Assessing the Information Content of Microarray Time Series

E. Yang
Rutgers University, USA

I.P. Androulakis
Rutgers University, USA

INTRODUCTION

While the rise of microarrays has heralded a new era in molecular biology with its ability to measure the expression level of thousands of genes at once, the usefulness of microarrays is exigent upon the ability to obtain accurate gene expression data for the individual genes (Bowtell, 1999; Brown & Botstein, 1999; Cheung, Morley, Aguilar, Massimi, Kucherlapati, & Childs, 1999). However, there has been significant criticism as to how meaningful the information derived via microarrays is. In cases where one has attempted to find genes that correlated to types of cancer or survival rate, it was found that different analysis techniques would often times yield radically different set of genes, calling into question the validity of the overall experiment itself (Dupuy & Simon, 2007). It is our contention that part of the problem associated with microarrays is that there does not exist a coherent method for dealing with data quality, and if a coherent method for dealing with data quality existed, many of the criticisms of microarrays could be addressed.

BACKGROUND

"Fishing Expedition" is normally used in a negative connotation in the legal field. The negativity has carried over to the scientific field and describes an experiment in which the researcher does not know precisely what one is looking for. However, the promise of microarrays is the fact that they allow for just such experiments, and coupled with different algorithms allow for a data driven approach to science (Nakai & Vert, 2002). By identifying the possible targets of gene regulation, researchers can then formulate more specific experiments to validate such a hypothesis. While the technology behind microarrays consistently advances in terms of the density of microarrays as well as the repeatability between each individual microarray, there still exists the major issue of noise. Whilst the technological improvements themselves are able to minimize the technical noise, there still exists significant *biological noise* which for complex multitissue organisms cannot be easily overcome with technology. For instance, despite the standardization of rat/mice lines, there still exists significant variation in any reading taken from a population of animals. Due to this noise, it is difficult to identify the genes which actually respond to a given treatment, and those genes whose fluctuations are due to random noise.

Therefore the most important algorithms for the processing of microarray data are those that select meaningful genes from the thousands that are measured via the microarray. The most common metric used by these algorithms is that of *statistical significance* (Smyth, Yang, & Speed, 2003). There is one caveat with the use of statistical significance primarily in the fact that not all genes that are statistically significance are biologically relevant. The set of biologically relevant genes is dependent wholly upon the biological ground truth, whilst the set of statistically significant genes are dependent upon the number of replicates, the quality of the microarray platform, inherent SNR, as well as biological significance. This is a recognized problem which has been addressed by many researchers as a *feature selection* problem under the assumption that the set of biologically relevant genes ought to be able to work as classifiers between the different states being tested (Wu, 2005).

While not all genes that are statistically significant are biologically relevant and vice versa, there does exist a tendency for biologically relevant genes to be statistically significant as well. Therefore, by selecting statistically significant genes, one increases the likelihood of identifying biologically relevant genes

Copyright © 2008, IGI Global, distributing in print or electronic forms without written permission of IGI Global is prohibited.

as well. However for one to have confidence in these initial results, care must be taken in the selection of statistically significant genes, paying special attention to normalization and the setting of statistically significant cutoffs.

