Application Notes Genomics & Informatics Vol. 2(4) 191-194, December 2004

The BIOWAY System: A Data Warehouse for Generalized Representation & Visualization of Bio-Pathways

Min Kyung Kim[1], Young Joo Seol[2,3], Sang Ho Lee[1], Eun Ha Song[1], Ho Il Lee[2,3], Chang Shin Ahn[3], Eun Chung Choi[1] and Hyun Seok Park[1,2*]

[1]Department of Computer Science and Engineering, Ewha University, Seoul 120-750, Korea, [2]Institute of Bioinformatics, Macrogen Inc., Seoul 153-023, Korea, [3]School of Computer Engineering, Sejong University, Seoul 143-747, Korea

Keywords: visualization, pathway, data warehouse, modeling
Availablity: http://203.241.70.162

Abstract

Exponentially increasing biopathway data in recent years provide us with means to elucidate the large-scale modular organization of the cell. Given the existing information on metabolic and regulatory networks, inferring biopathway information through scientific reasoning or data mining of large scale array data or proteomics data get great attention. Naturally, there is a need for a user-friendly system allowing the user to combine large and diverse pathway data sets from different resources.

We built a data warehouse - BIOWAY - for analyzing and visualizing biological pathways, by integrating and customizing resources. We have collected many different types of data in regards to pathway information, including metabolic pathway data from KEGG/LIGAND, signaling pathway data from BIND, and protein information data from SWISS-PROT.

In addition to providing general data retrieval mechanism, a successful user interface should provide convenient visualization mechanism since biological pathway data is difficult to conceptualize without graphical representations. Still, the visual interface in the previous systems, at best, uses static images only for the specific categorized pathways. Thus, it is difficult to cope with more complex pathways. In the BIOWAY system, all the pathway data can be displayed in computer generated graphical networks, rather than manually drawn image data. Furthermore, it is designed in such a way that all the pathway maps can be expanded or shrinked, by introducing the concept of super node. A subtle graphic layout algorithm has been applied to best display the pathway data.

Summary

General-purpose database integration systems aim at integrating data from remote heterogeneous sources. Compiled systems can serve as a platform for various special-purpose systems. EnsEMBL is an excellent example of a successful integration of data and tools for the genome browsing (Wong., 2002; Clamp *et al.*, 2003).

In case of pathway data, PFBP (Protein Function and Biochemical Pathway) project is an on-going project, lead by EBI (van Helden *et al.*, 2001). Its main output, aMAZE database that is constructed by parsing data from BRENDA, KEGG/LIGAND, and EMP primarily focuses on the metabolic pathways. Each database has its own ontology and data model and is suitable for the representation of metabolism, gene regulation, and signal transduction. But it has limitations; it does not yet provide signaling and protein interaction data in their alpha version.

Although KEGG contains metabolic and regulatory pathway information, it contains manually drawn image data rather than machine executable formats. To make matters worse, it is hard to understand interconnection among pathways because EC(Enzyme Classification) numbers are used in metabolic pathways instead of protein names in regulatory pathways. There is a consortium, called BIOPAX, which aims to standardize data exchange formats, data models and ontology for biopathways.

By combining the merits of previous systems, we have built a data warehouse, named BIOWAY by integrating and customizing pathway data from various resources (especially focused on pathways regarding agricultural microorganisms). The BIOWAY system is designed as a three-layer application as shown Fig. 1: a layer for data repository(Database Layer), the second layer for analyzing biopathway data(Analysis Layer), and the third level for providing interfaces layer (Visualization Layer). The BIOWAY system has been uniformly written in Java with ORACLE9i. For data exchange between analysis and visualization, the system offers an XML export.

