomic data by Sharan and Shamir [2000], which can generate highly unbalanced clusters and require elaborate post-processing. Shi and Malik, [2000] provide a comparison.

While finding the **A-B** partition that minimizes the NCut criterion is an NP-hard optimization problem, it is possible to relax the constraints in order to obtain a closed form eigenproblem that yields high quality approximations. The problem is formulated in terms of minimizing the Rayleigh quotient,

$$\frac{y^T(D-W)y}{y^T Dy}$$

where **W** is the matrix whose entries are $\mathbf{W_{ij}}$, **D** is a diagonal matrix with $\mathbf{D_{ii}} = \Sigma\ \mathbf{W_{ij}}$ and **y** is a partition indicator vector. If we allow **y** to take on continuous values then the minimum is obtained by the second leading eigenvector of the generalized eigenvalue problem **(D-W)y=λDy.**

2.2 K-Way Partitioning

The NCut bi-partitioning technique has been applied to genomic expression data by Xing and Karp [2001] for a data set containing two clusters. However, for the analysis of a large compendium of expression data, we would expect there to exist far more than two clusters. Generalization to the case of more than two groups can be obtained in a number of ways. One method is to apply bi-partitioning recursively on **A** and **B**. Another method is to compute k leading eigenvectors instead of just the second one. This leads to a k-dimensional *embedding* that is amenable to clustering with simple central methods such as k-means [Duda and Hart, 1973]. The approach taken in our present work is a combination of these two methods. We perform a recursive k-way clustering where k is automatically chosen at each level to minimize the k-way NCut criterion defined as

$$NCut_k(A_1, A_2, \ldots, A_k) = \frac{1}{k}\sum_{i=1}^{k}\frac{cut(A_i, V - A_i)}{assoc(A_i, V)}$$

We find that this criterion constitutes an effective form of *model selection* and yields natural clusters while avoiding the artificial constraint of bi-partitioning or pairwise merging schemes.

2.3 Clustering Large Datasets

Our algorithm was prototyped in MATLAB where it takes less than a minute to cluster the 560 genes used in our experiments. For very large problems, the computation and memory requirements to solve the eigenproblem can become a limiting factor for interactive data analysis. To avoid these costs, we can exploit the *Nyström Approximation* which allows one to extrapolate the solution to a large clustering problem using a small subset of the data [Fowlkes *et al.*, 2001].

This approximation exploits redundancy between rows of the $\mathbf{W_{ij}}$ matrix by choosing a small subset of the genes and computing their similarity to every other gene in the dataset. This thin strip of the matrix is then used to compute a direct numerical approximation to the eigenvectors needed for partitioning. The memory and processing expenses grow in proportion to the number of samples rather than the total number of data points so by using this approximation, our method should extend efficiently to the analysis of complete genomes with thousands of experiments.

3. RESULTS

We have built a system for interactively browsing the results of the NCut algorithm called GENECUT. The clustering results presented in this paper along with prototype software are available from our website at **http://www.cs.berkeley.edu/~fowlkes/bio/**. In this section we present some results that indicate our algorithm is capable of finding clusters that exist in the data. A robust algorithm is extremely important since true clusters in a data set are unknown and poor clustering results could easily be misleading. While it is difficult to evaluate the performance of clustering algorithms quantitatively, we are able to point to clusters of well characterized genes which have closely related functions, suggesting that the algorithm is effective.

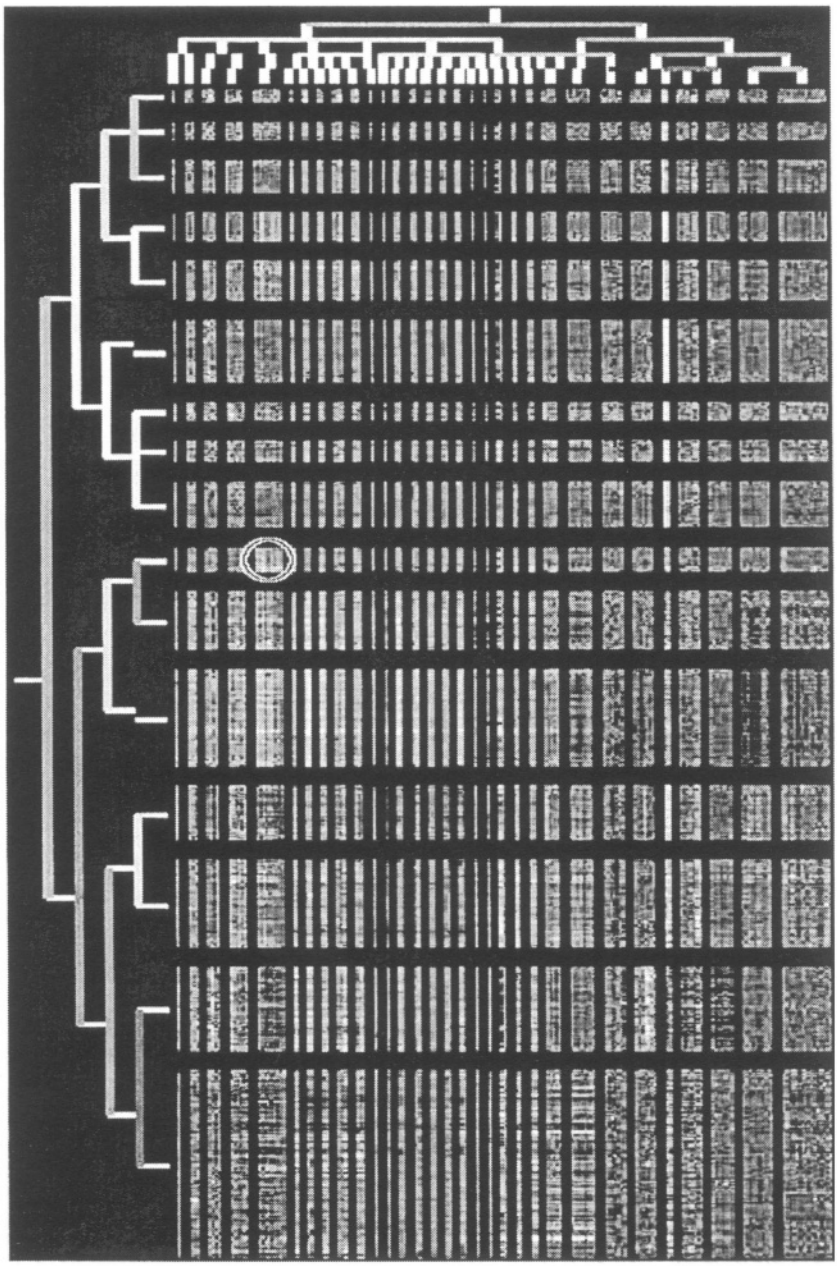


Figure 1. The result of performing a recursive, k-way partitioning on a set of 560 genes and 123 experiments. Genes are arrayed along the x-axis and experiments along the y-axis. The contents of the cluster indicated by the white circle are listed in Table 1 and Table 2. The color-coding on the tree indicates the cost of the associated k-way cut. The contents of other clusters are available for interactive exploration: **http://www.cs.berkeley.edu/~fowlkes/bio/**

Figure 1 gives a visual overview of the clustering analysis presented by GENECUT for the Rosetta gene expression dataset [Hughes *et al.*, 2000]. The output of the clustering algorithm is presented in the form of a web page that allows the user to traverse up and down through the layers of the tree structure in both the experimental and gene dimensions. The user can click on clusters in the overview image in order to view the genes and experiments in that cluster. Gene descriptions include links to detailed descriptions and a link that invokes a BLAST search of the *Saccharomyoes* Genome Database using the 500 bp upstream sequence.

We expect that clusters of genes showing similar expression patterns are likely to share some conservative regulatory motif. The ability to do a BLAST query quickly is a first step towards seeking similar transcription factor binding sites. We are currently exploring DNA motifs associated with several of these clusters. Automatic identification of these putative motifs would clearly be helpful in experimental design.

Experiment #	Description	
9	erg2	Deletion
10	erg3	Deletion
107	hmg1	Deletion
61	Yer044c (haploid)	Deletion
29	ERG11 (tet promoter)	Shutdown
35	HMG2 (tet promoter)	Shutdown
73	Lovastatin	drug treatment
82	Terbinafine	drug treatment
71	Itraconazole	drug treatment

Table 1. Experiment cluster #5, an interesting group of experiments found by GENECUT (shown circled in Figure 1). This cluster contains experimental conditions relating to the sterol synthesis pathway.

Table 1 shows a cluster along the experimental axis that groups together a set of experiments that all involve perturbations of sterol biosynthesis. To extract global features from an experimental cluster like these sterol synthesis experiments, we sort the gene clusters by their normalized variances. We reason that the makeup of gene clusters with high variance across a particular experiment cluster is likely to be biologically relevant.

Table 2 lists the gene cluster that has the highest mean variance in expression level for the sterol synthesis experiments cluster. This gene cluster makes biological sense and also agrees with a visual examination of the dataset.

Gene	System Name	Description
1	YHR007C	[*ERG11*] Cytochrome P450 (lanoterol 14 alpha-demethlase), essential for biosynthesis of ergosterol
110	YDR530C	[*APA2*] ATP adenylyltransferase II
169	YGL001C	[*ERG26*] C-3 sterol dehydrogenase, C-4 decarboxylase, required for ergosterol biosynthesis
195	YGR049W	[*SCM4*] Protein that suppressed temperature-sensitive allele of CDC4 when overexpressed
197	YGR060W	[*ERG25*] C-4 sterol methyl oxidase: enzyme of the ergosterol biosynthesis pathway
210	YGR175C	[*ERG1*] Squalene monooxygenase (squalene epoxidase), an enzyme of the ergosterol biosynthesis pathway
279	YJL113W	Unknown
337	YKRO53C	[*YSR3*] Sphingoid base-phosphate phosphatase, putative regulator of sphingolipid metabolism and stress response
344	YLL0112W	Protein with similarity to human triacylglycerol lipase
380	YML008C	[*ERG6*] S-adenosylmethionine delta-24-sterol-C-methyltransferase, carries out methylation of zymosterol as part of the ergosterol biosynthesis pathway
392	YMR015C	[*ERG5*] Cytochrome P450, delta 22(23) sterol desaturase, catalyses an intermediate pathway step in the biosynthesis pathway
434	YNL111C	[*CYB5*] Cytochrome b5
491	YOR237W	[*HES1*] protein implicated in ergosterol biosynthesis, member of the KES1/HES1/OSH1/YKR003W family of oxysterol-binding (OSBP) proteins
511	YOR394W	Member of the seripauperin (PAU) family (YPL282C and YOR394W code for identical proteins)
523	YPL272C	Unknown

Table 2. Gene cluster #10 found by GENECUT contains genes related to sterol biosynthesis. This cluster had the largest variance across experimental conditions for the set of experiments in experiment cluster #5

Many easily identified clusters discussed in [Hughes *et al.*, 2000] were also found by the GENECUT algorithm. This is notable since the two algorithms employed take quite different approaches (local agglomerative vs. global divisive). Figure 2 contrasts the genes found by our algorithm with those of Hughes *et al.* [2000] for the sterol gene cluster (our cluster #10). Genes that appear in the intersection of the two clusters are presumed to be related with high confidence while those which only appear in a single cluster require more experiments to pin down. Since the agglomerative clustering algorithm produces a dendrogram whose leaves are individual genes, the cluster shown is actually a manually selected sub-tree.

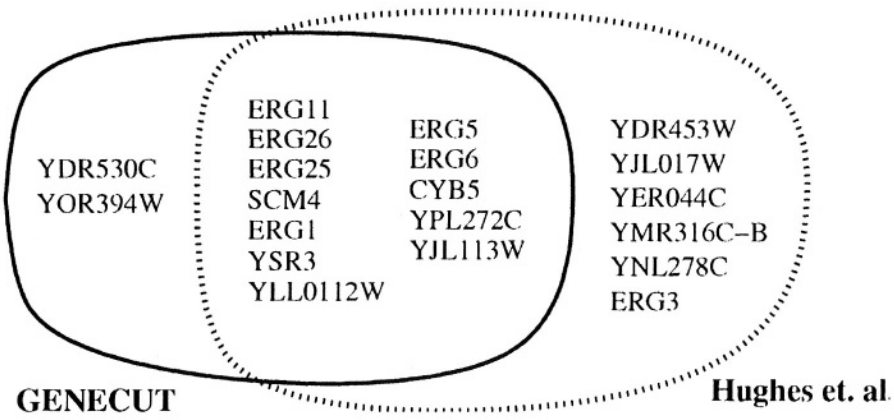


Figure 2. A comparison of the "sterol" cluster found by Hughes *et al.* [2000] (dotted circle) and that found by the GENECUT algorithm (solid circle). As with many other clusters, there is significant overlap.

4. CONCLUSIONS

In this report, we developed a novel application of the NCut algorithm to the problem of gene expression profile analysis. The algorithm performs favourably by focusing on the global features and recursively partitioning the dataset into clusters. We demonstrate the utility of NCut in extracting global features from an experiment cluster, and further explore regulatory sequences within the representative gene clusters. It may be possible to use this algorithm effectively in conjunction with hierarchical clustering tools in order to perform "harvesting" of dendrograms and allow rapid exploration of genomic data sets. We envision that this algorithm can ultimately be used as a general clustering tool in various areas of genomics research such as protein classification, DNA sequence data, and drug sensitivity profiling.

REFERENCES

Buhmann, JM. Data Clustering and Learning. In: Arbib, MA, ed. The Handbook of Brain Theory and Neural Networks. *MIT Press,* 1995.

Chung, FRK. Spectral Graph Theory. *American Mathematical Society* (1997).

Duda, R, Hart, P. *Pattern Classification and Scene Analysis.* John Wiley & Sons (1973).

Eisen, MB et al. Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci 95* (1998): 14863-14868.

Fowlkes, C, Belongie, S, Malik, J. Spatiotemporal grouping using the Nyström approximation. *Proc. IEEE Comput. Soc. Conf. Comput. Vision and Pattern Recogn* (2001).

Hughes, TR, Marton, MJ et al. Functional discovery via a compendium of expression profiles. *Cell* 102 (2000): 109-126.

Ripley, BD. Pattern Recognition and Neural Networks. *Cambridge* (1996).

Sharan, R, Shamir, R. Click: A clustering algorithm with applications to gene expression analysis. *Proc. Of ISMB. AAAI Press,* 2000.

Shi, J, Malik, J. Normalized cuts and image segmentation. *Proc IEEE Conf. Computer Vision and pattern Recognition* (1997): 731-737.

Shi, J, Malik, J. Normalized cuts and image segmentation. *IEEE Trans. PAMI 22* (2000): 888-905.

Tamayo, P et al. Interpreting patterns of gene expression with self-organizing maps: Methods and applications to hematopoietic differentiation. *Proc. Natl. Acad. Sci.* 96 (1999): 2907-2912.

Xing, EP, Karp, RM. Cliff: Clustering of high-dimensional microarray data via iterative feature filtering using normalized cuts. *Proc. Of the Nineteenth ISMB* (2001).

6

SUPERVISED NEURAL NETWORKS FOR CLUSTERING CONDITIONS IN DNA ARRAY DATA AFTER REDUCING NOISE BY CLUSTERING GENE EXPRESSION PROFILES

Alvaro Mateos, Javier Herrero, Javier Tamames[1] and Joaquín Dopazo
Bioinformatics Unit, National Center for Cancer Research (CNIO), Melchor Fernández Almagro 3, 28029 Madrid. Spain

[1] *ALMA Bioinformatics SL, c/ Ronda de Poniente 4, 28760 Tres Cantos, Madrid. Spain*

Abstract: In this paper we compare various applications of supervised and unsupervised neural networks to the analysis of the gene expression profiles produced using DNA microarrays. In particular we are interested in the classification of samples or conditions. We have found that if gene expression profiles are clustered at the optimal level, the classification of conditions obtained using the average gene expression profile of each cluster is better than that obtained directly using all the gene expression profiles. If a supervised method (a back propagation neural network) is used instead of an unsupervised method, the efficiency of the classification of conditions increases. We studied the relative efficiencies of different clustering methods for reducing the dimensionality of the gene expression profile data set and found that the Self-Organising Tree Algorithm (SOTA) is a good choice for this task.

Key words: SOTA, perceptron, clustering, linear runtime, gene expression, noise reduction.

1. INTRODUCTION

DNA microarray technology opens up the possibility of measuring the expression level of thousands of genes in a single experiment [Brown and Botsein, 1999]. Serial experiments measuring gene expression at different conditions or times, or distinct experiments with diverse tissues, patients,

etc., allow us to obtain gene expression profiles under the different experimental conditions studied. Initial experiments suggest that genes with similar expression profiles tend to play similar roles in the cell. Aggregative hierarchical clustering has been extensively used for finding clusters of co-expressed genes [Eisen *et al.*, 1998; Wen *et al.*, 1998]. Nevertheless, several authors [Tamayo *et al.*, 1999] have noted that aggregative hierarchical clustering suffers from a lack of robustness. In addition, typical aggregative hierarchical clustering methods have runtimes that can range from N^2 to N^4 [Hartigan, 1975], which makes them very slow when thousands of items are to be analyzed. In an attempt to overcome these problems, some authors have proposed the use of neural networks as an alternative to aggregative hierarchical cluster methods [Tamayo *et al.*, 1999; Törönen *et al.*, 1999; Herrero *et al.*, 2001]. Unsupervised neural networks, such as Self-Organising Maps (SOM) [Kohonen, 1997] or the Self-Organising Tree Algorithm (SOTA) [Dopazo and Carazo, 1997], provide a more robust framework appropriate for clustering large amounts of noisy data. Neural networks have properties that make them suitable for the analysis of gene expression patterns. They can deal with real-world data sets containing noisy, ill-defined items with irrelevant variables and outliers, and whose statistical distributions do not need to be parametric. Moreover, they are much faster and can easily be scaled to large data sets. Additionally, supervised methods like support vector machines (SVM) that are able to use prior information on the classes studied, have been applied to the analysis of functional classes of genes [Brown *et al.*, 2000].

On the other hand, clustering of samples has been used extensively for the classification of different types of cancers, where the molecular signature of the different tumoral tissues has been demonstrated to be a valuable diagnostic tool. Initial work has used classical hierarchical methods (see for example Alizadeh *et al.*, [2000]; Alon *et al.*, [1999]), but recent papers have proposed the use of supervised methods like SVM [Furey *et al.*, 2000] or supervised neural networks [Khan *et al.*, 2001]. Neural networks, in contrast to SVM, are able to discriminate amongst many different classes, and this is preferable for multi-class problems.

The objective of the present work is to compare the relative merits of different supervised and non-supervised clustering approaches for the classification of samples (here different cancer cell lines) based on their different gene signatures. A study of performance in terms of runtimes and accuracy of classification for classical and neural-network-based alternatives for clustering genes is given.

The problems of noise and non-informative gene expression profiles are also discussed. Here we give a combined approach in which the gene expression patterns are clustered into a reduced set of co-expressed genes,

and the clusters' average values are then used to train a supervised neural network. This approach provides superior accuracy of classification of samples when compared to the alternative unsupervised classification. The additional advantage of this approach is that the resulting entities used for the classification are not simple genes, but sets of co-expressed genes. Consequently, various data-mining techniques can be applied to assign some form of identity to them.

2. COMPARATIVE PERFORMANCES OF CLUSTERING METHODS

2.1 Data set used

The data set corresponding to the NCI-60 cancer cell lines with drug treatments [Scherf *et al.*, 2000] has been used. Gene expression levels were expressed as log(red/green). Ratios of fluorescence measurement were corrected by computational balancing of the two channels [Scherf *et al.*, 2000]. The data set includes expression values for 1,376 genes plus 40 assessed molecular targets for the drugs (a total of 1416 clones), in sixty different cell lines corresponding to nine different types of cancers.

2.2 Comparative runtimes

Since the analysis of DNA array data usually implies management of thousands of genes, the runtime of a method may constitute a real bottleneck for its application. Many of the classical methods used for clustering are based on iterative processing of a distance matrix obtained from "all-against-all" comparisons. If the most time-consuming operations are performed on such a distance matrix then runtimes must be at least proportional to the square of the number of items. This is the case for the family of aggregative hierarchical methods. Aggregative hierarchical methods, like average linkage and related methods, have runtimes in the range of N^2 to N^4 [Hartigan, 1975]. On the other hand, in the case of SOM or SOTA, the most time-consuming comparison operations are performed amongst the data and the nodes in the network (fixed in SOM and limited in SOTA at each step). The obvious advantage derived from this fact is that the number of comparisons needed for the classification depends principally on the number of items. Runtimes are therefore approximately linear [Dopazo *et al.*, 2001].

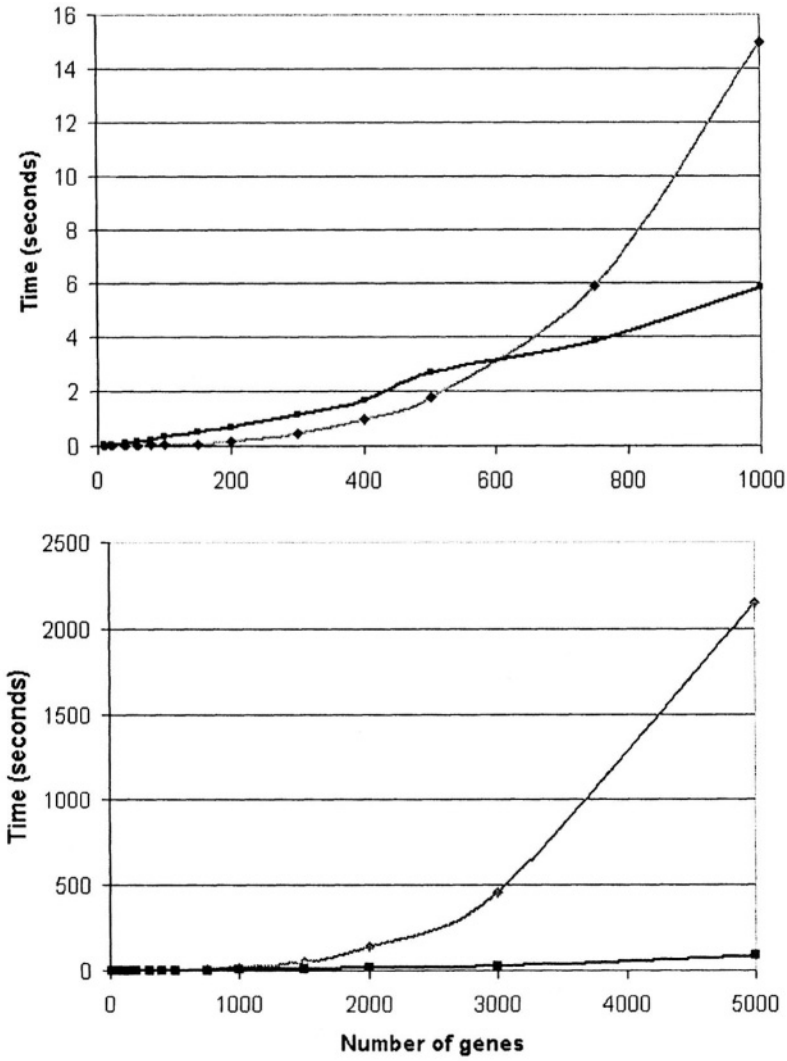


Figure 1. Comparison of runtimes of average linkage (grey line) and SOTA (black line). Top: detail of runtimes for up to 1000 genes, Bottom: runtimes for up to 5000 genes. The runtimes were obtained using an SGI Origin200. The data used were subsets randomly sampled from the complete data set of the study of gene expression in a synchronised cell cycle in yeast. [Eisen *et al.*, 1998].

When runtimes of both approaches are compared (see Figure 1) we can observe that average linkage is faster only when a few items (less than 600, see the top part of Figure 1) are to be analyzed. Otherwise, SOTA is clearly faster. Average linkage runtime is, at least, quadratic, whereas SOTA runtime is approximately linear. SOM behaviour (data not shown) is linear too.

2.3 Comparative accuracy

The silhouette statistic [Hand, 1981] was used to study the accuracy of the classification obtained by using each of the various methods. Silhouette measures how well the items are assigned to their corresponding clusters by comparing the distance from each item to the centre of its cluster, with the distance of the item to the centre of the closest cluster. The silhouette statistic is therefore defined for a cluster A as:

$$s(A) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}$$

where a(i) is the average of all the distances within cluster A, b(i) is the minimum of the distances $d(x_i,B)$, this distance being the average distance of the element x_i in cluster A to all the elements in cluster B, the nearest cluster to A. In this case, the higher the statistic the better.

Figure 2 shows how SOTA performs better than average linkage, irrespective of the split criterion used (see Herrero *et al.*, [2001]). The criterion used by SOTA to decide whether a node should be further divided so as to go on to a higher resolution relies on the intra-profile distances of the genes within the node. Due to this, SOTA is able to recover all the different patterns of expression profiles present in the data set analyzed, no matter how many genes display each pattern. In this aspect SOTA is superior to SOM, which holds more neurons for the more populated patterns. This is because SOM clustering is density-dependent [Kohonen, 1997], and this is not a desirable property when the aim is to discover all the different types of patterns present in the data.

SOTA and average linkage are always superior to SOM in terms of accuracy (data not shown).

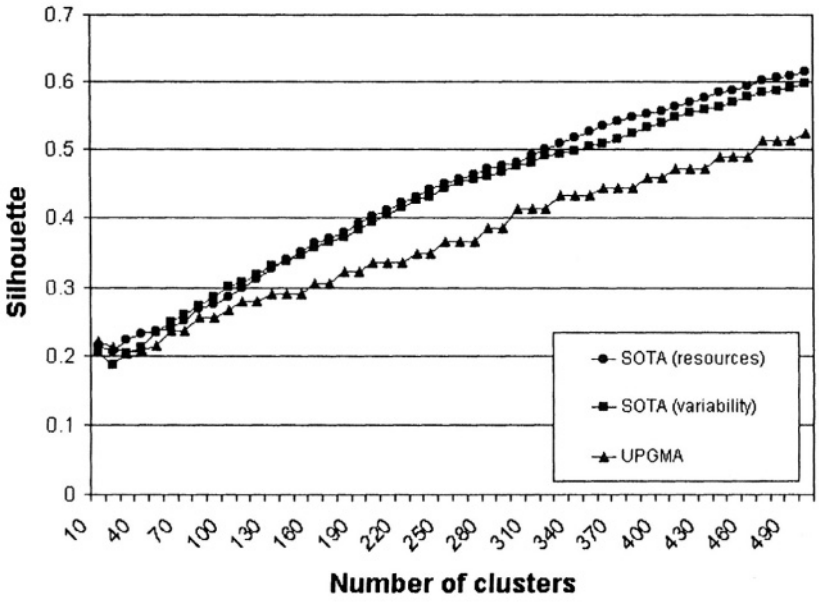


Figure 2. Silhouette statistic for the aggregative hierarchical method UPGMA (average linkage) and SOTA with two different criteria for growing the tree. In SOTA resources (circles), the decision for splitting a node is based on the mean distance to the average value of the cluster (centroid). In SOTA variability (squares) this decision is based on the maximum of the distances between all the genes in the cluster and the cluster's centroid.

2.4 Conclusions on comparative performances

Obviously the benchmark given here is far from exhaustive, but it focuses on the comparative efficiency of: (i) average linkage, as one of the most commonly-used distance matrix-based methods; (ii) SOM, as the alternative based on a neural network; and (iii) SOTA, as a hierarchical version of SOM, based on a growing topology [Dopazo and Carazo, 1997]. From the point of view of efficiency of the classification (as measured by the silhouette statistic), SOTA performs a little better than average linkage. SOTA also presents additional advantages with respect to runtime. Both SOTA and SOM presents an additional advantage: they can deal in a natural way with missing values that frequently occur in the DNA array data sets. Since the comparison operations are performed amongst the data and the average profiles in the nodes, the absence of some points (missing values) in a vector corresponding to a particular gene expression profile will have a negligible effect on the whole process of the network training. This avoids the use of methods for estimating missing values [Troyanoskaya *et al.*, 2001]

necessary if average linkage or similar methods are used. Table 1 summarizes some properties of the methods compared.

Table 1. Comparison of properties of Average linkage, SOM and SOTA.

Property	Average linkage	SOM	SOTA
Hierarchical	Yes	No	Yes
Growing	Aggregative (from tips to root)	Fixed size	Divisive (from root to tips)
Resolution	Full	Number of clusters fixed beforehand	Configurable level
Statistical definition of cluster	No	No	Built-in
Noise effect	Sensitive	Robust	Robust
Unequal cluster size effect	Sensitive	Very sensitive	Robust
Accept missing values	No	Yes	Yes
Runtime	$>N^2$	N	~N

3. CLUSTERING OF CONDITIONS

3.1 The problem of noisy patterns

The unequal distribution of genes among clusters is not only a problem from the point of view of clustering of gene expression patterns. In some cases it can even be a problem for the classification of conditions. Attempts at classifying cancer types based on the molecular signatures provide a clear example of this. Depending on the composition of a particular DNA microarray, there will be a number of genes that will display high or low expression values depending on the physiological circumstances of the patient, and are uncorrelated with the type of the cancer. They therefore introduce noise into the classification because they tend to produce an alternative clustering unrelated to the class-based clustering that is sought. If noisy classes are abundant and overpopulated they may make classification extremely difficult by drastically reducing the signal component in the data set.

Since the relative composition of clusters has no clear biological meaning, a significant part of the contribution of noise to the final classification could be removed if the average patterns of the gene

expression profiles were used for classification of the conditions, instead of using the gene profiles themselves.

3.2 Clustering of conditions and noise reduction

In recent work, a perceptron was trained to identify four different round blue cell tumours [Khan *et al*., 2001]. Principal component analysis (PCA) was used to reduce the number of items to analyze. One reason for this reduction is the consequent reduction in the number of parameters that the perceptron has to infer from the data, which depends on the size of the input layer (in this particular case this would be the number of genes). Generally speaking, fewer parameters means more generalisation power in the network. However, the quality of the input data was further improved by extracting from the data the components with most variance. Nevertheless, using this approach the biological meaning of the entities analyzed is then lost to some extent.

In the approach proposed here, the data are first clustered at gene level. A perceptron is then trained using the average values of the clusters found. We can study the accuracy of the classification obtained for the cell lines by using data clustered at different levels of resolution.

Figure 3 shows the learning rate of the perceptron when trained with the data set clustered at different levels of resolution. Learning rates at each point were obtained using "leave-one-out" validation. This means that for each point, sixty perceptrons with 59 cell lines are trained, and the remaining cell line is used as a test for studying the predictive ability of the perceptron. The value at each point is the number of cell lines properly classified in each of the sixty different training processes. The number of clusters (patterns) used for training the perceptron range from 1,416 (the actual number of clones) to only 13 different patterns.

The predictive power of the perceptron is very low if the training is performed with a high number of elements in the input layer. This reflects a combination of a high level of noise and the problem of overtraining of the network, due to the high number of parameters that are learned from the data. As the number of patterns approaches an optimum value for the learning process of the network, the number of true positives increases. When the optimum value of 161 patterns (in this case) is reached, the perceptron is able to identify 43 out of the sixty cell lines. For a number of patterns below this optimum value, the predictive power decreases, although performance is not as poor as when a number of profiles over the optimum value is used. For a small number of patterns, the performance of the perceptron is much better in terms of generalisation (opposed to overtraining), because the number of parameters to be learned from the data

set is lower. On the other hand, the information content decreases when various different clusters are collapsed into a single cluster.

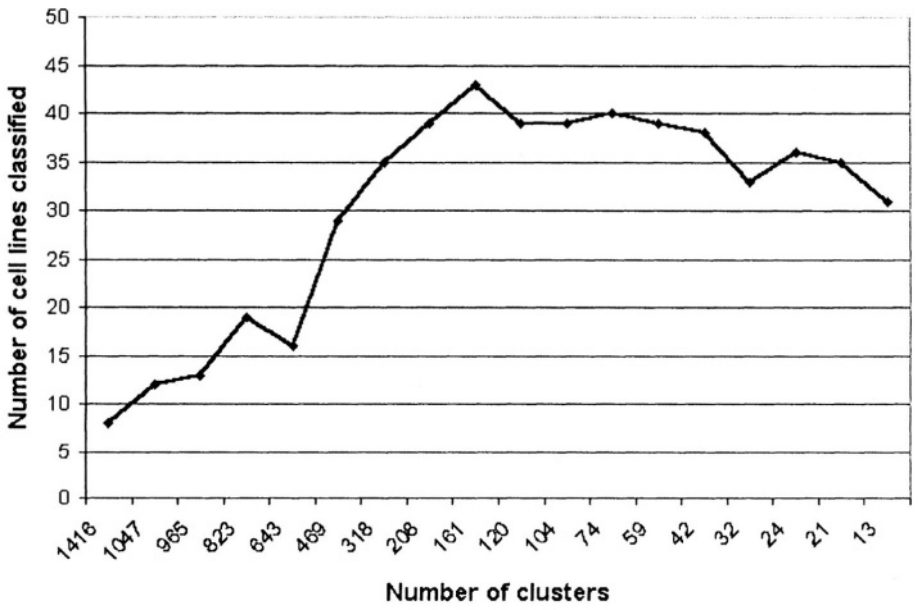


Figure 3. Ratio of success in the classification of cell lines at different levels of resolution.

