W466–W471 *Nucleic Acids Research, 2006, Vol. 34, Web Server issue*
doi:10.1093/nar/gkl308

BiologicalNetworks: visualization and analysis tool for systems biology

Michael Baitaluk[1],*, Mayya Sedova[2], Animesh Ray[3] and Amarnath Gupta[1]

[1]San Diego Supercomputer Center, University of California San Diego, 9500 Gilman Drive, La Jolla, CA 92093, USA, [2]7564 Charmant Drive, #1814, La Jolla, CA 92122, USA and [3]Keck Graduate Institute, 535 Watson Drive, Claremont, CA 91711, USA

Received February 15, 2006; Revised March 2, 2006; Accepted March 11, 2006

ABSTRACT

Systems level investigation of genomic scale information requires the development of truly integrated databases dealing with heterogeneous data, which can be queried for simple properties of genes or other database objects as well as for complex network level properties, for the analysis and modelling of complex biological processes. Towards that goal, we recently constructed PathSys, a data integration platform for systems biology, which provides dynamic integration over a diverse set of databases [Baitaluk *et al.* (2006) *BMC Bioinformatics* 7, 55]. Here we describe a server, BiologicalNetworks, which provides visualization, analysis services and an information management framework over PathSys. The server allows easy retrieval, construction and visualization of complex biological networks, including genome-scale integrated networks of protein–protein, protein–DNA and genetic interactions. Most importantly, BiologicalNetworks addresses the need for systematic presentation and analysis of high-throughput expression data by mapping and analysis of expression profiles of genes or proteins simultaneously on to regulatory, metabolic and cellular networks. BiologicalNetworks Server is available at http://brak.sdsc.edu/pub/BiologicalNetworks.

INTRODUCTION

Networks of molecular interactions are widely studied to reveal the complex roles played by genes, gene products and the cellular environments in biological processes. In these networks (or graphs), the nodes represent genes or gene products and the edges represent specific interactions. In a protein–DNA network, an edge may represent the binding of a transcription factor to a promoter region, while in a protein–protein physical interaction network it might represent a recorded evidence of co-immunoprecipitation or a two-hybrid interaction. The nodes of the network are typically associated with additional information about the genes (or gene products), such as positions in the chromosome (or localization sites), or their Gene Ontology (GO) classification. A number of specialized and publicly accessible databases are available, which contain data about the nodes [the SGD database (1)] and the interactions [BIND database (2)], and some databases contain information about both [such as KEGG (3)]. In addition, individual researchers often publish their data as part of their publications or project web sites. Currently, a number of analysis and visualization tools have been developed by different groups for assimilating, visualizing and for the analysis and modeling of these molecular interaction network data, of which the most notable are Cytoscape (4), Osprey (5), PathwayAssist (6), Pathways Database System (7), GeneGO (www.genego.com), VisANT (8,9). Yet another direction of effort is the storage and analysis of large-scale gene and protein expression data (10–14). While some of these store and display expression data, they do not allow query or analysis of such data or integration of novel gene expression data with existing network models.

Ideally, an analysis tool for molecular interaction networks should enable a user to import, efficiently store, effectively retrieve and perform analysis on single genes, gene families, patterns of molecular interactions, as well as on the global structure of the network. The tool needs to be sufficiently flexible for both micro-scale and macro-scale analysis using heterogenous data, and extract the data from a large number of disparate databases; it should allow one to construct interaction networks by curation as well as computation [e.g. using algorithms that convert a time-series microarray dataset into an influence network (15)]; it should enable the users to retrieve different interaction graphs through on-demand queries and construct new graphs by assembling them in a variety of ways. It should also allow the incorporation of novel datasets locally, such as the user's own microarray

*To whom correspondence should be addressed. Tel: +1 858 8225418; Fax: +1 858 5345152; Email: baitaluk@sdsc.edu expression data, and/or overlay these on biological networks to explore novel relationships among genes. These critical needs are the minimal requirements of a systems level analysis of biological pathways.

