Mining Microarray Datasets Aided by Knowledge Stored in Literature

Rob Jelier, MSc; *Guido Jenster, PhD; **Lambert C. Dorssers, PhD;
Erik M. van Mulligen, PhD Barend Mons, PhD; Jan A. Kors, PhD
Departments of Medical Informatics, *Urology and **Pathology, Erasmus MC, Rotterdam, The Netherlands

Abstract
DNA microarray technology produces large amounts of data. For data mining of these datasets, background information on genes can be helpful. Unfortunately most information is stored in free text. Here, we present an approach to use this information for DNA microarray data mining.

Background
DNA microarray technology allows the parallel analysis of the mRNA expression level of thousands of genes in a single experiment. In these datasets hundreds of genes may show significant changes in expression levels. The poignant matter at hand is to extract useful information efficiently. Tools for the evaluation of DNA-microarray data are many, though mainly based on statistical approaches. Clearly, exploitation of background knowledge available for genes can facilitate the evaluation of DNA microarray data. Examples of relevant information are listed in Table 1.

Table 1 Relevant gene information

Gene function
Part of regulatory pathway?
Tissue/cell type specific expression
Protein localization
Chromosomal gene location
Role in cancer
Known co-expressed genes

Currently, large datasets are available that contain specific information on genes and gene products, such as SwissProt, KEGG, OMIM, and Medline. Unfortunately, most information is stored in unstructured free text, such as in Medline and OMIM. We are developing tools to make free text information available for DNA microarray data mining, in order to link genes to literature, mine the literature, and visualize the found information.

Methodology
A thesaurus-based approach is used for mining free text. The thesaurus used is a combination of MeSH and a large set of concepts specifically compiled for medical and molecular biological purposes, e.g., by containing all known human gene names. Using Collexis® technology concepts are identified in the text and translated into fingerprints[1]. A fingerprint is a list of concepts with a weight that relates to the importance of the concept in the text. The weight is based on characteristics such as specificity of the concept and its frequency in the text. Fingerprints allow rapid evaluation of large amounts of literature and are the basis of our information retrieval system for genes. For higher level analysis, information discovery tools will be used, building forth on foundations laid by the Associative Concept Space (ACS) system[2,3]. This system places concepts in a virtual space based on their co-occurrence in the fingerprints. Evaluation of the developed tools is performed as part of an ongoing collaboration with cancer researchers within our institute. Predictions from our system will be verified experimentally.

Results
Using the ACS system, we have built concept spaces for several DNA array experiments. Qualitative evaluation of the ACS has yielded interesting connections between concepts known to be biologically sound. Results also show that homonymy for gene symbols, i.e., one symbol referring to multiple genes, is a problem. In some sets 40% of the gene symbols are homonyms.

Discussion
Currently a more quantitative evaluation is taking place for the ACS. Also, the thesaurus and the ACS system are being extended to extract and represent information about genes, such as listed in Table 1. This should allow, for example, to discover genes involved in the same regulatory pathway to be represented in the ACS, even when they are never mentioned together in a single article. Further evaluation will be presented.

References
1. Van Mulligen EM, Diwersy M, Schmidt M, Buurman H, Mons B. Facilitating networks of information. Proc AMIA Symp 2000;868-872.
2. Van der Eijk CC, Van Mulligen EM, Van den Berg J. Finding Complementary Scientific Concepts Using a Conceptual Associative Spatial Graph. ISAS-SCI (XIII) 2002;46-50.
3. Van Mulligen EM, Van Der Eijk CC, Kors JA, Schijvenaars BJ, Mons B. Research for research: tools for knowledge discovery and visualization. Proc AMIA Symp 2002;835-839.