Genomic Pathways Database and Biological Data Management

Z. M. Ozsoyoglu[1,2], G. Ozsoyoglu[1,2], and J. Nadeau[1,3]

[1] Center for Computational Genomics, Case Western Reserve University (CWRU)
[2] Department of Electrical Engineering and Computer Science, CWRU Case School of Engineering
[3] Department of Genetics, CWRU School of Medicine

Summary

In this paper we discuss the properties of biological data, and challenges it poses for data management, and argue that, in order to meet the data management requirements for "digital biology", careful integration of the existing technologies and the development of new data management techniques for biological data are needed. Based on this premise, we present PathCase: Case Pathways Database System. PathCase is an integrated set of software tools for modeling, storing, analyzing, visualizing, and querying biological pathways data at different levels of genetic, molecular, biochemical and organismal detail. The novel features of the system include: (a) genomic information integrated with other biological data and presented starting from pathways, (b) design for biologists who are possibly unfamiliar with genomics, but whose research is essential for annotating gene and genome sequences with biological functions, (c) database design, implementation and graphical tools which enable users to visualize pathways data in multiple abstraction levels, and to pose exploratory queries, (d) a wide range of different types of queries including, "path" and "neighborhood queries", and graphical visualization of query outputs, and, (e) an implementation that allows for web(XML)-based dissemination of query outputs (i.e., pathways data in BIOPAX format) to researchers in the community, giving them control on the use of pathways data.

Keywords: metabolic pathways**,** database system**,** biological data management, path queries, neighborhood queries, graphical visualization of pathways.

Biology and Life Sciences have become increasingly "data rich" over the past decade. Due to investments on public and private resources, there is a rapid growth of distributed and heterogeneous biological data; significant advances in data generation, storage, analysis, web-based availability, and sharing technologies; and the emergence of large-scale biological data gathering technologies. There are also significant investments in developing large biological information resources, and assembling this information into public databases. Many such resources and tools are available, including NCBI's Genbank, PubMed, Blast, and MGI's tools and databases. And, continued explosive growth in the amount and the diversity of biological and biochemical data is expected into the century. Finally, medical informatics and physiological data is also very large, diverse, non-standard, and distributed, and growing at a fast pace.

Biological data has three important characteristics: (i) complexity, (ii) heterogeneity, and (iii) highly dynamic data and schema. Biological data is *complex* in the sense that it is very rich in metadata, and requires metadata management techniques. It has inherently deeply-nested hierarchical structures (e.g., ontologies), some best modeled as graph structures at the conceptual level (e.g., metabolic pathways, or signaling pathways). Biological data is *heterogeneous* in the sense that it involves a wide array of data types, including text, image, sequence data, as well as streaming data (e.g., medical sensors data), temporal data, and incomplete and missing data. It is also gathered from heterogeneous sources and is in different formats.

Biological data is highly *dynamic*, not only in content, but also in schema (i.e., structure). Data management techniques can effectively handle dynamic data content, but a highly dynamic schema poses challenges for data management, and applications and software tools that employ database schema need to be frequently revised. As also discussed in the literature [NIH03], using off-the-shelf data management software tools are not sufficient for data management needs of "digital biology", and careful integration of the existing technologies for biological data as well as the development of new data management techniques are needed.

Figure 1. Pathway-Level Visualizations. Rectangles represent pathways, circles represent connecting metabolites

The three characteristics of biological data, namely, complexity, heterogeneity, and dynamic schema, are taken into account in the design of *PathCase: Case Pathways Database System*, an integrated software tool for storing, visualizing, querying, and analyzing, biological pathways at different levels of genetic, molecular, biochemical and organismal detail (http://nashua.case.edu/pathways). In the rest of this paper, we summarize the pathways data model, querying paradigms, capabilities, and the architecture of PathCase, along the way illustrating how we have attempted to handle issues resulting from pathways data complexity, heterogeneity, and dynamic schema.

