Design and Analysis of Microarray Experiments: Synthesizing Data for Research Questions

Elissa J. Chesler, PhD
Center for Genomics and Bioinformatics, Department of Anatomy and Neurobiology
University of Tennessee Health Science Center, Memphis, TN

© 2003 Chesler

Overview

Gene expression microarray technology is rapidly becoming an important technique in genomic neuroscience. Despite the tantalizing promise of massively parallel analysis of gene expression, or more appropriately, steady-state transcript abundance, many are left scratching their heads at heat-maps, gene-lists, and other attempts at distillation of rather complex results. Microarray data analysis is often a hypothesis-generating tool, creating as many questions as answers, and methods for the proper treatment of data are still catching up to the technology with which it is acquired. For the typically trained molecular biologist, deploying this method requires a major shift in thinking toward statistical approaches to experiments, reliance on converging evidence, and the potential for subjectivity in the analysis. Growing pains aside, we are beginning to tap into the full potential of this technique, with advanced novel strategies for the use of the microarray to address a large variety of research questions ranging from the simple two-group comparisons, to the dissection of gene-regulatory networks and mapping of regulatory loci modulating expression. This chapter is intended to highlight major issues in array analysis and provide a practical starting point for the use of microarrays in the neuroscience lab.

I. Experimental Design for Research Questions

"To consult a statistician after an experiment is finished is often merely to ask him to conduct a post-mortem examination. He can perhaps say what the experiment died of." R.A. Fisher, 1938

Individual microarrays were initially marketed as experiments, a controlled experiment on a single chip allowing simultaneous expression quantification for thousands of transcripts simultaneously. The array is more properly viewed as a measuring device for collection of multiple simultaneous observations from a single experimental unit, i.e., an individual mRNA sample. Basic principles of experimental design and analysis can be applied to the chip considered in this manner. These include the need for replication, within conditions of interest, to avoid confound of the units, with the effect of interest.

Good experimental design begins with a careful formulation of the research question PRIOR to the collection of data. Though most of the early array software and designs were used for two group comparisons, diverse experimental questions are now being addressed with these methods. Here are some examples:

Which genes are up regulated or down regulated in my mutant mouse?

How can I discriminate tumor cells from normal cells?

How can I differentiate between several different types of tumors?

What genes are differentially expressed in different brain regions and how does this differ by sex or strain?

What is the time course of gene expression following retinal injury?

Which genes are co-regulated?

Does variation in gene expression correlate with variation in other traits?

What are the genetic determinants of gene expression?

Each of these research questions requires different experimental designs and analyses. Well-characterized statistical procedures can be adapted to address these questions more directly than many of the ad-hoc methods that are often used for microarray analysis. Furthermore, these methods allow one to estimate and/or control statistical error rates. The influence of factors that are not of explicit interest to the experimenter can be reduced or eliminated using block designs and replication. Consider whether your goal is detection, i.e., to find a few "signature genes" or dissection, to characterize the involvement of a large set of differentially expressed genes. This strongly affects the number of replicate arrays you will need. Understand what types of replication are needed. Technical replicates (hybridizations of the same samples to multiple arrays, using different dyes for each sample in the case of the two color system) do not allow one to make claims about biological variation, which requires biological replicates (sampling from multiple individuals within the experimental groups).

Pooling of samples is a means of reducing environmental nuisance factors. Many frequently ignored variations in the laboratory environment including handling, housing, temperature, humidity, season, and time of day interact with sex and genetic factors, particularly as they can affect behavior *(1)* and thus presumably gene expression in the brain. It is very important to note that even though one does not explicitly consider these variables, they can systematically

© 2003 Chesler

NOTES

confound an experiment. In other words, unplanned does not equal random. The occurrence of these confounding conditions must be equally likely in each of the experimental groups. Once the samples are pooled, they cannot be taken apart. Consider whether you will later want to explore other factors, i.e., if the experiment is to be expanded to include other groups. For example, if one is interested in sex differences in gene expression, but has chosen to pool samples from across the estrous cycle, the information from cycle phase differences will be lost. If one were interested in the correlation of gene expression within individuals, rather than that which results from the group effect, again, pooling would be inappropriate. Pools also can be biased by individual samples, particularly if some of the samples in the pool are degraded. However, the risks of pooling have been found to be minimal compared to the gains of running the samples on individual arrays *(2)*.

Consult a statistician before undertaking costly data collection to be sure that the experimental design is valid, randomized to avoid confounds, sufficiently powered, and can properly address your research question. Having a clear understanding of how the data will be used once collected can help avoid costly and often irretrievable design errors!

