PathBank: Web-based Querying and Visualization of an Integrated Biological Pathway Database

Joshua Ho Tristan Manwaring Seok-Hee Hong Uwe Roehm David Cho Yau Fung
School of Information Technologies
University of Sydney, Australia
{joshua.ho,tman0127,shhong,uroehm,dfun2647}@usyd.edu.au

Kai Xu
NICTA (National ICT Australia)
Australia
KaiKevin.Xu@nicta.com.au

Tim Kraska
University of Münster
Münster, Germany
tim.kraska@uni-muenster.de

David Hart
Axogenic
Australia
dhart@axogenic.com

Abstract

PathBank is a web-based query and visualization system for biological pathways using an integrated pathway database. To address the needs for biologists to visualize and analyze biological pathways, PathBank is designed to be user-friendly, flexible and extensible. It is, to the best of our knowledge, the first web-based system that allows biological pathways to be visualized in three dimensions. PathBank demonstrates the ability to automatically generate and layout biological pathways in response to web-based database query about proteins, genes, a gene ontology and small molecules. Using a novel OWL-to-relational database schema generation approach, it can automatically integrate biological data from different sources that support the BioPAX exchange format (e.g. KEGG, BioCyc). The system's web interface allows both simple keyword and complex query-based searches in the database. The pathway visualization capability is embedded in a Java applet. PathBank makes extensive use of client sides' technology to reduce computational load of the server. It also makes extensive use of open-source technology.

Keywords— **Bioinformatics, Biological Network Visualization, Integrated Database**

1 Introduction

Large amount of biological data are available from the public domain. Storing, mining and visualizing these data present great challenges and opportunities to the life science research and bioinformatics communities. With the emergence of Systems Biology [11], scientists are increasingly interested in studying the large-scale interactions between genes, RNA proteins, and the other biomolecules. Through modeling and simulating these molecular interactions, we start to gain an understanding of complex cellular behaviours.

Molecular interactions are most commonly modeled as networks [5], in which nodes represent biomolecules and edges represent reactions or interactions. Examples include metabolic networks, gene regulatory networks, protein-protein interaction networks, and signal-transduction networks. By analyzing the the topology of these networks, functional and evolutionary information can be gained. Further, by measuring the topological features (e.g. average path length, clustering coefficient and node degree distribution) in different cellular states, we are able to unravel the molecular network dynamics behind a certain cellular behaviour [14].

To fully understand the biological processes underlying these networks, visualization can be a good analysis tool. Visualization provides an intuitive means for biologists to explore biological pathways of their interest [15]. Good visualization reveals the underlying network topology and hence provides biological insight. Currently, a number of bioinformatics tools are available for pathway visualization, including BioPath [16, 6], Cytoscape [17], Osprey [2], BioTapestry [13], and VisANT [9]. These similar tools differ in the variety of graph layout options and the analytical tools provided. A number of biological pathway visualization systems are evaluated in [15].

While these tools provide a large number of methods for pathway visualization and analysis, learning to use them require steep learning curve. Our target users are biologists who are primarily interested in accessing pathway data from existing databases, and visually analyze these pathways. They require simple and intuitive interface with fast and reliable results from the system. Motivated by the need to fill the gap between the user requirement and existing tools, we developed PathBank.