Previously we reported a general-purpose scalable warehouse of biological information, PathSys (16). PathSys is a comprehensive data warehouse resulting from the integration of molecular interaction data with other graph-structured data, such as ontologies, e.g. GO (17) and taxonomies, e.g. enzyme classification system and functional classification of yeast proteins (3), and state data such as gene expression profiles, from over 20 curated and publicly contributed data sources, biological experimental and PubMed data for the eight representative organisms (*Saccharomyces cerevisiae Drosophila melanogaster, etc.* for full list see website). It contains more than 100 000 events of regulation, interaction and modification among genes, proteins, cell processes and small molecules. Here we present BiologicalNetworks, the web-based query tool built on top of PathSys, and show how it enables a user to derive novel biological insights at the single gene level and functional relationships at the systems level.

Data integration

PathSys's data integration model achieves the following:

(i) Integration of object and property types from over 20 databases (for a full list see the website), thus creating a controlled vocabulary (ontology) of object/attribute types.
(ii) Integration of nomenclature for genes/proteins. Naming conventions between different datasets can be different, and the server-side parser translates between standard nomenclatures and an automated reconciliation procedure assigns multiple names as synonyms to the same ORF.
(iii) Integration of different types of networks. BiologicalNetworks supports an arbitrary number of interaction types. Users can upload different types of interactions by specifying different evidence codes that are supported in the BiologicalNetworks (see software documentation for a full list of interaction types).

To illustrate the novelty and capabilities of BiologicalNetworks, in Table 1 we compare BiologicalNetworks against Cytoscape and VisANT. A limitation of Cytoscape and VisANT is that the query capability of these systems is exactly the same as the graph-data manipulation and filtering capability visible on the interface. Thus, the visual integration tools do not have the capacity to take any combination of operations in any order and yet have the system retrieve the specified data in a fashion that optimizes memory and disk operations. We have circumvented this limitation of visual integration by a database-level integration method using a query evaluation engine implementing a query algebra.

Data representation

Pathways are represented as a graph with three types of nodes. The nodes of the first type are reserved for genes, proteins, small molecules, cellular processes, etc. The nodes of the second type (controls) represent events of functional regulation, chemical reactions or protein–protein interactions; they can have physical meaning, may denote general associations; they can represent shared characteristics between components. The nodes of the third type represent complex objects, such as macromolecular complexes, functional groups, pathways etc. In this case components are made up of subcomponents, being compound or modular, and the connections between modular components (or modules) exist along with interconnections between their subcomponents. Interactions in BiologicalNetworks can also be defined as successively higher-level connections between groups of proteins, complexes, pathways or sub-networks.

Most importantly, a classification scheme of Property Types representation (about 2000 unique Property Types and 25 000 nodes of the Property Types tree) allows BiologicalNetworks to represent detailed micro-level information. For example, a protein/process is not only localized in the 'nucleus', but it is represented as the starting point of finer subdivisions, such as 'within the nuclear membrane' and ending in 'outer surface of the nuclear membrane' as a 'component' or as simply 'peripherally associated' (Figure 1). Such a data model and integration environment align well with data representations of existing databases, such as BIND (2), KEGG (3), TransPath (18), eMaze (19) and significantly extends the concepts of other tools, such as Cytoscape and VisANT. To aid biological understanding, interaction networks and protein complexes can be viewed within the context of GO (17) annotations or KEGG (3) pathway assignments.

Interactions have not only information about the relevant literature, but also the experimental system used and a rich array of details on the evidence and classification of the biological properties. For example, a 'genetic interaction' between two genes may have information on the wild type/mutant forms, 'phenotype' (invasiveness, etc.), mutant 'allele', the number of gene copies, etc. Such a rich degree of annotation should play a vital role in understanding the nature of the interaction (see manual for description).

