Minireview Genomics & Informatics Vol. 3(2) 47-51, June 2005

Higher Order Knowledge Processing: Pathway Database and Ontologies

Ken Ichiro Fukuda

Computational Biology Research Center (CBRC), National Institute of Advanced Industrial Science and Technology (AIST), AIST Tokyo Waterfront Bio-IT Research Bldg. 10F. 42 Aomi, Koutou-ku, Tokyo 135-0064, Japan

Abstract

Molecular mechanisms of biological processes are typically represented as "pathways" that have a graph-analogical network structure. However, due to the diversity of topics that pathways cover, their constituent biological entities are highly diverse and the semantics is embedded implicitly. The kinds of interactions that connect biological entities are likewise diverse. Consequently, how to model or process pathway data is not a trivial issue. In this review article, we give an overview of the challenges in pathway database development by taking the INOH project as an example.

Keywords: pathway database, ontology, signal transduction, textual knowledge

Introduction

As the genome of many species has now been sequenced, the target of biological knowledge acquisition has shifted from elucidating the features of genes and proteins to discovering the underlying mechanisms of biological functions, an ensemble of collaborating proteins and other molecules. Not only the combinations of interacting proteins and chemicals but also the relations of each molecule to cellular, physiological or organism level functions have to be identified. Each biological function has a particular combination of these interactions or relations that explains the underlying molecular mechanisms. This set of interactions form a network structure called *pathway*. In this review article, we give an overview of the challenges in pathway database development by taking the INOH (Integrating Network Objects with Hierarchies) project as an example.

*Corresponding author: E-mail fukuda.cbrc@aist.go.jp, Tel +81-3-3599-8049, Fax +81-3-3599-8081
Accepted 30 May 2005

Biological Pathways

Pathway data are "processed" rather than "raw" knowledge or data, and are integrated from multiple knowledge sources. A pathway is a model of the mechanism of a biological function. The model is derived from biologists' interpretation of different experiment results and other pathways reported in the scientific literature. Therefore sharing pathway knowledge means sharing the current understanding of molecular biology.

Signal transduction pathway, which describes the mechanisms of various life phenomena in terms of interacting proteins and other bio-molecules, is a good example of this type of knowledge. Their constituent biological entities are highly diverse and range from metal ions to proteins to biological processes in general. Likewise, the kinds of interactions that connect biological entities are diverse. To unveil a new signal transduction pathway, a biologist has to integrate various types of distilled knowledge.

Since pathway relates physical entities such as proteins and chemicals to biological functions, they have the potential to serve as a new infrastructure for high-throughput experiment analysis, drug target screening, and fundamental molecular biology.

However, pathway data continue to reside primarily in the scientific literature and providing this knowledge in a computable form is crucial. This is the reason why pathway database development and pathway data format standardization becomes very important. It is said that there are over 150 pathway databases currently in the world.

Pathway databases

Higher-order knowledge is a knowledge that has to be described by multiple objects. While an entry in a sequence database describes the attributes of a single biological entity, an entry in a pathway database consists of many different biological entities and interactions. Therefore, implementation of a pathway database is not straight forward. There are two simple way of doing this but neither of them is sufficient. One is to decompose all pathways into their interactions and aggregate them to a

set of binary relations. The other is to prepare a hand-drawn clickable map for each pathway. In binary relation set based databases, a pathway is literally a graph and is defined as a connected component of the interactions. You can apply many graph theoretical queries but this model makes it difficult to define sub-pathways or processes. In other words, it is difficult to annotate contextual information. In illustration based databases, each illustration represents its contextual information and hyperlinks to other databases are embedded in the map. Although you can present a drawing for each pathway as it appears in the literature, this results in very limited computability.

The type of knowledge we want to represent in a pathway database and the type of query we want to conduct on a pathway database are not apparent. Consequently, there are many ongoing pathway database projects, each of which targets different types and levels of knowledge.

