BMC Bioinformatics

BioMed Central

INTEGRATOR: interactive graphical search of large protein interactomes over the Web

BMC Bioinformatics 2006, 7:146 doi:10.1186/1471-2105-7-146

Aaron N Chang (anchang@u.washington.edu)
Jason McDermott (mcdermottj@compbio.washington.edu)
Zachary Frazier (zfrazier@u.washington.edu)
Michal Guerquin (mikeg@compbio.washington.edu)
Ram Samudrala (ram@compbio.washington.edu)

ISSN 1471-2105

Article type Software

Submission date 31 Oct 2005

Acceptance date 16 Mar 2006

Publication date 16 Mar 2006

Article URL http://www.biomedcentral.com/1471-2105/7/146

INTEGRATOR: interactive graphical search of large protein interactomes over the Web

Aaron N. Chang,[1,2,3] Jason McDermott,[2] Zachary Frazier,[2] Michal Guerquin,[2] and Ram Samudrala[2,4]

[1] Dept. of Biomedical and Health Informatics, University of Washington, Seattle, Washington 98195, USA

[2] Dept. of Microbiology, University of Washington, Seattle, Washington 98195, USA

[3] Current Address: Rosetta Inpharmatics LLC, A wholly owned subsidiary of Merck & Co. Inc., Seattle, Washington 98109, USA

[4] Corresponding author. Email: ram@compbio.washington.edu, FAX: 206-732-6055

Abstract

Background: The rapid growth of protein interactome data has elevated the necessity and importance of network analysis tools. However, unlike pure text data, network search spaces are of exponential complexity. This poses special challenges for storing, searching, and navigating this data efficiently. Moreover, development of effective web interfaces has been difficult.

Results: We present Integrator, a web-integrated graphical search tool for protein-protein interaction networks across 50+ genomes.

Conclusions: Integrator provides single and multiple protein searches of the Bioverse database containing experimentally-derived and predicted protein-protein interactions. The interface provides animated local network views, rapid subgraph manipulation, and cross-referencing of functional annotations. Integrator is available at *http://bioverse.compbio.washington.edu/integrator*.

Background

High-throughput technologies that monitor cellular components on a large scale are becoming ubiquitous in the post-genomic era. An important analytical paradigm in systems biology is the molecular interaction network [1]. Networks provide an intuitive visualization of component relationships and are amenable to quantitative graph analysis. Constituents include genes, proteins, small molecules or combinations thereof [2-5]. In particular, public repositories of protein-protein interaction (PPI) data collected from yeast two-hybrid arrays, affinity chromatography, and manual curation methods have grown significantly in recent years [6-8].

Building search tools that effectively navigate these interaction networks remains a significant informatics challenge. This is due in large part to the exponential complexity of the search space, rapid data turnover, decentralized storage of primary data, and diversity in data models [9]. Taking this into consideration, we present Integrator, a tool for the analysis of PPI networks using a centralized data model. Integrator is composed of a highly interactive, low-memory overhead network viewer with an enterprise-level application server back-end accessing data from the Bioverse project [10, 11]. This database contains a large collection of experimentally-derived and predicted PPI data for over 50 genomes based in part by applying the Interolog prediction method [12]. Interologs are interactions predicted between proteins in one species using experimental interaction evidence and the relative sequence homologies to proteins in an orthologous species. Such predictions have been used to

extrapolate novel functional annotations for previously unannotated proteins with high accuracy [13].

In contrast to stand-alone network viewer applications, including one previously released by our group [14], the Integrator interface is completely intertwined with a server-based web application. This means several million PPIs stored in a relational database can be explored quickly over the web through common web browsers with minimal additional software. Integrator provides several new graph manipulation features that significantly improve upon tools previously released. It also performs multiple protein searches where protein identifier sets can be compared or contrasted by connected graph components. Integrator is a simple, all-in-one graphical search solution for large interactomes across several genomes.

Implementation

Integrator is based on a three-tier web application architecture using the Java-based Struts web application framework [15]. This design partitions the client, server, and database into three separate information layers. Advantages to this approach include the avoidance of having users install memory-intensive client programs, especially whenever an upgrade becomes available [16] and placing computational load away from clients and onto high-performance servers. The Struts model-view-controller (MVC) paradigm is used to organize tasks including node and edge searches, identifier synonym resolution, viewer

assembly, and database query. The JUNG graph analysis Java library is used for connected component analysis in multiple protein searches [17]. The network viewer is a modification of the Touchgraph Java applet viewer [18]. The data layer contains non-redundant pairwise PPI data (experimentally derived and predicted) from the Bioverse project warehoused in a MySQL database as described previously [10, 11].

