A Unified Object Database for Biochemical Pathways

T. S. Jung[1], J. S. Oh[2], H. K. Jang[2], M. S. Ahn[2], D. H. Roh[3], and W. S. Cho[2]
[1] *Dept. of Information Industrial Engineering, Chungbuk National University, Choungju, Chungbuk, Korea*
[2] *Dept. of Management Information Systems, Chungbuk National University, Choungju, Chungbuk, Korea*
[3] *Schools of Life Science, Chungbuk National University, Choungju, Chungbuk, Korea*
Email: {mispro, ofang, lodeston, epita55, dhroh, wscho}@chungbuk.ac.kr

ABSTRACT One of the most important issues in post-genome era is identifying functions of genes and understanding the interaction among them. Such interactions form complex biochemical pathways, which are very useful to understand the organism system. We present an integrated biochemical pathway database system with a set of software tools for reconstruction, visualization, and simulation of the pathways from the database. The novel features of the presented system include: (a) automatic integration of the heterogeneous biochemical pathway databases, (b) gene ontology for high quality of database in the integration and query (c) various biochemical simulations on the pathway database, (d) dynamic pathway reconstruction for the gene list or sequence data, (e) graphical tools which enable users to view the reconstructed pathways in a dynamic form, (f) importing/exporting SBML documents, a data exchange standard for systems biology.

1. Introduction

Systems biology is a synergistic application of experiment, theory, and modeling towards understanding biochemical processes as a whole system instead of isolated parts [4].

Biochemical processes can be represented by biochemical pathways, which are being accumulated in many databases. There are three kinds of biochemical pathways: *metabolic pathways, gene regulation pathways, and signal transduction pathways*. Metabolic pathways are responsible for carrying out the chemical reactions that provide basic biochemical functions such as DNA, RNA, protein synthesis and degradation, energy metabolism, fatty acid synthesis, etc. Gene regulation pathways are responsible for converting genetic information into proteins (gene products). Signal transduction pathways coordinate metabolic processes having transcription and protein synthesis.

In the previous approaches, each of these pathways having distinct attributes, bug representing the same phenomena of life kept and managed in a separate database.

In systems biology, biochemical pathway databases and tools are very important because they support virtual simulation of the life with in a computer. Pathway databases raise many important and challenging computational and bioinformatics issues, such as (1) querying and visualization of pathway database (2) seamless integration of different kinds of data distributed in diverse sources (3) graph-based querying and navigation.

Well-known systems such as KEGG [7], BioCyc [5], and WIT[OV02] have predefined reactions in the fixed metabolic maps, and thus support static retrieval to the pathway database. Recent systems such as PathBlazer [17], public domain systems [8] [10][13], Cytoscape [SH03], Osprey [1], and TopNet[14] support dynamic retrieval and visualization of metabolic pathways.

Note that separated database are used in these systems; actually each of the previous system treats just on kind of pathway.

Recently, developing a unified biochemical pathway database integrating three kinds of pathways is becoming an important issue [8]. SBML (System Biology Markup Language) designed for the exchange of all kinds of biochemical pathway data in a standard form [18], requires a unified pathway database. The unified pathway database can be a basic infrastructure for E-Cell [20]. E-Cell is aiming to model and reconstruct biological phenomena *in silico*.

In this paper, we present a unified biochemical pathway database supporting SBML data model. SBML is designed for the exchange of biochemical pathway data in a standard form. Our system can store and manage three kinds of pathways by importing and exporting SBML documents. For this, mediators and SBML converters for the object database have been developed. The result of query(a part of the pathways) can be visualized dynamically and the users can edit and simulate the pathways by using tools. The system dynamically reconstructs expected pathways from gene list or sequence data. For precise data integration and query result, we use gene ontology (GO) [3] which allows standard description of gene products and genome annotation with consistent terminologies.

The paper is organized as follows. In Section 2, we discuss the SBML and related work. In Section 3, we present the system architecture. In Section 4, we describe the novel features of the system. In Section 5, we present conclusions and future work.

2. Related Work

In this section, we explain SBML, pathway database systems, gene ontology, and Orthologs that are used in our system, in more detail.