Historically, pathway database development started from metabolic pathways (EcoCyc[1], KEGG[2]). Metabolism is largely conserved among species and we have matured consensus knowledge on these pathways, which can be described by chemical reactions and their catalysts in a unified way. Subsequently, several databases attempted to encode other levels of biological functions, such as transcription regulation pathway (TRANSPATH[3]), protein-protein interaction maps (BIND, http://bind.ca/, DIP, http://dip.doe-mbi.ucla.edu/, HPRD, http://www.hprd.org, IntAct, http://www.ebi.ac.uk/intact/index.jsp). This required wider coverage of concepts compared to biochemical pathways.

To relate physical entities to various levels of biological phenomena, for example, cellular processes and disease pathways, we need a framework that is able to handle multiple processes in different granularities. INOH (http://www.inoh.org), PATIKA (http://www.patika.org), aMaze (http://www.amaze.uib.ac.be), and Reactome (http://www.reactome.org) focus on biological processes at various levels, with INOH and Reactome performing manual curation from the scientific literature. These databases are process-oriented in the sense that they use a compound-graph structure. Compound graph is a hierarchical graph in which each node can contain a graph inside itself. This feature makes compound graph suitable for sub-pathways or sub-processes annotations.

Also worth mentioning is BioPAX[1], which is not a database but a data exchange format for pathway data, its current level 1 release is limited to the exchange of metabolic pathway data. However, the roadmap includes higher levels of pathways in level 3 and 4.

1) http://www.biopax.org

Ontologies

In any of the above cases, the curator has to annotate meanings to each pathway object. To provide machine-accessible pathway knowledge that resides in the scientific literature, encoding topological structure of pathways is not sufficient.

For example, it is not easy to specify a single sequence identity automatically for each molecule name that appears in the scientific literature. A molecule name may stand for concepts of various granularities, from concrete objects such as H-Ras and ERK1 to abstract concepts (generic concepts) or categories such as Ras and MAPK or kinase. Usually, biologists have the appropriate background knowledge and know that H-Ras is one of the Ras, and ERK1 is one of the MAPKs. However, computer systems have no such background knowledge. By annotating only the names of molecules, the relation between H-Ras and Ras is lost. Hence, background knowledge that biologists use to interpret pathway diagrams has to be made explicit and available to computers.

To accomplish this, the INOH project provides a set of ontologies for pathway annotation. Each of our ontologies is used to annotate certain types of objects or attributes of objects (biological processes, proteins, chemicals, localization, etc.) in a pathway.

INOH project

INOH is a manually curated signal transduction pathway database of model organisms including human, mouse, rat and others. INOH focuses on biological processes at various levels. It is based on a compound-graph data model and has a hierarchical and recursive structure.

In INOH, a pathway is called an "event". A flat event consists of input/output entities, conversion reaction and its controllers. A compositional event consists of other events (sub-pathways). Fig. 1 is an example of an INOH pathway. Each rectangle is a node (compound node) of this compound graph and these compound nodes have a graph inside themselves. The thick rectangle nodes represent protein complexes and the ovals represent proteins. The dotted rectangles represent biological processes. For example the pathway "JAK-STAT cascade" consists from four sub-processes (four rectangles) and one of these sub-processes has its own sub-process. Each event (dotted rectangle) has inputs which are converted to outputs. The outputs are then "passed" to the next event as inputs.