Results and discussion

Single protein identifier search

To search networks around a specific protein, Integrator first attempts to locate exact or similar identifier matches to a given query. If a single match is found, the user is returned a graph around the query protein. If similar identifier matches exist, the user is given a list from which to narrow the search. Links to sequence and functional annotation data for each protein are provided to aid this process. Integrator currently recognizes a number of identifiers including those from Genbank, Flybase, Wormbase, and the Saccharomyces Genome Database [10, 19-23].

To provide a visual interface for traversing network results, we implemented an interactive, frame-by-frame navigation solution (**Figure 1**). A network neighborhood (depth = 3) around a query protein is initially generated by a breadth-first search. Within this network, a user can interactively expand, contract, add, or subtract nodes and edges from the viewer. This allows for

dynamic manipulation of network components to aid visual analysis. When a user wishes to expand the network in greater detail around a specific node, contextual menus (right-click on nodes) can be used to re-center the graph around it. By repeating this process, a user can explore an entire connected network starting from any node.

Networks are represented in graph and table formats in the client browser (**Figure 2**). The graph viewer is based on the open-source Touchgraph viewer which provides several built-in features like pan, zoom, rotation, neighborhood view adjustment, and tool tips [18]. Graph components consist of proteins as nodes and undirected edges between them for physical interactions. The edges are color-coded by confidence value as assessed by the Bioverse project. The nodes are labeled with gene symbol identifiers or Interpro functional annotations, depending on availability [23]. These identifiers are also suffixed with unique Bioverse ID numbers to distinguish between potential isoforms or splice variants. Hovering over a node yields a tool tip containing hyperlinked functional annotations from the Gene Ontology (GO) and Interpro classification systems [23, 24]. Edge tool tips contain database source information and confidence values.

Two tables are also provided below the main viewer, which list the proteins and their interactions. The protein (node) table can be used to sort and manipulate node size, shape, or color. Similarly, the interaction (edge) table can be used to sort columns and select specific interactions to render subgraphs in the graph window. A detailed navigation tutorial is provided online.

Multiple protein identifier search

Integrator also provides the option to search multiple protein identifiers simultaneously. These batch searches are constrained to direct PPIs only (depth = 1). The resulting network is used to determine connected component profiles, or unbroken edge clusters, among them (**Figure 3**). Each individual cluster is then made available for display using the graphical interface. Additionally, users can compare or contrast PPIs between two different sets of proteins.

Conclusions

Integrator provides a simple, unified interface for interactome data using familiar web technology. The learning curve required is relatively low and installation of client software is minimal. As a result, users can focus quickly on network analysis. By restricting searches to varying ranges of PPI (edge) confidence values, a user can compare different networks for a given set of proteins. Users have the added flexibility of manipulating subgraphs within each network using graphical and tabular interfaces.

One restriction for users is the method of PPI prediction scoring currently employed in the Bioverse database that is based on the Interolog method [12]. However, users do have the option to restrict their analysis to original source PPIs by filtering for interactions with the highest confidence score (1.0) using the

table interface. As new PPI data sets become available, we will continue to update the database, meaning search results are likely to change over time. Although the current database is limited to intra-genomic PPIs, efforts are underway to compile inter-genomic PPIs (i.e. host-pathogen PPIs).

Like other two-dimensional graph viewers, there is a practical upper-bound limit on the number of nodes in a network (~500) that can be effectively viewed at once. This is commonly due to spatial crowding produced by existing layout algorithms. Integrator attempts to overcome this limitation by utilizing multiple-frame searches. An alternative solution for viewing large global networks might be the use of three-dimensional hyperbolic layout viewers that can interactively display >100,000 simultaneous nodes [25]. Such viewers are promising approaches to network visualization with the caveat that they are dependent on well-defined minimum spanning trees and require significant computational overhead. Two-dimensional viewers will likely remain the preferred interface of choice for related, web-based search technologies.

