The Development of Ontology for Metabolic Pathways Using METHONTOLOGY

Yuslina Zakaria, Safaai Deris, Muhammad Razib Othman
Faculty of Computer Science and Information Systems
Universiti Teknologi Malaysia, Malaysia
Email: lina.zakaria@gmail.com; safaai@fsksm.utm.my; razib@fsksm.utm.my

ABSTRACT

Understanding the biological function of an organism and how they interact to yield a living cell is the major challenge of the post genome-sequencing era. To understand the cellular function, most scientists and biologists study parts of metabolic pathway which contains various kinds of pathways and large volume of knowledge including genes, gene products, chemical compounds, enzyme reactions that interlinked each other. Moreover, the identification and validation of drug targets in drug discovery depends critically on knowledge of the metabolic pathways in which potential target molecules operate within cells. For this reason, the study of biochemical pathways is the focus of numerous drug discovery researches and is central to the stage of many biopharmaceutical and genomic companies. In order to provide better access to relevant knowledge, the knowledge of domain should be conceptualized and formalized by providing a domain-specific vocabulary using ontology-based approach. The methodology for ontology development, METHONTOLOGY is used as the methodology in developing and constructing the ontology for metabolic pathway. This paper presents how to build a domain ontology following METHONTOLOGY and using the ontology conceptualization tools, Protégé.

KEYWORDS

Metabolic pathways, Ontology, METHONTOLOGY, Protégé

1. Introduction

Metabolism is a chemical aspect of life that produces metabolites and energy by chemical reactions catalyzed by enzymes. Metabolic reactions that synthesize basic metabolites such as nucleic acids, carbohydrates, and amino acids are shared by various organisms. Those that synthesize secondary metabolites such as antibiotics, toxins, dyes and hormones are specific to individual organisms. Most of the known secondary metabolites have biological activities that are valuable for medical and agricultural uses [1]. While basic metabolic pathways are well studied and collected in the LIGAND, PATHWAY, REACTION databases in Kyoto Encyclopedia of Genes and Genomes [2, 3], pathways for the biosynthesis of secondary metabolites are only partly or scarcely studied.

Metabolic pathway is reconstructed by predicting intermediate reactions stepwise in the direction reverse to biosynthesis; from a given secondary metabolite to a starting basic metabolite. At each prediction step, more than one reactions are usually proposed as the candidates for the next intermediate reaction because each intermediate metabolite usually have several possible reaction sites in its chemical structure. To select the most probable reaction, knowledge is required to evaluate the chemical reactivity of the candidates [4]. We have analyzed metabolic reactions to obtain chemical knowledge on the reactions in the cells. We are accumulating two different types of chemical knowledge for the prediction of metabolic pathways. The one is related to the substrate and reaction specificity of the enzymes. The other type of chemical knowledge is related to the preference in the chemical reactions in the cells. These two are supplemental to each other.

2. The Challenge of Building Ontologies

Ontology building is a craft rather than an engineering activity. Each development team usually follows its own set of principles, design criteria, and phases. The absence of structured guidelines and methods hinders the development of shared and consensual ontologies within and between teams, the extension of an ontology by others, and its reuse in other ontologies and final applications. We believe that the source of these problems is the absence of an explicit and fully documented conceptual model upon which to formalize the ontology.

Unlike knowledge-based system (KBS) developers, ontology developers (ontologists) lack sufficiently tested and generalized methodologies recommending what activities should be performed and at what stage of ontology development. Ontology developers often switch directly from knowledge acquisition to implementation, which poses these problems:

First, the conceptual models are implicit in the implementation codes. Making the conceptual models explicit usually requires reengineering. Second, ontological commitments [5] and design criteria are implicit and explicit in the ontology code. Third, domain experts and human end users have no understanding of formal ontologies codified in ontology languages. Research has shown that, using the Ontology Server browser tools, experts and users could gain a full understanding of and validate taxonomies and partially understand instances. However, they were unable to understand abstract definitions of concepts, relations, functions, and axioms. Fourth, as with traditional knowledge bases, direct coding of the knowledge-acquisition result is too abrupt a step, especially for complex ontologies. Fifth, ontology-developer preferences in a given language condition the implementation of the acquired knowledge. So, when people code ontologies directly in a target language, they are omitting the minimal encoding bias criterion defined by Tom Gruber [4].

Finally, ontology developers might have difficulty understanding implemented ontologies or even building new ontologies. This is because traditional ontology tools focus too much on implementation issues rather than on design questions. For example, someone who knows how to build ontologies but is unfamiliar with the language in question might have difficulties working at the implementation level. If we are very familiar with the language, understanding existing definitions and writing new definitions are almost impossible. The problem is not to understand that density is equal to mass divided by volume at the knowledge level, but writing this in a target language. Therefore, something that is apparently very simple at the conceptual level is extremely complicated when expressed at the implementation level, if you're not familiar with the language.

This means that ontologies are built exclusively by developers who are perfectly acquainted with the languages in which the ontologies will be implemented. Because these ontologists are not necessarily experts in the domain for which the ontology is built, they spend a lot of time and resources on knowledge acquisition.