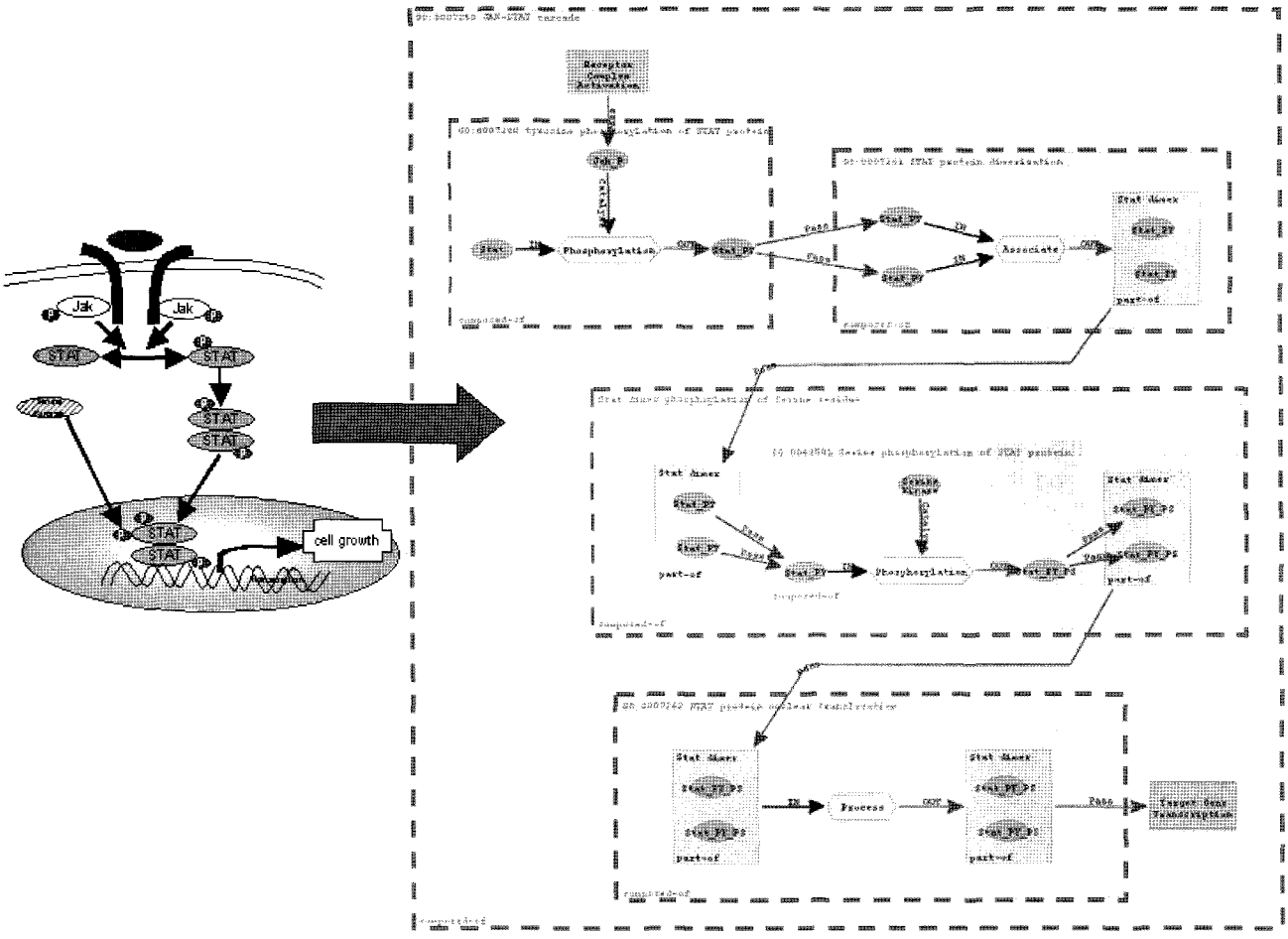

Fig. 1. An INOH pathway. Each rectangle represents a process of different granularity.

As mentioned before, these pathway objects are annotated by a set of ontologies, such as protein name ontology (MoleculeRole Ontology) and event ontology.[2)]

MoleculeRole ontology and Event ontology

The MoleculeRole Ontology is a bio-ontology of protein names and protein family names.

Molecule names in the scientific literature can be classified into the following categories. (a) Concrete-Names: Names specific enough to identify each of their sequences (Grb2), (b) Generic-Names: Names that stand for several sequences (Ras, Raf, MEK, ERK), (c) Complex-Names: Names that refer to Complexes (PI3K), (d) Function-Names: Names that describe only functions without specifying a concrete or abstract molecule name (tyrosine kinase receptor).

