

## BioNetwork Bench: Database and Software for Storage, Query, and Analysis of Gene and Protein Networks

Oksana Kohutyuk<sup>1,3</sup>, Fadi Towfic<sup>2,3</sup>, M. Heather West Greenlee<sup>2,4</sup> and Vasant Honavar<sup>1-3</sup>

<sup>1</sup>Department of Computer Science, Iowa State University, Ames, Iowa. <sup>2</sup>Bioinformatics and Computational Biology Program, Iowa State University, Ames, Iowa. <sup>3</sup>Artificial Intelligence Research Laboratory, Iowa State University, Ames, Iowa. <sup>4</sup>Department of Biomedical Sciences, Iowa State University, Ames, Iowa.

Corresponding author email: [honavar@cs.iastate.edu](mailto:honavar@cs.iastate.edu)

**Abstract:** Gene and protein networks offer a powerful approach for integration of the disparate yet complimentary types of data that result from high-throughput analyses. Although many tools and databases are currently available for accessing such data, they are left unutilized by bench scientists as they generally lack features for effective analysis and integration of both public and private datasets and do not offer an intuitive interface for use by scientists with limited computational expertise. We describe BioNetwork Bench, an open source, user-friendly suite of database and software tools for constructing, querying, and analyzing gene and protein network models. It enables biologists to analyze public as well as private gene expression; interactively query gene expression datasets; integrate data from multiple networks; store and selectively share the data and results. Finally, we describe an application of BioNetwork Bench to the assembly and iterative expansion of a gene network that controls the differentiation of retinal progenitor cells into rod photoreceptors. The tool is available from <http://bionetworkbench.sourceforge.net/>

**Background:** The emergence of high-throughput technologies has allowed many biological investigators to collect a great deal of information about the behavior of genes and gene products over time or during a particular disease state. Gene and protein networks offer a powerful approach for integration of the disparate yet complimentary types of data that result from such high-throughput analyses. There are a growing number of public databases, as well as tools for visualization and analysis of networks. However, such databases and tools have yet to be widely utilized by bench scientists, as they generally lack features for effective analysis and integration of both public and private datasets and do not offer an intuitive interface for use by biological scientists with limited computational expertise.

**Results:** We describe BioNetwork Bench, an open source, user-friendly suite of database and software tools for constructing, querying, and analyzing gene and protein network models. BioNetwork Bench currently supports a broad class of gene and protein network models (eg, weighted and unweighted, undirected graphs, multi-graphs). It enables biologists to analyze public as well as private gene expression, macromolecular interaction and annotation data; interactively query gene expression datasets; integrate data from multiple networks; query multiple networks for interactions of interest; store and selectively share the data as well as results of analyses. BioNetwork Bench is implemented as a plug-in for, and hence is fully interoperable with, Cytoscape, a popular open-source software suite for visualizing macromolecular interaction networks. Finally, we describe an application of BioNetwork Bench to the problem of assembly and iterative expansion of a gene network that controls the differentiation of retinal progenitor cells into rod photoreceptors.

**Conclusions:** BioNetwork Bench provides a suite of open source software for construction, querying, and selective sharing of gene and protein networks. Although initially aimed at a community of biologists interested in retinal development, the tool can be adapted easily to work with other biological systems simply by populating the associated database with the relevant datasets.

**Keywords:** network analysis, software, network construction, network integration

*Bioinformatics and Biology Insights* 2012:6 235–246

doi: [10.4137/BBI.S9728](https://doi.org/10.4137/BBI.S9728)

This article is available from <http://www.la-press.com>.

© the author(s), publisher and licensee Libertas Academica Ltd.

This is an open access article. Unrestricted non-commercial use is permitted provided the original work is properly cited.



## Introduction

Understanding how the basic molecular building blocks work together to form dynamic functional units (eg, gene and protein networks that orchestrate development, aging, and response to disease) is one of the central goals of modern biology. The emergence of high-throughput techniques for measuring the expression of thousands of genes, interactions between proteins, genes, regulatory RNAs, and other signaling agents, has made possible system-wide measurements of biological variables.

Network models offer a powerful approach for the representation, integration, and analysis of the resulting data. Hence, the construction and analysis of genetic regulatory networks,<sup>1</sup> protein-protein interaction networks, metabolic networks,<sup>2</sup> and their combinations<sup>3,4</sup> are central concerns of systems biology.<sup>1–10</sup> Recent advances in systems biology have led to substantial progress on problems such as understanding the essential macromolecular sequence and structural features of molecular interactions;<sup>11</sup> extracting signaling pathways from gene and protein interaction networks;<sup>12,13</sup> discovering topological and other characteristics of these networks;<sup>14–18</sup> integration of disparate types of data (microarrays, proteomics, physical interaction, subcellular localization, etc.);<sup>13,19,20</sup> predicting the most important nodes in networks;<sup>21</sup> and finding functional modules in networks.<sup>22–24</sup> Such efforts necessarily involve extracting, often in an iterative fashion, meaningful information from large amounts of disparate—often noisy—data and then storing, modifying, and annotating the results of such analyses (eg, hypothesized networks or pathways). There is an urgent need for user-friendly software tools to assist bench scientists to efficiently navigate and manage such a discovery process. Consequently, a number of efforts have focused on development of databases or data warehouses to support specific types of analysis of high throughput gene expression or protein-protein interaction datasets.

Gene expression databases such as the Gene Expression Database (GXD),<sup>25</sup> the Stanford Microarray Database (SMD),<sup>26</sup> and ArrayExpress<sup>27</sup> provide excellent resources for disseminating gene expression data. However, they provide limited support for advanced querying of expression datasets or selective sharing of private datasets within or among research groups.

A variety of software packages have recently become available for constructing and enriching genetic and regulatory networks from gene expression data (eg, ARACNe<sup>28</sup> is a command line tool for reconstructing gene regulatory networks from expression data using information gain; ExpressionCorrelation Cytoscape<sup>29</sup> is a plug-in for constructing gene-correlation networks using Pearson correlation), binding data (eg, MINDy<sup>18</sup> is an algorithm for finding the influence of a modulator gene on the regulatory activity of a transcription factor gene given a set of target genes; MatrixREDUCE<sup>30</sup> is a tool for predicting binding specificity and concentration of transcription factors in the nucleus), and information from literature searches (eg, Agilent Literature Search,<sup>31</sup> GSNet,<sup>32</sup> and Biology Networks Gene Ontology tool (BINGO)<sup>33</sup> construct gene networks based on combining information parsed from paper abstracts/literature), and for visualizing and analyzing networks (eg, Cytoscape,<sup>34</sup> Osprey,<sup>35</sup> BioMiner,<sup>36</sup> Pathway Editor,<sup>37</sup> GenePath,<sup>38</sup> Genetic Network Analyzer,<sup>39</sup> GenMapp2,<sup>40</sup> GeneWays,<sup>41</sup> and geWorkbench<sup>42</sup> are all platforms for visualization and analysis of gene/protein networks). Currently, no single plug-in exists that automates the tasks of constructing, querying, and analyzing the network models for both weighted and un-weighted networks based on gene expression data within Cytoscape while also providing a user-friendly interface for all the functionality, and a repository for storing and retrieving the networks.

