Databases and ontologies

PathBuilder—open source software for annotating and developing pathway resources

Kumaran Kandasamy[1,2,3], Shivakumar Keerthikumar[1,2], Rajesh Raju[1,2], T. S. Keshava Prasad[1], Y. L. Ramachandra[2], Sujatha Mohan[1,4] and Akhilesh Pandey[3,*]

[1]Institute of Bioinformatics, International Tech Park, Bangalore 560066, [2]Department of Biotechnology and Bioinformatics, Kuvempu University, Shankarghatta, Karnataka, India, [3]McKusick-Nathans Institute of Genetic Medicine and the Departments of Biological Chemistry, Pathology and Oncology, Johns Hopkins University, Baltimore, Maryland 21205, USA and [4]Research Unit for Immunoinformatics, Research Center for Allergy and Immunology, RIKEN Yokohama Institute, Kanagawa, Japan 230-0045

ABSTRACT

Summary: We have developed PathBuilder, an open-source web application to annotate biological information pertaining to signaling pathways and to create web-based pathway resources. PathBuilder enables annotation of molecular events including protein–protein interactions, enzyme–substrate relationships and protein translocation events either manually or through automated importing of data from other databases. Salient features of PathBuilder include automatic validation of data formats, built-in modules for visualization of pathways, automated import of data from other pathway resources, export of data in several standard data exchange formats and an application programming interface for retrieving existing pathway datasets.

Availability: PathBuilder is freely available for download at http://pathbuilder.sourceforge.net/ under the terms of GNU lesser general public license (LGPL: http://www.gnu.org/copyleft/lesser.html). The software is platform independent and has been tested on Windows and Linux platforms.

Contact: pandey@jhmi.edu

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

1 INTRODUCTION

Experimental research to elucidate biological pathways in detail has generated large amounts of data that are scattered across the published literature. Because of the complexity of pathway data, there is a need for trained biologists to manually collect and curate biological information. A major issue that needs to be addressed is to store, retrieve and visualize the collected data in a simple fashion with provision for integration with other pathway resources. Though software like cPath is available for storing, visualizing and analyzing biological pathways (Cerami *et al.*, 2006), there is currently no publicly available open-source software that allows biologists to rapidly deploy a web-based pathway resource. The importance of
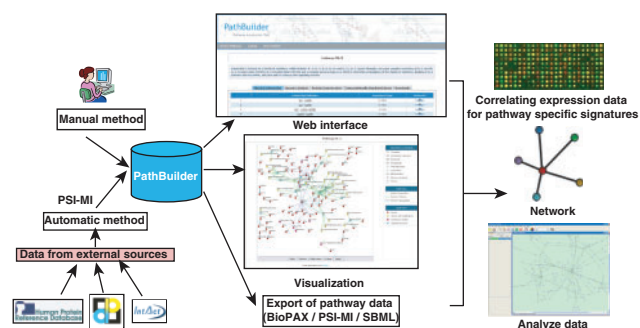

Fig. 1. PathBuilder architecture. Data can be populated manually or automatically. The stored data can be viewed directly through a web browser or can be exported to standard exchange formats for visualization and analysis using other software.

pathways is underscored by the fact that over 200 biological pathway related resources are currently available (Bader *et al.*, 2006).

We have developed PathBuilder, an open-source application which enables annotation of signaling pathways (Fig. 1). Biological characteristics of signaling pathways including protein–protein interactions, enzyme–substrate relationships and protein translocation events can be catalogued using this software. These events occur upon stimulation with a specific ligand or activation of its specific receptor. In addition, the tool has provision for cataloging of genes that are transcriptionally regulated by pathways. Thus, PathBuilder can facilitate pathway data collection as well as rapid deployment of pathway resources.

2 IMPLEMENTATION

PathBuilder is developed using Zope web application framework (http://www.zope.org/). The data is stored in a MySQL database, processed in an application layer implemented in Python programming language and published to the web using DTML, a Zope HTML templating language.

Data stored in PathBuilder can be accessed via standard web-based application programming interface (API) which allows third

---

*To whom correspondence should be addressed.

party software to access data, thus enabling interoperability. The API can be controlled by specifying the URL parameters. For more information on the use of API, please read the documentation available on the project web site.

3 CREATING AND ANNOTATING PATHWAYS

3.1 Annotating pathway data

The annotation pipeline in PathBuilder (Supplementary Fig. 1) has four central steps—annotation of data, automatic validation of logical and typographical errors, initial review and review by Pathway Authorities. The installation of PathBuilder provides an unpopulated functional database with default parameters. The two modes of populating PathBuilder include manual entering of data through a series of web forms and automated import of data. Currently, PathBuilder successfully imports physical interaction datasets as PSI-MI (Hermjakob *et al*., 2004) files from HPRD (Keshava Prasad *et al*., 2009), IntAct (Kerrien *et al*., 2007) and DIP (Salwinski *et al*., 2004). This would allow researchers aggregate data from disparate resources to create custom databases.