The Molecule Role Ontology encodes (1) relations between Function-Names and Generic-Names (e.g. protein serine/threonine kinase and MAPK), (2) relations between Generic-Names and Concrete-Names (e.g. MAPK and ERK1), (3) Complex and its subunit (e.g. PI3-kinase and PI3-kinase p110 subunit), (4) Concrete-Names and UniProt accession numbers. All the terms are arranged manually in a hierarchical structure (a DAG). The UniProt (SwissProt/TrEMBL) accession numbers are identified manually for each Concrete-Name.

The Event Ontology is a DAG structured controlled vocabulary for events. It has four top categories for biological events: (1) "Molecular event" includes molecular level phenomena, (2) "Cellular event", (3) "Organism event", and (4) "Physiological event". The term in the Event Ontology covers terms that is currently not included in the Gene Ontology's bioprocess ontology.

INOH pathway queries

The combination of a hierarchical data model with a set of ontologies for annotation has several advantages

2) INOH ontology can be accessed at http://www.inoh.org

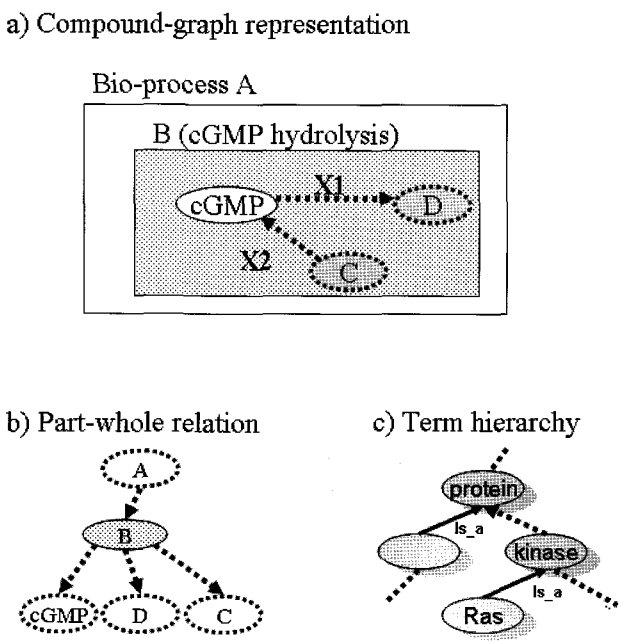

Fig. 2. Translation of a biological query to a compound graph structure. a) Which molecule interacts with cGMP? (C and D via X1 and X2). b) Which biological process requires cGMP? (B), c) Is Ras a kinase? (yes)

compared to a plain graph model.

With ontologies, the user can specify the meaning of the query string and the system can expand the query based on the ontologies. With a compound graph representation, the user can search part-whole relations of sub-pathways as well as node attributes, edge attributes, and their connectivity (Fig. 2).

As a result, our user can query any pathways or biological components of pathways, including sub-pathways, by specifying attributes of nodes, edges, pathways, and their values.

Query search in INOH with ontological support

The ontological knowledge can be utilized during a pathway query. For example, if one finds a protein entity annotated as "MAP kinase", she knows from the ontology that it may indicate ERK1 of a human, JNK1 of a mouse, p38-alpha of a rat, etc, and vice versa.

The unique feature of INOH is its ontology-based query relaxation search. If the user enters "ubiquitin ligase" as a query, the system firstly search for pathway entities that is directly annotated by the character string "ubiquitin ligase" and then relaxes the query by expanding the query term according to the MoleculeRole ontology and will get pathway entities annotated by, for example, "Smurf". Conducting this type of query by a simple keyword search without ontological annotation would be difficult because the system does not know that the "Smurf" is a type of "ubiquitin ligase". The reader should note that a system without ontological annotations will fail to provide a semantically complete result.

Discussion

Compared to sequence data and other conventional biological data, implementing or modeling a pathway database is not straight forward. A clickable-map of a pathway diagram is easy for human to understand, but provides only a very limited computability. On the other, a binary relation based pathway representation is easy for computers to process but difficult for human to interpret.

A modern pathway database should have a mechanism to encode contextual information, i.e. when and where the combination of interactions of a particular pathway emerges.

By adopting a hierarchical and recursive representation model and annotating every biological entity in the model with ontologies, one can provide rich semantics and a powerful querying facility.