A critical new feature in network navigation introduced by Integrator, with respect to other published network viewers, is the interactive table interface. Graph-centric viewers are powerful but generally lack the capacity to render subgraphs rapidly to specific nodes or edges. The table interface in Integrator remedies this shortcoming by allowing users to sort columns on various properties and subselect nodes and edges. To illustrate this point, we compared a graph-only viewer, previous released by our group, versus the Integrator interface (**Figure 4**). We found that graph-only viewers restricted subgraphing to

a node-by-node or edge-by-edge operation, causing significant delay in analysis. The table interface within Integrator far exceeded the graph-only viewer in terms of speed and ease-of-use. The complex visual aspect of network analysis requires that these operations occur quickly, especially when users wish to filter networks against specific molecular criteria. In practice, decomposition of networks in this manner appeared to aid hypothesis generation for most end-users we have encountered.

Currently, our group is working towards providing an advanced text search solution through the Bioverse search homepage (http://bioverse.compbio.washington.edu). This will include significant enhancements in Boolean queries and complex query handling. We are working towards a tighter integration of this interface with Integrator's network analysis tools. Furthermore, by making the Integrator codebase available to the public, we hope that integration with similar or derivative projects will provide interesting new features in the future.

Availability and requirements

Integrator is freely available to all users through any Java-enabled web browser at *http://bioverse.compbio.washington.edu/integrator*. All visits to the web application preserve user and data anonymity. A source code distribution is available at this site.

Authors' contributions

ANC designed and implemented the Integrator web application. JM computed and updated the Bioverse dataset. ZF designed and implemented the Bioverse relational database. MG assisted with database design and database optimization. RS oversaw original design specifications, coordinated all relevant projects and assessed development milestones. All authors read and approved the final manuscript.

Acknowledgements

A.N.C. was supported by a National Library of Medicine Medical Informatics Training Grant (1T15LM07441-01). This work is supported in part by a Searle Scholar Award and NSF Grant DBI-0217241 to R.S.

References

1. Barabasi AL, Oltvai ZN: **Network biology: understanding the cell's functional organization**. *Nat Rev Genet* 2004, **5**:101-13.

2. Ito T, Chiba T, Yoshida M: **Exploring the protein interactome using comprehensive two-hybrid projects**. *Trends Biotechnol* 2001, **19**:S23-7.

3. Tong AH, Evangelista M, Parsons AB, Xu H, Bader GD, Page N, Robinson M, Raghibizadeh S, Hogue CW, Bussey H, Andrews B, Tyers M, Boone C: **Systematic genetic analysis with ordered arrays of yeast deletion mutants**. *Science* 2001, **294**:2364-8.

4. Ideker T, Thorsson V, Ranish JA, Christmas R, Buhler J, Eng JK, Bumgarner R, Goodlett DR, Aebersold R, Hood L: **Integrated genomic and proteomic analyses of a systematically perturbed metabolic network**. *Science* 2001, **292**:929-34.

5. Lee TI, Rinaldi NJ, Robert F, Odom DT, Bar-Joseph Z, Gerber GK, Hannett NM, Harbison CT, Thompson CM, Simon I, Zeitlinger J, Jennings EG, Murray HL, Gordon DB, Ren B, Wyrick JJ, Tagne JB, Volkert TL, Fraenkel E, Gifford DK, Young RA: **Transcriptional regulatory networks in Saccharomyces cerevisiae**. *Science* 2002, **298**:799-804.

6. Xenarios I, Rice DW, Salwinski L, Baron MK, Marcotte EM, Eisenberg D: **DIP: the database of interacting proteins**. *Nucleic Acids Res* 2000, **28**:289-91.

7. Rigaut G, Shevchenko A, Rutz B, Wilm M, Mann M, Seraphin B: **A generic protein purification method for protein complex characterization and proteome exploration**. *Nat Biotechnol* 1999, **17**:1030-2.

8. Uetz P, Giot L, Cagney G, Mansfield TA, Judson RS, Knight JR, Lockshon D, Narayan V, Srinivasan M, Pochart P, Qureshi-Emili A, Li Y, Godwin B, Conover D, Kalbfleisch T, Vijayadamodar G, Yang M, Johnston M, Fields S, Rothberg JM: **A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae**. *Nature* 2000, **403**:623-7.

9. Stein LD: **Integrating biological databases**. *Nat Rev Genet* 2003, **4**:337-45.

10. McDermott J, Samudrala R: **Bioverse: Functional, structural and contextual annotation of proteins and proteomes**. *Nucleic Acids Res* 2003, **31**:3736-7.

11. McDermott J, Samudrala R: **Enhanced functional information from predicted protein networks**. *Trends Biotechnol* 2004, **22**:60-2; discussion 62-3.

