BIOINFORMATICS APPLICATIONS NOTE Vol. 22 no. 13 2006, pages 1665–1667
doi:10.1093/bioinformatics/btl163

Gene expression

Integrative Array Analyzer: a software package for analysis of cross-platform and cross-species microarray data

Fei Pan[1,†], Kiran Kamath[1,†], Kangyu Zhang[1], Sudip Pulapura[1], Avinash Achar[1], Juan Nunez-Iglesias[1], Yu Huang[1], Xifeng Yan[2], Jiawei Han[2], Haiyan Hu[1], Min Xu[1], Jianjun Hu[1] and Xianghong Jasmine Zhou[1,*]

[1]Program in Molecular and Computational Biology, University of Southern California, Los Angeles, USA and [2]Department of Computer Science, University of Illinois at Urbana-Champaign, Urbana, IL 61801, USA

Received on February 16, 2006; revised on March 23, 2006; accepted on April 24, 2006
Advance Access publication May 3, 2006
Associate Editor: Joaquin Dopazo

ABSTRACT

Summary: The rapid accumulation of microarray data translates into an urgent need for tools to perform integrative microarray analysis. Integrative Array Analyzer is a comprehensive analysis and visualization software toolkit, which aims to facilitate the reuse of the large amount of cross-platform and cross-species microarray data. It is composed of the data preprocess module, the co-expression analysis module, the differential expression analysis module, the functional and transcriptional annotation module and the graph visualization module.

Availability: http://zhoulab.usc.edu

Contact: xjzhou@usc.edu

1 INTRODUCTION

Microarray gene expression profiling has been conducted in many laboratories, resulting in a rapid accumulation of data in public repositories. Although there are many advantages to combine microarray datasets for integrative analysis, it is not a trivial task to integrate cross-platform microarray datasets. The existence of different microarray technology and alternative experimental parameters (e.g. use of direct or indirect labeling, choice of controls, choice of different scanner and image analysis software) results in systematic variations among datasets that are often beyond the capability of statistical normalization. Recently, several studies have begun to address these issues and have proposed statistical and computational methods to assess expression patterns concordant among several microarray datasets (Choi *et al*., 2003; Hu *et al*., 2005; Lamb *et al*., 2003; Lee *et al*., 2004; Rhodes *et al*., 2002; Segal *et al*., 2004; Zhou *et al*., 2005). Yet, there lack software tools for biologists to perform integrative analysis of multiple microarray datasets. Here, we report the first comprehensive analysis software toolkit, Integrative Array Analyzer (‘*iArray*’ in short), for platform-independent integration of microarray datasets.

iArray employs a meta-analysis approach to first derive expression patterns from individual microarray dataset and then discovers those patterns frequently occurring across multiple datasets. Identifying recurrent expression patterns not only significantly enhances signal/noise separation, but also provides the context information for expression signals, e.g. under which conditions/datasets the expression patterns are activated. Furthermore, *iArray* can be used to identify conserved expression patterns across different species.

2 CORE FUNCTIONS AND FEATURES

iArray includes the data preprocessing module, the co-expression analysis module, the differential expression analysis module, the functional and transcriptional annotation module and the graphical visualization module.

