Chapter 2
Pathway Databases

Pathway information is available through a large number of databases ranging from high-quality databases created by professional curators to massive databases, covering a vast number of putative pathways, created through natural language processing and text mining of abstracts. Because of the various differences in size, quality, and/or property, it is necessary to use the right database for the user's purpose, regardless of whether it is for commercial or for public use. In this chapter we introduce some of the major pathway databases. These databases can display pathway diagrams, which combine metabolic, genetic, and signal networks based on the literature. This chapter also covers some software applications for the production, editing, and analysis of such pathways.

2.1 Major Pathway Databases

Pathway databases are being created all around the world. Each database strongly reflects its builder's intent and purpose. There are databases with detailed metabolic pathways, while others have detailed signaling pathways. Most databases are created by curators who read papers and extract pathway information which will be organized together with pathway diagrams in the databases. Others are created using natural language processing and text mining, which extract from papers various biological relations such as gene regulatory relations and organize them into databases. This chapter covers those databases focused on metabolic and signaling pathways.

Pathway information is often described in the XML (eXtensible Markup Language) data format, which varies from database to database. This format can be easily read by both computers and humans. The following example shows the information "The lecture with Id "5" will be given on 4/1/2007 by a person named "masao nagasaki" in XML format:

```
<lecture id="5">
      <date>2007-04-01</date>
```

```
          <person>masao nagasaki</person>
    </lecture>
```

In the following chapters, we use acronyms ending with "... ML". This ending simply indicates that the pathway information is stored in some variant of XML. In this book, we do not go into the details of XML.

COLUMN 2

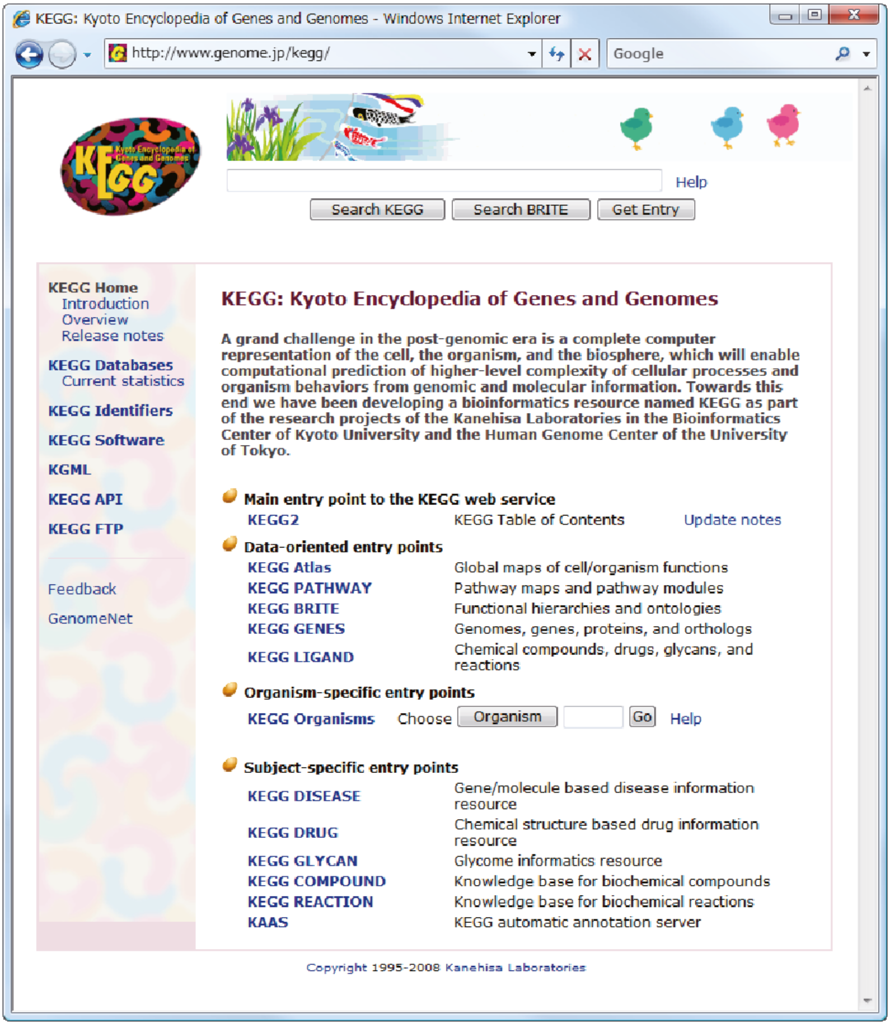
What's XML?