12. Matthews LR, Vaglio P, Reboul J, Ge H, Davis BP, Garrels J, Vincent S, Vidal M: **Identification of potential interaction networks using sequence-based searches for conserved protein-protein interactions or "interologs"**. *Genome Res* 2001, **11**:2120-6.

13. McDermott J, Bumgarner R, Samudrala R: **Functional annotation from predicted protein interaction networks**. *Bioinformatics* 2005, **21**:3217-26.

14. Chang AN, McDermott J, Samudrala R: **An enhanced Java graph applet interface for visualizing interactomes**. *Bioinformatics* 2005, **21**:1741-2.

15. **Struts Framework** [http://struts.apache.org/]

16. Kurniawan B: **Java for the Web with Servlets, JSP, and EJB**. Indianapolis: New Riders; 2002.

17. **JUNG Graph Library** [http://jung.sourceforge.net]

18. **Touchgraph Viewer** [http://touchgraph.sourceforge.net]

19. Gelbart WM, Crosby M, Matthews B, Rindone WP, Chillemi J, Russo Twombly S, Emmert D, Ashburner M, Drysdale RA, Whitfield E, Millburn GH, de Grey A, Kaufman T, Matthews K, Gilbert D, Strelets V, Tolstoshev C: **FlyBase: a Drosophila database. The FlyBase consortium**. *Nucleic Acids Res* 1997, **25**:63-6.

20. Benson DA, Boguski MS, Lipman DJ, Ostell J, Ouellette BF, Rapp BA, Wheeler DL: **GenBank**. *Nucleic Acids Res* 1999, **27**:12-7.

21. Christie KR, Weng S, Balakrishnan R, Costanzo MC, Dolinski K, Dwight SS, Engel SR, Feierbach B, Fisk DG, Hirschman JE, Hong EL, Issel-Tarver L, Nash R, Sethuraman A, Starr B, Theesfeld CL, Andrada R,

Binkley G, Dong Q, Lane C, Schroeder M, Botstein D, Cherry JM: **Saccharomyces Genome Database (SGD) provides tools to identify and analyze sequences from Saccharomyces cerevisiae and related sequences from other organisms**. *Nucleic Acids Res* 2004, **32 Database issue**:D311-4.

22. Stein L, Sternberg P, Durbin R, Thierry-Mieg J, Spieth J: **WormBase: network access to the genome and biology of Caenorhabditis elegans**. *Nucleic Acids Res* 2001, **29**:82-6.

23. Mulder NJ, Apweiler R, Attwood TK, Bairoch A, Barrell D, Bateman A, Binns D, Biswas M, Bradley P, Bork P, Bucher P, Copley RR, Courcelle E, Das U, Durbin R, Falquet L, Fleischmann W, Griffiths-Jones S, Haft D, Harte N, Hulo N, Kahn D, Kanapin A, Krestyaninova M, Lopez R, Letunic I, Lonsdale D, Silventoinen V, Orchard SE, Pagni M, Peyruc D, Ponting CP, Selengut JD, Servant F, Sigrist CJ, Vaughan R, Zdobnov EM: **The InterPro Database, 2003 brings increased coverage and new features**. *Nucleic Acids Res* 2003, **31**:315-8.

24. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G: **Gene ontology: tool for the unification of biology. The Gene Ontology Consortium**. *Nat Genet* 2000, **25**:25-9.

25. Munzner T: **Exploring Large Graphs in 3D Hyperbolic Space**. *IEEE Computer Graphics and Applications* 1998, **18**:18-23.

Figure legends

Figure 1

Frame-by-frame network navigation. A search begins at node 1 in the left-most window. Once node 2 is reached in the middle window, a new search is performed to center the network around that node. This process is repeated at node 3 in the right-most window. A connected network can be fully traversed using this method. This localized navigation approach works well to reconcile the exponential complexity of networks and the limitations of two-dimensional viewers.

Figure 2

The Integrator network viewer interface. Shown here is a representative network search result around a selected node (yellow) with various other nodes modified for color, size, or shape. The center frame contains the main interactive graph viewer. Users can traverse a network by clicking on nodes. Double-clicking on a node opens a window containing detailed sequence and annotation information. Hovering over nodes also brings up tool tips containing GO and Interpro annotations. Hovering over edges shows tool tips with edge confidence data. Right-click contextual menus also exist for nodes and edges which allow for search, hiding, showing, and changing their visual properties. The slider bar at the top of the graph viewer modulates zoom, rotation, or viewable neighborhood size. Below this are two interactive network tables, nodes on the left, edges on

the right. These tables can be used to sort and modify various node and edge properties in the viewer.

