Data Mining and Meta-Analysis on DNA Microarray Data

Triantafyllos Paparountas, Biomedical Sciences Research Center "Alexander Fleming," Greece

Maria Nefeli Nikolaidou-Katsaridou, Biomedical Sciences Research Center "Alexander Fleming," Greece

Gabriella Rustici, European Molecular Biology Laboratory-European Bioinformatics Institute, UK

Vasilis Aidinis, Biomedical Sciences Research Center "Alexander Fleming," Greece

ABSTRACT

Microarray technology enables high-throughput parallel gene expression analysis, and use has grown exponentially thanks to the development of a variety of applications for expression, genetics and epigenetic studies. A wealth of data is now available from public repositories, providing unprecedented opportunities for meta-analysis approaches, which could generate new biological information, unrelated to the original scope of individual studies. This study provides a guideline for identification of biological significance of the statistically-selected differentially-expressed genes derived from gene expression arrays as well as to suggest further analysis pathways. The authors review the prerequisites for data-mining and meta-analysis, summarize the conceptual methods to derive biological information from microarray data and suggest software for each category of data mining or meta-analysis.

Keywords: *Biological Information, Data Mining, Gene Networks, Meta-Analysis, Microarray*

INTRODUCTION

The ability to investigate an organism's entire genomic sequence has revolutionized biological sciences. One aspect of this phenomenon was the fabrication of gene microarrays in the late 1980s (Fodor et al., 1991). Array based high-throughput gene expression analysis is widely used in many research fields; gene expression microarrays have been used in numerous applications, including the identification of novel genes associated with diseases, most notably cancers (Lee, 2006; Kim et al., 2005; Al Moustafa et al., 2002; Lancaster et al., 2006), the tumors classification (Perez-Diez, Morgun, & Shulzhenko, 2007; Nguyen & Rocke, 2002; Ray, 2011; Dagliyan, Uney-Yuksektepe, Kavakli, & Turkay, 2011; Best et al., 2003) and the prediction of patient outcome (Mischel, Cloughesy, & Nelson, 2004; Simon, 2003; Futschik, Sullivan, Reeve, & Kasabov, 2003; Michiels, Koscielny, & Hill, 2005; Liu, Li, & Wong, 2005), as well

DOI: 10.4018/ijsbbt.2012070101

as the -cell line related- drug chemosensitivity identification (Amundson et al., 2000; Dan et al., 2002; Kikuchi et al., 2003; Sax & El-Deiry, 2003; Ikeda, Jinno, & Shirane, 2007; Baggerly & Coombes, 2009; Ory et al., 2011).

Typically, a microarray experiment generates a list of genes that have been identified as statistically significant differentially expressed (DEGs). Following this ensues the real challenge of assigning biological significance to the results and reconstructing pathways of interactions among DEGs. Several software tools for pathway analysis, gene ontology analysis and gene prioritization are routinely used for identifying common features in lists of DEGs.

As the quantity and size of microarray datasets continues to grow (Table 2, Microarray repositories), researchers are provided with a rich data resource, but also face interoperability and data management issues. The primary data should be stored in a MIAME (Minimum Information About Microarray Expression) compliant format, which is a set of guidelines outlining the minimum information that should be included when describing a microarray experiment. It is required in order to facilitate the interpretation of the experimental results unambiguously and to potentially reproduce the experiment (Brazma et al., 2001). Complimentary to the standardization of data storage, workflows (School of Computer Science, 2008) (Table 3, Holistic Approaches) offer a solution to data management and analysis issues as they enable the automated and systematic use of distributed bioinformatics data and applications from the scientist's desktop. In order to address reliability concerns as well as other performance, quality, and data analysis issues, the National Center for Toxicological Research, NCTR, has initiated the MAQC, MicroArray Quality Control project, (Shi et al., 2006, 2010), in response to the FDA's (U.S. Food and Drug Administration, n.d.) Critical Path Initiative (Coons, 2009; Mahajan & Gupta, 2010; Woodcock & Woosley, 2008). The main target of this initiative is to develop guidelines for microarray data analysis and provide the public with large reference datasets.

1. PREREQUISITES FOR DATA MINING

Generating high quality microarray data requires applying stringent quality control measures and best practices at each individual step of the process, starting with choosing the most appropriate experimental design for the study, the correct experimental platform, the protocols for sample preparation, processing, and ultimately ending with the data analysis approach for normalization and statistical analysis. (Chuaqui et al., 2002) provides a short review on the validation of primary analysis methods, (Allison, Cui, Page, & Sabripour, 2006; Dupuy & Simon, 2007; Ioannidis et al., 2009; Shi et al., 2010) inform on reasons of result discrepancies after reanalysis of raw data across different teams, while (Troester, Millikan, & Perou, 2009) provide a short list of guidelines for statistical analysis and reporting of microarray studies.

1.1. Experimental Design

Experimental design is one of the most important aspects of a successful experiment related to the identification of differential gene expression patterns. Proper experimental design is crucial to ensure that the biological questions of interest can be answered and that this can be done accurately. Appropriate experimental design (Churchill, 2002; Festing & Altman, 2002; Qiu, 2007; Shaw, Festing, Peers, & Furlong, 2002) allows a more accurate identification of DEGs and prediction of false positives (Benjamini & Hochberg, 1995; Reiner, Yekutieli, & Benjamini, 2003; Wolfinger et al., 2001). Fundamental principles of experimental design are simplicity, replication & statistical power (Festing & Altman, 2002) and bias prevention through randomization & blocking (Damaraju, 2005; Johnson & Besselsen, 2002).

1.1.1. Replication

The effects of the: Treatment-group, subject, sample, gene, probe and noise are the major sources of variability in microarray experiments. Ideally to estimate the statistically significant