Table 2 shows how the perceptron performs as well as or better than an unsupervised classification obtained by SOTA (or average linkage, data not shown), using the same data set. When all the gene expression profiles are used the unsupervised classification is still inferior to the supervised classification, except in the case of the breast cancer line. For some reason, the information that leads to the proper classification of this cell line is best represented in the original, unclustered set of profiles. Nevertheless, the ratio of success in the classification is still too low, and this putative increase in the efficiency of the classification might just be an artefact.

Table 2. Comparison performances of supervised vs unsupervised classifications.

Cell line	Total	Supervised	Unsupervised (161)	Unsupervised (1,416)
Breast	8	2	2	4
Melanoma	8	7	7	7
Prostate	2	0	0	0
Renal	8	7	6	7
Lung	8	5	3	3
CNS	6	4	4	4
Ovary	6	5	3	3
Leukemia	6	6	6	5
Colon	7	7	7	7

Figure 4 shows the unsupervised classification of the different cell lines obtained by SOTA using (Figure 4A) the patterns corresponding to the 161 clusters with optimal information content, and (Figure 4B) the original 1,416 gene expression profiles. In both cases the tree was grown up to nine nodes, corresponding to the nine different cell lines. The relative position in the tree changes for some cell lines, but the efficiency in the classification is slightly better in the case of the 161 patterns. Some of the cell lines are well defined, such as colon cancer or renal carcinoma, but others cannot properly be discriminated, and the number of false positives is high too. The results obtained with average linkage are very similar (data not shown).

In other experimental systems (e.g. [Alizadeh *et al.*, 2000; Alon *et al.*, 1999]) where each phenotype can easily be discriminated (data not shown), this approach performs considerably better.

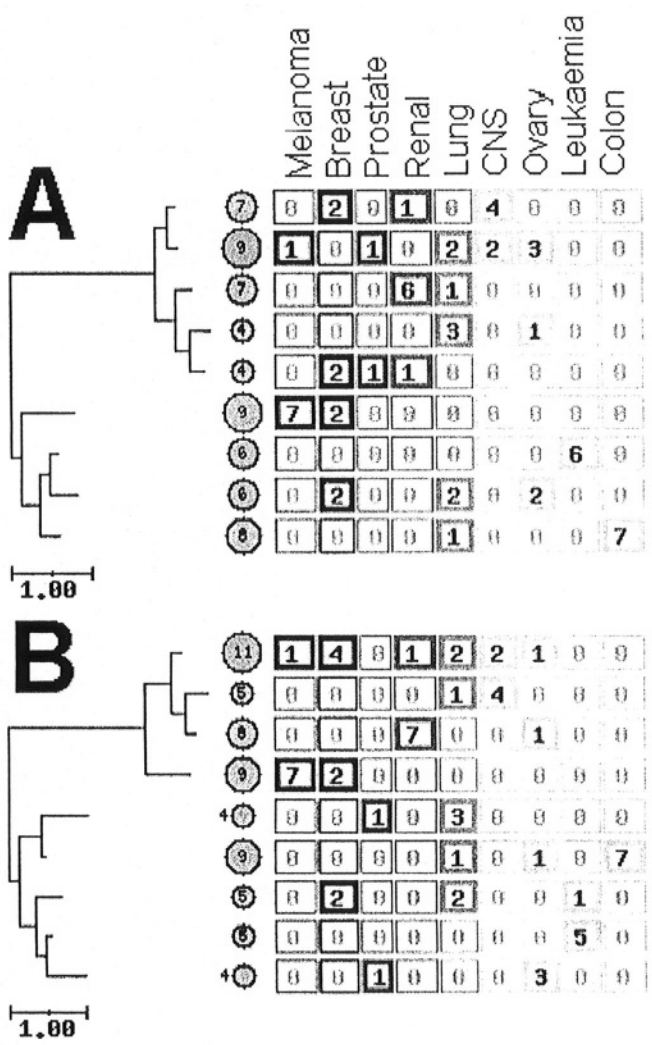


Figure 4. Unsupervised clustering of the 60 cell lines into nine clusters by SOTA using as data: **A** 161 patterns and **B** all 1,416 gene profiles. The number of cell lines in each cluster is represented by a circle of proportional size. The numbers in the squares represent the number of cell lines of each type in the corresponding cluster.

4. CONCLUSIONS

SOTA is a clustering method with linear runtimes and superior accuracy compared to its widely-used counterparts, average linkage and SOM. The possibility provided by the method for obtaining clusters of co-expressed genes at different levels of the hierarchy can be used to study the relative information content of this hierarchy at different levels. When different "slices" of the hierarchy are used to produce a classification of samples based on the average values of the gene expression profiles of the various clusters at different levels, we found that there exists an optimal information level at which the classification obtained is the best one possible. An explanation for this is that the optimal information level corresponds to the best signal-to-noise ratio in the data when these are subject to a process of compression based on the divisive segregation produced by SOTA. The classification can be improved by using a supervised method, such as a perceptron. Unlike other approaches, such as using PCA for reducing the dimensionality of the data [Khan *et al.*, 2001], the classification obtained here depends upon groups of co-expressed genes, which probably play a related role in the cell.

Web interfaces to the programs used in the present work (SOTA, SOM and average linkage) can be found on the web server at: http://bioinfo.cnio.es/dnarray/analysis/

ACKNOWLEDGEMENTS

JH is supported by a CNIO fellowship. AM is supported by an IBM fellowship.

REFERENCES

Alizadeh, AA, Eisen, MB, Davis, RE, Ma, C, Lossos, IS, Rosenwald, A, Boldrick, JC, Sabet, H, Tran, T, Yu, X, Powell, JI, Yang, L, Marti, GE, Moore, T, Hudson, J Jr., Lu, L, Lewi,s DB, Tibshirani, R, Sherlock, G, Chan, WC, Greiner, TC, Weisenburger, DD, Armitage, JO, Warnke, R, Levy, R, Wilson, W, Grever, MR, Byrd, JC., Botstein, D, Brown, PO, Staudt LM. Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature* 403 (2000):503-511

Alon, U, Barkai, N, Notterman, DA, Gish, K., Ybarra, S, Mack, D, Levine, AJ. Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed with oligonucleotide arrays. *Proc Natl Acad Sci USA* 96 (1999): 6745-6750.

Brown, PO, Botsein, D. Exploring the new world of the genome with DNA microarrays. *Nat Biotechnol* 14 (1999): 1675-1680.

Brown, MPS, Grundy, WN, Lin, D, Cristianini, N, Sugnet, CW, Furey, TS, Ares, M, Haussler, D. Knowledge-based analysis of microarray gene expression data using support vector machines. *Proc Natl Acad Sci USA* 97 (2000): 262-267.

Dopazo, J, Carazo, JM. Phylogenetic reconstruction using a growing neural network that adopts the topology of a phylogenetic tree. *J Mol Evol* 44 (1997): 226-233.

Dopazo, J, Zanders, E, Dragoni, I, Amphlett, G, Falciani, F. Methods and approaches in the analysis of gene expression data. *J. Immunol Meth* 250 (2001): 93-112.

Efron, B, Tibsirani, R. Statistical data analysis in the computer age. *Science* 253 (1991): 390-395.

Eisen, M, Spellman, PL, Brown, PO, Botsein, D. Cluster analysis and display of genome-wide expression patterns. *Proc Natl Acad Sci USA* 95 (1998): 14863-14868.

Furey, TS, Cristianini, N, Duffy, N, Bednarski, DW, Schummer, M, Haussler, D. Support vector machine classification and validation of cancer tissue samples using microarray expression data. *Bioinformatics* 16 (2000): 906-914.

Hand, DJ. *Discrimination and classification,* NY: Wiley, 1981.

Hartigan, JA. *Clustering algorithms.* New York: Wiley, 1975.

Herrero, J, Valencia, A, Dopazo, J. A hierarchical unsupervised growing neural network for clustering gene expression patterns. *Bioinformatics* 17 (2001): 126-136.

Khan, J, Wei, JS, Ringnér, M, Saal, LH, Ladanyi, M, Westermann, F, Berthold, F, Schwab, M, Antonescu, CR, Peterson, C, Meltzer, PS. Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. *Nature Med 7* (2001): 673-579.

Kohonen, T. *Self-organizing maps.* Berlin: Springer-Verlag, 1997.

Scherf, U, Ross, DT, Waltham, M, Smith, LH, Lee, JK, Tanabe, L, Kohn, KW, Reinhold, WC, Myers, TG, Andrews, DT, Scudiero, DA, Eisen, MB, Sausville, EA, Pommier, Y, Botstein, D, Brown, PO, Weinstein, JN. A gene expression database for the molecular pharmacology of cancer. *Nat Genet* 24 (2000): 236-44.

Tamayo, P, Slonim, D, Mesirov, J, Zhu, Q, Kitareewan, S, Dmitrovsky, E, Lander, ES, Golub TR. Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation. *Proc Natl Acad Sci USA* 96 (1999): 2907-2912.

Törönen, P, Kolehmainen, M, Wong, G, Castrén, E. Analysis of gene expression data using self-organizing maps. *FEBS letters* 451 (1999): 142-146.

Troyanskaya, O , Cantor, ML , Sherlock, G , Brown, P, Hastie, T, Tibshirani, R , Botstein, D, Altman, RB. Missing value estimation methods for DNA microarrays. *Bioinformatics* 17 (2001): 520-525.

Wen, X, Fuhrman, S, Michaels, GS, Carr, DB, Smith, S, Barker, JL, Somogyi, R. Large-scale temporal gene expression mapping of central nervous system development. *Proc.Natl Acad Sci USA* 95 (1998): 334-339.

This page intentionally left blank

7

BAYESIAN DECOMPOSITION ANALYSIS OF GENE EXPRESSION IN YEAST DELETION MUTANTS

Ghislain Bidaut[1,2], Thomas D. Moloshok[1], Jeffrey D. Grant[1], Frank J. Manion[1], and Michael F. Ochs[1*]

[1]Biomedical Informatics, Fox Chase Cancer Center, Philadelphia, PA 19111, [2]Structural and Genetic Information Laboratory, CNRS-AVENTIS, Marseille, France

Abstract: Many methods have been proposed for the analysis of microarray data. In general, these methods are borrowed from statistics and data mining, and they ignore the underlying biology that gives rise to the data. Biological systems, such as cells, are complex, with constant activation and deactivation of multiple pathways in response to external and internal stimuli. Thus, of particular concern is the failure of many analysis methods to allow expression levels for a single gene to be explained as arising from multiple, different stimuli. Bayesian Decomposition, originally developed for spectral mixture analysis, overcomes this problem by permitting the discovered patterns within the expression data to overlap, allowing genes to belong to multiple groups. We present results of the application of Bayesian Decomposition to the deletion mutation data, demonstrating its ability to assign genes that are regulated by multiple pathways to multiple coexpression groups, allowing identification of changes to specific signalling pathways.

Key words: Microarray, gene expression, Bayesian methods, cellular signalling

* Author to whom correspondence should be addressed

1. INTRODUCTION

1.1 The Development of Cancer

Human cancer is the second leading cause of death in the United States and throughout the Western world [Alison *et al.,* 1997]. Unlike heart disease and diabetes, the fundamental cellular biology underlying the development of cancer is poorly understood, at least partly because cancer arises from a myriad of different cellular malfunctions [Cooper, 1992; Macdonald *et al.,* 1997]. In order to understand cancer development in individual malignancies, the recovery of the process that led to the specific cellular malfunction present in the cancer cells must be identified. A key feature of such development must involve the cellular signalling pathways and metabolic pathways that control cell growth, differentiation, apoptosis (programmed cell death), and motility. Recovery of pathway information is presently possible only through complex experiments [Winzeler *et al.,* 1999; Hughes *et al.,* 2000], however new technologies such as microarrays and gene chips offer the possibility of more quickly and cheaply determining such pathways.

Pathway information is critical not only to the understanding of cancer development, but also to the design of effective therapeutics in the treatment of cancer. Present cancer treatments, such as radiotherapy and chemotherapy, result in substantial collateral damage to healthy tissues. Targeted therapies would try to alter behaviour in a cell specific manner, affecting only cancer cells. The creation of these therapies will require a detailed understanding of how disrupting specific cellular pathways affect downstream events in cells and an understanding of the signalling and metabolic networks in order to avoid unintended side effects in treatment (e.g. disrupting a pathway in a healthy cell leading to damage to healthy tissues).

1.2 Microarray Measurements and Analysis

Recent advances in microarray and gene chip technology have led to large amounts of data that potentially could aid in the understanding of the cellular function and pathways involved in human disease. While studies have already shown that it is possible in some cases to identify disease states more accurately using mRNA expression profiles than can be done using classic pathology methods [Golub *et al.,* 1999; Alizadeh *et al.,* 2000; Zhang *et al.,* 2001], the complexity of the underlying biological systems is reflected in difficulties in data analysis. The gene expression and proteomic data sets are expected to dwarf present sequence data in complexity [Bittner *et al.,*

1999] leading to the need for computer science technology to recover the maximal information [Lockhart *et al.,* 2000; Young, 2000].

Many attempts have been made to apply standard data mining algorithms to discover patterns within gene expression array data. Many algorithms are variations on standard methods for analysing matrices, since array data typically takes the form of two dimensional sets of numbers (e.g. expression levels for many genes at different conditions). Methods applied include the use of standard statistical methods [Claverie, 1999; Alter *et al.,* 2000; Ideker *et al.,* 2000; Kerr *et al.,* 2000; Kerr *et al.,* 2001], self-organising maps [Tamayo *et al.,* 1999], support vector machines [Brown *et al.,* 2000], and clustering [Eisen *et al.,* 1998; Getz *et al.,* 2000; Kerr *et al.,* 2001; Lukashin *et al.,* 2001; Yeung *et al.,* 2001], among other methods, reviewed by Brazma and Vilo [Brazma *et al.,* 2000]. Recently new statistical methods that maintain non-Euclidean relationships during reduction of the dimensionality of the data space have been reported [Roweis *et al.,* 2000; Tenenbaum *et al.,* 2000], which may help in defining relationships in complex expression data. In general, these methods do not incorporate knowledge of the underlying biological system, although there are methods that do take into account some experimental information [Heyer *et al.,* 1999]. However all these methods still lack an ability to recover fundamental behaviour, since each gene within the expression experiment can be assigned to only one coexpression group. This violates the underlying biological fact that many individual genes are coexpressed in multiple groups in response to different stimuli [Roberts *et al.,* 2000]. This fundamental flaw limits the usefulness of most algorithms as it leads inevitably to the loss of information related to behaviour arising from multiple inputs, which is critical for understanding cellular behaviour [Bittner *et al*., 2000].

Originally developed for use in multidimensional spectral imaging [Ochs *et al.,* 1999], Bayesian Decomposition is a matrix factorisation method that identifies physically meaningful basis vectors (patterns) simultaneously with their distributions in a data set. The basis vectors need not be orthogonal as in principal component analysis or obey other independence criteria. As used here, the method is similar to nonnegative matrix factorisation [Lee *et al.,* 1999], however Bayesian Decomposition allows negative basis vectors with properly encoded prior information (see below).

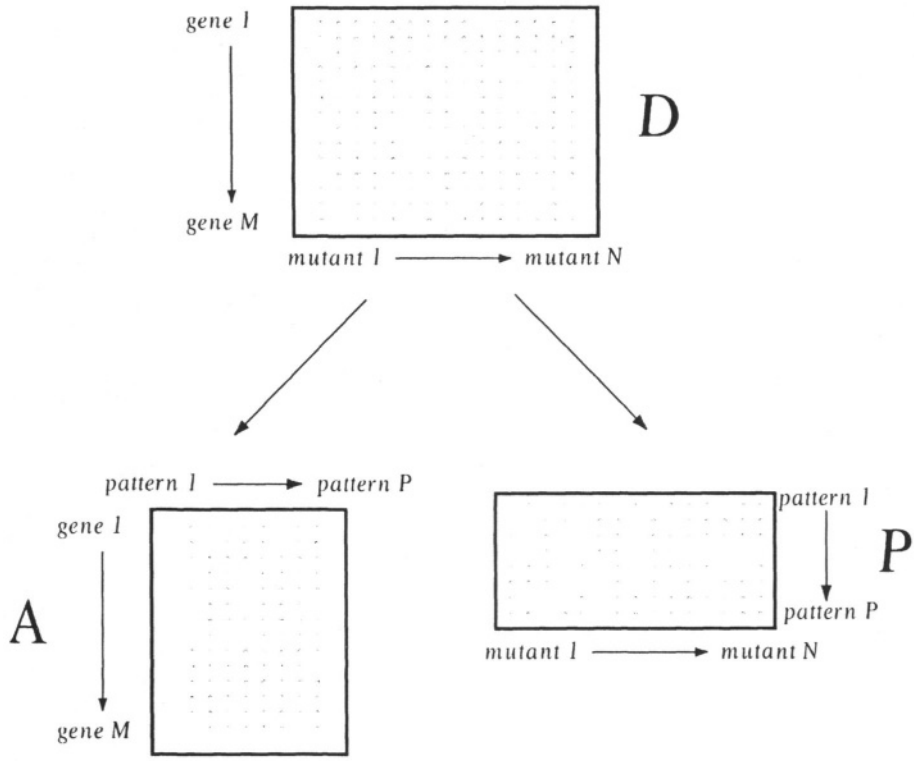


Figure 1. The matrix decomposition performed by Bayesian Decomposition. The data can be viewed as a matrix (***D***). The goal is to identify the matrices ***A*** and ***P***, the distribution and pattern (basis vector) matrices respectively, where the patterns have some physical or physiological meaning. The mock data (***M***) is the data that would result from the model (***A*** and ***P***) in the absence of noise.

2. METHODS

2.1 Bayesian Decomposition

The fundamental decomposition performed by Bayesian Decomposition is the recovery of a distribution matrix (***A***) and a pattern matrix (***P***) that combine to form a mock data matrix (***M***) that reproduces the data matrix (***D***) within the noise. This relationship can be written as

$$D \approx M = AP, \qquad [1]$$

and is shown diagrammatically in Fig. 1. This factorisation of $\boldsymbol{D}$ into $\boldsymbol{A}$ and $\boldsymbol{P}$ is generic, but can be specialised to specific situations by the incorporation of additional information in the analysis. Unlike a statistical approach, such as principal component analysis (PCA), the rows of $\boldsymbol{P}$ do not need to be orthogonal or fulfil other independence criteria. This allows the algorithm to model biological and physical systems, which typically have underlying processes that are nonindependent. For example, gene regulation through the cell cycle would have genes upregulated in G1 with tails of expression remaining in the S phase. Different rows of the $\boldsymbol{P}$ matrix would have peaks of expression in different phases, but overlaps between the rows would be present (i.e. the rows are nonorthogonal). Meanwhile, each column of $\boldsymbol{A}$ would contain significant values for genes upregulated within a single phase. Some rows of $\boldsymbol{A}$ would contain significant values in multiple columns, indicating peaks of expression in multiple parts of the cell cycle.

Since in the general problem, neither $\boldsymbol{A}$ nor $\boldsymbol{P}$ are known and there are no independence criteria, the problem is mathematically degenerate making an analytical solution impossible. However, a Markov chain Monte Carlo (MCMC) procedure can be used to sample the space of possible solutions (posterior distribution) to determine its properties, which provides a mean solution and uncertainty estimates. Furthermore, multiple possible solutions can be identified if supported by the data. The application of MCMC to stochastic image processes was initially demonstrated by Geman and Geman [Geman *et al.,* 1984] leading to exploration of a wide variety of sampling procedures [Metropolis *et al.,* 1953; Hastings, 1970; Kirkpatrick *et al.,* 1983] for solution of imaging problems, reviewed by Besag *et al.* [Besag *et al.,* 1995].

MCMC techniques require relative probability measurements at each sampled point in the solution space, which is provided here through a Bayesian approach. In the past decade Bayesian methods using MCMC techniques have been used in a wide variety of problems in data analysis, e.g. medical imaging, agricultural field studies, population studies, and economic forecasting [Besag, 1986; Besag *et al.,* 1993; Grenander *et al.,* 1994; Hill, 1994]. Bayesian statistical analysis starts with the apparently trivial statement,

$$p(Model, Data)\begin{cases} = p(Model \mid Data)p(Data) \\ = p(Data \mid Model)p(Model) \end{cases} \qquad [2]$$

where $p(Model,Data)$ is the probability of both the model and the data (the joint probability distribution), $p(Model|Data)$ is the conditional probability of the model given the data (the posterior), $p(Data)$ is the probability of the

data (the evidence), $p(Data\backslash Model)$ is the conditional probability of the data given the model (the likelihood), and $p(Model)$ is the probability of the model (the prior). The posterior distribution is the solution space for our problem, since it measures the probability of the present model (sample) in light of the data. Rearrangement of Eqn. 2 yields the posterior,

$$p(Model \mid Data) = \frac{p(Data \mid Model)p(Model)}{p(Data)}, \qquad [3]$$

which provides the MCMC algorithm with probabilities for determining steps during the sampling process. Since the evidence, $p(Data)$, usually acts as a scaling parameter, it can be ignored in this case since MCMC only needs relative probabilities. This means that the relative probability between points in the solution space is determined completely by the likelihood, which is easily determined by comparing the model to the data, and the prior, which is the probability of the model independent of the data. The prior allows for inclusion of domain knowledge about the problem under study (e.g. known coexpression).

Putting in the matrices $\boldsymbol{A}$ and $\boldsymbol{P}$ for the model leads to the specific form of Bayes' equation (Eqn. 3) for the bilinear problem (ignoring the scale factor),

$$p(A, P \mid D) \propto p(D \mid A, P)p(A, P). \qquad [4]$$

The sampling from the posterior distribution and the encoding of the prior are done using a bilinear form of the Massive Inference™ Gibbs sampler from Maximum Entropy Data Consultants (Cambridge, England) that also enforces positivity on the solutions. The sampler also encodes a prior probability distribution, $p(\boldsymbol{A},\boldsymbol{P})$. This is done by creating multiple domains with mappings between them.

The first domain (top line of Fig. 2) is an atomic domain that contains two infinitely divisible ($\boldsymbol{2^{32}}$ points *in silico)* one dimensional spaces (one corresponding to the $\boldsymbol{A}$ matrix and one to the $\boldsymbol{P}$ matrix) in which atoms (point masses) exist. These atoms are created and destroyed in accordance with a prior distribution comprising a uniform spatial distribution and a logarithmic flux distribution [Sibisi *et al.*, 1997]. This prior tends to remove atoms that are not forced to exist by the data, yielding a minimal structure in the model. The prior in this space also enforces positivity and additivity of atoms, which effectively reduces the search space for the sampler by a factor of 2^N, where N is the number of dimensions. The second domain is the model domain (bottom of Fig. 2) that contains the $\boldsymbol{A}$ and $\boldsymbol{P}$ matrices. The

transformation of atoms from the atomic domain to the model domain permits the inclusion of prior knowledge in the form of convolution functions (*f*'s in Fig. 2). These convolution functions encode prior knowledge by forcing correlations between elements in the ***A*** and ***P*** matrices. For example, an atom in the ***P*** atomic domain may be constrained to be an mRNA abundance curve with a known rise time and half-life, thereby making each row of ***P*** a measure of a type of transcriptional response. In addition the convolution functions can create negative values within the model, so long as the underlying atomic distributions remain positive and additive.

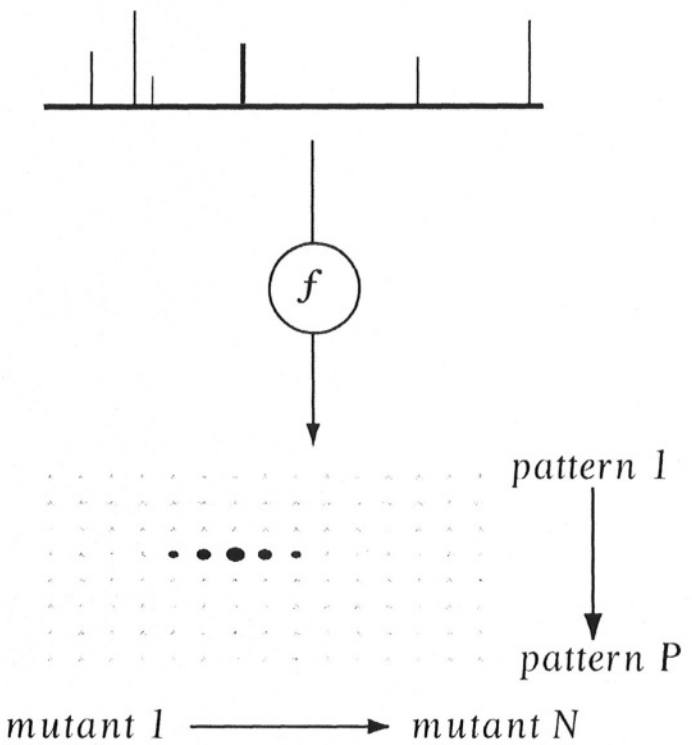


Figure 2. The atomic and model domains used by Bayesian Decomposition. The atomic domain (only the domain for the patterns shown here) contains an infinitely divisible line. Atoms (point masses) are created and placed onto the line. Each atom (for instance, the one in bold) is then mapped to the model domain (***A*** and ***P*** matrices) by a convolution function (*f*) that can distribute their flux in simple or complicated ways (above it is spread nonuniformly to five matrix elements as shown by the spots).

Such convolution functions were used to model inversion recovery curves [Ochs *et al.,* 2001], which have inherently negative components, allowing analysis in magnetic resonance relaxographic imaging [Labadie *et al.,* 1994].

Once the prior is chosen, the remainder of the problem is straightforward. A random model is generated in accordance with the prior as the starting point for the Markov chain. A change to this model is generated according to the prior, and a new likelihood is calculated using the sum of the squares of the residuals normalised by the standard deviation, σ, of the noise in the data, i.e. a normalised χ^2 distribution. Rather than calculate the full likelihood at each point, the change in the likelihood is calculated for the specific change in the model, so that the likelihood can be updated incrementally. The log likelihood, L, can be written in matrix notation as

$$L = \frac{1}{2\sigma^2} Tr\left[(AP - D)^T (AP - D)\right]. \qquad [5]$$

where $\boldsymbol{A}^T$ represents the transpose of $\boldsymbol{A}$ and Tr indicates the trace of the quantity in the brackets. The noise (σ) has been assumed to be equal at all points in order to simplify the form of Eqns. 5 and 6 (the actual calculation is done allowing independent noise estimates for each data point). The behaviour of the change in the log likelihood, ΔL, can be derived by looking at the effect of adding a small amount of flux, $\delta \boldsymbol{P}$, to the model. By inserting $\boldsymbol{P}+\delta\boldsymbol{P}$ for $\boldsymbol{P}$ in Eqn. 5 and subtracting Eqn. 5 from the result, the change in log likelihood is

$$\Delta L(\delta P) = \frac{1}{2\sigma^2} Tr\begin{bmatrix} (A\,\delta P)^T (AP - D) \\ +(AP - D)^T A\,\delta P + (A\,\delta P)^T (A\,\delta P) \end{bmatrix} \qquad [6]$$

where it is assumed that only changes to $\boldsymbol{P}$ are made. A similar equation governs calculations for changes in the model for $\boldsymbol{A}$. In order to simplify the calculations, we do not allow simultaneous changes in $\boldsymbol{A}$ and $\boldsymbol{P}$, since allowing such changes would require evaluation of terms involving $\delta\boldsymbol{A}\delta\boldsymbol{P}$. Note that barring such changes does not prevent the system from reaching any state and should have no effect on the final result, since the sampler can move $\delta\boldsymbol{P}$ followed by $\delta\boldsymbol{A}$ and reach the same point. As long as detailed balance is maintained, the sampler still samples the space correctly. For each step of the Markov chain, a random change is generated in the atomic domain in accordance with the prior. The algorithm then calculates the change in the likelihood using Eqn. 6 and determines whether to accept this change by comparing this with a randomly generated value. If the step is

taken, the likelihood is updated. MCMC samplers require a "burn-in" time to reach an area of high probability that is suitable for sampling. The sampler runs for an operator-specified time without recording samples and then continues while recording for an equal number of steps.

2.2 Issues in the Application of Bayesian Decomposition

One difficulty in Bayesian Decomposition (BD) analysis is the estimation of the correct dimensionality of the model space (i.e. the number of columns of $\boldsymbol{A}$ and rows of $\boldsymbol{P}$). This can be provided by estimation of the number of dimensions needed to fit the data through statistical analysis (such as PCA) or by making multiple runs with BD using different estimates. For many analyses, specific features appear indicating when the number of patterns is excessive. In spectroscopic studies and time domain modelling, these are often the emergence of patterns that appear to be unrelated to spectral features or to natural time behaviours. In data sets that have no likely correlated structures between points, such as the Rosetta data, the estimation of the dimensionality is more problematic.

Eqn. 1 is mathematically degenerate, so that multiple, analytical solutions exist. Bayesian Decomposition searches through these possible solutions for those that are most probable and samples the likely solutions. In general, if the number of elements in $\boldsymbol{D}$ is significantly larger than the number of elements in $\boldsymbol{A}$ and $\boldsymbol{P}$ combined or if there is a good mathematical model of the process underlying the generation of the data, these multiple solutions can be representations of the same solution, yielding a mean and standard deviation for each element of the matrix. In cases where there are multiple, significantly differing solutions, the sampler may move between these. This will generally yield significant uncertainties at the points that differ between the models, however BD saves snapshots of individual solutions that can be examined to verify that multiple solutions exist. In addition, by repeating the analysis using different random seeds in the Markov process, different Markov chains are generated and the results can be compared. This reduces the probability that the result obtained provides only one solution out of many that all fit the data equally well.

2.3 Application to the Rosetta Compendium

For gene expression analysis, each row of $\boldsymbol{D}$ represents the expression of a single gene with the columns representing different conditions (in this work, different deletion mutants). The matrix $\boldsymbol{M}$ would match the matrix $\boldsymbol{D}$ exactly if $\boldsymbol{A}$ and $\boldsymbol{P}$ were perfect models of the system and if there were no noise. The distribution matrix $\boldsymbol{A}$ contains rows that describe the amount of

each pattern within the corresponding row (gene) in ***D***, with each column being associated with a single pattern. The rows of ***P*** are the patterns that show the behaviour of the coexpressed genes among the deletion mutants (i.e. which coexpression groups are present in each mutant). Because of the symmetry of the decomposition, the columns of ***A*** also contain patterns. These patterns are the coexpressed genes within each mutational pattern from ***P***. This information can be used to infer pathway activation, suppression, and interaction in the deletion mutants, since many genes are known *a priori* to be transcribed in response to activation of specific pathways. Essentially the genes for which a great deal of information is known are used as guides for interpreting other genes in the coexpression groups. Previous work on the yeast cell cycle data sets [Cho *et al.,* 1998; Spellman *et al.,* 1998] has shown that the method can extract biologically significant, overlapping expression patterns from gene expression data [Moloshok *et al.,* In Press].

The deletion mutation data set was downloaded from Rosetta Inpharmatics and filtered to remove experiments where less than two genes underwent three-fold changes and to remove genes which did not change by three-fold across the remaining experiments. The resulting data set comprised 764 genes and 228 experiments. The Rosetta error model [Hughes *et al.,* 2000] provided the estimate of uncertainty in the data used in the calculation of the likelihood during sampling. Since Bayesian Decomposition presently is limited to mock data with positive values in gene expression analysis, all data were transformed from log ratios to ratios. PCA was used to estimate the dimensionality of the data (number of columns of ***A*** and rows of ***P***). Two "knees" appeared in the amount of variance explained by the principal components at three principal components and seven principal components. We focused our analysis on seven patterns. Multiple runs of Bayesian Decomposition were performed, each with the Markov chain process beginning at different, randomly generated points in the space of possible solutions. The individual patterns were cross-correlated between runs of the algorithm to identify those patterns that were consistent. Analysis of the data in terms of coregulation and pathway activation were focused on these consistent patterns. For each pattern, the genes that were significantly expressed in each pattern were identified (i.e. the amplitude of their assignment to the pattern was greater than three times the uncertainty in that amplitude). The largest amplitude genes in each pattern were analyzed in terms of cellular role as defined in the Yeast Database from Proteome [Costanzo *et al.,* 2000; Costanzo *et al.,* 2001] in order to assign a function to the pattern. The patterns with clear function were then validated by analysis of specific deletion mutants from the compendium.

3. RESULTS

3.1 Identification of the Patterns

The patterns (rows of ***P***) identified in the data were assigned functions based on the top scoring genes in each pattern (largest amplitude in the corresponding column of ***A***). Since cellular processes are complex and involve numerous activities, the assignment of a pattern to a cellular behaviour and thus to a signalling pathway or set of pathways is difficult. By using the Proteome database to identify cellular roles for the top 50 scoring genes in each pattern, a tentative cellular role for the pattern was defined.

The repeated runs of the Bayesian Decomposition algorithm were used to check the reproducibility of the identified patterns. It was determined that four of the patterns were tied consistently to specific deletion mutants, while three of the patterns showed more variation. This could be a result of the inherent structure in the data, allowing multiple methods of mixing together some of the patterns to explain the variation along certain directions in the multidimensional data set. It is also possible that the variation is an indication of a requirement for more basis vectors to explain the data, with each of the present, varying vectors actually being comprised of mixtures of these more fundamental, underlying vectors. However, the Rosetta error model already permits the normalised χ^2 to be unduly low, suggesting overfitting of the data. Although, this may instead indicate that the Rosetta error model has overestimated the actual error.