Figure 3

Multiple protein identifier search results. A representative search result where three clusters (defined by connected graph components) are shown with their constituent nodes. Each of the individual clusters can be viewed graphically by following their hyperlinks.

Figure 4

Comparison of a graph-only viewer versus the graph-plus-table interface as provided by Integrator. (A) Network viewer without a table interface, previously released by our group (http://bioverse.compbio.washington.edu/viewer). Subgraph operations are performed one node or edge at a time. (B) Integrator network before subgraphing. (C) Integrator network after subgraphing. Subgraphs can be specified on any number of node or edge properties just by sorting columns and selecting the desired rows.

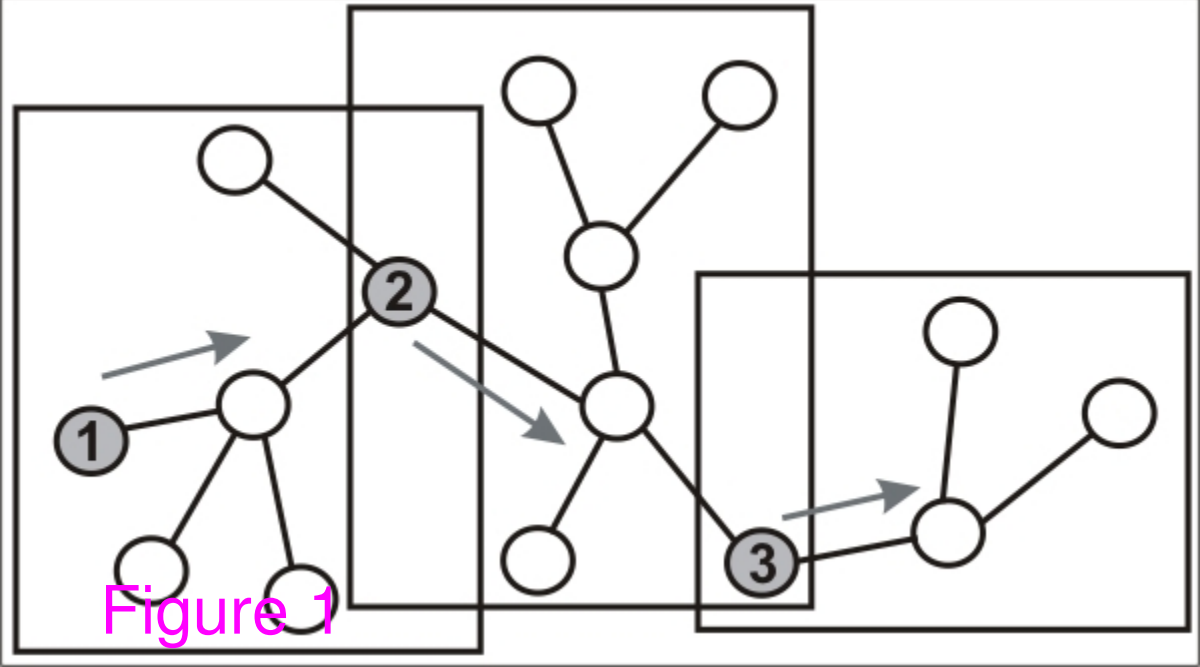

Figure 1

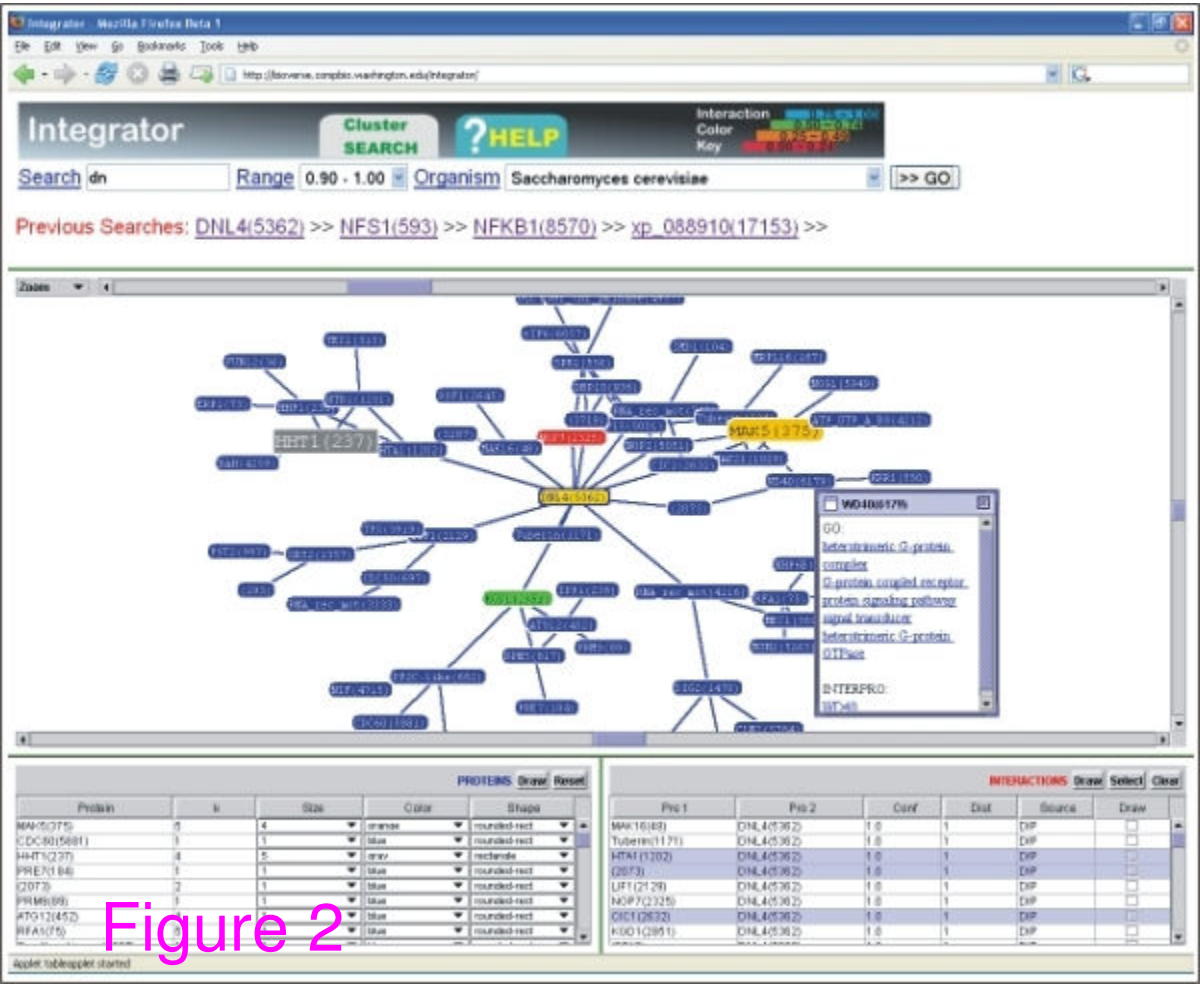
Integrator
Cluster SEARCH
HELP
Interaction Color Key
Search dn
Range 0.90 - 1.00
Organism Saccharomyces cerevisiae
>> GO
Previous Searches: DNL4(5362) >> NFS1(593) >> NFKB1(8570) >> xp_088910(17153) >>
Zoom
HHT1(237)
MAK5(375)
DNL4(5362)
WD40(617)
GO:
heterotrimeric G-protein complex
G-protein coupled receptor protein signaling pathway
signal transducer
heterotrimeric G-protein GTPase
INTERPRO:
PROTEINS Draw Reset
Protein
Id
Size
Color
Shape
INTERACTIONS Draw Select Clear
Pro 1
Pro 2
Conf
Dist
Source
Draw
Applet tableapplet started