2.1 SBML

The SBML [18] developed by the systems biology community is a free and open language for biochemical reaction networks in systems biology. It was started in the year 2000 and the first release was published in 2001. It uses XML as its description language and UML for modeling the components. Recently, existing biochemical database KEGG [5] support KEGG2SBML to convert metabolic pathway database the SBML format

2.2 Pathway database systems

A wide variety of systems exist for the storage and display of pathway information. Here, we review several of them. KEGG[5] is the most famous metabolic pathways database. While KEGG uses static approaches for querying data visualization, its value lies primarily in the breadth of its content coverage. The benefits of the first-generation systems such as WIT [OV00], MPW [12], and EcoCyc [5], which perform limited dynamic querying and visualization. General purpose systems for pathway rendering include BioJake [11], PathDB [6], PathFinder [7], Pathways Database System [8], PaVESy [15], VitaPad [9]. The BioJake Website is not operational currently: the paper indicates that the system was, at best, an early prototype. PathDB's data model, while rich, is not user-extensible: this limitation also applies to the Pathways Database System. Data model of PathFinder, PaVESy, and VitaPad is restricted for integration with other data sources.

2.3 Gene Ontology [3]

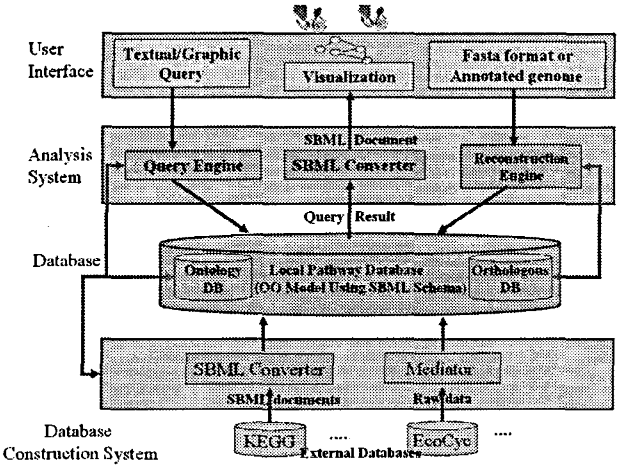
Biologists currently waste a lot of time and effort in searching for all of the available information about each small area of research. This is hampered further by the wide variations in terminology that may be common usage at any given time, and that inhibit effective searching by computers as well as people. For example, if you were searching for new targets for antibiotics, you might want to find all the gene products that are involved in bacterial protein synthesis, and that have significantly different sequences or structures from those in humans. But if one database describes these molecules as being involved in 'translation', whereas another uses the phrase 'protein synthesis', it will be difficult for you - and even harder for a computer - to find functionally equivalent terms. The Gene Ontology (GO) project is a collaborative effort to address the need for consistent descriptions of gene products in different databases. The project began as a collaboration between three model organism databases: FlyBase (*Drosophila*), the *Saccharomyces* Genome Database (SGD), and the Mouse Genome Database (MGD) in 1998. Since then, the GO Consortium has grown to include many databases, including several of the world's major repositories for plant, animal and microbial genomes.

2.4 Orthologs

The ancestral genes have been distributed in each organism's genome by speciation and duplication for the biochemical evolution process. Therefore each organism has biochemical same function and specialized function. Genes in two species that have directly evolved from a single gene in the last common ancestor are called Orthologs. The genes in same Orthologs are most likely to share the function and similar sequence[16]. So Orthologs are useful to predict automatically the function of unknown gene in incomplete genome and to study the gene having the same function in each organism [16].

3. System Architecture

In this section, we describe the system architecture which integrating three kinds of biochemical pathways in a database. The architecture of the system has four layers, as shown in <Figure 1>.



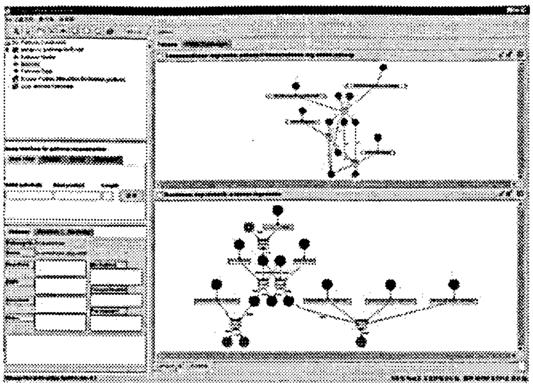
<Figure 1> System Architecture.

