

Poster K-12

FLAME: Frequency-Based Lexical Analysis of Microarray Gene Lists



Authors:

Anar K. Khan (*AgResearch Ltd., Invermay Agricultural Centre, Puddle Alley, Mosgiel, Otago, New Zealand*)

Alistair K. Knott (*Computer Science Department, University of Otago, 90 Union Place East, Dunedin, New Zealand*)

Jonathan W. Warren (*AgResearch Ltd., Invermay Agricultural Centre, Puddle Alley, Mosgiel, Otago, New Zealand*)

John C. McEwan (*AgResearch Ltd., Invermay Agricultural Centre, Puddle Alley, Mosgiel, Otago, New Zealand*)

Short Abstract: FLAME is a web-based Java program which provides a portal to the literature surrounding a differentially expressed (DE) gene list derived from a microarray experiment. FLAME leverages term usage information in PubMed abstracts to summarise and cluster a given DE gene list, thereby automatically revealing associations amongst DE genes.

Long Abstract:

Extracting biologically meaningful patterns from microarray datasets remains a difficult challenge. On obtaining a list of differentially expressed (DE) genes from a microarray dataset, the user is confronted with the question: what is the biological significance of this list of genes? Functional interpretation is often performed manually, for example through retrieving relevant papers from the biomedical literature, and filtering, reading and integrating this information. Considering there may be a few hundred DE genes identified in any given experiment, this task is onerous and represents a severe time cost to the researcher. Additionally, integrating this information manually may lead to conclusions which are biased toward researchers' prior expectations. FLAME (an acronym for Frequency-based Analysis of Microarray Experiments) is a web-based Java program which automatically analyses a specified DE gene list by leveraging term frequency information in PubMed abstracts. It takes as input a list of DE genes and a corresponding list of non-differentially expressed (NDE) genes; as output it produces a list of salient terms extracted from abstracts which describe the functions represented in the DE gene list, a clustering of the DE genes into functional subgroups, and a list of functionally informative sentences extracted from abstracts and hyperlinks to those abstracts in PubMed. FLAME is a fully automated, modular system, incorporating the following five modules: Gene Indexing, Term Selection, Gene Clustering, Cluster Annotation and User Interface. The Gene Indexing module identifies and retrieves a set of relevant PubMed abstracts for each gene on the microarray slide. The Term Selection Module extracts content-bearing terms, on which the subsequent clustering and annotation analyses are to be based. The Gene Clustering module uses document clustering techniques to cluster the DE genes on the basis of term frequencies in their associated abstracts, to derive a partition of the genes reflecting functional subgroups within the DE gene list. The Cluster Annotation module summarises each gene cluster, producing a list of terms and sentences which describe each cluster and hyperlinks to pertinent abstracts,

allowing rapid access to the literature describing the DE gene clusters. FLAME has been tested on a number of synthetic and real microarray datasets, using a combination of information retrieval and cluster quality metrics, coupled with manual inspection of results. Preliminary evaluation of FLAME demonstrates its usefulness in revealing interesting functional themes from the literature. FLAME provides a portal to the literature surrounding a given DE gene list, thereby reducing the amount of time invested in searching the literature for relevant associations amongst DE genes.