II. Experimental Designs for Spotted Microarrays

The experimental designs and approaches described here can be applied to both commercially manufactured oligonucleotide arrays and the two-color spotted microarrays. However, there are some additional design questions regarding the sample application to the arrays that occur in the use of the two-color system. Much has been written about improving the efficiency of experimental designs for spotted microarrays *(3)*. The simplest approach to spotted microarray experiments is to use a single large pool of common reference mRNA, which can be made or purchased from commercial vendors. The advantage to this approach is that the reference stock is always available, and thus, additional samples can be added easily to the sample design. More complex designs are aimed primarily at reducing chip costs. These are a bit more difficult to extend, and depend heavily on good hybridization results. The balance of an experiment, and thus the usefulness of the entire set of data can be compromised or lost if a single component fails. These designs have another practical drawback. Technicians must keep very careful track of which samples are to be hybridized together and which dye is to be used with each. Even with the best espresso on board, this is no mean feat! A common reference sample reduces this complexity, leaving only the need to counter-balance dyes between the reference and experimental groups. Half of the arrays in each group should be hybridized to Cy5 Red and Half to Cy3 Green. This will allow a good assessment of the array effects, dye effects, and group effects. Another common approach, which uses half the number of arrays, is the 'Dye Swap.' In this design, two mRNA samples are obtained from each group or condition in the experiment, and split into

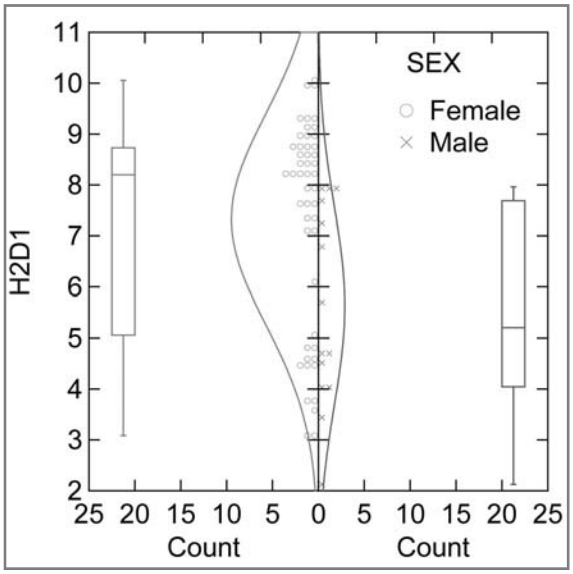


Fig. 1 A statistical analysis (t-test) for H2-D (Affymetrix Mouse U74Av2 Probe Set 197541_f_at), a transcript with a misleading fold change of 3.4. The t-test p-value is 0.005, non-significant when considering 12422 tests performed on the array. The high degree of overlap between the two data sets is not taken into account by the fold change approach. Box and whisker plots show the medians, 25% and 75% quartiles, and whiskers to the nearest point outside within 1.5 interquartile ranges. Projected normal distributions are mean centered. Note that the data are actually bi-modal within each sex, suggesting that another factor explaining gene expression variation is present in the data.

© 2003 Chesler

two aliquots. One of each is labeled with each dye, and the samples are then hybridized with the sample having the opposite group and dye. Though Affymetrix arrays are much more consistent, batch effects in processing can occur. Technical replicates can be used to verify consistency.

III. Group Comparisons

Early microarray analysis employed a simple quantification of "fold-changes" in gene expression levels between a pair of experimental conditions. This approach was developed when a microarray experiment consisted of a pair of manufactured arrays, or a single spotted slide, with one sample from each of two conditions. Because this approach contains no biological replication, an estimate of the noise in the experiment is lost. Were the two samples kept consistently at the same temperature at all times and for the same duration? Were the mutant mouse male and the wild type female? Now that replicate samples are becoming the norm, consideration of this noise is possible. While fold change may have intuitive appeal for biologists, the approach is problematic for several reasons. Fold change merely considers the effect size without respect for variability in that effect. An example of a misleading fold change is shown in Figure 1. This is particularly problematic in array experiments in which there is very low sample size and the precision of the mean estimates is poor. There is no ability to control the error rate, and arbitrary fold change cut-offs have emerged. The approach also extends poorly to more complex experimental designs. Many would agree that even small changes in signaling molecules can have massive biological effect, and the fold change cut-off prevents their detection.