METHONTOLOGY [6, 7] and Protégé alleviate some of these problems. METHONTOLOGY provides a user-friendly approach to knowledge acquisition by non-knowledge engineers and an effective, generally applicable method for domain-knowledge-model construction and validation. Protégé will support the ontology building at the implementation by implemented automatically using translators. So, ontologists don't need to know the ontology's implementation language.

3. Ontology Development Methodology

Ontological engineering requires the definition and standardization of a life cycle ranging from requirements specification to maintenance, as well as methodologies and techniques that drive ontology development. So, the METHONTOLOGY [8] framework includes

- The identification of the ontology development process;

- A life cycle based on evolving prototypes; and
- The methodology itself, which specifies the steps for performing each activity, the techniques used, the products to be output, and how the ontologies are to be evaluated.

In developing pathway ontology, our first step was to plan. Then, because we are not experts in the biochemistry domain, we acquired knowledge to put together a preliminary version of the requirements specification. We then simultaneously acquired and conceptualized more knowledge; conceptualization helped guide acquisition. Then, by doing the *intermediate representations*, the conceptualization can be evaluated by experts.

Our experience shows that the specification, conceptualization, integration, and implementation can be performed as often as required. Figure 1 presents the development cycle of the chemical ontology.

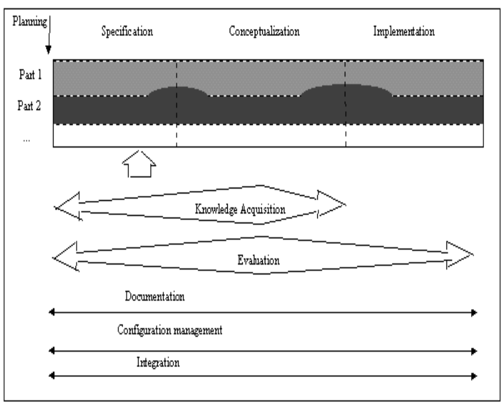

Figure 1. Ontology Development LifeCycle

3.1 Specification

Ontology specification's goal is to put together a document that covers the ontology's primary objective, purpose, granularity level, and scope. This specification should be as complete and concise as possible.

Figure 2 shows a short example of ontology requirements specification document in the chemical domains. We built this specification after acquiring domain knowledge.

Domain: Metabolic Pathway
Date: 21 June 2004
Developed by Yuslina Zakaria

Purpose: Ontology about metabolic pathway to be used when information about all elements in metabolic pathway such as chemical compounds, reactions, transformations and pathways is required in pathway analysis and simulations.

Level of formality: Semiformal.

Scope: List of elements: *purine, amino acid, galactogen, alanine, glycin*, ...
List of concepts: *chemical, generalized-reactions, enzymatic-reactions, pathways, compounds*, ...
Information about at least the following properties: *structure-atoms, structure-bonds, molecular-weight, chemical-weight, reaction-lists, predecessors*, ...

Sources of knowledge:
a) Interviews with expert, PM Dr Rosli Ilyas
b) The following books:
   a. Biochemistry, 5th edition, WH Freeman and Company, New York, 2004.
   b. Biology, 5th edition, Addison Wesley Longman, Inc., 1999.
c) The following websites:
   a. www.ecocyc.org
   b. www.genome.ad.jp/kegg

Figure 2: Ontology requirements specification document for the metabolic pathway ontology

3.2 Conceptualization

Conceptualization organizes and structures the acquired knowledge using external representations that are independent of the implementation languages and environments. Specifically, this phase organizes and converts and informally perceived view of a domain into a semiformal specification, using set of intermediate representations that the domain expert and ontologist can understand.

This set of IRs is based on those used in the conceptualization phase of the ideal methodology for knowledge-based systems development. Figure 3 illustrates the order we follow for conceptualization.

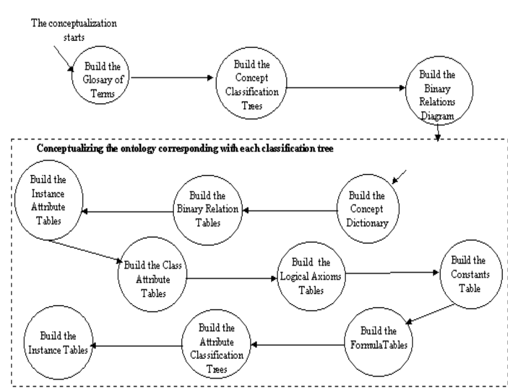

Figure 4: Conceptualization according to METHONTOLOGY

First, we built a *glossary of terms* that includes all the terms (concepts, instances, attributes, verbs, and so on) of the metabolic pathway domain and their descriptions.

When the glossary contained a sizable number of terms, we built *concept-classification trees* using relations such as is-a, part-of and instance-of. So, we identified this domain's taxonomies, and each taxonomy produced an ontology as prescribed by METHONTOLOGY.