An important goal of systems biology is to generate dynamical models of molecular interaction networks (20). To enable this capability we have stored kinetic parameters as properties of reactions, reactants and products. This has been possible to achieve because of our representation of all three above objects as nodes in the interaction graph. This allows the user to represent the process graphs in SBML (21) format for dynamical simulation using a variety of computational methods.

To accompany BiologicalNetworks, we have developed a preliminary standard for exchanging files that have visual markup and annotation of network layouts. Users of BiologicalNetworks can input several basic data types, including data in standardized network and interaction data exchange formats, such as PSI-MI (22), BioPAX (http://www.biopax.org) and SBML (21).

Data analysis

Once a network dataset has been imported or loaded into BiologicalNetworks, the genes or proteins within it can be queried for other known and predicted interactions from the PathSys's database. Additionally, a repository of curated

Table 1. Comparison of BiologicalNetworks against Cytoscape and VisANT

	BiologicalNetworks	Cytoscape	VisANT
Graph manipulation	Developed in house	Based on yFiles package graph engine	Developed in house
Project workspace	Project workspace; data sharing, through user/account/user privileges mechanism	Not available	Project workspace could be shared by e-mail
Data representation	Generic data model having three types of nodes (primary, connector and graph nodes representing modularity) and Node/Attribute types hierarchies	Ternary relations; no modularity	Ternary relations; modularity presented
Input	Local file, database load	.sif formatted file	Database load
Output	Local (tab delimited, xml, SBML, BN project) file; database edit/update; image printing	Local file; Image printing	.visML file
Data integration	Data integration engine performing data and property types integration, thus creating biological data and properties ontologies	GO database	SGD, KEGG, GO are integrated
Filtering	Filtering by any combination of Attribute/Node types from Attribute/Node type hierarchies	Flexible filters with different attributes of node and edge	Several 'select' filters available

Search	Analytical search tools;
	Keyword search;
	Build/expend pathways;
	Find direct interactions;
	Find covering pathways (all shortest paths);
	Find common targets/regulators;
	Find intersections with curated pathways;
		Search node name on the graph	Search by keyword and node name on the graph

Network operations	Various layouts, Network intersection/ union/subtraction, statistics, search for cycles, Networks comparison (Network BLAST)	Various layouts, several plug-ins for network operations available	Relaxing layout and statistical tool available

Microarray data	Import/Export microarray data;
	Expression patterns;
	Clustering analysis (different clustering algorithms);
	Visually display (static and dynamic time display) gene expressions on the pathways;
	Building pathways from expression values;
	Building correlation (e.g. Pearson correlation) networks;
	Run GO terms overrepresentation analysis (Fisher's test) on expression clusters, networks or group of genes
		Several plug-ins available	Not available

pathways (∼100 for *S.cerevisiae* and several hundreds from other organisms) is available for analysis. Imported interactions and components define a network 'workspace', which can be annotated and saved for sharing inside user groups working with BiologicalNetworks.

To enable data analysis, the following tools are available:

Search: find and display a list of objects based on a name or a keyword.

Expand: searches the database and displays objects functionally linked to a selected node or a set of nodes. Thus, by alternating expand and filtering options, users can browse through the database building their favorite pathways.

Build pathways: finds a set of links between two or more nodes by searching for the shortest path in the total network of all links in the database. This tool assists in finding regulatory paths between all selected objects.

Find common targets/regulators: searches for common targets or regulators for the group of molecules. This tool as well as Build Pathway can find functional links between proteins in the lists imported from other programs (e.g. gene expression clusters).