It is against this background that we developed BioNetwork Bench, a user-friendly open source suite of database and software tools for constructing, querying, and analyzing gene and protein network models as a Cytoscape plug-in featuring a common and intuitive user-interface for all the functionality. BioNetwork Bench currently supports a broad class of gene and protein network models (eg, weighted and un-weighted undirected graphs, multi-graphs). It enables biologists (especially, bench biologists with limited expertise in database query languages etc.) to manipulate public as well as private gene expression, macromolecular interaction and annotation data; interactively query gene expression datasets; integrate data from multiple networks; query multiple networks for interactions of interest; and store and selectively share the data as well as the results of analyses. BioNetwork Bench is implemented as



a plug-in for, and hence is fully interoperable with, Cytoscape—a popular open-source software suite for visualizing macromolecular interaction networks.

BioNetwork Bench has been successfully used by a team of bench biologists to assemble and iteratively expand an experimentally well-characterized seed network of genes that orchestrate the differentiation of retinal progenitor cells into rod photoreceptors. BioNetwork Bench offers a powerful and easy-to-use tool for bench biologists to explore multiple high throughput datasets in the context of their specialized biological knowledge, to help them generate and prioritize hypotheses for further investigation using traditional molecular or genetic approaches. Although initially aimed at a community of biologists interested in the retina, the tool can be adapted easily to work with other systems simply by populating the associated database with the relevant datasets.

BioNetwork Bench currently supports the storage, manipulation, and sharing of annotated networks and expression datasets; concurrent searching of multiple private and public datasets; transferring query results into a new search, and easy querying by Gene Ontology (GO) annotation<sup>43</sup> (including display and browsing of GO categories and automatic retrieval of GO annotations). It also offers features such as constructing correlation networks from expression data and advanced network merging.

## Program Features

BioNetwork Bench supports multiple types of users with different privileges (administrators, expert users, and standard users). Database administrators have rights to administer the database and user accounts (and in principle, have access to all datasets stored on the system). A script is provided for database administrators to install a MySQL database that will be utilized by BioNetwork Bench, as well as a script to update a local copy of the GO database<sup>44</sup> used for finding GO annotations of network genes.

All users, including database administrators, expert users, and standard users, are allowed to utilize public data and their own private data as well as query networks and expression datasets that they are allowed to access. All users are allowed to load any dataset of their choice into the database, annotate it, and if they so choose, selectively share data (including datasets that they have uploaded into the system, or results

of analyses that they do) with other users by granting them rights to view, update, or delete specific datasets. All users are allowed to create a private copy of a public dataset, modify their private copy and store it in their private space, which is invisible to others except the administrator and the specific users or groups of users that have been granted access rights.

## Manipulating networks and expression data

BioNetwork Bench provides a simple method to load networks or expression data into the database, save and delete networks from the database, or load expression datasets from a file. BioNetwork Bench requires the user to supply dataset citation, a record that includes some basic information about the source of the data (name of the dataset, title of the publication, journal, etc.) when saving a new dataset into the database. Users can query for the dataset using the dataset citation information. In the case of derived data, the citation can include details of the workflow or analysis steps that were used to generate the derived data (eg, selection of a subset of data based on some criteria).

## Building correlation networks from expression data

BioNetwork Bench currently supports the construction of gene expression correlation networks (undirected weighted graphs) from expression data using a Pearson or Spearman Rank correlation function and a user-specified positive or negative correlation threshold.

## Querying expression data

The search function enables users to find genes correlated with “any” or “all” user-provided “target” genes with a correlation coefficient higher than the selected threshold (using Pearson or Spearman functions). A set of networks is returned, where each is a correlation network extracted from a single dataset, trimmed to only include the target genes and genes they are correlated with, in a given dataset. Resulting networks can be loaded into Cytoscape for viewing, merged (with a Merge option in the Networks window) to identify intersection and/or overlap of the networks, or reused for further queries. An “Import selected names from Cytoscape” option imports a list of target gene names of user-selected network nodes

from networks loaded in Cytoscape. “Correlated in at least k datasets” option produces a network composed of target genes and genes correlated with them in at least k datasets, where k is a parameter specified by the user.

## Merging networks

The Merge function allows the creation of a new network merged from a set of selected networks. A resulting network will only include nodes and edges present in all of the selected networks if an “intersection” option is chosen, or all nodes and edges present in at least one of the selected networks if the “union” option is chosen. The users have the ability to force a certain set of nodes to be included in the intersection network, even if these nodes are not present in all of the networks used for merging.

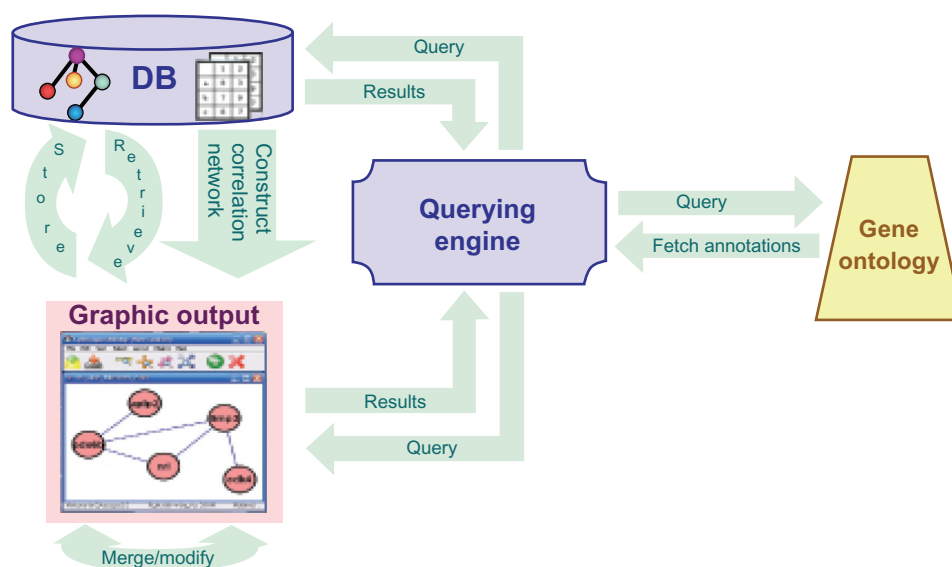
## Querying networks

BioNetwork Bench provides multiple querying capabilities for easily searching for genes in networks loaded into Cytoscape as well as networks stored in the database. The list of genes that satisfy the query criteria are highlighted in the networks being queried. The results can be used in multiple ways, such as creating a new network from highlighted nodes (utilizing Cytoscape’s functionality), storing resulting networks in the database, merging networks, and reusing the results by transferring names of highlighted nodes to new queries.