*Corresponding author: E-mail hspark@macrogen.com, Tel +82-2-2113-7007, Fax +82-2-2113-7016

Accepted 6 November 2004

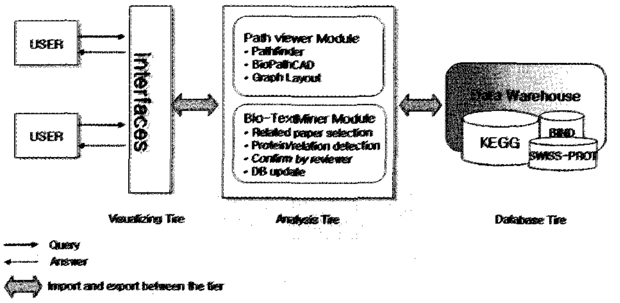

Fig. 1. The system Architecture of BIOWAY.

Database Layer

There are four representative entity types in BIOWAY: nodes, interactions, pathways, and networks. The four entity types have recursive hierarchical relationship. The relationship between nodes is interaction. A set of interactions is used for the description of pathways. A network is composed of a set of pathways. Pathways can be considered as a node in a network. According to these data model, we show the whole network as pathway of pathways. Any molecules in a cell are nodes in BIOWAY. It contains five sub-entity sets: DNA, RNA, protein, complex and compound. A DNA node contains only a small number of regulatory sequences (eg promoter regions) from BIND since Protein coding DNA data is stored in a separate GENE table. A RNA node contains snRNA that is a component of snRNP. Majority of the nodes are proteins whose annotations are extracted from SWISS-PROT, KEGG/LIGAND and BIND (Boeckmann *et al.*, 2003; Kanehisa *et al.*,2002; Bader *et al.*, 2003). The protein information is recognized by SWISS-PROT id and GI number.

Complexes are the result of interactions and ensemble of nodes that can be co-purified from a cell culture or tissue. Since complexes are usually joined pathways together, we consider it as a separate component of the entity. We applied a guideline that the entity found in SWISS-PROT is classified as protein and the entity found in BIND COMPLEX is classified as complex. Other nodes such as small molecules and compounds might be more difficult to define at present. We have checked every item for the integration of the molecules from KEGG/LIGAND and BIND databases. Small molecules such as ATP, NADPH, and NADH should be notified for the computation of biological processes because they participate in complicated interaction steps frequently. For example, if we compute biological processes A → B → C, given interaction A → B and B → C, interacting partner of small molecules would be too big to handle.

Interactions represent the operations in which nodes participate. There are two types of interactions: direct interaction (such as reaction, assembly, modification) and regulation (such as catalysis, control). The interaction between nodes belongs to direct interaction. Regulation is an interaction between node and interaction. For example, transcriptional regulation is the interaction between regulatory molecules (node) and transcriptional process(direct interaction between DNA and RNA). Each interaction has several attributes for example direction. If it was originated form chemical reaction, it should be contained to represent metabolic pathway.

Pathways are ordered sequences of binary interaction. BIOWAY aims to describe all kinds of pathways such as genetic network, signal transduction and metabolic pathways by sets of interactions between the nodes. For the construction of BIOWAY database, we convert the format of raw pathway data from KEGG/LIGAND database to metabolic pathways as shown figure 2 (Kim and Park, 2002). While a signaling pathway is represented as a series of interactions, metabolic pathway is not. We want to describe metabolic pathway as a series of interactions like signaling pathway and focus on the possibility that metabolic pathway could represent as a set of binary interactions.

Networks are described as a set of pathways and useful for the representation of crosstalk and interdependency among different kinds of pathways. Networks are designed for the proteins participating in metabolic and signaling pathway at the same time. These kinds of proteins are acting as a hub in a cell (Kim *et al.*, 2003). For example, Raf protein takes part in several kinds of metabolic pathways such as spingolipid, inositol phosphate, starch/sucrose, nicotinate/nicotinamide metabolism, benzoate degradation (in KEGG) and EGFR signaling pathway (in

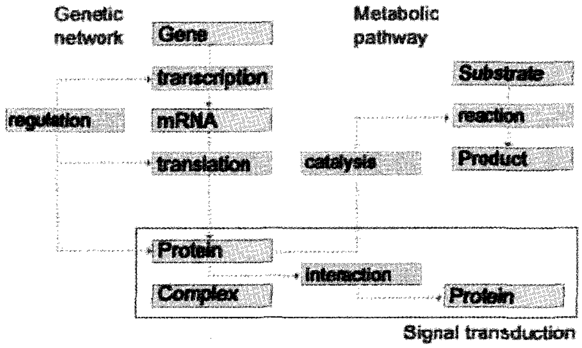

Fig. 2. The data model of BIOWAY. Protein-protein interaction, signal transduction, genetic network and metabolic pathway are represented as a single schema.