XML is one of many self-extensible markup languages. Its proper name is Extensible Markup Language. A markup language uses a sentence structure to list and categorize information. XML was developed in 1996 by the XML Working Group, part of the international standardization organization W3C. Because the creator can define and share a file format, a creator can use a standardized XML format for multiple applications, while allowing for a high degree of expression not constrained by the syntax.

2.1.1 KEGG

KEGG (Kyoto Encyclopedia of Genes and Genomes) (http://www.kegg.jp/) is a series of databases developed by both the Bioinformatics Center of Kyoto University and the Human Genome Center of the University of Tokyo. This database has been available for over 10 years. As the name encyclopedia suggests, the database includes information necessary for systems understanding of biology, such as genome sequences and chemical information (Figure 2.1). With its goal of collecting all knowledge relevant to biological systems, including the environmental information, KEGG will be a true encyclopedia. The "Pathway" section of KEGG consists mainly of metabolic pathways. For noncommercial uses, the license is free, while for commercial uses, the license is sold from Pathway Solutions Inc. (http://www.pathway.jp/).

KEGG is unique for its focus and coverage of yeast, mouse, and human metabolic pathways. Currently, signaling pathways for cell cycles and apoptosis are being expanded. New pathways are created by professionals (*curators*) who read and summarize the relevant literature. The information is displayed as a browser-viewable



KEGG: Kyoto Encyclopedia of Genes and Genomes

A grand challenge in the post-genomic era is a complete computer representation of the cell, the organism, and the biosphere, which will enable computational prediction of higher-level complexity of cellular processes and organism behaviors from genomic and molecular information. Towards this end we have been developing a bioinformatics resource named KEGG as part of the research projects of the Kanehisa Laboratories in the Bioinformatics Center of Kyoto University and the Human Genome Center of the University of Tokyo.

Main entry point to the KEGG web service

KEGG2	KEGG Table of Contents	Update notes

Data-oriented entry points

KEGG Atlas	Global maps of cell/organism functions
KEGG PATHWAY	Pathway maps and pathway modules
KEGG BRITE	Functional hierarchies and ontologies
KEGG GENES	Genomes, genes, proteins, and orthologs
KEGG LIGAND	Chemical compounds, drugs, glycans, and reactions

Organism-specific entry points

Subject-specific entry points

KEGG DISEASE	Gene/molecule based disease information resource
KEGG DRUG	Chemical structure based drug information resource
KEGG GLYCAN	Glycome informatics resource
KEGG COMPOUND	Knowledge base for biochemical compounds
KEGG REACTION	Knowledge base for biochemical reactions
KAAS	KEGG automatic annotation server

Copyright 1995-2008 Kanehisa Laboratories

Fig. 2.1

pathway diagram. For example, one could search for the existence of a metabolic pathway from substance A to B, or the required enzymes for such a reaction. In addition, the database has links to relevant information such as genome sequences, positions, and conditions. The database is stored in a format called *KEGGML*. Since the pathways are then displayed as GIF files, the user cannot easily edit the pathway information.

2.1.2 *BioCyc*

BioCyc is a pathway database provided by SRI International (http://www.biocyc.org/). The database is a high-quality database focused on metabolic pathways originally formed by SRI International's bioinformatics research group. Related to BioCyc are the EcoCyc, MetaCyc, HumanCyc databases. Licenses are free for academic and nonprofit uses. Humans and *E. coli* are the major organisms listed with a variety of others. EcoCyc is mainly a database of *E. coli* metabolic pathways. These reactions are shown in the form of chemical equations. EcoCyc also contains a small number of signaling pathways. Curators extracted the pathway knowledge from the literature. Pathways are described with a proprietary format.

In addition, gene regulatory information upstream of the metabolic pathways is also listed. In other words, there is a link from a metabolic pathway to the genes coding enzymes and its regulators. The pathway map displays are separated in levels of detail. At the most detailed level, the metabolic products are shown in terms of the chemical equations.

2.1.3 *Ingenuity Pathways Knowledge Base*

Ingenuity Pathways Knowledge Base (IPKB) is the pathway database created by Ingenuity Systems Inc. (http://www.ingenuity.com/). All licenses, including academic and nonprofit, require a fee. The database consists of gene regulatory and signaling pathways. Curators extract knowledge from the literature for this database, which currently contains human, mouse, and rat genetic information. (As of May 2008, the website claims 13,600 human genes, 11,000 mouse genes, and 6,600 rat genes cataloged.) The database uses the Ingenuity Pathways Analysis (IPA) software mentioned later to view and analyze pathway data and thus IPKB is inaccessible through a web browser. Like KEGG and BioCyc, IPKB uses its own internal format for storage. However, unlike KEGG and BioCyc, IPKB allows for the editing of pathways through IPA. This edited data can later be exported as a graphic format such as SVG.