Querying nodes by name allows the user to determine which networks in the database contain the gene(s) of interest or quickly locate gene(s) in a loaded network. Querying nodes by attribute locates nodes that possess a certain attribute, such as “CanonicalName.” Querying by interaction provides a way to find nodes/genes connected to a set of target genes in previously annotated networks. Querying genes by their GO annotation (see Fig. 4D) provides an easy way to find genes that belong to a certain GO category with respect to molecular function, biological process, or cellular location. A GO graph is loaded in a form of an expandable tree, where a user can select any category at any level to query a network (loaded in Cytoscape or stored in the database) for genes in that category.

## Implementation

BioNetwork Bench consists of a database for storage and manipulation of networks and expression data for multiple users, together with an interactive query engine. The query engine allows interactive probing of expression datasets and annotated networks, as well as merging and reuse of the results obtained (Fig. 1). The query engine retrieves information from networks and expression data loaded into Cytoscape, networks and expression data stored in the database, and GO annotations and term relationships stored in the local copy of the GO database. The BioNetwork Bench Database stores protein and gene expression datasets, as well as inferred or proposed genetic or



**Figure 1.** A diagram of the main components of BioNetwork Bench and a typical workflow.

interaction networks. Since our research is focused on mouse retinal development, the datasets currently present in the database reflect that interest. However, data for any organism or tissue can be stored in the database; the schema used is equally suitable for storing gene/protein expression data and networks for any biological system. A relational database was installed using MySQL Server 5.0 which was chosen due to its widespread use and non-commercial availability.<sup>45</sup> Taking into consideration a possible expansion of the database and a need for future improvements in speed, MySQL was also chosen among other RDBMS (Relational Database Management Systems) due to its support of range and hash partitioning (available in MySQL server 5.0.1 beta). The database consists of 11 tables containing network data, gene expression data, and information about users (Fig. 2).

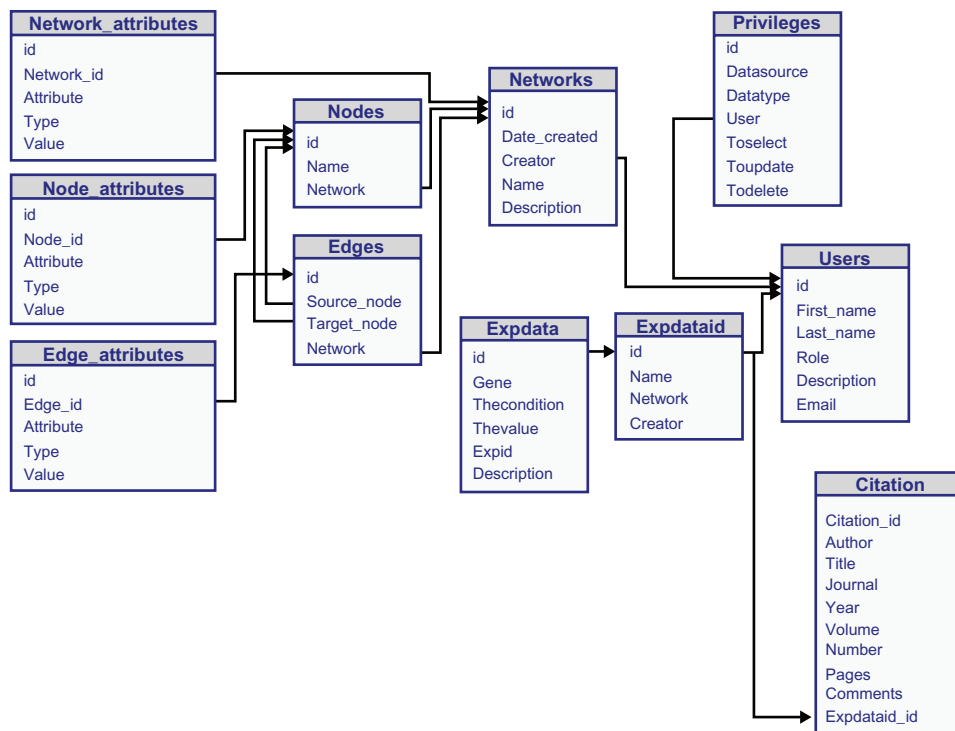
A MySQL copy of GO was obtained from the GO consortium website.<sup>46</sup> The Gene Ontology schema is used for finding gene annotations and determining which nodes (genes) in a network belong to a given GO category. A batch script simplifying the extraction of corresponding files was created to simplify the update process of Mygo, the local copy of the GO schema. A database administrator can easily

update the GO database copy on a regular basis by downloading the corresponding releases of the tool.

The database querying program was developed as a plug-in for Cytoscape, a widely used software package for visualization and analysis of genetic networks. The querying engine is written in Java, employing JDBC (Java Database Connector) to create database connections with the Retina and Mygo databases. Several Borland JBuilder<sup>47</sup> libraries were used in the creation of the graphical user interface. The querying engine translates user-selected options in the Query tab in Expression or Networks window into a set of corresponding SQL queries. Occasionally, extra computation is required in addition to querying, such as computing correlations between gene expression profiles.

