(White Paper)
January 2003

Pathways Database System

Z. Meral Ozsoyoglu

Center for Computational Genomics
Department of Electrical Engineering and Computer Science
Case Western Reserve University
mxo2@po.cwru.edu

Abstract

During the next phase of the Human Genome Project, research will focus on functional studies of attributing functions to genes, their regulatory elements, and other DNA sequences. To facilitate the use of genomic information in such studies, a new modeling perspective is needed to examine and study genome sequences in the context of many kinds of biological information. Pathways are the logical format for modeling and presenting such information in a manner that is familiar to biological researchers. In this paper we introduce an integrated system, called Pathways Database System, with a set of software tools for modeling, storing, analyzing, visualizing, and querying biological pathways data at different levels of genetic, molecular, biochemical and organismal detail.

Keywords: biological pathways; visualization; database; querying, web services.

1 Introduction

The conventional perspective for managing, analyzing, viewing and querying genomic information is in the context of DNA sequence. In this perspective, DNA sequences are annotated with the identity and location of genes, transcriptional motifs and other regulatory elements, repetitive DNA elements, and chromosome segments that have been conserved among various species during evolution. This perspective is appropriate for studying questions of genome organization and evolution, and for identifying mutated genes that are responsible for phenotypic variation including human diseases. However, DNA sequence does not reflect the context in which most genes act, i.e. functionally related genes are usually not physically clustered in DNA, but instead are distributed among distant sites. The protein products of these genes assemble at appropriate cellular locations to coordinate their biological functions. Thus an alternative to DNA sequence for studying genomic information is biological pathways. Pathways are the sequential and cumulative action of genetically distinct but functionally related molecules. Each reaction in each pathway begins with specific substrates, uses various combinations of molecules as cofactors, activators and inhibitors, and ends with products that are chemically modified substrates. Individual steps in every pathway involve at least one genetically unique gene product which catalyzes the reaction. Thus pathways are an appropriate format for representing the functional role of most genes in the genome.

The three general classes of biological pathways are (1) metabolic and biochemical, (2) transcription, regulation and protein synthesis, and (3) signal transduction. Metabolic pathways are responsible for carrying out the chemical reactions that provide basic biological functions such as DNA, RNA and protein synthesis and degradation, energy metabolism, fatty acid synthesis, and many others. Transcription and protein synthesis are responsible for converting genetic information into proteins (gene products). Signal transduction pathways are responsible for coordinating metabolic

processes with transcription and protein synthesis. Each of these three kinds of pathways has distinct attributes, to be kept and managed in the pathways database.

From this perspective, the functional relations between molecules can be illustrated in these three kinds of pathways. These annotations include, for example, the identity of the substrate(s), product(s), cofactors, activators, inhibitors, enzymes or other processing molecules, RNA and protein expression patterns, reaction kinetics and associated phenotypic variation and diseases. Ultimately, many other kinds of information can be incorporated. As we describe below, in our ongoing work, we are incorporating information about gene and protein sequence, RNA expression patterns, protein function, phenotypes associated with mutated genes, and others. This perspective provides a rich research resource that integrates genomic and biological information that can be managed, analyzed, queried and displayed in dynamic ways at various levels of biological and genetic detail to provide insight into diverse biological processes in health and disease.

Pathways databases raise many important and challenging computational and bioinformatics issues, such as querying and visualizing graph structured databases in multiple abstraction levels, seamless integration of data distributed in diverse sources, integrated and graph-based querying and navigation of data in multiple dimensions, i.e., from biological function to gene expression. Pathways Database System is an ongoing project which aims to address several of these problems.

In this paper, we summarize main features of the current version of Pathways Database System, which is an integrated software system for storing, managing, analyzing, visualizing and querying biological pathways at multiple abstraction levels of detail. At the computational level, Pathways Database System allows users to visualize pathways in multiple abstraction levels, and to pose a wide range of queries using a graphical user interface. By different abstraction levels, we refer to the representation of pathways at different levels of biological function. At one level, for example, all of the individual steps in methylation can be illustrated, while, at another level, the collection of steps are labeled methylation. Together this is an easy and intuitive way to query complex sets of genomic, genetic and biological information. Figure 1 illustrates, as another example, ***multiple abstraction levels*** at which pathways data can be queried, visualized and analyzed, using a hierarchy from individual molecules to pathways, involving structures of molecules, functional use of molecules in processes, pathways of processes, and complex networks of related pathways. Note that this is only an example abstraction hierarchy, and, there may be additional user defined and/or universal abstraction levels and classifications on pathways, and other groups of objects involved in studying pathways data.

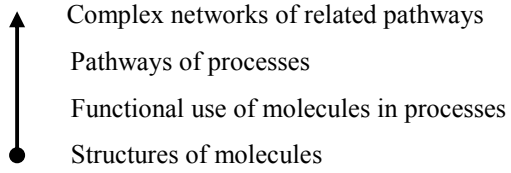

Figure1. An Example Multi-level Abstraction Hierarchy for Pathways Data

The novel features of the Pathways Database System include:

1. Genomic information integrated with other biological data and presented from a pathway, rather than the DNA sequence, perspective.
2. Design for biologists who are possibly unfamiliar with genomics, but whose research is essential for annotating gene and genome sequences with biological functions,
3. Database design, implementation and graphical tools which enable users to visualize pathways data in multiple abstraction levels, and to pose ad-hoc and predetermined queries, and
4. An implementation that allows for web (XML)-based dissemination of query outputs (i.e., pathways data) to researchers, giving them control on the use of pathways data.

References:

[1] L. Krishnamurthy, J. Nadeau, G. Ozsoyoglu, M. Ozsoyoglu, G. Schaeffer, M. Tasan, W. Xu, "Pathways Database System: An integrated set of tools for biological pathways", *ACM SAC-BIO*, 2003, (to appear), selected for publication in *Journal of Bioinformatics*, 2003 (to appear).