Data Model: Main objects of the PathCase database are *pathway*, *process* and *molecular entity*. Pathway is an interconnected arrangement of processes, representing functional roles of genes in the genome. Process is a reaction (or step) in a pathway possibly involving a gene product (e.g., a catalyzing enzyme for enzymatic reactions). (Substrates, products, co-factors, inhibitors, activators of a reaction are all molecular entities in this perspective). Molecular entity is the general name given to any entity participating in a process, such as a basic molecule, protein, enzyme, gene, amino acids, etc. Pathcase, also maintains information on genes, EC numbers, organisms, GO annotations, synonyms. Gene, while in our current design considered only as a specialization of molecular entity, is an important object in the PathCase database, and the next version of PathCase will have gene as a more central object, i.e., another entry point for querying the database, at the same level (but orthogonal to) Pathways, Processes and Molecular entities. Thus, the pathways data is highly complex (in the sense that it involves complex objects with set-valued and composite attributes, several specialization/generalization hierarchies, and is graph-structured) and dynamic (in the sense that the data evolves as more is learned about the functionality of genome). And, we are finding out that we have to routinely revise the schema of our data due to the additions of new objects and/or new applications requiring changes to the schema.

For the purpose of controlling the visualization complexity and heterogeneity of pathways data, we provide *views of pathways at different abstraction levels*. At the *molecular entity level*, pathways are represented in the form of graphs (more specifically, hypergraphs) where nodes are substrates/products (metabolites), and hyper-edges are processes (reactions). At the *pathways level*, pathways are represented as (composite) nodes, and two pathway nodes are connected by edges if there are molecular entities that are shared by the two pathways. (Actually, we also represent these shared molecular entities connecting patways as (simple) nodes in the graph representation). Pathway nodes in this graph can be expanded to view the details of the pathway so that the user can move from one abstraction level to another. At the *GO ontology level*, GO annotations of enzymes in pathways, and the relevant GO sub-hierarchies are represented as graphs, together with the relevant GO term annotation statistics. We are presently working on providing the GO-functionality-based *pathway template level*, where pathways are abstracted into GO-functionality templates, and visualized for users to perform pathways data mining and pathway (fragment) inferencing. While the PathCase database is maintained as a relational database, pathways querying is provided over graph-structured conceptual views.

Database content: Presently, the PathCase database contains 39 metabolic pathways, 37 from [M99], and two (Folate and Homocystine pathways) for human and mouse, provided by Joe Nadeou and Toshimori Kitami. The pathways from [M99] are uploaded to the database using the Pathways Editor tool [PE04]. At the present time, we have 876 processes for different organisms or organism classes, which are human, mouse, animals, prokarya, plants & yeasts, and unspecified.

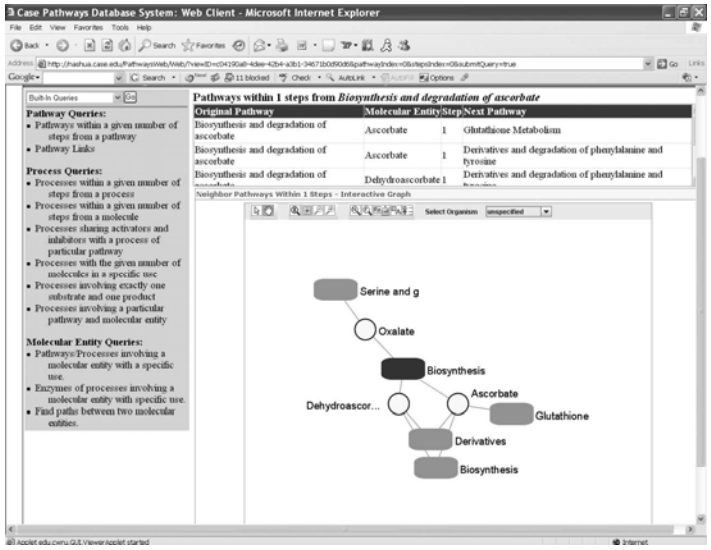

Figure 2. Connected pathways in the neighborhood of a given pathway query . Output is provided both in tabular and in graphical forms.