The next step was to build ad hoc *binary-relations diagrams* between concept-classification trees. These diagrams establish relationships between concepts of the same or different ontologies. They will set out the guidelines for integrating ontologies, because if a concept C1 is linked by a relation R to a concept C2, the ontology containing C1 includes the ontology containing C2, provided that C1 and C2 are in different concept-classification trees. For each concept-classification tree generated, we built these IRs:

The *concept dictionary* contains all the domain concepts, instances of such concepts, class and instance attributes of the concepts, and optionally, concept synonyms and acronyms.

A *binary-relations table* specifies the name, the names of the source and target concept, the inverse relation, and so on for each ad hoc relation whose source concept is in the concept-classification tree.

An *instance-attribute table* describes each instance attribute in the concept dictionary. Instance attributes are defined in the concept but that take values in its instances.

The other tables and diagrams that we have built are *class-attribute table*, *logical-axioms table*, *constants table*, *formula table*, *attribute-classification tree and finally*, *instance table.*

3.3 Knowledge Acquisition

This is an independent activity within ontology development. However, it coincides with other activities.

3.4 Integration

Throughout ontology development, we identified terms that could be included from other ontologies such as EcoCyc [9], KEGG [10] and Gene Ontology [11].

3.5 Implementation

Finally, based on conceptualization using IRs, Protégé formalize the ontology and automatically generated CLIPS to transform the conceptual model into an implemented model (see Figure 5).

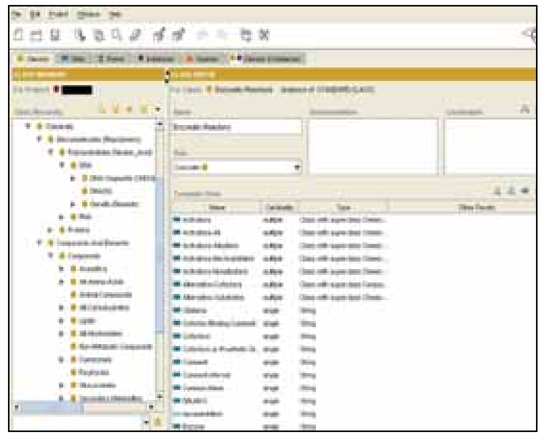

Figure 5: Ontology implementation using Protégé

4. Conclusion

Our experience shows that domain experts and human end users understand and validate the most of the METHONTOLOGY IRs. METHONTOLOGY is a user-friendly approach to knowledge acquisition by domain experts who are not knowledge-engineers. So, all the ontologies built using this approach are not hand-crafted; they rely on same conceptualization process; and they have been built independently of their end use.

5. References

[1] Karp, P.D., Krummenacker, M., Paley,S. and Wagg, J. (1999) Integrated pathway/genome databases and their role in drug discovery. *Trends Biotechnol.*, 17, 275-281.

[2] Goto, S., Bono, H., Ogata, H., Fujibuchi, W., Nishioka, T., Sato, K., and Kanehisa, M., Organizing and computing metabolic pathway data in terms of binary relations, *Pacific Symposium on Biocomputing '97*, (Ed. by Altman, R. B. and Dunker, A. K. and Hunter, L. and Klein, T. E.), World Scientific, 175-186, 1997.

[3] Goto, S., Nishioka, T., and Kanehisa, M., LIGAND database for enzymes, compounds, and reactions, *Nucleic Acids Research*, 27. pp 377-379, 1999.

[4] Karp, P. D. and Riley, M. (1994) Representations of metabolic knowledge: pathways. In Altman,R., Brutlag,D., Karp.,P., Lathrop,R. and Searls,D. (ed.) *Second International Conference on Intelligent Systems for Molecular Biology*. AAAI Press, Menlo Park, CA.

[5] Farquhar, A. et al. Collaborative ontology construction for information integration, T*ech. Report KSL-95-69*, Knowledge System Laboratory, Stanford Univ., Stanford, California, 1995.

[6] Uschold, M., and Gruninger, M.,. Ontologies: principles, methods and applications. *Knowledge Engineering Review*, Vol. 11, No. 2, June 1996.

[7] Gruninger, M., Fox, MS., Methodology for the design and evaluation of ontologies. *Proc. International Joint Conference AI Workshop on Basic Ontological Issues in Knowledge Sharing*, 1995.

[8] Fernandez, M., Gomez-Perez, A., and Juristo, N., Methontology: from ontological art towards ontological engineering. *Proc. AAAI Spring Symposium Series*, AAAI Press, Menlo Park, California, 1997, pp. 33-40.

[9] Karp, P. D., M. Riley, M. Saier, et al. (2000). The EcoCyc and MetaCyc databases. *Nucleic Acids Res* 28(1): 56-9.

[10] Ogata, H., Goto, S., Sato, K., Fujibuchi, W., Bono, H., and Kanehisa, M.; KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res*. 27, 29-34. 1999.

[11] Ashburner, M., C. A. Ball, et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature Genetics* 25(1): 25-9. 2000.