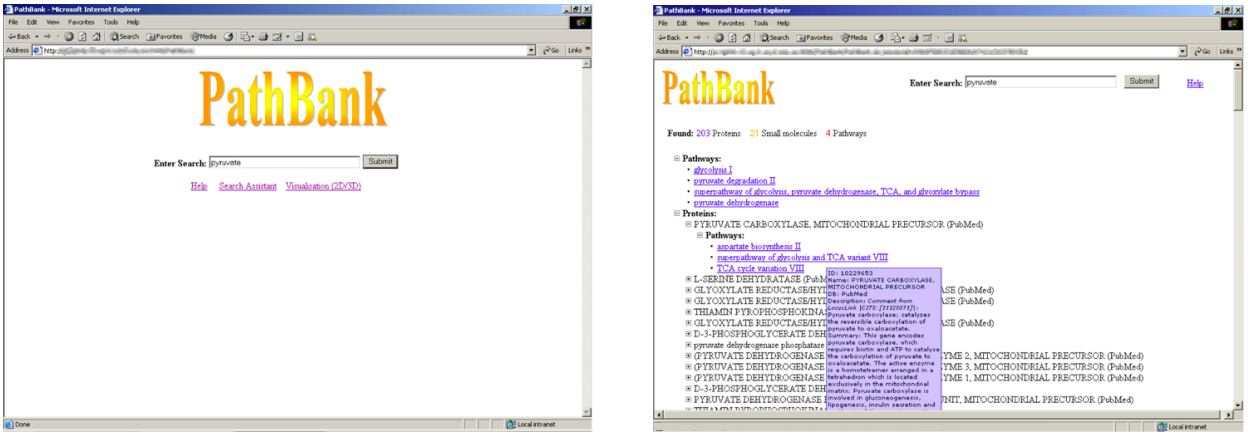

Figure 1: Query view of PathBank. The home page (left) of PathBank is designed to be Google -like. Query results (right) are displayed in a tree-like hierarchy. Additional details of pathways are shown in a floating window on mouse-over.

PathBank features an intuitive Google-like interface for database query, browsing, and pathway visualization (see Figure 1). Search modifiers and boolean operators are allowed for complex searches. Query results are displayed dynamically using AJAX technology. A Java applet is used for pathway visualization (see Figure 2). Different layout styles can be chosen to suit user's need.

This paper first surveys the related work on biological pathway query and visualization in section 2. Then, we present our approach on pathway visualization and database integration in sections 3 and 4, followed by system features in section 5. System design and implementation are discussed in section 6. We conclude in section 7.

2 Related Work

The number of systems designed for pathway visualization is increasing quickly. In this section, we survey some of them that generate visualization automatically. Therefore, we do not include systems that rely on manually-produced visualization such as KEGG [10] and BioCarta (www.biocarta.com). Given the large number of available systems, we selectively cover several of them according to availability, popularity and relevance to PathBank. Namely, they are BioPath [16, 6], Cytoscape [17], Osprey [2], BioTapestry [13], and VisANT [9]. Some of these are designed for pathway only, while others support general biological networks.

BioPath [16, 6] is designed specifically for pathway visualization. It modifies the traditional layered-drawing algorithm — the Sugiyama method [18] — to follow some of the conventions commonly used in hand-made drawings, so it will be easier for biologists to understand. BioPath also provides various search functions that include: search substances, search pathways and reactions, and search for a reaction net between two substances.

Cytoscape [17] provides visualization of molecular interaction networks and related information, such as expression profile and phenotypes. The layouts provided are force-directed methods and a circular layout, however in two dimensions only. It supports overlaying related data on the graph by mapping them to node/edge visual attributes such as size and color. Its visualization and analysis capabilities can be extended by adding new plug-ins.

Osprey [2] provides visualization of general biological network and supports comparison between networks. The visualization method is two dimensional and the available layout methods are circular layout and its few variations. It allows text search such as gene names and uses color to map data related to nodes and edges of the network. The users can also load a few networks and perform analysis such as finding the common part between them.

BioTapestry [13] is a visualization tool designed mainly for genetic regulatory network. Grid drawing algorithm is used to compute graph layout in two dimensions. BioTapestry also supports visualization of network changes over time. Users have the option to view the given network at a specific time using a time slider.

VisANT [9] is a system designed to visualize general biological networks. Besides the 2D graph visualization methods similar to previous systems, it also provides a few statistical analysis methods such as the node degree and clustering coefficients distribution, which can be shown in separate scatter plot. VisANT also adopts a plug-in framework so that new functions can be added easily.