### Case study

As noted earlier, BioNetwork Bench is intended to assist bench biologists in exploiting large scale gene expression datasets to iteratively refine a hypothesized gene network and to prioritize hypotheses for further experimental investigation. We describe a case study that demonstrates how BioNetwork Bench supports this type of analysis. Specifically, we illustrate the application of BioNetwork Bench using the analysis



**Figure 2.** Database tables and foreign key constraints in Retina database.

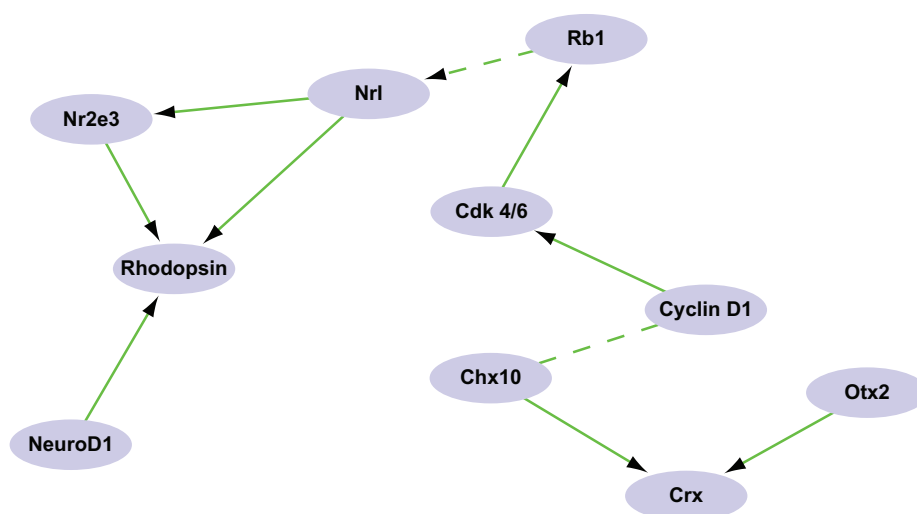
previously reported by Hecker et al,<sup>13</sup> which explored an approach to querying gene expression data from five previously published datasets from developing mouse retina.<sup>48–51</sup> This approach used a ‘seed network’ of genes (Fig. 3) that have been shown using detailed molecular and genetic experiments to govern rod photoreceptor development.<sup>52–60</sup>

Despite their reported low level of concordance across the different datasets, Hecker et al<sup>13</sup> showed that by integrating multiple datasets, they could reconstruct a majority of the links between the seed network genes simply on the basis of observed correlations between genes in multiple (at least 2 out of 5) gene expression datasets (recreated using BioNetwork Bench, Table 1). The resulting network showed positive correlations between several genes known to be expressed by dividing cells and positive correlations between genes known to be expressed by mature photoreceptors, with negative correlations between the two groups. Based on the premise that genes that are likely to play important roles in rod photoreceptor development are likely to be correlated with more than one seed-network gene, multiple datasets were queried to identify genes that were correlated with more than one seed-network member. Cell signaling pathway data (KEGG, <http://www.genome.jp/kegg/pathway.html>)<sup>61</sup> was then retrieved for each gene that was correlated with multiple members of the seed network. Using this procedure, 10 such genes were identified as part of the BMP/SMAD signaling pathway. BMP/SMAD

signaling has been implicated in rod photoreceptor development<sup>62</sup> and 22 proteins were identified as members of WNT/Frizzled signaling, which has been implicated in rod photoreceptor differentiation.<sup>63</sup> From the list of genes correlated with multiple seed-network members, Hecker et al reported 8 additional hypothesized candidates for addition to the seed-network using this approach.<sup>13</sup>

The analysis reported by Hecker et al utilized a combination of custom-written statistical analysis routines in R to compute the correlations and match ids across datasets. The basic steps used in this analysis are not especially complicated. However, perhaps because of the effort needed to glue together the individual steps, either manually or with user-written scripts, such analyses is not commonly used by bench biologists.

BioNetwork Bench was created primarily for use by biologists with limited exposure to database query languages such as SQL but who are nevertheless interested in acquiring and manipulating information about their genes/proteins/pathways of interest. The software can be used by scientists with diverse backgrounds and levels of expertise in informatics. Our goal was to build a software system that would allow users to ask biologically meaningful questions about their data and let them combine and reuse their results to construct and annotate genetic networks. We utilized BioNetwork Bench in our case study (which recreated an analysis of mouse retinal



**Figure 3.** The seed network utilized by Hecker et al for querying expression dataset for genes that have been shown to govern rod development.

**Notes:** The network was constructed based on published experimental evidence and is made up of ten genes. Solid lines indicate direct relationships between seed genes and dashed lines indicate indirect relationships.

**Table 1.** Datasets supporting each edge between all pairs of genes shown to be linked in the seed network.

	SAGE	MOE430.20	Mu74Av2_1	Mu74Av2_2	cDNA microarray	Satisfied in at least 2/5 datasets?	Original seed network
Ccnd1-Cdk4		X	X	X	X	*	Yes
Ccnd1-Chx10							Yes
Ccnd1-Rb1		X					No
Cdk4-Rb1		X	X			*	Yes
Cdk4-Chx10				X			No
Crx-Nrl	X		X	X		*	No
Nrl-Nr2e3	X	X				*	Yes
Nrl-Rhodopsin	X	X			X	*	Yes
Crx-Rhodopsin	X	X		X		*	Yes

**Notes:** BioNetwork Bench was used to construct the correlation network for each dataset while only including nodes whose correlations were equal to or above a 0.65 Spearman Rank correlation cutoff (see Fig. 4A). Each link in the table was then verified using the Network Query functionality of BioNetwork Bench (see Fig. 4B).

development), specifically to identify and prioritize experimental targets for analysis of rod photoreceptor differentiation. BioNetwork Bench, a plug-in for Cytoscape, is freely available for download at <http://bionetworkbench.sourceforge.net>.

We now proceed to describe how all of the steps in the analysis described by Hecker et al can be carried out using the BioNetwork Bench. We start by populating the BioNetwork Bench database with gene expression data from developing mouse retinas using the following datasets: an Affymetrix microarray containing gene expression measurements in developing retina from Dorrell et al,<sup>64</sup> another Affymetrix microarray of retinal genes from Liu et al,<sup>65</sup> Serial Analysis of Gene Expression (SAGE) of developing retina from Blackshaw et al,<sup>66</sup> a cDNA microarray of whole retina from Zhang et al,<sup>67</sup> and an Affymetrix microarray with measurements limited to rod progenitor cells only from Akimoto et al.<sup>68</sup>