Statistical approaches essentially estimate the signal to noise ratio. The two group t-test or multi-group ANOVA models compare variability between groups to variability within, effectively asking the question, is expression of this gene more similar between individuals in the same group than it is different from individuals in the other groups? Other sources of variation in feature intensity that relate to the hybridization, array feature, dye, or other technical factors can all be included in the statistical model of gene expression, so that only the signal actually attributable to the conditions of interest can be assessed *(4)*. Non-parametric statistical methods have also been proposed to evaluate group differences, including permutation based tests and Wilcoxon or Kruskal-Wallace tests. It should be noted that the method of normalization (discussed by Dr. Miles) is more influential than the use of parametric versus non-parametric analysis *(5)*. A statistic is calculated to test whether the null hypothesis of no expression difference is true, or if it should be rejected in favor of the alternative hypothesis of an expression difference. A probability of getting a chance result of the same magnitude of that observed is calculated (p-value). This is estimated from theoretical distributions or via computational approaches such as the permutation test, which estimates the exact p-value obtainable from the sample data. A threshold, alpha (typically 0.05 by historical convention), is chosen for an acceptable p-value, and the resulting p-value is compared to this threshold to determine whether to retain or reject the null hypothesis. These approaches can be extended to analysis of factorial designs, in which several treatments (e.g., sex and brain tissue type) are considered simultaneously. Contrast analysis (Figure 2) can be employed to evaluate particular comparisons of interest while maintaining statistical power. For

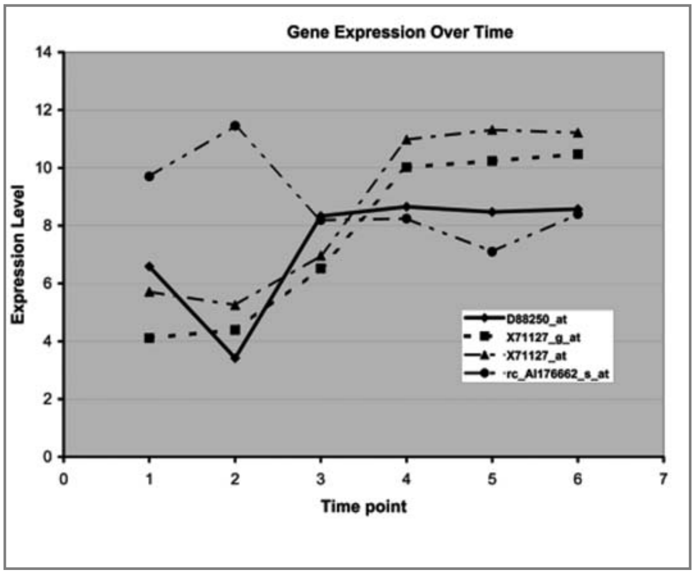


Fig. 2 Expression of rat transcripts at various time points after injury. Contrast analysis allows categorization of transcripts by response profile.

© 2003 Chesler

NOTES

example, in an experiment measuring gene expression at five time points following injury, one may wish to contrast expression in the early time points vs. the later time points.

Limitations on problem sizes imposed by computational resources often demand that only a subset of interesting genes be analyzed. Often, the fold change is applied to further reduce the set of "interesting" results after a statistical test is applied. Particularly in the event that one is interested in genetic dissection, this can be a hazardous approach! Many important genes have statistically detectable expression differences in different groups, with very small but consistent effects. The notion that a doubling of gene expression is required for an interesting biological effect is questionable to say the least, particularly for genes residing in early stages of signal transduction cascades, or that regulate transcription of many other genes. Despite this major drawback, the approach has some practical value, only the large effect size differences are likely to be validated using other confirmatory methods such as RT-PCR. If experimental validation is not the primary concern, it is quite reasonable to consider more than the top few transcripts on the 'gene list'. This is particularly so if comparisons of up- or down- regulated transcripts to gene ontologies, clustering, or further modeling approaches capable of using larger sets of genes will be employed. Particularly if follow-up studies using text mining, sequence analysis, or other bioinformatics approaches are to be used, the gene list might be best pruned using known gene-gene relationships, or the analyses can be done on repeated draws of the differentially expressed genes.

