Pathway Construction and Extension Using Natural Language Processing

Hong-Woo Chun, Sung-Jae Jung, Mi-Nyeong Hwang, Chang-Hoo Jeong, Sa-Kwang Song, Seungwoo Lee, Sung-Pil Choi, and Hanmin Jung

Korea Institute of Science and Technology Information
{hw.chun,sjjung,mnhwang,chjeong,esmallj,swlee,spchoi,jhm}@kisti.re.kr

Abstract. Construction and maintenance of signaling pathway is a time-consuming and labor-intensive task. In addition, integration of various pathways is also ineffective since several markup languages are used to express pathways. To overcome these limitation, automatic pathway construction and extension with a standard format may provide a solution. The proposed approach has constructed a gold standard corpus that describes the signaling pathways, and it has been used to training and evaluating the automatic pathway construction and extension. Moreover, a standard format to express the signaling pathways has been developed and has been used to express the previous major 10 signaling pathways. An effective visualization tool has been also developed for the standardized format as well. The visualization tool can help to construct pathways and extend the current pathways using all articles in PubMed.

1 Introduction

The signaling pathway indicates a group of molecules in a cell that work together to control one or more cell functions. Such pathways specify mechanisms that explain how cells carry out their major functions by means of molecules and reactions. Since many diseases can be explained by defects in pathways, information from pathways provides useful source to develop new drugs. There are some limitations in construction of pathways. (1) Mostly pathways are constructed by manual methods. Biologists have to read many articles and construct a pathway. (2) The curation of a constructed pathway also requires regularly monitoring of up-to-date publications. (3) The number of publicly available pathways is not sufficient and many useful pathways can be used after paying expensive license fee. (4) Since there is no standard format to express pathways even though there exist popular description formats, a total search is difficult. Automatic pathway construction and extension may overcome these limitations in the pathway construction and maintenance.

2 Related Work

Kyoto Encyclopedia of Genes and Genomes (KEGG), one of the major pathway databases, provides a lot of metabolic and signaling pathways and all pathways

S. Yamamoto (Ed.): HIMI/HCII 2013, Part II, LNCS 8017, pp. 32–38, 2013.

Fig. 1. Event ontology

are expressed by KGML format [2]. Reactome contains 1,300 human pathways related with metabolism, signaling, gene regulation, and other biological processes and pathways in Reactome are provided by SBML and BioPAX format [3]. SRI's BioCyc consists of EcoCyc, MetaCyc HumanCyc and BSubCyc, and the number of pathways is 361, 2142, 303 and 279 respectively. SBML, BioPAX are the method to describe pathways [4].

All pathway databases in the previous work are popularly used in various applications and the most process of pathway construction have been done by a manual method. In addition, since all pathway databases have not been integrated, they cannot be searched at one time and they have constructed the pathways that are partially same pathways with other pathway databases. We aim to integrate pathways that are described with various expression methods and develop an automatic pathway construction and extension system.

3 Construction of a Gold Standard Corpus

To construct and extend pathways automatically, a gold standard corpus has been constructed. Types of annotation are as follows: (1) Target entities contain protein, gene, chemical compound and complex. (2) 24 target relations (1) have been selected from Systems Biology Ontology with respect to relations in the signaling pathway.