3 Network Visualization

Network visualization concerns two main aspects: the visual representation of nodes and edges, and the layout of the underlying graph. The visual representation (e.g. the size, colour and shape) of nodes and edges is useful for

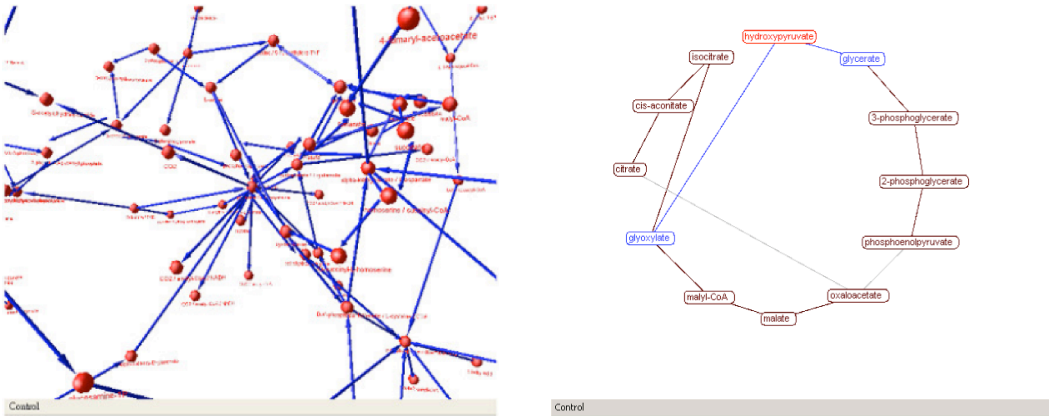

Figure 2: Pathway visualization in PathBank: Selected pathway can be visualized in 3D (left) or 2D (right).

conveying certain information about the pathway. On the other hand, the layout of the graph is important to reveal the underlying topology of the network. In our work, we concentrate on the layout of the network.

In the past decade, a large number of graph drawing algorithms and methods are invented to automatically layout node-edge diagrams [3]. Well known algorithms include tree drawing algorithms, force directed methods, Sugiyama method, multi-dimensional scaling methods and spectral graph drawing methods. For the details, see the recent proceedings of Graph Drawing and Information Visualization conferences.

There are a number of layout algorithms specifically designed for biological pathways (e.g. [1] and [20]). These graph layout methods vary in their ability in highlighting underlying network topology. Users may be more accustomed to a certain kind of layout style, and hence allowing user to choose the layout style is important.

One of the main advantages of PathBank is that user can choose from a variety of graph layouts. Currently, users can choose from either force-directed layout, both in two and three dimensions, or circular layout, while the force-directed 2D layout is the default visualization (see Figure 3).

Metabolic network is often modeled as a hypergraph, where each node represents a collection of metabolites or enzymes. In order to reduce the complexity of the model, hypernodes are concatenated by considering the collection of biomolecules as one node (see Figure 3). As a result, a hypergraph can be represented by a simple directed graph.

PathBank, to the best of our knowledge, is the first web-based visualization tool that draws biological pathway in three dimensions interactively. Note that HCI research suggested that visualization in three dimensions can be more effective in conveying complex information, such as networks [19]. Recently, three dimensional visualization of biological pathway using *virtual reality* has been suggested in [20].

4 Database Integration

PathBank relies on relational database technology in its backend system; we use PostgreSQL 8 in our current prototype, but PathBank in fact supports any JDBC-compliant relational database. The pathway database is optimized for fast browsing, querying, and graph visualization. We achieve this by using appropriate indexes and a layer of relational views that provide fast access to all data needed for visualizing a biological pathway; this way, the visualization component can fetch a complete pathway with just one database access.

An important design goal was the capability to integrate biological data from different sources. We have developed a novel storage approach for biological pathway data, called *Genea*, that allows to automatically convert pathway data, given in BioPAX format, into a corresponding relational schema [12]. Our approach is very generic. It is capable of mapping any given OWL ontology (OWL: Web Ontology Language), e.g. the BioPAX format, into a human-readable relational schema, and it also can automatically load corresponding OWL instance data. Because part of the reasoning over the ontology is done during those schema-generation and instance-data-loading phases, most searches on the backend database can be answered with simple and hence fast SQL queries. This perfectly integrates with the web-frontend and the graph visualization, providing quick interactive response time and good scalability.

