INOH: A Textual Knowledge Based Pathway Database

Satoko Yamamoto[1] syamamot@hgc.jp
Tatsuya Kushida[1] kushidat@hgc.jp
Naotaka Ono[1] onon@hgc.jp
Yuki Yamagata[1] snowfox@hgc.jp
Toshihisa Takagi[2] tt@k.u-tokyo.ac.jp
Ken-ichiro Fukuda[3] fukuda-cbrc@aist.go.jp

[1] Institute for Bioinformatics Research and Development (BIRD), Japan Science and TechnologyAgency (JST), 3-14-4 Shirokane-dai, Minato-ku, Tokyo 108-0071, Japan
[2] Graduate School of Frontier Sciences, University of Tokyo, 5-1-5 Kashiwanoha, Kashiwa-shi, Chiba 277-8562, Japan
[3] Computational Biology Research Center (CBRC), National Institute of Advanced IndustrialScience and Technology (AIST), 2-43 Aomi, Koutou-ku, Tokyo 135-0064, Japan

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

1 Introduction

Completions of the human and mouse genomes have shifted the target of biological knowledge acquisition from features of each bio-molecule to higher order knowledge such as relationships among multiple bio-molecules that constitute signal transduction pathways (STPs). Most part of this knowledge resides in scientific articles, and making it computer-accessible becomes increasingly important. However, these kinds of biological entities, and the relations between them, are highly diverse, e.g. metal ions, chemicals, proteins and phenomena such as “apoptosis” and “transport”, “control” and “reactions”. To make this knowledge computable, a strongly structured ontology-aware system has to be developed.

In this study, we report our ontology of signaling molecule functions and a GUI (Graphical User Interface) for complex data input and the data curation process. Further, a STP database named INOH based on these studies is presented.

2 Signal Molecule Family Ontology

In the scientific literature, especially in review articles, some biological terms represent abstract, conceptual molecules that are used for unspecified organisms. Usually, by relating these terms to instances of the molecule and to other terms, biologists can interpret the biological context. For example, the term ”MAP-kinase” indicates ERK1 of a human, JNK1 of a mouse, p38alpha of a rat, etc. The term also means that it is one kind of ser/thr protein kinase. Such background knowledge should be stored in the computer before we develop a pathway database based on the literature. We collected the names of abstract molecules from the literature, and developed an ontology for the signal molecule family (sig-family). Its classification is not identical to the families derived from sequence analysis; our ontology is constructed from conceptual classification of molecular functions. It contains the class name (for example, “ligand” and “adaptor protein”), which is not concerned with similarity of sequence but, rather, is grouped based on the same role in signal transduction. In addition, each term of the sig-family is linked to SWISS-PROT [2] entries and Gene ontology (GO) [1]. In our database model, each node that represents a conceptual molecule in the pathway graph is associated with a sig-family. Therefore, it is possible to search the pathways and compare similarities between pathways without deviating from the context of biology.

3 Graphical Editor GOEMON

In order to classify many biological concepts related to signal transduction reported in the literature by using compound graph expression [3], we developed a graphical editing tool, the Graphical Ontology Editor for MOlecular Network (GOEMON). In response to a user's request, objects required for compound graph expression can be freely defined in the GOEMON system. Using this tool, we can express the relevance of many concepts of an object field in the compound graph scheme, and, at the same time, the classification of the vocabulary required for each graph expression is attained. We try to describe the graph expression of some typical STPs and to classify the biological vocabulary accompanying them with the GOEMON system.

4 Curation Process

First, molecule-molecule interactions described in review articles of the scientific literature were extracted. Next, every sub-process, which constitutes a STP, was arranged into a hierarchy to form a compound graph. Finally, we annotated every object in the compound graph with biological ontologies. To do this, important concepts relevant to STP knowledge such as localization, biological function/process and chemicals were integrated. Each concept has links to other databases or ontologies such as GO and KEGG (Kyoto Encyclopedia of Genes and Genomes) [5]. Abstract molecule functions required for biological processes over multiple species can be annotated by the sig-family ontology. By grouping smaller sub-processes, e.g. enzyme reactions into a larger STP process, it became possible to formalize and describe the previously fragmented and incomplete STP knowledge. In addition, with the extensive use of ontologies, one can query the various relationships among different concepts such as processes, abstract molecules and concrete ones.

5 Discussion

To encode textual knowledge in a computable form, it is important to encode the intention of each description explicitly with ontologies, especially in the signal transduction domain where heterogeneous concepts form complex relationships. The database described herein achieves a powerful querying facility that is difficult to realize in a traditional clickable-map or keyword-search based database. In addition, various inference mechanisms to infer conflicts or cross-talk would become possible to implement [4].

6 Acknowledgments

This work was supported in part by BIRD of Japan Science and Technology Agency (JST), and Grant-in-Aid for Scientific Research on Priority Areas "Genome Information Science" from the Ministry of Education, Culture, Sports, Science and Technology of Japan.

References

[1] Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., Harris, M.A., Hill, D.P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J.C., Richardson, J.E., Ringwald, M., Rubin, G.M., and Sherlock, G., Gene ontology: tool for the unification of biology. The Gene Ontology Consortium, *Nat. Genet.*, 25:25–29, 2000.

[2] Boeckmann, B., Bairoch, A., Apweiler, R., Blatter, M.C., Estreicher, A., Gasteiger, E., Martin, M.J., Michoud, K., O'Donovan, C., Phan, I., Pilbout, S., and Schneider, M., The Swiss-Prot protein knowledgebase and its supplement TrEMBL in 2003, *Nucleic Acids Res.*, 31:365–70, 2003.

[3] Fukuda, K. and Takagi, T., Knowledge representation of signal transduction pathways, *Bioinfomatics*, 17:829–837, 2001.

[4] Fukuda, K. and Takagi, T., Signal transduction pathways and logical inferences, *Proc. of the 2001 International Conference on Mathematics and Engineering Techniques in Medicine and Biological Science, METMBS' 2001*, 297–303, 2001.

[5] Kanehisa, M., Goto, S., Kawashima, S., and Nakaya, A., The KEGG databases at GenomeNet, *Nucleic Acids Res.*, 30:42–46, 2002.