STATISTICAL SELECTION OF GENES FROM MICROARRAYS

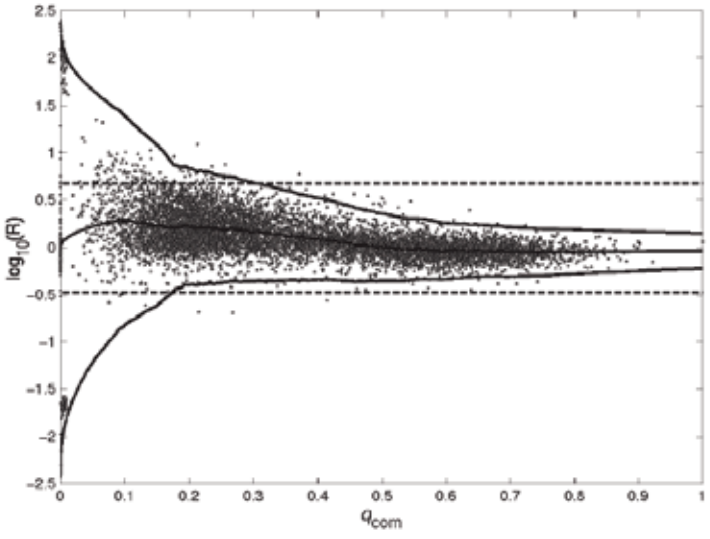
Normalization

The use of normalization is important because many of the selection algorithms that look for over/under expressed genes in a two state experiment base their assumption off of the fact that the initial distribution log is normalized and compute their confidence intervals according to that distribution. Therefore by transforming the data so it does conform to the log-normal distribution, one is able to use standard statistical tests such as the t-test to ascertain whether or not the variations are due to noise or due to some intrinsic change in the expression level of the gene.

One of the challenges with analyzing microarray data is the problem of translating the recorded intensity level obtained by the detection equipment and determining the true expression value of the given probe. Generally, for genes which show a high intensity value, there exists a good correlation between the intensity between the two samples follows a roughly linear trend. However, at low intensity levels, the correlation between the two samples deviates from this (see Figure 1). The justification for the majority of genes being linearly correlated is that under most situations, only a small fraction of genes are responding to the overall treatment and even with the addition of noise, they should be consistent over multiple chips in *temporal gene expression* experiments. The small fraction of genes that do deviate from this linear relationship are then the ones that deviate by a given statistically significant level. Techniques such as the LOESS, dCHIP, and PDNN (Cleveland, 1979; Li & Hung Wong, 2001; Millenaar, Okyere, May, van Zanten, Voesenek, Peeters, 2006; Nielsen, Gautier, & Knudsen, 2005), attempt to normalize the data in such a manner in which the correlation between two samples becomes consistent, thereby allowing for easier identification of statistical outliers.

LOESS (Cleveland, 1979) is a local nonparameter method which attempts to fit a low order polynomial, normally linear or quadratic, to the scatter-plot attempting to minimize the random variations in the data. It is most often used for the *normalization* of two dye experiments, primarily to account for the slight difference in affinity between the two dyes at low expression levels, but can be used generally to correct for the nonlinearities found at lower intensity levels. It is similar to a nonlinear regression fit, except it performs a local regression upon blocks of data. The blocks of

Figure 1. The deviation from the log normal distribution at low intensity levels. The LOESS curve centers the distribution and forces log-normality (Wang, Hessner, Wu, Pati, & Ghosh, 2003)



6 more pages are available in the full version of this document, which may be purchased using the "Add to Cart" button on the publisher's webpage:
www.igi-global.com/chapter/assessing-information-content-microarray-time/12931

Related Content

Semi-Automatic Systems for Exchanging Health Information: Looking for a New Information System at Fixed E-Healthcare Points for Citizens in Greece
Dimitrios Emmanouil, Antonia Mourtzikouand John Mantas (2014). *International Journal of Reliable and Quality E-Healthcare (pp. 38-54).*
www.irma-international.org/article/semi-automatic-systems-for-exchanging-health-information/124947

Statistical Inference to Develop Budgets From Activity-Based Funding Costing Data
Harry Chiam (2019). *Clinical Costing Techniques and Analysis in Modern Healthcare Systems (pp. 120-127).*
www.irma-international.org/chapter/statistical-inference-to-develop-budgets-from-activity-based-funding-costing-data/208280

Prevalence in MSM Is Enhanced by Role Versatility
Andrés J. Cortés (2018). *Big Data Analytics in HIV/AIDS Research (pp. 140-148).*
www.irma-international.org/chapter/prevalence-in-msm-is-enhanced-by-role-versatility/202917

Human and Organizational Factors of Healthcare Data Breaches: The Swiss Cheese Model of Data Breach Causation And Prevention
Faouzi Kamounand Mathew Nicho (2014). *International Journal of Healthcare Information Systems and Informatics (pp. 42-60).*
www.irma-international.org/article/human-and-organizational-factors-of-healthcare-data-breaches/110185

Assessing Physician and Nurse Satisfaction with an Ambulatory Care EMR: One Facility's Approach
Karen A. Wager (2011). *Developments in Healthcare Information Systems and Technologies: Models and Methods (pp. 54-64).*
www.irma-international.org/chapter/assessing-physician-nurse-satisfaction-ambulatory/46668