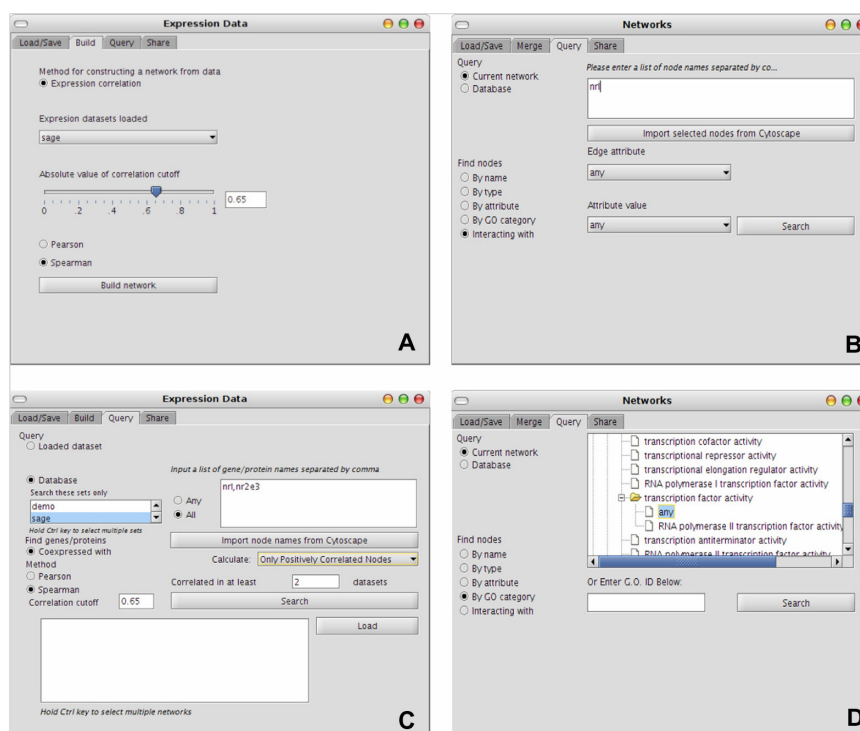
Datasets were pre-processed prior to the analysis. Expression profiles of unidentified genes were discarded. The datasets were processed in the same manner described by Hecker et al.<sup>13</sup> Specifically, in cases where multiple SAGE tags or 2D PAGE spots mapped to a single gene, the total expression for the gene was obtained by summing the values of the SAGE tags/2D PAGE spots. In cases where multiple microarray probes mapped to a single gene, the total expression for the gene was obtained by taking the median of the probes' expressions.<sup>13</sup> As this pre-processing step is highly dataset dependent (eg, SAGE datasets/spots are treated differently than microarray datasets), BioNetwork Bench does not currently offer

a streamlined approach to automate the aforementioned pre-processing steps. However, BioNetwork Bench offers the option to normalize any imported dataset. Thus, the datasets were normalized with respect to a mean of 0 and variance of 1 by BioNetwork Bench to ensure that the changes in expression levels were represented on the same scale.

Currently, several methods are available for dataset normalization. For example, RMAExpress (<http://rmaexpress.bmbolstad.com>) is a program available for Microsoft Windows and Mac OS X platforms for normalizing Affymetrix microarray datasets. Similarly, R's Bioconductor (<http://www.bioconductor.org>) package provides several functions including RMA, MAS 5.0, Quantile and LOESS normalization.

BioNetwork Bench was used to construct a gene network based on each dataset by establishing a link between a pair of genes if the magnitude of Spearman Rank correlation between their expression values was greater than or equal to a threshold of 0.65 (Fig. 4A). Table 1 shows the links between pairs of genes in the seed network that can be recovered from the resulting gene network.

The Expression Dataset Query window (Fig. 4C) was used to search for genes that are positively or negatively correlated with at least two genes in the seed network at a |0.65| Spearman Rank correlation cutoff in at least two out of the five datasets. To do this, all 21 possible pairwise links between nrl, nr2e3, chx10, rho, neurod1, crx, and rb1 (ie, nrl-nr2e3, nrl-chx10, nrl-rho ... etc.) were used as queries. The generated networks from each query were then searched for genes that correlated (either positively



**Figure 4.** (A) The Expression data window and settings used to construct each of the networks for constructing Table 1 is shown. BioNetwork Bench was used to construct the correlation network for each dataset while only including nodes whose correlations was equal to or above a 0.65 Spearman Rank correlation cutoff. (B) The Networks Query window used to query for nodes connected to each gene in the seed network is shown. Each link in table 1 was verified using the Network Query functionality of BioNetwork Bench. (C) An example of the settings used to construct the positively correlated nodes with two or more genes of the seed network supported by at least two datasets with a correlation cutoff of 0.65. (D) A screenshot of the Network Query window demonstrating BioNetwork Bench's feature to query a loaded or stored network based on the GO category of the genes in the network.

or negatively) with both query genes using the Network Query window (Fig. 4B). The specific procedure within BioNetwork Bench is as follows: using the Expression Dataset Query window (example is shown in Fig. 4C) each of the seed network genes was entered, the “Any” radio button was selected (instead of “All”), and the “calculate” field was set to “All Correlated Nodes” (instead of “Only Positively Correlated Nodes” or “Only Negatively Correlated Nodes”). We then queried the network using each of the 21 possible pairwise links between *nrl*, *nr2e3*, *chx10*, *rho*, *neurod1*, *crx*, and *rb1* (ie, *nrl-nr2e3*, *nrl-chx10*, *nrl-rho* ... etc.). The generated networks from each query were then searched for genes that correlated (either positively or negatively) with the seed network genes using the Network Query window (Fig. 4B). The gene list of correlated genes was then exported in tabular format through Cytoscape’s built-in “Export Table” functionality.

The above analysis demonstrates BioNetwork Bench’s capability to ease the construction and querying of gene networks without requiring the user

to possess background knowledge in SQL, scripting languages or other specialized software packages. The time required to perform this analysis was around 4 hours of computing time (total time required for the construction of the networks from the expression data conducted on a 2.4 GHz Quad-core machine with 4 GB of memory). The results produced through this case study reflected the same results presented in Hecker et al’s paper while abiding by a streamlined process that is consistent and less prone to errors that may result from exporting data from one software package format to another, as was necessary for the analysis presented in Hecker et al.<sup>13</sup>

## Discussion

A number of databases for multi-organism data augmented with front-end querying applications have been created in the recent years. The Biological Networks tool<sup>69</sup> offers support for arbitrary queries, clustering and Gene Ontology enrichment analysis of nodes, searching over 20 public databases utilizing known annotations for the stored pathways,





and presents several network construction options. However, it does not allow users to easily include their own networks or datasets in a database search or to easily store and reuse query results. Furthermore, constructing complex queries requires a grasp of database querying language at a level that is often beyond the expertise of a bench biologist. The Biozon database contains sequence, structure, and interaction data and supports querying and fuzzy searches using sequence, expression, or structural similarity.<sup>70,71</sup> However, it does not offer support for user-defined attributes or visualizing search results in the form of a network. Other public databases or data warehouses with limited querying capabilities include: Protein, Signaling, Transcriptional Interactions and Inflammation Networks Gateway (pSTIING), which provides support for generating pictorial representations of protein-protein interactions and transcriptional regulatory networks and includes CLADIST, a tool for clustering gene or protein expression data;<sup>72</sup> BioWarehouse, a data warehouse populated with biological data from public databases;<sup>73</sup> the ONDEX framework, which supports analysis of protein-protein interactions, transcription factors, analysis of relationships between expressed genes, and some text mining;<sup>74</sup> and cPath, a cancer pathway database.<sup>75</sup>

BioNetwork Bench provides the ability to store, share, and modify biological networks and expression datasets seamlessly through Cytoscape. In addition to the correlation network construction functionality that BioNetwork Bench offers, it also allows the users to query the genes in the stored networks and datasets based on Gene Ontology categories (for networks only), custom annotations, interaction partners, and correlation of their expression patterns.