2.1.4 *TRANSPATH*

TRANSPATH is a gene regulatory and signaling pathway database created by BIOBASE (http://www.biobase-international.com/). The most recent version of the data requires a fee for both nonprofit and commercial uses. However, some parts of the old data are provided to academic users as a trial version (http://www.gene-regulation.com/). In addition to TRANSPATH, BIOBASE offers the TRANSFAC database of transcription factors and PROTEOME database of protein. It also provides a software ExPlain which combines and analyzes these databases.

TRANSPATH is formed similarly to those listed above through curators and therefore maintains high quality. Pathways are listed using a proprietary format. If the user has a license, the pathways are viewable from a web browser. In addition, it is possible to download the data stored as text file. For example, the phosphorylation of I-κB is shown below:

```
IkappaB-alpha, IkappaB-beta:p50:RelA +
ATP-IKK-alpha{p}:IKK-beta{p}:(IKK-gamma)2
-> IkappaB-alpha, IkappaB-beta{pS}:p50:RelA +
ADP (phosphorylation)
```

Each reaction has a link to the literature that confirms its existence. Therefore it is easy to understand what each biochemical reaction means. Figure 2.2 shows the IL-1 pathway displayed via a web browser, while Figure 2.3 displays the reaction information from TRANSPATH shown through a web browser. (As of May 2008, the website claims a total of 135,563 reactions mainly for human, mouse, and rat.)

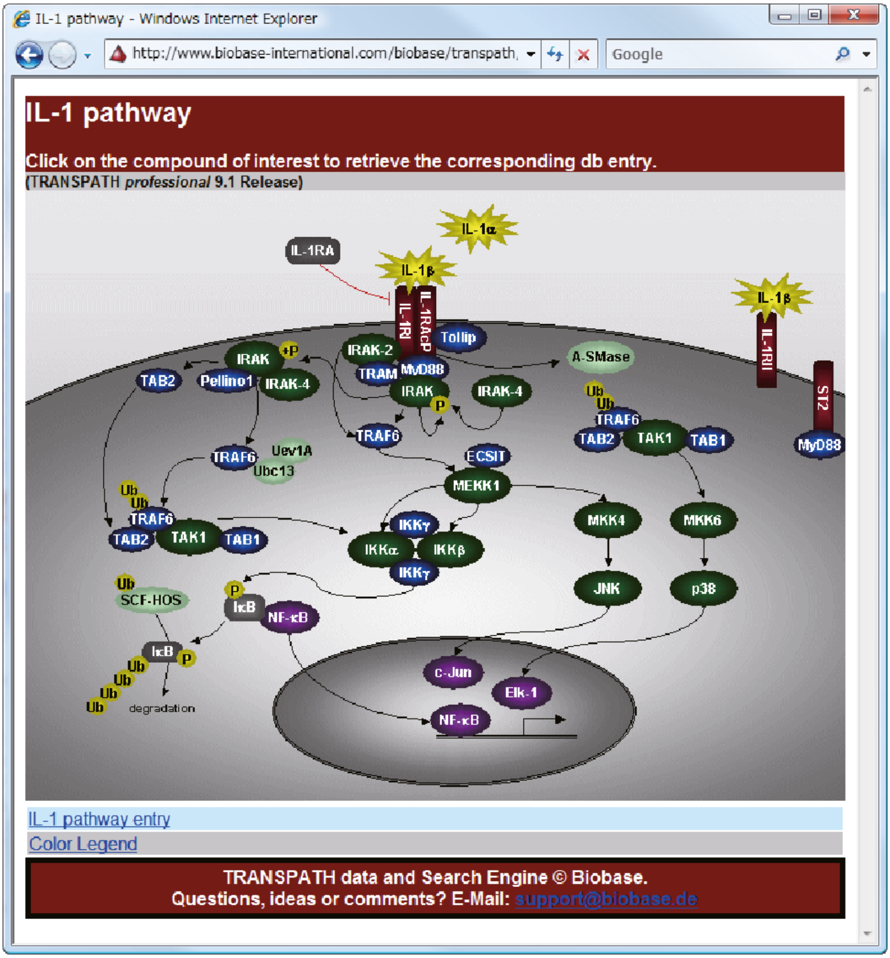
2.1.5 ResNet

ResNet (http://www.ariadnegenomics.com/) is the pathway database created by Ariadne Genomics. Academic and commercial licenses require a fee. The pathways of ResNet consist mainly of gene regulatory and signaling pathways. Unlike other databases, ResNet is constructed through computer analysis. In other words, the pathways and networks are created through natural language processing of relevant literature. MedScan is used for this natural language processing procedure. The database is constructed mainly from abstracts in PubMed, but some entries make use of the full text. In addition, there are a small number of entries created by curators.