Of the consistent patterns, the first is clearly linked to amino acid metabolism with 22 genes linked to this function out of 36 of known function (cellular role in the Proteome database) in the top 50 scoring genes. The second is more difficult to identify as it has a mixture of metabolic, RNA processing, and DNA processing genes in the top 50 scoring genes. The third is also hard to identify, with 30 of 50 genes having no known function. The fourth is clearly linked to the mating response with 13 genes linked to mating response and 5 to meiosis out of 23 genes with known function in the top 50 scoring genes. In addition, 8 of the genes of mating function appear in the top 50 scoring genes of this pattern across all runs of Bayesian Decomposition.

The analysis allows a further confirmation of the roles of these patterns. By analysing the deletion mutants that have or lack the pattern, the assignment of the pattern to the specific cellular role can be verified. Clearly only the mating pathway allows this to be done easily, as the other stable patterns are not clearly linked to a signalling pathway. This is likely to be a result of the fact that not all deletion mutants affect signalling pathways, but

instead have effects on metabolic behaviour and other cellular processes. However, the focus of our application of the algorithm is on the ability to identify changes to signalling pathways from expression data, so we focus on the mating response.

3.2 Validation of a Pattern

In order to validate the mating response pattern, we explored deletion mutants related to the mating pathway [Posas *et al.,* 1998] by focusing on those specific mutants in the mating pattern (i.e. the amplitude for these mutants in the row of ***P*** related to the mating pattern). The mating response in *S. cerevisiae* is mediated via a MAPK signalling cascade initiated by binding to the *Ste2* or *Ste3* membrane receptors. The signal is transduced through *Ste11, Ste7,* and *Fus3* with *Ste5* serving as a scaffolding protein. The signal activates the *Ste12* transcription factor, leading to transcription of mating response genes. In addition, the signal is transduced to the MAPK cascade from the membrane by a G protein complex or through the *Ste20* protein. Fig. 3 shows the amount of the behaviour (amplitude normalized by column in matrix ***P***) of the overall gene expression attributable to the mating pathway in the experiments with the genes for these proteins knocked out. Note that in every case, the behaviour is exactly as would be expected, with the exception of the *Fus3* deletion mutant. This is because *Kss1* can substitute for *Fus3,* yielding a mating response in the absence of *Fus3* [Posas *et al*., 1998]. However, the double knockout *Kss1/Fus3* does show the loss of the mating response. The fact that the response stays active in the *Ste20* knockout is a result of the alternative activation of the signalling cascade directly by the G protein complex.

Similar graphs can be shown for the other patterns within the data, however interpretation is problematic since the function is difficult to deduce. Nevertheless, this example demonstrates the potential for Bayesian Decomposition to identify gene expression changes related to changes in signalling pathways, even in the midst of complex behaviour. In fact, the *Dig1*,*Dig2* dual deletion mutant scored highest in this pattern, with 37% of the behaviour explained by the mating response. This indicates that the gene expression pattern related to mating response was identified within a data set where no more than ~1/3 of any set of gene expression levels could be related directly to it.

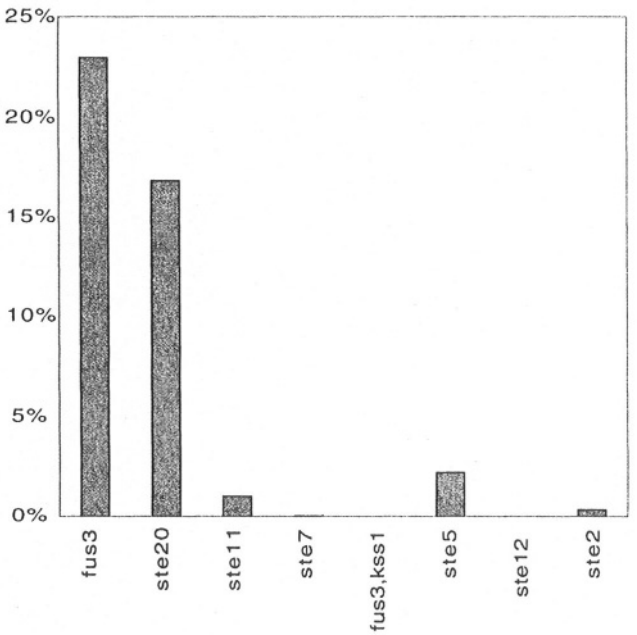


Figure 3. The percentage of the behaviour of an individual deletion mutant that is explained by the mating response based on gene expression.

4. CONCLUSIONS

The Rosetta compendium provides a vast amount of information for the study of gene expression in yeast. Analysis of this data presents special problems seen in few other data sets presently available. The effect of a single deletion can be extremely complex, due to the ability of yeast to provide similar function with different proteins, the ability of multiple separate pathways to yield similar transcription, and the inherent cross-talk present in signalling pathways within yeast.

The analysis presented here shows both the potential of and difficulties with identifications of pathway modifications from gene expression data. While some patterns were consistent across multiple runs of the algorithm, others varied significantly. This is probably a reflection of flexibility in the underlying model with multiple solutions of seven patterns being equally capable of explaining the expression profiles of the genes in the compendium. However, some patterns were consistent across multiple runs, allowing insight into some pathways. It is not necessarily immediately obvious what the function controlled by the pathway (or linked pathways) is for each of these patterns. This can be a result of the identified genes in the

coregulation group having unknown function or of a consistent cellular role being unidentifiable with our present knowledge.

Despite the difficulty of analysing data such as the Rosetta compendium, Bayesian Decomposition has successfully isolated the transcriptional response corresponding to the mating pathway. No single gene is transcribed solely in response to this pathway activation. Nevertheless, a group of coregulated genes has been identified related to activation of the pathway. Changing together, these genes provide a fingerprint of pathway activation. The identification of this pathway is verified by the fact that the mating pattern is absent in deletion mutants of proteins integral to the pathway (*Ste5, Ste7, Ste11, Ste12,* and *Fus3/Kss1*). In addition, the analysis identifies a series of genes that either knock out mating due to loss of critical cellular function or play an important role in the mating pathway, since they also have total loss of the mating fingerprint. One advantage of this method is that it is fairly tolerant of false positives in the data. Since each pattern consists of many genes showing significant expression, the role of the pattern is not determined by only a few genes. Since false positives should arise from random variations, it is unlikely that a pattern will be falsely assigned as it would require coordinated false positives.

A key unsolved question in the use of Bayesian Decomposition is the number of dimensions to use in the analysis. This is a different question from the number of expression units, as Bayesian Decomposition is attempting to identify a minimal set of basis vectors for the data within the constraints. These are unlikely to be identical to groups of genes that together provide some cellular function, but instead may be groups of cellular functions that within the experiment are activated simultaneously. As such, the number of patterns sought by Bayesian Decomposition may be significantly less than the number of groups required to map out cellular function. However, the patterns identified should be the number of independent sets of these groups of cellular functions, with each set being turned on or off together throughout the experiment being analyzed. Since this does not provide an *a priori* method to choose the number of patterns (i.e. to match a known or suspected number of cellular functions), other means need to be developed to identify the dimensionality. In this work PCA was used, however it is not a reliable tool for estimating dimensionality in gene expression data.

Bayesian Decomposition offers significant advantages for the analysis of microarray data. Often mere identification of coregulation is not of great interest, while identification of changes in pathway activation and behaviour is. This is logical since a broad spectrum of diseases is known to be induced by errors in proteins whose primary function is signal transduction (e.g. p53, abl, c-kit). Such errors may include loss of ATP binding sites, other

mutations, or full loss of function due to loss of heterozygosity. If such changes get reflected in levels of gene expression, which is overwhelmingly likely, then analysis of expression data with the goal of identifying modifications in pathway behaviour is an important step in identifying targets for treatment of disease. Bayesian Decomposition has been designed in order to untangle interacting signals in other areas, and it offers a promising method for doing the same in expression analysis, thus allowing recovery of pathway information. BD can also be used in combination with more complicated methods that attempt to identify signalling and metabolic networks, including reverse engineering of networks using various modelling techniques [D'Haeseleer *et al.,* 2000] or use of gene expression data to test existing network models [Hartemink *et al.,* 2001].

In the future we intend to encode additional prior knowledge into the analysis system. Such knowledge will include linkages between genes which are known to be coregulated through correlations within the columns of the $\boldsymbol{A}$ matrix. In addition, for time series data, we have recently introduced correlations in the rows of the $\boldsymbol{P}$ matrix corresponding to modelled rise times and half-lives for mRNA species. With these additions, Bayesian Decomposition should be able to discover significant information within gene expression data sets.

ACKNOWLEDGEMENTS

This work was supported by the National Institutes of Health, National Cancer Institute (CA06927 to R. Young, pilot grant to mfo under CA83638 to R. Ozols) and the Pew Foundation.

REFERENCES

Alison, M, Sarraf, C. *Understanding Cancer.* Cambridge: Cambridge University Press, 1997.

Alizadeh, AA, Eisen, MB, Davis, RE, Ma, C, Lossos, IS, Rosenwald, A, Boldrick, JC, Sabet, H, Tran, T, Yu, X, Powell, JI, Yang, L, Marti, GE, Moore, T, Hudson, J, Jr., Lu, L, Lewis, DB, Tibshirani, R, Sherlock, G, Chan, WC, Greiner, TC, Weisenburger, DD, Armitage, JO, Warnke, R, Staudt, LM *et al.* Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature* 403(6769) (2000): 503-11.

Alter, O, Brown, PO, Botstein, D. Singular value decomposition for genome-wide expression data processing and modeling. *Proc Natl Acad Sci U S A 97(18)* (2000): 10101-6.

Besag, J. On the statistical analysis of dirty pictures. *J. R. Statist. Soc. B* 48 (1986): 259-302.

Besag, J, Green, P, Higdon, D, Mengersen, K. Bayesian computation and stochastic systems. *Statistical Science* 10(1) (1995): 3-66.

Besag, J, Green, PJ. Spatial statistics and Bayesian computation. *J. R. Statist. Soc. B* 55 (1993): 25-37.

Bittner, M, Meltzer, P, Chen, Y, Jiang, Y, Seftor, E, Hendrix, M, Radmacher, M, Simon, R, Yakhini, Z, Ben-Dor, A, Sampas, N, Dougherty, E, Wang, E, Marincola, F, Gooden, C, Lueders, J, Glatfelter, A, Pollock, P, Carpten, J, Gillanders, E, Leja, D, Dietrich, K, Beaudry, C, Berens, M, Alberts, D, Sondak, V. Molecular classification of cutaneous malignant melanoma by gene expression profiling. *Nature* 406(6795) (2000): 536-40.

Bittner, M, Meltzer, P, Trent, J. Data analysis and integration: of steps and arrows. *Nat Genet* 22(3) (1999): 213-5.

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

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

Cho, RJ, Campbell, MJ, Winzeler, EA, Steinmetz, L, Conway, A, Wodicka, L, Wolfsberg, TG, Gabrielian, AE, Landsman, D, Lockhart, DJ, Davis, RW. A genome-wide transcriptional analysis of the mitotic cell cycle. *Mol Cell* 2(1) (1998): 65-73.

Claverie, JM. Computational methods for the identification of differential and coordinated gene expression. *Hum Mol Genet* 8(10) (1999): 1821-32.

Cooper, GM. *Elements of Human Cancer.* Boston: Jones and Bartlett Publishers, 1992.

Costanzo, MC, Crawford, ME, Hirschman, JE, Kranz, JE, Olsen, P, Robertson, LS, Skrzypek, MS, Braun, BR, Hopkins, KL, Kondu, P, Lengieza, C, Lew-Smith, JE, Tillberg, M, Garrels, JI. YPD, PombePD and WormPD: model organism volumes of the BioKnowledge library, an integrated resource for protein information. *Nucleic Acids Res* 29(1)(2001): 75-9.

Costanzo, MC, Hogan, JD, Cusick, ME, Davis, BP, Fancher, AM, Hodges, PE, Kondu, P, Lengieza, C, Lew-Smith, JE, Lingner, C, Roberg-Perez, KJ, Tillberg, M, Brooks, JE, Garrels, JI. The yeast proteome database (YPD) and Caenorhabditis elegans proteome database (WormPD): comprehensive resources for the organization and comparison of model organism protein information. *Nucleic Acids Res* 28(1) (2000): 73-6.

D'Haeseleer, P, Liang, S, Somogyi, R. Genetic network inference: from co-expression clustering to reverse engineering. *Bioinformatics* 16(8) (2000): 707-26.

Eisen, MB, Spellman, PT, Brown, PO, Botstein, D. Cluster analysis and display of genome-wide expression patterns. *Proc Natl Acad Sci USA* 95(25) (1998): 14863-8.

Geman, S, Geman, D. Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence PAMI* 6(6) (1984): 721-741.

Getz, G, Levine, E, Domany, E. Coupled two-way clustering analysis of gene microarray data. *Proc Natl Acad Sci USA* 97(22) (2000): 12079-84.

Golub, TR, Slonim, DK, Tamayo, P, Huard, C, Gaasenbeek, M, Mesirov, JP, Coller, H, Loh, ML, Downing, JR, Caligiuri, MA, Bloomfield, CD, Lander, ES. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science* 286(5439) (1999): 531-7.

Grenander, U, Miller, MI. Representations of knowledge in complex systems. *J. R. Statist. Soc.B* 56 (1994): 549-603.

Hartemink, AJ, Gifford, DK, Jaakkola, TS, Young, RA. Using graphical models and genomic expression data to statistically validate models of genetic regulatory networks. *Pac Symp Biocomput* (2001):422-33.

Hastings, WK. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika* 57 (1970): 97-109.

Heyer, LJ, Kruglyak, S, Yooseph, S. Exploring expression data: identification and analysis of coexpressed genes. *Genome Res* 9(11) (1999): 1106-15.

Hill, BM. Bayesian forecasting of economic time series. *Econometric Theory* 10 (1994): 483-513.

Hughes, TR, Marton, MJ, Jones, AR, Roberts, CJ, Stoughton, R, Armour, CD, Bennett, HA, Coffey, E, Dai, H, He, YD, Kidd, MJ, King, AM, Meyer, MR, Slade, D, Lum, PY, Stepaniants, SB, Shoemaker, DD, Gachotte, D, Chakraburtty, K, Simon, J, Bard, M, Friend, SH. Functional discovery via a compendium of expression profiles. *Cell* 102(1) (2000): 109-26.

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

Kerr, MK, Afshari, CA, Bennett, L, Bushel, P, Martinez, J, Walker, NJ, Churchill, GA. Statistical analysis of a gene expression microarray experiment with replication. *Statistica Sinica* (2001).

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

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

Kirkpatrick, S, Gelatt, CD, Vecchi, MP. Optimization by simulated annealing. *Science* 220 (1983): 671 - 680.

Labadie, C, Lee, J-H, Vetek, G, Springer CS Jr. Relaxographic imaging. *Journal of Magnetic Resonance B* 105(2) (1994): 99 - 112.

Lee, DD, Seung, HS. Learning the parts of objects by non-negative matrix factorization. *Nature* 401 (6755) (1999): 788-91.

Lockhart, DJ, Winzeler, EA. Genomics, gene expression and DNA arrays. *Nature* 405(6788) (2000): 827-36.

Lukashin, AV, Fuchs, R. Analysis of temporal gene expression profiles: clustering by simulated annealing and determining the optimal number of clusters. *Bioinformatics* 17(5) (2001): 405-14.

Macdonald, F, Ford, CHJ. *Molecular Biology of Cancer.* Oxford: BIOS Scientific Publishers Ltd., 1997.

Metropolis, N, Rosenbluth, A, Rosenbluth, M, Teller, A, Teller, E. Equations of state calculations by fast computing machines. *J. Chem. Phys.* 21 (1953): 1087-1091.

Moloshok, TD, Klevecz, RR, Grant, JD, Manion, FJ, Speier IV, WF, Ochs, MF. Application of Bayesian Decomposition to microarray data. *Bioinformatics* (In Press).

Ochs, MF, Stoyanova, RS, Arias-Mendoza, F, Brown, TR. A new method for spectral decomposition using a bilinear Bayesian approach. *J Magn Reson* 137(1) (1999): 161-76.

Ochs, MF, Stoyanova, RS, Brown, TR, Rooney, WD, Springer CS Jr. *A Bayesian Markov chain Monte Carlo solution of the bilinear problem.* Edited by JT Rychert, GJ Erickson, CR Smith. Bayesian Inference and Maximum Entropy Methods in Science and Engineering: 19th International Workshop. Melville: American Institute of Physics, 2001.

Posas, F, Takekawa, M, Saito, H. Signal transduction by MAP kinase cascades in budding yeast. *Curr Opin Microbiol* 1(2) (1998): 175-82.

Roberts, CJ, Nelson, B, Marton, MJ, Stoughton, R, Meyer, MR, Bennett, HA, He, YD, Dai, H, Walker, WL, Hughes, TR, Tyers, M, Boone, C, Friend, SH. Signaling and circuitry of multiple MAPK pathways revealed by a matrix of global gene expression profiles. *Science* 287(5454) (2000): 873-80.

Roweis, ST, Saul, LK. Nonlinear Dimensionality Reduction by Locally Linear Embedding. *Science* 290(5500) (2000): 2323-2326.

Sibisi, S, Skilling, J. Prior distributions on measure space. *Journal of the Royal Statistical Society, B* 59(1) (1997): 217-235.

Spellman, PT, Sherlock, G, Zhang, MQ, Iyer, VR, Anders, K, Eisen, MB, Brown, PO, Botstein, D, Futcher, B. Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization. *Mol Biol Cell* 9(12) (1998): 3273-97.

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

Tenenbaum, JB, Silva, VD, Langford, JC. A global geometric framework for nonlinear dimensionality reduction. *Science* 290(5500) (2000): 2319-2323.

Winzeler, EA, Shoemaker, DD, Astromoff, A, Liang, H, Anderson, K, Andre, B, Bangham, R, Benito, R, Boeke, JD, Bussey, H, Chu, AM, Connelly, C, Davis, K, Dietrich, F, Dow, SW, EL Bakkoury, M, Foury, F, Friend, SH, Gentalen, E, Giaever, G, Hegemann, JH, Jones, T, Laub, M, Liao, H, Davis, RW *et al.* Functional characterization of the S. cerevisiae genome by gene deletion and parallel analysis. *Science* 285(5429) (1999): 901-6.

Yeung, KY, Haynor, DR, Ruzzo, WL. Validating clustering for gene expression data. *Bioinformatics* 17(4) (2001): 309-18.

Young, RA. Biomedical discovery with DNA arrays. *Cell* 102(1) (2000): 9-15.

Zhang, H, Yu, CY, Singer, B, Xiong, M. Recursive partitioning for tumor classification with gene expression microarray data. *Proc Natl Acad Sci U S A* 98(12) (2001): 6730-5.

8

USING FUNCTIONAL GENOMIC UNITS TO CORROBORATE USER EXPERIMENTS WITH THE ROSETTA COMPENDIUM

Simon M. Lin*[1], Xuejun Liao*[2], Patrick McConnell*[1], Korkut Vata*[3], Lawrence Carin[2], and Pascal Goldschmidt[3]

*[1]Duke Bioinformatics Shared Resource, Duke University Medical Center, [2]Department of Electrical Engineering and Computer Engineering, Duke University, [3] Department of Cardiology, Duke University Medical Center. * Authors contributed equally to work.*

Abstract: The Rosetta data set opens the possibility of comparing an experimental microarray data set with a reference profile from the compendium. However, explaining this comparison in terms of individual genes could be a daunting task because of the sheer number of genes. Thus, we postulate a new strategy of modeling microarray data in terms of functional genomic units (FGUs). A functional genomic unit is a group of genes that carries out a certain biological function. We explored the possibility of defining the functional genomic units from the Gene Ontology (GO) annotation of the yeast genome. To visualize the tree structure of the GO, we have written a yeast genomic knowledge browser in Java, and integrated it with the microarray data. The pitfall of using the GO is that only a portion of the genes in the genome are functionally known or inferred. Thus, we further investigated an unsupervized learning method to identify those functional genomic units in the yeast genome. We have applied an established analysis method from digital signal processing, Independent Component Analysis (ICA), to the Rosetta data set. To further validate the utility of the Rosetta compendium, we have designed an experiment to investigate the yeast cells transfected with human Rac1, a small GTPase protein of the Rho family, and demonstrated that functional genomic units helped us to corroborate our own microarray experiment with the Rosetta data set.

Key words: Functional Genomic Units (FGUs), Independent Component Analysis (ICA), Principle Component Analysis (PCA), Gene Ontology (GO), Rac1, Rosetta Compendium.

1. INTRODUCTION

With the accumulation of large amounts of gene expression data, the scientific community hopes to use compendiums to corroborate individual experiments and elucidate functional changes. The Rosetta compendium [Hughes *et al.,* 2000] describes the genomic responses of *S. cerevisae* under 300 conditions, providing a large reference data set to make this kind of comparison possible. Various tools from the statistics and computer science communities have been developed to simplify comparing microarray experiments. However, biologists still have a difficult time comprehending the results of these approaches because the results are expressed in terms of individual genes, and there are such a large number of individual genes in the genome. Furthermore, a standard for comparing results from different array platforms has yet to be developed, so corroborating results can be technically difficult.

Here we postulate a new strategy of modeling the microarray data: instead of using individual genes as building blocks for modeling, we use functional genomic units. A functional genomic unit (FGU) is a group of genes that carries out a particular biological function. These genes can be in the same pathway or can span several pathways, but together they achieve a certain biological task. Functional genomic units can be identified by taking advantage of existing knowledge or by developing new tools. We demonstrate both approaches in this paper.

The Gene Ontology Consortium has produced a hierarchical, structured vocabulary for describing the molecular function, biological process, and cellular component of gene products (see http://www.geneontology.org/). These categorizations are further delineated as one traverses the hierarchy, becoming more specific as the lower levels are approached. When this hierarchy is applied to a set of genes, each GO term describes a subset of functionally coordinated genes, so a GO node at a certain level can be thought of as a functional genomic unit. We took this approach, which is based on existing knowledge, as a method for identifying functional units in microarray data.

Principle Component Analysis (PCA) is an excellent method for extracting linear combinations of the latent components that underlie observed data when the observable is governed by a Gaussian distribution. However, the observed quantities in biology are frequently non-Gaussian, in which case the principal components extracted are no longer statistically independent because zero second-order statistics (covariance) cannot guarantee independence of non-Gaussian quantities [Vigario *et al.,* 2000]. Therefore, we utilize an alternative technique, independent component analysis (ICA), for extracting the underlying components that have

biological significance. Our initial research results show that ICA provides a very promising tool for extracting the functional genomic units without utilizing existing knowledge about the data.

To demonstrate the utility of the Rosetta data set to corroborate other experiments, we used the Affymetrix platform to profile the yeast cells transformed with constitutively active human Rac1. Rac1 is a small GTPase protein of the Rho family [Ridley, 2001] regulating the organization of the actin cytoskeleton, cell migration, cell proliferation, vesicle trafficking, redox system, and gene transcription. However, the downstream mechanisms of Rac1 are not clear. We used the Rosetta data set to support the explanation of our Rac1 results.

2. METHODS

2.1 GO Browser

We implemented a Java-based program for browsing the GO hierarchy, which is a tree where each node is a GO term. Each gene in the data set that has one or more GO terms associated with it was mapped to a node in the tree. We were able to discern which terms have genes associated with them by coloring each node based on whether it maps to a gene or its descendants map to a gene, as well as labeling each node with its gene contents. We were able to get a finer quantitative picture for which terms have gene association by visualizing the data in a "treemap" (see http://www.cs.umd.edu/hcil/treemaps/). Here, each node is drawn as a rectangle, and the children of each node are drawn as rectangles within the parent. The size of each rectangle is directly related to the number of genes attached to that branch. Thus, GO terms with many gene associations can be quickly identified because they (and their parent) are drawn larger. Furthermore, in treating GO terms as functional units, we developed tools to extract the genes from a functional unit and compile the cDNA data from the Rosetta compendium, as well as the user-supplied data.

2.2 ICA Model of the DNA Microarray

Suppose we have a K-dimensional random signal $\boldsymbol{x}_i = [x_1, x_2, \cdots, x_K]^T$, $i = 1,2,\cdots,N$, with N being the number of observations and T the matrix transpose. We seek to find M latent components $\boldsymbol{s} = [s_1, s_2, \cdots, s_M]^T$ satisfying the following linear statistical model,

$$\boldsymbol{x} = F\boldsymbol{s} \qquad [1]$$

where F is a mixing matrix, each column of which gives the contributions of a latent component to the K observed components. Generally, without posing any constraints, the equation in [1] cannot be solved because both F and s are unknown. However, in many scenarios including that considered here, the components in 5 can be assumed independent of each other. The model with the independent constraint on s is thus referred to as an ICA model.

If the distribution of x is Gaussian (and therefore $\boldsymbol{s}$), the columns of F are nothing but the eigenvectors of the covariance matrix of $\boldsymbol{x}$, and the PCA can be used to find F and $\boldsymbol{s}$. In our application, since the gene expression profiles are typically non-Gaussian, the solution involves more complex techniques.

The basic idea behind the techniques to solve [1] is to minimize the Kullback-Leibler (KL) distance between the joint probability density function (pdf) and product of marginal pdf's of $\boldsymbol{s}$. If we restrict ourselves to a standardized version of $\boldsymbol{x}$ (that is, with zero mean and unit covariance matrix), this is equivalent to maximising the summation of the marginal negentropics of $\boldsymbol{s}$ [Comon, 1994]. Because negentropy is in fact the KL distance between the pdfs of $\boldsymbol{s}$ and a Gaussian distribution with the same covariance matrix, the problem boils down to maximising the non-Gaussianity of each component.

One way to measure non-normality of a random variable is to estimate its cumulants. In [2] a cumulant-based ICA algorithm is developed based on pair wise processing. To avoid the noise sensitivity of the cumulant, an alternative ICA algorithm called FastICA [Hyvärinen and Oja, 2000] was devised based on a new approximation of differential entropy.

Suppose the given data matrix has in each column the gene expression levels for one experiment (corresponding to a given environment for the cells of interest). The expression levels of a particular gene in different experiments are then given in the corresponding rows of the data matrix. We aim to extract a certain number of independent components, with each component representing the genes sharing common biological functions. Since the components are mutually independent, each such component should represent a particular biological function that is distinct from all other functions.

The independent component $\boldsymbol{s}$ and the mixing matrix F have interesting interpretations from biology. To see this, let us rewrite [1] as

$$\boldsymbol{x} = \sum_{i=1}^{N} s_i \mathbf{f}_i \qquad [2]$$

where $\mathbf{s}_i$ is the *i*-th component in s and $\mathbf{f}_i$ the *i*-th column of *F*. In [2] the gene expression profile $\boldsymbol{x}$ is expressed as a linear combination of $\mathbf{f}_i$ with the combination coefficient $\mathbf{s}_i$. This is in fact a linear transformation that allows $\boldsymbol{x}$ to be expressed with respect to $\mathbf{f}_i$, with the expression level $\mathbf{s}_i$. Then the $\mathbf{f}_i$ here serves as characteristic vectors, which are responsible for the interpretation of the new expression profile $\mathbf{s} = [s_1, s_2, \cdots, s_N]$. Remember [2] is an ICA model, which means the components of s are mutually independent. This implies that characteristic vectors $\mathbf{f}_i$ may represent some self-defined independent concept. This assumption is reasonable in biology: there are many genes co-expressed in one experiment and these co-expressed genes are usually responsible for some common biological function. Then it is natural for us to define a functional genomic unit based on each characteristic vector $\mathbf{f}_i$. Since the positive or negative sign of values in gene expression only indicates the impact of experiments in different directions (increase or decrease), it is reasonable to use only absolute values of $\mathbf{f}_i$. Let $|\mathbf{f}_i|$ (| | denotes taking absolute values) be normalized to unit norm. The normalized $|\mathbf{f}_i|$ defines a functional genomic unit with the *j*-th value in $|\mathbf{f}_i|$ indicating the fuzzy membership of the *j*-th gene belonging to the *i*-th functional unit.

2.3 Profiling the yeast cells transfected with constitutive active human Rac1 gene

GC1945 yeast cells are transformed with the constitutive active human Rac1 gene. The experimental details are reported in a separate paper [Vata *et al.,* 2002]. Briefly, three biological replicates were measured under the transformed and control conditions using Affymetrix yeast S98 oligoarray that contains about 6,400 yeast ORFs. Logarithmic ratios of expression levels between the transfected and vector-only control cell lines are taken and normalized.

3. RESULTS

3.1 GO mapping of yeast genes

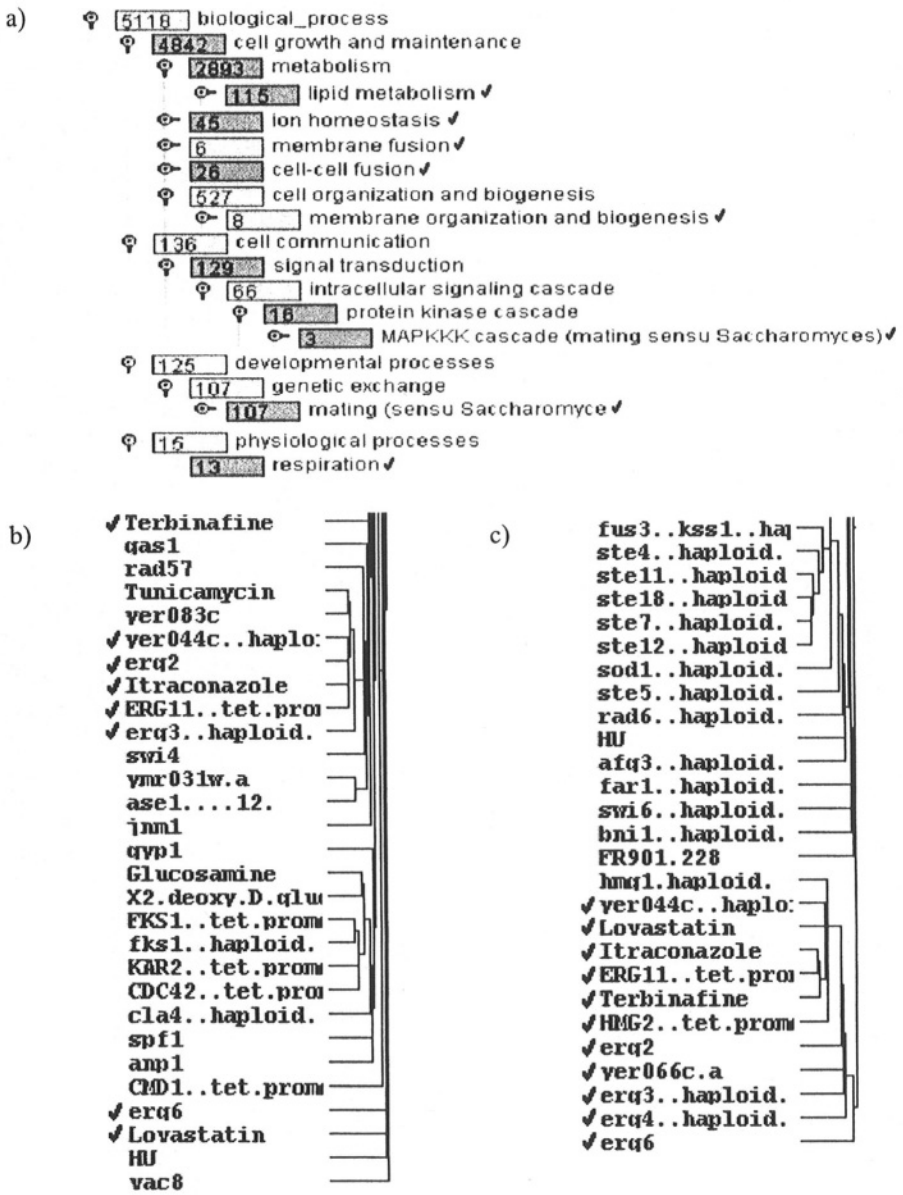


Figure 1. a) Relevant genes selected with Yeast GO Browser. Functional categories related with ergosterol biosynthesis are checked. b) Rosetta experiments grouped by average-linkage clustering [Eisen *et al.*, 1998] with 6136 gene features. Experiments with perturbation of ergosterol synthesis are checked. Only the interesting portion of the results shown. c) Same clustering method as applied in Fig 1 (b), but using 238 gene features selected from Fig 1 (a).

From previous knowledge, we know the perturbation of ergosterol biosynthesis will compromise ion homeostasis, membrane functions, mating behaviour, and respiration [Parks *et al.,* 1995]. Accordingly, we selected 238 relevant yeast genes with the GO Browser (Figure 1a). By using this relevant set of genes for clustering, we can see tighter clustering of ergosterol-related experiments (Fig 1c) as compared to Fig 1b, which used all gene features for clustering.

3.2 ICA Results

3.2.1 Simulated data

Suppose we have three genomic functional units whose simulated expression levels versus experiments are given in the left portion of Figure 2, and assume the expression levels of the observed genes are in the right portion of Figure 2. The expression levels of the extracted independent components versus experiments are shown in the left portion of Figure 3, and, for comparison, the corresponding PCA results are shown in the right portion of Figure 3. Under an ideal situation, the results in Figure 3 should recover the information in the left portion of Figure 2. Figure 3 demonstrates that ICA does a better job of recovering the original genomic function units than PCA.

