

## Systems biology

## BioNetBuilder: automatic integration of biological networks

Iliana Avila-Campillo<sup>3,†</sup>, Kevin Drew<sup>1,2,†</sup>, John Lin<sup>1</sup>, David J. Reiss<sup>3</sup> and Richard Bonneau<sup>1,2,\*</sup><sup>1</sup>Department of Biology, <sup>2</sup>Courant Institute, Department of Computer Science, New York University, New York, NY, USA and <sup>3</sup>Institute for Systems Biology, Seattle, WA, USA

Received on September 28, 2006; revised and accepted on November 21, 2006

Advance Access publication November 30, 2006

Associate Editor: Trey Ideker

## ABSTRACT

BioNetBuilder is an open-source client-server Cytoscape plugin that offers a user-friendly interface to create biological networks integrated from several databases. Users can create networks for ~1500 organisms, including common model organisms and human. Currently supported databases include: DIP, BIND, Prolinks, KEGG, HPRD, The BioGrid and GO, among others. The BioNetBuilder plugin client is available as a Java Webstart, providing a platform-independent network interface to these public databases.

**Availability:** <http://err.bio.nyu.edu/cytoscape/bionetbuilder/>**Contact:** [iliana\\_avila-campillo@merck.com](mailto:iliana_avila-campillo@merck.com)

## 1 INTRODUCTION

Access to large amounts of molecular interaction data is available for many organisms through public and private databases. However it is currently difficult for many users to integrate interactions from these databases so that the resulting molecular networks can be visualized and analyzed. PSI-MI (Orchard *et al.*, 2005) and BioPAX (Luciano, 2005) are data exchange formats that will standardize interaction databases but they are not used by all major public databases as of yet. Furthermore, interaction databases use different identifiers to identify the same gene (GI, SwissProt, internal identifiers, etc.) requiring the resolution of synonymous names/IDs across databases. There are commercial tools available which handle some of these difficulties but they are expensive, proprietary, have limited database sets and/or have limited architecture support (Ariadne Genomics, 2006, [www.ariadnegenomics.com](http://www.ariadnegenomics.com); Ingenuity Systems, 2006, [www.ingenuity.com](http://www.ingenuity.com)).

For these reasons we have developed a freely available, open-source software tool that integrates molecular interactions and other types of high-throughput data from different public databases to build biological networks automatically for all species for which such data can be found. BioNetBuilder, is a plugin for Cytoscape (Shannon *et al.*, 2003), an open-source network visualization platform, allowing for access to features of this well developed visualization tool. BioNetBuilder allows for the creation of networks composed of metabolic relationships, protein and protein–DNA interactions, and associations from comparative genomics regardless of what database the gene product originally came from or

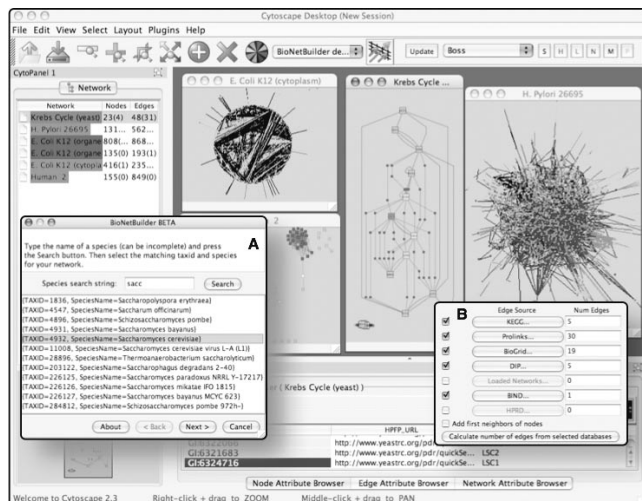
what data format the integration databases support. Another Cytoscape plugin that uses a similar strategy of retrieving biological information is the InteractionFetcher (Reiss, 2005).

BioNetBuilder has an intuitive ‘network creation wizard,’ used to build networks of interacting genes and proteins. We detail the main steps by which users create networks:

- (1) Organism: the user selects an organism among 1523 tax-ids (organisms and species) all of which have entries in at least one interaction database (Fig. 1A).
- (2) Network nodes: the user selects gene products from: user generated lists, on the basis of GO (Gene Ontology, 2000) annotations, all genes matching a selected taxonomy ID, or genes from a previously saved Cytoscape network. While selecting genes through a user-defined list, users can specify in their lists different identifiers from different databases by pre-pending their genes IDs with a prefix such as ‘RefSeq:’ or ‘ORF:’, BioNetBuilder will then automatically interpret and translate the prefix and ID. Other sources of genes include a query tool that returns gene names that match a user defined string pattern, and nodes from currently loaded Cytoscape networks. In all cases users are also presented with the option of growing out gene sets to include neighboring nodes in the following step.
- (3) Edges/Interactions: BioNetBuilder supports different types of interaction databases to create biological networks: functional linkages inferred from evolutionary methods [Prolinks (Bowers *et al.*, 2004)]; protein–protein, protein–DNA and protein–RNA interactions [(HPRD; authorization required; Peri *et al.*, 2003), BioGrid (Stark *et al.*, 2006), BIND (Gilbert, 2005) and DIP (Xenarios *et al.*, 2002)]; metabolic pathways [KEGG (Kanehisa, 2002)]. Users can select databases and set database parameters at this step of the network creation wizard (Fig. 1B).
- (4) Connection to annotations, last steps: the first finishing step allows a user to specify the priority of identifiers (i.e. synonyms/names selected for genes) to visually label the network’s nodes. Next, users attach web resources for annotation to the nodes. For example, genes are linked to protein annotation URLs displaying each protein’s structure-based annotation via Human Proteome Folding Project (HPF, 2006, [www.worldcommunitygrid.org/projects\\_showcase/viewHpf2Research.do](http://www.worldcommunitygrid.org/projects_showcase/viewHpf2Research.do)). Finally, the network is named.
- (5) Cytoscape-Network: once the network is created by BioNetBuilder it can be output, saved, viewed, annotated or analyzed

\*To whom correspondence should be addressed.