5 Features

PathBank is a web-based pathway query and visualization system. A number of features are available to aid the

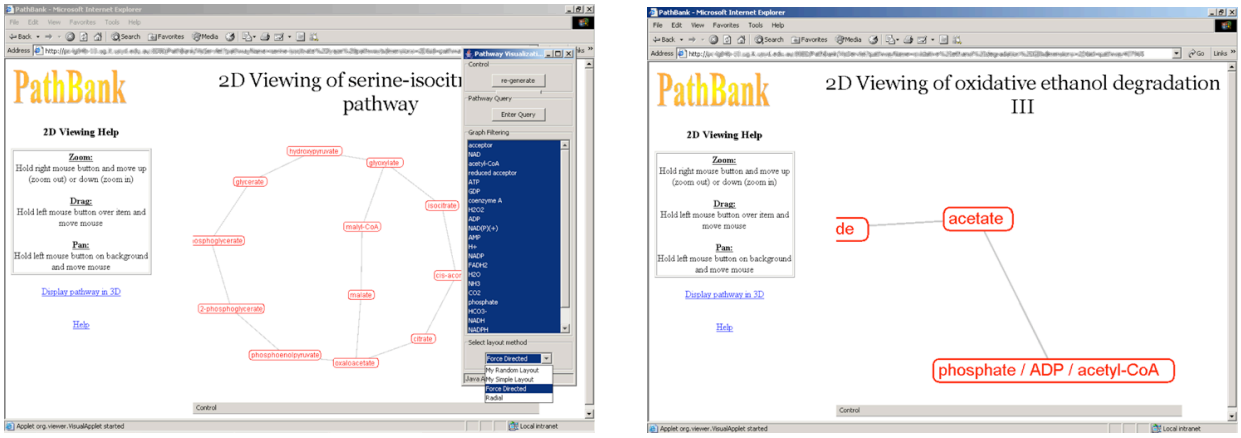

Figure 3: Visualization of Biological Pathway. A Java applet is embedded in the web page. The pathway can be interacted using mouse. Layout style (left) and chemicals to be filtered (right) can be controlled from a separate panel.

understanding of biological pathways:

- Google-like web search interface: On top of the user view, PathBank also adopts the use of identifiers (e.g. PID:name:kinase for all proteins which have name containing string *kinase*). Boolean operations on search entities (e.g. AND, OR ,NOT) are also possible.
- DHTML search assistant to help construction of complex queries.
- Expandable search result page: Matching entities are grouped by their type (e.g. Protein, Gene, Gene Ontology, Small molecules). AJAX technology was used to fetch data dynamically (see Figure 1).
- Selected pathway can be visualized in two or three dimensions: We make use of open source Java packages prefuse [8] and WilmaScope [4] for automatic graph drawing in 2D and 3D respectively (Figure 2).
- Commonly occuring small molecules (such as H_2O, H^+, NADP, ATP, and Acetyl-CoA) can be filtered from the graph to simplify the view and reduce visual complexity.
- Simple interaction and navigation methods: The pathway can be translated, rotated, dragged and zoomed by mouse actions.
- A number of well known graph layout methods are made available to users.
- New visualization and analysis plug-ins can be incorporated into the visual applet easily. This makes PathBank extensible.
- A flexible API for fast and reliable access to the integrated database.
- Integrated relational database using BioPAX schema.

6 System Design and Implementation

The targeted users of PathBank are mainly biologists who may not have a lot of experience in complex bioinformatics tools. These users prefer simple user interface that allows searches to be performed quickly and reliably. As PathBank is currently a proof of concept, research prototype system, features offered by PathBank are still limited. We, however, envisage this web tool to incoroporate data mining and network analysis capabilities in the future. As a result, the main design goals of PathBank are user-friendliness, flexibility and extensibility. Open source technologies are used extensively in order to meet short development time and low cost. The overall three-tier architecture is illustrated in Figure 4.

The database backend consists of a relational database storing data from multiple sources. The schema of this relational database can be automatically generated via web ontology language (OWL) of BioPAX (version 1.4 was used in our work). The ability of automatically generating database schema from ontology is a great way to respond to the rapidly changing BioPAX standard. PostgreSQL database version 8.0 was used as a RDBMS. Material views are constructed to aid database access via the API.