2.1 Data preprocessing module

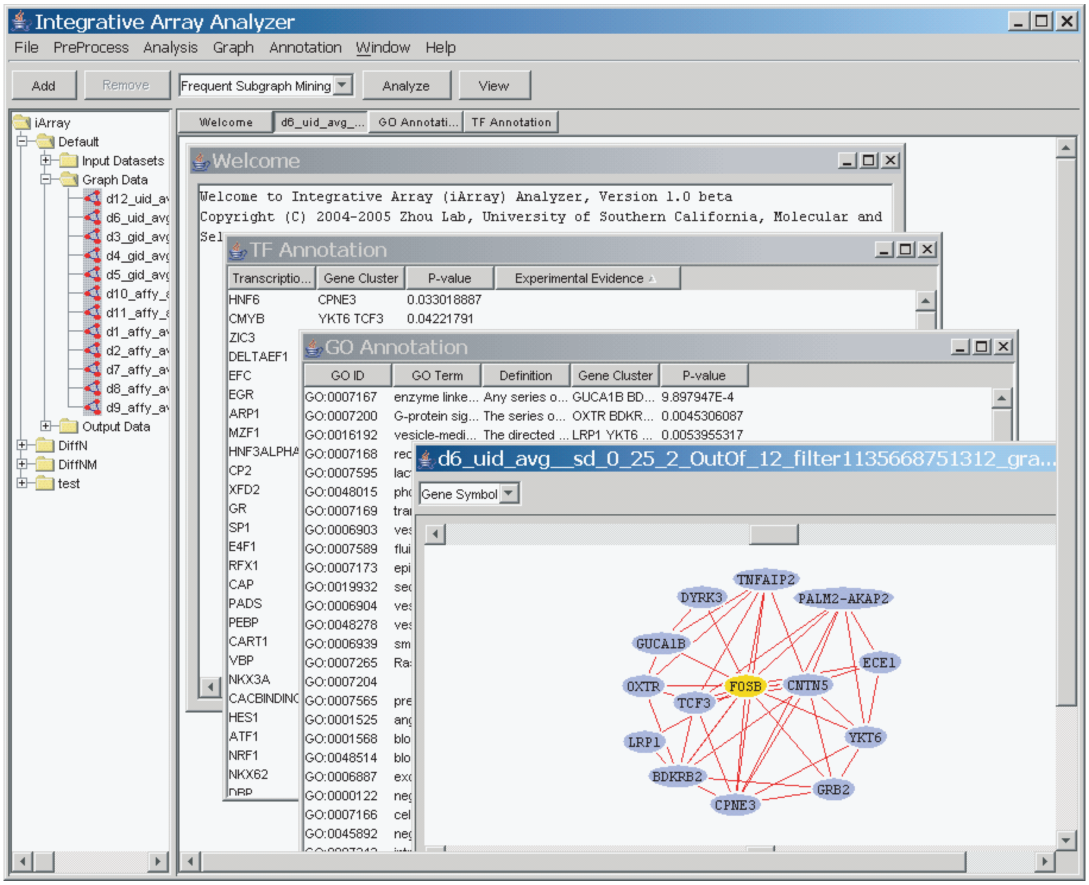
iArray can accept microarray expression datasets from any platforms as input, as long as the data have been summarized into a matrix of normalized expression values. The input files of expression data matrices should be delimited text files, with tab, comma or space as separators. Genes on the different array platforms can be linked via their Unigene ID, and genes of different organisms can be linked based on their homologous relationships using the NCBI HomoloGene database. Users may select different filtering criteria to narrow down the gene lists to only those genes, which are presented in most samples and/or demonstrate large variations across samples.

2.2 Co-expression analysis module

This module is used to derive sets of genes simultaneously co-expressed in multiple datasets. We employ a graph-theoretical approach to perform integrative analyses on multiple microarray datasets. First, we model a microarray dataset as an unweighted and undirected graph. In those graphs, each gene is represented by one node, and if two genes show expression correlation higher than a

[*] To whom correspondence should be addressed.

[†] The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.



```
Welcome to Integrative Array (iArray) Analyzer, Version 1.0 beta
Copyright (C) 2004-2005 Zhou Lab, University of Southern California, Molecular and
Sel
```

Transcriptio...	Gene Cluster	P-value	Experimental Evidence
HNF6	CPNE3	0.033018887	
CMYB	YKT6 TCF3	0.04221791	
ZIC3			
DELTAEF1			
EFC			
EGR			
ARP1			
MZF1			
HNF3ALPHA			
CP2			
XFD2			
GR			
SP1			
E4F1			
RFX1			
CAP			
PADS			
PEBP			
CART1			
VBP			
NKX3A			
CACBINDINC			
HES1			
ATF1			
NRF1			
NKX62			
DBP			

GO ID	GO Term	Definition	Gene Cluster	P-value
GO:0007167	enzyme linke...	Any series o...	GUCA1B BD...	9.897947E-4
GO:0007200	G-protein sig...	The series o...	OXTR BDKR...	0.0045306087
GO:0016192	vesicle-medi...	The directed ...	LRP1 YKT6 ...	0.0053955317
GO:0007168	rec			
GO:0007595	lac			
GO:0048015	pho			
GO:0007169	trai			
GO:0006903	ves			
GO:0007589	flui			
GO:0007173	epi			
GO:0019932	sec			
GO:0006904	ves			
GO:0048278	ves			
GO:0006939	sm			
GO:0007265	Ra			
GO:0007204				
GO:0007565	pre			
GO:0001525	ang			
GO:0001568	blo			
GO:0048514	blo			
GO:0006887	exc			
GO:0000122	neg			
GO:0007166	cel			
GO:0045892	neg			

Fig. 1. Graph visualization of a co-expression module from *iArray*.

given threshold, they are connected with an edge. Given *k* microarray datasets measuring the expression of the same genes, we construct *k* co-expression graphs with the same nodes but different topologies, which will be subjected to comparative network analysis, e.g. searching for frequent subgraph patterns. A subgraph that occurs repeatedly in a sufficient number of datasets is likely to be associated with biological significance in terms of functional modules or biological pathways. Starting from each of the frequent subgraphs, *iArray* can further search for its densely connected subgraphs by applying a recursive min-cut partitioning algorithm, until the subgraphs satisfy the min-cut threshold or reach the user-specified density criterion. Because edges in our graph represent high co-expression, a densely connected subgraph corresponds to a tight co-expression cluster, and a frequent dense subgraph corresponds to a frequent tight co-expression cluster.