IV. Multiple Testing Considerations

The simultaneous evaluation of thousands to tens of thousands of transcripts can result in a high rate of detection of chance effects or False Positives. This is called the Type I error rate, the probability of rejecting the null hypothesis given that it is true. With a significance threshold alpha = .05 and 100 genes being tested, the probability of a false positive on a single test is still .05, but the probability of at least one false positive is much greater, exceeding 99%. Thus, while the comparison-wise error rate is only 5%, the "Family-wise" error rate is much higher. Strict procedures for controlling the family-wise Type I error rate such as the Bonferroni adjustment are too conservative for the thousands of tests being employed, leading to a high rate of false negatives. This can be detrimental if dissection is the goal, and subsequent analysis of large sets of genes is to be performed. Often these procedures assume that tests are independent, but many transcripts are co-regulated and thus individual expression comparisons are highly correlated. More flexible and appropriate error control procedures are being developed with the demands of high-throughput biology in mind. The false discovery rate *(6)* controls the rate of false discoveries, i.e., the fraction of false positives among the rejected null hypotheses. This is more powerful (less conservative) than a strict Bonferroni adjustment, and much more intuitive. For 100 gene expression values declared to be significantly different between groups, 10% are likely to be false positives. Keep in mind, we cannot say which ones they are! The positive false discovery rate, pFDR, and estimation of q-values *(7)* (Figure 3), is a novel approach to the multiple testing problem which does not involve the determination of a rejection threshold or control of Type I error, but rather the estimation of the rate of

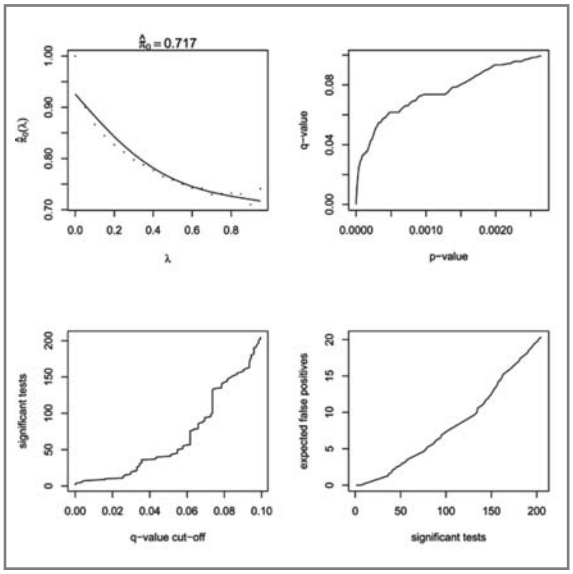


Fig. 3 q-value plots from *qvalue,* a program for estimation of gene specific false discovery rates. **A.** Estimation of the proportion of true null hypotheses, non-significant results. **B.** The q-value, proportion of false positives observed at a given p-value. **C.** The number of significant tests observed at a particular false discovery rate (q-value cut-off). **D.** The number of false positives among a given number of significant tests.

© 2003 Chesler

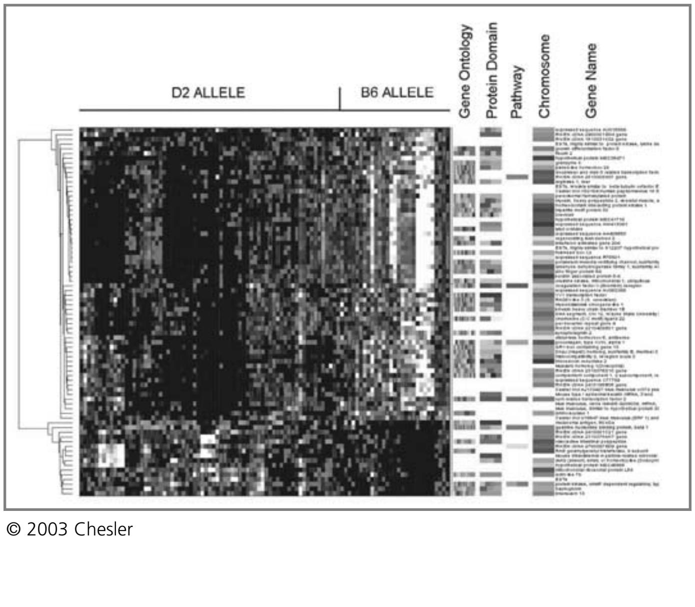
false positives that one would encounter when rejecting the null hypothesis at the level of each observed statistic. These applications are promising and commensurate with the goals of most microarray applications. Software for q-value estimation is freely available at http://faculty.washington.edu/~jstorey/qvalue/ (R is required, see the section on free software).

V. Multivariate Analysis

Gene by gene analysis often leads to a list of significantly differentially expressed genes of arbitrary, yet overwhelming length. The high-dimensionality of microarray data can be daunting, and not conducive to visual presentation or interpretation of results. Reducing the dimensionality through principal component or singular value decomposition, clustering, or multidimensional scaling can greatly facilitate data interpretation. However, these analytic methods often produce results that may not be readily interpreted with existing biological knowledge.