Figure 2

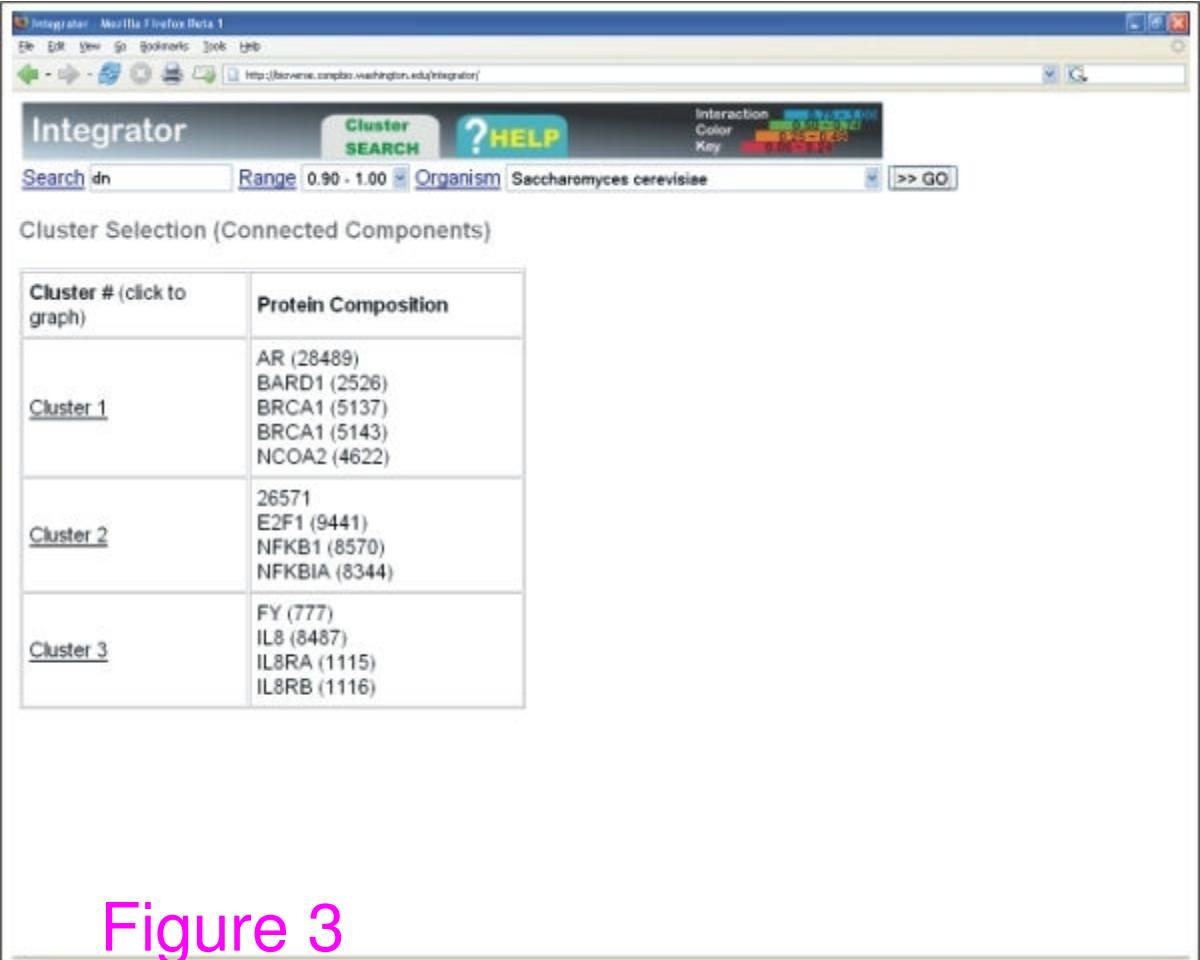
Integrator - Mozilla Firefox Beta 1
File Edit View Go Bookmarks Tools Help
Integrator
Cluster SEARCH
HELP
Interaction Color Key
Search dn
Range 0.90 - 1.00
Organism Saccharomyces cerevisiae
>> GO
Cluster Selection (Connected Components)
Cluster # (click to graph)
Protein Composition
Cluster 1
AR (28489)
BARD1 (2526)
BRCA1 (5137)
BRCA1 (5143)
NCOA2 (4622)
Cluster 2
26571
E2F1 (9441)
NFKB1 (8570)
NFKBIA (8344)
Cluster 3
FY (777)
IL8 (8487)
IL8RA (1115)
IL8RB (1116)
Done

Figure 3


A
Edge Color Key
0.75 ~ 1.00
0.50 ~ 0.74
0.25 ~ 0.49
0.00 ~ 0.24
Bioverse
Find
in M. tuberculosis (4178 molecules)
Search
Preferences | About | Help
This version supercedes Bioverse v1 - Status: Compatibility issues with Internet Explorer for Windows addressed. Under active development.
Last update January 22, 2005 19:30 PDT
M. tuberculosis molecule 953
Mycobacterium tuberculosis (CDC1551) molecule 953
Section
Subsections
Sequence
Structure - secondary
Structure - tertiary
Function
c-terminal domain of b transposition protein + c-terminal effector domain of the bipartite response regulators +
18 functional annotations: Top 3: response regulator receiver + heavy metal response regulator + transcriptional