The pathway data created by MedScan can be viewed through the viewing tool Pathway Studio. Similarly to other databases, MedScan uses its own proprietary format. ResNet employs arrows with various labels to show the relationships between molecules. '$+$' indicates activation, while '$-$' indicates suppression. Relationships which cannot be determined are indicated with '?'. In addition, comments are attached to the relation for nontrivial biological information. All such data are completely user editable.

2.1.6 Signal Transduction Knowledge Environment (STKE): Database of Cell Signaling

The database of Cell Signaling, a part of Signal Transduction Knowledge Environment (STKE) (http://stke.sciencemag.org/), is an online service provided by Science. This is a high-quality signaling pathway database created and maintained by curators. The database can be accessed by subscribing to the online service of Sci-



IL-1 pathway

Click on the compound of interest to retrieve the corresponding db entry.
(TRANSPATH *professional* 9.1 Release)

IL-1 pathway entry
Color Legend

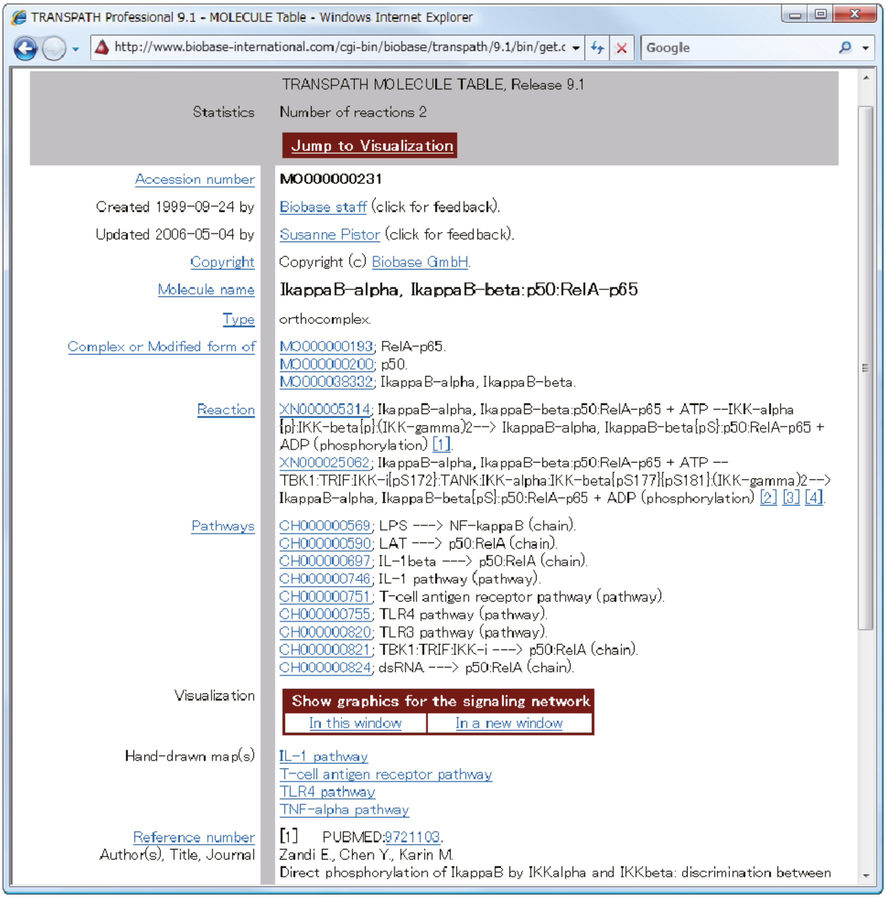
TRANSPATH data and Search Engine © Biobase.
Questions, ideas or comments? E-Mail: support@biobase.de

Fig. 2.2

ence although user registration does grant limited functionality such as pathway viewing. This database is accessible in GIF or SVG format through a web browser. Similarly to KEGG and BioCyc, this makes the pathway uneditable in browser. Similarly to ResNet, this database makes use of the labels ‘+’ for stimulatory relations, ‘−’ for inhibitory relations, ‘0’ for neutral relations, and ‘?’ for undefined relations. A feature of this database is the separation of pathways into “specific” and “canonical”. Specific pathways are those which are unique to an organism, while canonical pathways are those which are common. Unlike TRANSPATH or ResNet, however, the user cannot specify a list of genes (proteins) and create a network on that selection.

