BIOINFORMATICS APPLICATIONS NOTE
Vol. 26 no. 2 2010, pages 285–286
doi:10.1093/bioinformatics/btp656

Systems biology

EGAN: exploratory gene association networks

Jesse Paquette* and Taku Tokuyasu

Helen Diller Family Comprehensive Cancer Center, University of California, San Francisco, CA 94143, USA

Received on September 4, 2009; revised on November 4, 2009; accepted on November 19, 2009

Advance Access publication November 23, 2009

Associate Editor: Martin Bishop

ABSTRACT

Summary: Exploratory Gene Association Networks (EGAN) is a Java desktop application that provides a point-and-click environment for contextual graph visualization of high-throughput assay results. By loading the entire network of genes, pathways, interactions, annotation terms and literature references directly into memory, EGAN allows a biologist to repeatedly query and interpret multiple experimental results without incurring additional delays for data download/integration. Other compelling features of EGAN include: support for diverse -omics technologies, a simple and interactive graph display, sortable/searchable data tables, links to external web resources including ≥240 000 articles at PubMed, hypergeometric and GSEA-like enrichment statistics, pipeline-compatible automation via scripting and the ability to completely customize and/or supplement the network with new/proprietary data.

Availability: Runs on most operating systems via Java; downloadable from http://akt.ucsf.edu/EGAN/

Contact: jesse.paquette@cc.ucsf.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 INTRODUCTION

Integration of existing knowledge about genes is a key facilitator for interpretation of results from exploratory -omics assays. But the benefits of knowledge integration do not come from flooding the user with the information; they come when there is an interface present to effectively funnel the data patterns and biological context from the assay into the mind of a biologist. Graph visualization methods employed by software such as Cytoscape (Shannon *et al*., 2003), GenMAPP (Dahlquist *et al*., 2002), Ingenuity IPA (http://www.ingenuity.com), PubGene (Jenssen *et al*., 2001) and VisANT (Hu *et al*., 2007) are becoming increasingly useful in this capacity, helping the user to understand how significant genes in an assay are interconnected by pathways, pairwise relationships, annotation terms and literature.

Performing knowledge integration for graph-based visualization on the -omics scale is not trivial; it requires a combination of programming, software use and processing time. A biologist with limited programming experience is entirely dependent on the availability and efficiency of collaborating statisticians/programmers and data integration software. Also, due to the immense volume of publicly available gene-centric information, collaborating statisticians/programmers and software applications (Cline *et al*., 2007) typically integrate only the subnetwork of knowledge surrounding a specific subset of genes chosen by the biologist (e.g. 'genes passing a false discovery rate cutoff' or 'genes in cluster x'). This process can be defined as *query-dependent* knowledge integration; only those data related to the biologist's gene set are included.

Query-dependent knowledge integration has a major drawback. Each time the biologist formulates a new query about a different set of genes (e.g. 'genes in cluster y'), a non-negligible amount of time is required to download/integrate the new information. In situations where the biologist is not using a software tool directly, but is instead collaborating with a multi-tasking statistician/programmer, the limitations to this paradigm become critical. The biologist might have to wait days or even weeks to receive a network image in response to a single query set of genes.

Computational biologists are continuously developing faster and better algorithms to analyze bioinformation. However, the rate-limiting step in the process of discovery from exploratory analysis is rarely the computational algorithm but rather the communication of algorithm results to a biologist. EGAN allows a biologist to loosen this bottleneck by providing a full precollated knowledge network onto which the results of high-throughput analyses can be mapped. Subnetworks of interesting gene sets are easily queried, displayed, adjusted, combined, compared, output to PDF, saved and retrieved; all via mouse clicks. EGAN enables a powerful alternative paradigm for exploratory assay analysis: one where a biologist can quickly and independently ask, answer and generate questions about the biology-driven patterns in their own data.

EGAN was not developed to be a one-stop analysis shop for all high-throughput experimental platforms. It is an analysis module to be placed downstream of statistical analysis: *t*-tests, ANOVA, clustering, classification, etc. EGAN takes the results of those and other analyses as input and maps them onto its internal gene association network. This allows EGAN to work with results from multiple -omics platforms; so long as the entities measured (peptides, SNPs, transcripts, etc.) can be mapped to genes.

EGAN is also a downstream module for the knowledge integration process. The knowledge network in EGAN is stored and loaded as a set of human- and computer-readable text files; a user can specify network data in EGAN using the same simple wizard that is used to load experimental result files. Alternately, an EGAN configuration file specifying all data (network and experiments) could be automatically generated by another software program, allowing the EGAN to seamlessly integrate with data repositories, automated analysis pipelines and/or other software.

*To whom correspondence should be addressed.

© The Author 2009. Published by Oxford University Press. All rights reserved. For Permissions, please email: journals.permissions@oxfordjournals.org

2 USING EGAN

EGAN has been utilized by research groups at the Helen Diller Family Comprehensive Cancer Center at UCSF to visualize and combine results from expression microarrays, aCGH, MS/MS proteomics and genome-wide association studies. EGAN can also be used to analyze results from ChIP-chip experiments, DNA methylation assays, high-throughput sequencing and computational algorithms including sequence comparison (transcription factor/miRNA target predictions, protein interaction predictions, homology, etc.) and natural language processing of literature.