The results in Figure 2 and 3 are re-plotted in Figure 4 and 5, respectively, in the format of histograms, which demonstrate in a clearer manner that ICA recovers the original three non-Gaussian sources (i.e., genomic function units) while PCA does not.

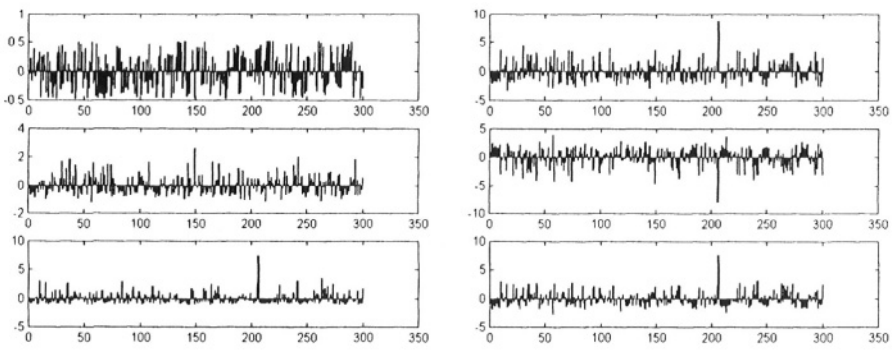


Figure 2. Left: Simulated expression levels of three genomic functional units as a function of experiment. Right: Expression levels of the observed three genes as a function of experiment.

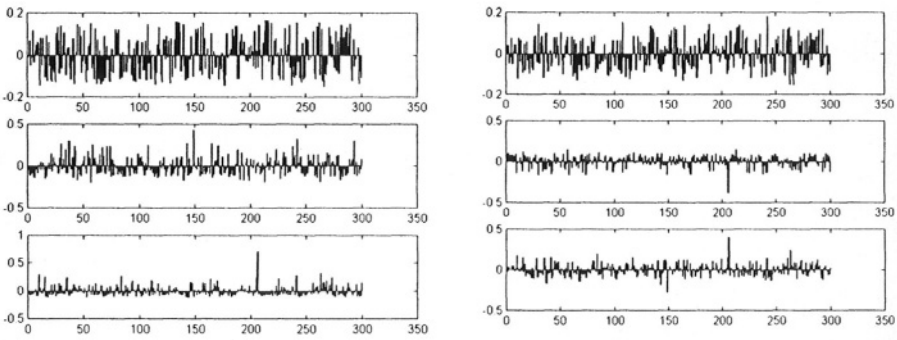


Figure 3. ICA and PCA results on the simulated gene expression data. Left: Expression levels of the independent components as a function of experiment. Right: Expression levels of the principal components as a function of experiment.

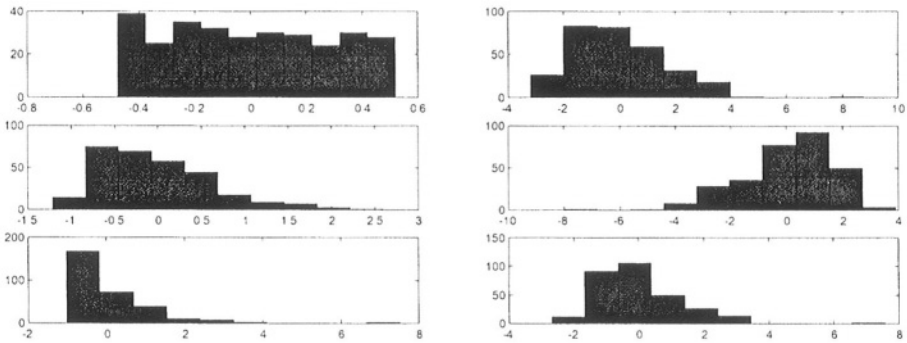


Figure 4. Left: Histogram of the simulated expression levels of the three genomic functional units in Figure 2 (left). Right: Histogram of the expression levels of the observed three genes in Figure 2 (right).

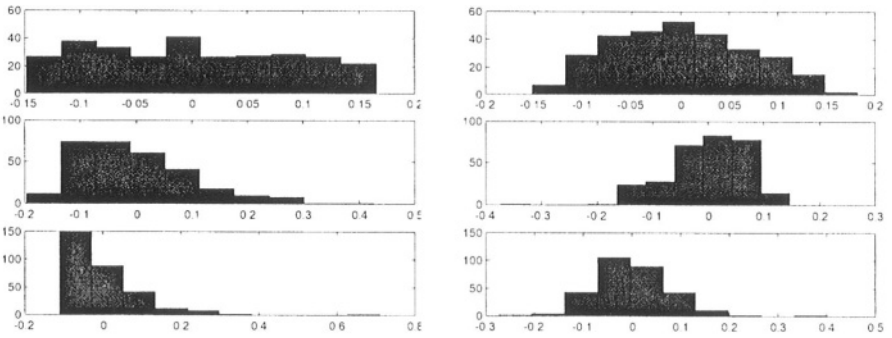


Figure 5. Histogram of the ICA and PCA results on the simulated data. Left: Histogram of the expression levels of the independent components in Figure 3 (left). Right: Histogram of the expression levels of the principal components in Figure 3 (right).

3.2.2 The compact Rosetta dataset

We then applied ICA to a compact Rosetta dataset in which we have good a priori knowledge on the genes and experiments. It consists of 122 genes measured in 126 experiments. The independent components extracted are shown in the left portion of Figure 6. The characteristic vectors of each independent component are shown in the right portion of Figure 6. These characteristic vectors give the fuzzy membership of genes belonging to each independent component and therefore define genomic function units. A manual check of the results in Figure 6 with the a priori knowledge demonstrates that they are consistent. For example, the prominent genes in FGU #7 include genes in:

- the mating response:
  - YJL157C (FAR1, cell cycle arrest, mating response)
  - YOR212W (STE4, beta subunit of G protein coupled to mating factor receptor)
  - YNL145W (MFA2, mating a-factor pheromone precursor)
  - YPL256C (CLN2, G1 cyclin)
- lipid and ergosterol biosynthesis pathway
  - YMR015C (ERG5)
  - YLR056W (ERG3)
  - YGR175C (ERG1)
  - YNL111C (CYB5, cytochrome b5)

- YHR007C (ERG11)
- YJL196C (ELO1, elongation enzyme 1)

- carbohydrate metabolism
  - YPR160W (GPH1)
  - YEL011W (GLC3)
  - YGR032W (GSC2)
  - YGL256W (ADH4)
  - YJL153C (INO1)
  - YDR074W (TPS2)
  - YFL014W (HSP12)
  - YCL040W (GLK1)
  - YJR009C (TDH2)

Accordingly, FGU #7 was downregulated in experiments of ergosterol perturbations (erg2, erg11, YER044C, itraconazole) and mating (ste4, ste11, ste12, stel8, fus3); but upregulated in experiments perturbing the cell wall function (tet-KAR2, tet-CDC42, tet-FKSl, fksl, tet-RHOl). This indicates new insights to how the biological system works, and corresponds with our previous knowledge [Parks *et al.,* 1995; Kitajna *et al.,* 2000]. Similarly, we found FGU #13 (corresponding to protein synthesis); FGU #4, 5, 9, 10, and 20 (corresponding to various aspects of energy and carbohydrate metabolism); and FGU #9 (corresponding with general metabolism).

3.2.3 The Complete Rosetta dataset

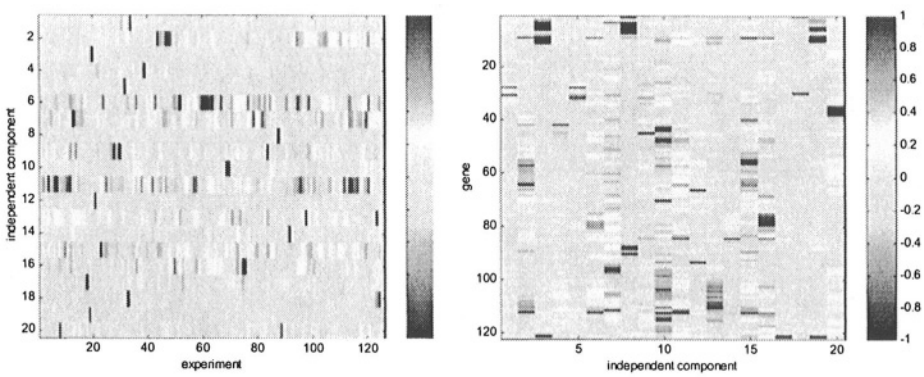


Figure 6. ICA results on the Compact Rosetta dataset. Left: Expression levels of the independent components (column) versus experiment (row). Right: Characteristic vector (column) of each independent component (row).

The full Rosetta dataset consists of 6316 genes measured in 300 experiments. To save space, the complete ICA results are not shown here. Instead, we show two of the functional units that have been identified in the complete Rosetta dataset by the ICA algorithm.

The fuzzy membership function (defined in section 2.2) of the first identified unit is shown in Figure 7. By setting an appropriate threshold at 0.06, we find there are about 10 genes in this unit, including carbohydrate metabolism: YFL053W (DAK2), YLR307W (CDA1); and cell growth, division and DNA synthesis: YFL026W (STE2), YDR218C (SPR28), YLR307W (CDA1), YPL121C (MEI5).

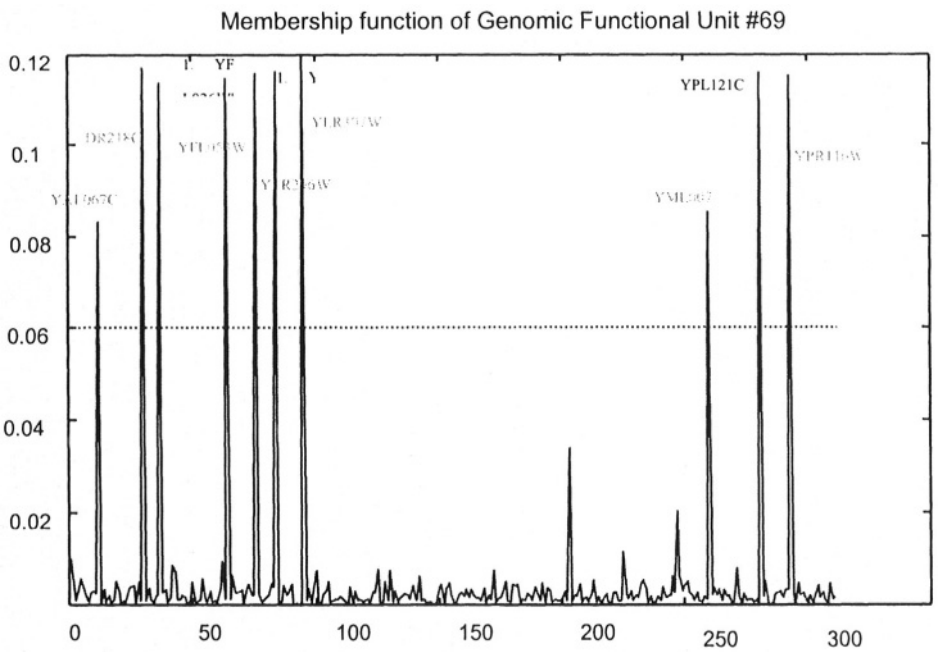


Figure 7. Fuzzy membership function of the first genomic function unit (FGU #69) identified in the complete Rosetta dataset.

By similar procedures, we analyze a second FGU with 15 genes. Six of these genes are coding for isoforms of α- glucosidase (MAL62, MAL32, MAL12, FSP2, YIL172c, and YJL216c). Four of the genes are directly associated with cell-wall synthesis and sporulation (YER096w, YHR139c, YDR403w, and YJR150c). Five genes are involved in glucose metabolism (TUPlYDL245c, YEL069c, YNR072w, and YJR158w). ICA analysis suggests these three groups of genes are working coordinately. It reflects the underlying sequential process of biology: i) glucose uptake, ii) the

intracellular transport and metabolism of glucose, and iii) utilization of sugars in the cell-wall biosynthesis as building blocks. In conclusion, this FGU includes 15 genes directing sugar metabolism into cell-wall synthesis in a coordinated manner.

3.3 Using the Rosetta data set to corroborate the Rac1 Experiment

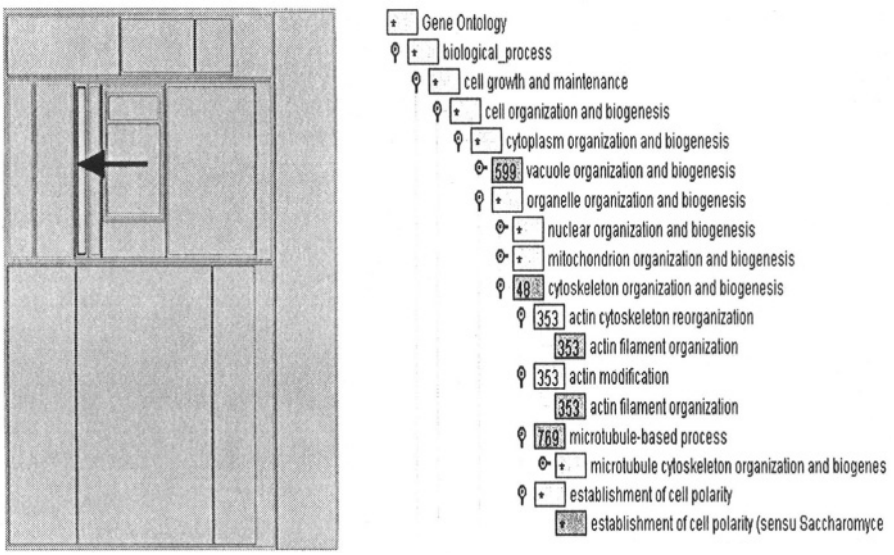


Figure 8. Right: Tree representation of the GO hierarchy where nodes are labeled by an internal gene index. Left: Treemap of "cell organization and biogenesis" where "actin filament organization" is highlighted with an arrow.

Using SAM statistics [Tusher *et al.,* 2001], we identified 792 genes from the Rac1 experiment that are differentially regulated more than 1.5 fold. These genes are mapped to 1685 GO terms. 726 of these gene-to-term mappings were found under the "cell growth and maintenance" branch, the largest expression in the tree. Within this branch, we identified "cell organization and biogenesis" as being both pertinent to our experiment and having a substantial portion of the genes at 71 mappings. Further in the tree, we found "actin filament organization" to be expressed (see Figure 8), which is relevant to the Rac1 biology. We took its parent, "organelle organization and biogenesis," to be a functional unit and visualized the microarray data from these genes in Eisen's Cluster and TreeView (see http://rana.lbl.gov/index.htm and [Eisen *et al.,* 1998]). Our Rac1 experiment is close to the far1 and rvs161 mutant experiment of the Rosetta data set.

Far1 inhibits cdc28, and thus controls the cell cycle arrest. Rvs161 is a cytoskeletal protein binding protein. Mutations of rvs161 result in a delocalization of the actin cytoskeleton. Both of them provide insight on howRac1 works.

When we analyze the Rac1 experiments in terms of the FGUs identified by the ICA algorithm, we can see Rac1 has upregulation of FGU #7 (see section 3.2.2 for discussion of its characteristic genes). This effect is similar to the Rosetta experiment findings tet-KAR2, tet-CDC42, tet-FKS1, fks1, and tet-RHO1. It suggests the close relationship between the Rac1 effects and tet-Rho1 effects, since Rac is in the Rho protein family [Ridley, 2001].

4. DISCUSSION

Traditionally, microarray data is interpreted on a gene-by-gene basis. In this paper, we proposed a new strategy by using a group of functionally related genes called functional genomic units (FGUs) to interpret the complex data. Two complementary approaches have been explored to define the FGUs.

First, we construct the FGUs by using the Gene Ontology. This approach is actually a feature selection process based on expert opinions as defined in the GO annotation of genes. A supervised learning approach for feature extraction has been discussed in Chapter 9 of this volume.

As demonstrated in section 3.1, selecting pertinent gene features involved in ergosterol biosynthesis helps us to answer relevant questions of ergosterol pertubation, whereas a "kitchen-sink" clustering with all the variables has limited explanatory capabilities. This is because clustering is highly dependent on the context of analysis. Large numbers of irrelevant features will degrade the results [John *et al.,* 1994]. For example, in the context of clustering auto insurance customers, one would use features such as driving accidents and driver's age, but not features such as driver's body weight. Similarly in microarray analysis, we want to select relevant gene features with the biological focus of our interest.

Unfortunately, FGUs defined by selecting appropriate nodes in the gene ontology tree limit us to what is already known. As a complementary method, we explored an *ab initio* data modeling approach by using ICA for feature extraction. ICA has been shown to be a good algorithm for blind source separation [Comon, 1994], EEG signal processing [Hyvärinen and Oja, 2000], and brain imaging [Tzyy-Ping *et al.,* 2001]. To our knowledge, this is the first time it has been applied to microarray data. To better understand the behavior of the ICA algorithm, we started with simulated data, and then moved on to a compact Rosetta dataset before we applied it to

the full dataset. ICA exploits the assumption that the functional units of genes are mathematically independent of each other. This independence, achieved by linear combination of genes, may or may not overlap with existing knowledge of molecular pathways. In addition, the linear combination of genes, which typically includes large numbers of uncharacterised genes, imposed problems for biological explanations. In section 3.2 and 3.3, we are only able to explain some of the FGUs; the remains are still of interest and should be explored further.

With so many sources of variation in microarray experiments, many investigators have argued against the usefulness of large-scale gene expression databases to deposit data generated from different labs with various technology platforms. In this paper, we use the Rosetta dataset to corroborate our Rac1 experiment data. It generates an interesting hypothesis of the Rac signalling pathway awaiting further biological validation. This suggests that, with caution, it is possible to use a large reference database to corroborate other experiments.

Finally, it should be noted that the fact that the microarray data compiled from our GO Browser corroborates designed experiment data lends support to a correct GO annotation of the genes. Although many genes have yet to be annotated, the GO database may be a powerful tool in genomics research.

REFERENCES

Comon, P. Independent component analysis - a new concept? *Signal Processing* 36 (1994): 287-314.

Eisen, MB, Spellman, PT, Brown, PO, Botstein, D. Cluster analysis and display of genome-wide expression patterns. *PNAS* 95(25) (1998): 14863-14868.

Hughes, TR, *et al.* Functional discovery via a compendium of expression profiles. *Cell* 102 (2000): 109-126.

Hyvärinen, A, Oja E. Independent Component Analysis: Algorithms and Applications. *Neural Networks*, 13(4-5) (2000): 411-430.

John, GH, Kohavi, R, Pfleger, K. Irrelevant features and the subset selection problem. In *Proceedings of the 11th International Conference on Machine Learning* (1994): 121-129.

Kitajma, Y. Structural and biochemical characteristics of pathogenic fungus; cell walls, lipids and dimorphism, and action modes of antifungal agents. *Japanese Journal of Medical Mycology.* (2000) 41(4):211-217.

Parks, LW, Smith, SJ, Crowley, JH. Biochemical and physiological effects of sterol alterations in yeast--a review. *Lipids* 30(3) (1995): 227-30.

Ridley, AJ. Rho family proteins: coordinating cell responses. *Trends Cell Biol* 11(12) (2001): 471-477.

Ridley, AJ. Rho proteins: linking signaling with membrane trafficking. *Traffic* 2(5) (2001): 303-310.

Tusher VG, Tibshirani R, Chu G. Significance analysis of microarrays applied to the ionizing radiation response. *Proc Natl Acad Sci U S A* 98(9) (2001): 5116-5121.

Tzyy-Ping J, *et al.* Imaging brain dynamics using independent component analysis. *Proceedings of the IEEE* 89(7) (July 2001): 1107 -1122.

Vata, K, Lin, SM, Dressman, H, Nevins, J, Goldschmidt-Clermont, PJ. Genome-wide profiling of the S. cerevisae Transcriptional Response to Constitutively Active Human Rac1 GTPase. *Manuscript in preparation* (2002).

Vigario, R, Sarela, J, Jousmiki, V, Hamalainen, M, Oja, E. Independent component approach to the analysis of EEG and MEG recordings. *IEEE Transactions on Biomedical Engineering* 47(5) (May 2000): 589–93.

This page intentionally left blank

9

FISHING EXPEDITION - A SUPERVISED APPROACH TO EXTRACT PATTERNS FROM A COMPENDIUM OF EXPRESSION PROFILES

Zhen Zhang[1,3], Grier Page[2], Hong Zhang[4]
[1] Center for Biomarker Discovery, Dept. of Pathology, Johns Hopkins Medical Institutions, Baltimore, MD 21231, [2] Dept. of Biometry and Epidemiology Medical University of SC, Charleston, SC 29425, [3] 3Z Informatics, LLC., Mt. Pleasant, SC 29464, [4] Dept. of Computer Science, Armstrong Atlantic State University, Savannah, GA 31419

Abstract: Reference databases of expression profiles from diverse mutations and chemical treatments of a single assay offer a bird's-eye view of changing expression patterns due to multiple perturbations. Such a compendium of expression profiles has been used to ascertain the roles of previously uncharacterised genes and infer the pathways through which their impact may take place. However, many genes have multiple molecular functions and are involved in different biological processes. The interaction patterns between genes and profiles from two-dimensional hierarchical clustering of such compendium of data could be very complex and often scattered. This makes it difficult to identify and extract all the genes and profiles whose variation in expression levels is closely associated with a particular target function. In this paper, a supervised component analysis approach is proposed in which a small number of profiles and/or genes of known properties are used as "bait" to help "fish out" other profiles and genes from a reference database that are relevant to a particular function of interest. The final cluster analysis and pattern match is then done using a much-reduced data set.

Key words: Expression data analysis, microarray data analysis, supervised analysis, support vector machine, SVM, unified maximum separability analysis, UMSA, discriminant analysis, cluster analysis.

1. OBJECTIVES

A reference database of expression profiles from diverse mutations and chemical treatments of a single assay offers a bird's-eye view of changing expression patterns due to multiple perturbations. In such a compendium of expression profiles, a particular perturbation would most likely result in collective changes among multiple genes. The individual genes could also be implicated in profiles from different experimental conditions. Because of such many-to-many relationships between genes and expression profiles, the direct application of 2-dimensional hierarchical cluster analysis on a large number of genes and profiles often results in noisy and scattered patterns that are difficult to interpret.

There are situations in which the objective of analysis is to identify genes and perturbations that are pertinent to a predetermined set of molecular functions or biological processes of interest. The proposed approach in this paper is to incorporate this information into a supervised algorithm to select a subset of the original data upon which further clustering and pattern matching can be performed more effectively and efficiently.

2. METHODS

2.1 Data Sets

The publicly available reference database of expression profiles of yeast mutants and chemical treatments [Hughes *et al.,* 2000] is used as test data for the proposed algorithm. A subset of 136 experiment profiles and 551 ORFs have been selected from the original data of 300 experiment profiles and 6298 ORFs based on the criteria of including only experiments with 2 or more genes up- or down-regulated at greater than or equal to 3 fold, and a p-value ≤ 0.01 based on the error model in Hughes *et al.* [2000]; and only genes up- or down-regulated at greater or equal to 3 fold, and p-value ≤ 0.01 in 2 or more experiments. In addition, from the same source, profiles of 63 negative controls where also used in the analysis.

One of the purposes of this paper is to show that the proposed supervised method to select a subset of data for further analysis will not result in significant loss of useful information. The above data pre-processing steps were chosen to closely match the data selection criteria used in the original publication [Hughes et. al, 2000] so that results from the two approaches could be compared.

2.2 The Algorithms

In Zhang *et al.*, [200la], the Unified Maximum Separability Analysis (UMSA) procedure was reported for computing a projection vector in a high-dimensional space along which two classes of data are optimally separated. The UMSA procedure incorporates partial data distribution information into the construction of an optimal soft-margin hyper-plane similar to the ones described in the Support Vector Machine (SVM) literature [Vapnik, 1998]. In Zhang *et al.,* [200la], UMSA was used in a backward stepwise algorithm to assign significance scores to individual genes according to their collective contributions to the separation of classes of experiments.

The UMSA classifier for a set of m training samples $x_1, x_2, \ldots, x_m$ drawn from distributions D^+ and D^- with the corresponding class membership labels $l_1, l_2, \ldots, l_m \in \{-1,1\}$ is determined by solving the following constrained optimisation problem:

$$\text{Minimize} \quad \frac{1}{2}\upsilon \cdot \upsilon + \sum_{i=1}^{m} p_i \xi_i \qquad [1]$$

$$\text{subject to} \quad l_i(\upsilon \cdot x_i + b) \geq 1 - \xi_i, i = 1, 2, \ldots, m,$$

where the non-negative variables $\xi_1, \xi_2, \ldots, \xi_m$ represent errors in the constraints that are penalized in the object function, and the coefficients $p_1, p_2, \ldots, p_m$ are positive constants reflecting the relative "importance" of the m individual data point. In UMSA, $p_i = \phi(x_i, D^+, D^-) > 0$ is used to incorporate prior knowledge about data distribution into the optimisation procedure so that the resultant classifier does not rely solely on boundary points (support vectors). In the current implementation, p_i is typically related to the level of disagreement of a sample x_i to a classifier derived based on distributions of D^+ and D^- estimated from the m training samples (e.g., classifier from linear discriminant analysis). Let this level of disagreement be δ_i, the following positive decreasing function is used to compute p_i:

$$p_i = \phi(\delta) = C \cdot e^{-\delta_i^2/\sigma^2}, \text{ where } C > 0. \qquad [2]$$

In this paper, for processing microarray expression data where the large number of variables and the small sample size make the direct estimation of conditional distributions difficult, δ_i is defined to be the shortest distance between the data point x_i and the line that goes through the two class means. The two parameters, σ and C modulate the amount of influence an

individual sample may have upon the solution of υ in the optimisation problem above. One may notice that for a very large σ relative to the range of ∂_i, p_i would essentially become a constant close to *C*. The UMSA algorithm then becomes equivalent to the optimal soft-margin classifier in SVM.

The UMSA procedure may also be used for component analysis [Zhang *et al.,* 200lb]. The basic algorithm iteratively computes a projection vector d along which two classes of data are optimally separated for a given set of UMSA parameters. The data are then projected onto a subspace perpendicular to d. In the next iteration, UMSA is applied to compute a new projection vector within this subspace. The iteration continues until a desired number of components have been reached. For interactive 3D data visualisation, often only three components are needed. Depending on the shape of data distribution, for many practical problems, three dimensions appear to be sufficient to "extract" all the significant linear separation between two classes of data. The following is the actual UMSA component analysis algorithm for a data set of *m* samples and *n* variables:

inputs:
  UMSA parameters C and σ;
  number of components $q \le \min(m, n)$;
  data $X = (x_1, x_2, \ldots, x_m)$; and
  class labels $L = (l_1, l_2, \ldots, l_m)$, $l_i \in \{-1, +1\}$.
initialisation:
  component set $D \leftarrow \{\}$;
  $k \leftarrow 1$.
while $k \le q$
  1. applying UMSA(σ, C) on $X = (x_1, x_2, \ldots, x_m)$ and L;
  2. $d_k \leftarrow \upsilon / \|\upsilon\|$; $D \leftarrow D \bigcup \{d_k\}$;
  3. $x_i \leftarrow x_i - (x_i^T d_k) d_k$, $i = 1, 2, \ldots, m$;
  4. $k \leftarrow k + 1$.
return D.

The UMSA component analysis method is similar to the commonly used principal component method (PCA) or Singular Value Decomposition (SVD) in that they all reduce data dimension. The difference is that in PCA/SVD, the components represent directions along which the data have maximum variations while in UMSA component analysis, the components correspond to directions along which two predefined classes of data achieve maximum separation. PCA/SVD are for data representation; UMSA

Component Analysis is for data classification (which is also why in many cases, a three dimensional component space is sufficient for linear classification analysis).

2.3 The approach

In microarray expression data processing, a particular phenotypic variation may be associated with multiple genotypic changes. Similarly, individual genes are often implicated in multiple biological functions and pathways. The commonly used 2D hierarchical cluster analysis approaches require a gene or an experiment profile to "take a stand" and be grouped into one and only one of the clusters. As with many unsupervised methods, there is no guarantee that a gene or an experiment profile with multiple associations would necessarily be grouped into a cluster that is meaningful for the purpose of a particular analysis.

The approach proposed in this paper uses UMSA component analysis to project the entire expression data onto a 3D component space. The projection is determined based on a selected subset of data with known properties important for the purpose of analysis. The user would then be able to discard a significant (albeit conservative) portion of the data that are not relevant before applying a regular cluster analysis procedure.

To analyze the compendium of yeast mutants and chemical treatments data, a small number of experiments with conditions associated with the molecular functions or biological processes of interest are used as "bait" forming one of the two classes of data points. The control experiments are used to form the other class of data. In the absence of control data, the unselected large number of profiles may serve as the control group. Once the projection to a 3-dimensional UMSA component space is determined, the entire data set is projected onto this space. Data points (profiles) that are close to the few selected "bait" in the UMSA component space are selected for further analysis. Genes that correspond to large projection coefficients (loading factors), especially of the first component, are also selected for further analysis.

The above procedure projects profiles as data points in the gene space. One may also first select a few genes that are known to be involved in the biological processes of interest and carry out the selection process by projecting genes as data points in a profile space.

UMSA component analysis selects genes that are essential in differentiating the "bait" profiles from the large compendium of profiles. The selected profiles are most similar to the "bait" profiles only in the subspace formed by these essential genes. Irrelevant genes and profiles with strong patterns and forming large clusters lose their significance in this

subspace. In contrast, under PCA/SVD, genes and profiles that are underrepresented in terms of expression variations and absolute numbers may not be selected at all. Many that do get selected due to repetition and large variance may have nothing to do with the purpose of the current analysis.

3. RESULTS

Mutants erg2Δ, and erg3Δ, and tet-ERG11 are used as the "bait" class and the 63 negative controls as the control class. UMSA parameters s=10.0 and K=5.0 [Zhang *et al.,* 200la] are used for the component analysis, resulting in a subset of 78 profiles and 200 ORFs. The UMSA component analysis results and profile selection are demonstrated in Figure 1. Results from 2-dimensional hierarchical cluster analysis (absolute uncentered) using the Cluster software package from Stanford University [Eisen *et al.,* 1998] are shown in Figure 2, which are compared to clustering results using the entire dataset of 136 profiles and 551 ORFs in Figure 3.

The cluster of ORFs identified as related to yeast ergosterol biosynthesis from both the original data set and the reduced data set are listed in Table 1. It shows that the reduced data set contains most ORFs except erg3, CAF120/ynl278w, and ysr3/ykr053c. It has four additional ORFs, however, that are not in the results from the large data set: erg25, ymr134w, yll012w, and cyb5/ynl111c. The omission of erg3 in the results from the reduced data set is explained by the fact that erg3Δ is used as one of the three "bait" profiles. The UMSA component representation of erg3Δ hence has very little to do with erg3. Two of the additional ORFs identified from the reduced data set are hypothetical ORFs with unknown functions.

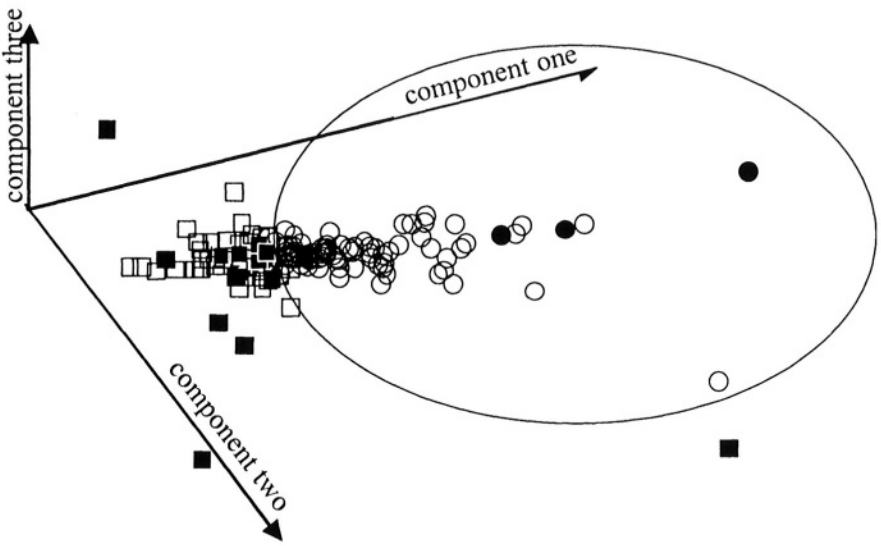


Figure 1. 3D plot of individual profiles in UMSA component analysis. The three filled circles are mutant erg2Δ, and erg3Δ, and tet-ERG11 used as the "bait." The filled squares are the 63 controls used with the "bait" to determine the UMSA component projections. The hollow circles and squares are the projections of the remaining profiles in the UMSA component space. Profiles within the neighbourhood of the "bait" (the hollow circles) are selected for further analysis. During the same process, genes that contribute the most to the separation between the "bait" and the controls are selected for further analysis. This is a Matlab 3D scatter plot based on the original Java 3D API plot in colour.