The following information is available in this database (as of March 2007):

- Cell Biology (46 pathways)


TRANSPATH Professional 9.1 - MOLECULE Table - Windows Internet Explorer
http://www.biobase-international.com/cgi-bin/biobase/transpath/9.1/bin/get.c
Google

TRANSPATH MOLECULE TABLE, Release 9.1

Statistics	Number of reactions 2
	Jump to Visualization
Accession number	MO000000231
Created 1999-09-24 by	Biobase staff (click for feedback).
Updated 2006-05-04 by	Susanne Pistor (click for feedback).
Copyright	Copyright (c) Biobase GmbH.
Molecule name	IkappaB-alpha, IkappaB-beta:p50:RelA-p65
Type	orthocomplex.
Complex or Modified form of	MO000000193; RelA-p65. MO000000200; p50. MO000038332; IkappaB-alpha, IkappaB-beta.
Reaction	XN000005314; IkappaB-alpha, IkappaB-beta:p50:RelA-p65 + ATP --IKK-alpha{p}:IKK-beta{p}:(IKK-gamma)2--> IkappaB-alpha, IkappaB-beta{pS}:p50:RelA-p65 + ADP (phosphorylation) [1]. XN000025062; IkappaB-alpha, IkappaB-beta:p50:RelA-p65 + ATP --TBK1:TRIF:IKK-i{pS172}:TANK:IKK-alpha:IKK-beta{pS177}{pS181}:(IKK-gamma)2--> IkappaB-alpha, IkappaB-beta{pS}:p50:RelA-p65 + ADP (phosphorylation) [2] [3] [4].
Pathways	CH000000569; LPS ---> NF-kappaB (chain). CH000000590; LAT ---> p50:RelA (chain). CH000000697; IL-1beta ---> p50:RelA (chain). CH000000746; IL-1 pathway (pathway). CH000000751; T-cell antigen receptor pathway (pathway). CH000000755; TLR4 pathway (pathway). CH000000820; TLR3 pathway (pathway). CH000000821; TBK1:TRIF:IKK-i ---> p50:RelA (chain). CH000000824; dsRNA ---> p50:RelA (chain).
Visualization	Show graphics for the signaling network In this window In a new window
Hand-drawn map(s)	IL-1 pathway T-cell antigen receptor pathway TLR4 pathway TNF-alpha pathway
Reference number Author(s), Title, Journal	[1] PUBMED:9721103. Zandi E., Chen Y., Karin M. Direct phosphorylation of IkappaB by IKKalpha and IKKbeta: discrimination between

Fig. 2.3

- Developmental and Reproductive Biology (32 pathways)
- Immune, Inflammatory, and Defense Signaling (17 pathways)
- Microbiology (6 pathways)
- Neurobiology (5 pathways)
- Plant Biology (15 pathways)
- Stress, Death, and Survival Signaling (9 pathways)
- Pathways Implicated in Human Disease (11 pathways)

2.1.7 Reactome

Reactome is a pathway database containing cell metabolic and signaling pathways (http://www.reactome.org/). Cold Spring Harbor Laboratory, European Bioinformatics Institute, and Gene Ontology Consortium—which specifies Gene Ontology mentioned later—are the main developers of the project. Although humans are the main organism catalogued, it has data for 22 other species such as mouse and rat. Pathway knowledge is extracted by curators.

Reactome's pathways and reactions can be viewed but not edited through a web browser. Though the storage format is proprietary, a large number of pathways can be obtained in multiple formats. Human reactions are distributed through SBML format, human protein relations are given through TSV format, and cellular event information is given through the BioPAX format listed in Section 2.3.5. All data can easily be downloaded and edited.

2.1.8 Metabolome.jp

Metabolome.jp (http://metabolome.jp/) is a metabolic pathway-focused database created by some research labs led by the University of Tokyo Graduate School of Frontier Sciences. Using an applet called ARM, pathways can be viewed and edited through a browser. Pathways are created by curators. Each metabolic product is shown with an atomic structural formula and it is possible to display a pathway which considers atom movements. Unlike KEGG, it is possible to track the movement of atoms in metabolic reactions. Pathway storage uses a proprietary format.

