

Text Data Mining for Automatic Gene Ontology Extension^a

Jin-bok Lee and Jong C. Park^b
Computer Science Department and AITrc
Korea Advanced Institute of Science and Technology (KAIST)
373-1 Kusong-dong, Yuseong-gu, Daejeon 305-701 South KOREA
{jblee,park}@nlp.kaist.ac.kr

Bioinformatics is the research discipline that deals with the vast amount of discovered knowledge in a biomedical domain through computational techniques. Many researchers in bioinformatics are investigating ways of combining the construction of useful knowledge bases and the extraction of information from biomedical literature.

We have earlier developed a system for pathway identification, that automatically extracts events including protein-protein interactions from biomedical literature (Park et al.¹; Park²). The system extracts events including interactions such as interact, associate, bind and inhibit. It utilizes a bidirectional incremental parsing technique with validation for the grammaticality of noun phrases in a Combinatory Categorical Grammar (CCG) framework with relatively high precision. The extracted results are binary relations that can be directly utilized, as discussed in the present paper, to extend gene ontology, among many other uses.^c

We have then realized that the extracted results must be informative enough for a more valuable use in the discovery of non-trivial and novel information in biology, including the relationships at the molecular, biochemical and cellular levels. In other words, in extracting relations, we must also extract the conditions on such relations. Consider the following sentences.

- (1) HEC inhibits the proteolysis of metotic cyclin B *in vitro*. (PMID: 9295362)
- (2) *Autophosphorylated* DNA-PK dissociates from Ku:DNA. (PMID: 8621537)

In sentence 1, ‘*in vitro*’ is the condition that makes the inhibition possible. Conditions of this kind appear mainly in adverbial phrases, prepositional phrases, noun phrases or subordinate clauses within the matrix sentences.^d In sentence 2, ‘phosphorylation’ is the condition, or ‘condition of modification’ (cf. Kohn⁴), that presents the characteristic of ‘DNA-PK’. Conditions of this kind appear mainly

^a The present work is funded by the Korea Science and Engineering Foundation through AITrc.

^b presenting and corresponding author

^cIn this regard, our system is different from that presented by Pustejovsky et al.³, which extracts only inhibit relations and accepts even those binary relations with a single argument, such as ‘the inhibition of X’ or ‘the X inhibitor’. In contrast, our system extracts only those binary relations with both arguments explicitly asserted in the original sentence, excluding those exclusively in the noun phrases and those with one argument missing.

^dFor the data analysis of this and of the following claim, we have examined more than 100 sentences with various types of conditions, to be detailed in an extended version and the presentation of the paper.

in adjectives, prepositional phrases or relative clauses. In order to extract these additional types of conditional information, we need to extend the bidirectional incremental parsing to full parsing in a CCG framework, or in any grammar-based framework for that matter. In order to control the accompanying overhead in parsing, however, our modified system identifies the conditions by first locating the phrases that modify either the extracted relation itself or any of its arguments. Due to the complexity of the types of conditions and of the locations that they appear in the written text, the implemented system works at present on fairly limited types of conditions, but we enlist some of the lessons below.

We are also working on making use of the *conditioned* relations to extend Gene Ontology (GO), among other ontologies, that contains descriptions on molecular function, biological process and cellular component of gene products from eukaryotes and uses mainly the ‘is-a’ and ‘part-of’ relationships between concepts. Biomedical ontologies such as GO are semantic models that encode domain-specific knowledge, including shared vocabularies, classification of concepts, relationships among objects and further information in a biomedical domain. However, existing biomedical ontologies (Stevens et al.⁵; Altman et al.⁶; GO Consortium⁷) do not yet contain extensive enough knowledge primarily due to the overhead related to manual construction. We believe that it is thus very important to be able to extend existing ontologies automatically, yet reliably, with the help of the information extracted from biomedical literature, in order to cope with the explosive growth of the extent of the knowledge in a biomedical domain. The extracted events in our system work as key relations in extending the given ontology.

The most important feature of an automatic ontology extension system would be the guarantee for consistency in the resulting body of ontology. It is not straightforward to maintain consistency in the present setting, however, due to the variations in the use of biomedical terms and inconsistencies among the findings in the literature. We propose below how to enrich the gene ontology in a consistent manner in four steps.

First, information from GO is incorporated into a relational database, which includes only GO terms and hierarchical relations. Second, our system incorporates the extracted events including not only protein-protein interactions but also protein-gene interactions, cell signaling and so on, thus providing useful extension points for GO that deals mainly with gene-products. At this step, our system finds identity relations (acronyms/synonyms) by analyzing vocabularies using pattern matching with regular expressions. Third, our system attempts to further extend the database of the extended ontology by finding implicit relations in the extended database to increase the degree of inter-connectedness among the entities in the extended Gene Ontology. At this step, it performs a morphology analysis for each entry name in the database to find ‘functor-of’ or ‘function-of’ relations – *e.g.* ‘*peptidase*’ is functor (*enzyme*) of ‘*peptide*’ and a structure analysis between entry names in the database to find ‘is-a’, ‘part-of’ or identity relations – *e.g.* ‘*an IGF-I receptor monoclonal antibody*’ is the same expression as ‘*a monoclonal antibody to*

the IGF-I receptor'.