## Conclusions

We have developed BioNetwork Bench, an open source, user-friendly suite of database and software tools for constructing, querying, and analyzing gene and protein network models. BioNetwork Bench currently supports a broad class of gene and protein network models (eg, weighted and un-weighted undirected graphs, multi-graphs). It enables biologists to manipulate public as well as private gene expression, macromolecular interaction and annotation data; interactively query gene expression datasets using a network of seed genes of interest;

integrate data from multiple networks; query multiple networks for interactions of interest; as well as store and selectively share both the data and the Results of analysis. BioNetwork Bench is implemented as a plug-in for, and hence is fully interoperable with Cytoscape, a popular open-source software suite for visualizing macromolecular interaction networks.

Our case study has demonstrated the usefulness of the BioNetwork Bench to bench biologists interested in exploiting high throughput datasets to identify candidate genes and generate testable hypotheses to take back to the bench. Work in progress is aimed at extending the BioNetwork Bench to support a broad class of network representations (eg, Boolean networks,<sup>76</sup> temporal Boolean networks,<sup>77</sup> and their probabilistic counterparts),<sup>78,79</sup> additional algorithms for network construction, topological analysis;<sup>15,16</sup> discovery of network motifs<sup>22,23</sup> and network alignment;<sup>80,81</sup> and support for capturing, storing, publishing, and sharing workflows that include a complex pipeline of analysis or queries. Although initially aimed at a community of biologists interested in the retina, the tool can be adapted easily to work with other biological systems simply by populating the associated database with the relevant datasets.

## Availability and requirements

BioNetwork Bench is downloadable from <http://bionetworkbench.sourceforge.net/>. Minimum requirements for running BioNetwork Bench include:

- 512 Mb of RAM or higher
- 1GHz CPU or better
- Windows 2000/XP/Vista, Mac OS X 10.4, Linux with Java SE 5 or 6 installed (required by CYTOSCAPE)
- Screen resolution of 1024×768 or higher (required by CYTOSCAPE)
- CYTOSCAPE 2.5.0 or above installed
- An active internet connection

## Acknowledgements

This research was funded in part by a grant by a National Institutes of Health grant (EY014931) to Heather West Greenlee and Vasant Honavar, a National Science Foundation Integrative Graduate Education and Research Training (IGERT) grant to Iowa State University (DGE 0504304), and by the



Center for Integrative Animal Genomics and the Center for Computational Intelligence, Learning, and Discovery at Iowa State University. The work of Vasant Honavar while working at the National Science Foundation was supported by the National Science Foundation. Any opinion, finding, and conclusions contained in this article are those of the authors and do not necessarily represent the views of the National Science Foundation. The authors are grateful to Tim Alcon and Laura Hecker for helpful discussions on the research described in this paper.

## Authors' Contributions

Conceived the BioNetwork Bench: VH and HWG. Designed, implemented, and developed the documentation for the BioNetwork Bench: OK and FT. Carried out the case study demonstrating the integrative analysis of multiple datasets from the developing retina using the BioNetwork Bench: FT. Prepared the manuscript for publication: OK, FT, HWG, and VH.

## Funding

This research was funded in part by a grant by a National Institutes of Health grant (EY014931) to Heather West Greenlee and Vasant Honavar, a National Science Foundation Integrative Graduate Education and Research Training (IGERT) grant to Iowa State University (DGE 0504304), and by the Center for Integrative Animal Genomics and the Center for Computational Intelligence, Learning, and Discovery at Iowa State University.

## Competing Interests

Author(s) disclose no potential conflicts of interest.

## Disclosures and Ethics

As a requirement of publication author(s) have provided to the publisher signed confirmation of compliance with legal and ethical obligations including but not limited to the following: authorship and contributorship, conflicts of interest, privacy and confidentiality and (where applicable) protection of human and animal research subjects. The authors have read and confirmed their agreement with the ICMJE authorship and conflict of interest criteria. The authors have also confirmed that this article is unique and not under consideration or published in

any other publication, and that they have permission from rights holders to reproduce any copyrighted material. Any disclosures are made in this section. The external blind peer reviewers report no conflicts of interest.