2.1.9 Summary and Conclusion

As described above, a variety of databases are available. The databases vary in the types of information offered; there are metabolic pathway databases and signaling pathway databases. In addition, there are differences in the organisms covered by the databases. However, a common problem is that these databases do not have enough information to permit simulating the pathways.

Pathway databases are constructed by curators or through the use of natural language processing and text mining tools via computer. This difference affects the characteristics of the databases significantly. Through methods such as natural language processing, one has the advantage of a large breadth of literature which curators are unable to cover. In addition to the quality problem, however, there is usually the problem of lacking specific biological or experimental facts listed in the database. Although it is likely that this technology will be improved in the future, such databases are currently ancillary to those created by curators (such as IPKB or TRANSPATH). Databases created by curators are on the whole more reliable and detailed. Each pathway database has its own proprietary format. Although there are formats such as SBML and BioPAX (mentioned later) which aim at standardizing these formats, the current situation is not satisfactory in practice.

In addition to the databases introduced here, there are many other good pathway databases. Some of them are:

- BioCarta: Signaling pathways (http://www.biocarta.com/)
- INOH: Signaling pathways (http://www.inoh.org/)
- iPath: Signaling pathways (http://www.invitrogen.com/content.cfm?pageid=10878)
- Molecular Interaction Map: Signaling pathways as well as gene regulatory networks (http://discover.nci.nih.gov/mim/index.jsp)

There are a myriad of databases which are not listed here. It is likely that databases—whether or not they are listed here—will develop or disappear for a variety of reasons: "Research fund is terminated."; "The government fully supports the database."; "The database is commercialized." When using a database, the following items will be a useful guideline for assessment.

- Is the database viewable through a web browser?
- Is there a licensing fee?
- What is the data type (metabolic, gene regulatory, signaling, etc.)?
- Is the database developed through computer or curator?
- Is there any software for editing pathways?
- Is it possible to simulate the pathway?

2.2 Software for Pathway Display

Pathway information must somehow be displayed. In this section, we introduce software applications that help visualize pathways.

2.2.1 Ingenuity Pathway Analysis (IPA)

Ingenuity Pathway Analysis (IPA) is the software used to display pathway data from the Ingenuity Pathway Knowledge Base (IPKB) by Ingenuity Systems Inc. For a given gene set, IPA automatically generates the pathways that are related to those genes. This means that, for example, if one finds a set of genes with large gene expression variance as a result of microarray analysis, IPA automatically generates the pathway which involves those genes. The pathway is generated with a mixture of human, mouse, and rat data. Therefore, it should be cautioned that there can be no pathway in the real organism of the user's interest even if IPA generates some pathway.

2.2.2 Pathway Builder

Pathway Builder is a viewer that automatically generates pathways from the TRANSPATH database (http://www.biobase-international.com/). Pathway Builder can find the pathways related to a set of genes and connect them to display as one pathway. This allows to search and display genes upstream and downstream of the genes in the set. Using this feature, one can find the genes whose transcriptions are activated by a gene (downstream search) or find the genes which regulate a particular gene (upstream search).

2.2.3 Pathway Studio

Pathway Studio is the viewer for Ariadne Genomics' ResNet. Pathway Studio has a function to add new molecules and user's information into the pathway. The automatic layout feature is one of the unique parts of this viewer. Like IPA and Pathway Builder, Pathway Studio can search with gene names and create a pathway of genes related to any given gene (or protein).

2.2.4 Connections Maps

Connections Maps is a viewer for Signal Transduction Knowledge Environment (STKE): Database of Cell Signaling. This program creates the GIFs and SVGs of the pathways according to the data created by curators called "Pathway Authorities". Genes and proteins have specific set symbols and colors, and the relations are indicated with '$+$' (activation), '$-$' (repression), and '?' (undefined). In addition, the graphics have embedded links, which make it simple to get more detailed information. Because of the SVG format, the user is free to magnify any level of the pathway. However, Connections Maps is unable to generate custom pathways from a list of genes, unlike IPA and Pathway Builder.

2.2.5 Cytoscape

Cytoscape is a software tool designed to visualize the molecular interactions as a network diagram (http://www.cytoscape.org/). It was developed mainly by the Institute for Systems Biology and University of California San Diego as well as some other institutions such as the Pasteur Institute, MSKCC, Agilent, and UCSF as an open source project. The program is free to download and it requires the use of Java; the current version (as of April 2008) is 2.6.0.