Web-based Pathways Query and Visualization Sub-System Architecture: PathCase system has four basic subsystems. The *PathCase Desktop Client*, our first-generation system, runs at the client site, and is a stand-alone, rich desktop application for browsing, querying and interacting with the PathCase database; its deployment at the client site requires the .Net framework and a visualization software [TS] (to perform the visualizations at the client side) [K+03], a significant effort on the part of the users to set up the software environment. As a result, we have built a web-based version of PathCase, our second generation system, which can be used through a web browser.

The web-based PathCase version is built around two subsystems, namely, the PathCase Server and the PathCase Web Client. The *PathCase Server* runs on a server at Case Western Reserve University, and is responsible from maintaining the PathCase database as well as providing web services. The PathCase web services allows us to (a) handle data and schema changes in pathways data gracefully, and to (b) add, new visualization levels for users as needed. When PathCase database is changed or upgraded, the software changes are limited to changes to the functions provided by the PathCase Web services. And, different visualizations of pathways data to users can be provided by changing (a) the pathway visualization code and (b) the PathCase Web Services components of the PathCase Web Server.

The *PathCase Web Client* is a web-based online interface to the PathCase database, written in Java and used from within a web browser, and requires only the Java plug-in to the web browser in order to expose PathCase functionalities to users.

The fourth PathCase subsystem, *Pathways Editor*, allows for the creation, modification, and addition of new pathways, processes, and molecular entities into the PathCase database [PE04]. This paper only summarizes the features of PathCase Web Client, and its interactions with the PathCase Server.

The present functionalities of PathCase Web Client include: (a) browsing and visualization of the PathCase database entities, namely, pathways, processes, and molecular entities (basic molecules, proteins, and genes), (b) various built-in queries about pathways (that provide different levels of pathway visualizations), processes, and molecular entities, (c) posing a large class of queries dynamically constructed during a user session (i.e., *dynamic querying*) using query forms for parametrized queries, and the *Advanced Query Interface*, (d) visualizing the results at different levels, and multiple forms (i.e., both as a graph, and in tabular form), (e) allowing users to pose queries seamlessly on the query results, and hence providing an environment for "exploratory" querying, and (f) connecting the enzymes in the PathCase database to Gene Ontology (GO) terms and vice versa, as well as pathway-centric or GO-centric visualizations of results, using the *Ontology Viewer*. To achieve these functionalities, various XML objects (SOAP XML, Graph XML, and HTML documents) are exchanged between the PathCase Server and the PathCase Client.

The PathCase Web Client does not directly communicate with the database; instead, it requests and gets its data using PathCase Web Services. This allows for a graceful extension of PathCase capabilities (i.e., handling the complexity, heterogeneity, and schema dynamics of PathCase data in a flexible manner). At

the present time, we have about 60 PathCase web service functions, adding new functions with each new capability incorporated into the PathCase system.

One of the novel features of the PathCase Web Client is the way in which client-side PathCase visualizations are materialized. If a given PathCase query needs a visualization, the PathCase Server prepares a *graph XML object*, specifying the needed visualization with explicit layout specifications, and