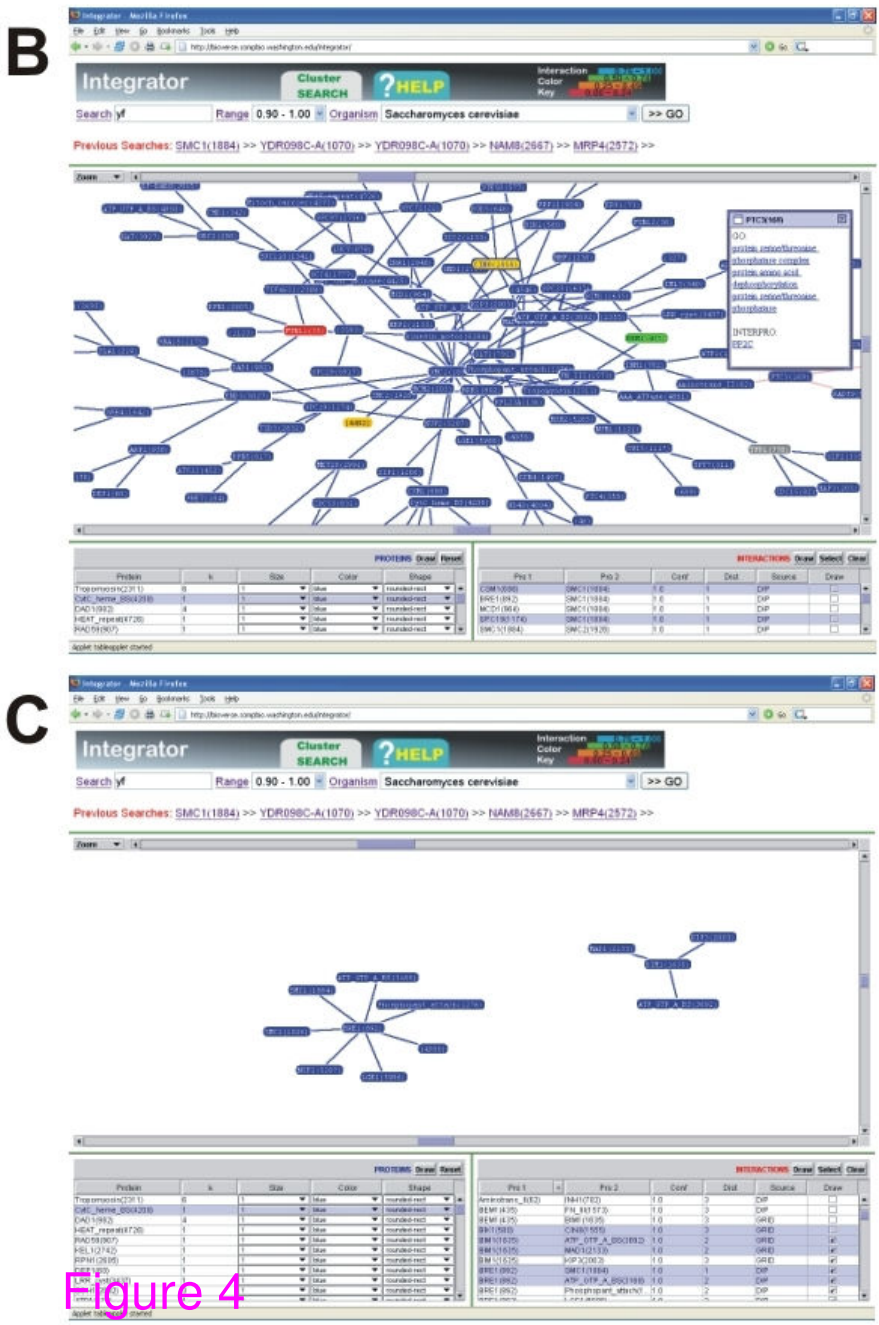
B
Integrator
Cluster SEARCH
HELP
Search yf
Range 0.90 - 1.00
Organism Saccharomyces cerevisiae
>> GO
Previous Searches: SMC1(1884) >> YDR098C-A(1070) >> YDR098C-A(1070) >> NAM8(2667) >> MRP4(2572) >>
C
Integrator
Cluster SEARCH
HELP
Search yf
Range 0.90 - 1.00
Organism Saccharomyces cerevisiae
>> GO
Previous Searches: SMC1(1884) >> YDR098C-A(1070) >> YDR098C-A(1070) >> NAM8(2667) >> MRP4(2572) >>

Figure 4