## References

- de Jong H. Modeling and simulation of genetic regulatory systems: a literature review. *J Comput Biol.* 2002;9(1):67–103.
- Jeong H, Tombor B, Albert R, Oltvai ZN, Barabási AL. The large-scale organization of metabolic networks. *Nature.* 2000;407(6804):651–4.
- Auffray C, Imbeaud S, Roux-Rouquie M, Hood L. From functional genomics to systems biology: concepts and practices. *C R Biol.* 2003;326(10–1):879–92.
- Baitaluk M, Qian X, Godbole S, Raval A, Ray A, Gupta A. PathSys: integrating molecular interaction graphs for systems biology. *BMC Bioinformatics.* 2006;7:55.
- Bruggeman FJ, Westerhoff HV. The nature of systems biology. *Trends Microbiol.* 2007;15(1):45–50.
- Ideker T. Systems biology 101—what you need to know. *Nat Biotechnol.* 2004;22(4):473–5.
- Klipp E, Herwig R, Kowald A, Wierling C, Lehrach H. *Systems Biology in Practice: Concepts, Implementation and Application.* Weinheim, Germany: Wiley-VCH; 2005.
- Special Issue: Systems Biology. *Science.* 2002;295:5560.
- Ge H, Walhout AJ, Vidal M. Integrating 'omic' information: a bridge between genomics and systems biology. *Trends Genet.* 2003;19(10):551–60.
- Liu ET. Systems biology, integrative biology, predictive biology. *Cell.* 2005;121(4):505–6.
- Walhout AJ. Unraveling transcription regulatory networks by protein-DNA and protein-protein interaction mapping. *Genome Res.* 2006;16(12):1445–54.
- Scott J, Ideker T, Karp RM, Sharan R. Efficient algorithms for detecting signaling pathways in protein interaction networks. *J Comput Biol.* 2006;13(2):133–44.
- Hecker LA, Alcon TC, Honavar VG, Greenlee MH. Using a seed-network to query multiple large-scale gene expression datasets from the developing retina in order to identify and prioritize experimental targets. *Bioinform Biol Insights.* 2008;2:91–102.
- Farkas IJ, Jeong H, Vicsek T, Barabasi AL, Oltvai ZN. *The Topology of the Transcription Regulatory Network in the Yeast, Saccharomyces Cerevisiae* [online manuscript]. Chicago: Northwestern University Medical School; 2003. Available from: <http://www.ingentaconnect.com/content/els/03784371/2003/00000318/00000003/art01731>.
- Ravasz E, Somera AL, Mongru DA, Oltvai ZN, Barabási AL. Hierarchical organization of modularity in metabolic networks. *Science.* 2002;297(5586):1551–5.
- Yook SH, Oltvai ZN, Barabasi AL. Functional and topological characterization of protein interaction networks. *Proteomics.* 2004;4(4):928–42.
- Khanin R, Wit E. How scale-free are biological networks. *J Comput Biol.* 2006;13(3):810–8.
- Basso K, Margolin AA, Stolovitzky G, Klein U, Dalla-Favera R, Califano A. Reverse engineering of regulatory networks in human B cells. *Nat Genet.* 2005;37(4):382–90.
- Sharan R, Ideker T. Modeling cellular machinery through biological network comparison. *Nat Biotechnol.* 2006;24(4):427–33.
- Bernard A, Hartemink AJ. Informative structure priors: joint learning of dynamic regulatory networks from multiple types of data. *Pac Symp Biocomput.* 2005:459–70.
- Jeong H, Mason SP, Barabási AL, Oltvai ZN. Lethality and centrality in protein networks. *Nature.* 2001;411(6833):41–2.
- Segal E, Shapira M, Regev A, et al. Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. *Nat Genet.* 2003;34(2):166–76.
- Milo R, Shen-Orr S, Itzkovitz S, Kashtan N, Chklovskii D, Alon U. Network motifs: simple building blocks of complex networks. *Science.* 2002;298(5594):824–7.



