

BioArrayMine: A software package for integrative analysis of cross-platform and cross-species microarray data

Fei Pan^{1*}, Kiran Kamath^{1*}, Haiyan Hu¹, Yu Huang¹, Kangyu Zhang¹, Min Xu¹, Xifeng Yan¹, Jiawei Han², and X. Jasmine Zhou^{1#}

¹Program in Molecular and Computational Biology, University of Southern California, Los Angeles, USA

²Department of Computer Science, Univ. of Illinois at Urbana-Champaign, Urbana, IL 61801, USA

*These two authors contributed equally to this work.

#Contact: X. Jasmine Zhou, xjzhou@usc.edu

Microarray gene expression profiling is performed in many laboratories, resulting in the rapid data accumulation in public repositories. However, due to the existence of different microarray platforms and the lack of standard experimental protocols, systematic variation among data sets often exceeds the capability of statistical normalization. Currently, there is an urgent need for methods and tools to integrate the enormous amount of microarray data. Recently, we have designed several methods to address this need^{1,2}. Here, we demonstrate our newly developed software package *BioArrayMine*.

BioArrayMine is data mining and visualization software for the integrative analysis of many cross-platform microarray data sets. We employ a meta-analysis approach to derive expression patterns from each microarray data set, and then discover those patterns frequent occurring across multiple data sets. Typical expression patterns include: co-expression networks and differentially expressed gene lists. Due to the noisy nature of microarray data, identifying recurrent signals across multiple datasets can enhance signal/noise separation, and allows us to draw biological inference with higher confidence. *BioArrayMine* can also be used to identify conserved expression patterns across different species.

We will demonstrate the following utility modules of *BioArrayMine*:

- ***Data preprocessing module***: includes user-specified gene selection criteria to narrow down the list of genes to study. For example, a user may select only those genes that show significant variation in at least m out of the total n data sets. Such strategies are often used by standard microarray analysis software such as the clustering programs. Genes of different data sets will be linked via their Unigene IDs.
- ***Data Analysis module***: This includes two analysis approaches: (1) *Co-expression analysis*: we use a graph theoretical approach to identify a set of genes co-expressed across multiple data sets¹. (2) *Differential expression analysis*: We use a frequent pattern mining algorithm to identify a set of genes differentially expressed m out of n given data sets.
- ***Functional analysis***: Given a recurrent expression pattern (a co-expression subnetwork or a list of differentially expressed genes), we can determine the over-represented GeneOntology functional categories and the potential transcription regulators.

In order to allow biologists to assimilate and explore the data and analyze results in an intuitive manner, an interactive graph visualization module is developed as an important part of *BioArrayMine*. *BioArrayMine* provides various options to handle graph files, including loading/saving graph from or into external text files, expanding a subgraph from any node, and our previously developed method of transitive functional annotation based on shortest path analysis³.

BioArrayMine is implemented using both Java and C++ that combine the cross-platform capability of Java and efficiency of C++ so that *BioArrayMine* is able to run on both Windows and Unix System efficiently. *BioArrayMine* is publicly freely downloadable for academic usage at <http://zhoulab.usc.edu>. We expect that *BioArrayMine* will significantly facilitate the re-use of the vast amount of existing microarray data, reduce the necessity to generate new data, and improve our understanding of cellular functions and networks under a variety of perturbations.

Reference

1. Hu, H., Yan, X., Huang, Y., Han, J. & Zhou, X. Mining coherent dense subgraphs across massive biological networks for functional discovery. *ISMB 2005, accepted* (2005).
2. Zhou, X.J. et al. Functional annotation and network reconstruction through cross-platform integration of microarray data. *Nat Biotechnol* **23**, 238-243 (2005).
3. Zhou, X., Kao, M.C. & Wong, W.H. Transitive functional annotation by shortest-path analysis of gene expression data. *Proc Natl Acad Sci U S A* **99**, 12783-12788 (2002).