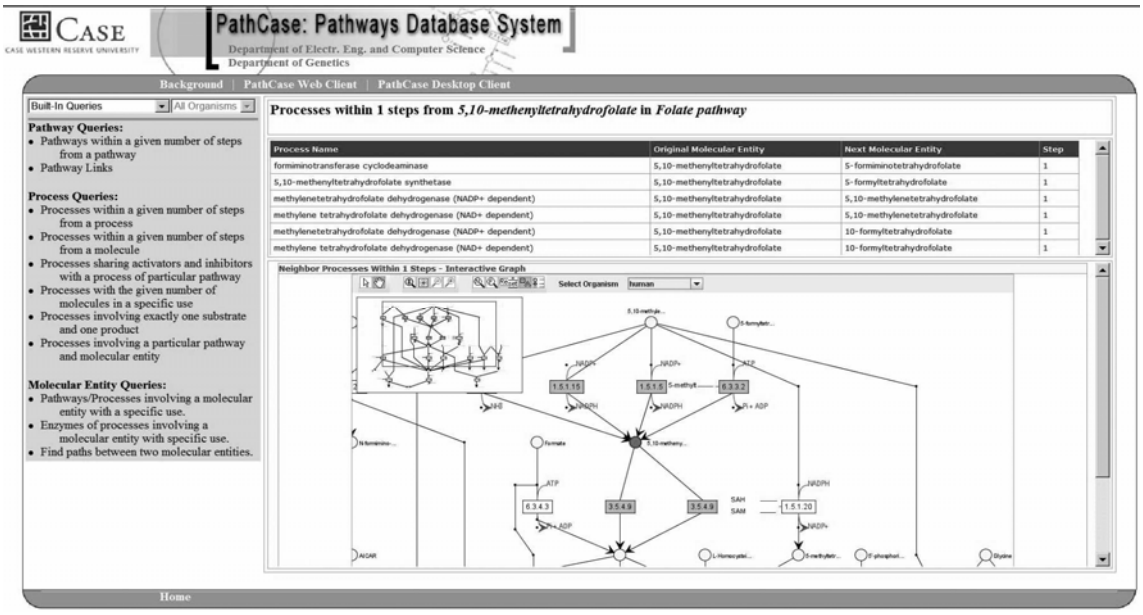

Figure 3. Processes in the neighborhood of a given process. Query output is given in tabular and graphical forms.

passes the object to the PathCase Client. The PathCase Client then caches the graph XML object at the client site, and calls its *PathCase Java Viewer* to parse the graph XML object and to simply regenerate the visualization from the cached graph XML object. The main time cost here is only the transmission of the graph XML object, not the client-side re-generation of the visualization (which is very fast). After the caching of the graph XML object, all subsequent re-visualizations (e.g., zooming-in, zooming-out, re-coloring of objects, etc) only employ the cached graph XML object, do not need any communication with the PathCase Web Server, and thus are very fast. Moreover, once a pathway is visualized (regardless of the current user session, until the user closes the web browser), the visualization stays in the local cache, and is always fetched from the local cache, making visualizations very efficient.

Other pathways systems and resources on the web: A comprehensive list of Pathways databases and systems along with other databases that are of value to the biologist are given in [Ga 2005]. Here we briefly discuss the pathways databases that we consider are the most relevant to PathCase. Reactome is a curated resource of core pathways and reactions in human biology [Re], and uses a frame-based knowledge representation model. Reactome model's key concepts (data classes) include physical entity (molecules, complexes, sets of molecules), CatalystActivity (molecular functions and involved PhysicalEntities), and events (reactions and pathways). The most recent version of Reactome Pathways are dynamically visualized (i.e., constructed from a database on the spot), and Reactome includes a search tool, named Pathway Finder, that allows limited dynamic querying capabilities for pathways. Kegg [Kegg], a significant pathways data source on the web, provides extensive visualizations of pathways with EC numbers (sometimes, referring to a family of EC numbers) and links to related pathways. Kegg visualizations are static (i.e., pre-generated images), and do not support dynamic querying. Kegg does not provide co-factor, activator, inhibitor, regulator information. Recently, Kegg pathways have been made available in XML (and HTML).

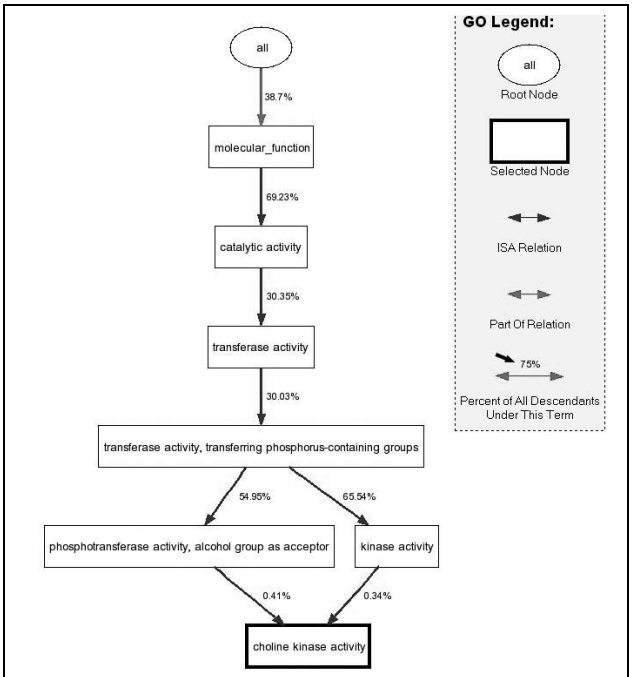

Figure 4. GO Term Hierarchy for Annotation choline kinase activity of Enzyme choline kinase as output of a parametrized query which is activated from the process node with the specified enzyme on a pathway graph.