The following subsections describe some general steps that are commonly performed while using EGAN to interpret experiment results. The term *association node* refers to a predefined set of genes in the network. Pathways, annotation terms, transcription factor targets, miRNA targets, conserved domains, clusters, modules, i.e. anything that can represent a group of genes, is represented in EGAN as an association node. In EGAN there are only two types of graph information in the network: pairwise relationship edges between gene nodes (e.g. protein–protein interactions) and association nodes—also commonly referred to as meta-nodes (Hu *et al*., 2007) or hyper-edges (Berge, 1985).

2.1 Launch EGAN

EGAN is launched from a webpage; the user clicks on a hyperlink to a .jnlp file, and Java WebStart (http://java.com) uses the information in the .jnlp file to launch EGAN. Experiments and network data can be prespecified in a configuration file and/or the user can specify additional data via the EGAN data loading wizard. As soon as the experiments are loaded, the EGAN window appears and the user is ready to begin; all other data are loaded in the background.

2.2 Select gene nodes of interest

There are three common methods for selecting interesting genes in EGAN. The user can sort the gene table by one or more experiment result columns (say, by p-value) then select the top gene rows. The user can alternatively use the "Multi-select…" dialog to perform a query across multiple columns (e.g. all genes with up-regulated expression and frequently amplified copy number). Finally, the user can locate an association node of interest (e.g. predicted targets of the transcription factor AP-1) and select all of its gene neighbors. It is recommended that the EGAN user immediately save their selected set of genes by creating a custom association node.

2.3 Show selected gene nodes on the graph

Once the genes of interest are selected, the user can display them as nodes on the graph and arrange them with a layout algorithm (EGAN employs Cytoscape libraries for graph rendering/layout). Every gene node in EGAN contains summary information from Entrez Gene (Maglott *et al*., 2005) and a link to its Entrez Gene web page as well as its chromosomal context in the UCSC Genome Browser (Kent *et al.*, 2002). Gene node borders are colored and sized according to the statistics and *P*-values from a selected experiment, and pairwise relationships between visible genes are automatically shown as edges. Many pairwise relationship edges are supported with literature references; those edges contain links to the corresponding articles in PubMed (Wheeler *et al*., 2000).

2.4 Activate enrichment calculations for association nodes

Given a set of interesting genes, EGAN can go beyond pairwise interactions and also visualize the association nodes (i.e. other gene sets) that are overrepresented (or enriched) in the set. EGAN employs a standard one-tailed Fisher exact (hypergeometric) test for enrichment calculations given a set of visible genes.

2.5 Show association nodes of interest

With calculated enrichment scores for all association nodes, the user then browses through the different types of association node tables (see Supplementary Table 1 for precollated types), sorts them by the 'Visible enrichment' column and selectively shows enriched association nodes on the graph. Because there are often more enriched association nodes than are comfortably visualized on each graph, discriminate selection of association nodes for display by the biologist is a critical feature of EGAN.

2.6 Explore network and export image

Many edges between association and gene nodes also contain supporting literature references. The user can link to these articles in PubMed as well as create an HTML report file that displays a screenshot of the graph with a list of reference links. The standard format for exporting the network image is PDF, which is written in scalable vector format so subsequent viewers are able to zoom in and out without distorting the graph image.

2.7 Repeat

Finally, the most compelling feature of EGAN: the ability to allow the user to quickly wipe the slate and restart the process. The user just needs to think of a different set of interesting genes: genes from another cluster, a sub- or super-set of the previous query, kinases that are also AP-1 targets, etc.

ACKNOWLEDGEMENTS

The authors would like to thank Donna Albertson, Mike Baldwin, Debo Das, Debra Goldberg, Scot Federman, Stephan Gysin, Graeme Hodgson, Ajay Jain, Katerina Kechris, Ben Kopman, Alan Kuchinsky, Dennis Lezotte, Mau-ting Lin, Scooter Morris, Adam Olshen, Alex Pico, Dan Pinkel, David Quigley, Ingrid Revet, Ritu Roydasgupta, Joachim Silber, Mark Turner and Sook Wah Yee.

Funding: National Institutes of Health (grant number P30 CA92103).

Conflict of Interest: none declared.

REFERENCES

Berge,C. (1985) Graphs and Hypergraphs Elsevier Science Ltd. The Blvd., Langford Ln., Kidlington, Oxford, OX5 1GB, United Kingdom.

Cline,M.S. *et al*. (2007) Integration of biological networks and gene expression data using Cytoscape. *Nat. Protocols*, **2**, 2366–2382.

Dahlquist,K.D. *et al*. (2002) GenMAPP, a new tool for viewing and analyzing microarray data on biological pathways. *Nat Genet.*, **31**, 19–20.

Hu,Z. *et al*. (2007) VisANT 3.0: new modules for pathway visualization, editing, prediction and construction. *Nucleic Acids Res.*, **35**, W625–W632.

Jenssen,T. *et al*. (2001) A literature network of human genes for high-throughput analysis of gene expression. *Nat Genet.*, **28**, 21–28.

Kent,W.J. *et al*. (2002) The Human Genome Browser at UCSC. *Genome Res.*, **12**, 996–1006.

Maglott,D. *et al*. (2005) Entrez Gene: gene-centered information at NCBI. *Nucleic Acids Res.*, **33**, D54–D58.

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

Wheeler,D.L. *et al*. (2000) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.*, **28**, 10–14.