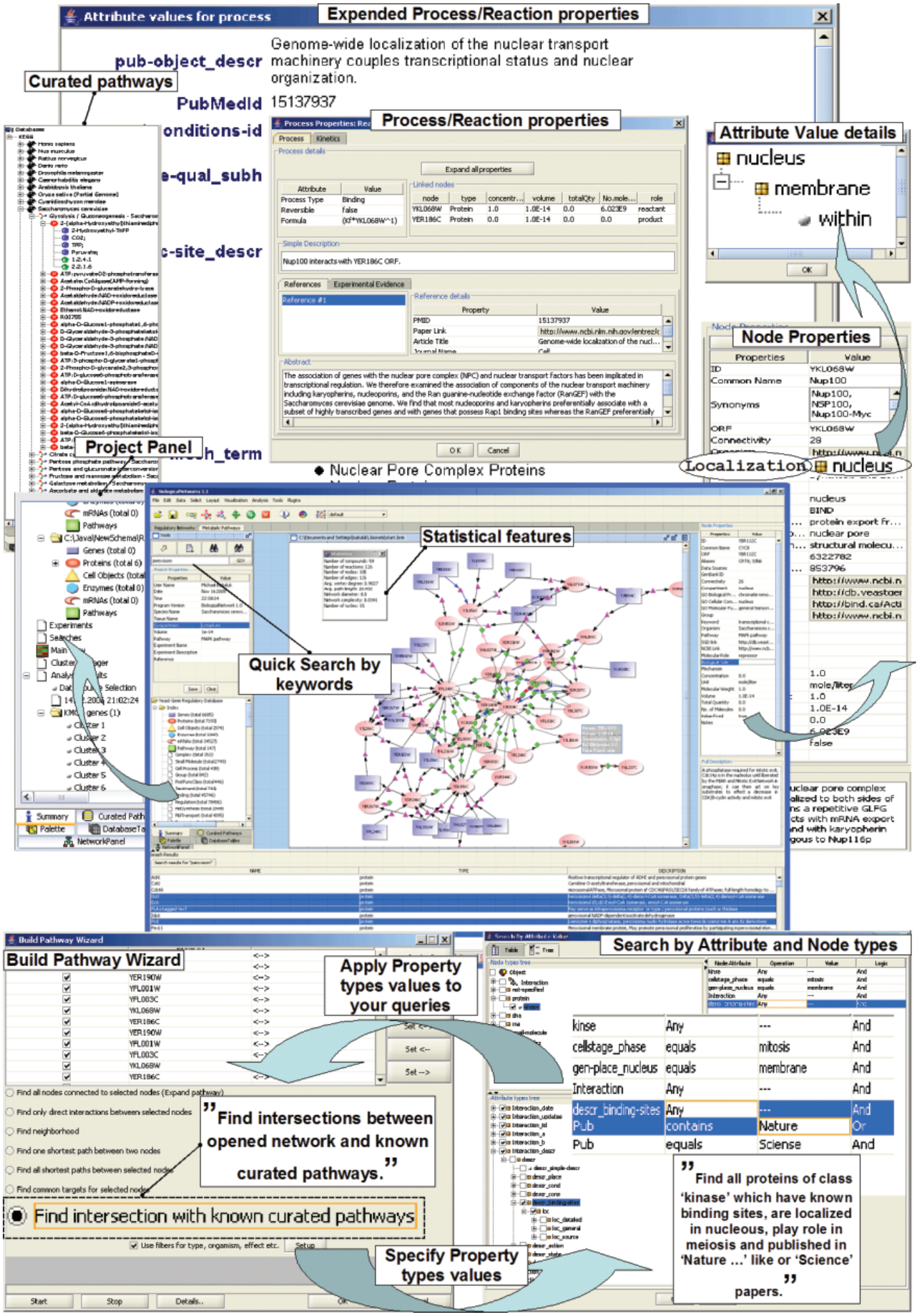
Find intersection with curated pathways: searches a group of nodes for other known and predicted interactions from the PathSys's repository of curated pathways.

BiologicalNetworks provides an advanced querying facility for retrieving the data of user's interest by querying Nodes and Properties types. User friendly querying interface allows user to make query with any logical combination of conditions both on Node and Property trees (Figure 1, see User's Tutorial for details).

Networks can also be analyzed for graph topological properties, such as degree distributions, path lengths, shortest paths or clustering coefficients.