BioCyc [BC] is a collection of pathway databases, namely, EcoCyc (*Escherichia coli* K12) [EC], MetaCyc (Metabolic pathways and enzymes from 300 organisms) [MC], and has Pathway Tools software. BioCyc has a query page that provides several different mechanisms for querying Pathway (and Genome) Databases. Pathway Tools software allows users to observe pathway variations in different organisms. However, pathways graphs of MetaCyc and EcoCyc, once generated, are static, and there are no dynamic querying capabilities. PATIKA [PT] provides pathway visualization and editing tools with an ontology for efficient querying, and a microarray data analysis component to help relate expression data to pathways. PATIKA has dynamic pathway visualization, and recently introduced dynamic querying capabilities. Cytoscape [Cyto] is an open source bioinformatics software platform for *visualizing* molecular interaction networks and *integrating* these interactions with gene expression profiles and other state data. Cytoscape has dynamic pathway visualization, but not querying, tools. BioCarta [BioC] source provides HTML-formatted static biological pathway charts, and graphical pathways of BioCarta do not have user-manipulated multiple abstractions, nor are they queriable using graphical querying features. The remaining pathways sources on the web, including ExPASy (Expert Protein Analysis System), Cell Signaling Networks Database, Enzymes and Metabolic Pathways, Metabolic Pathways of Biochemistry, Roche Apoptosis Pathway, The University of Minnesota Biocatalysis/Biodegradation Database, Soybean metabolic pathways, Nicholson minimaps web source have no dynamic pathway querying and drawing capabilities.

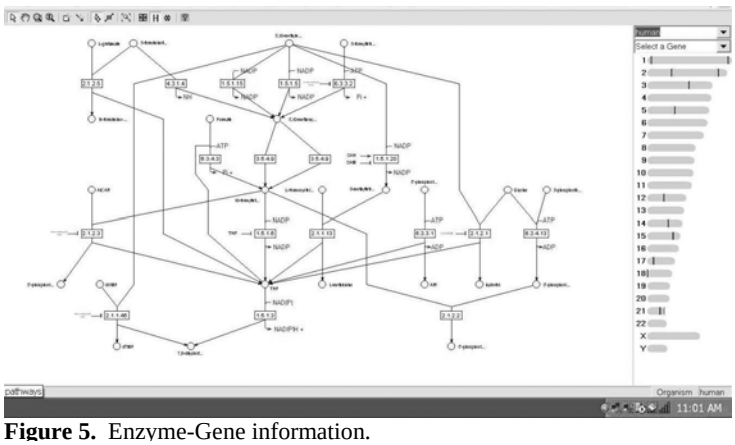

Figure 5. Enzyme-Gene information.

Exploratory Querying and Visualization: PathCase provides several functionalities for exploratory querying pathways including: (a) viewing whole network of pathways available in the PathCase database (Figure 1), (b) viewing at multiple levels of abstractions, i.e., at the level of processes (Figure 3), at the level of pathways and the molecular entities that connect those pathways (Figure 2), and at the level of the GO term hierarchies associated with the GO annotations of enzymes in pathways (Figures 4 and 6), (c) querying specific properties of any pathway component at any level of granularity, (d) path queries, (e) neighborhood queries (Figure 3), (f) different forms of queries and output displays, namely, textual versus graphical queries and outputs (Figure 3, and 5), built-in and parameterized queries, tabular versus graphical query outputs, and the Advanced Query Interface [NO04] .

