Functional Gene Group Summarization by Clustering MEDLINE Abstract Sentences

Jianji Yang, M.S., Aaron M. Cohen, M.D., M.S., William R. Hersh, M.D.
Oregon Health & Science University, Portland, OR, USA

ABSTRACT: Tools to automatically summarize functional gene group information from the biomedical literature will help genomics researchers both better interpret gene expression data and understand biological pathways. In this study, we built a system that takes in a set of genes and MEDLINE records and outputs clusters of genes along with summaries of each cluster by sentence extraction from MEDLINE abstracts. Our preliminary use-case evaluation shows that this approach can identify gene clusters similar to manually generated groupings.

BACKGROUND: Interpretation of gene expression data is an important task involving both exploratory analysis of the expression profile and comparison against manually compiled knowledge resources. Even though these resources can map genes to functional groups or certain biological pathways, the information provided for each group is usually the Gene Ontology term or pathway names. In order to gather more information on the genes, researchers need to search for them in the immense amount of unstructured biomedical literature. Searching this free text literature to find information on a large set of genes is a time consuming task for human beings. Therefore, a tool to summarize the buried information on gene groups and their interactions in free text will be a valuable tool for scientists. Here we present a basic system for clustering genes into functional groups and use sentences extracted from MEDLINE abstracts as summaries.

METHODS: For higher accuracy in gene name recognition, we focused on mouse genes. The ten-year Medline corpus (from 1994 to 2003) used in the TREC2005 Genomics Track was filtered with the MeSH Heading *Mice*. Using a gene and protein name entity recognition and normalization system we previously developed, gene and protein names in this subset were tagged and identified by unique identifiers. Sentences within the abstract or title containing at least one reference to a gene were extracted and grouped by the gene they contain. The sentence groups for each gene were modeled as word vectors weighted by term frequency after stop-word removal and Porter-stemming, with MESH headings as optional features. Centroid-based clustering was applied on vector similarity calculated as the cosine of the angle between the two gene vectors. Rank of the sentences emphasized the similarity to the centroid, the title sentences, and sentences containing references to more than one gene. The top-ranking sentences in a cluster were extracted as summaries. The authors also manually constructed a use-case from a published research on atherosclerosis to evaluate the system.

RESULTS: The results suggest that this approach may be useful in identifying functional gene groups and provide more in-depth information through summarization of the literature. The gene clusters found were similar to the use-case. For example, the automated system identified the manually created matrix metalloproteinase and inhibitors group, the ATP-binding cassette group, and the cytoplasmic signaling group.

CONCLUSION: Our approach using only plain text and MeSH terms may serve as the first step in producing more detailed summaries for gene groups and relations from the literature. Our preliminary result is encouraging. It can be used to help interpretation of expression data without the labor overhead of curated resources. In addition, the extracted sentence summary can serve as an entry point for additional literature search and review.