BioBibliometrics: information retrieval and visualization from co-occurrence of gene names in medline abstracts.

BENJAMIN J. STAPLEY[1] & GERRY BENOIT[2].

[1] Kentucky Center for Structural Biology and Kentucky Center for Computational Science, University of Kentucky, Lexington, KY, USA, 50436-0298.

[2] School of Library and Information Science, University of Kentucky, Lexington, KY.

Email: ben@sophocles.gws.uky.edu

Successful information retrieval from biomedical literature databases is becoming increasingly difficult. In an attempt to relieve the problem, we have developed a prototype system for retrieving and visualizing information from literature and genomic databases using gene names. The premise of our work is that, if two genes have a related biological function, the co-occurrence of two gene names (or aliases of those genes) within the biomedical literature is more likely.

From a document collection of medline records publish between 1997 and 1998 with the MeSH term '*Saccharomyces cerevisiae*', we have extracted the number of co-occurrences of every pair of *cerevisiae* genes. The query is automatically conflated to include SGD gene aliases as well. In addition, the retrieved document set can be filtered by the user with a MeSH term regular expression. From this co-occurrence data we construct a matrix that contains similarity measurements of every pair of genes, based on their joint and individual occurrence statistics. A graph is generated from this matrix, with node and edge inclusion being determined by a user-defined threshold. Nodes of the graph represent genes, while edge lengths are a function of the joint occurrence probability of the two genes within the literature.

Implemented using JAVA and Perl over the WWW, nodes can be hypertext-linked to sequence databases, while edges are linked to those medline documents that generated them. The system is a powerful tool for comprehending literature/genome relationships and for efficiently exploring the biomedical information landscape.

Andrade, M. A. and A. Valencia 1998. Automatic extraction of keywords from scientific text: application to the knowledge domain of protein families. *Bioinformatics 14*:600-7.

Cherry, J. M., C. Adler, et al. 1998. SGD: Saccharomyces Genome Database. *Nucleic Acids Research 26*:73-9.

Schuler, G. D., J. A. Epstein, et al. 1996. Entrez: molecular biology database and retrieval system. *Methods In Enzymology 266*:141-62.