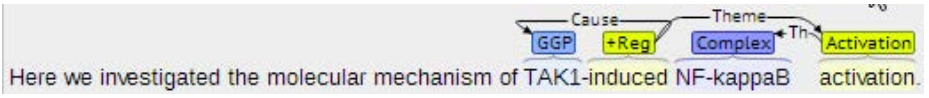

Fig. 2. An annotation example

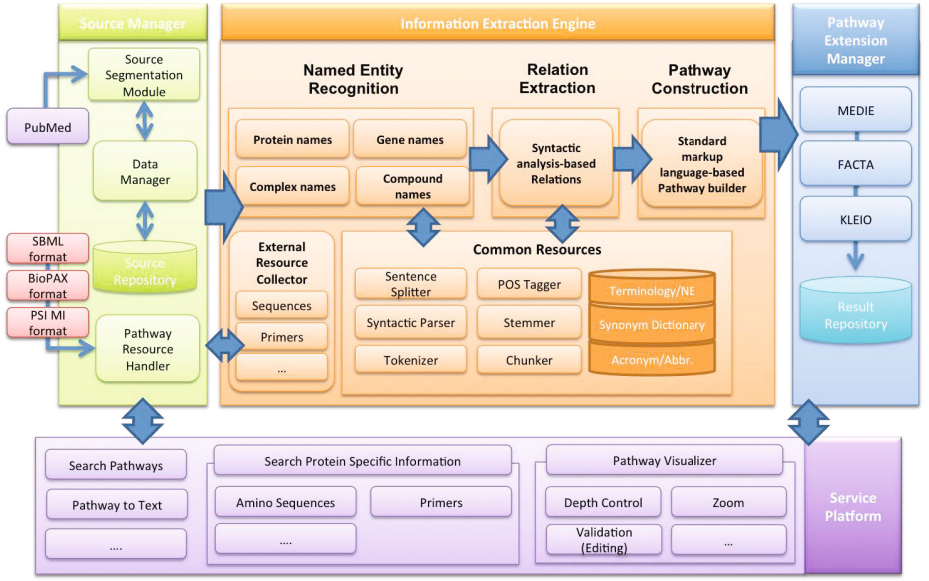

Fig. 3. Architecture of Pathway Construction and Extension

Figure 2 describes an annotation example. Brat version 1.3 had been used to annotate the gold standard. It is popular in the field of natural language processing and effective to express entities and their relations. 570 PubMed abstracts were selected with respect to p53, NF Kappa B that were popular issues in the biomedical domain. Four annotators have participated in the tasks of recognition of entities and their relations. The inter-annotator agreement were calculated by F-measure that is one of the popularly used method for the structured data, and showed 61.0%. One annotation result was regarded as the gold standard and other three annotation results were evaluated in each iteration. Four F-measures were calculated by four times and their average was calculated.

4 Pathway Construction and Extension

To construct and extend pathways automatically, natural language processing-based text mining techniques were applied. Figure 3 describes the system architecture for pathway construction and extension. Meaningful biomedical information was recognized and extracted from PubMed abstracts. Machine