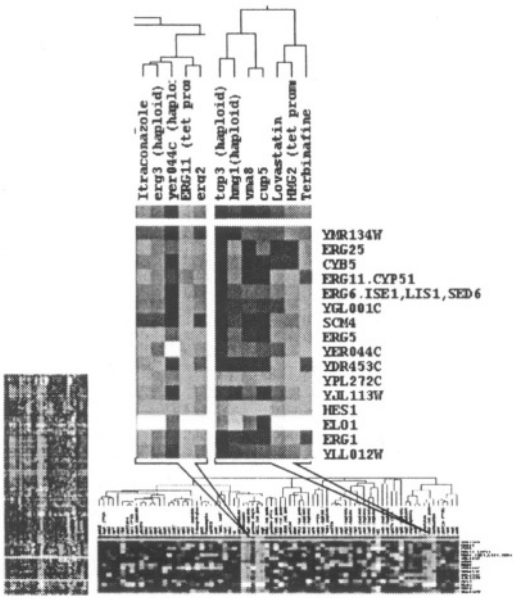


Figure 2. Two-dimensional hierarchical cluster analysis of the reduced data set of 78 profiles and 200 ORFs with zoomed image of ORFs up-regulated in ergosterol biosynthesis related mutants and chemical treatments.

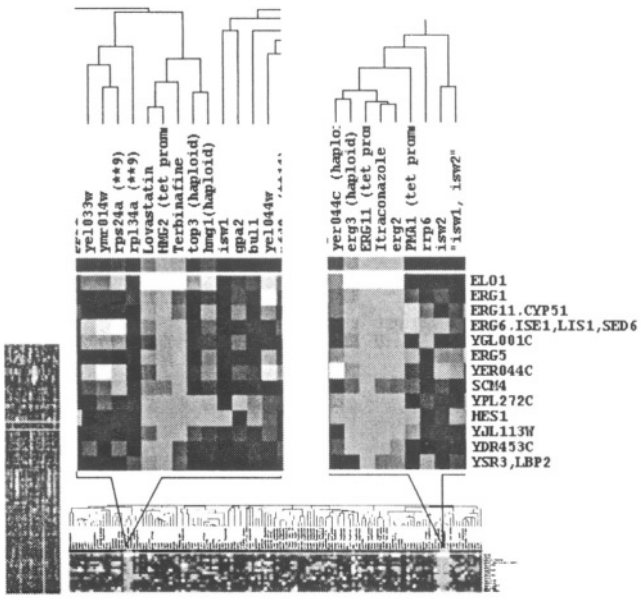


Figure 3. Two-dimensional hierarchical cluster analysis of the original data set of 136 profiles and 551 ORFs with zoomed image of ORFs up-regulated in ergosterol biosynthesis related mutants and chemical treatments.

Table 1. Comparison of ORFs identified using large and reduced data sets. Identified ORFs are indicated by "*".

ORFs	Large Set	Reduced Set
YDR453C	*	*
YER044C	*	*
YGL001C	*	*
SCM4/YGR049W	*	*
ERG25/YGR060W		*
ERG1/YGR175C	*	*
ERG11/YHR007C	*	*
YJL113W	*	*
ELO1/YJL196C	*	*
YSR3/YKR053C	*	
ERG3,SYR1/YLR056W	*	
YLL012W		*
ERG6//YML008C	*	*
ERG5/YMR015C	*	*
YNL278W	*	
YMR134W		*
CYB5/YNL111C		*
HES1/YOR237W	*	*
YPL272C	*	*

4. CONCLUSIONS

Analysis of large databases often requires careful balance between efficiency through data reduction and minimizing the risk of losing useful information. The main advantage of the proposed approach is that by using a supervised method, known properties of experiments and genes are incorporated into the data selection process, which in turn improves the effectiveness and efficiency of cluster analysis and pattern matching and detection. However, this approach is most useful for "fishing out" unknown relationships amongst genes and profiles that have something in common with the pre-selected "bait" profiles or genes.

REFERENCES

Eisen, MB, Spellman, PT, Brown, PO, Botstein, D. Cluster Analysis and Display of Genome-Wide Expression Patterns. *Proc. Natl. Acad. Sci. USA* 95 (1998): 14863–14868.

Hughes, TR *et al.* Functional Discovery via a Compendium of Expression Profiles. *Cell* 102 (July 2000): 109-126.

Vapnik, VN. *Statistical Learning Theory.* New York: John Wiley & Sons, 1998.

Zhang, Z, Page, G, Zhang, H. Applying Classification Separability Analysis to Microarray data. In: Lin, SM, Johnson, KJ, eds. *Methods of Microarray Data Analysis.* Norwell, MA: Kluwer Academic Publishers (2001): 125-136.

Zhang, Z, Zhang, H. A Unified Maximum Separability Analysis Procedure with Applications to Biological Expression Data Processing. *Manuscript under review.*

This page intentionally left blank

10

MODELING PHARMACOGENOMICS OF THE NCI-60 ANTICANCER DATA SET: UTILIZING KERNEL PLS TO CORRELATE THE MICROARRAY DATA TO THERAPEUTIC RESPONSES

Nilanjan Dasgupta[1], Simon M. Lin[2] and Lawrence Carin[1]
[1]*Department of Electrical Engineering. Duke University*
[2]Duke Bioinformatics Shared Resource, Duke University Medical Center

Abstract: Modeling the relationship between genomic features and therapeutic response is of central interest in pharmacogenomics [Musumarra *et al.,* 2001]. The NCI-60 cancer data set with both gene expression and drug activity measurements provides an excellent opportunity for this modeling exercise. To correlate the gene expression profile with the drug activity pattern, we utilized a soft modeling technique called Partial Least Squares (PLS) [Tobias, 2000]. Soft modeling requires less stringent assumptions about the data than other modeling techniques [Falk *et al.,* 1992], A high level of collinearity in multi-dimensional gene expression profiles motivates us to undertake the PLS approach, which not only trims data redundancy but also exposes the underlying hidden functional units as latent features. It is believed that these functional gene groups play a key role in determining the efficacy of the cancer drugs to different cell lines (types of cancer). We have shown the efficacy of PLS in identifying drug resistant and drug sensitive genes. We have also investigated techniques to exploit the non-linear dependence between individual gene expressions in order to explain variations in the drug activity pattern. This is facilitated by a kernel function that implicitly carries out the regression in a higher-dimensional space where the data is linear [Christiannini *et al.,* 2000]. The kernel-based non-linear approach is shown to be more effective in defining the correlation between the drug response and the gene expressions. The PLS approach, as implemented here, could be used to differentiate cancer cell lines between renal cancer and melanoma, for example, or different drug groups like Alkylating agents and Tubulin-active anti-mitotic agents.

Key words: Pharmacogenomics, NCI-60 microarray data set, anti-cancer therapeutics, PCA, Multiple Linear Regression (MLR), kernel, kernel-PLS, soft modeling.

1. INTRODUCTION

DNA microarray technology provides an enormous opportunity to analyze the behavior of several thousand genes in a cell or a tissue. The massive scale of parallelism using DNA-chip technology quantifies the concentration of an array of predefined genes in a single experiment. The focus of our research is to identify the underlying (hidden) functional gene units from gene expression data and correlate the behavior of these latent features with drug activity patterns. The NCI-60 data set provides an excellent opportunity for an exercise of modeling pharmacogenomics. In this data set [Ross *et al.,* 2000], the A matrix represents the activity pattern of 118 drugs over 60 different human cancer cell lines. The activity of any drug for a given cell line is defined as the log-concentration of the drug required to reduce the growth rate to 50% ($\log GI_{50}$). The T matrix shows the expressions (concentrations) of 1375 genes over the same cell lines. Since the gene expression pattern observed in the NCI-60 data set corresponds to untreated cells, we limit our analyses on sensitivity to therapy rather than on molecular consequences of therapy. In our research, we are only looking for intrinsic genomic components rather than individual genes contributing to drug resistance.

The objective of our research is to model the correlation between the gene expression profile and the drug activity pattern based on NCI-60 anticancer data. We have proposed a soft modeling approach called Partial Least Squares (PLS) to that effect. PLS is shown to be effective in handling high dimensionality and requires relatively few observations to model the underlying correlation between latent features and drug groups. Ordinary PLS searches only in the linear space of genomic features, and hence ignores non-linear relationships between the genes, if any. The kernel-based non-linear PLS approach [Ranner *et al.,* 1994] incorporates the non-linearity and is shown in our modeling to be highly efficient and effective in expressing the interdependence between the latent components. These non-linear components explain the drug activities in a more compact fashion. The paper is structured in the following manner: The motivation behind the PLS approach is discussed in Section 2 followed by the methodology in Section 3. The performance analyses of proposed algorithms is shown in Section 4 followed by conclusions and acknowledgements in Section 5 and 6.

2. MOTIVATION

Since our primary objective is to quantify the correlation between the gene expression and the drug response over a set of cell lines, one intuitive approach is to investigate the Multiple Linear Regression (MLR) method. MLR assumes the individual features in the input set to be linearly independent. Hence, it explains the variation in the output data as a weighted sum of individual correlations. We believe that individual gene-drug correlations do not have independent biological interpretation. Rather, it is a group of genes which play a key role in any biochemical reaction. A high level of interdependence between individual genes in a cell makes MLR inappropriate for analyzing gene expression data. Another intuitive approach to reduce redundancy in the multivariate data is the eigenvalue decomposition used in Principal Component Analysis (PCA) [Janne *et al.,* 2001]. For a matrix X, the principal components are represented by the eigenvectors of the square matrix $X^T X$. Though PCA is shown to be effective for noise reduction, it does not address the notion of finding relevant features (gene groups) that are responsible for the variation in drug responses. In other words, PCA is effective in reducing the dimensionality by identifying the eigenvector directions in the multidimensional gene space, but these directions do not necessarily explain the variation in the output space (drug response) in a most effective way. Hence, we aim for modeling scheme that reduces the dimensionality of the gene expression data while taking the variation in drug response into consideration. The PLS approach presented here is a soft modeling technique to analyze the gene expression pattern to extract the underlying gene groups that correlate maximally to the drug response. In PLS, the optimal linear predictive relationships between input and output variables are created with an objective to minimize the generalization error. Generalization error is defined statistically as modeling error on the entire input data space.

The motivation behind the PLS approach is twofold:

1. It is known in the scientific community that every gene in a human genome does not express itself independently of each other in their roles for malfunctioning of tumor cells. Rather, it has been verified that in most of the complex biochemical reactions, a small subset of genes work in cohesion. This phenomena leads to high multi-collinearity among the variables (gene expressions) in microarray data. Hence, the algorithm should be capable of extracting underlying features governing the biochemical reactions from a high-dimensional correlated data set.
2. The dimensionality of the feature space in the NCI-60 data set is much higher than the number of observations (cell lines) available for training. Hence, the modeling approach should handle overfitting and minimization of generalization error.

Partial Least Squares (PLS) is shown to be very effective in chemometrics [Ranner *et al.,* 1994] and econometrics under similar constraints.

3. METHODOLOGY

PLS was developed by Herman Wold in 1966 as a general model to estimate the latent constructs from multiple indirect observations [Wold *et al.,* 1984]. PLS models are completely described by two sets of linear equations; one represents the relationship between the latent features extracted from the observables and the other connects the latent variables with the observed quantities.

The gene expression profiles serve as "descriptors" for PLS modeling and the drug activity patterns are "response" vectors. Ordinary PLS, as described by Wold, addresses only the linear dependence between the features, whereas a Kernel-based PLS approach incorporates non-linearity. For brevity, we shall only focus on the methodology of a kernel-based PLS technique, though we have implemented both techniques and have shown the Kernel-PLS to be more effective in our scenario.

Kernel PLS was initially proposed by Roman Rosipal *et al.* [Rosipal *et al.,* 2001]. It has been proved to be very successful in expressing the interdependence between the multivariate input and the output data set in a very generalized and concise manner [Janne *et al.,* 2001]. This method is particularly useful in situations where the dimension of the input space is significantly greater than the number of training samples used for modeling. The main advantage of a Kernel based PLS method is that it can incorporate the non-linear relationship between the input and the output data set. We have shown the performance of Radial Basis Function (RBF) kernel in our algorithm and compared its performance with linear (no-kernel) PLS techniques.

Kernel-PLS belongs to the class of kernel-based non-linear regression techniques. Any linear regression method finds the optimum linear regression surface in the multidimensional input and output space through minimization of the least square error. The optimality is defined only in the linear space. Hence, any non-linear relationship between the features are ignored (approximated by a linear surface in the multidimensional space). In a kernel-based scheme, input space is transformed into a higher dimensional feature space and linear regression analysis is performed in higher dimensional feature space which projects to non-linearity in the original space of observables. The correlation in feature space is represented by a kernel defined entirely in terms of input space variables. The advantage of a

kernel method is the execution of the regression analysis in the higher dimensional space without explicitly defining the transformation. A non-linear transformation from input variable space X to feature space F is done through a non-linear mapping $\phi : x_i \in R^n \rightarrow \phi(x_i) \in F, i=1,...,n$ where n is the size of the input data set. Our objective is to construct a linear PLS regression model in F. This is achieved in terms of the latent variables in the feature space. The method is described as follows:

1. intialize u
2. $t=\phi\phi^T$
3. $c=Y^T t$
4. $u=Yc, u \leftarrow u/\|u\|$
5. *repeat steps* 2.–5.*until convergence*
6. *deflate* $\phi\phi^T$, *Y matrices* : $\phi\phi^T \leftarrow (\phi - tt^T)(\phi - tt^T\phi)^T$,
   $Y \leftarrow Y - tt^T Y$

Note that u,t and c are latent variables in the feature space.

Applying the so-called 'kernel-trick' where $K(x_i, x_j) = \phi(x_i)\ \phi(x_j)$, we can effectively bypass the explicit transformation of input variables to the higher dimensional feature space. The equivalent equation in terms of kernel matrix K is given by

$$K \leftarrow (I - tt')K(I - tt') \qquad [1]$$

For prediction based on testing point $\{x_i\}_{i=1}^{m}$, the predicted output vectors are calculated as given by the following equation:

$$Y_t = \phi_t B = K_t U(T' KU)^{-1} T' Y \qquad [2]$$

In any kernel method, one transforms the input space onto a higher dimensional feature space and subsequently tries to achieve classification or regression in that space. For most non-separable data there exists a higher dimensional mapping for which transformed data fall on the linear regression surface. The motivation behind kernel-PLS is the inherent non-linearity that exists in the basic mechanism of most of the biochemical reactions in living cells.

4. PERFORMANCE ANALYSES

The aim of our research is to extract the underlying functional gene units that explain variation in drug response. PLS is shown to be efficient in extracting the latent features from both the gene expression profile and the drug response pattern. PLS is an iterative greedy algorithm that extracts at each step the latent component pair from the input and the output data set that exhibits maximum correlation. The latent components are either a linear or non-linear (kernel method) combination of the measured observables. Figure 1 shows the percentage reconstruction of gene expression and drug activity patterns using linear PLS (no-kernel) and that 20 PLS components are sufficient to explain 95% variation in drug response.

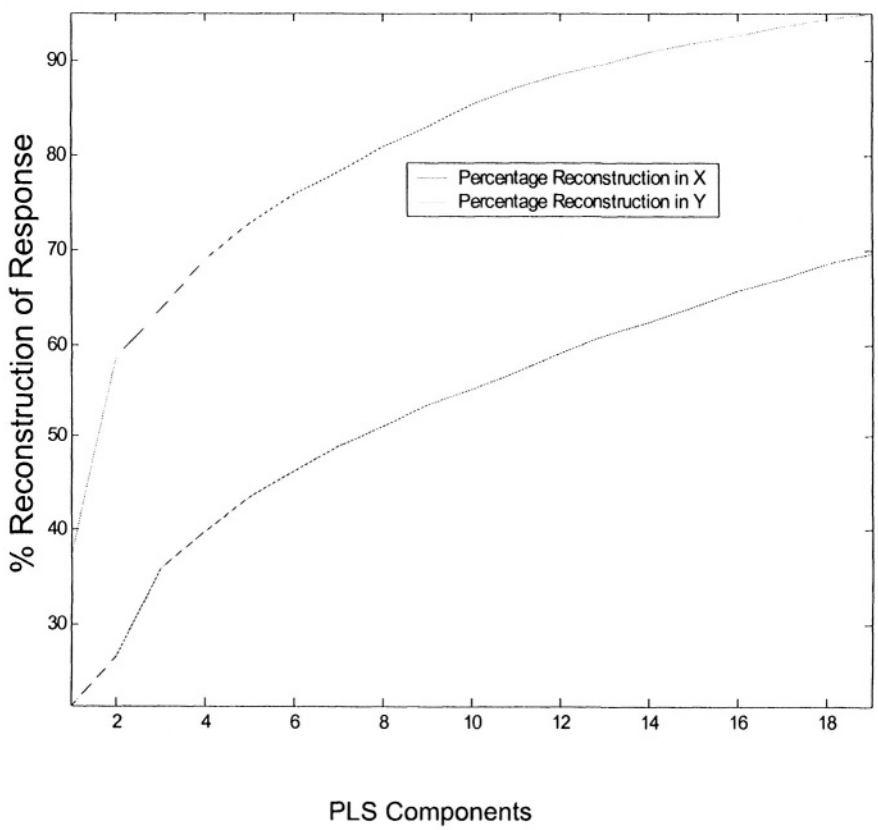


Figure 1. Percentage reconstruction of drug response as a function of ordinary (linear) PLS components.

This provides evidence that a small group of genes (functional units) play key roles in explaining a biochemical reaction, hence the drug responses. With this insight, we focused our attention to a particular class of drugs, called Taxols. These drugs form a subset of the entire drug activity pattern data set. Figure 2 represents the linear regression plot between the most significant latent feature (1st PLS component) extracted from the gene expression data set with the average Taxol response over 60 cell lines. A strong correlation suggests that Taxol drugs act preferentially on a group of genes that play a significant role inhibiting the growth of microtubules.

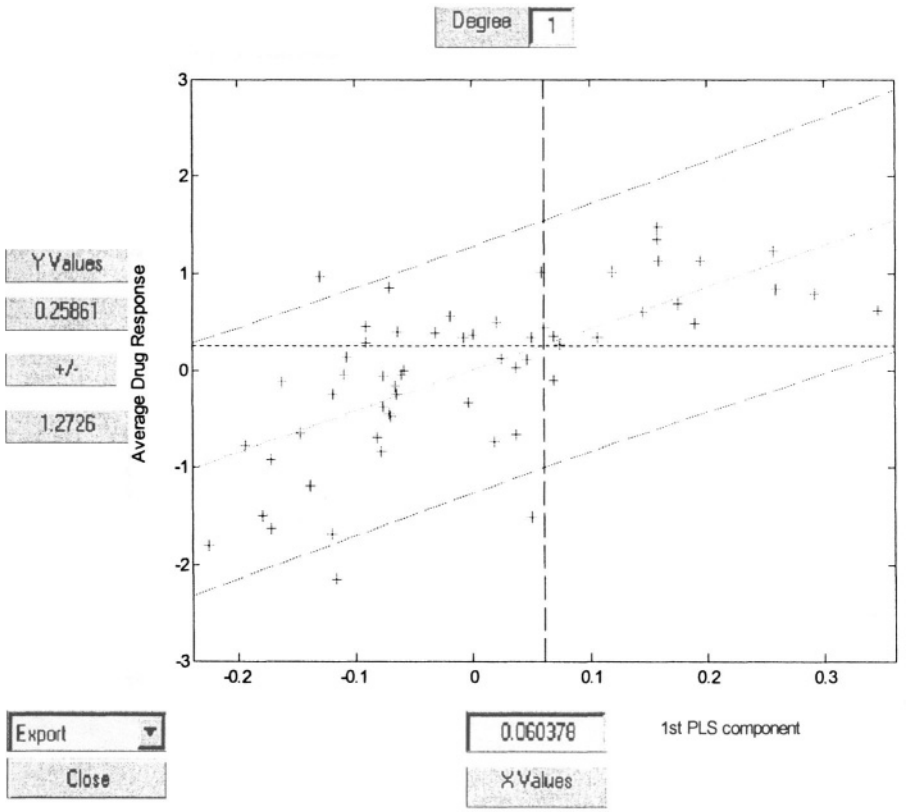


Figure 2. Linear Regression between average Taxol response and 1st PLS component.

Figure 3 shows a regression plot of average Taxol response as a function of the first two PLS components, and these results support our belief. The extracted PLS components are linearly independent of each other. Hence we could shift our attention from the whole set of genes to individual groups that work independently. Each PLS component refers to a weight vector in

the multidimensional gene space that represents the significance of each gene in the corresponding PLS component.

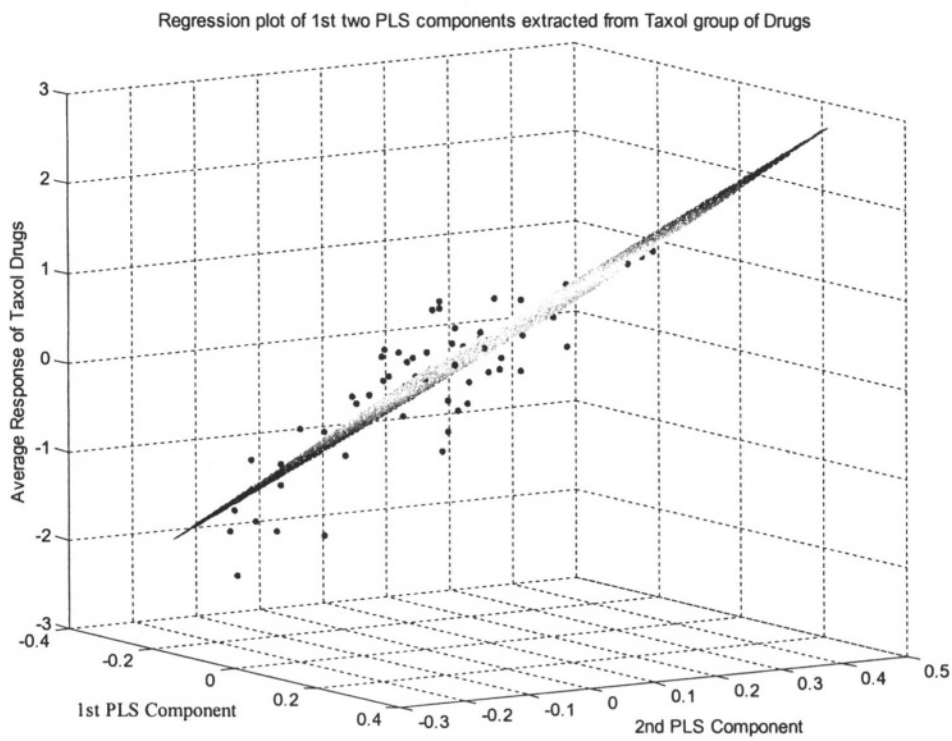


Figure 3. Regression surface between average Taxol response and first two PLS components.

Figure 4 shows the plot of eigenvalues as a function of the PLS component index. It is evident from the plot that one could terminate the PLS extraction process after 5 iterations. This also gives us an idea of how many PLS components should be considered in explaining a drug behavior.

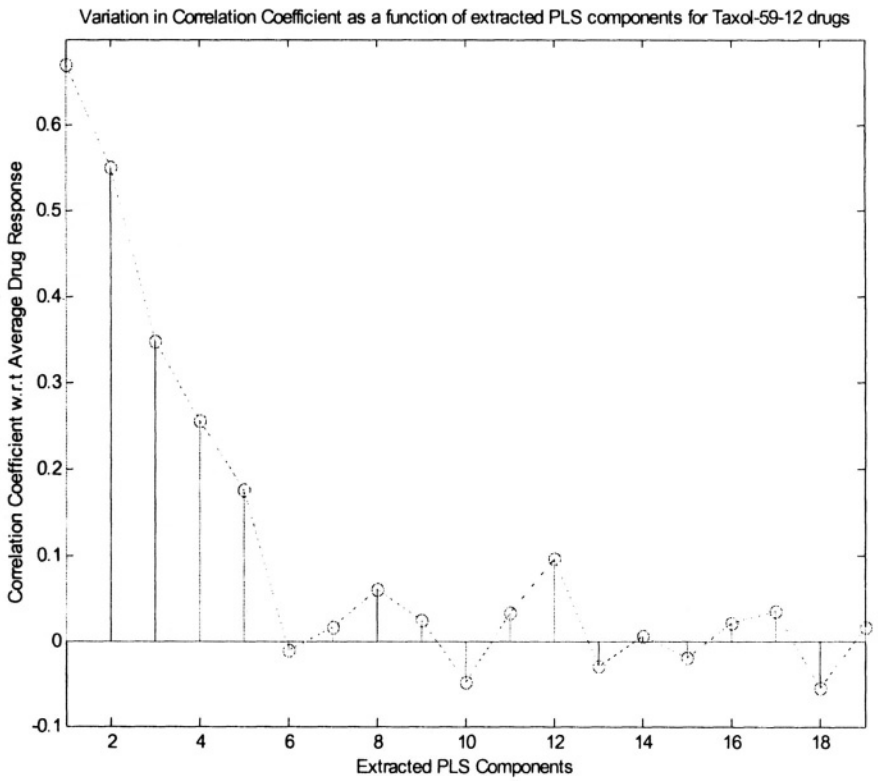


Figure 4. Plot of eigenvalues as a function of PLS component index.

We have extracted the significant genes from the first five PLS components and plotted the percentage reconstruction in Figure 5 using only the extracted genes. The plot shows little difference, confirming our belief that PLS could be used to extract important genes from a pool of genes expressions. Figure 6 represents a regression plot similar to Figure 3 using only the extracted genes.

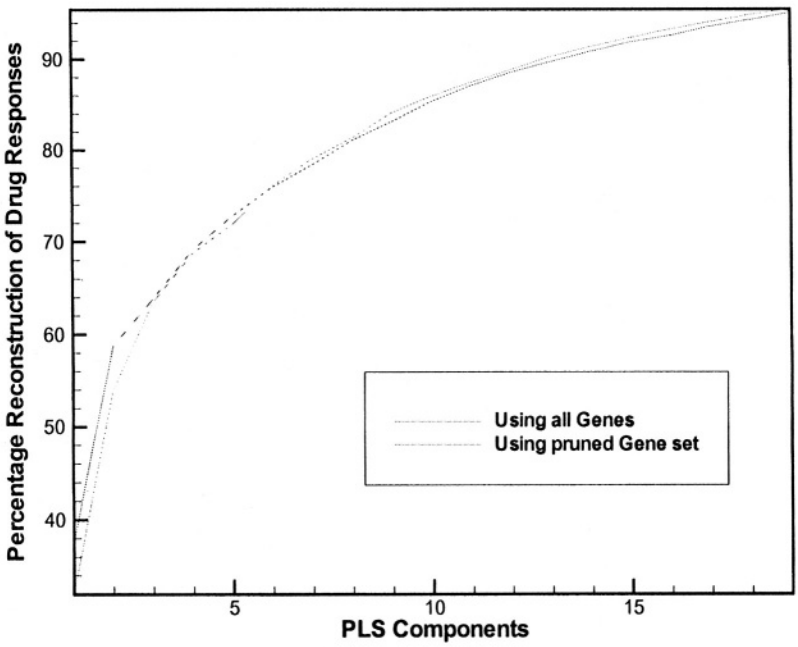


Figure 5. Percentage reconstruction using only extracted genes.

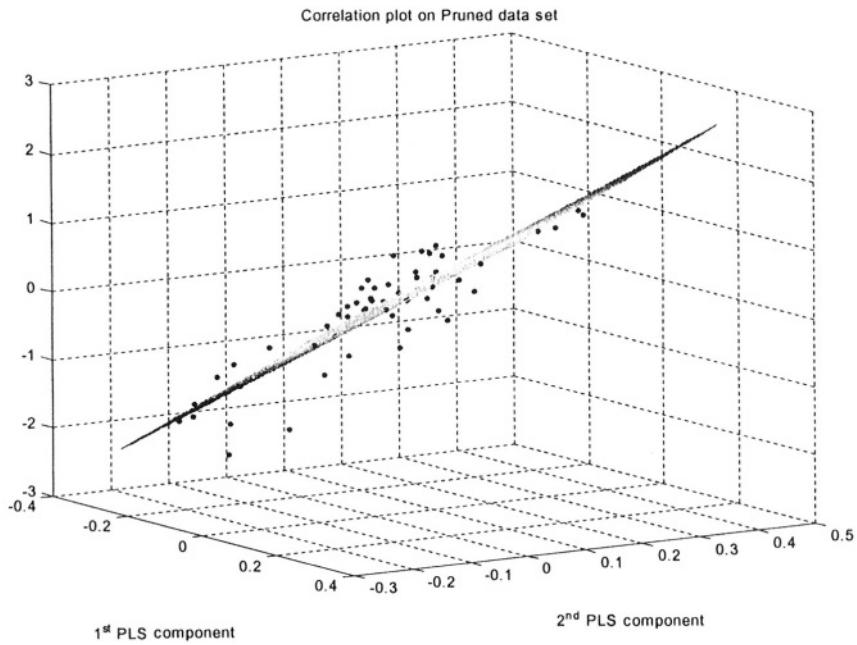


Figure 6. Plot similar to Figure 3 using only extracted genes.

The kernel method is thought to be more effective in explaining pharmacogenomics in a quantitative manner. Figure 7 shows the percentage reconstruction of the drug response using RBK kernel and no kernel. It is apparent from the plot that, by using non-linear methods, one could explain the variations in the drug response in a more concise manner. Figure 8 shows the linear regression plot of the 1st PLS component extracted using an RBF kernel as a function of average Taxol response. It shows a strong positive correlation, whereas we find almost no correlation with subsequent PLS components.

We believe that since non-linear interaction is allowed in the kernel-based approach, it could explain the complex interdependence between genes using a single PLS component. Hence the rest of the PLS components are redundant. Therefore, we infer that the complex correlation between different genes and drug activity pattern can be explained by kernel-PLS in a more accurate and efficient way. The PLS approach, as described here, is more biologically motivated and considered to be superior to other clustering algorithms. First of all, PLS tries to extract the latent genomic components which are maximally decorrelated. Thus, they are expected to explain independent biological events in a cell or a tissue. Secondly, any particular gene could be a member of more than one genomic component or class that is not allowed by other similarity criteria-based clustering algorithms. Also, the number of clusters are predefined in traditional clustering schemes whereas PLS defines the genomic clusters “on the fly.”

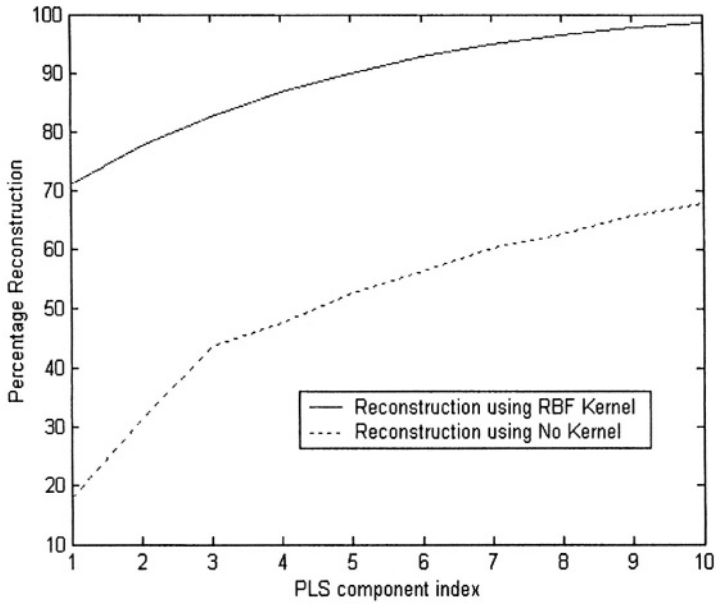


Figure 7. Percentage reconstruction comparison between RBF and no kernel.

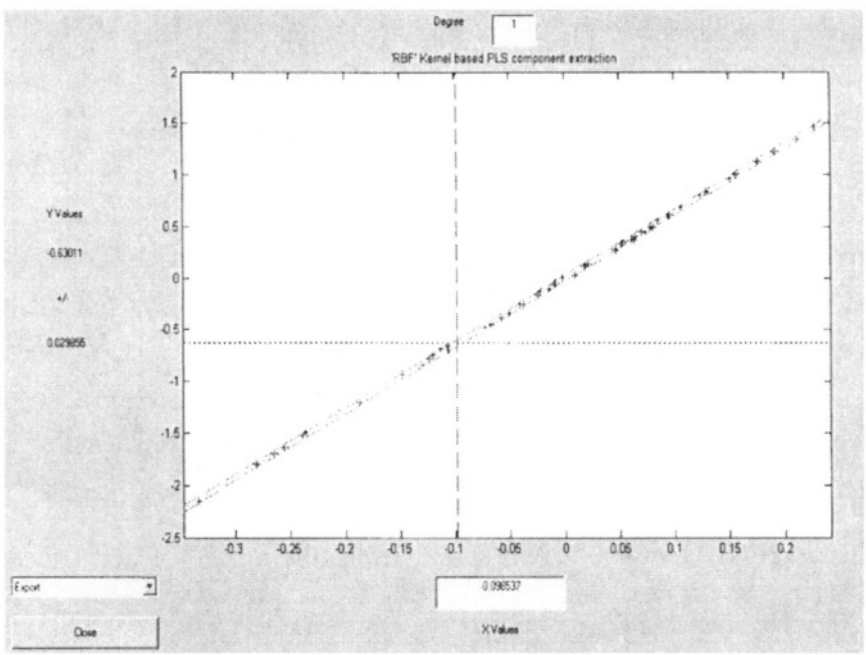


Figure 8. Linear Regression between average Taxol response with the first PLS component using RBF kernel.