The software can use protein–protein binding information, protein–DNA binding information, and microarray data to provide a network view. The network visualization is proprietary. Proteins and genes are shown as circles, triangles, and squares (called nodes), while relationships are shown as lines (called edges). In addition to nodes and edges, various attributes such as Gene Ontology or wet lab measurements of expression can be added. Cytoscape has a filtering feature to show only the network of interest. By using Gene Ontology in combination with filtering, it is possible to show all the genes with a certain function.

In addition, analysis functionality can be provided as plugin. A number of plugins have been developed for a variety of purposes. For example, a plugin provided by Agilent allows Cytoscape to extract protein and genome information from textual abstracts and display the results as a network.

A variety of storage formats can be imported, such as Simple Interaction File (SIF), Graph Markup Language (GML), Extensible Graph Markup and Modeling Language (XGMML), SBML, BioPAX, and PSI MI. Of these, GML and XGMML are standard XML formats for graph (a set of vertices connected with edges) formation. SBML, BioPAX, and PSI MI will be mentioned later. The SIF format is, as the name states, a simple format for showing interactions. For example, if protein A and protein B act upon each other, one would simply put the interaction type between the names and write in the following way:

`A pp B` (`pp` stands for protein-protein interaction)

2.3 File Formats for Pathways

2.3.1 Gene Ontology

Gene Ontology (GO) defines a common framework to organize biological concepts (http://www.geneontology.org/). Ontology was originally studied in Artificial Intelligence and is defined as “a hierarchical taxonomy of terms for a certain area of knowledge”. The GO project began in the 1990s, and seeks to record genetic and functional information in the same syntax to simplify database comparison. The terms defined by GO are called GO terms and can be divided into the following three categories:

- Biological processes
- Cellular components
- Molecular functions

These categories have terms such as “nuclear chromosome”, “chromosome”, “nucleus”, and “cell”. Between these terms are relationships such as “*is_a*” as in “nuclear chromosome *is_a* chromosome” or “*part_of*” as in “nucleus *part_of* cell”. These relationships are called *ontologies*. The relationships between such terms are listed in a directed acyclic graph (DAG). A consortium has been formed adopting GO and there are a large number of databases contributing to the project.

2.3.2 PSI MI

Proteomics Standards Initiative (PSI) began around 2002 and attempts to standardize data from mass spectrometry and protein–protein interaction experiments, in order to facilitate data comparison and transfer (http://psidev.sourceforge.net/). PSI MI is defined to handle information on protein-protein interactions.

2.3.3 CellML

CellML is the first Systems Biology XML format to integrate cellular level molecular dynamics as a part of its format. Over 300 models have already been submitted and displayed at the CellML Repository (http://www.cellml.org/). It is a format developed by the University of Auckland in New Zealand under the auspices of the International Physiome Project. CellML 1.0 was published in 2000 and CellML1.1 is currently proposed. CellML is structured to include model structures, differential equations-based dynamics information, and additional comments. To store all these, CellML utilizes MathML, a math typesetting format for XML. The format seeks to describe everything from the cellular to organ level by combining with FieldML (http://www.physiome.org.nz/xml_languages/fieldml/).

2.3.4 SBML

SBML (Systems Biology Markup Language) is one of the XML formats designed to model biological reactions (http://www.sbml.org/). In 2001, SBML level 1 was released, and in 2003, SBML level 2 was released. Like CellML, SBML was expanded to include MathML support, spatial position and physical size information.

As of May 2008, SBML 2.3 is the current version. Currently, this format is actively heading towards level 3 release. An open source application called *SBW* (Systems Biology Workbench) has been developed to combine with other simulation and analysis software for use with SBML. In addition, a database called *BioModels* (http://www.ebi.ac.uk/biomodels/) based upon SBML, though small, has been under development.