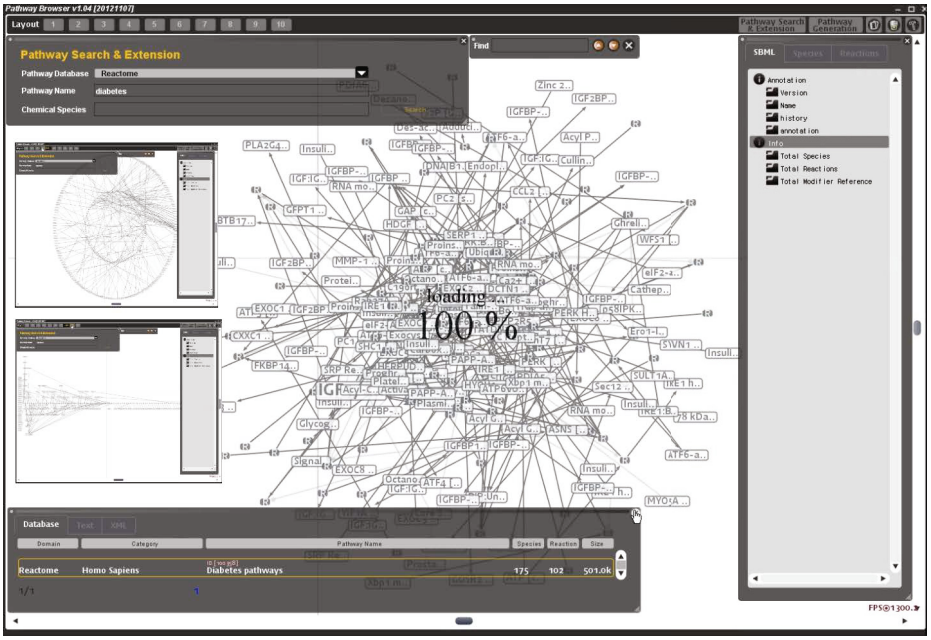

Fig. 4. Pathway Browser

learning technique-based *Named Entity Recognition* [5] and *Relation Extraction* [6] [7] [8] are used, and pathways have been constructed based on the extracted entities and their relations. All pathways constructed in the proposed approach are expressed by a standard markup language. To extend a part of a current pathway, all current pathways were converted the standard markup language. *Pathway extension manager* provides the evidential contexts that support to explain a part of pathways or a whole pathway. The collection methods of the evidential contexts include MEDIE [9], FACTA [10], KLEIO [11]. *Common Resources* are useful techniques and data to recognize entities and their relations, and Natural Language Processing (NLP) tools such as a sentence splitter, a syntactic parser [12] as well as language resources such as technology dictionaries, synonym dictionaries, verb dictionaries, acronym dictionaries, relation pattern dictionaries. The common resources are used in both NER and RE part. Pathway search, advanced search for proteins, and pathway visualization would be possible services in the *Service Platform*.

To visualize pathways, a effective network browser were developed (Figure 4). There are two functions: (1) Pathway search and extension; (2) Pathway generation. A keyword search is possible in the pathway search and extension service. To improve readability and in the point of information delivery, 10 visualization types were applied. Information of species and their hierarchy is displayed in the right panel.

Figure 5 describes the automatic pathway extension with two species (ATP and ADP) in the current pathway. More species selection makes to search more

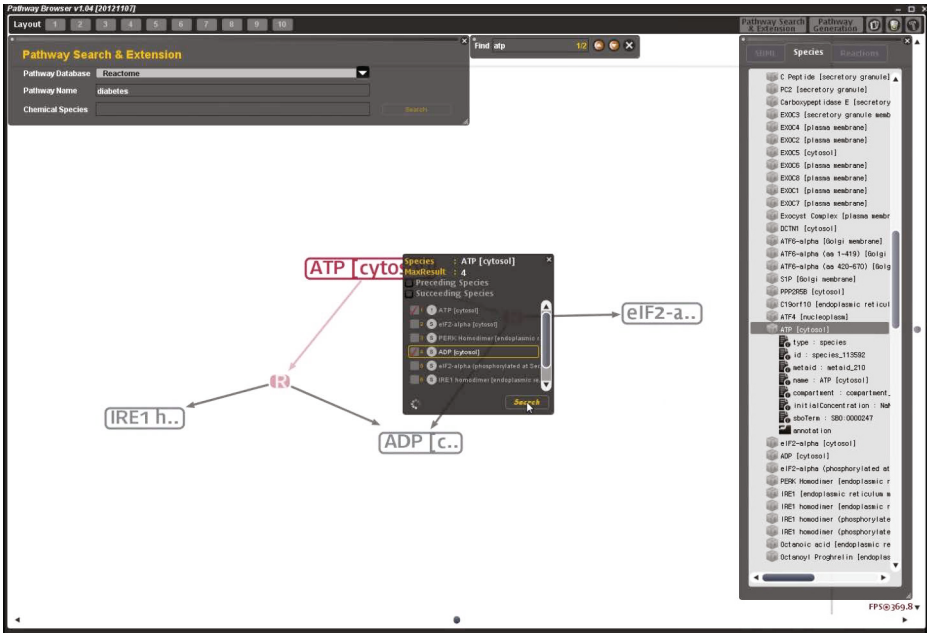

Fig. 5. Architecture of Pathway Construction

specific and perfect candidate pathways. Figure 6 shows the results of the pathway extension in Figure 5. Four groups indicate four different PubMed abstracts and the each abstract can be checked. Each species or each abstract can be easily removed by "Delete" button in the keyboard. From the provided candidates, users can select the right pathways.