The application programming interface (API) acts as a central access point to the database. The API is written in Java and is further divided into the server-side and client-side components. The server-side component is responsible for validating queries, processing queries (i.e.

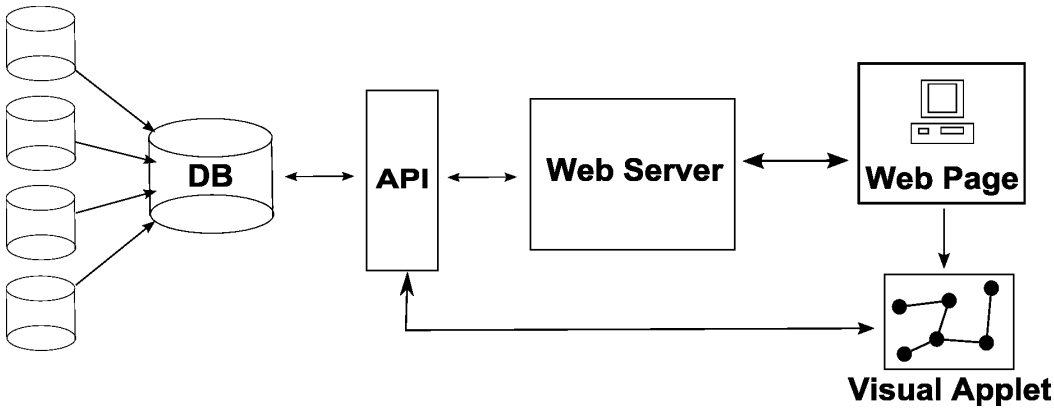

Figure 4: System architecture of PathBank. The relational database at the backend stores data imported from multiple sources via the BioPAX format. Access to the database is managed by the Application Programming Interface (API) at the middle tier. The web server generates web pages that embeds a Java applet for pathway visualization. The Java applet directly communicate with the API to gain access to the pathway graph from the database.

decomposing complex queries), constructing SQL statements, processing returned results, and generating pathway graphs. Java Database Connectivity (JDBC) and Apache Common DBCP are used to implement connection pooling between API and the database. The client-side carries out simple query validation, and acts as a wrapper for the underlying communication channel to the server-side component. Java Remote Method Invocation (RMI) is used for communication between server- and client-side components.

In our relational database, pathways are represented as a collection of chemical reactions, and each reaction is related to two sets of chemicals (i.e. reactants and products). By joining related reactions together, we are able to generate biological pathway from relational data.

The web-frontend is implemented using a Model-View-Controller (MVC) architecture. It makes use of the JavaServer Pages (JSP) and Apache Struts framework. The advantage of using MVC is to allow the web-frontend to be extended easily in the future to incorporate new pages and functionalities. The web application is currently deployed on an Apache Tomcat server. In addition, we also make extensive use of the web-browser power to implement the dynamic behaviour of our system. In particular, dynamic data fetching is implemented by using Asynchronous JavaScript and XML (AJAX) technology. It enables query result tree to be retrieved at a level-by-level basis. This avoids the long loading time for large amount of results. DHTML and JavaScript are used to build the search assistant. Layout of the web pages are controlled by a cascading style sheet (CSS).

The visualization engine of PathBank is implemented as a reasonably sized Java applet. We make use of two open source visualization libraries, *prefuse* [8] for 2D visualization and *WilmaScope* [4] for 3D visualization. This allows high quality interactive visualization to be incorporated into the system in the least amount of time. Our system makes use of the applet caching mechanism to ensure the applet is preloaded and is only reloaded if there is an update.

An example of a movie which shows the functionalities of PathBank system is available from *http://www.ug.it.usyd.edu.au/∼tman0127/PathBank/doc/*.

7 Conclusions and Future Work

PathBank is a web-based biological pathway query and visualization system using an integrated database that supports systems-level investigation of biological data.