Microarray data analysis

Expression data are easily imported through the Import Expression Data Wizard with a minimal amount of data


Attribute values for process
Expended Process/Reaction properties
Curated pathways
Process/Reaction properties
Attribute Value details
Node Properties
Project Panel
Statistical features
Quick Search by keywords
Localization
Build Pathway Wizard
Apply Property types values to your queries
Search by Attribute and Node types
Specify Property types values
Set <--
Set -->
Start
Stop
Details..
OK
Cancel
Expand all properties
Nuclear Pore Complex Proteins

pub-object_descr: Genome-wide localization of the nuclear transport machinery couples transcriptional status and nuclear organization.

PubMedId: 15137937

Attribute	Value
Process Type	Binding
Reversible	false
Formula	(Kf*YKL068W^1)

node	type	concentr...	volume	totalQty	No.mole...	role
YKL068W	Protein	1.0	1.0E-14	0.0	6.023E9	reactant
YER186C	Protein	0.0	1.0E-14	0.0	0.0	product

Simple Description: Nup100 interacts with YER186C ORF.

Property	Value
PMID	15137937
Paper Link	http://www.ncbi.nlm.nih.gov/entrez/c
Article Title	Genome-wide localization of the nucl...

Abstract: The association of genes with the nuclear pore complex (NPC) and nuclear transport factors has been implicated in transcriptional regulation. We therefore examined the association of components of the nuclear transport machinery including karyopherins, nucleoporins, and the Ran guanine-nucleotide exchange factor (RanGEF) with the Saccharomyces cerevisiae genome. We find that most nucleoporins and karyopherins preferentially associate with a subset of highly transcribed genes and with genes that possess Rap1 binding sites whereas the RanGEF preferentially

nucleus
membrane
within

Properties	Value
ID	YKL068W
Common Name	Nup100
Synonyms	Nup100, NSP100, Nup100-Myc
ORF	YKL068W
Connectivity	28
Localization	nucleus

Node Attribute	Operation	Value	Logic
kinse	Any	---	And
cellstage_phase	equals	mitosis	And
gen-place_nucleus	equals	membrane	And
Interaction	Any	---	And
descr_binding-sites	Any	---	And
Pub	contains	Nature	Or
Pub	equals	Sciense	And