The result of the automatic pathway generation using one or more species is the same with Figure 6. The difference is that the pathway generation is started with one or more species that users are interested in.

5 Standardization of Pathway Markup Language

The proposed approach uses Natural Language Processing (NLP) techniques to recognize and extract knowledge from biomedical literature. From news articles, magazines, patents as well as PubMed articles relations among protein, gene, chemical compound, complex are extracted and integrated those knowledge into the existing pathways that are expressed by various formats containing SBML, BioPAX, PSI-MI formats. Based on relational knowledge pathways can be constructed. Currently, 10 pathway databases are integrated: *Reactome, Human Cyc, Panther, Signal Link, SABIO-RK, PharmGKB, KEGG, BioModels, HPRD* and *DIP*. Table 1 describes various pathway markup languages for pathways.

The various pathway description methods is ineffective to integrate knowledge in the various pathways. Many similar pathways have been constructed since the

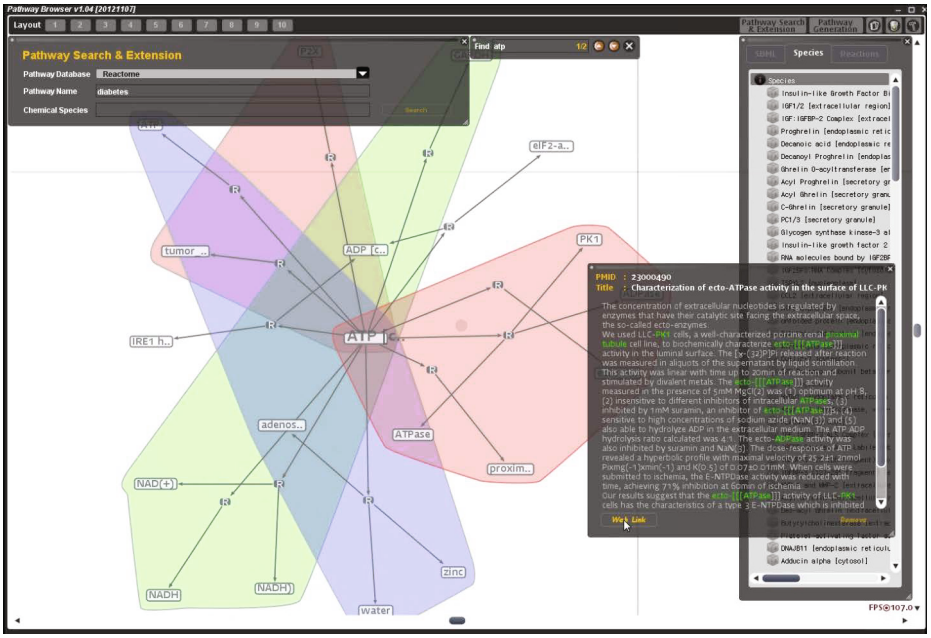

Fig. 6. Architecture of Pathway Construction

Table 1. Various Pathway Markup Languages

Pathway	Markup Language
Reactome	BioPAX, SBML
Human Cyc	BioPAX, SBML
Panther	BioPAX, SBML
Signal Link	SMBL
PharmGKB	PharmGKBML
KEGG	KGML, BioPAX
BioModels	BioPAX, SBML, CellML
HPRD	PSI-MI
DIP	PSI-MI

integration of the current pathways was complicated. The proposed approach has developed a standard format and converters from various markup languages to the standard format. Since BioPAX has more information compared with other markup languages, some information was omitted in the proposed standard format based on pathway construction.

6 Conclusion

Pathways are definitely meaningful information to analyze protein functions or protein process. The analysis can be applied to develop new treatments for such

diseases. In the viewpoint of construction and maintenance of pathways, the process is labor-intensive and time consuming. The proposed approach aims to develop an automatic pathway construction and extension system. Auto generation and extension of pathways might overcome the limitations, and the performance of pathway auto-generation and auto-extension showed encouraging results. We have planned to integrate more pathway databases and to apply SBGN (Systems Biology Graphical Notation) for displaying pathways.