Mathematically, each PLS component is a combination of individual genes. To test if these combinations correspond to any previous biological knowledge, we dissect each component by matching them with the gene co-occurrence networks obtained by a PubGene literature search [Jenssen *et al.*, 2000]. Although these reconstructed gene networks from literature co-occurrence differ from the real networks operating in the cells, it is a feasible model. If the PLS components are able to discover the co-regulated genes, we would expect to reveal a highly connected network from a PubGene search. As shown in Figure 9, the first PLS component includes signalling genes such as MYC and NDRG1 and cytoskeleton related genes such as COL1A1 and COL4A1. The second PLS component in Figure 10 includes another set of connected genes such as VEGF, ERG1 and SPOCK.

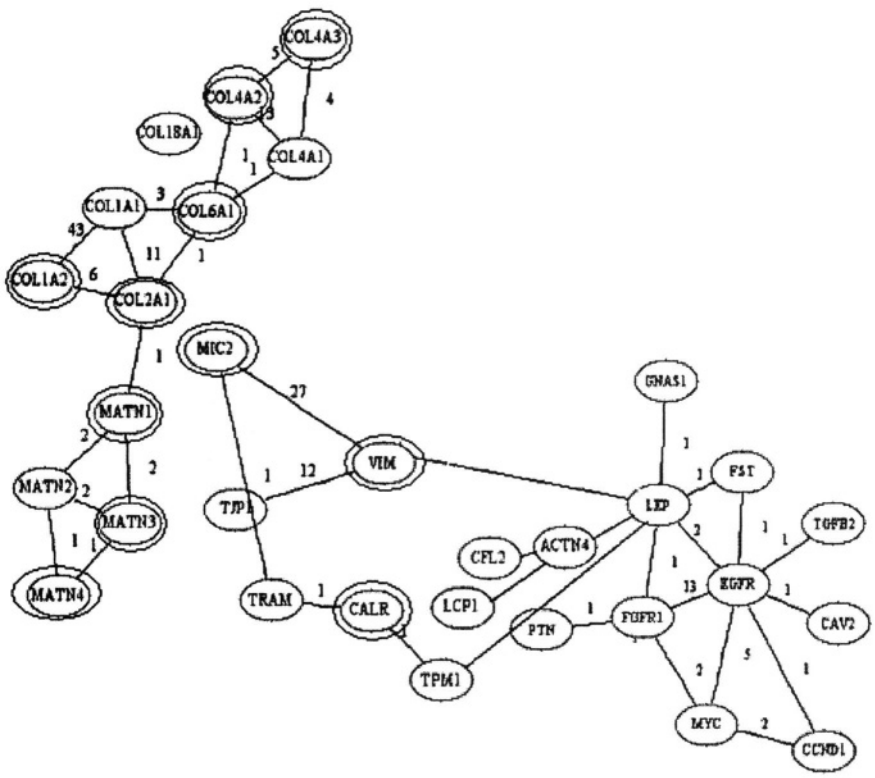


Figure 9. Literature Network of genes with genes extracted from the first PLS component. Each node in Figure 9 and 10 is a gene by the HUGO symbols. The connections between the nodes is based on literature co-occurrence calculated by PubGene. All the nodes (genes) except those within double ellipse are revealed by the PLS modeling. COL1A1: collagen, type I, alpha 1; COL4A1: collagen, type IV, alpha 1; COL18A1: collagen, type XVIII, alpha

1; COL2A1: collagen, type II, alpha; COL1A2: collagen, type I, alpha 2; COL4A2 collagen, type IV, alpha 2; COL6A1: collagen, type VI, alpha 1; COL4A3: collagen, type IV, alpha 3; MATN1: matrilin 1, cartilage matrix protein; MATN2: matrilin 2; MATN3: matrilin 3; MATN4: matrilin 4; TRAM: translocating chain-associating membrane protein; CALR: calreticulin; MIC2: antigen identified by monoclonal antibodies 12E7, F21 and O13; VIM: vimentin; TJP1: tight junction protein (zona occludens 1); AMBP: alpha-1-microglobulin/bikunin precursor; TPM1 tropomyosin (alpha); ACTN4: actinin, alpha 4; LCP1 lymphocyte cytosolic protein (actin-binding); CFL2 cofilin; LEP leptin receptor gene related protein; EGFR: epidermal growth factor receptor (avian erythroblastic leukemia viral (v-erb-b) oncogene homolog); CCND1: cyclin; MYC: v-myc avian myelocytomatosis viral oncogene homolog; NDRG1: N-myc downstream regulated gene; CAV2: caveolin 2; PTN: pleiotrophin (heparin binding growth factor 8, neurite growth-promoting factor 1); TGFB2: transforming growth factor, beta 2; FGFR1: fibroblast growth factor receptor; GNAS1: guanine nucleotide binding protein (G protein), alpha stimulating activity polypeptide.

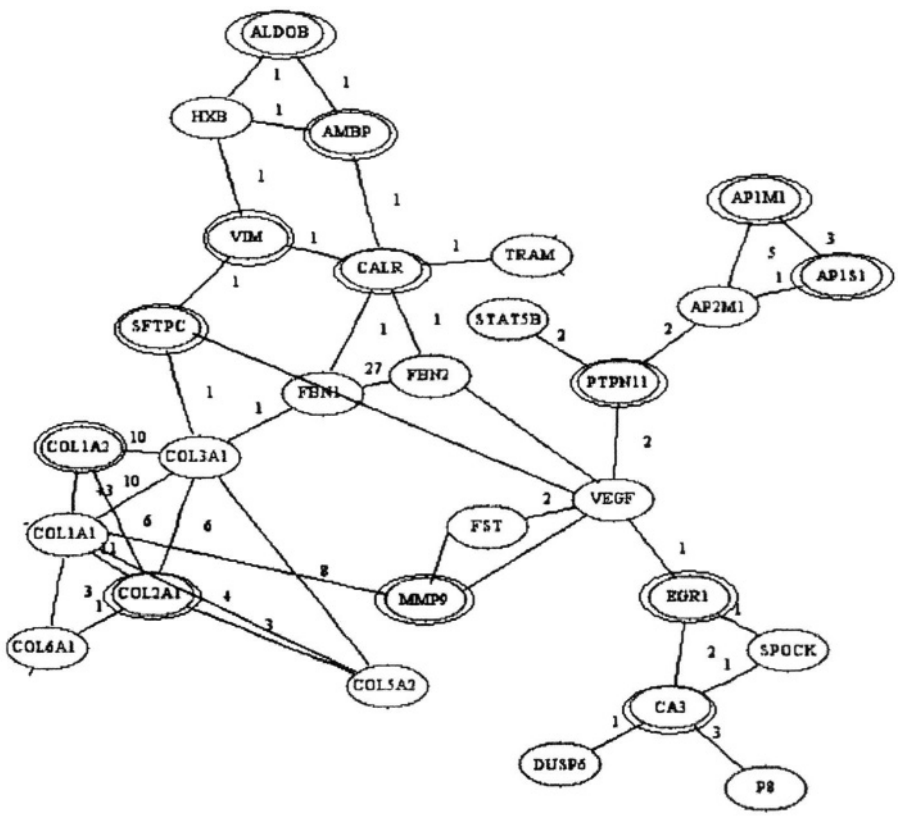


Figure 10. Literature Network of genes with extracted genes from the second PLS component. All the genes except those under double ellipse are revealed by the second PLS component. COL1A1: collagen, type I, alpha 1; COL2A1: collagen, type II, alpha; COL1A2 collagen, type I, alpha 2; COL3A1: collagen, type III,

alpha; COL5A2: collagen, type V, alpha 2; COL6A1: collagen, type VI, alpha 1; FBN1: fibrillin; FBN2: fibrillin; VIM: vimentin; AMBP: alpha-1-microglobulin/bikunin precursor; HXB: hexabrachion (tenascin C, cytotactin); CALR: calreticulin; ALDOB: aldolase B, fructose-bisphosphate; TRAM: translocating chain-associating membrane protein; SFTPC: surfactant, pulmonary-associated protein C; VEGF: vascular endothelial growth factor; FST: follistatin; STAT5B: signal transducer and activator of transcription 5B; MMP9: matrix metalloproteinase 9; PTPN11 protein tyrosine phosphatase, non-receptor type 11; AP2M1: adaptor-related protein complex 2, mu subunit; EGR1: early growth response 1; CA3: carbonic anhydrase III, muscle specific; SPOCK: sparc/osteonectin, cwcv and kazal-like domains proteoglycan (testican); P8: p8 protein (candidate of metastasis 1); DUSP6 dual specificity phosphatase6.

The role of many of these genes in taxol treatment has been indicated previously. For example, MYC is a previously identified oncogene regulating cell growth and proliferation. The interaction between MYC and microtubules has been shown by immunocoprecipitation and the regulation of MYC expression by microtubule system was suggested [Khyari *et al.*, 1997]. Further analysis of these results in a biological context could generate new insights and hypotheses. For example, some of the double ellipse genes, which indicate genes not revealed by PLS modeling, are not included in the microarray. Based on the tight connections with other involved genes, it should be interesting to investigate them on an individual basis by RT-PCR. In conclusion, PLS components not only offer good predictability but also offer satisfying explanations of the complex biological system.

5. DISCUSSION

Kernel-based techniques to achieve Principal Component Analysis (PCA) and Partial Least Squares (PLS) prove to be appropriate for microarray data analyses. In the situations of high multi-collinearity among regressors, Ordinary Least Squares (OLS) produces unbiased estimates of regression coefficients with high variance. This is also true in cases where numbers of observations are far less than the number of observed variables in the input space. The main advantage of the kernel-based PLS is the projection of the original regressors to the 'real' latent variables. This increases the noise immunity of the proposed modeling scheme. Using the standard 'kernel' trick, one effectively bypasses the actual transformation of the input vectors to the feature space. We have shown the Kernel-PLS approach to be more effective in explaining the drug variation and extracting underlying functional gene units. There are many potential applications to our research:

1. To recognize the method of action for a new drug.

2. To classify the cell lines and recognize a new unknown cell line based on its drug response.
3. To cluster genes in small working groups that work independently in a cell.
4. To make an informed decision on combination therapy to maximize the cytotoxic effect and minimize the drug-resistance potential.

ACKNOWLEDGEMENTS

Our thanks to Drs. Michael Colvin and David Adams of Duke University Medical Center for their helpful discussion on anticancer therapy, and Dr. John Weinstein of National Cancer Institute for the inspiration for kernel-based methods.

REFERENCES

Cristianini, N, Shaw-Taylor, J. *Support Vector Machines.* Cambridge University Press, 2000.

Falk, RF, Miller, NB. *A Primer for Soft Modeling.* The University of Akron Press, 1992.

Frank, IE, Friedman, JH. A Statistical View of Some Chemometric Regression Tools. *Technometrics* 35(2) (May 1993).

Hilko, Van der Voet. Comparing the predictive accuracy of models using a simple randomised test. *Chemometrics and Intelligent Laboratory Systems* 25 (1994): 313-323.

Hoskuldsson, A. PLS Regression Methods. *Journal of Chemometrics* 2 (1998): 211-228.

Janne, K, Pattersen, J, Lindberg, NO, Lundstedt, T. Hierarchical principal component analysis (PCA) and projection to latent structure (PLS) technique on spectroscopic data as a data pretreatment for calibration. *Journal of Chemometrics* 15 (2001): 203-213.

Jenssen, TK, Laegreid, A, Komorowski, J, Hovig, E. A literature network of human genes for high-throughput analysis of gene expression. *Nature Genetics* 28(1) (May 2000).

el Khyari, S, Bourgarel, V, Barra, Y, Braguer, D, Briand, C. Pretreatment by tubulin agents decreases C-MYC induction in human colon carcinoma cell line HT29-D4. *Biochemical & Biophysical Research Communications.* 231(3) (1997): 751-4.

Musumarra, G, Condorelli, DF, Scire, S, Costa, AS. Shortcuts in genome-scale cancer pharmacology research from multivariate analysis of the National Cancer Institute gene expression database. *Biochemical pharmacology 62* (2001): 547-553.

Ranner, S, Lindgren, F, Geladi, P, Wold, S. A PLS Kernel Algorithm for data sets with many variables and fewer objects - Part I: Theory and Algorithm. *Journal of Chemometrics* 8 (1994): 111-125.

Rosipal, R, Trejo, LJ. Kernel Partial Least Squares Regression in RKHS: Theory and Empirical Comparison. *Technical Report, University of Paisley* (March 2001).

Ross, TD, Scherf, U *et al.* Systematic variation in gene expression patterns in human cancer cell lines. *Nature Genetics* 24 (March 2000).

Tobias, R. *An Introduction to Partial Least Squares Regression.* TS-509. Gary, NC: SAS Institute Inc, April 1997

Wold, S, Ruhe, A, Wold, H, Dunn III, JW. The Collinearity Problem in Linear Regression. The Partial Least Squares (PLS) approach to generalized inverses. *Siam J. Sci. Stat. Comput.* 5(3) (Sept 1984).

This page intentionally left blank

11

ANALYSIS OF GENE EXPRESSION PROFILES AND DRUG ACTIVITY PATTERNS BY CLUSTERING AND BAYESIAN NETWORK LEARNING

Jeong-Ho Chang, Kyu-Baek Hwang, and Byoung-Tak Zhang
Biointelligence Laboratory, School of Computer Science and Engineering, Seoul National University, Seoul 151-744, Korea

Abstract: High-throughput genomic analysis provides insight into a complicated biological phenomena. However, the vast amount of data produced from up-to-date biological experimental processes needs appropriate data mining techniques to extract useful information. In this paper, we propose a method based on cluster analysis and Bayesian network learning for the molecular pharmacology of cancer. Specifically, the NCI60 dataset is analyzed by soft topographic vector quantization (STVQ) for cluster analysis and by Bayesian network learning for dependency analysis. Our results of the cluster analysis show that gene expression profiles are more related to the kind of cancer than to drug activity patterns. Dependency analysis using Bayesian networks reveals some biologically meaningful relationships among gene expression levels, drug activities, and cancer types, suggesting the usefulness of Bayesian network learning as a method for exploratory analysis of high-throughput genomic data.

Key words: Gene expression pattern, drug activity pattern, molecular pharmacology, soft topographic vector quantization (STVQ), Bayesian networks

1. INTRODUCTION

Recent developments in the technology for biological experiments have made it possible to produce massive biological datasets. For example, microarrays obtained from cDNA chips or oligonucleotide chips provide a parallel view of the expression pattern of tens of thousands of genes in a

sample. These massive datasets provide an opportunity to broaden the knowledge of the complex biological phenomena, but also require appropriate analysis techniques different from conventional methods for the traditional one-gene-in-one-experiment paradigm. Until now, diverse methods from the statistics and machine learning fields, such as hierarchical clustering [Eisen *et al.,* 1998], principal component analysis (PCA) [Raychaudhuri *et al.,* 2000], neural networks [Khan *et al.*, 2001], and Bayesian networks [Friedman *et al.,* 2000; Hartemink *et al.*, 2001; Hwang *et al.*, 2001], have been applied to high-throughput genomic analysis. In data analysis, it is most important to adopt the appropriate methods to the purpose of the analysis.

In this paper, the NCI60 dataset [Scherf *et al.*, 2000] is analyzed for the molecular pharmacology of cancer. The NCI60 dataset consists of 60 human cancer cell lines from 9 kinds of cancers, which are colorectal, renal, ovarian, breast, prostate, lung, and central nervous system origin cancers, as well as leukemias and melanomas. On each cell line, the gene expression pattern is measured by a cDNA microarray of 9,703 genes including ESTs. Also, 40 molecular targets other than mRNA are assessed. And 1,400 chemical compounds are tested on the 60 cell lines. These compounds include some anticancer drugs that are currently in clinical use. The drug activity on the cell line is measured by the growth inhibition assessed from changes in total cellular protein after 48 hours of drug treatment using sulphorhodamine B assay [Scherf *et al.*, 2000].

We use soft topographic vector quantization (STVQ) [Graepel, 1998] for cluster analysis and Bayesian network learning for dependency analysis. In the cluster analysis, 60 cell lines are clustered based on the gene expression patterns and drug activity patterns. Dependency analysis aims to model the probabilistic relationships among the expression level of each gene, the activity of each drug, and the kind of cancer.

The paper is organized as follows. In Section 2, the cluster analysis by STVQ is described. The dependency analysis by Bayesian network learning is described in Section 3. Finally, the conclusion and some directions for further research are given in Section 4.

2. CLUSTER ANALYSIS OF THE NCI60 DATASET

We have clustered the 60 human cancer cell lines of the NCI60 dataset based on gene expression patterns and drug activity patterns, respectively. In the experiments, we investigate if there is a common pattern in gene expression and drug activities of the cell lines from the same tissue of origin, and thus if cell lines of the same cancer type can be clustered appropriately.

2.1 Soft Topographic Vector Quantization

Soft topographic vector quantization (STVQ) [Graepel, 1998] is a clustering algorithm based on principles from statistical physics. It can provide not only a stable and good clustering solution, but also a topographic map of the clustered data.

In this algorithm, clustering is defined in terms of an optimisation problem. The cost function to be optimised is given as

$$E = \sum_{i=1}^{N} \sum_{j=1}^{M} m_{ij} e_{ij}, \qquad [1]$$

where N is the number of samples and M is the number of clusters. m_{ij} is a binary variable indicating whether the i^{th} sample belongs to the j^{th} cluster, and e_{ij} is the error occurred by assigning the i^{th} sample to the j^{th} cluster. The error term is defined as

$$e_{ij} = \frac{1}{2} \sum_{k=1}^{M} h_{jk} \left\| \mathbf{x}_i - \mathbf{z}_k \right\|^2, \quad \sum_{k=1}^{M} h_{jk} = 1 \quad (\forall j), \qquad [2]$$

where $\mathbf{x}_i$ is a sample vector and $\mathbf{z}_k$ is a cluster centre whose value is determined by the average of the sample vectors assigned to it. h_{jk} is a neighbourhood function between j^{th} and k^{th} clusters. By introducing h_{jk} for every pair of clusters, STVQ is able to visualize the cluster structure in the same way as the self-organizing map (SOM) does in the one- or two-dimensional space.

STVQ provides an efficient procedure to find a good solution to the minimization of Equation 1 based on the maximum entropy principle and the idea of deterministic annealing. It is initialised with a random configuration as a K-means algorithm and proceeds using an iterative optimisation method, the EM algorithm [Dempster *et al.,* 1977], with some annealing schedule. In the *E*-step, the expectation value of m_{ij}, that is the probability that the sample $\mathbf{x}_i$ belongs to the j^{th} cluster, is estimated for each pair of samples and clusters. Then, all the cluster centres are calculated in the *M*-step. These two steps are iteratively alternated until convergence. More details about STVQ can be found in [Graepel, 1998].

2.2 Clustering of the NCI60 Cell Lines Using STVQ

The NCI60 dataset comprises two matrices, called the **T** matrix and the **A** matrix. In the **T** matrix, each cell line is represented by 1,416 attributes that

include 1,376 genes and 40 molecular characteristics. The 1,376 genes are those with strong patterns of variation among the cell lines and with less than or equal to 4 missing values [Scherf *et al.,* 2000]. Each cell line in the **A** matrix is represented by the activity values of 1,400 chemical compounds.

For each cell line in the **T** matrix, all of its attribute values were standardized (mean value is 0 and the standard deviation is 1) across 1,416 attributes, including genes and individual targets. Likewise, all the drug activity values of each cell line in the **A** matrix were standardized. Now, each cell line is represented as a vector, where the vector $\mathbf{x}_i$ corresponds to the i^{th} cell line.

First, we have clustered the 60 cell lines based on the gene expression profiles. For each cluster centre $\mathbf{z}_k$, all of its attribute values are standardized after every update. Then the squared Euclidean distance in Equation 2 is closely related with the Pearson correlation coefficient. That is,

$$\begin{aligned} \|\mathbf{x}_i - \mathbf{z}_k\|^2 &= (\mathbf{x}_i^T\mathbf{x}_i + \mathbf{z}_k^T\mathbf{z}_k - 2\mathbf{x}_i^T\mathbf{z}_k) \\ &= \left(2D - 2D \times \frac{\mathbf{x}_i^T\mathbf{z}_k}{D}\right) = 2D(1 - r_{ik}), \end{aligned} \quad [3]$$

where D is the number of attributes of $\mathbf{x}_i$ and $\mathbf{z}_k$, and r_{ik} is the Pearson correlation coefficient for $\mathbf{x}_i$ and $\mathbf{z}_k$. Based on this relation, we have used the squared Euclidean distance scaled by $1/D$ as the distance between $\mathbf{x}_i$ and $\mathbf{z}_k$, and the error term in Equation 2 is equivalent to

$$e_{ij} = \frac{1}{2}\sum_{k=1}^{M} h_{jk} \frac{\|\mathbf{x}_i - \mathbf{z}_k\|^2}{D} = \sum_{k=1}^{M} h_{jk}(1 - r_{ik}) \quad [4]$$

The cell lines have been clustered with varying number of clusters, that is 9, 16, and 25. The result with 16 clusters is shown in Figure 1(a). It can be seen that each cluster or nearby clusters appropriately reflect the organ of origin of its constituent, especially for the leukemias (LE), the colon cancer lines (CO), the CNS lines, the renal carcinoma lines (RE), and the melanoma lines (ME).

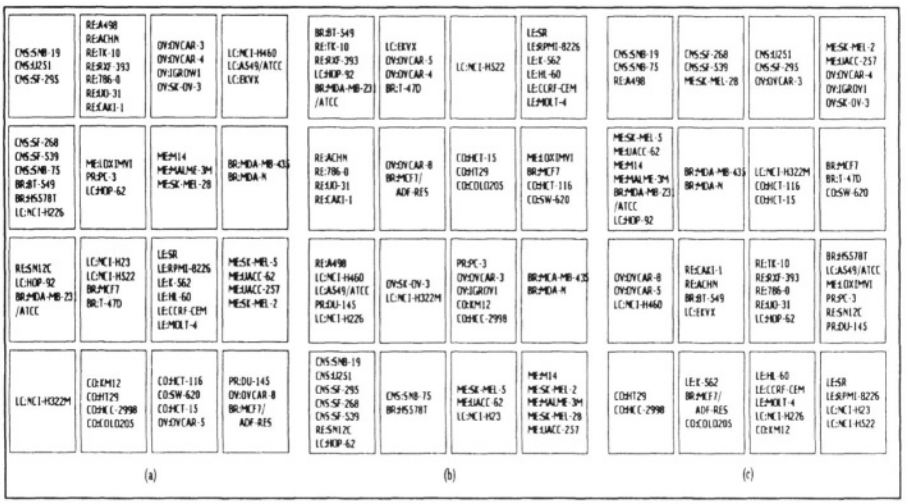


Figure 1. The results of cell line clustering: (a) based on gene expression profiles (α = 0.0), (b) based on interpolated distance (α = 0.7), and (c) based on drug activity patterns (α = 1.0). The value of h_{jk} is inversely proportional to the Euclidean distance between j^{th} and k^{th} clusters, where each cluster is represented as a discrete position in the two-dimensional lattice. In this 4×4 lattice, the cluster in the upper-left corner is encoded as (0, 0) and that in the lower-right corner as (3, 3). Clusters at the corners and ends are not neighbouring each other in view of Euclidean distance between the coordinates in the lattice.

We then ask, will the cell lines from the same tissue of origin show similar patterns in drug activities, such that they appear in the same or nearby clusters? To investigate if this is the case, we have clustered the cell lines based on both gene expression profiles and drug activity patterns. The error occurred by assigning a cell line to a particular cluster is defined as

$$e_{ij} = \frac{1}{2}\sum_{k=1}^{M} h_{jk}\left[(1-\alpha)\left\|\mathbf{x}_i^g - \mathbf{z}_k^g\right\|^2 + \alpha\left\|\mathbf{x}_i^d - \mathbf{z}_k^d\right\|^2\right], \qquad (0 \le \alpha \le 1) \qquad [5]$$

where the cell line $\mathbf{x}_i^g$ and the cluster $\mathbf{z}_k^g$ are related with gene expression profiles, and $\mathbf{x}_i^d$ and $\mathbf{z}_k^d$ with the drug activity patterns. The constant α is used to interpolate two distances based on the gene expression profiles and drug activity patterns.

Two criteria were used to measure the quality of the clustering results: the average Pearson correlation coefficient R and the average entropy H across all the clusters. They are defined as

$$R = \sum_{j=1}^{M} \frac{N_j}{N}\left[\frac{2}{N_j(N_j-1)}\sum_{i<k} r_{ik}^j\right], \qquad [6]$$

$$H = \sum_{j=1}^{M} \frac{N_j}{N} \left[-\sum_{k=1}^{C} \frac{N_{jk}}{N_j} \log\left(\frac{N_{jk}}{N_j} \right) \right], \quad [7]$$

where N is the number of cell lines, M is the number of clusters, and C is the number of tissues of origin. N_j represents the number of cell lines assigned to the j^{th} cluster, and N_{jk} the number of cell lines from the k^{th} organ of origin in the j^{th} cluster. The value in the bracket in Equation 6 is the average Pearson correlation coefficient across all the pairs of cell lines in the same cluster and that in Equation 7 represents the entropy in a cluster. When the cluster size is fixed, the higher value of entropy H means that the cluster structure is less reflective of the tissue of origin of the cell lines. In the case of the Pearson correlation coefficient R, the higher value means a better quality of clustering result in terms of inner cluster similarity.

Figure 2 shows the variation of the values of R and H in clustering of the cell lines, respectively, with varying α values in Equation 5. It can be seen that, with the higher value of α, the value of R based on gene expression profiles gets lower and the value based on drug activity patterns gets higher, showing the opposite trends between the two cases. In the case of the average entropy, as the value of α increases, the entropy has a tendency of being higher (for 16 clusters, from 0.40 to 0.72), and thus the quality of clustering becomes worse.

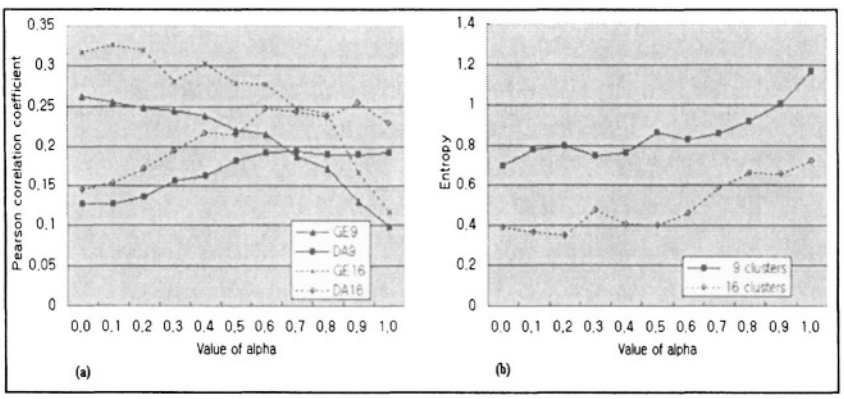


Figure 2. Values of the two measures of clustering quality over varying α. (a) Averaged Pearson correlation coefficients for 9 and 16 clusters. GE: The Pearson correlation coefficient based on gene expression profiles. DA: The coefficient based on drug activity patterns. (b) Averaged clustering entropies for 9 and 16 clusters. Only cancer types of the constituents in a cluster are considered, so just one graph suffices for each experiment.

From these two results, we can see that, in general, the similarity in gene expression profiles among a set of cell lines does not necessarily relate to a

similarity in drug activity patterns among the cell lines. Also, the drug activity patterns are less related to the organ of origin, when compared with the gene expression profiles.

The cluster structure of the 60 cell lines on the basis of drug activity patterns only, that is $\alpha = 1.0$ in Equation 5, is shown in Figure l(c). As also indicated by the value of average entropy, the cluster structure of the cell lines can be seen to be more heterogeneous than the result based on the gene expression profiles only. And Figure l(b) shows a compromised solution with $\alpha = 0.7$. In [Scherf *et al.,* 2000], it has been proposed that this heterogeneity might be partly due to the activity of genes related to drug sensitivity and resistance, which has been supported by the fact that several cell lines with a relatively high expression level of multi-drug resistance gene *ABCB1* have been clustered in the same group. Inspired by our clustering results and the proposal, we have tried analysing the relationships among the activities of anticancer drugs and the expression levels of the genes by Bayesian network learning.

3. DEPENDENCY ANALYSIS USING BAYESIAN NETWORK LEARNING

3.1 Bayesian Networks

A Bayesian network [Heckerman, 1999] is a probabilistic graphical model that represents the joint probability distribution over a number of random variables. For an efficient representation, conditional independencies among the variables are exploited. These conditional independencies are encoded by a DAG (directed acyclic graph) structure in which a node corresponds to a random variable. The joint probability distribution over a set of n random variables $\mathbf{X} = \{X_1, X_2, \ldots, X_n\}$, given the Bayesian network for $\mathbf{X}$, is described as follows:

$$P(\mathbf{X}) = \prod_{i=1}^{n} P(X_i \mid \mathbf{Pa}(X_i)), \qquad [8]$$

where $\mathbf{Pa}(X_i)$ is the set of parents of node X_i in the Bayesian network structure. $P(X_i \mid \mathbf{Pa}(X_i))$ in the above equation is called the local probability distribution for X_i. Typically, the linear Gaussian model for continuous variables and the multinomial model for discrete variables are used for modeling the local probability distribution.

Learning Bayesian networks from data consists of two parts: learning the network structure and learning the local probability distribution for each node in the given structure. The second part corresponds to a simple calculation under some reasonable assumptions [Heckerman, 1999]. A popular approach to structural learning is the score-based search. The search space is nevertheless super-exponential in the number of variables. Because it is nearly impossible to find the best-scoring network structure even in a moderate case (7 or 8 variables), several search heuristics such as greedy search, greedy search with random restarts, and simulated annealing are used in practice [Heckerman, 1999]. In this paper, the greedy search algorithm and another search heuristic for hundreds of variables with the BD (Bayesian Dirichlet) scoring metric [Heckerman *et al.,* 1995] are used to learn Bayesian networks from the NCI60 dataset.

3.2 Applying Bayesian Networks to the Analysis of NCI60 Dataset

The NCI60 dataset contains gene expression patterns (T matrix) and drug activity patterns (A matrix) for 9 different cancer types [Scherf *et al.,* 2000]. To model the probabilistic relationships among them, we use a Bayesian network where each node corresponds to each variable. In the Bayesian network learning, the **T** matrix and **A** matrix are combined together, so that each cell line sample has gene expression levels and drug activities as its attributes.

3.2.1 Pre-Processing of the Dataset

The experiments focus on the 1,376 genes and 118 drugs as in the analysis of gene-drug correlations by Scherf *et al.* [2000]. Furthermore, genes and drugs that have more than 3 missing values across 60 samples, as well as unknown ESTs, were eliminated for robust analysis. Consequently, the analyzed NCI60 dataset includes 60 samples with 890 attributes (805 gene expression levels, 84 drug activities, and one additional variable for the kind of cancer).

The number of attributes is extremely large compared to the number of samples. This might cause problems, such as a seriously slow learning speed, low confidence in learned models, and infeasibility of probabilistic inference. To cope with these problems, the number of attributes is reduced in two ways. One is to use prototypes of attributes. Genes and drugs are clustered respectively and the centre of each cluster is regarded as an attribute. The other is attribute selection. Here, all the genes and drugs are clustered together and all the members of some adjacent clusters are selected to construct the Bayesian network for the specific purpose of the analysis.

The soft topographic vector quantization (STVQ) described in Section 2 is used for clustering.

All the continuous attribute values were discretized into three levels (low, normal, and high) for the multinomial local probability distribution model of the Bayesian network. The multinomial model is chosen because of its expressive power although discretization might cause some information loss. Two discretization boundary values for each attribute are calculated as $\mu + c \cdot \sigma$ and $\mu - c \cdot \sigma$. Here, μ is the mean value and σ is the standard deviation of the attribute across 60 samples. c is a constant, which determines the distribution ratio of the original values in low, normal, and high.

3.2.2 A Fast Search Heuristic for Bayesian Network Learning

A general greedy search algorithm [Heckerman, 1999] is nearly inapplicable to learning Bayesian networks which consist of hundreds of nodes. Friedman *et al.* [1999] suggest a fast search heuristic for such cases and a similar approach is adopted in the experiments. The "local to global" heuristic is a kind of greedy search algorithm. Here, the search space is reduced by learning the structure around each node within small bounds before performing the greedy search procedure. The bounds are based on the concept of a Markov blanket [Pearl, 1988]. The Markov blanket of a variable satisfies the following.

$$P(X_i \mid \mathbf{X} - X_i) = P(X_i \mid \mathbf{BL}(X_i)), \qquad \mathbf{BL}(X_i) \subseteq \mathbf{X} - X_i, \qquad [9]$$

where $\mathbf{X}$ is the set of all the variables and $\mathbf{BL}(X_i)$ is the Markov blanket of X_i. Because the Markov blanket size of each node is unknown, the maximum size is pre-specified. Although the "local to global" heuristic is not guaranteed to find a good-scoring network in all cases, the learning speed is much faster than a general greedy search algorithm in the case of learning Bayesian networks with hundreds of nodes.