Principal components analysis is a method for identifying linear combinations of variables that explain the most overall variability in the data. Typically, the first two dimensions of these new variables are considered, but if others explain a large portion of the variance, they may also be used. By examining the weights on the variables in the measure, one can often identify meaningful patterns in genes that are differentially expressed, and can determine which samples have elevated and down-regulated levels of these genes. The analysis can be performed in either the gene dimension, to reduce the number of gene variables that explain array variation. For example, the weighted sum of a set of 1000 highly correlated genes is a single new variable derived from 1000 gene variables. In the sample dimension, the goal is to reduce the number of array variables to explain variation in gene expression.

Clustering methods *(8)* have been very popular for the analysis of expression data, particularly when there are more than two groups of samples being compared. Again, ANOVA methods can be applied in this situation, but the intuitive visual displays that can be generated from cluster analyses serve as a ready aid in interpretation. Filtering of results prior to clustering using parametric statistical tests such as the t-test or ANOVA is often recommended. Various methods of clustering are possible. Hierarchical clustering calculates distances between genes using a variety of distance metrics, and assigns them to clusters based on the distance between each gene and the point nearest to it. Nearest neighbor clustering assigns genes to clusters based on the distance to the nearest neighboring gene within the cluster. The centroid method assigns genes based on their distances to the geometric center of the clusters, and is a more unbiased method. The result of this type of analysis is a 'dendrogram.' K-means clustering is a computationally efficient approach that assigns genes randomly to a user-selected number of clusters, and iteratively reassigns them until the distances between the genes and their centroids is minimized. Beware, you will get the result you ask for! Try several k-values. Bootstrapping and other procedures have been proposed to ensure that the result obtained is reliable *(9)*. The results of cluster analysis are presented as "heat maps" showing genes that are up or down-regulated together, along with a dendrogram in the margins showing the proximity of the genes (Figure 4). Once clusters are identified, comparisons to known ontology, sequence databases, and other resources can be performed.



© 2003 Chesler

Fig. 4 Hierarchical clustering analysis performed in dChip. Genes were first filtered by t-test p-values < .0001 (a permutation based false discovery rate of 1.3%) for the comparison between allelic statuses at a chromosome 9 genetic marker. Allelic status is that of the majority of chips in each portion of the figure. This cluster plot reveals that for several genes, C57BL/6J and DBA/2J allele at this location is not the only factor, probably due to the influence of other genetic loci on gene expression. This software also identifies highly represented Gene Ontology, Protein Domain, Pathway or Chromosome categories among the clusters.

NOTES

VI. Regression, Correlation, and QTLs

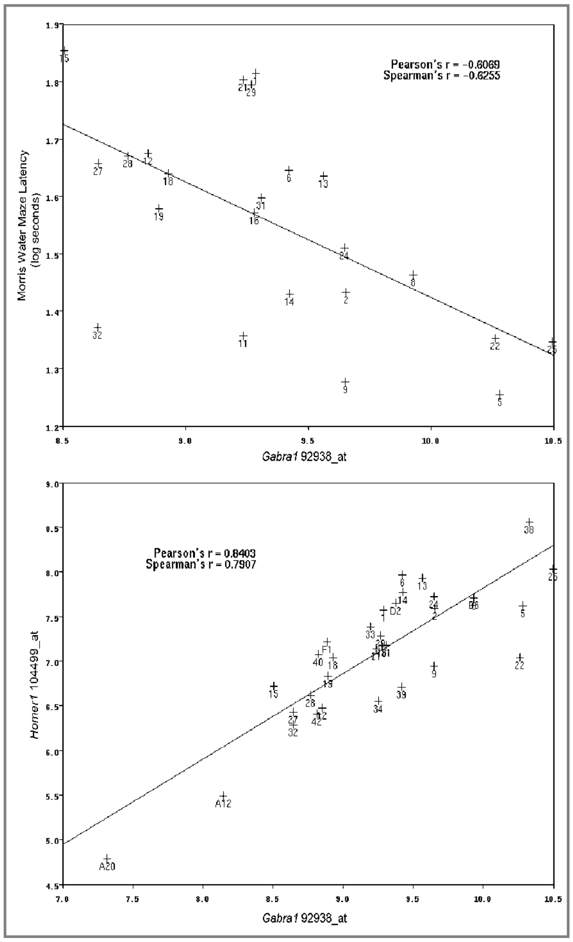
The statistical approach to microarray data allows the opportunity for the application of this technology to novel ways. Microarray data collected in reference populations such as inbred mice can be compared to other data collected in those same mice. This ‘ relational data model’ allows a direct connection between disparate databases, and allows for powerful analysis of the relation between genotypes, neurobehavioral phenotypes, and molecular phenotypes such as microarray measurements of gene expression. Rather than simply comparing groups to make calls of up or down regulation in expression, in this approach, the gene expression levels obtained in the reference population are correlated with either categorical values, such as genotype at known genetic markers, or phenotypes spanning the range from other gene expression levels to brain structure, function, and behavior. Figure 5 shows a gene-to-gene correlation, and a gene to mouse water maze latency correlation *(10)*. These genetic correlation analyses can be used to map quantitative trait loci (QTLs), which are regions of the genome that determine level of expression of a trait (Discussed by Dr. Williams). This approach was initially performed in yeast *(11)* and has been done in the mouse forebrain, cerebellum *(12)*, hematopoietic stem cell lines *(13)*, and liver *(14)*. By performing this analysis in a well-characterized recombinant inbred mapping panel, correlation with over 20 years of neurobehavioral data is possible (www.webqtl.org). To access this data, and examine gene expression correlations with any other phenotype, the user need only search for genes or phenotypes of interest or enter their own phenotypic data collected in the BXD recombinant inbred strain set. From there, identification of loci influencing levels of the trait and determining which gene expression levels correlate with the trait is possible.