2.3 Differential expression analysis module

This module is developed to derive sets of genes frequently differentially expressed in multiple microarray datasets, which employs similar experimental designs, e.g. all on comparison between cancer and normal tissues. The analysis module includes two steps: (1) performing differential analysis for each individual dataset (with Bonferroni or false discovery rate adjustment for multiple comparisons) and (2) identify sets of genes frequently differentially expressed in multiple datasets from results obtained in Step (1) For Step (1), two statistical methods to identify differentially expressed genes are implemented: Student's *t*-test and Mann–Whitney test. For Step (2), we will use the frequent itemset mining algorithm to identify gene sets which occur in at least *n* out of the total *m* differentially expressed gene lists. Our approach will be sensitive to signals that occur only in a subset of the datasets.

2.4 Functional and transcriptional annotation module

Gene network patterns derived from co-expression analysis or gene sets from differential expression analysis can be subjected to functional and transcriptional annotation module. In the functional annotation module, *iArray* uses hypergeometric distribution to assess the statistical significance of the enrichment of genes from particular functional categories or pathways. Users can choose to check the gene enrichment from GeneOntology categories (Ashburner *et al*., 2000), BioCarta pathway annotations (Galperin, 2004) or KEGG pathways annotations (Kanehisa, 1997).

Transcriptional annotation module is developed to predict potential transcription regulators for recurrent co-expression clusters or differentially expressed gene sets. For all genes in the six model organisms including human, mouse, rat, fruit fly, nematode and yeast, we obtained the 1 kb upstream sequence as putative promoter sequences from public genome resources and then screened those sequences for transcription factor binding sites by position-weighted scoring matrices obtained from the TRANSFAC database (Heinemeyer *et al*., 1999).

2.5 Graphic visualization module

All data and results including expression values, co-expression network, differentially expressed gene sets, etc. can be visualized with the Visualization module. It has an interactive graph interface to allow biologists to explore the data and results in an intuitive manner (Fig. 1). Network data can also be exported to other software such as Cytoscape (Shannon *et al*., 2003) for further analysis.

3 CONCLUSION

We presented a software package, Integrative Array Analyzer (*iArray*), for integrative analysis of cross-platform microarray datasets. *iArray* aims at facilitating the reuse of the vast amount of public microarray datasets, reducing the necessity to generate new data and enhancing our understanding of cellular functions under a variety of conditions. Not only do we provide a set of novel approaches and tools for the analysis of multiple microarray datasets, but also we integrate various types of existing knowledge database to facilitate the interpretation of the analysis results.

The capacity of the *iArray* depends on the size of the input datasets and memory/speed of users' computers. The memory and CPU requirements of *iArray* are equivalent to those of hierarchical clustering, since the maximum memory and time-consuming functionality in *iArray* is the generation of correlation matrix in order to construct the co-expression graphs.

In the future, we will incorporate more meta-analysis methods reported in the literature and further improve the functional and transcriptional analysis modules.

ACKNOWLEDGEMENTS

Funding to pay the Open Access publication charges for this article was provided by the NIH grant 1R01GM074163-01A1, the NSF grant 0515936, a pilot grant from the Seaver foundation, and a faculty setup grant from USC.

Conflict of Interest: none declared.

REFERENCES

Ashburner,M. *et al*. (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet*., **25**, 25–29.

Choi,J.K. *et al*. (2003) Combining multiple microarray studies and modeling interstudy variation. *Bioinformatics*, **19** (Suppl. 1), i84–i90.

Galperin,M.Y. (2004) The Molecular Biology Database Collection: 2004 update. *Nucleic Acids Res*., **32**, D3–D22.

Heinemeyer,T. *et al*. (1999) Expanding the TRANSFAC database towards an expert system of regulatory molecular mechanisms. *Nucleic Acids Res*., **27**, 318–322.

Hu,H. *et al*. (2005) Mining coherent dense subgraphs across massive biological networks for functional discovery. *Bioinformatics*, **21** (Suppl. 1), i213–i221.

Kanehisa,M. (1997) A database for post-genome analysis. *Trends Genet*., **13**, 375–376.

Lamb,J. *et al*. (2003) A mechanism of cyclin D1 action encoded in the patterns of gene expression in human cancer. *Cell*, **114**, 323–334.

Lee,H.K. *et al*. (2004) Coexpression analysis of human genes across many microarray datasets. *Genome Res*., **14**, 1085–1094.

Rhodes,D.R. *et al*. (2002) Meta-analysis of microarrays: interstudy validation of gene expression profiles reveals pathway dysregulation in prostate cancer. *Cancer Res*., **62**, 4427–4433.

Segal,E. *et al*. (2004) A module map showing conditional activity of expression modules in cancer. *Nat. Genet*., **36**, 1090–1098.

Shannon,P. *et al*. (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res*., **13**, 2498–2504.

Zhou,X.J. *et al*. (2005) Functional annotation and network reconstruction through cross-platform integration of microarray data. *Nat. Biotechnol*., **23**, 238–243.