3.3 Experimental Results

Experimental results on the original dataset (Dataset 1), one reduced dataset with prototypes (Dataset 2), and another reduced dataset with selected attributes (Dataset 3) are given here. Table 1 shows the properties of these three datasets with applied learning methods, learning time, and the applicability of probabilistic inference. This table describes the properties of three datasets with respect to the applied learning methods, learning time, and the applicability of probabilistic inference. Samples in Dataset 2 have

gene prototypes and drug prototypes as attributes. Dataset 1 is too large to apply the general greedy search algorithm. Dataset 3 is so small that the "local to global" heuristics are not required. Microsoft MSBN software - http://research.microsoft.com/research/dtg/msbn/OldMSBN.htm - was used for probabilistic inference in the analysis. The average learning time is measured on a Pentium III 1GHz machine.

Table 1. The properties of three datasets with respect to the applied learning methods, learning time, and the applicability of probabilistic inference. The numbers in the parentheses of the forth column represent the number of runs of the greedy search algorithm with random initialisations. The numbers in the parentheses of the fifth column represent the used maximum Markov blanket sizes. The rightmost column shows the applicability of probabilistic inference to the Bayesian networks learned from each dataset.

	# of genes	# of drugs	Greedy search	"Local to global" heuristics	Learning time in avg. (secs)	Prob. inference
Dataset 1	805	84	"—"	O (5 ~ 8)	3233.7	no
Dataset 2	40	5	O (20)	O (5 ~ 15)	123.9	yes
Dataset 3	12	4	O (100)	"—"	15.6	yes

3.3.1 Experimental Results on the Original Dataset

Three Bayesian networks were learned from the original dataset according to three different discretization boundaries (c = 0.43, 0.50, and 0.60). Probabilistic inference from the Bayesian network with 890 nodes is nearly impossible. Hence, only the number of edges connected to each node is analyzed here. An edge represents direct probabilistic dependency and the node with many edges is considered to be related to many other nodes. Table 2 lists the top ten nodes that are most related to others on average in three Bayesian networks. The most related one is the cancer type node. The other nine nodes are all for genes. The results seem to be reasonable since the strong relationship between gene expression patterns and the kind of cancer is discovered from the cluster analysis in Section 2.

To investigate the influence of different discretization boundaries on the analysis, the Pearson correlation coefficient (r_{ij}) among the numbers of edges of all the nodes in two Bayesian networks was calculated as follows:

$$r_{ij} = \frac{\sum_{k=1}^{890} n_{ki} n_{kj} - \frac{1}{890} \sum_{k=1}^{890} n_{ki} \sum_{k=1}^{890} n_{kj}}{\sqrt{\sum_{k=1}^{890} n_{ki}^2 - \frac{1}{890}\left(\sum_{k=1}^{890} n_{ki}\right)^2} \cdot \sqrt{\sum_{k=1}^{890} n_{kj}^2 - \frac{1}{890}\left(\sum_{k=1}^{890} n_{kj}\right)^2}}, \qquad [10]$$

where n_{ki} is the number of edges of node k in Bayesian network i and n_{kj} is the number of edges of the same node in Bayesian network j. The average value of r_{ij} among three Bayesian networks is 0.841. The number of edges of each node does not seem to be so much influenced by different discretization boundary values.

Table 2. Top ten nodes that are closely related to the others. The first is cancer type and the other nine nodes are all for genes. The average number of edges of each node over all 890 nodes is 5.21.

Description of node	The average number of edges
The kind of cancer	125
SID W 487878, SPARC/osteonectin [5':AA046533, 3':AA045463]	25
Homo sapiens Cyr61 mRNA, complete cds Chr.1 [486700, (DIW), 5':AA044451, 3':AA044574]	18.3
SID W 162479, Homo sapiens epithelial-specific transcription factor ESE-1b (ESE-1) mRNA, complete cds [5':H27938, 3':H27939]	16
CDH2 Cadherin 2, N-cadherin (neuronal) Chr. [325182, (DIRW), 5':W48793, 3':W49619]	13.7
H.sapiens mitogen inducible gene mig-2, complete CDS Chr.14 [488643, (IW), 5':AA045936, 3':AA045821]	13.3
SID W 429623, Homo sapiens clone 24659 mRNA sequence [5':AA011634, 3':AA011635]	13.3
SID W 290871, Integrin alpha-3 subunit [5':N99380, 3':N71998]	13
COL4A1 Collagen, type IV, alpha 1 Chr.13 [145292, (EW), 5':R78225, 3':R78226]	12.7
COL4A1 Collagen, type IV, alpha 1 Chr.13 [489467, (IEW), 5':AA054624, 3':AA054564]	12.7

3.3.2 Experimental Results on the Reduced Dataset with Prototypes

In the Bayesian network learned from the reduced dataset with 40 gene prototypes and 5 drug prototypes, the negative correlation between *ASNS* (Asparagine synthetase Chr.7 [510206, (IW), 5':AA053213, 3':AA053461]) and L-asparaginase, as well as the negative correlation between *DPYD* (SID W 278125, Dihydropyrimidine dehydrogenase [5':N94809, 3':N63511]) and 5FU (fluorouracil) are examined [Scherf *et al.,* 2000]. Figure 3 shows two parts of the Bayesian network. In Figure 3(a), *G*4 is the gene prototype which includes *ASNS* and *D*2 is the drug prototype which includes L-asparaginase. *G*4 and *D*2 are dependent on each other directly. This suggests that these two nodes are strongly correlated with each other. In Figure 3(b), *G*8 is the gene prototype that includes *DPYD* and *D*5 is the drug prototype that includes 5FU. *G*8 and *D*5 do not directly depend on each other.

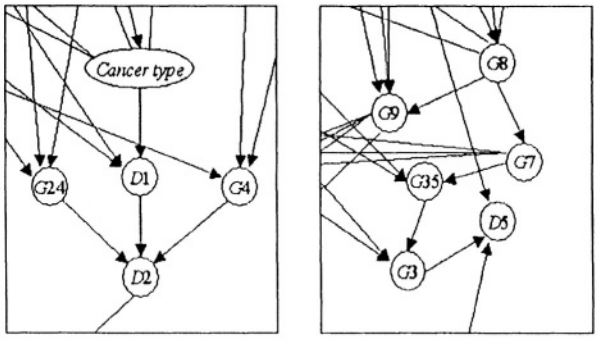


Figure 3. Two parts of the Bayesian network with 46 nodes. $G1 \sim G40$ are gene prototypes. $D1 \sim D5$ correspond to drug prototypes. In (a), $D2$ directly depends on $G4$ and vice versa. $D5$ is not directly dependent on $G8$ in (b).

Table 3 presents the results of the probabilistic inference from the Bayesian network. The inferred conditional probabilities do not show the expected negative correlation between $D2$ and $G4$ clearly. For example, $P(D2 = \text{low} \mid G4 = \text{high})$ should be greater than $P(D2 = \text{high} \mid G4 = \text{high})$. As a consequence, the Bayesian network with 46 nodes has failed to reveal some biologically known facts clearly. It might be due to the information loss induced from discretization, the use of prototypes, or both of these.

Table 3. The conditional probability table for $P(D2 \mid G4)$ inferred from the Bayesian network in Figure 3. The negative correlation is not apparent here.

	D2 = low	D2 = normal	D2 = high
G4 = low	0.32096	0.27086	0.40818
G4 = normal	0.31387	0.41247	0.27366
G4 = high	0.32167	0.34920	0.32913

3.3.3 Experimental Results on the Reduced Dataset with Selected Attributes

To investigate the probabilistic relationships around L-asparaginase, 12 genes and 4 drugs were selected through clustering. Figure 4 shows the part of the Bayesian network with 17 nodes. In this figure, the direct probabilistic dependency is observed between the cancer type and L-asparaginase. L-asparaginase and *ASNS* are also dependent on each other directly. In addition, *ASNS* directly depends on *P5CR* (SID W 484773, PYRROLINE-5-CARBOXYLATE REDUCTASE [5':AA037688, 3':AA037689]). Tables 4

and 5 show the results of some probabilistic inferences from the Bayesian network. The conditional probabilities in Table 4 coincide with the negative correlation between *ASNS* and L-asparaginase. Moreover, when the cancer type is known to be leukemia, the negative correlation is stronger.

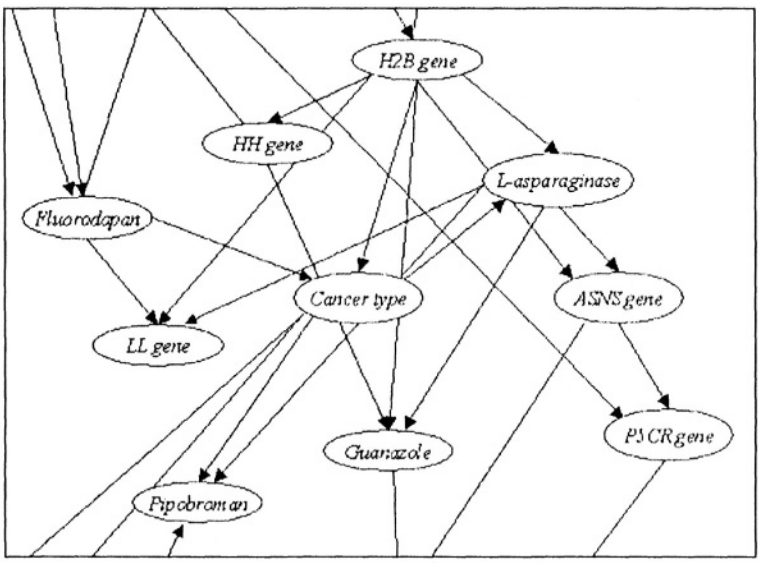


Figure 4. The Bayesian network with 17 nodes. Gene nodes are represented by acronyms. Following is the list of full names of the acronyms P5CR, ASNS, H2B, HH, and LL: SID W 484773, PYRROLINE-5-CARBOXYLATE REDUCTASE [5':AA037688, 3':AA037689] (P5CR), ASNS Asparagine synthetase Chr.7 [510206, (IW), 5':AA053213, 3':AA053461] (ASNS), SID 470936, Homo sapiens mRNA for histone H2B, clone pjG4-5-14 [5':AA034106, 3':AA032092] (H2B), SID W 376009, HISTONE H1D [5':AA040305, 3':AA040326] (HH), SID W 430196, LACTOYLGLUTATHIONE LYASE [5':AA010331, 3':AA010332] (LL).

In addition, *P5CR* and L-asparaginase are highly negative-correlated in Table 5. *P5CR* is involved in the alanine and aspartate metabolism. *ASNS* is involved in the arginine and proline metabolism. These two metabolisms are closely located in the metabolic and regulatory pathway in the Kyoto Encyclopaedia of Genes and Genomes (KEGG) located on the web at (http://www.genome.ad.jp/kegg). And the similarity of *P5CR* and *ASNS* in relation to the negative correlation with L-asparaginase seem to indicate a meaningful relationship.

Table 4. The conditional probability table for *P(L-asparaginase | ASNS)* and *P(L-asparaginase | ASNS, Cancer type = Leukemia)* (the values in the parentheses). The quantified probabilistic dependency between the expression level of *ASNS* and the activity of L-asparaginase coincides with the known biological fact (the negative correlation).

	L-asparaginase = low	**L-asparaginase = normal**	**L-asparaginase = high**
ASNS = low	0.19857 (0.17536)	0.27471 (0.22838)	0.52672 (0.59626)
ASNS = normal	0.31110 (0.27128)	0.49795 (0.53790)	0.19095 (0.19081)
ASNS = high	0.42159 (0.38500)	0.36279 (0.42437)	0.21561 (0.19063)

Table 5. The conditional probability table for *P(L-asparaginase | P5CR)*. The quantified probabilistic dependency between the expression level of *P5CR* and the activity of L-asparaginase is similar to that between *ASNS* and L-asparaginase.

	L-asparaginase = low	**L-asparaginase = normal**	**L-asparaginase = high**
P5CR = low	0.27510	0.35226	0.37263
P5CR = normal	0.31621	0.41072	0.27307
P5CR = high	0.33837	0.39664	0.26499

4. CONCLUSION AND FUTURE WORK

In this paper, the NCI60 dataset was analyzed for the molecular pharmacology of cancer. First, the 60 cell lines were clustered using the STVQ algorithm. While the hierarchical clustering algorithm used in [Scherf *et al.,* 2000] operates in an agglomerative way and provides the tree-like cluster structure, the STVQ algorithm, starting from a coarse global structure, successively refines the cluster structure with some annealing schedule. And it finally represents the cluster structure in a two- or three-dimensional lattice.

We have performed cluster analyses based on the gene expression pattern and the drug activity pattern, respectively. The differences of the cluster structures were shown quantitatively in terms of the averaged Pearson correlation coefficient and the clustering entropy. The drug activity pattern less reflects the tissue of origin than the gene expression pattern, and it is suggested that this might be partly due to the expression of particular genes related to some drug activities. From these results, the drug activity pattern is analyzed with gene expression patterns and cancer types for more detailed information, and Bayesian network learning was applied for this purpose.

In the experiments, a fast search heuristic was applied to learning the Bayesian network with hundreds of nodes. Among hundreds of attributes,

only a few of them, including the cancer type and some genes, show notable relations to others. In order to perform the probabilistic inference, we reduced the dimensionality of attributes by clustering. By using prototypes, the known biological facts could not be discovered clearly. This might be due to the loss of useful information in the original data by the use of gene prototypes and drug prototypes. Hence, the dimensionality reduction by attribute selection was performed. Focusing on the discovery of relationships around L-asparaginase, we selected 12 genes and 4 drugs by clustering. The results of the analysis coincide with the known biological facts: the negative correlation between L-asparaginase and *ASNS,* as well as the influence of the kind of cancer on this negative correlation. In addition, the positive correlation between *ASNS* and *P5CR* was discovered. Biologically, *ASNS* and *P5CR* are located closely in the metabolic pathway. To summarize, the relationships among genes, drugs, and cancer types could be modelled by Bayesian network learning. This suggests that Bayesian network learning and clustering are appropriate for the exploratory analysis of high-throughput genomic data.

Directions for further research are as follows: In a complex domain such as DNA microarray analysis, the learned results are prone to be unreliable because of the small sample size compared with the number of attributes. The eMCMC (evolutionary Markov chain Monte Carlo) method [Zhang *et al.,* 2001] might be an appropriate solution. The more efficient and robust learning and inference algorithms for large Bayesian networks should also be studied. In addition, combining knowledge from biomedical literature with data analysis is a candidate for the improvement of the quality of results.

ACKNOWLEDGEMENTS

The authors would like to thank Sirk June Augh for thorough discussions on the experimental results. The authors also thank anonymous reviewers and editors for helpful comments that greatly improve the paper. This work was supported in part by BK21-IT, IMT-2000, BrainTech, and AITrc programs.

REFERENCES

Dempster, AP, Laird, NM, Rubin, DB. Maximum likelihood from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society B* 39 (1977): 1-38.

Eisen, MB, Spellman, PT, Brown, PO, Botstein, D. Cluster analysis and display of genome-wide expression patterns. *Proceedings of the National Academy of Sciences of the United States of America* 95(25) (1998): 14863-14868.

Friedman, N, Nachman, I, Pe'er, D. Learning Bayesian network structure from massive datasets: the "sparse candidate" algorithm. In *Proceedings of the 15th Conference on Uncertainty in Artificial Intelligence (UAI 1999)* (1999): 206-215.

Friedman, N, Linial, M, Nachman, I, Pe'er, D. Using Bayesian networks to analyze expression data. In *Proceedings of the 4th Annual International Conference on Computational Molecular Biology* (*RECOMB 2000*) (2000): 127-135.

Graepel, T, Burger, M, Obermayer, K. Self-organizing maps: Generalizations and new optimization techniques, *Neurocomputing* 21 (1998): 173-190.

Hartemink, AJ, Gifford, DK, Jaakkola, TS, Young, RA. Using graphical models and genomic expression data to statistically validate models of genetic regulatory networks. *Pacific Symposium on Biocomputing* 6 (2001): 422-433.

Heckerman, D. *A tutorial on learning with Bayesian networks.* Edited by MI Jordan. Learning in Graphical Models. MIT Press, 1999.

Heckerman, D, Geiger, D, Chickering, DM. Learning Bayesian networks: the combination of knowledge and statistical data. *Machine Learning* 20(3) (1995): 197-243.

Hwang, K-B, Cho, D-Y, Park, S-W, Kim, S-D, Zhang, B-T. Applying machine learning techniques to analysis of gene expression data: cancer diagnosis. In: Lin, SM, Johnson, KF, *Methods of Microarray Data Analysis,* Norwell, MA, Kluwer Academic Publishers, (2001):167-182.

Khan, J *et al.* Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. *Nature Medicine* 7(6) (2001): 673-679.

Pearl, J. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference.* Edited by M Kaufmann. (1988).

Raychaudhuri, S, Stuart, JM, Altman, RB. Principal components analysis to summarize microarray experiments: application to sporulation time series. *Pacific Symposium on Biocomputing* 5 (2000): 452-463.

Scherf, U *et al.* A gene expression database for the molecular pharmacology of cancer. *Nature Genetics* 24 (2000): 236-244.

Zhang, B-T, Cho, D-Y. System identification using evolutionary Markov chain Monte Carlo. *Journal of Systems Architecture* 47(7) (2001): 587-599.

12

EVALUATION OF CURRENT METHODS OF TESTING DIFFERENTIAL GENE EXPRESSION AND BEYOND

Yi-Ju Li[1], Ling Zhang[1,2], Marcy C. Speer[1], and Eden R. Martin[1]
[1] *Center for Human Genetics, Duke University Medical Center, Durham, NC.,*
[2]*Bioinformatics Group, Statistics Department, North Carolina State University, Raleigh, NC*

Abstract: One frequent question in the study of microarrys concerns the number of replicates required to obtain vaild data. We used the T-matrix data from the NCI-60 cancer cell lines dataset to investigate this question. Five testing methods were evaluated. We selected two cancer groups for comparisons, ovarian (OV) vs. breast (BR) and leukemias (LE) vs. renal carcinoma (RE), to perform hypothesis testing for detecting the genes expressed differentially between cancer groups. Our goal is to examine the pattern and performance of each testing method and the required sample size. The first four testing methods are t-test based methods with different strategies of computing sampling variance, including the uses of sampling variance, pooled variance, and common variance. The 5th test is a permutation test based on the t-test with pooled variance. Our results show that there are more genes with statistically significant differences in expression in the LE vs. RE comparison than between the OV vs. BR. The permutation works similarly to the t-test itself. Overall, the pooled variance approach proved a better strategy. For sample size, as expected, the number of significant genes increased as the number of cell lines increased for the same testing method. However, we found that the results derived from 3 cell lines are very different from the other results. It may imply that more than three cell lines or replicates are needed in the microarray study in order to attain enough power to detect the differential gene expression

Key words: microarry, replicates, t-test, permutation test, sample size, power

1. INTRODUCTION

An exciting biological achievement in the last few years is the utilization of microarray technologies to measure simultaneously the expression levels of thousands of genes. The image data from the arrays lead to gene-specific intensities representing relative expression levels. A comparison of gene expression of cells or tissues from two experimental conditions, which may refer to samples drawn from two types of tissues, tumors or cell lines, may provide useful information on important biological processes or functions [Botstein and Brown, 1999; Lander, 1999]. It has been found that due to high noise-signal ratios, a single microarray may not provide enough information to be reliable for analysis [Lee *et al.*, 2000]. Replication of array experiments is often recommended. Replication also makes it possible to assess variability of expression within genes between replicates. One of the important statistical issues in microarray study is the hypothesis testing for detecting differentially expressed genes between two experimental conditions. Due to the high cost of each replicate, it is impractical to produce many replicates of arrays. On the other hand, a small number of replicates make it difficult to perform statistical tests. The balance between maintaining reasonable cost and maximizing statistical robustness requires more thorough investigation.

With this question in mind, we made use of the gene expression data from the T-matrix of NCI-60 cancer cell lines dataset [Scherf *et al.*, 2000] and evaluated five simple t-test based methods for evaluating differential gene expression. We also examined the effect of the number of cell lines for each testing method. Although the data here are not exactly in the same format and property as the replicates of microarrays, our goal is two fold: (1) to find the pattern and evaluate the performance of each testing method and, (2) to infer the number of replicates needed in the microarray if the testing method is applied.

2. MATERIALS AND METHODS

The testing methods that we evaluated in this study are t-test based methods, differing from one another because of different estimating procedures for the variance. The details of five testing methods are described below.

Let Y_{ijA} and Y_{ijB} be the intensity measurement for gene i ($i=1, \ldots, n$) at the jth cancer cell line for group A and B cancer (e.g. ovarian and breast cancer), respectively, where $j_A=1,\ldots, r_A$ and $j_B = 1,\ldots, r_B$ .We use type A as an example to describe the sampling variance estimation of each gene; that is,

$$S_{iA}^2 = \frac{\sum_{j=1}^{rA} (Y_{ijA} - \bar{Y}_i)^2}{r_A - 1}$$

The pooled variance for each gene is computed by:

$$S_{iP}^2 = \frac{(r_A - 1)S_{iA}^2 + (r_B - 1)S_{iB}^2}{(r_A + r_B - 2)}$$

In addition, we also investigated the t-tests based on the common variance, which is defined as the average of sampling variance over all genes [Nadon *et al.*, 2001]. The use of the common variance strategy in the t-test was suggested as more powerful than the conventional way of estimating sampling variance when the number of replicates is small. The basic idea is to combine all sampling variances across all genes to form a common variance; that is, variance will be constant through all genes. The common variance can be obtained as below.

$$S_A^2 = \frac{\sum_{i=1}^{n} S_{iA}^2}{n}$$

We evaluated the following five tests:

Test1: use sampling variance for each gene in each cancer group.

$$T_i = \frac{\bar{Y}_{iA} - \bar{Y}_{iB}}{\sqrt{\frac{S_{iA}^2}{r_A} + \frac{S_{iB}^2}{r_B}}}$$

Test2: use pooled sampling variance for each gene.

$$T_i = \frac{\bar{Y}_{iA} - \bar{Y}_{iB}}{\sqrt{S_{ip}^2 (\frac{1}{r_A} + \frac{1}{r_B})}}$$

Test3: use the common variance for each cancer group.

$$T_i = \frac{\bar{Y}_{iA} - \bar{Y}_{iB}}{\sqrt{\frac{S_A^2}{r_A} + \frac{S_B^2}{r_B}}}$$

Test4: use the pooled common variance.

$$S_P^2 = \frac{(r_A - 1)S_A^2 + (r_B - 1)S_B^2}{(r_A + r_B - 2)}$$

$$T_i = \frac{\bar{Y}_{iA} - \bar{Y}_{iB}}{\sqrt{S_p^2(\frac{1}{r_A} + \frac{1}{r_B})}}$$

Test5: a permutation test within each gene based on Test2. We first computed Test2 for each gene in the observed dataset. We mixed all expression data from all cell lines in each gene together and randomly withdrew r_A and r_B cell lines into each disease without replacement. Then, performed Test2. We repeated the same procedure 1000 times and computed the number of times (N) that the test from permuted samples is more extreme than the one from the observed sample. The p-value is, therefore, computed by *N/1000.* If the p-value is less than 0.05, we reject the null hypothesis of equal expression levels between group A and B.

All tests were compared to the standard t-distribution with a threshold of 5% significance level.

Two sets of comparisons were performed in this study: ovarian (OV) vs. breast (BR) and leukemias (LE) vs. renal carcinoma (RE). One reason to choose these two pairs for comparison is that more cell lines are available for these four disease categories in the data set. In the original data set, we had 6 cell lines for OV, 8 for BR, 6 for LE, and 8 for RE. We used this original data set to perform all five tests. For examining the sample size, we used computer generated random numbers to choose 3, 4, and 5 cell lines from each disease and then performed the first four testing methods, that is, Testl to Test4.

3. RESULTS

We used C++ and Splus for data management and computations. Table 1 summarizes the number of genes demonstrating significant differential expression in the OV vs. BR and LE vs. RE comparisons for each of the 5 testing procedures. The comparison between OV vs. BR produced fewer differentially expressed genes than the comparison of LE vs. RE. The permutation test obtained 99 genes in the OV vs. BR group and 560 genes in the LE vs RE group, which is almost the same as Test2. Since the permutation test is based on Test2, our results indicate that the permutation test did not improve the outcome of Test2 in this set of data. Overall, the strategy of using the common variance (Test3 and Test4) identified fewer genes than tests using sampling variance (Test1 and Test2). From the view of the pooled variance approach, we found that pooled variance strategies (Test2 and Test4) detected fewer genes expressed differentially than the tests not using pooled variance (Test1 and Test3) in the OV vs. BR group, but vice versa in the LE vs. RE group.

Table 1: The number of genes expressed significantly different obtained from Test1-Test5 for OV vs. BR and LE vs. RE.

	OV vs. BR	LE vs. RE
Test1	105	551
Test2	96	561
Test3	82	526
Test4	75	541
Test5: Permutation test based on Test2	99	560

To compare the performance between each testing method, we examined whether the same gene was detected by any pair of testing methods. Table 2 summarizes the number of genes showing significantly different expression in each pair of testing methods. The upper triangle is for OV vs. BR and the lower triangle is for LE vs. RE. For instance, we found that 90 genes had different expression levels between OV and BR by Testl and Test2, and 537 genes had different expression levels between LE and RE by Testl and Test2.

We also summarized the number of genes that were found to show significantly different expression levels in one test, but not in the other test. The results of these pairwise comparisons are summarized in Table 3. Each row presents the testing method with significant results and each column is for the testing method with non-significant results. For instance, we found 2

genes with significant results from Test5, but nonsignificant in Testl for the comparison between OV vs. BR (Table 3a)

Table 2: The number of common genes detected by each pair of testing methods.

Comparison		OV vs. BR				
		Test1	Test2	Test3	Test4	Test5
LE vs. RE	Test1		90	61	57	90
	Test2	537		60	56	91
	Test3	477	487		75	59
	Test4	485	495	526		55
	Test5	555	553	485	493	

Note: The upper triangle is for the OV vs. BR group and the lower triangle is for the LE vs. RE group.

Table 3: Pairwise comparison between testing methods for (a) OV vs. BR (b) LE vs RE.

(3a)

	Test1	Test2	Test3	Test4	Test5
Test1*		15	44	48	51
Test2*	6		36	40	42
Test3*	21	22		7	23
Test4*	18	19	0		37
Test5*	2	2	16	18	

(3b)

	Test1	Test2	Test3	Test4	Test5
Test1*		14	74	66	52
Test2*	24		74	66	58
Test3*	49	39		0	71
Test4*	56	46	15		79
Test5*	5	1	49	42	

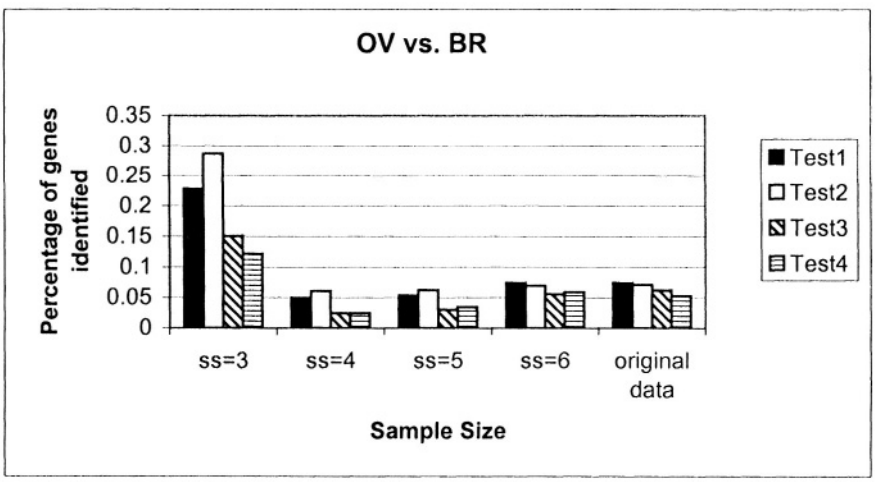
Note: Each row (Test1*-Test5*) is for significant testing results and each column (Test1-Test5) is for nonsignificant testing results.

The examination of sample size is summarized in Figure 1, which shows the percentage of genes showing a significantly different expression level

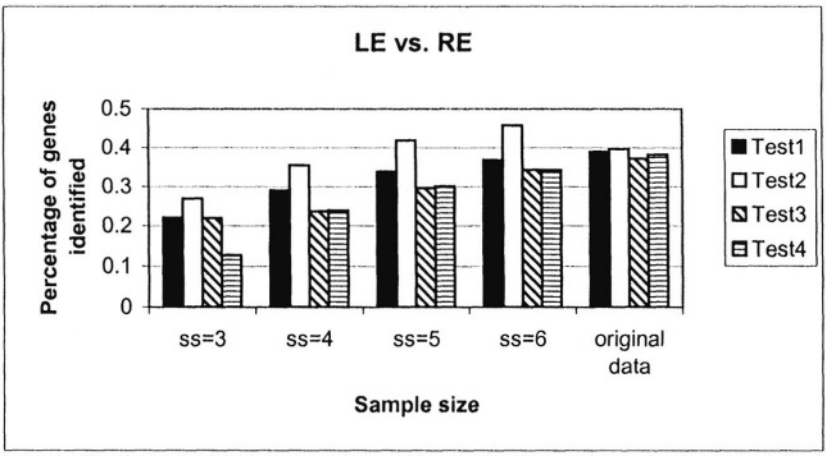
between diseases based on each testing method for each sample size that we examined. We used computer generated random numbers to select the sample (cell lines). The sampled cell lines were then applied to each testing method. It is clear that a similar pattern was observed for the cases of sample size 4 or greater, but it is very clear that wrong results were produced under sample size 3. As can be seen, the histogram pattern under sample size 3 is much different from those produced by other sample sizes for both the OV vs. BR group and the LE vs. RE group (Figure 1a, 1b). Furthermore, we found that Test2 detected more genes in the OV vs. BR group (29%) than in the LE vs. RE group (27%) when sample size is 3, which is not what we expected.

Figure 1. The percentage of significant differential expressed genes under different sample sizes (ss=3, 4, 5, and 6 cell lines from each disease) by Test1—Test4. (1a) Comparison of OV and BR (1b) Comparison of RE and LE.

(1a)



(1b)



4. DISCUSSION

Microarray analysis has become a popular laboratory tool for studying changes in expression across a large number of genes. Image data from the arrays lead to gene-specific intensities representing relative expression levels, which are difficult to interpret without proper statistical testing. Replication of microarray experiments is needed to assure the quality of each spot measurement. The importance of replication has been nicely illustrated in recent work by [Lee *et al.*, 2000]. Here, we investigated one of the interesting statistical issues in microarray replication, the statistical testing methods and the number of replicates, by using the T-matrix of NCI-60 cancer cell lines dataset. We evaluated five t-test based testing methods. In the future, it is possible to extend our scope to other testing methods such as methods based on a mixed model [Wolfinger *et al.*, 2001]. Although a t-test is a relatively simple statistical testing method, we have often seen it implemented in the computer packages specialized for microarray data analysis, for instance, Partek software (http://www.partek.com). It will be useful if we have a more thorough understanding of the performance of t-test based methods when they are applied to microarray data. It should be noted that the dataset used here does not represent exact microarray replicates, so the results should be interpreted with caution. There is no standard to quantify which testing method is close to the true answer; that is, the true number of genes with different expression level in two diseases that we compared. We can only compare the similarity and differentiation between

testing methods, not the robustness of the testing methods. The numbers presented here (Table 1, 2, and 3) are not meaningful if interpreted alone. For microarray data analysis, in practice, we test multiple genes at the same time. Therefore, it is necessary to correct for multiple testing using an adjusted p-value (e.g. Bonferroni correction) to declare significant results. In this study, our purpose is to compare testing methods with the same set of data. Therefore, we did not apply the Bonferroni correction because it will not affect our conclusion.

As the results demonstrate, ovarian and breast cancer show fewer gene expression differences than leukemias and renal carcinoma. These results are consistent with the biological similarities between breast and ovarian cancer: for instance, both are of epithelial origin and both have similar oncogenic changes (e.g., p53 over-expression). This finding corresponds to the results of cluster analysis [Scherf *et al*., 2000]. Although we do not know which method detects the closest number of genes to the true answer, in general, we found that common variance strategy (Test3 and Test4) will detect fewer genes than the sampling variance strategy (Testl and Test2). We also saw a consistent pattern of the testing methods with pooled variance (Test2 and Test4) in both groups of comparison; that is, both Test2 and Test4 showed either an increasing (in LE vs. RE) or a decreasing (in OV vs. BR) number of genes when we compare them to Testl and Test3. We interpret this finding as an indication that a pooled variance strategy can adjust the results toward the true answer better than the one without a pooled variance strategy. Furthermore, we saw no improvement by using a permutation test as Test5 is almost exactly the same as Test2. The pairwise comparison (Table 2 and 3) nicely corresponds to the theoretical derivation. For instance, Test3 and Test4 are similar tests based on the common variance, so most of the genes identified in Test3 and Test4 are exactly the same (only 7 genes were identified by Test3, but not by Test4). This finding is also consistent with the outcome of Table 2, for instance, more genes were detected at the same time when Testl was compared to Test2 than when Testl was compared to other testing methods (Test3 and Test4). In summary, we recommend Test2 and Test4 for future analyses.

