LA-UR-04-0545

Title: MESH KEY TERMS FOR VALIDATION AND ANNOTATION OF GENE EXPRESSION CLUSTERS

Author(s): Andreas Rechtsteiner, 174608, CCS-3
Luis M. Rocha, 121983, CCS-3

Submitted to: Eighth Annual International Conference on Research in Computational Molecular Biology

MeSH Key Terms for Validation and Annotation of Gene Expression Clusters

Andreas Rechtsteiner and Luis M Rocha [1]

Keywords: gene expression analysis, validation, information retrieval, automated functional annotation

Integration of different sources of information is a great challenge for the analysis of gene expression data, and for the field of Functional Genomics in general. As the availability of numerical data from high-throughput methods increases, so does the need for technologies that assist in the validation and evaluation of the biological significance of results extracted from these data. In mRNA assaying with microarrays, for example, numerical analysis often attempts to identify clusters of co-expressed genes. The important task to find the biological significance of the results and validate them has so far mostly fallen to the biological expert who had to perform this task manually.

One of the most promising avenues to develop automated and integrative technology for such tasks lies in the application of modern Information Retrieval (IR) and Knowledge Management (KM) algorithms to databases with biomedical publications and data. Examples of databases available for the field are bibliographic databases containing scientific publications (e.g. MEDLINE/PUBMED), databases containing sequence data (e.g. GenBank) and databases of semantic annotations (e.g. the Gene Ontology Consortium and Medical Subject Headings (MeSH)).

We present here an approach that uses the MeSH terms and their concept hierarchies to validate and obtain functional information for gene expression clusters[2]. The controlled and hierarchical MeSH vocabulary is used by the National Library of Medicine (NLM) to index all the articles cited in MEDLINE. Such indexing with a controlled vocabulary eliminates some of the ambiguity due to polysemy (terms that have multiple meanings) and synonymy (multiple terms have similar meaning) that would be encountered if terms would be extracted directly from the articles due to differing article contexts or author preferences and background. Further, the hierarchical organization of the MeSH terms can illustrate the conceptual/functional relationships of genes associated with MeSH terms. MeSH terms can be associated with genes through co-occurrence of these in MEDLINE citations, i.e. the genes occur in titles or abstracts and the MeSH terms are assigned by experts.

To identify MeSH terms associated with a group of genes we used the tool MESHGENE developed at the Information Dynamics Lab at HP Labs (http://www-idl.hpl.hp.com/meshgene/). When presented with a list of human genes, MESHGENE uses some sophisticated techniques to search for these gene symbols in the titles and abstracts of all MEDLINE citations. MeSH terms and the number of co-occurrences can be retrieved[3]. Gene symbols that are aliases of each other are pooled from several databases. This addresses the problem of synonymy, the fact that several symbols can refer to the same gene. MESHGENE employs some sophisticated algorithms that disregards symbols that are likely to be acronyms for other concepts than a gene. This addresses the problem of polysemy, i.e. possible multiple meanings of a gene symbol[4].

We applied our approach to gene expression data from herpes virus infected human fibroblast cells [1]. The data contains 12 time-points, between 1/2 hrs and 48 hrs after infection. Singular Value Decomposition was used to identify the dominant modes of expression. 75% of the variance in the expression data was captured by the first two modes, the first exhibiting a monotonly increasing expression pattern and the second a more transient pattern. Projection of the gene expression vectors onto this

[1] CCS-3, Los Alamos National Laboratory. E-mail: andreas@lanl.gov

[2] Masys et al. [2] presented a proof-of-concept utility related to this approach.

[3] MESHGENE also returns 'scores', a statistical measure of association that attempts to take into account how often one would find a certain gene-MeSH term co-occurrence by chance.

[4] Both of these techniques improve on some of the short-comings of the approach by Masys et al. [2]

first two modes identified 3 statistically significant clusters of co-expressed genes[5]. 500 genes from cluster 1 and 300 genes from clusters 2 and 3 each were uploaded to MESHGENE and the MeSH terms and co-occurrence values were retrieved. MeSH terms were also obtained for 5 groups of randomly selected genes with similar numbers of genes. The log was taken of the co-occurrence values and for each MeSH term these log co-occurrence values were summed for each group over the genes in that group. A matrix with 8 columns for the 8 groups of genes and with 14,000 rows with the MeSH terms was obtained. To analyze this association matrix we used a Latent Semantic Analysis (LSA) approach. We applied SVD to this gene-group vs. MeSH term association matrix. The first 2 modes that capture most of the variation (and therefore most times also information) in the association matrix were highly associated with MeSH terms that occurred uniquely or disproportionally in the 3 gene clusters. MeSH terms highly associated with the 5 groups of randomly selected genes were associated with the lower modes. These modes seem to just capture 'noise' in the association matrix. This result by itself is of great interest for gene expression analysis. We were able to show that the 3 clusters of genes not only separated in 'expression space' but also in the MeSH term space with which they are associated through the literature. Further, our results indicate that the genes within the 3 clusters are not only similar in expression space (i.e. co-expressed) but are also more coherent in MeSH space (and therefore probably also functionally coherent) than the randomly grouped genes. These two observations support, and in some way validate, the clustering results obtained from the expression data.

We also inspected the MeSH terms most associated with the 3 clusters for functional information and compared the results to a manual annotation of clusters 1 and 2 that was done by biological experts[6]. Many MeSH terms highly associated with cluster 1 were related to viruses, indicating that many genes in cluster 1 must have been reported in the literature in connection with these. A disproportionate number of MeSH terms for cluster 1 were also related to oncogenesis, transcription regulation, antigene processing, antibody formation, and Major Histocompatibility Complexes I and II. The manual annotation by the biological experts identified a disproportionate high number of genes for cluster 1 related to all of the above biological processes or cellular components suggested by the MeSH terms. MeSH terms highly associated with cluster 2 were related to immune response, T cells, macrophages, infections, inflammation, cytokines and cytokine receptors. Manual annotation supported the above findings by identifying a disproportionate number of genes in cluster 2 related to these concepts. For cluster 3 highly associated MeSH terms were related to connective tissue and connective tissue diseases (note the herpes infected cells were human fibroblast cells), enzymes involved in apoptosis, immune response and oncogene proteins. (No manual annotation of cluster 3 genes was performed.) The results indicate that the MeSH terms associated with the genes in the 3 clusters are informative about functions of many of the genes in these clusters, and would at least be a good guide for the biological expert trying to investigate the functions of these in more detail.

We are working on further exploration and improvements on the above methodology. We are exploring, for example, different measures of association between MeSH terms and genes. We also explore the usefulness of our approach for other areas of Functional Genomics, like protein function inference.

References

[1] E.P. Browne, B. Wing, D. Coleman, and T. Shenk. Altered cellular mRNA levels in human cytomegalovirus-infected fibroblasts: Viral block to the accumulation of antiviral mrnas. *Journal of Virology*, 75(24):12319–30, 2001.

[2] D.R. Masys, J.B. Welsh, J. Lynn Fink, M. Gribskov, I. Klacansky, and J. Corbeil. Use of keyword hierarchies to interpret gene expression patterns. *Bioinformatics*, 17(4):319–26, 2001.

[5] A manuscript describing this SVD based algorithm is in preparation.

[6] A manuscript of this study is in preparation.