(a) **Database Construction System (DCS)**: DCS constructs a unified integrated database from external databases. It has two tools – the SBML Converter and Mediator - for the integration. SBML Converter transforms SBML documents from external sources such as KEGG into the local object database. Mediator receives various kinds of biochemical data and converts them into the local object database. We have developed several mediators for various external data source. Gene ontology, an essential tool for heterogeneous bio data integration, is used in the database construction and query for accuracy.

(b) **Local Database**: It consists of a biochemical pathway database, an ontology database, and an orthologs database. An SBML-based object database is used for storing various biochemical pathway data. Predefined or ad-hoc queries are provided for biochemists.

(c) **Pathway Simulation System** : It provides query processing, dynamic pathway reconstruction, and simulation. Pathway reconstruction can be done either by gene list or sequence data; In the case of non-annotated sequence data, the reconstruction engine uses orthologs database for increased precision. Simulation system is useful to the biochemists who want to know the quantitative changes by time in the biochemical pathways. As the biochemists modify the initial quantities and the kinetic laws for the pathways, corresponding results are displayed in the graphical forms.

(d) **Graphical User Interface** : It supports user-friendly queries and result visualization. <Figure 2> shows the graphical user interface.



<Figure 2> Graphical User interface.

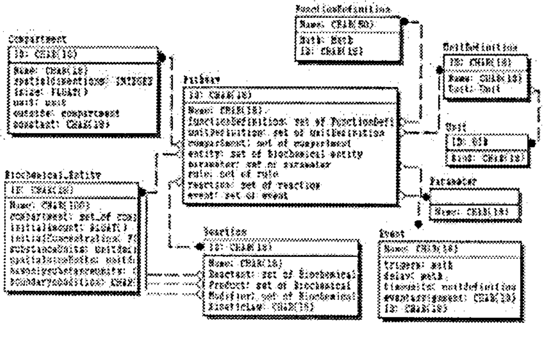
4. Novel features of the system

In this section, we present the novel features of the system in detail. The most important feature is the integration of three kinds of biochemical pathways into a unified object databases. From the unified database, various applications on metabolic, signal transduction, and gene regulation pathways can be made by the biologist.\

4.1 Integrated data model

A whole biochemical pathway is composed of signal transduction, metabolic and regulatory pathways. Each pathway has its own properties, so databases have been constructed for storing each pathway respectively even though they have common properties. But in systems biology, we have to integrate all these pathways into one database with coherence and then we can analyze an organism as a whole thing and manage all pathway-related data. For these purposes, we need an integrated data model which defines all entities in three pathways and stores efficiently them.

In our proposal, we use object-relational model because it integrates heterogeneous pathway data in an efficient way. Also we use SBML(Systems Biology Markup Language) that represents all pathway entities in an XML based object schema is used in our system. Figure 3 shows integrated pathway database schema. It consists of 9 classes including pathway, biochemical_entity, compartment, reaction, etc. Note that these classes are used for representing integrated three kinds of biochemical pathways.



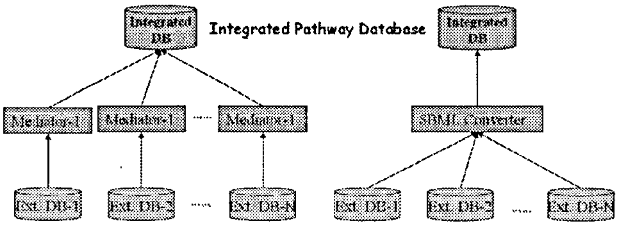
<Figure 3> Abstraction of integrated pathway database schema

4.2 Data integration: SBML Converter and Mediators

To integrate various types of pathway data in a unified object database, data integration tools are indispensable. We devised mediators for several external databases (KEGG, Ecocyc, dbEST, UnigeneDB, etc.). The mediators parse the flat files then convert them into the object database schema.

Another useful integration tool is the SBML Converter. Since SBML is a de facto standard format to exchange biochemical pathways, the SBML Converter is very useful to import/export SBML documents via the integrated object database. The well-known pathway databases such as KEGG starting to export their contents in the SBML format.

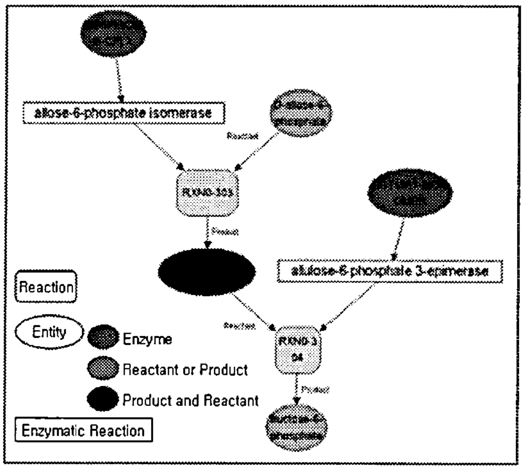
Figure 4 shows the difference between mediator and SBML Converter in the integration of the heterogeneous pathway databases.