VII. From Genes to Gene Networks

Going beyond the perspective of individual genes is the exciting potential of high throughput biology. There is a wealth of techniques being developed for the simultaneous use of expression data from thousands of transcripts to study networks of gene-gene interactions, and relationships to other biological and behavioral traits. Visualization of these results, comparisons to known pathways, and ultimately validating new findings are at the forefront of array analysis today.

Many approaches to network analysis are also in development. Bayesian network analysis allows inference of gene expression networks from expression data and other causal factors. Relevance networks formed from associations of gene expression levels within chips can also be identified. Genetic correlations of gene expression can be used to find genes that share common regulation. Reverse engineering of transcription regulatory networks has also been performed from time series array experiments. A major issue in the development of network models is the determination of the flow of causality. Perturbations of the network either naturalistically via genetic approaches, time changes, or other manipulations are experimental approaches that can be used to determine the direction of gene-gene relationships.

GenMAPP (Discussed by Dr. Miles) allows users to overlay their expression data onto known biological pathways, and users are encouraged to submit their own network and pathway data *(15)*. This tool is available at www.GenMAPP. org, and can now be integrated with Gene Ontology annotations using the tool MAPPFinder *(16)*. The quality of these analyses depends on the annotation of existing



© 2003 Chesler

Fig. 5 Genetic correlations of GABA-A receptor alpha 1 subunit expression and Top) Log latency to find a hidden platform in the Morris water maze, Bottom) Homer1 the Neuronal transcription factor expression in BXD RI strains.

genomic databases. Earlier notions of pathways are also bit unsophisticated. For example, in considering neurotransmission, the classical view of a few synthesis enzymes, possibly some vesicle proteins, receptors, and re-uptake pumps must now be expanded to include the entire cytoskeleton, anchoring proteins, motor proteins, post-translational modifiers of all of the aforementioned gene products, etc.

Text mining allows the incorporation of expression data with known biology, giving new meaning and interpretability to array experiments. First generation literature mining such as PubGene, publicly available at www.pubgene.org, constructs networks for co-occurrence of gene names in Medline abstracts *(17)*. While several good text-mining approaches are being developed, the literature is confounded with many gene symbols used to refer to different gene products. This is especially true of gene products named by size, and short gene names. Text mining efforts are now aimed at discovering relationships between sets of literatures, for example, http://arrowsmith.psych.uic.edu/arrowsmith_uic/index.html (discussed by Dr. Smalheiser). Newer approaches to these types of analysis use latent semantic indexing to discover hidden relationships between genes. For example if gene A is frequently reported to be associated with gene B, and gene C is also frequently associated with gene B, then an association between genes A and C can be inferred. An example of this can be found at http://shad.cs.utk.edu/sgo *(18)*. Relational text mining is another promising approach, which detects not only the co-occurrence of terms, but also parses text to find the direction of the relationship between them (for example, determining that gene A upregulates gene B). Most of these analyses are still largely prototypical.