Ontological annotations assure that the system returns semantically complete answers.

Since pathway is a processed knowledge, there could be different or conflicting pathways for the same biological phenomena. Pathway database will serve as an infrastructure to overview the current biological knowledge, including contradicting or controversial knowledge. Induction or inference of new pathways, conflict detection would be ambitious and interesting challenges for the Semantic Web and other AI techniques for strongly structured knowledge.

Acknowledgements

This work was supported in part by BIRD of Japan Science and Technology Agency (JST).

References

Karp, P.D., Arnaud, M., Collado-Vides, J., Ingraham, J., Paulsen, I.T., and Saier, M.H. Jr. (2004). The E. coli EcoCyc Database: No Longer Just a Metabolic Pathway Database. *ASM News* 70, 25-30.

Kanehisa, M., Goto, S., Kawashima, S., Okuno, Y., and Hattori, M. (2004). "The KEGG resource for deciphering the genome", *Nucleic Acids Res.* 32, D277-D280.

Krull, M., Voss, N., Choi, C., Pistor, S., Potapov, A., and Wingender, E. (2003). TRANSPATH®: An integrated

database on signal transduction and a tool for array analysis. *Nucleic Acids Res.* 31, 97-100.

Bader, G.D., Donaldson, I., Wolting, C., Ouellette, B.F.F., Pawson, T., and Hogue, C.W.V. (2004). "BIND - Biomolecular Interaction Network Database" *Nucleic Acids Res.* 32, 248-250

Salwinski, L., Miller, C.S., Smith, A.J., Pettit, F.K., Bowie, J.U., and Eisenberg, D. (2004). The Database of Interacting Proteins: 2004 update. *Nucleic Acids Res.* 32, Database issue, D449-D451.

Peri, S., Navarro, J.D., Amanchy, R., Kristiansen, T.Z., Jonnalagadda, C.K. *et al* (2003). Development of human protein reference database as an initial platform for approaching systems biology in humans. *Genome Research* 13, 2363-2371.

Hermjakob, L., Montecchi-Palazzi, C., Lewington, S., Mudali, S., Kerrien, S., Orchard, M., Vingron, B., Roechert, P., Roepstorff, A., Valencia, H., Margalit, J., Armstrong, A., Bairoch, G., Cesareni, D., Sherman, I and R., Apweiler. (2004). IntAct - an open source molecular interaction database. *Nucleic Acids Res.* 32, D452-D455.

Fukuda, K. Yamagata, Y. and Takagi, T. (2003). FREX: a query interface for biological processes with a hierarchical and recursive structures. *In Silico. Biol.* 4, 0007 .

Demir, E., Babur, O., Dogrusoz, U. *et al.* (2004). An Ontology for Collaborative Construction and Analysis of Cellular Pathways. *Bioinformatics* 20, 349-356.

Van Helden, J., Naim, A., Mancuso, R., Eldridge, M., Wernisch, L., Gilbert, D., and Wodak, S.J. (2000). Representing and analysing molecular and cellular function using the computer. *Biol. Chem.* 381, 921-935.

Joshi-Tope, G., Vastrik, I., Gopinathrao, G., *et al.* (2003). The Genome Knowledgebase: A Resource for Biologists and Bioinformaticists. In *Cold Spring Harbor Symposia on Quantitative Biology, Volume LXVIII.* Stillman, B., Stewart, D. ed. (Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York) pp.237-244.

Fukuda, K. and Takagi, T. (2001). Knowledge Representation of Signal Transduction Pathways. *Bioinformatics* 17, 829-837.

Yamamoto, S., Asanuma, T., Takagi, T., and Fukuda, K. (2004). An Ontology for Annotation of Signal Transduction Pathway Molecules in the Scientific Literature: Molecule Role Ontology. *Comp. Funct. Genom* 5, 528-536.

Harris, M.A., Clark, J., Ireland, A., Lomax, J., Ashburner, M., Foulger, R., Eilbeck, K., Lewis, S. *et al.* The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Res.* 32, D258-D261.