PathBuilder was developed primarily for creation of a pathway resource for which the data was entered manually. There are separate web forms available for different data types that allow the user to annotate data through a web browser which permits the annotation process to be carried out at different geographic locations simultaneously.

Data contained in PathBuilder can be reviewed. Any change suggested by an initial reviewer is sent automatically to the respective curator for further changes and the entry is not finalized. Once the reviewer approves an entry, it is marked as 'reviewed' and is finalized in the database. It also allows a final review and editing by designated scientists who are experts in specific pathways called 'Pathway Authorities'. The 'Pathway Authorities' report errors, if any, or specify additional information about a pathway that can be included.

3.2 Browse, lookup, display and export of pathway data

PathBuilder provides browse and lookup options for the annotated pathways. The curator or reviewer can lookup using identifiers such as gene symbol, protein name, Entrez Gene ID or PubMed ID. The pathway home page contains a brief description, a list of molecules involved and hyperlinks to view details of downstream signaling reactions annotated in the pathway. All downstream signaling reactions are displayed under separate tabs and also allow export of pathway data (Supplementary Fig. 2).

3.3 Dynamic generation of network graphs

PathBuilder dynamically generates network graphs that can be viewed through a web browser using Medusa applet (Hooper and Bork 2005). PathBuilder also provides pathway data that can be visualized using downloadable software such as Pajek (Batagelj 1998), Cytoscape (Shannon *et al*., 2003) and Osprey (Breitkreutz *et al*., 2003). Supplementary Figure 3 shows the network graphs of the IL-1 pathway generated using Medusa, Pajek, Cytoscape and Osprey.

3.4 Development of NetPath, a resource for human signaling pathways, using PathBuilder

We used PathBuilder to develop NetPath (http://www.netpath.org/) as a resource for human signaling pathways (S. Mohan *et al*., submitted for publication). Pathway data were populated manually using the web forms in PathBuilder. Supplementary Figure 4 shows various fields for annotating physical interactions. Importantly, the use of PathBuilder for developing NetPath allowed annotation and review by experts in different countries, most of whom had no bioinformatics expertise. Supplementary Table 1 gives a comparison of various features in PathBuilder with other software available for pathway annotation such as cPath (Cerami *et al*., 2006), PATIKA (Demir *et al*., 2002), PathCase (Krishnamurthy *et al*., 2003) and GenMAPP (Dahlquist *et al*., 2002).

4 CONCLUSIONS

PathBuilder is a simple software for creation of pathway resources. PathBuilder facilitates manual entry of biological pathway data in addition to supporting XML-based import of data from other publicly available databases. PathBuilder aims to facilitate storage, retrieval, organization and visualization of biological pathway data in an efficient manner. Future developments in PathBuilder will focus on addition of modules that facilitate integration of transcriptomic data over the current network-based visualization of pathways.

REFERENCES

Bader,G.D. *et al*. (2006) Pathguide: a pathway resource list. *Nucleic Acids Res*., **34**, D504–D506.

Batagelj V,M.A. (1998) Pajek - program for large network analysis. *Connections*, **2**, 47–57.

Breitkreutz,B.J. *et al*. (2003) Osprey: a network visualization system. *Genome Biol*., **4**, R22.

Cerami,E.G. *et al*. (2006) cPath: open source software for collecting, storing, and querying biological pathways. *BMC Bioinformatics*, **7**, 497.

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

Demir,E. *et al*. (2002) PATIKA: an integrated visual environment for collaborative construction and analysis of cellular pathways. *Bioinformatics*, **18**, 996–1003.

Hermjakob,H. *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.

Hooper,S.D. and Bork,P. (2005) Medusa: a simple tool for interaction graph analysis. *Bioinformatics*, **21**, 4432–4433.

Kerrien,S. *et al*. (2007) IntAct–open source resource for molecular interaction data. *Nucleic Acids Res*., **35**, D561–D565.

Keshava Prasad,T.S. *et al.* (2009) Human Protein Reference Database–2009 update. *Nucleic Acids Res.*, **37**, D767–D772.

Krishnamurthy,L. *et al.* (2003) Pathways database system: an integrated system for biological pathways. *Bioinformatics*, **19**, 930–937.

Salwinski,L. *et al.* (2004) The database of interacting proteins: 2004 update. *Nucleic Acids Res.*, **32**, D449–D451.

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