<Figure 4> Difference between Mediator and SBML Converter

Currently, biochemical pathway databases are exists in on-line. But, each databases are not contains all type pathways. So, to analysis the all of biochemical pathways must to integrate the biochemical pathway from other many sources into an integrated database. In this paper, we provide the SBML converter and mediator to integrate the biochemical pathway data. The SBML converter can convert the SBML documents from others source into the local database.

And the mediator can parse and convert the flat file from other many sources. We developed the mediator for the several external databases (ex, Ecocyc, KEGG, dbEST, Unigene DB, etc).



<Figure5> Visualization of the pathway D-Allose degradation.

4.3 Graphical User Interface: Visualization

The system provides GUI (Graphical User Interface) is tools for user's query and visualizing biochemical pathways. Visualizing biochemical pathway is essential for user to analyze biochemical data efficiently. <Figure 5> shows the visualization of the biochemical pathways of D-allose degradation. Reactions connect three kinds of entities (distinguished by the color in the system) via edges Note that in the object model, each object can be expressed differently according to the objects type (polymorphism).

4.4 Gene-ontology database

Biologists may describe the some biochemical entities in a different terminology. So there are many data duplication and inconsistency in the conventional databases. Gene Ontology is an essential tool to integrate biochemical data scattered all around consistently. In the integration of various biochemical pathways from remote heterogeneous sites, we use gene ontology to unify terminologies. When users querying to the database, the terminologies used in the query are transformed into a standard terminologies by using the gene ontology.

4.5 Querying service

Users must have access to the data of biochemical pathway database and have to retrieve the data in a various formats. Querying service is the environment that provides users with a lot of predefined queries or ad-hoc queries. For example, if the users want to search a specific pathway, then users click the pathway name, then the system generates a query for the selected pathway. The querying service supports retrievals through pathway name, enzyme name, enzyme number, reaction product name, etc. The queries are transformed into object queries before accessing the databases.

4.6 Pathway reconstruction service

Pathway reconstruction can be done either by inserting gene list or by inserting into the sequence data that is not annotated. The system predicts the pathways corresponding to the gene list or the genome sequence by using the local pathway database. The local pathway database contains more than 20000 pathways. For each pathway P in database, the reconstruction engine evaluates the evidence that P occurs in organism S by computing how many enzymes in P are present in S, based on the existing set of functional annotations for S. The algorithmic evaluation of pathway evidence differentiates enzymes that are unique signatures for a given pathway from enzymes that are used in multiple pathways and thereby provide weaker evidence for the presence of the pathway. And the local orthologs database contains the sample orthologs group for 20 complete genomes including eukaryote.

For the sequence that was not annotated sequence, the reconstruction engine works like <Figure 6>.

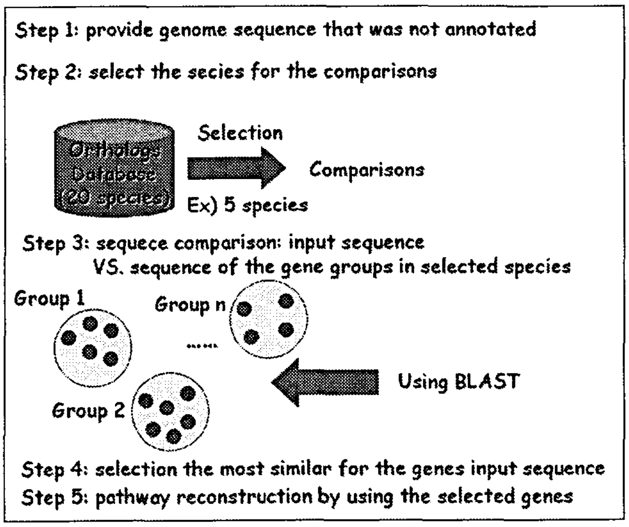
[Step 1] User provides the sequence for the pathway.

[Step 2] User selects the species for the comparisons with input sequence.

