Gene expression

MeSHer: identifying biological concepts in microarray assays based on PubMed references and MeSH terms

Amira Djebbari1,2, Svetlana Karamycheva1, Eleanor Howe5 and John Quackenbush1,3,4,5,*

1The Institute for Genomic Research, 9712 Medical Center Drive, Rockville, MD 20850, USA, 2Department of Computer Science and 3Department of Biochemistry, The George Washington University, Washington, DC, USA, 4Department of Statistics, Bloomsberg School of Public Health, The Johns Hopkins University, Baltimore, MD, USA and 5Department of Biostatistics and Computational Biology, Dana-Farber Cancer Institute, Boston, MA, USA

Received on April 4, 2005; revised and accepted on May 16, 2005
Advance Access publication May 26, 2005

ABSTRACT
Summary: MeSHer uses a simple statistical approach to identify biological concepts in the form of Medical Subject Headings (MeSH terms) obtained from the PubMed database that are significantly overrepresented within the identified gene set relative to those associated with the overall collection of genes on the underlying DNA microarray platform. As a demonstration, we apply this approach to gene lists acquired from a published study of the effects of angiotensin II (Ang II) treatment on cardiac gene expression and demonstrate that this approach can aid in the interpretation of the resulting ‘significant’ gene set.

Availability: The software is available at http://www.tm4.org
Contact: johnq@jimmy.harvard.edu
Supplementary information: Results from the analysis of significant genes from the published Ang II study.

INTRODUCTION
DNA microarrays have been widely used to survey large numbers of genes and to identify those that correlate with the particular biological processes under study. Laboratory-based protocols for generating gene expression data have greatly improved in recent years and methods for data analysis and the identification of ‘significant’ genes have evolved substantially, the interpretation of the functional roles played by these genes remains an ongoing challenge. Tools such as EASE (Hosack et al., 2003) and GO Miner (Zeeberg et al., 2004) allow Gene Ontology assignments (GO terms) (Ashburner et al., 2000) and pathway assignments to be used to identify general classes of genes that are overrepresented in a particular dataset. While such approaches have been widely used, these cannot provide direct associations with disease states or other related phenomena. The primary source of such information remains the corpus of biological literature and mining the literature remains a laborious process. Although a number of utilities, such as HAPI (Masys et al., 2001) and MedMiner (Tanabe et al., 1999), use the Medical Subject Headings (MeSH terms, http://www.nlm.nih.gov/mesh/meshhome.html) and literature references from the PubMed database (http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=PubMed) to offer some insight into what might be represented in a particular gene set, they do not provide estimates of the significance of any particular association derived from such analysis. MeSHer was developed to extend these methods using the Fisher’s exact test to estimate the likelihood that a particular MeSH term occurs in a selected gene set by chance relative to its association with the overall representation of MeSH terms associated with genes on the array and to provide a list of literature references that link these terms to specific genes.

METHODOLOGY
MeSH is the National Library of Medicine’s controlled vocabulary for indexing articles in the PubMed database. MeSH terms are assigned by expert curators who attempt to summarize the information presented in each indexed article and the genes described therein. There are 19 000 MeSH terms organized in a hierarchy based on 15 top-level categories that provides a consistent way to retrieve information regarding manuscripts that may use different terminology to describe the same biological or medical concepts. MeSH terms have previously been used effectively to provide insight into results from microarray studies. Masys et al. (2001) applied MeSH terms associated with genes that had been used to discriminate between leukemia subtypes (Golub et al., 1999) and discovered that these genes were not simply markers of hematopoietic lineage but were also involved in cancer pathogenesis, confirming the result uncovered by Golub et al. (1999) in their analysis.

MeSHer builds on the available annotation for microarray and other genomic resources captured in the RESORCERER database (Tsai et al., 2001). RESORCERER represents the vast majority of commercial and cDNA microarray resources used for human, mouse, rat, zebrafish and Xenopus, as well as RefSeq genes for these species and other resources such as collections of mouse embryonic stem cell knock-out lines and databases of sequences with identified coding single nucleotide polymorphisms; a version of RESORCERER has recently been released for plants. For each resource represented in the RESORCERER, a variety of annotation is provided, including gene names, genomic locations, links to orthologous genes in other species and relevant PubMed references, which MeSHer uses as the source of MeSH terms for analysis.

In building the RESORCERER, we faced a number of possible alternatives for associating PubMed references with individual elements represented in any resource. An obvious approach might be to use gene names. While gene names are very useful in manuscripts, computationally they are quite difficult to use, in part because they suffer from the problems of synonymy and polysemry; synonymy means that one gene can be called by several names and polysemry means that the same name can refer to several genes, and both of these problems are common. Consequently, we chose a sequence-based approach to link the individual elements in each resource...
to well-characterized proteins and through those to the PubMed references and their associated MeSH terms. The basis for the links are the TIGR Gene Index (TGI) databases (Quackenbush et al., 2000, 2001; Lee et al., 2005). The TGI are constructed for each of the 83 species represented by collecting high-quality expressed sequence tag and gene sequences, clustering them based on sequence similarity and assembling them into high-confidence Tentative Consensus (TC) sequences. The resulting TCs are then searched against a non-redundant amino acid database, populated with sequences from Swiss-Prot (Boeckmann et al., 2003, http://us.expasy.org/sprot/), GenPept (ftp://ftp.ncbi.nih.gov/genbank/), PIR (Wu et al., 1998), and other resources.

Among the most significant MeSH terms were those linked to tissue injury. Ang II is known to induce hypertrophy of the heart and vasculature, and recent studies have suggested a role for Ang II in the development of heart disease, to be significant in this analysis. This is of particular note as the study by Larkin et al. (2004) was the first to suggest a potential mechanistic link between hypertension and Alzheimer’s, an association found previously in a number of clinical studies (Sparks et al., 2000; Sparks, 1997; Kang et al., 2005). A list of the 15 most significant upregulated and downregulated genes and the associated significant MeSH terms and their P-values, as well as the MeSHer hierarchical output, can be found in the Supplemental Material.

**SUMMARY AND CONCLUSIONS**

MeSHer uses a relatively straightforward approach to using the literature to shed light on the biological processes that may be associated with a particular list of ‘significant’ genes selected from a larger population. The advantage of MeSHer relative to previously described approaches is that it provides statistical support for the individual MeSH terms and displays them in the context of the MeSH hierarchy, allowing them to be more easily interpreted and providing an important supplement to the existing tools for the interpretation of array data. The software supporting MeSHer is freely available with source code distributed under the open-source Artistic License and available on http://www.tm4.org. In the future, we plan to integrate MeSHer into the MeV (Saeed et al., 2003) microarray data mining software to provide biological concepts found in the given gene lists side-by-side with MeV’s clustering algorithms as we believe that such integration would help biologists interpret their microarray datasets more seamlessly.

**Conflict of Interest:** none declared.

**REFERENCES**