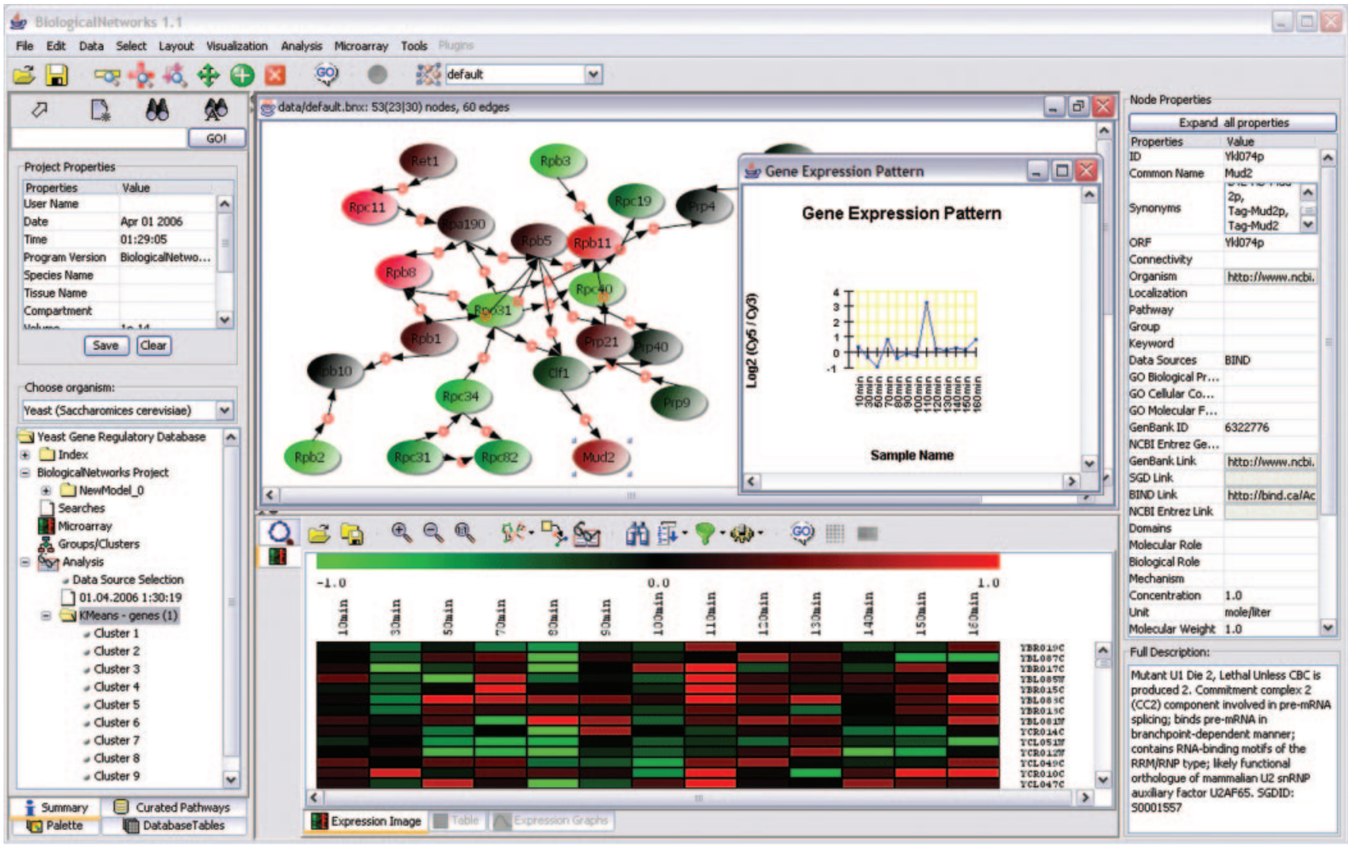
☑	YER190W	<-->
☑	YFL001W	<-->
☑	YFL003C	<-->
☑	YKL068W	<-->
☑	YER186C	<-->
☑	YER190W	<-->
☑	YFL001W	<-->
☑	YFL003C	<-->
☑	YKL068W	<-->
☑	YER186C	<-->

○ Find all nodes connected to selected nodes (Expand pathway)
○ Find only direct interactions between selected nodes
○ Find neighborhood
○ Find one shortest path between two nodes
○ Find all shortest paths between selected nodes
○ Find common targets for selected nodes
● Find intersection with known curated pathways
☑ Use filters for type, organism, effect etc.

"Find intersections between opened network and known curated pathways."

"Find all proteins of class 'kinase' which have known binding sites, are localized in nucleous, play role in meiosis and published in 'Nature …' like or 'Science' papers."

Figure 1. BiologicalNetworks data representation and querying.



Project Properties

Properties	Value
User Name	
Date	Apr 01 2006
Time	01:29:05
Program Version	BiologicalNetwo...
Species Name	
Tissue Name	
Compartment	

Choose organism: Yeast (Saccharomices cerevisiae)

Node Properties

Properties	Value
ID	Ykl074p
Common Name	Mud2
Synonyms	2p, Tag-Mud2p, Tag-Mud2
ORF	Ykl074p
Connectivity	
Organism	http://www.ncbi.
Localization	
Pathway	
Group	
Keyword	
Data Sources	BIND
GO Biological Pr...	
GO Cellular Co...	
GO Molecular F...	
GenBank ID	6322776
NCBI Entrez Ge...	
GenBank Link	http://www.ncbi.
SGD Link	
BIND Link	http://bind.ca/Ac
NCBI Entrez Link	
Domains	
Molecular Role	
Biological Role	
Mechanism	
Concentration	1.0
Unit	mole/liter
Molecular Weight	1.0

Full Description:

Mutant U1 Die 2, Lethal Unless CBC is produced 2. Commitment complex 2 (CC2) component involved in pre-mRNA splicing; binds pre-mRNA in branchpoint-dependent manner; contains RNA-binding motifs of the RRM/RNP type; likely functional orthologue of mammalian U2 snRNP auxiliary factor U2AF65. SGDID: S0001557

Figure 2. Microarray data analysis in BiologicalNetworks.

preprocessing. BiologicalNetworks can interpret files of several types, including Tab Delimited (Stanford) Multiple Sample format, the Affymetrix file format, the TIGR format and GenePix file format.

The Expression Experiment Viewer is designed to display a graphical representation of processed gene expression data. It provides a workspace and a suite of algorithms for data analysis, sorting and searching, clustering and normalization, etc. These allow the user the flexibility in creating meaningful views of the expression data.

Results of the clustering analysis are represented in the form of tables and heat maps, and graphically as expression graphs. These viewers appear as a subtree under the Analysis Result within the main Project Properties tree.

Functionalities available from Microarray submenu and Microarray Experiment Manager Menu bar, allows the user to:

- Open an expression experiment in a form of a table and heat map;
- Sort the experiment by a particular sample;
- Expression data can be visually displayed on an existing pathway diagram by showing different shades of green/red depending on the fold change of expression;
- Build pathways from expression values;
- Build correlation networks (e.g. Pearson correlation);
- Run GO terms overrepresentation analysis (Fisher's test) on expression clusters, networks or group of genes.

In Figure 2 a sample pathway incorporating expression data from a microarray experiment has been assembled. On the left panel is a hierarchical tree of analysis workspace, where different types of microarray data as well as analysis and associated results can be accessed.

On the pathway diagram genes that are up-regulated in a particular experiment are shown in shades of red, while genes that are down-regulated are shown in shades of green; if no match is found the color gray is used. Using these data it is now possible to provide further annotations of edge property such as positive or negative regulation or to provide new edges between nodes.

In conclusion, the BiologicalNetworks web server allows a systems level analysis of genomic scale information as well as single object queries over a variety of databases for integrative views of biological function and for hypothesis generation.

ACKNOWLEDGEMENTS

Part of this work was funded by the National Science Foundation grants EIA-0205061 and EIA-0130059 to A.G. and A.R. The authors also thank the National Biomedical Computational Resource for support. Funding to pay the Open Access publication charges for this article was provided by NSF grant FIBR-0527023.

Conflict of interest statement. None declared.

