**Bioinformation** by Biomedical Informatics Publishing Group *open access*

www.bioinformation.net **Software**

# GT-Miner: a graph-theoretic data miner, viewer, and model processor

**Douglas E. Brown, Amy J. Powell, Ignazio Carbone and Ralph A. Dean***

Center for Integrated Fungal Research (CIFR), Department of Plant Pathology, Box 7251, North Carolina State University, Raleigh, NC 27695 7251; Ralph A. Dean* - Email: radean2@ncsu.edu; * Corresponding author

received September 24, 2008; accepted October 10, 2008; published December 15, 2008

**Abstract:**
Inexpensive computational power combined with high-throughput experimental platforms has created a wealth of biological information requiring analytical tools and techniques for interpretation. Graph-theoretic concepts and tools have provided an important foundation for information visualization, integration, and analysis of datasets, but they have often been relegated to background analysis tasks. GT-Miner is designed for visual data analysis and mining operations, interacts with other software, including databases, and works with diverse data types. It facilitates a discovery-oriented approach to data mining wherein exploration of alterations of the data and variations of the visualization is encouraged. The user is presented with a basic iterative process, consisting of loading, visualizing, transforming, and then storing the resultant information. Complex analyses are built-up through repeated iterations and user interactions. The iterative process is optimized by automatic layout following transformations and by maintaining a current selection set of interest for elements modified by the transformations. Multiple visualizations are supported including hierarchical, spring, and force-directed self-organizing layouts. Graphs can be transformed with an extensible set of algorithms or manually with an integral visual editor. GT-Miner is intended to allow easier access to visual data mining for the non-expert.

**Keywords**: graph theory; data mining; visualization; information visualization

**Availability:** The GT-Miner program and supplemental materials, including example uses and a user guide, are freely available from http://www.cifr.ncsu.edu/bioinformatics/downloads/.

**Background:**
Contemporary biology faces challenges of analyzing and integrating the ever-accumulating high-throughput datasets to derive a coherent systems-based view of organisms **[1]**. Important challenges include relating genomic, transcription, proteomic, and other data for inference of metabolic and regulatory networks embodying complex processes such as disease phenotypes. Graphs, structures containing nodes and edges linking the nodes, can be used to model biological systems **[2]** wherein entities such as genes, proteins, RNA elements, and metabolites can serve as the nodes and experiment-specific relationships serve as the edges. Attributes, defined properties or additional information associated with the nodes and edges, of the graph form additional dimensions of information. Exploration of a graph's properties and network topology can provide insight into a biological system's architecture and or functioning.

From a systems biology perspective, software applications supporting visualization, exploration, integration, and analysis of disparate datasets are available, such as cytoscape, VisANT, Osprey, PathwayStudio **[3, 4, 5, 6]**; however, they can be economically and computationally expensive, restricted to specific computing platforms, require significant specialist knowledge or have narrow utility, and may be constrained to handle information in specific forms. In the context of visual data mining for bioinformatics, frameworks for discovering and interpreting relationships, characterization of graphs, and graph based visualizations have been developed **[7]**.

**Implementation:**
GT-Miner **[8]** integrates a graphical user interface (GUI), transformational analyses for modifying the graph structure and information content, visualization layout of the graph, direct editing of the graph, and storage access for graph representations of data sets. The program accepts data from text files, applications like Microsoft Excel, or from databases like MySQL and Postgresql. The GUI supports user interactivity and graph visualization through multiple visual layouts, as well as multiple transformations for element filtering, merging of labeled graphs, and cluster analysis.

GT-Miner forms a lightweight, parsimonious framework wherein the graph and its associated attributes is the primary means for coupling information flow between software components. Much of the functionality is implemented as modules focusing on one part of an overall iterative analytical process. Extension with new transformations and layouts is through a simple programming interface, giving direct access to the graph structure and to the Java Swing graphic display, and the extensions incorporate into the framework through run-time configuration files.



ISSN 0973-2063 (online) 0973-8894 (print)
Bioinformation 3(5): 235-237 (2008)

The base program and most of the plug-ins are written in the Java language. Visual layouts in the distributed software are based on GraphViz and in-memory modeling of the graph is based on a modified version of Grappa. Database queries are performed using JDBC, thus enabling access to an unbounded suite of database technologies, and result-table columns are mapped to graph elements by interpretation of the table's meta data. Since data base access is critical for handling large volumes of information, a copy of Apache Derby, a SQL-92 compliant database, is included with the software distribution.

**Utility and caveat:**

Flexibility arises from maintaining a distinction between the visualization and analytical processes. The user can utilize a given visualization and apply multiple transformations or, conversely, utilize multiple visualizations for a given transformation. An unbounded set of attributes can be associated with the graph elements and used with the transforms to modify the graph's structure or visual presentation. Modifications can be saved for incorporation into additional analytical processes. Combining attribute based transformations with the programs' built-in support for visual editing of the graph through simple mouse gestures can greatly facilitate the discovery process (see Figure 1).

Bioinformatics source data is often represented in a variety of potentially incompatible formats requiring a burdensome reformatting of the information into an acceptable form. Our solution partially addresses the problem by decoupling the acquisition and preparation of data from the analytical data mining and visualization processes through two approaches, both external to the application, for loading information: 1) support for common graph file formats like DOT, PHYLIP NEWICK, or GXL; and 2) acquiring the information in tabular formatted adjacency-lists describing the nodes, edge relationship, and attributes. This allows the user to convert raw data, typically via a SQL selection expression, into a graph format without the need for extending the application program. Consequently, the information can originate from specialized applications such as phylogenetic analysis programs, or more general sources like databases and spreadsheets. The final result is saved in the above file formats or in a database. The distribution includes a user guide and three complete tutorials covering phylogenetic ancestral recombination graphs **[9]**, networks of gene duplications **[10]**, and visualization of Gene Ontology annotations.