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



**Fig. 1.** Cytoscape networks built with BioNetBuilder plugin. Inset (A) depicts the organism selection panel of the plugin's wizard. Inset (B) depicts the edge selection panel of the wizard.

by a large array of Cytoscape features and/or plugins (Fig. 1). For example, the webstart we have provided is bundled with the CyGaggle plugin, providing access to numerous non-Cytoscape analysis tools.

## 2 METHODS

BioNetBuilder consists of a client, described above, and a secure Java servlet. XML-RPC (Apache Software Foundation, 2006) is used for communication between the client and servlet. The servlet consists of several database handlers, which make queries to read-only interaction MySQL databases. There is also a handler for a synonym-resolution system, which is a mapping database for gene identifiers.

The synonym-resolution system maintains all of the translations for different supported identifiers. For example, one can translate from a RefSeq accession to a SwissProt number. This system allows BioNetBuilder to integrate data from databases that identify their genes with different ID types. Much of our synonym database was populated by the IPI database (Kersey *et al.*, 2004).

BioNetBuilder does not require a rigid database schema, file-format or data-model that new data sources must conform to. This allows us to quickly add new database interfaces to the server with source data from several possible formats being used with little formatting cost. In order to access the independent data sources, bioinformaticians can write database handlers in Java that are aware of a particular database's schema, and of the kind of information contained therein.

As part of this tool, we maintain a server that responds to requests made by users/clients. Additionally, we provide database initialization and updating tools (for the supported data sources) so that users can install their own mirror BioNetBuilder servlet and databases. This gives users full control of

backend database updating and the ability to add additional data types to the system; this extensibility is important as several useful databases do not currently have interfaces to the tool [such as MIPS (Pagel *et al.*, 2005), etc.].

BioNetBuilder is a robust and scalable solution for building and visualizing biological networks for all species for which such network data can be found publicly. Users can create connected networks for any species with a NCBI tax-id supported by at least one of the interaction databases. This allows the creation of networks for 1523 different tax-ids.

We provide a Java WebStart for immediate use by users, which includes CyGoose, access to the Gaggle (Shannon *et al.*, 2006). For additional Cytoscape plugins see [www.cytoscape.org](http://www.cytoscape.org). Cytoscape, BioNetBuilder and CyGoose are all coded in Java and are freely available. The BioNetBuilder source code, client executable, servlet Web Archive and a user tutorial are also available from our website.

## ACKNOWLEDGEMENTS

We would like to thank Lee Hood, Peter Bowers and Junghwan Park.

*Conflict of Interest:* none declared.

## REFERENCES

- Apache Software Foundation (2006) Apache XML-RPC.
- Ariadne Genomics (2006) PathwayStudio.
- Bowers,P.M. *et al.* (2004) Prolinks: a database of protein functional linkages derived from coevolution. *Genome Biol.*, **5**, R35.
- Gene Ontology Consortium (2000) The Gene Ontology: tool for the unification of biology. *Nat. Genet.*, **25**, 25–29.
- Gilbert,D. (2005) Biomolecular interaction network database. *Brief. Bioinformatics*, **6**, 194–198.
- HPF: Human Proteome Folding (2006) IBM.
- Ingenuity Systems (2006) Ingenuity Pathways Analysis.
- Kanehisa,M. (2002) The KEGG database. *Novartis Found. Symp.*, **247**, 91–101, discussion 101–103, 119–128, 244–152.
- Kersey,P.J. *et al.* (2004) The International Protein Index: an integrated database for proteomics experiments. *Proteomics*, **4**, 1985–1988.
- Luciano,J.S. (2005) PAX of mind for pathway researchers. *Drug Discov. Today*, **10**, 937–942.
- Orchard,S. *et al.* (2005) The use of common ontologies and controlled vocabularies to enable data exchange and deposition for complex proteomic experiments. *Pac. Symp. Biocomput.*, 186–196.
- Pagel,P. *et al.* (2005) The MIPS mammalian protein–protein interaction database. *Bioinformatics*, **21**, 832–834.
- Peri,S. *et al.* (2003) Development of human protein reference database as an initial platform for approaching systems biology in humans. *Genome Res.*, **13**, 2363–2371.
- Reiss,D.J. *et al.* (2005) Tools enabling the elucidation of molecular pathways active in human disease: application to Hepatitis C virus infection. *BMC Bioinformatics*, **6**, 154.
- Shannon,P. *et al.* (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.*, **13**, 2498–2504.
- Shannon,P.T. *et al.* (2006) The Gaggle: an open-source software system for integrating bioinformatics software and data sources. *BMC Bioinformatics*, **7**, 176.
- Stark,C. *et al.* (2006) BioGRID: a general repository for interaction datasets. *Nucleic Acids Res.*, **34**, D535–539.
- Xenarios,I. *et al.* (2002) DIP, the Database of Interacting Proteins: a research tool for studying cellular networks of protein interactions. *Nucleic Acids Res.*, **30**, 303–305.