2.3.5 BioPAX

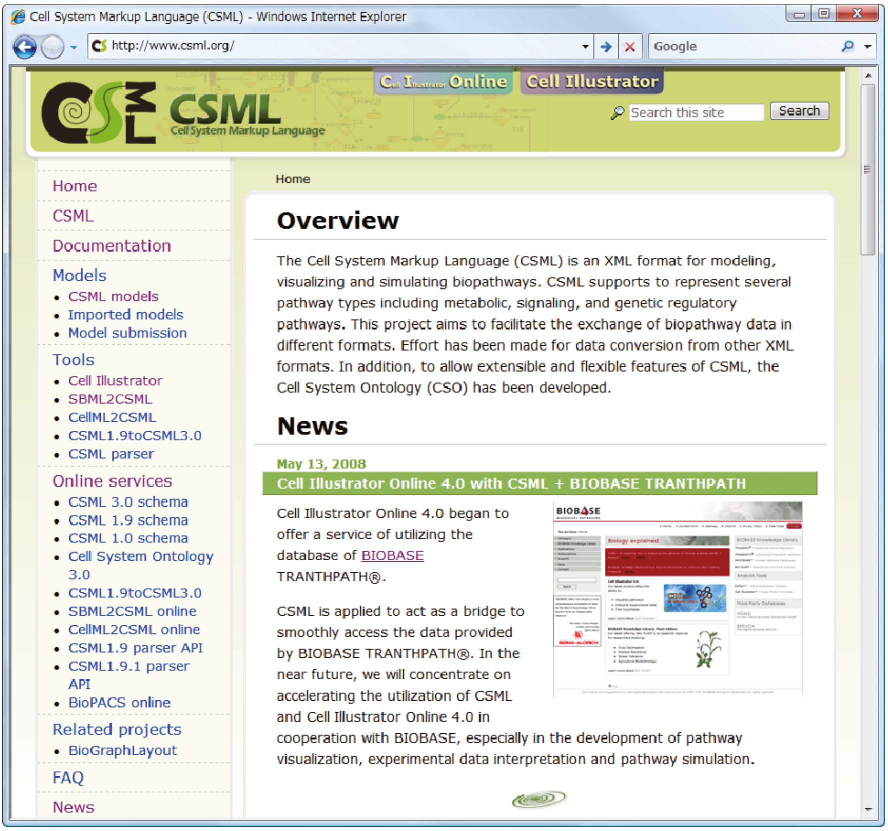
BioPAX was started in 2002 in order to encourage open source formats for pathway information (http://www.biopax.org/). The format is defined by using OWL (an XML type language used to define ontologies). BioPAX level 1 targets information regarding compounds and metabolism. BioPAX level 2 targets molecular relationships and includes information on molecular bindings, phosphorylation sites, posttranslation modifications, as well as experimental data and pathway structures. Discussions on BioPAX level 3 are developing so that it will include gene regulatory and signaling pathways.

2.3.6 CSML/CSO

CSML (Cell System Markup Language) is an XML format designed to define gene regulatory, metabolic, and signaling pathways with regard to system dynamics (http://www.csml.org/). It has been developed at the Human Genome Center of the University of Tokyo.

As of May 2008, CSML 3.0 is the newest version. In addition, CSML is widely extensible and can import the CellML and SBML formats introduced in Sections 2.3.3 and 2.3.4.

Furthermore, in order to achieve a high level of compatibility with other data formats, CSML defines and uses its own ontology format, Cell System Ontology (CSO). CSO is an ontology which effectively describes dynamics and signal pathways not expressible by BioPAX introduced in Section 2.3.5. In addition, CSO defines a large number of standardized icons (over 350) to be used for defining necessary terms and relations (see Chapter 4, Figure 4.37). CSML pathways are displayed in Cell Illustrator—software which will be described in Chapter 3—which uses these icons. CSML models can be downloaded from the above URL (Figure 2.4).



Overview

The Cell System Markup Language (CSML) is an XML format for modeling, visualizing and simulating biopathways. CSML supports to represent several pathway types including metabolic, signaling, and genetic regulatory pathways. This project aims to facilitate the exchange of biopathway data in different formats. Effort has been made for data conversion from other XML formats. In addition, to allow extensible and flexible features of CSML, the Cell System Ontology (CSO) has been developed.

News

May 13, 2008

Cell Illustrator Online 4.0 with CSML + BIOBASE TRANTHPATH

Cell Illustrator Online 4.0 began to offer a service of utilizing the database of BIOBASE TRANTHPATH®.

CSML is applied to act as a bridge to smoothly access the data provided by BIOBASE TRANTHPATH®. In the near future, we will concentrate on accelerating the utilization of CSML and Cell Illustrator Online 4.0 in cooperation with BIOBASE, especially in the development of pathway visualization, experimental data interpretation and pathway simulation.

Fig. 2.4

Springer

http://www.springer.com/978-1-84882-022-7

Foundations of Systems Biology
Using Cell Illustrator and Pathway Databases
Nagasaki, M.; Saito, A.; Doi, A.; Matsuno, H.; Miyano, S.
2009, XII, 155 p. With CD-ROM., Hardcover
ISBN: 978-1-84882-022-7