24. Sen TZ, Kloczkowski A, Jernigan RL. Functional clustering of yeast proteins from the protein-protein interaction network. *BMC Bioinformatics*. 2006;7:355.
25. Constance M, Smith J, Jacqueline H, Finger J, Hayamizu TF, McCright IJ, Eppig JT, Kadin JA, Richardson JE, Ringwald M. The mouse Gene Expression Database (GXD): 2007 update. *NAR*. 2006;35(1):D618–D623.
26. Sherlock G, Hernandez-Boussard T, Kasarskis A, Binkley G, Matese JC, Dwight SS, Kaloper M, Weng S, Jin H, Ball CA, Eisen MB, Spellman PT, Brown PO, Botstein D, Cherry JM. The Stanford Microarray Database. *NAR*. 2001;29(1):125–155.
27. Parkinson H, Kapushesky M, Shojatalab M, Abeygunawardena N, Coulson R, Farne A, Holloway E, Kolesnykov N, Lilja P, Lukk M, Mani R, Rayner T, Sharma A, William E, Sarkans U, Brazma A. ArrayExpress—a public database of microarray experiments and gene expression profiles. *NAR*. 2006;35(1):D747–D750.
28. Margolin A, Nemenman I, Basso K, et al. ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. *BMC Bioinformatics*. 2006;7(Suppl 1):S7.
29. Niissalo A. *Cytoscape and its Plugins*. Finland: Department of Computer Science, University of Helsinki; 2007.
30. Foat BC, Morozov AV, Bussemaker HJ. Statistical mechanical modeling of genome-wide transcription factor occupancy data by MatrixREDUCE. *Bioinformatics*. 2006;22(14):e141.
31. Agilent Technologies (2007). Agilent Literature Search [Computer software]. Santa Clara, CA. Retrieved July 22, 2007.
32. Choi YJ (2006). GSNNet [Computer software]. Daejeon, Korea. Retrieved June 23, 2006.
33. Maere S, Heymans K, Kuiper M. BiNGO: a Cytoscape plug-in to assess overrepresentation of gene ontology categories in biological networks. *Bioinformatics*. 2005;21(16):3448–9.
34. Shannon P, Markiel A, Ozier O, et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res*. 2003;13:2498–504.
35. Breitkreutz BJ, Stark C, Tyers M. Osprey: a network visualization system. *Genome Biol*. 2003;4(3):R22.
36. Sirava M, Schafer T, Eiglsperger M, et al. BioMiner—modeling, analyzing, and visualizing biochemical pathways and networks. *Bioinformatics*. 2002;18 Suppl 2:S219–30.
37. Sorokin A, Paliy K, Selkov A, et al. The Pathway Editor: a tool for managing complex biological networks. *IBM Journal of Research and Development*. 2006;50(6):561–73.
38. Zupan B, Bratko I, Demsar J, et al. GenePath: a system for inference of genetic networks and proposal of genetic experiments. *Artif Intell Med*. 2003;29(1–2):107–30.
39. de Jong H, Geiselman J, Hernandez C, Page M. Genetic network analyzer: qualitative simulation of genetic regulatory networks. *Bioinformatics*. 2003;19(3):336–44.
40. Salomonis N, Hanspers K, Zambon AC, et al. GenMAPP 2: new features and resources for pathway analysis. *BMC Bioinformatics*. 2007;8:217.
41. Rzhetsky A, Iossifov I, Koike T, et al. GeneWays: a system for extracting, analyzing, visualizing, and integrating molecular pathway data. *J Biomed Inform*. 2004;37(1):43–53.
42. Califano A, Floratos A, Smith K, Ji Z, Watkinson J. geWorkbench: an open-source platform for integrated genomics. *Bioinformatics*. 2010;26(14):1779–80.
43. Ashburner M, Ball CA, Blake JA, et al. Gene Ontology: tool for the unification of biology. *Nat Genet*. 2000;25(1):25–9.
44. Harris MA, Clark J, Ireland A, et al. The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Res*. 2004;32(Database issue): D258–61.
45. Oracle Corporation (2012). MySQL Server V5.0 [Computer software]. Redwood Shores, CA. Retrieved June 1, 2012.
46. *Gene Ontology Downloads* [website]. 2007. Available from: <http://www.godatabase.org/dev/database/archive/>.
47. Borland (2007). Borland JBuilder 5 [Computer software]. Cupertino, CA. Retrieved May 1, 2007.
48. Akimoto M, Cheng H, Zhu D, et al. Targeting of GFP to newborn rods by Nrl promoter and temporal expression profiling of flow-sorted photoreceptors. *Proc Natl Acad Sci U S A*. 2006;103(10):3890–5.
49. Blackshaw S, Harpavat S, Trimarchi J, et al. Genomic analysis of mouse retinal development. *PLoS Biol*. 2004;2(9):E247.
50. Dorrell MI, Aguilar E, Weber C, Friedlander M. Global gene expression analysis of the developing postnatal mouse retina. *Invest Ophthalmol Vis Sci*. 2004;45(3):1009–19.
51. Liu J, Wang J, Huang Q, et al. Gene expression profiles of mouse retinas during the second and third postnatal weeks. *Brain Res*. 2006;1098(1):113–25.
52. Ahmad I, Acharya HR, Rogers JA, Shibata A, Smithgall TE, Dooley CM. The role of NeuroD as a differentiation factor in the mammalian retina. *J Mol Neurosci*. 1998;11(2):165–78.
53. Chen S, Wang QL, Nie Z, et al. Crx, a novel Otx-like paired-homeodomain protein, binds to and transactivates photoreceptor cell-specific genes. *Neuron*. 1997;19(5):1017–30.
54. Cheng H, Khanna H, Oh EC, Hicks D, Mitton KP, Swaroop A. Photoreceptor-specific nuclear receptor NR2E3 functions as a transcriptional activator in rod photoreceptors. *Hum Mol Genet*. 2004;13(15):1563–75.
55. Green ES, Stubbs JL, Levine EM. Genetic rescue of cell number in a mouse model of microphthalmia: interactions between Chx10 and G1-phase cell cycle regulators. *Development*. 2003;130(3):539–52.
56. Mears AJ, Kondo M, Swain PK, et al. Nrl is required for rod photoreceptor development. *Nat Genet*. 2001;29(4):447–52.
57. Nishida A, Furukawa A, Koike C, et al. Otx2 homeobox gene controls retinal photoreceptor cell fate and pineal gland development. *Nat Neurosci*. 2003;6(12):1255–63.
58. Pennesi ME, Cho JH, Yang Z, et al. BETA2/NeuroD1 null mice: a new model for transcription factor-dependent photoreceptor degeneration. *J Neurosci*. 2003;23(2):453–61.
59. Rutherford AD, Dhomen N, Smith HK, Sowden JC. Delayed expression of the Crx gene and photoreceptor development in the Chx10-deficient retina. *Invest Ophthalmol Vis Sci*. 2004;45(2):375–84.
60. Zhang J, Gray J, Wu L, Leone G, et al. Rb regulates proliferation and rod photoreceptor development in the mouse retina. *Nat Genet*. 2004;36(4):351–60.
61. Kanehisa M, Araki M, Goto S, et al. KEGG for linking genomes to life and the environment. *Nucleic Acids Res*. 2008;36(Database issue): D480–4.
62. Murali D, Yoshikawa S, Corrigan RR, et al. Distinct developmental programs require different levels of Bmp signaling during mouse retinal development. *Development*. 2005;132(5):913–23.
63. Yu J, He S, Friedman JS, et al. Altered expression of genes of the Bmp/Smad and Wnt/calcium signaling pathways in the cone-only Nrl<sup>-/-</sup> mouse retina, revealed by gene profiling using custom cDNA microarrays. *J Biol Chem*. 2004;279(40):42211–20.
64. Dorrell MI, Aguilar E, Weber C, Friedlander M. Global gene expression analysis of the developing postnatal mouse retina. *Invest Ophthalmol Vis Sci*. 2004;45:1009–9.
65. Liu J, Wang J, Huang Q, et al. Gene expression profiles of mouse retinas during the second and third postnatal weeks. *Brain Res*. 2006;1098(1):113–25.
66. Blackshaw S, Harpavat S, Trimarchi J, et al. Genomic analysis of mouse retinal development. *PLoS Biol*. 2004;2(9):E247.
67. Zhang SSM, Xuming X, Liu MG, et al. A biphasic pattern of gene expression during mouse retina development. *BMC Dev Biol*. 2006;6:48.
68. Akimoto M, Cheng H, Zhu D, et al. Targeting of GFP to newborn rods by Nrl promoter and temporal expression profiling of flow-sorted photoreceptors. *Proc Natl Acad Sci U S A*. 2006;103(10):3890–5.
69. Baitaluk M, Sedova M, Ray A, Gupta A. Biological Networks: visualization and analysis tool for systems biology. *Nucleic Acids Res*. 2006;34(Web Server issue):W466–71.
70. Birkland A, Yona G. BIOZON: a hub of heterogeneous biological data. *Nucleic Acids Res*. 2006;34(Database issue):D235–42.
71. Birkland A, Yona G. Biozon: a system for unification, management and analysis of heterogeneous biological data. *BMC Bioinformatics*. 2006;7:70.
72. Ng A, Busrteinas B, Gao Q, Mollison E, Zvelebil M. PSTIING: a ‘systems’ approach towards integrating signaling pathways, interaction and transcriptional regulatory networks in inflammation and cancer. *Nucleic Acids Res*. 2006;34:D527–34.



73. Lee TJ, Pouliot Y, Wagner V, et al. BioWarehouse: a bioinformatics database warehouse toolkit. *BMC Bioinformatics*. 2006;7:170.
74. Köhler J, Baumbach J, Taubert J, et al. Graph-based analysis and visualization of experimental results with ONDEX. *Bioinformatics*. 2006; 22(11):383–90.
75. Cerami EG, Bader GD, Gross BE, Sander C. CPath: open source software for collecting, storing, and querying biological pathways. *BMC Bioinformatics*. 2006;7:497.
76. Akutsu T, Miyano S, Kuhara S. Identification of genetic networks from a small number of gene expression patterns under the Boolean network model. *Pac Symp Biocomput*. 1999:17–28.
77. Silvescu A, Honavar V. Temporal boolean network models of genetic networks and their inference from gene expression time series. *Complex Systems*. 2001;13:54–75.
78. Shmulevich I, Dougherty ER, Kim S, Zhang W. Probabilistic Boolean networks: a rule-based uncertainty model for gene regulatory networks. *Bioinformatics*. 2002;18(2):261–74.
79. Santos E, Young JD. Probabilistic temporal networks: A unified framework for reasoning with time and uncertainty. *Int J Approx Reason*. 1999;20(3):263–91.
80. Kalaev M, Bafna V, Sharan R. Fast and accurate alignment of multiple protein networks. *J Comput Biol*. 2009 Aug;16(8):989–99.
81. Towfic F, Greenlee MHW, Honavar V. Aligning biomolecular networks using modular graph kernels. To Appear In *Lecture Notes in Bioinformatics*. 2009.