The sample sizes do affect the outcome of a test. Our results clearly show that a sample size of 3 cell lines shows a completely different pattern of results (Figure 1). For instance, Testl and Test2 detected more genes in the OV vs. BR group than in the LE vs. RE group, which we know is incorrect based on the biological knowledge. It implies that the variance estimates by using 3 cell lines (or replicates) may not be robust. We strongly suggest that more than 3 replicates are necessary in a microarray study.

ACKNOWLEDGEMENTS

The authors gratefully acknowledge support from NIH grants NS26630, ES11375, Parkinson Udall P50 NS39764-03, and Parkinson Udall supplement P50 NS39764-02S2.

REFERENCES

Botstein, D, Brown, P. Exploring the new world of the genome with DNA microarrays. *Nature Genetics (Suppl)* 21 (1999): 33-37.

Lander, ES. Array of hope. *Nature Genetics (Suppl)* 21 (1999): 3-4.

Lee, M-L T, Kuo, FC, Whitmore ,GA, Sklar J. Importance of replication in microarray gene expression studies: statistical methods and evidence from repetitive cDNA hybridizations. *Proc Nat Acad Sci* 97 (2000): 9834-9839.

Nadon, R, Shi, P, Skandalis, A, Woody, E, Hubschle, H, Susko, E, Rghei, N, Ramm, P. Statistical inference methods for gene expression arrays. *http://www.imagingresearch.com* (2001).

Scherf, U, Ross, DT, Waltham, M, Smith, LH, Lee, JK, Tanabe, L, Kohn, KW, Reinhoid, WC, Myers, TG, Andrews, DT, Scudiero, DA, Eisen, MB, Sausville, EA, Pommier, Y, Botstein, D, Brown, PO, and Weinstein, JN. A gene expression database for the molecular pharmacology of cancer. *Nature Genetics* 24 (2000): 236-244.

Wolfinger, RD, Gibson, G, Wolfinger, ED, Bennett, L, Hamadeh, H, Bushel, P, Afshari, C, Paules, RS. Assessing gene significance from cDNA microarray expression data visa mixed models. *J Comput Biol* 8(6) (2001): 625-637.

13

EXTRACTING KNOWLEDGE FROM GENOMIC EXPERIMENTS BY INCORPORATING THE BIOMEDICAL LITERATURE

James P. Sluka
InPharmix Inc.

Abstract: We present a technique to extract relevant information from the literature to aid in the analysis of a typical genomics data set. Analysis was conducted using PDQ_MED, a program based on the assumption that if two genes are found to be related under an experimental paradigm, such as a gene chip experiment, then any literature which relates the two genes is of interest. PDQ_MED searches MEDLINE for abstracts that contain two or more of the terms in the user's query set. For this paper, we have used PDQ_MED to analyze 160 genes up-regulated in acute myeloid leukemia (AML) from the NCI-60 dataset. PDQ_MED executed 12,880 queries to MEDLINE and identified nearly 300,000 abstracts that refer to at least one of the 160 terms. PDQ_MED identified and analyzed a set of 81 terms that can be grouped together via the literature. In addition, there is literature directly linking 52 of the terms with AML. Overall, the literature analysis identified 1028 sentences that directly relate two or more of the query genes.

Key words: gene expression analysis, literature, DNA microarray, PDQ_MED, text mining

1. OBJECTIVE

As the use of genomic tools increases, there is a growing need for tools to effectively exploit the resulting data. Lists of genes that are related under an experimental paradigm are a common result of genomics techniques such as subtracted libraries, differential display, 2D protein gels and gene chip (DNA microarrays) or protein array experiments. Currently, there are only a few tools for extracting useful information from the scientific literature in

conjunction with these large data sets. Two such tools are MedMiner[1] [Tanabe *et al.,* 1999] and PubGene[2] [Jenssen *et al.,* 2001].

2. ANALYTICAL METHODS

PDQ_MED (Pair-wise Data Query to MEDLINE) exhaustively searches MEDLINE for abstracts that contain two or more of the terms in the user's data set. This pair-wise approach allows the researcher to effectively mine the nearly 11 million abstracts in MEDLINE for information relevant to their genomics projects.

PDQ_MED is based on the assumption that if two genes are found to be related under some experimental paradigm, such as in a gene chip experiment, then any literature which relates the two genes is of interest. A "co-occurrence" is defined as any abstract that contains two or more of the query terms. The simplest embodiment of this idea is to search MEDLINE (or other databases) with all possible pairwise combinations of the query terms. For N terms, $\sim N^2/2$ searches are required. For small values of N, this can be done manually. For larger values, the number of searches quickly becomes impractical.

2.1 Data Sets

We have chosen to analyze a subset of the NCI-60 cancer gene expression database [Scherf *et al.,* 2000]. The initial set consisted of the expression data for the full set of 9,703 genes for the three leukemia cell lines, CCRF-CEM, MOLT-4 and K-562, in the NCI database. CCRF-CEM and MOLT-4 are from acute lymphoblastic leukemias (ALL) whereas K-562 represents acute myelogenous leukemia (AML). The K-562/AML data was divided by the average for the two ALL lines in order to reduce the influence of genes characteristic of leukocytic cell lines. The resulting expression data is similar to the Golub data set [Golub *et* al., 1999] used for CAMDA-2000. The resulting modified expression values were then sorted and the 250 most highly expressed genes used as the gene list. For these 250 genes we then removed unnamed genes including ESTs, KIAAs and genes annotated as "similar to" another gene, resulting in a final list of 160 named genes. In addition, we included a term for the disease (AML).

As our literature database (knowledge domain), we used MEDLINE accessed through Entrez via the web (http://www.ncbi.nlm.nih.gov/entrez/).

[1] http://discover.nci.nih.gov/textmining/filters.html
[2] http://www.pubgene.org/

MEDLINE currently contains more than 11 million abstracts and, in terms of the total number of characters, is approximately the same size as the GENBANK nucleotide database.

2.2 Software

PDQ_MED is a web-based Perl program that searches MEDLINE for abstracts that contain two or more of the terms in the user's data set.

2.2.1 Input

The first step in the analysis is to assign names to each gene that are suitable for searching in MEDLINE. In this case, the original names are those that appear in the NCI-60 database. Since these names tend to be brief, cryptic or outdated, some work was needed to verify or correct the names. To assign the best possible name to each gene we used keyword and/or BLAST searches across a combination of publicly available databases. These included GENBANK, OMIM, GDB and GeneCards. Typical original and corrected names are shown in Table 1.

Table 1. Typical corrected names for the NCI-60 dataset as used in this study.

NCI-60 "Name"	Corrected Name(s)
SID W 293514, Human 54 kDa progesterone receptor-associated immunophilin FKBP54 mRNA, partial cds [5':N98804, 3':N63715]	FKBP54 "54 kDa progesterone receptor-associated immunophilin"
SID W 361787, Human guanine nucleotide-binding regulatory protein (Go-alpha) gene [5':W96534, 3':W96428]	GNAO1 "guanine nucleotide-binding regulatory protein"
Hemoglobin, alpha 1 Chr. [469647, (E), 5':AA027875, 3':AA027832]	HBA1 "Hemoglobin, alpha 1"
SID 81641, H.sapiens mRNA for Nup88 protein [5':T64514, 3':T65939]	Nup88 "nucleoporin 88kD"
SID W 509700, Ornithine aminotransferase (gyrate atrophy) [5':AA058461, 3':AA058361]	OAT "Ornithine aminotransferase" "ORNITHINE OXO-ACID AMINOTRANSFERASE"
PRKCB1 Protein kinase C, beta 1 Chr.16[284459, (IEW), 5':N75108, 3':N52338]	PKCB PRKCB PRKCB2 "Protein kinase C, beta 1" PKC-b1
PNMT Phenylethanolamine N-methyltransferase Chr.17[289857, (R), 5':, 3':N63192]	PNMT "Phenylethanolamine N-methyltransferase" PENT

The basic input to PDQ_MED is a list of query terms encompassing the genes, proteins, diseases or other concepts under investigation (see Figure 1). An individual query term can consist of more than one version of a particular name. For example, a query can consist of a full name and an abbreviated name; “Interleukin-1b IL-1b”, or alternative names; “proteasome iota

macropain iota". PDQ_MED automatically inserts ORs between the individual terms, or quoted phrases, contained on a single line of the input representing a single gene, gene product, disease or other concept. In addition, the user may explicitly join terms by any of the Boolean operators or use any of the field or date operators supported by MEDLINE.

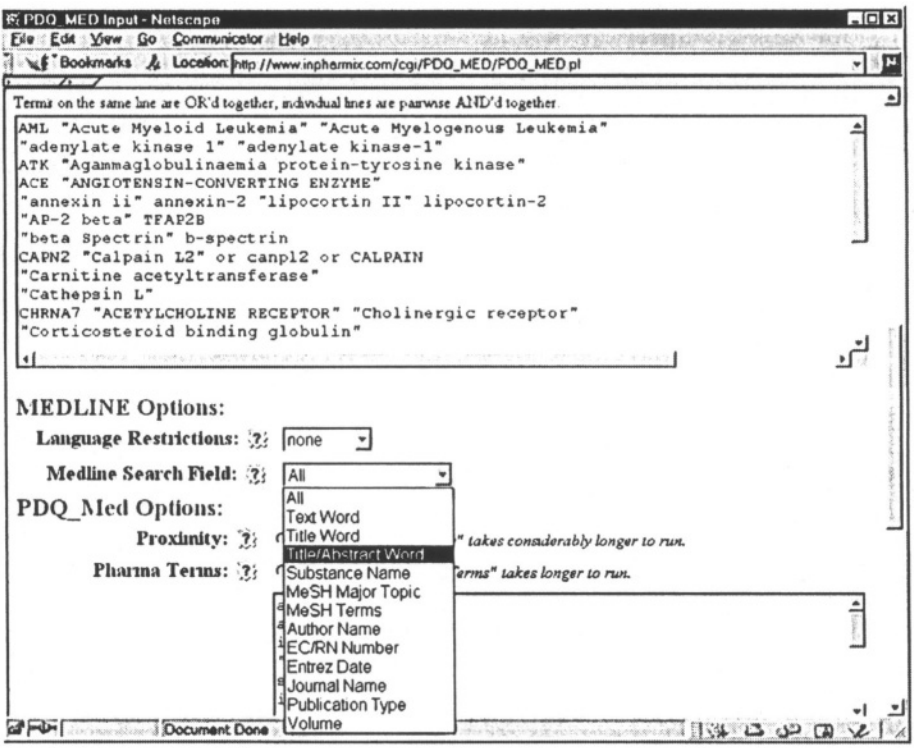


Figure 1. Part of the PDQ_MED input page.

2.2.2 Search

Searches are carried out by constructing individual Entrez URLs for all possible pairwise combinations of the query terms joined by AND. The URLs are then submitted via the internet and the search results captured and analyzed by PDQ_MED.

2.2.3 Local Acronyms and Proximity Searching

A refinement to the basic search strategy is to require a higher degree of dependence, i.e., closer proximity within the document, between two query terms. In "Proximity" searching, PDQ_MED examines all abstracts

containing two terms and determines if the terms co-occur in the same sentence. Sentence level proximity searching is not directly supported by MEDLINE.

One challenge to effectively use proximity searching in the scientific literature is the highly variable nature of the names of genes, proteins and small molecules. As mentioned above, PDQ_MED allows the user to enter multiple names for the same entity. However, acronyms that either are common words, or used for more than one concept, are problematic. For example, a common acronym of "Acute Lymphoblastic Leukemia" is ALL. Since ALL is a common English word MEDLINE will not search for abstracts containing it. In addition, it is common for more than one gene, protein or concept to use the same acronym. These problems with acronyms make proximity searching in the biomedical literature difficult. Consider, for example, the abstract:

> *"In acute lymphoblastic leukemia (**ALL**), the cell surface ... (followed by several sentences). GPRE also decreased the fraction of CD11-bearing **ALL** M2 and M5 cells."*

In this case, the use of a "local acronym" (ALL) destroys proximity between the terms "acute lymphoblastic leukemia" and CD11. To circumvent this problem, PDQ_MED identifies local acronyms on a per abstract basis. Briefly, a local acronym is defined as a short parenthetical character string following a query term as in the ALL example above. A local acronym is only used for the abstract in which it was found. These local acronyms allow PDQ_MED to identify the CD11 plus ALL (a local acronym for "acute lymphoblastic leukemia") sentence shown above as a proximity sentence.

2.2.4 Analysis

After PDQ_MED has identified all of the abstracts containing two or more of the query phrases, it uses a greedy clustering algorithm to organize the terms into groups. These groups represent sets of terms that co-occur in the literature. For example, if query-A and query-B co-occur in a set of abstracts and query-B and query-C co-occur in a different set of abstracts, then queries-A, B and C are clustered together in the same group (Figure 2). Groups may suggest relationships between terms that are not explicitly present in MEDLINE. In the example in Figure 2, grouping would suggest a possible relationship between query-A and query-D because of their common linkage to query-B, even though query-A and query-D do not explicitly co-occur in any abstracts.

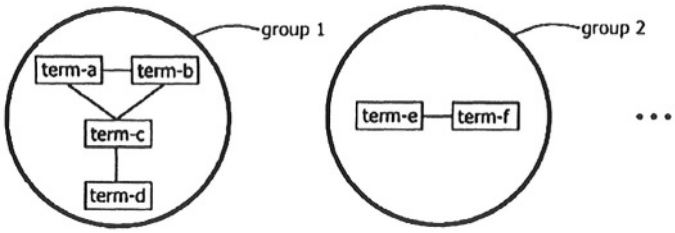


Figure 2. Grouping of query terms.

The user may also search for "Pharma Terms" such as "agonist", "antagonist" or "drug" (Table 2). The "Pharma Term" search results are used to rank and highlight the proximity sentences for each term pair and provide additional practical information about the individual query terms.

Table 2. Default "Pharma Terms" used by PDQ_MED.

antagonis*	down-regulat*
agonis*	regulat*
inhibit inhibit*	X-ray "crystal structure"
bind* bound	therapy therapeutic
stimulat*	drug
interact*	target target*
up-regulat*	efficacy efficacious

3. RESULTS

For a complete search of MEDLINE with the AML dataset including proximity checking, PDQ_MED executed 12,880 queries and identified nearly 300,000 abstracts that refer to at least one of the 160 query phrases (gene names). Total run time for this analysis was three hours. The run time is essentially independent of the computer used since the majority of the time (>90%) is spent waiting for the MEDLINE responses to the queries. The query term that occurred most frequently in MEDLINE was "angiotensin-converting enzyme" (23,588 abstracts). AML occurred in 21,564 abstracts.

For the 161 terms in this data set, PDQ_MED identified a group of 81 terms (which includes AML) that can be linked together (grouped) via the literature. For these 81 terms, there were a total of 1028 sentences

representing 204 term pairs. No co-occurrences were found for the remaining 80 terms.

Figure 3 shows a distance geometry representation of the simplified co-occurrence data for the terms in the 81-member group. In Figure 3, each box represents a query term. Connected boxes represent terms that co-occur in at least one abstract. The length of the interconnection is inversely proportional to the co-occurrence frequency. cFos, AML, VEGF, ACE, IGF1, IL8 and cadherin were the most extensively cross-referenced terms in this set with 27, 25, 21, 20, 19, 18, 18 co-occurring terms respectively. To simplify the graph in Figure 3, only the three strongest links from each node are shown.

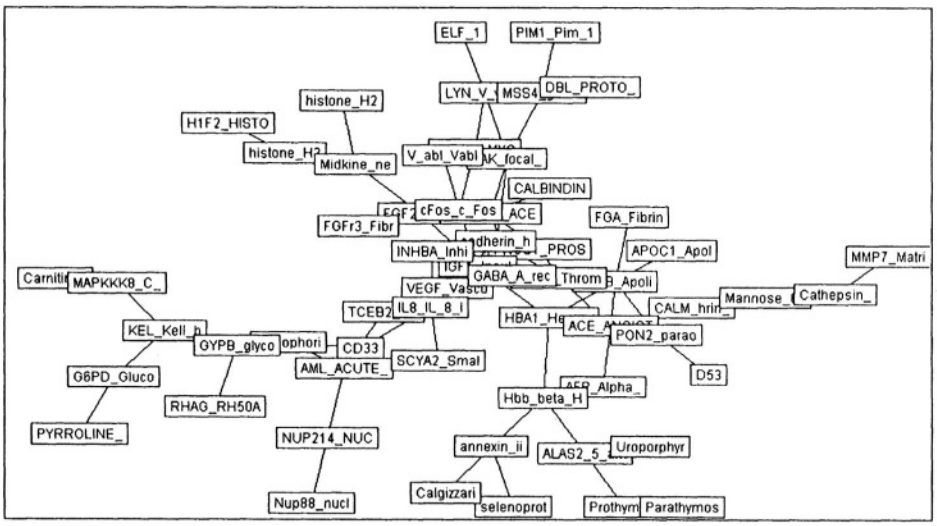


Figure 3. Distance geometry representation of the relationships found in MEDLINE for the terms in the 81-term group. In this display only the three strongest links from each node are shown.

4. DISCUSSION

The PDQ_MED analysis of the 300,000 abstracts covering this set of 161 terms resulted in selecting 1028 sentences, a more than 1000 fold reduction in data. The 1028 sentences are partitioned across 204 term pairs, with an average of five proximity sentences per term pair. Though examination of the 1028 sentences is a formidable task, it is a practical undertaking.

There are several analyses of the results provided by PDQ_MED that may be used, depending upon research needs. In the sections that follow we examine several of these.

4.1 Title Proximity

For highly cross-linked data sets, such as the AML data (Figure 3), it is useful to first examine only the strongest links found in MEDLINE. One way to do this is to use PDQ_MED's ability to search only the titles of papers for co-occurrences of query terms. If two query terms occur together in the title of a paper then there is a very good chance that the paper says something significant about the relationship between the two terms. Figure 4 shows the distance geometry analysis of the terms from the AML dataset which co-occur in the titles of papers. As can be seen, the number of relationships is significantly fewer than in the full abstract search (compare Figures 3 and 4). AML (marked by an arrow) is directly linked to seven other terms (the limit used for the generation of the graph).

It is interesting to note the "constellation" of 8 terms all linked to both IL8 and VEGF (marked by an arrow), consisting of Cadherin-H, Pros1, ACE, cFos, AFP, IGF-1, Inhibin-A and TCEb2, which may suggest a particular pathway or regulatory network is operating. Examination of the proximity sentences for these terms suggests their involvement in angiogenesis, tumour development and various carcinomas.

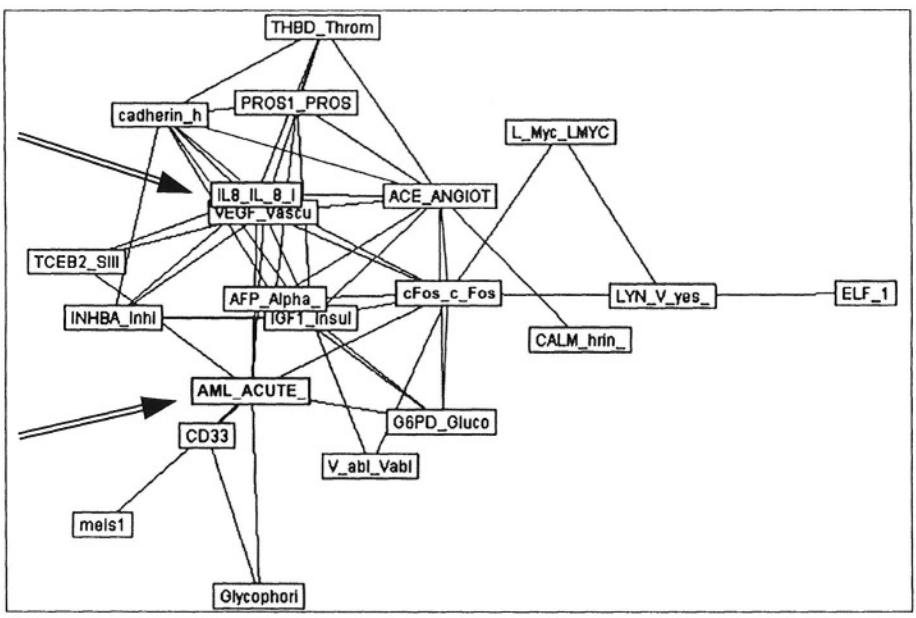


Figure 4. Title Proximity for the AML dataset. In this representation, only the seven strongest links per node are shown.

4.2 Genes Linked to the Disease

A second analysis of the PDQ_MED results is to ask *for which of these genes does the literature provide a precedence for their involvement in AML?* AML co-occurs in abstracts with 52 of the query genes and co-occurs in sentences with 25 of the query genes. A listing of the query terms that co-occurred with AML two or more times (with proximity checking) is shown in Table 3. In Table 3, the number of abstracts containing both terms and the number of proximity sentences are given by the "Abstract" and "Proximity" columns, respectively.

Table 3. Terms (gene or protein names) with >1 co-occurrence with AML in MEDLINE.

Abstract	Proximity	Gene / Protein Name
>250	83	CD33
14	9	Vegf
22	8	Il-8
9	8	Meis1
35	5	Glycophorin A
29	5	G6pd
43	4	cFos
6	3	Calm
6	3	Thbd
5	3	Cadherin
5	3	Mss4
4	2	Asparagine Synthetase
3	2	Lyn
2	2	Inhibin Beta A

The query term that co-occurs most frequently with AML is CD33 and out of a total of 83 proximity sentences, the two top ranked sentences were (query terms in bold face):

1. Blast cells from most patients with **acute myelogenous leukemia** express **CD33**, whereas normal stem cells necessary for maintenance of hematopoiesis do not.
2. Two anti-**CD33** monoclonal antibody conjugates, Y90-HuM195 and CMA-676, have been used in **acute myelogenous leukemia** (**AML**) and have shown some efficacy.

From these two sentences, the user quickly learns something about the relationship between CD33 and AML. In this case, that CD33 is a characteristic marker of AML cells and that it has been used as a therapeutic target for intervention in AML.

Overall, there is literature precedence for some relationship between AML and about one third of the high expression genes from the AML cell line in the NCI-60 database.

4.3 Genes That Cannot Be Linked to the Disease

A third useful analysis of the PDQ_MED results is examination of the list of *genes that cannot be linked to AML.* As mentioned above, 52 of the genes can be linked at the abstract level, an additional 29 genes fall in the same group as AML, leaving 80 genes that could not be linked, directly or indirectly, to AML. For some of these "un-linked" genes there is simply very little literature available. However, others occur frequently in MEDLINE. For example, MAP3K5 occurred in 2657 abstracts but never with AML or any of the 80 terms that grouped with AML. This suggests a research opportunity with several attractive features including;

1. Experimental observation of increased levels of MAP3K5 in AML cells.
2. Significant quantity of literature describing the function of MA3K5 in other systems.
3. The apparent novelty of the idea that MAP3K5 is related in any way to AML.

Table 4 shows a portion of the "Pharma Term" sentence output for MAP3K5 (MEK1) that identifies two small molecule inhibitors, U0126 and PD98059, of this kinase. It may be worthwhile to investigate the affect of these inhibitors on AML cells. Similar results are found for several other of the genes in the AML dataset (data not shown).

Table 4. Selected "Pharma Sentences" for MAP3K5 (MEK1). Query terms are in bold, "Pharma Terms" are bold italics, and the underlined number is the MEDLINE abstract ID.

MAP3K5 OR "mitogen-activated protein kinase kinase kinase 5" OR "MAP/ERK kinase kinase 5" OR ASK1 OR MAPKK1 OR MAPKKK5 OR MEK1 OR MEKK5
11423913 Pretreatment with either the MEK1 inhibitor U0126 or PI3-kinase inhibitor LY294002 sensitized BAE cells to TNF-induced apoptosis.
11431469 Three different inhibitors of MEK1/2 abolished PE-induced activation of S6K2 whereas expression of constitutively active MEK1 activated S6K2, without affecting the p38 mitogen-activated protein kinase and JNK pathways, indicating that MEK/ERK signaling plays a key role in regulation of S6K2 by PE.
11437382 To determine the involvement of MEK1-p42/p44 MAPK pathway in mediating DAB2 gene expression, we have performed the following experiments and found that (i) there was sustained activation of p42/p44 MAPK, but not p38 MAPK, upon K562 cells differentiation; (ii) application of MEK1 inhibitor U0126 reduced the expression of DAB2 protein, mRNA and promoter activity, as well as cell differentiation; (iii) constitutively active MEK1 increased DAB2 promoter activity; and (iv) dominant negative ERK2 abolished constitutively active MEK1-induced DAB2 promoter activity.
11440832 PD98059, a specific inhibitor of ERK kinase (MEK1), reduced H(2)O(2)-induced AR expression.
11444915 The MEK1/2 inhibitor PD098059 abrogated ISO-stimulated ERK activity, albeit the increase in protein synthesis was unaffected.
11454948 In the present study, we examined the effects of PD098059 and U0126, two structurally dissimilar inhibitors of MAP kinase kinase (MEK1/2), on the activation of ERK and Akt stimulated by human 5-hydroxytryptamine(1B) (serotonin) (5-HT1B) receptors.

4.4 Terms That Cannot Be Linked to Any Other Term

No proximity co-occurrences were found for 80 of the genes in the AML dataset. For some of these genes, co-occurrences do occur at the abstract level (data not shown). A trivial explanation for unlinked terms is simply that they were incorrectly named in the query list. This highlights the most difficult aspect of searching the biomedical literature with gene names derived from sequence based databases.

4.5 Types of Errors

When examining the types of errors that a search tool may produce it is convenient to differentiate two types, false negatives (errors of omission) and false positives (errors of inclusion). With PDQ_MED, and similar tools, false negatives can be caused by several factors. These include the use of an incomplete list of name variants for a particular gene in the input list, spelling errors in the query list or the target database (MEDLINE) and "name drift" in the literature. Of these, "name drift" is the most problematic.

For example, from 1986 until 1996, what is now called "estrogen receptor 1" (ESR1) was simply the "estrogen receptor". From 1996 to about 2000, it was called "estrogen receptor alpha" before being changed to the current excepted name. Throughout this period the alternate spelling "oestrogen" was also used.

False positives are generally caused by multiple usages of the same name or acronym. In the ESR1 case, the literature prior to 1996 contains many references which use only the acronym ER. Unfortunately, ER is also frequently used for other uses such as "endoplasmic reticulum" and "emergency room".

Ultimately, it is up to the user to verify the suitability of particular names and to extract the relevant information.

4.6 Other Uses for the "Pharma Sentences"

As mentioned earlier, the "Pharma Sentences" provide a quick method of filtering the literature and highlighting particularly interesting sentences containing one or more of the query terms. Table 5 shows the "Pharma Sentences" for Cathepsin L and Annexin II (Lipocortin II). For Cathepsin L, several inhibitors were found. For Annexin II, regulatory information is found such as regulation of Annexin II by AnV and stimulation of Annexin II translocation to the plasma membrane by phorbol esters.

Table 5. Partial listing of "Pharma Sentences" for Annexin II (Lipocortin II) and Cathepsin L. Query terms are in bold, "Pharma Terms" are in bold italics, and the underlined number is the MEDLINE identifier (PMID).

Cathepsin L;
10698261 This activation was not inhibited by CA-074, a specific inhibitor of cathepsin B, but was strongly inhibited by CLIK-066 and CLIK-181, specific inhibitors of cathepsin L.
10713271 The propeptide of cathepsin S was observed to inhibit cathepsin L with a K(i) of 0.08 nM, yet cathepsin L propeptide inhibited cathepsin S only poorly.
10748021 The mushroom protein is a tight binding inhibitor of papain (K(i) = 0.59 nm), cathepsin L (K(i) = 0.41 nm), cathepsin B (K(i) = 0.48 micrometer), and bromelain (K(i) = 0.16 micrometer) but is inactive toward cathepsin H, trypsin, and pepsin.
10748022 Saxiphilin is now characterized as a potent inhibitor of three cysteine proteinases: papain, human cathepsin B, and cathepsin L.
Annexin II or Lipocortin II;
10084978 However, the immunolocalized tPA protein was most strongly associated with the amnion and chorion, as was its receptor annexin II, suggesting that the amnion and chorion are the targets for decidual tPA.
10213612 These observations furthermore suggest that AnV may regulate the fusogenic function of annexin II.
10376803 With the use of immunofluorescence, annexin II was found to translocate from cytoplasm to plasma membranes in type II cells upon stimulation with phorbol 12-myristate 13-acetate.

4.7 Comparison To Other Tools

There are relatively few tools available for mining the biomedical literature with lists of genes such as those obtained in many types of genomics research. Two of the tools that are available are PubGene [Jenssen *et al.,* 2001] and MedMiner [Tanabe *et al.,* 1999]. Both of these tools are useful for analysing lists of genes. However, both suffer from the fact that they are dependent upon a pre-calculated index. Both PubGene and the GeneCards portion of MedMiner rely upon a gene index which contains the links to MEDLINE. This limits their usefulness in that they can only process terms which have been indexed and, in both cases, only human gene and gene product names are included. Since it does not use a pre-calculated index, PDQ_MED has the advantage of no limitations on the terms that can be searched, other than those imposed by MEDLINE itself. In addition, PDQ_MED does not require an index be maintained and periodically updated. Perhaps the biggest advantage of using a pre-calculated index is speed. Both PubGene and MedMiner are considerably faster than PDQ_MED.

Table 6. Comparison of PDQ_MED with PubGene and MedMiner

Package:	PDQ_MED	PubGene	MedMiner (MEDLINE + GeneCards)
Types of query terms	genes, gene products, drugs, diseases …	genes & gene products	genes, gene products, drugs, diseases …
Requires updating of index	no	yes	yes
Sentence level proximity checking	yes	no	no
Shows matching sentences	yes	no	no
Provides links back to MEDLINE	yes	no	yes
Allows use of "holo" names †	yes	no	yes
Allows use of non-gene terms	yes	no	yes
Species restrictions	none	human only	human only ††
Relative speed	slow	fast	fast

† For example, cFos and cJun together make AP-1.

†† The GeneCards portion of MedMiner is limited to human genes.

5. CONCLUSIONS

We have demonstrated PDQ_MED, a new tool for the search and analysis of the scientific literature. PDQ_MED allows researchers to effectively mine the more than 11 million abstracts in MEDLINE for information that will allow them to fully exploit the results of their genomics experiments. PDQ_MED quickly provides a framework, based on the biomedical literature, which helps to organize and explain why certain sets of genes are co-regulated. PDQ_MED also identifies pairs of genes or gene-disease relationships for which there is no literature precedence. In total, this information can suggest avenues of further research. Overall, PDQ_MED ensures that the researcher can effectively gather and analyze the relevant literature for large sets of genes, proteins and disease terms hence providing a key capability for a successful genomics research project.

REFERENCES

Golub, TR, Slonim, DK, Tamayo, P, Huard, C, Gaasenbeek, M, Mesirov, JP, Coller, H, Loh, ML, Downing, JR, Caligiuri, MA, Bloomfield, CD, Lander ES. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science* 286 (1999):531-537.

Jenssen, T-K, Laegreid, A, Komorowski, J, Hovig, E. A literature network of human genes for high-throughput analysis of gene expression. *Nature Genetics* 28(1) (2000): 21-28.

Scherf, U, Ross, DT, Waltham, M, Smith, LH, Lee, JK, Tanabe, L, Kohn, KW, Reinhold, WC, Myers, TG, Andrews, DT, Scudiero, DA, Eisen, MB, Sausville, EA, Pommier, Y, Botstein, D, Brown, PO, Weinstein, JN. A gene expression database for the molecular pharmacology of cancer. *Nature Genetics* 24(3) (2000): 236-44.

Tanabe, L, Smith, LH, Lee, JK, Scherf, U, Hunter, L, Weinstein, JN. MedMiner: An internet tool for filtering and organizing bio-medical information, with application to gene expression profiling. *BioTechniques* 27 (1999): 1210-1217.

This page intentionally left blank

Glossary

ENTREZ - Internet based search engine for the GENBANK, PUBMED, MEDLINE and OMIM databases maintained by the National Center for Biotechnology Information, the National Library of Medicine and the National Institute of Health. See www.ncbi.nlm.nih.gov/Entrez/.

GDB - The Genome Database (GDB) is the official central repository for genomic mapping data resulting from the Human Genome Initiative. See www.gdb.org.

GENBANK - Database of DNA and protein sequences maintained by National Center for Biotechnology Information. See also ENTREZ.

GeneCards - GeneCards (TM) is a database of human genes, gene products and their involvement in biological processes. It offers concise information about the functions of many human genes culled from multiple sources on the internet. See http://www.dkfz-heidelberg.de/GeneCards/.

MEDLINE - Database of citations, including abstracts, from the biomedical literature maintained by National Center for Biotechnology Information. See also ENTREZ.

OMIM - "Online Mendelian Inheritance in Man" database is a catalog of human genes and genetic disorders. The database contains textual information and references. It also many links to other databases maintained by National Center for Biotechnology Information. See also ENTREZ.

p-value - A measure of statistical significance for a hypothesis test. The p-value for a test is the probability of observing a value for a statistic that is as extreme or more extreme than the observed value if the null hypothesis is true.

Power - The probability that a statistical test rejects the null hypothesis given that the null hypothesis is false.

PUBMED - Database of citations, including abstracts, from the biomedical literature maintained by National Center for Biotechnology Information. See also ENTREZ.

significance level - The probability of rejecting the null hypothesis given that the null hypothesis is true; also referred to as the Type I error of a statistical test.

Index