Meeting Report

Microarray Analysis: A Comparative Approach[1]

Kimberly F. Johnson[2] and Simon M. Lin

Duke Bioinformatics Shared Resource, Duke University Medical Center, Durham, North Carolina 27710

The second annual CAMDA[3] conference (2001) was sponsored recently by the Duke Bioinformatics Shared Resource in Durham, North Carolina, at Duke University on October 15–16, 2001. One hundred and fifty researchers attended from nine countries including Germany, France, Japan, Spain, Korea, Canada, United Kingdom, Israel, and the United States. The CAMDA conferences (1), organized by Kimberly F. Johnson and Simon M. Lin at Duke University, strive to bring together biologists, statisticians, mathematicians, and computer scientists to evaluate different data analysis methodologies for microarrays. Additional information on past and future CAMDA conferences can be found at the CAMDA website.[4]

To provide a basis for comparison, two datasets published previously were selected in March of 2001. Each participant chose one of the datasets and presented his or her methods of analysis. The datasets chosen for CAMDA '01 were the Rosetta Compendium, from a study of 300 expression profiles of yeast mutants and chemical treatments (2), and the NCI60 Cancer Cell Lines with Drug Treatments, a pharmacogenomic database (3).

The conference started with a keynote talk by Dr. Roland Stoughton, Vice President of Bioinformatics at Rosetta Inpharmatics (Kirkland, WA), titled "Microarray Data Analysis: Lessons Learned." The presentation by Dr. Stoughton reviewed the Rosetta yeast compendium experiment conducted 1 year ago, and highlighted experimental design and quality control issues. The second keynote address by Dr. David Lockhart, visiting scholar at the Salk Institute and President and CSO of Aventa Biosciences (San Diego, CA), was titled "Quantitative Expression Profiling in the Brain." Dr. Lockhart provided an excellent overview of microarrays, from fabrication to analysis.

Mining a large reference database such as Rosetta is not only of interest to biologists but also a challenge for new algorithm development. The Rosetta data set was originally analyzed by clustering methods. At CAMDA 2001, Fowlkes *et al.*[5] additionally investigated the clustering approach with a new algorithm called normalized cut. This method identifies clusters not identifiable previously by known methods that indicate biological functionality of the expression data.

Bidaut *et al*.[6] made an interesting contribution by introducing Bayesian Decomposition to the field. They showed that complex microarray data could be decoded into much more simple patterns that are associated with certain biological events.

Lin *et al.*[7] proposed a new concept of "functional genomic units." They demonstrated that the limitations of functional annotation by gene ontology could be complemented using an independent component analysis algorithm. This algorithm deduces the coordinately regulated genes. Their goal is strikingly similar to Bidaut *et al.*,[6] but the algorithmic approaches are different. Lin *et al.*[7] also described an interesting attempt to use the Rosetta data set (cDNA platform) to corroborate a Rac1 transfection experiment (Affymetrix platform).

Zhang *et al.*[8] described a novel approach of supervised component analysis to find the optimum projection of experiments into the gene space. By a supervised reduction of the gene dimensionality, the analyses of the authors provided an intriguing observation: additional experiments such as erg25Δ had been "fished out" when using erg2Δ and erg3Δ as "baits."

The NCI-60 data set provided an excellent exercise opportunity for various computational approaches. Coombes *et al.*[9] presented "Biology-Driven Clustering of NCI-60 Microarray Data." The authors addressed the issues of correcting the cell line annotations and updating the spot annotations of the spots before their analysis. Then with the help of Gene Ontology and Locus Link, the authors clustered the cancer samples by using gene subsets from the same chromosome or from the same functional category. In contrast with the paper presented by Zhang *et al.,*[8] on feature selection by machine learning, the approach by Coombes relies heavily on the human expert opinion in the gene ontology for select-

Received 11/16/01; revised 11/27/01; accepted 11/28/01.
[1] Supported in part by grants from the National Science Foundation (Grant DBI-0136915), North Carolina Biotechnology Center, the Center for Bioinformatics and Computational Biology at Duke University, and the Office of the Vice Provost of Duke University. Paradigm Genetics and Bristol Myers provided additional support. The Critical Assessment of Microarray Data Analysis 2001 conference was held in Durham, NC 27710. It was organized by K. F. J. and S. M. L.
[2] To whom requests for reprints should be addressed, at Duke Bioinformatics Shared Resource, Box 3958, Duke University Medical Center, Durham, NC 27710. Phone: (919) 681-5426; Fax: (919) 681-8028; E-mail: johns001@mc.duke.edu.
[3] The abbreviation used: CAMDA, Critical Assessment of Microarray Data Analysis.
[4] Internet address: http://www.camda.duke.edu.

[5] Fowlkes, C., Shan, Q., Belongie, S., and Malik, J. UC Berkeley, Berkeley, CA. Extracting Global Structure from Gene Expression Profiles.
[6] Bidaut, G., Grant, J., Moloshok, T., Manion, F., and Ochs, M. Fox Chase Cancer Center, Philadelphia, PA. Application of Bayesian Decomposition to Gene Expression Analysis of Deletion Mutation Data.
[7] Lin, S., Liao, X., McConnell, P., Vata, K., Carin, L., and Goldschmidt, P. Duke University, Durham, NC. Using functional genomic units to corroborate user experiments with the rosetta compendium.
[8] Zhang, Z., Page, G., and Zhang, H. Medical University of South Carolina, Charleston, SC. Fishing expedition: a supervised approach to extract patterns from a compendium of expression profiles.
[9] Coombes, K., Baggerly, K., Stivers, D., Wang, J., Gold, D., Sung, H., and Lee, S. M.D. Anderson Cancer Center, Houston TX. Biology-driven clustering of microarray data: applications to the NCI60 data set.

ing a subset of genes before clustering. Coombes clearly demonstrated that selecting appropriate genes is critical for the separation of cancer samples. This technique introduced many biologically interesting questions that were discussed after the paper was presented. Coombes *et al.*[9] show that cancer types differ greatly in their degree of heterogeneity, ranging from homogeneous (colon and leukemia) through moderately heterogeneous (renal and melanoma) to extremely heterogeneous (breast and lung) types based on their expression profiles.