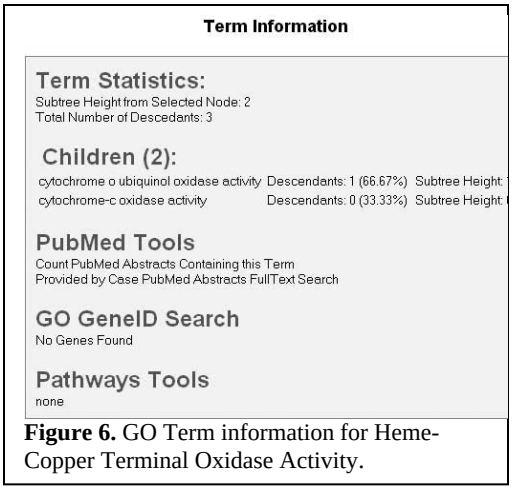

Figure 6. GO Term information for Heme-Copper Terminal Oxidase Activity.

Data Management and Path Query Evaluation for Pathways Analysis: PathCase database is relational, with users provided with an object-relational representation of database objects. XML is currently used as the data exchange format between the PathCase Server and the PathCase client. The users can download and save visualized pathways in BioPAX-formatted XML documents. Using XML as the storage format for pathways data is being explored. We consider using encoding, labeling schemes for the structure of XML documents, for efficient evaluation of queries [AEO 05]. While such encoding schemes improve the performance of path query evaluation traversing the nested structure of the data, there is not much research on the efficient evaluation of graph traversal queries on pathway graphs, such as neighborhood queries. Since exploratory queries typically have multiple steps and use the query results in a previous step to specify the query in the next step, there is a great potential to improve the query performance by cashing the results of previous queries, and utilizing them by query rewriting techniques instead of recomputing the result at each step from the database [XO05].

Acknowledgment: This research has been partially funded by a gift from Charles Wang Foundation and NSF research grant DBI-0218061.

References

[AEO05] Akgul, S.f, Elliott, B., Ozsoyoglu, Z.M., " ToPP-Labeling: An Insert Friendly Labeling for XML Query Processing", Technical Report, EECS Dept., CWRU, 2005.

[BC] BioCyc, available at http://www.biocyc.org

[BioC] BioCarta Web site, available at http://www.biocarta.com

[Cyto] Cytoscape, available at http://www.cytoscape.org

[EC] EcoCYC: Encyclopedia of Escherichia coli K12 Genes and Metabolism, available at http://ecocyc.org

[ExPASy] Digitized versions of Boehringer Mannheim Biochemical Pathway Charts, available at http://www.expasy.ch/cgi-bin/search-biochem-index

[Ga05] Galperin, M. Y., The Molecular Biology Database Collection: 2005 update, Nucleic Acids Research, 2005, Vol. 33, Database Issue, D5-24.

[K+03] Krishnamurthy, L., et al, "Pathways Database System: An Integrated Set of Tools for Biological Pathways", *BioInformatics*, Vol. 19, No. 8, pp. 930-937, 2003.

[M99] Michal, G., *Biochemical Pathways*, John Wiley & Sons Inc., 1999.

[MC] MetaCyc, available at http://metacyc.org

[NIH03] NIH Digital Biology 2003, NIH BISTI workshop on Digital Biology, http://www.bisti.nih.gov/2003meeting/

[PT] PATIKA, available at http://www.patika.org

[Re] Reactome - a knowledgebase of biological processes, at http://www.reactome.org

[Kegg] Kyoto Encyclopedia of Genes and Genomes, available at http://www.genome.ad.jp/kegg

[NO04] Newman, S., and Ozsoyoglu, Z.M., "A Tree-Structured query interface for Querying Semi-Structured Data", SSDBM Conf., 2004.

[PE04] "Pathway Editor Tool--Release #2", User Manual, Computational Genomics Center, EECS and Genetics Depts, CWRU, July 2004, available at http://nashua.case.edu/PathwaysWeb/Desktop/UserManuals.aspx

[XO05] Xu, W. and Ozsoyoglu, Z.M., "Answering Queries Using Materialized XPath views", VLDB 2005.

[TS] Tom Sawyer Software, http://www.tomsawyer.com