**Acknowledgment:**

Research results supported by North Carolina State University, the University of North Carolina Office of the President, the National Science Foundation, the National Institutes of Health, and the National Research Initiative of the USDA Cooperative State Research, Education and Extension Service.

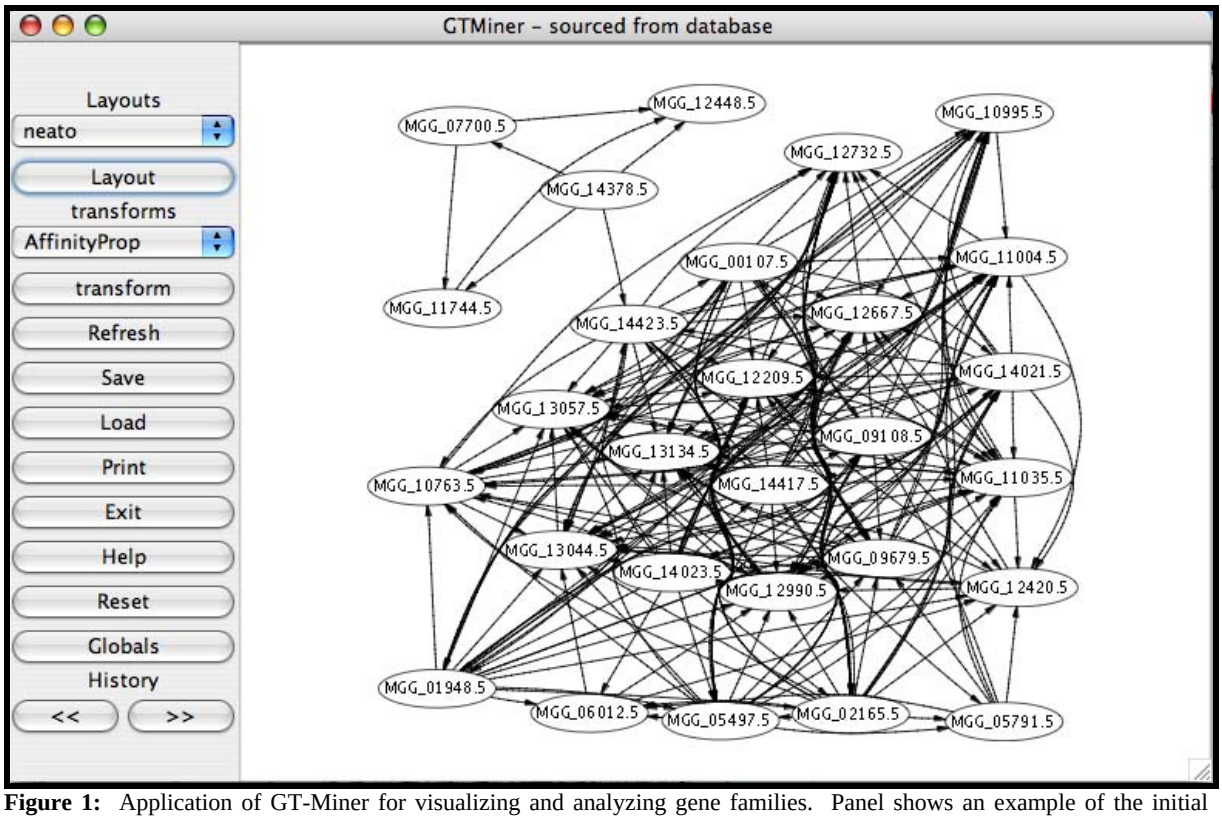


**Figure 1:** Application of GT-Miner for visualizing and analyzing gene families. Panel shows an example of the initial relationships for a family of homologous genes from the plant pathogenic fungus *Magnaporthe grisea*, strain 70-15, determined using the NCBI blastp program. Iterative analysis of the gene family using GT-Miner rapidly reveals that the linkage between MGG_14378.5 and MGG_14423.5 may be erroneously linking two different families.



ISSN 0973-2063 (online) 0973-8894 (print)
Bioinformation 3(5): 235-237 (2008)

**References:**

**[01]** H. Ge *et al*., *Trends in Genet*., 19: 551 (2003) [PMID: 14550629]

**[02]** T. Aittokallio and B. Schwikowski, *Brief Bioinform*., 7: 243 (2006) [PMID: 16880171]

**[03]** P. Shannon *et al*., *Genome Res*., 13: 2498 (2003) [PMID: 14597658]

**[04]** Z. Hu *et al*., *Nucleic Acids Res*., 33: W352 (2005) [PMID: 15980487]

**[05]** B. Breitkreutz, *Genome Biol*. (2003) 4: R22 [PMID: 12620107]

**[06]** Nikitin *et al*., *Bioinformatics*, 19: 2155 (2003) [PMID: 14594725]

**[07]** D. Keim, *IEEE Trans Vis Comput Graph*, 08: 1 (2002)

**[08]** http://www.cifr.ncsu.edu/bioinformatics/downloads

**[09]** I. Carbone *et al*., *Mol Ecol*., 16: 4401 (2007) [PMID: 17725568]

**[10]** A. Powell *et al*., *BMC Genomics*, 9: 147 (2008) [PMID:18373860]

**Edited by P. Kangueane**

**Citation: Brown *et al.*,** Bioinformation 3(5): 235-237 (2008)