The Coombes paper was ultimately presented with the Best Presentation Award at the end of the conference. This award was determined by a vote of the attendees and Scientific Committee members and carries an award of $1000 to the winning authors(s). This year's award presentation indicated the need to not only analyze the data but to show a compelling biological relevance of the analysis.

Mateos *et al.*[10] described a Self-organizing Tree Algorithm for clustering the samples. This algorithm is a fast and efficient method for clustering gene expression profiles. In addition, the average values of the clusters provided by the Self-organizing Tree Algorithm can be used for obtaining a classification of samples either by unsupervised or supervised methods.

The work by Li *et al.*[11] contributed to additional understanding of the minimum number of replicates needed to reach enough statistical power to detect differential gene expression. Sluka[12] presented a refreshing investigation of extracting knowledge from microarray data by incorporating MEDLINE. A new text data-mining algorithm was demonstrated to retrieve relevant information from the literature. 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 that relates the two genes is of interest.

Finding the relationship between gene expression profiles and drug response is a key element of pharmacogenomics. Dubitzky *et al.*[13] address this issue by detecting broad-band and selective correlation patterns. Chang *et al.*[14] reported an analysis by Bayesian network learning to deduce gene-gene, gene-drug, and drug-drug interactions.

A second paper by Berrar *et al.*[15] described the attempt of using association-rule mining to find the interaction patterns. The work by Dasgupta *et al.*[16] suggested the Partial Least Square method could be used to predict the anticancer therapeutic response and to identify marker genes.

A lively discussion of the merits and flaws of a particular technique followed each presentation. The Monday meeting ended with a reception and poster session to allow participants time to meet informally with others.

Closing remarks were presented by Dr. John Weinstein, highlighting the need for bioinformatics to unleash the power of genomics. Dr. Weinstein also discussed the need for more training for bioinformaticians and the need to distinguish between applied bioinformatics *versus* theoretical bioinformatics.

By the end of the conference, it was apparent that microarray data analysis is still in its infancy. Participants left Durham, NC, with both practical information and increased appreciation for new algorithm developments. The papers presented in CAMDA 2001 will be published by Kluwer Academic Publishing (Norwell, MA) in early 2002. Clearly, we have come a long way from the cluster analysis conducted by the Brown Lab (Stanford, Palo Alto, CA), and certainly conference participants will be looking forward to the third CAMDA conference, which will be held in 2002.

Acknowledgments

We thank the scientific committee for suggesting the data sets and for reviewing the submissions. We also thank the advisory committee for their input and critiques: John Harer (Duke University), Michael Colvin (Duke University Medical Center), John Weinstein (NIH).

References

1. Johnson, K. F., and S. M. Lin. Critical assessment of microarray data analysis. The 2001 challenge. Bioinformatics (Oxf.), *17:* 857–858, 2001.

2. Hughes, T. R., Marton, M. J., Jones, A. R., Roberts, C. J., Stoughton, R., Armour, C. D., Bennett, H. A., Coffey, E., Dai, H., He, Y. D., *et al*. Functional discovery via a compendium of expression profiles. Cell, *102:* 109–126, 2000.

3. Scherf, U., Ross, D. T., Waltham, M., Smith, L. H., Lee, J. K., Tanabe, L., Kohn, K. W., Reinhold, W. C., Myers, T. G., Andrews, D. T., *et al*. A gene expression database for the molecular pharmacology of cancer. Nat. Genet., *24:* 236–244, 2000.

[10] Mateos, A., Herrero, J., Tamames, J., and Dopazo, J. CNIO Bioinformatics Unit, Madrid, Spain. Supervised and hierarchical unsupervised neural networks for clustering both gene expression profiles and samples.
[11] Li, Y., Zhang, L., Speer, M., and Martin, E. Duke University, Durham, NC. Evaluation of current methods of testing differential gene expression and beyond.
[12] Sluka, J. Inpharmix, Greenwood, IN. Extracting knowledge from genomic experiments by incorporating the biomedical literature.
[13] Dubitzky, W., Berrar, D., Granzow, M., Eils, R. German Cancer Research Center, Heidelberg, Germany. Detecting broad-band and selective correlation patterns among gene expression and drug activity data.
[14] Chang, J., Hwang, K., and Zhang, B. Seoul National University, Seoul, Korea. Analysis of gene expression profiles and drug activity patterns for the molecular pharmacology of cancer.

[15] Berrar, D., Dubitzky, W., Granzow, M., and Eils, R. German Cancer Research Center, Heidelberg, Germany. Analysis of gene expression and drug activity data by knowledge-based association mining.
[16] Dasgupta, N., Lin, S., and Carin, L. Duke University, Durham, NC. Modeling pharmacogenomics of the NCI-60 anticancer data set: using kernel PLS to correlate the microarray data to therapeutic responses.