Sequence mining analyses are used to identify shared motifs in the regulatory regions of genes that cluster together. Upstream sequences of a set of genes can be retrieved and sent to MEME for comparison MEME (http://meme.sdsc.edu/meme/website/meme-intro.html) *(19)*. QTL analysis (or a search of www.webqtl.org) can be used to determine whether co-expressed genes share a common regulatory region. These approaches can provide the needed structure to limit the computational demands of network analyses, by reducing the need to test the relationships between all possible genes.

VIII. Data Management

Microarray data analysis and other high throughput biological techniques have resulted in an explosion of data. Managing the modern biology laboratory is a task that goes beyond the lab notebook and even the large spreadsheets that we are becoming accustomed. Relational databases, ranging from user-friendly varieties such as FilemakerPro or Microsoft Access to Oracle or open source relational software such as MySQL or PostgreSQL, are becoming essential for data sharing and collaboration within and between labs (see the SfN 2002 Short Course Syllabus at www.nervenet.org). These are also powerful tools for assembling data for analysis across experiments. An effective relational database can be used to track array experiments from the birth of the organism to the collection of tissues, sample processing, and data analysis. The push for common data formats for deposit and publication has led to the creation of standards for preparation and annotation of data.

The Minimal Information About a Microarray Experiment (MIAME) document created by the Microarray Gene Expression (MGED) outlines the information required *(20)*. Gene expression indexes should be reported with reliability information for these measurements, information on the probe sequences on the arrays and information on the target samples hybridized to the arrays. Note that with commercial arrays, much of this information is already well-compiled. Fully electronic record keeping will facilitate the incorporation of analysis results, raw array data, and sample information with other array information. MAGE-ML is the MicroArray Gene Expression Markup Language and is based on XML. The language is intended to facilitate the process of sharing and communicating microarray expression data. The full MAGE-ML specification is located at http://cgi.omg.org/cgi-bin/doc?lifesci/01-11-02. User-friendly tools for the laboratory scientist to actually create MAGE-ML documents are under development.

© 2003 Chesler

NOTES

IX. Recommended Books

This chapter was intended to highlight some of the major issues in microarray analysis. Each of these topics is an area of active development, and more extensive treatment of the issues raised here is necessary. There is a wealth of new books about microarray analysis, ranging from simple to quantitatively complex. These books are a few recommendations, and are not based on a comprehensive evaluation of those available.

a. Getting Started

A Biologists Guide to the Analysis of DNA Microarray Data. Steen Knudsen. (2002) This book is inexpensive and gives a highly readable introduction to microarray analysis. Slender on quantitative detail, this book outlines a wide breadth of topics in a highly accessible form. Great foothold for the novice with limited statistical expertise. John Wiley and Sons.

Microarrays for the Neurosciences: An Essential Guide Daniel Geschwind and Jeffery Gregg, Eds. (2002) Written by an SFN short course instructor, this book is specifically geared toward the Neurosciences, with treatment of advanced topics and field relevant examples. MIT Press.

b. Thorough Treatment

Statistical Analysis of Gene Expression Microarray Data. Terry Speed, Ed. (2003) This text is much more detailed, and highly applied. It covers normalization, experimental design, group comparisons and classification algorithms, presented clearly with a good measure of statistical detail while maintaining readability. Chapman and Hall/CRC.

DNA Microarrays and Gene Expression From Experiments to Data Analysis and Modeling (2002). Pierre Baldi and G. Wesley Hatfield. This book is heavier on biological and statistical theory than the others, however it retains practical focus with software reviews, and discussion of algorithms. Cambridge University Press.

X. Free Software

Much of the freely available software for microarray analysis is versatile, user-friendly, well documented, and generally well understood by research community. Other ‘black-box’ analytic tools may be costly, and have a much smaller community of experienced users. These tools may use ad hoc procedures and computational heuristics that are not well known. A few of the popular free softwares are listed below.

a. dChip

This software for Affymetrix arrays performs Model Based Expression Index normalization *(21)*, group comparisons by several methods, clustering, integration of data with gene ontology, and has great graphical capabilities (Fig 5). Perfect match or whole probe set data can be analyzed. The software can be obtained at www.biostat.harvard.edu/complab/ dchip/dchip.exe

b. Bioconductor

The R statistical language is a GNU licensed, open source version of the S statistical language, and can be obtained at http://www.r-project.org. Users contribute packages to perform basic statistical functions that are available at the R archive www.cran.r.project.org. Specific packages for biological data are available at http://www.bioconductor.org, which is features open source/open development software for genomic data analysis and integration with other biological information. Note that the most recent versions of packages and fullest documentation are often obtained directly from the authors' Web sites. The marray group of packages and Affy package are two popular microarray analysis tools for cDNA and Affymetrix Arrays respectively, and other related packages can be used to annotate, and filter results via analysis with respect to other databases. These are programmable and can be readily modified and incorporated by sophisticated users into data pipelines through other scripting languages.