[Step 3] The engine does sequence alignment for the input sequence with the sequence of the gene groups in the selected species by using BLAST.

[Step 4] Select the most similar gene list for the input sequence.

[Step 5] If the results contain the gene-list, then reconstruct the pathway by using the genome information of selected gene-list.



Step 1: provide genome sequence that was not annotated

Step 2: select the secies for the comparisons

Step 3: sequece comparison: input sequence
VS. sequence of the gene groups in selected species

Step 4: selection the most similar for the genes input sequence

Step 5: pathway reconstruction by using the selected genes

<Figure 6> Pathway reconstruction steps for the not annotated sequence

4.7 Biochemical pathway simulation supports

Biochemists want to know the quantitative change by time in the biochemical/biochemical pathways. To do it, the system provides the biochemical pathway simulation. User can modify the initial quantity and the kinetic law for the pathway, and the result are exported into the SBML documents. Then user can simulate to the pathway using public simulation software such as Gepasi [2] or COPASI [19] by importing the SBML documents.

5. Conclusion and future work

Research on biochemical pathway an important area in system biology. In the paper, we presented an integrated pathway database system and its novel features. The integrated database is needed to manage all kinds of the biochemical pathway data including metabolic, signal transduction, and gene regulation pathways. Several tools are provided for the analysis of the biochemical pathway data. The presented system provides dynamical visualization and querying tools for the pathway database. The system also provides SBML converter for the exchange the biochemical pathways. To solve the terminology ambiguity, gene ontology is used in the data integration and query processing. For genome information (Gene-list or FASTA format file), the system generates expected pathways by using reconstruction service.

In the future, we will accommodate all kinds of pathways data in the unified database, and provide integrated analysis services to the users. We are applying our system to the pathway reconstruction of vibrio and *sphingomonas chungbukensis* DJ77.

Acknowledgements

This work was supported by grant No. R01-2003-000-11723-0 from the Basic Research Program of the Korea Science & Engineering Foundation.

References

[1] B. J. Breitkreutz, et al., "Osprey: a network visualization system," *Genome Biol., 3, PREPRINT0012*, 2002.
[2] Gepasi, *http://www.gepasi.org*
[3] Gene Ontology, http://www.geneontology.org
[4] Hiroaki Kitano, *Foundations of Systems Biology*, MIT Press, 2001.
[5] P. D. Karp, et al., "The MetaCyc database," *Nucleic Acids Res.*, 30, 59-61, 2000.
[6] R. M. Kuffner, et al., "PathDB," *The Molecular Biology Database*, 2004
[7] S. Goto, et al., "LIGAND: database of chemical compounds and reactions in biochemical pathways," *Nucleic Acids Res.*, 30, 402-404, 20002.
[8] L. Krishnamurthy, et al., "Pathways database system: an integrated system for biochemical pathways," *Bioinformatics*, 19, 930-937, 2003.
[9] Matthew Holford, et al., "VitaPad: Visualization Tools for the Analysis of Pathway Data," *Bioinformatics Advance Access published*, November 25, 2004
[10] D. Pan, et al., "PathMAPA: a tool for displaying gene expression and performing statistical tests on metabolic pathways at multiple levels for Arabidopsis," *BMC Bioinformatics*, 4, 56, 2003.
[11] W. Salamonsen, et al. "BioJAKE: a tool for the creation, visualization and manipulation of metabolic pathways," *Pac. Symp. Biocomput.*, 392-400, 1999.
[12] E. Selkov, et al., "MPW: the Metabolic Pathways Database," *Nucleic Acids Res.*, 26, 43-45.
[13] E. Trost, et al., "Java editor for biochemical pathways," *Bioinformatics*, 19, 786-787, 2003.
[14] H. Yu, et al., "TopNet: a tool for comparing biochemical subnetworks, correlating protein properties with topological statistics," *Nucleic Acids Res.*, 32, 328-337, 2004.
[15] A. Ludemann, et al. "PaVESy: pathway visualization and editing system," *Bioinformatics*, Eprint, 2004.
[16] R. L. Tatusov, et al., "A genomic perspective on protein families," *Science* 278, 631637, 1997.
[17] Valery Reshetnikoy, et al., "Vector PathBlazer: A New Pathway Analysis and Visualization Tool," *ISMB*, 2003.
[18] SBML (System Biology Markup Language), http://www.sbml.org
[19] Copasi, http://www.copasi.org
[20] What is E-Cell, http://www.e-cell.org