BIND). Whereas the crosstalk between pathways in KEGG and BIND database is limited to its own pathway domain, BIOWAY could represent all kinds of pathway under a single schema as illustrated in Fig. 2.

Analysis Layer

This layer consists of several modules: PathViewer, BioTextMiner, 3DViewer, BioPathCad, and PathFinder.

PathFinder is a new tool for computing the entire biopathway based on the binary representation of each reaction. It has been tested on the 130 metabolic pathways and 8 regulatory pathways. Given two nodes in the pathway maps, this module will return all the possible paths between nodes, sorted in the order of scoring parameters.

BioPathCad module makes it possible to edit the computer-generated pathway maps. Much of the design of this module is based on the general CAD system which is usually used in architecture field. Important background knowledge in bioinformatics is often buried in textual documents.

The role of BioTextMiner module is to search the literature and automatically extract information from abstracts and papers, to provide two essential research support services: accelerating user's task by partially automating the process to find the relations between genes and gene products, and providing a convenient environment for researchers, annotating (or tagging) the biological literature in response to queries from biological experts. Text mining methods range from term recognition to extraction of complex relationships of interaction between proteins. It is expected that text mining in general will provide tools to facilitate the annotation of vast amounts of biological pathways.

Visualization Layer

This layer is directly related to PathViewer module in Analysis Layer.

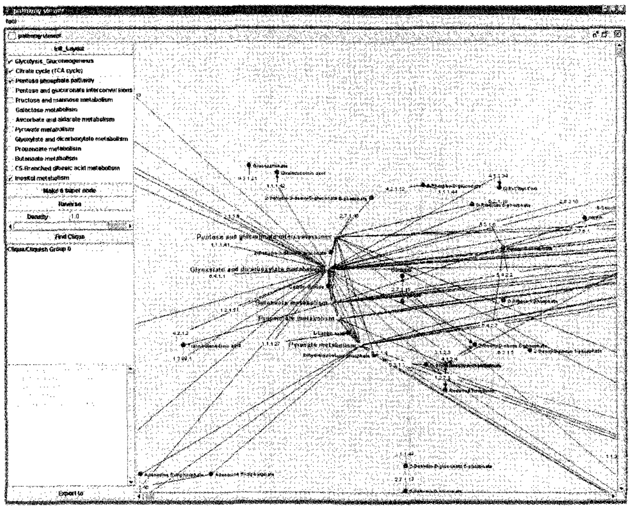
pathway viewer
Init_Layout
✓ Glycolysis_Gluconeogenesis
✓ Citrate cycle (TCA cycle)
✓ Pentose phosphate pathway
Pentose and glucuronate interconversions
Fructose and mannose metabolism
Galactose metabolism
Ascorbate and aldarate metabolism
Pyruvate metabolism
Glyoxylate and dicarboxylate metabolism
Propanoate metabolism
Butanoate metabolism
C5-Branched dibasic acid metabolism
✓ Inositol metabolism
Make a super node
Reverse
Density 1.0
Find Clique
Clique/Cliquish Group 0

Fig. 3. The result of layout. Users can select pathways in the left pathway list. This figure shows the case of Glycolysis, TCA cycle, Pentose Phosphate, and Inositol metabolism pathways.

GraphLayout module is designed to display pathway data. A subtle graphic layout algorithm has been applied to best display the pathway data. We used yFiles which provides a powerful framework for visualization applications. In the BIOWAY system, users can select the pathways to show more than 2 pathways at a time (Fig. 3). And each specialized pathways could be represented as a single node and this abstraction process has been made to be reversible. Combining several pathways to make a bigger pathway is possible, due to our layer-based GUI design. All objects inherit from one super-class. Each subclass inherits attributes from all of its super-classes.