Last, our system validates the extended ontology for consistency. It scans all the relations in the accumulated database to report possibly erroneous information including different relations with respect to the same pairs of terms in the extended ontology. Consider the following pair of sentences in the literature.

- (3) In conclusion, our results suggest that *PKC epsilon* stimulates *Raf-1* indirectly by inducing the production of autocrine growth factors. (PMID: 9416835)
- (4) However, using coexpression experiments in Sf-9 cells and transiently transfected A293 cells we did not obtain any evidence for a direct activation of *Raf-1* by *PKC epsilon*. (PMID: 9416835)

While the extracted relations between 'PKC epsilon' and 'Raf-1' are 'stimulate' and 'activate' from sentences 3 and 4, respectively, sentence 4 does not actually support the extraction of the latter, due to the negative mode of the matrix sentence '*did not obtain*'. In this particular case, we can curate the extended database by noticing the fact that the finding from sentence 3 overrides that from sentence 4, though, again, in this particular case, no real post-action should be taken, due to the incorrect extraction of the positive relation from the second sentence. When the system encounters both positive and negative events with respect to the same terms, such as *active* and *inhibit*, the system consults further information, such as the basic sentence structure, confidence levels of journals or authors, frequencies of reports about that relation and so on, and makes suggestions for correction.

This validation step is necessary to build a usable resource because our system extracts about 10% of relation information incorrectly. If domain experts curate the automatically extended ontology conveniently through the validation step, the ontology becomes consistent and usable. However, if there is a lot of faulty information in the extended ontology, it is no use to validate it in such a manner, i.e., in a batch style, because of possibly confusing criteria. There is another way to enforce consistency during the extension process of the given ontology. GO is growing very slowly because domain experts check manually for the relevance of information to be added before adding it. If the system helps the users to make an incremental modification, the system has the effect of processing the incorporation step and the validation step in tandem to populate reliable information into the database. We believe that this method provides more convenience to the user.

The extended ontology includes relations between objects on proteins or genes, where the information is obtained from a large body of biomedical literature that reports experimental results by a much larger number of domain experts in various experimental environments. Inconsistencies in the literature are bound to occur. First, there may be many fragmentary reports on relations between groups of objects in a biomedical domain.^e For example, when there are relations such as

^eZabell and Post⁸ point out that each complex is conformationally relaxed by molecular mechanics to optimize the interaction. It is thus natural to expect that each biomedical research group reports only some part of interactions according to their experimental environment.

‘*A* inhibit *P*’, ‘*B* inhibit *P*’ and ‘*C* inhibit *P*’ in the extended ontology, it may be possible that a group of *A*, *B* and *C* together inhibits *P*.^f Second, there is the possibility that, since a protein may have multiple functions with respect to conditions, the protein may be classified as different kinds of proteins.^g We expect that these problematic cases can be dealt with by constructing informative and consistent knowledge bases such as the extended ontology from various kind of useful information extracted from biomedical literature.

In this abstract, we have described a possibility of applying text data mining techniques to the automatic extension of a biomedical ontology with a system implemented in Perl and MySQL. We are currently working on focusing on a particular group of interesting interactions, as a way of proving that the presented method works reliably and fruitfully. For this purpose, we are also looking into utilizing a version of Description Logic in order to encode the extracted interactions so that consistency becomes a derivative of the soundness of the logic-based inference system.

References

1. J. Park, H. Kim, and J. Kim. Bidirectional Incremental Parsing for Automatic Pathway Identification with Combinatory Categorical Grammar. In *Proceedings of PSB*, 2001.
2. J. Park. Using Combinatory Categorical Grammar to Extract Biomedical Information. *IEEE Intelligent Systems*, 16(6):62–67, 2001.
3. J. Pustejovsky, J. Castano, and J. Zhang. Robust Relational Parsing over Biomedical Literature: Extracting Inhibit Relations. In *Proceedings of PSB*, 2002.
4. Kurt W. Kohn. Molecular Interaction Map of the Mammalian Cell Cycle Control and DNA Repair Systems. *Molecular Biology of the Cell*, 10:2703–2734, 1999.
5. R. Stevens, C. Goble, I. Horrocks, and S. Bechhofer. Building a Bioinformatics Ontology Using OIL. *Special issue of IEEE Information Technology in Biomedicine on Bioinformatics*, 2001.
6. R. Altman, M. Bada, X. Chai, M. Carillo, R. Chen, and N. Abernethy. RiboWeb: An Ontology-Based System for Collaborative Molecular Biology. *IEEE Intelligent Systems*, 14(5):68–76, 1999.
7. The Gene Ontology Consortium. Gene Ontology: tool for the unification of biology. *Nature Genetics*, 25:25–29, 2000.
8. A. Zabell and C. Post. Docking multiple conformations of a flexible ligand into a protein binding site using NMR restraints. *Proteins: Structure, Function, and Genetics*, 46(3):295–307, 2002.

^fThe extended ontology shows an example at hand: ‘*Calphostin C [inhibit] PKC*’, ‘*Ro 31-8220 [inhibit] PKC*’ and ‘*bis-indolylmaleimide GF 109203X [inhibit] PKC*’.

^gIt is sometimes reported that some proteins reported to behave differently in different locations are, in fact, same one with respect to DNA sequence and behavior in a same condition.