PathBank uses relational database technology in its backend to provide fast and efficient access to pathway graph data. To support the integration of different biological data sources, such as KEGG or BioCyc, that provide data in the ontology-based BioPAX exchange format, we developed a database creation tool, called Genea. Genea automatically maps and loads ontology data into a relational database using generic mapping rules.

PathBank provides an intuitive, Google-like web-interface for database querying and browsing. The pathway visualization capability is embedded in a Java applet. PathBank makes extensive use of client side's technology to reduce computational load of the server. It also makes extensive use of open-source technology. PathBank can be extended by adding new visualization and network analysis plugins to the system. Network analysis and data mining tools can be added to the system, to complement the pathway visualization.

References

[1] Becker M, Rojas I (2001) A Graph Layout Algorithm for Drawing Metabolic Pathways. Bioinformatics, 17, 461-464.

[2] Breitkreutz BJ, Stark C, Tyers M (2003) Osprey: a network visualization system, Genome Biology, 4:R22.

[3] Di Battisa G, Eades P, Tamassia I, Tollis I (1999) Graph Drawing: Algorithms for the visualization of graphs, Prentice Hall.

[4] Dwyer T, Eckersley P (2003) Graph Drawing Software, Mathemetics and Visualization, Springer, chapter WILMASCOPE - a 3D graph visualization system, 55-76.

[5] Endy D, Brent R (2001) Modelling cellular behaviour, Nature, 409, 391-395.

[6] Forster M, Pick A, Raitner M, Schreiber F, Brandenburg FJ (2002) The system architecture of the BioPath system. In Silico Biology, 2(3), 415-426.

[7] Gopalacharyulu PV, Lindfors E, Bounsaythip C, Kivioja T, Yetukuri L, Hollmen J, Oresic M (2005) Data integration and visualization system for enabling conceptual biology, Bioinformatics, 21, i177-i185.

[8] Heer J, Card S, Landay J (2005) prefuse:a toolkit fr interactive information visualization. CHI2005. Portland, Oregon, USA.

[9] Hu Z, Mellor J, Yamada T, Holloway D, DeLisi C (2005) VisANT: data-integrating visual framework for biological networks and modules, Nucleic Acids Research, 33, doi:10.1093/nar/gki431.

[10] Kanehisa M, Goto S (2000) KEGG: Kyoto Encyclopedia of Genes and Genomes, Nucleic Acids Research, 28, 27-30.

[11] Kitano H (2002) Computational Systems Biology, Nature, 420, 206-210.

[12] Kraska, T. and Roehm, U.: Genea: Automated Mapping of Ontologies into Relational Schema. Submitted for publication, March 2006.

[13] Longabaugh WJR, Davidson EH, Bolouri H (2005) Computational representation of developmental genetic regulatory networks, Developmental Biology, 283, 1-16.

[14] Luscombe NM, Babu MM, Yu H, Snyder M, Teichmann SA, Gerstein M (2004) Genomic analysis of regulatory network dynamics reveals large topological changes, Nature, 431, 308-312.

[15] Saraiya P, North C, Duca Karen (2005) Visualizaing biological pathways: requirements analysis, systems evaluation and research agenda, Informationa Visualization, advanced online publication, doi:10.1057/palgrave.ivs.9500102.

[16] Schreiber F (2002) High quality visualization of biochemical pathways in BioPath. In Silico Biology, 2(2):59-73.

[17] Schwikowski B, Ideker T, Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N (2003) Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks, Genomoe Research, 13, 2498-2504, doi:10.1101/gr.1239303.

[18] Sugiyama K, Tagawa S, Toda M (1981) Methods for visual understanding of hierarchical system structures, IEEE Transactions on Systems, Man, and Cybernetics, 11(2):109-125.

[19] Ware C, Franck G (1994) Viewing a Graph in a Virtual Reality Display is Three Times as Good as a 2-D Diagram. IEEE Conference on Visual Languages.

[20] Yang Y, Engin L, Wurtele ES, Cruz-Neira C, Dickerson JA (2005) Integration of metabolic networks and gene expression in virtual reality, Bioinformatics, 21, 3645-3650.