References

1. Ananiadou, S., Kell, D.B., Tsujii, J.: Text mining and its potential applications in systems biology. Trends Biotechnology 24, 571–579 (2006)
2. Kanehisa, M., Goto, S., Sato, Y., Furumichi, M., Tanabe, M.: KEGG for integration and interpretation of large-scale molecular data sets. Nucleic Acids Research 40, D109–D114 (2012) (Database issue)
3. Joshi-Tope, G., Gillespie, M., Vastrik, I., DEustachio, P., Schmidt, E., Bono, B., Jassal, B., Gopinath, G.R., Wu, G.R., Matthews, L., Lewis, S., Birney, E., Stein, L.: Reactome: a knowledgebase of biological pathways. Nucleic Acids Research 33, D428–D432 (2005) (Database issue)
4. Karp, P.D., Ouzounis, C.A., Moore-Kochlacs, C., Goldovsky, L., Kaipa, P., Ahrén, D., Tsoka, S., Darzentas, N., Kunin, V., López-Bigas, N.: Expansion of the BioCyc collection of pathway/genome databases to 160 genomes. Nucleic Acids Research 33(19), 6083–6089 (2005)
5. Song, S.-K., Choi, Y.-S., Chun, H.-W., Jeong, C.-H., Choi, S.-P., Sung, W.-K.: Multi-words Terminology Recognition Using Web Search. In: Kim, T.-H., Adeli, H., Ma, J., Fang, W.-C., Kang, B.-H., Park, B., Sandnes, F.E., Lee, K.C. (eds.) UNESST 2011. CCIS, vol. 264, pp. 233–238. Springer, Heidelberg (2011)
6. Chun, H.-W., Jeong, C.-H., Song, S.-K., Choi, Y.-S., Choi, S.-P., Sung, W.-K.: Relation Extraction Based on Composite Kernel Combining Pattern Similarity of Predicate-Argument Structure. In: Kim, T.-h., Adeli, H., Ma, J., Fang, W.-c., Kang, B.-H., Park, B., Sandnes, F.E., Lee, K.C. (eds.) UNESST 2011. CCIS, vol. 264, pp. 269–273. Springer, Heidelberg (2011)
7. Choi, S.P., Myaeng, S.H.: Simplicity is better: revisiting single kernel PPI extraction. In: Proceedings of the 23rd International Conference on Computational Linguistics, pp. 206–214 (2010)
8. Chun, H.-W., Jeong, C.-H., Song, S.-K., Choi, Y.-S., Jeong, D.-H., Choi, S.-P., Sung, W.-K.: Smart Searching System for Virtual Science Brain. In: Zhong, N., Callaghan, V., Ghorbani, A.A., Hu, B. (eds.) AMT 2011. LNCS, vol. 6890, pp. 324–332. Springer, Heidelberg (2011)
9. Miyao, Y., Ohta, T., Masuda, K., Tsuruoka, Y., Yoshida, K., Ninomiya, T., Tsujii, J.: Semantic Retrieval for the Accurate Identification of Relational Concepts in Massive Textbases. In: Proceedings of COLING-ACL, pp. 1017–1024 (2006)
10. Tsuruoka, Y., Tsujii, J., Ananiadou, S.: FACTA: a text search engine for finding associated biomedical concepts. Bioinformatics 24(21), 2559–2560 (2008)
11. Nobata, C., Cotter, P., Okazaki, N., Rea, B., Sasaki, Y., Tsuruoka, Y., Tsujii, J., Ananiadou, S.: Kleio: a knowledge-enriched information retrieval system for biology. In: Proceeding of ACM SIGIR, pp. 787–788 (2008)
12. Miyao, Y., Tsujii, J.: Feature Forest Models for Probabilistic HPSG Parsing. Computational Linguistics 34(1), 35–80 (2005)