Future Direction

We have parsed the core information from flat files of KEGG/LIGAND, SWISS-PROT and XML type BIND data. The information allows data retrieval through user interface (Fig. 4). Current prototype represents range of protein-protein interaction, signaling and metabolic pathway, and network information by simple binary interactions (see Table 1.). The database will be extended to transcriptional regulation(TRANSFAC), 3D structure database(CATH), protein motif(PROSITE) and disease related gene database(OMIM). Some of our future plans are to:

- design a mining method for finding a meaningful pathway.
- fine tune for data model
- suggest data submission form.

Identification of appropriate knowledge discovery problems and development of evaluation methods for data mining results are also our ongoing efforts. For future versions, the concept of quality(for example, time and condition) and quantity(for example, flux, kinetics,

Table 1. The Statistics of BIOWAY

Node	157471	Interaction	62519	Pathway	204	Network	2
DNA	5	Translation	35005	Metabolic	196		
RNA	2	Reaction	5429	Signaling	8		
Protein	145966	Catalysis	5436				
Complex	851	Interaction	16649				
Compound	10647						

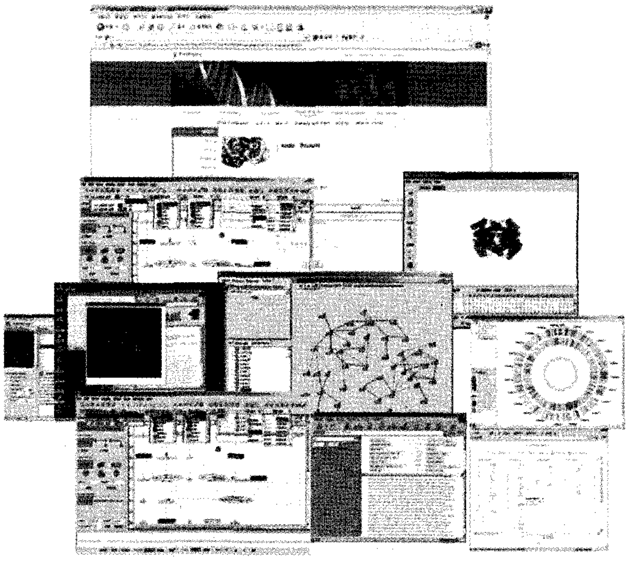

Fig. 4. The snapshot of BIOWAY interfaces. Some of them are standalone software tools developed for the BIOWAY System.

and signaling cascade) would be implemented into the system.

Acknowledgement

The BIOWAY project has been partially supported by the Ministry of Information and Communication and the Ministry of Science and Technology of Korea under the grant number of IMT2000-ab05 and 21st Century Frontier R&D Program of Stemcell Research.

References

Bader, G.D., Betel, D., and Hogue, C.W. (2003). BIND: the Biomolecular Interaction Network Database. *Nucleic Acids Res.* 31, 248-250.

Boeckmann, B., Bairoch, A., and Apweiler, R.M. (2003). The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res.* 31, 365-370.

Clamp, M., Andrews, D., and Baker, D. (2003). Ensembl 2002: accommodating comparative genomics. *Nucleic Acids Res.* 31, 38-42.

Kanehisa, M., Goto, S., Kawashima, S., and Nakaya, A. (2002). The KEGG databases at GenomeNet. *Nucleic Acids Res.* 30, 42-46.

Kim, M.K. and Park, H.S. (2002). Generalized Representation of metabolic and regulatory pathways. *Genome Informatics* 13, 351-352

Kim, M.K., Park, H.S., and Yoo, S.J. (2003). Crosstalk between metabolic and regulatory pathways. *Genome informatics* 14, 372-373

van Helden, J., Naim, A., Lemer, C., Mancuso, R., Eldridge, M., and Wodak, S.J. (2001). From molecular activities and processes to biological function. *Brief Bioinform.* 2, 81-93.

Wong, L. (2002). Technologies for integrating biological data. *Brief Bioinform.* 3, 389-404.