REFERENCES

1. Ball,C.A., Jin,H., Sherlock,G., Weng,S., Matese,J., Andrada,R., Binkley,G., Dolinski,K., Dwight,S., Harris,M. *et al*. (2001) *Saccharomyces* Genome Database provides tools to survey gene expression and functional analysis data. *Nucleic Acid Res*., **29**, 80–81.
2. Bader,G., Betel,D. and Hogue,C. (2001) BIND–The Biomolecular Interaction Network Database. *Nucleic Acid Res*., **29**, 242–245.
3. Kanehisa,M. and Goto,S. (1999) KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acid Res*., **27**, 29–34.
4. Shannon,P., Markiel,A., Ozier,O., Baliga,N.S., Wang,J.T., Ramage,D., Amin,N., Schwikowski,B. and Ideker,T. (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res*., **13**, 2498–2504.
5. Breitkreutz,B.J., Stark,C. and Tyers,M. (2003) Osprey: a network visualization system. *Genome Biol*., **4**, R22.
6. Nikitin,A., Egorov,S., Daraselia,N. and Mazo,I. (2003) Pathway Studio–the analysis and navigation of molecular networks. *Bioinformatics*, **19**, 2155–2157.
7. Krishnamurthy,L., Nadeau,J., Ozsoyoglu,G., Ozsoyoglu,M., Schaeffer,G., Tasan,M. and Xu,W. (2003) Pathways database system: an integrated system for biological pathways. *Bioinformatics*, **19**, 930–937.
8. Hu,Z., Mellor,J., Wu,J. and DeLisi,C. (2004) VisANT: an online visualization and analysis tool for biological interaction data. *BMC Bioinformatics*, **5**, 17.
9. Hu,Z., Mellor,J., Wu,J., Yamada,T., Holloway,D. and DeLisi,C. (2005) VisANT: data-integrating visual framework for biological networks and modules. *Nucleic Acid Res*., **33**, W352–W357.
10. Mlecnik,B., Scheideler,M., Hackl,H., Hartler,J., Sanchez-Cabo,F. and Trajanoski,Z. (2005) PathwayExplorer: web service for visualizing high-throughput expression data on biological pathways. *Nucleic Acids Res*., **33**, W633–W637.
11. Pandu,R., Guru,R.K. and Mount,D.W. (2004) Pathway Miner: extracting gene association networks from molecular pathways for predicting the biological significance of gene expression microarray data. *Bioinformatics*, **20**, 2156–2158.
12. Chung,H.J., Kim,M., Park,C.H., Kim,J. and Kim,J.H. (2004) ArrayXPath: mapping and visualizing microarray gene-expression data with integrated biological pathway resources using Scalable Vector Graphics. *Nucleic Acids Res*., **32**, W460–W464.
13. Goesmann,A., Haubrock,M., Meyer,F., Kalinowski,J. and Giegerich,R. (2002) PathFinder: reconstruction and dynamic visualization of metabolic pathways. *Bioinformatics*, **18**, 124–129.
14. Pan,D., Sun,N., Cheung,K.H., Guan,Z., Ma,L., Holford,M., Deng,X. and Zhao,H. (2003) PathMAPA: a tool for displaying gene expression and performing statistical tests on metabolic pathways at multiple levels for *Arabidopsis*. *BMC Bioinformatics*, **32**, W460–W464.
15. Soinov,A., Krestyaninova,M. and Brazma,A. (2003) Towards reconstruction of gene networks from expression data by supervised learning. *Genome Biol*., **4**, R6.
16. Baitaluk,M., Qian,X., Godbole,S., Raval,A., Ray,A. and Gupta,A. (2006) PathSys: integrating molecular interaction graphs for systems biology. *BMC Bioinformatics*, **7**, 55.
17. Harris,M.A., Clark,J., Ireland,A., Lomax,J., Ashburner,M., Foulger,R., Eilbeck,K., Lewis,S., Marshall,B., Mungall,C. *et al*. (2004) The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Res*., **32**, D258–D261.
18. Schacherer,F., Choi,C., Gotze,U., Krull,M., Pistor,S. and Wingender,E. (2001) The TRANSPATH signal transduction database: a knowledge base on signal transduction networks. *Bioinformatics*, **17**, 1053–1057.
19. van Helden,J., Naim,A., Lemer,C., Mancuso,R., Eldridge,M. and Wodak,S.J. (2001) From molecular activities and processes to biological function. *Brief Bioinform*., **2**, 81–93.
20. Chen,K.C., Calzone,L., Csikasz-Nagy,A., Cross,F.R., Novak,B. and Tyson,J.J. (2004) Integrative analysis of cell cycle control in budding yeast. *Mol. Biol. Cell*, **15**, 3841–3862.
21. Hucka,M., Finney,A., Sauro,H.M., Bolouri,H., Doyle,J.C., Kitano,H., Arkin,A.P., Bornstein,B.J., Bray,D., Cornish-Bowden,A. *et al*. (2003) The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models. *Bioinformatics*, **9**, 24–31.
22. Hermjakob,H., Montecchi-Palazzi,L., Bader,G., Wojcik,J., Salwinski,L., Ceol,A., Moore,S., Orchard,S., Sarkans,U., von Mering,C. *et al*. (2004) The HUPO PSI's molecular interaction format—a community standard for the representation of protein interaction data. *Nat. Biotechnol*., **22**, 177–183.