c. Cluster & Tree View

These are widely used softwares developed by Eisen *et al* (1998) for cluster analysis. Tree View is used to visualize the output of Cluster. Both are available at: http://rana.lbl.gov/EisenSoftware.htm

d. ClusFavor

(Gene Expression Cluster and Factor Analysis using Varimax Orthogonal Rotation) is an easy to use and well-documented package for performing clustering and principal components analysis on large datasets. It is particularly oriented toward microarray data, but can be used for clustering of other massive data sets.

Acknowledgments

Data sets from which figures were produced were contributed by Monica Jablonski, Eldon Geisert, and Robert W. Williams.

References

1. Chesler, E.J., Wilson, S.G., Lariviere, W.R., Rodriguez-Zas, S.L., Mogil, J.S. *Neuroscience and Biobehavioral Reviews*, **26**(8): 907-23,2002.; Chesler, E.J., Wilson, S.G., Lariviere, W.R., Rodriguez-Zas, S.L., Mogil, J.S., *Nature Neuroscience*, 5:1101-1102, 2002
2. Han *et al.* cited in *Statistical Analysis of Gene Expression Microarray Data*. Terry Speed, Ed. Chapman and Hall/CRC, 2003.
3. Kerr, M.K and Churchill, G.A. *Biostatistics*, **2**:183-201, 2001.
4. Kerr, M.K., Martin, M., and Churchill, G.A. J. *Computational Biology*, **7**:819-837, 2000.
5. Hoffmann, R., Seidl, T., Dugas, M. *Genome Biology* 3(7):research0033.1-0033.11, 2002.
6. Benjamini, Y. and Hochberg, V., *Journal of the Royal Statistical Society*, B., 1995; Benjamini, Y. and Yeuketeli, D. *Annals of Statistics* **29**(4):1165-1188, 2002.
7. Storey, J.D. *Journal of the Royal Statistical Society, B.*, **64**:479-498, 2003; Storey J.D. and Tibshirani, R. *Proc Natl Acad Sci*, 2003.
8. Eisen, M.B., Spellman, P.T., Brown, P.O., Botstein, D. *Proc Natl Acad Sci* **95**:14863-14868, 1998.
9. Kerr, M. K. and Churchill, G. A. *Proc Natl Acad Sci* **98**(16):8961-8965, 2001.
10. Milhaud, J.M. Halley, H., Lassalle, J.M. *Behav Genet* **32**(1):69-78, 2002.
11. Brem, R.B., Yvert, G., Clinton, R., Kruglyak, L. *Science* **296**:752-755, 2002.
12. Chesler, E. J., Wang, J., Lu, L., Qu, Y., Manly, K. F., Williams, R. W. *Neuroinformatics*. 2003; Wang, J., Chesler, E.J., Williams, R.W., Manly, K. *Neuroinformatics*, 2003.
13. DeHaan *et al*, in prep.
14. Schadt, E.E. *et al*, Nature **422**(6929);297, 2003.
15. Dahlquist, K.D., Salomonis, N., Vranizasn, K., Lawlor, S.C., Conklin, BR. *Nat Genet*, **31**:19-20, 2002.
16. Doniger, S.W., Salomonis, N., Dahlquist, K.D., Vranizan, K., Lawlor, S.C., Conklin, B.R. *Genome Biology*, **4**:R7, 2003.
17. Jenssen, T-K., Laegreid, A., Komorowski, J., Hovig, E. *Nature Genetics* **28**: 21-28, 2001.
18. Homayouni, R., Heinrich, K., Wei, L., Berry, M. J. *Bioinformatics*, 2003.
19. Grundy, W.N., Bailey, T.L., *et al Compu Appli Biosci* **13**(4):397-406, 1997.Brazma, A., Hinghamp, P. *et al*, *Nat Genet* **29**(4):365-71, 2001.
21. Li, C. and Wong, W.H. *Genomebiology* **2**(8):research0032.1-0032.11, 2001; Li, C. and Wong, W.H. *Proc Natl Acad Sci* 98(1):31-36, 2001.
22. Irizarry RA, Bolstad BM, Collin F, Cope LM, Hobbs B, Speed TP. *Nucleic Acids Res*.**15**;31(4):e15, 2003