Data and text mining

GenInfoMiner—a web server for exploring biomedical literature using batch sequence ID
Weijian Xuan, Stanley J. Watson and Fan Meng*
Department of Psychiatry and Molecular and Behavioral Neuroscience Institute, University of Michigan, Ann Arbor, MI 48109, USA

ABSTRACT
Summary: GenInfoMiner is a web-based system for searching Medline abstracts using sequence ID lists such as GenBank accession numbers derived from high-throughput experiments. It will map query results to MeSH topics to facilitate the exploration of the biological significance of the sequence ID lists. GenInfoMiner is based on a custom gene and protein name identification engine that can map gene and protein names to important molecular biology databases.

Availability: GenInfoMiner is freely available over the Internet at http://brainarray.mbi.med.umich.edu/GIM.asp

INTRODUCTION
The Medline database is the most utilized database for finding biological functions for gene lists derived from high-throughput studies. However, sequence IDs such as accession numbers or UniGene IDs usually cannot be used in Medline search as only 1.09% of Medline records contain sequence ID annotations and a large portion of them are related to gene cloning rather than functional studies. Furthermore, due to the redundant nature of the GenBank, one gene can usually be represented by several sequences with different sequence IDs deposited by different research groups. This redundancy issue is greatly amplified by expressed sequence tag (EST) sequencing projects for various species. Consequently, dozens, if not hundreds, of sequences may represent the same gene under many circumstances. Nonetheless, even Medline records with accession numbers do not list all possible accession numbers that can represent the same gene. As a result, Medline searches based on existing accession number annotations in Medline records usually miss a significant percentage of functional study papers.

While there are tools that aim at filtering relevant citations, e.g. MedMiner (Tanabe et al., 1999) and PubGene (Jenssen et al., 2001), they do not have the ability to explore literature by using batch sequence ID directly. In order to search Medline using sequence ID lists effectively, we developed the GeneInfoMiner based on our gene/protein name identification engine. Our implementation establishes a translation mechanism between sequence IDs and gene/protein names or abbreviations in the Medline records. Query results are mapped onto Medical Subject Heading (MeSH) terms and grouped under the hierarchical MeSH structure. As a result, users can quickly get an overview of the main functions associated with gene or probe lists derived from high-throughput experiments and easily retrieve the relevant records.

ALGORITHM AND IMPLEMENTATION
One critical part of the system is to identify gene and protein entities in biomedical text. This is a difficult task due to the lack of naming consistency. Previous attempts can be divided into two categories: systems using statistical and machine learning models and systems using hand-craft rules along with knowledge resources. GenInfoMiner incorporates a gene and protein name identification engine developed in our group, which uses a combination of heuristics and statistical models (Xuan et al., 2003). In addition to the identification of gene/protein names, our program can further map extracted gene and protein names to human Entrez Gene (previously LocusLink) IDs. If a gene name stands for multiple identifiers, we use feature context (e.g. gene descriptions and abstracts cited in the Entrez Gene database) for disambiguation. We run our gene name identification program on all Medline citations and index extracted gene entities with mapped Entrez Gene IDs.

Sequence IDs submitted to the GeneInfoMiner, such as GenBank accession numbers and Affymetrix probe IDs, will first be mapped onto human Entrez Gene IDs through our ProbeMatchDB (Wang et al., 2002). The GeneInfoMiner then retrieves all papers containing these Entrez Gene IDs using the precompiled index mentioned before. Sequence IDs that our engine could not find match in Medline are listed in a separate table with the corresponding database link(s). It should be pointed out that the GeneInfoMiner is primarily for human, mouse and rat sequence ID batch queries since both our gene/protein name identification engine and the ProbeMatchDB are targeted for these three species.

Rather than simply providing a linear list of hundreds or even thousands of Medline abstracts, we utilize the MeSH (Coletti and Bleich, 2001) to organize the search results for easier literature exploration. As of this writing, over 98% Medline citations include an average of 11 MeSH terms per citation judged to be most relevant to the main topics of each citation. Our system will count the number of Medline records associated with each MeSH term and sort results according to the number of hits for each MeSH term in descending order. The GeneInfoMiner also allows the display of MeSH term relationship for individual genes by clicking the name of the gene in the result. Consequently, it is very easy for a researcher to identify the most frequent topics associated with a sequence ID list or individual genes. Because MeSH terms are assigned using full-length literature,

*To whom correspondence should be addressed.
Search medline using sequence ID

CONCLUSION

GeneInfoMiner is designed to simplify the task of mapping sequence ID lists derived from genome-wide high-throughput studies to the Medline database. It uses our custom gene and protein name identification engine to find relationship between sequence IDs and gene/protein names in Medline records. In addition, the ability of the GeneInfoMiner to present search results according to their MeSH hit count as well as flexible sorting functions can greatly speed up the retrieval of the relevant literature related to genome-wide high-throughput results.

ACKNOWLEDGEMENTS

The authors are members of the Pritzker Neuropsychiatric Disorders Research Consortium, which is supported by the Pritzker Neuropsychiatric Disorders Research Fund L.L.C. Part of this work is also supported by National Institute on Drug Abuse R21 DA13754-01 to F.M.

Conflict of Interest: none declared.

REFERENCES


