Induced Extension of Gene Ontology from Biomedical Resources with Flexible Identification of Candidate Terms

Jin-Bok Lee [a] **Jung-jae Kim** [a] **Jong C. Park** [a]

[a] Computer Science Division and AITrc, Korea Advanced Institute of Science and Technology (KAIST)

Abstract

We present a novel method to predict more detailed terms than those in the present Gene Ontology (GO). We apply this method to semantic tagging for natural language expressions that denote potential GO terms even when there is no direct mapping of such expressions into GO terms. The terms that are newly identified in this process can be used to extend GO by utilizing semantic relations such as hyponyms or synonyms. Finally, we suggest how to find a suitable direction for the possible extension of an ever-growing ontology such as GO.

Summary

We have identified meta-level rules among the components of Gene Ontology (GO) terms. For example, when we analyze GO terms that belong to *chemokine binding*, we can utilize meta-level rules to the effect that *C-C chemokine*, *C-X-C chemokine* and *C-X3-C chemokine* all belong to the *chemokine* family. To identify such rules automatically, our system utilizes context-sensitive rules that are designed to prevent overgeneration. After predicting new terms, the system validates them with information collected from the biomedical literature, where the system parses the relevant sentences to identify the syntactic dependencies among the components of a given term, and provides validity checking for each newly introduced term.

We have compared the automatically extended GO with a manually extended GO with a year interval. In the automatically extended GO, 112 kinds of concepts are found newly introduced by domain experts in the more recent version of GO; for example, *tricarboxylic acid transporter activity*, *citrate transporter activity*, and *monocarboxylic acid transporter activity*. While this set of overlapping concepts is rather small, we find that it is due to the way in which context-sensitive restrictions are designed to control overgeneration. We expect that such a set will grow much bigger when the system utilizes rules with advanced restrictions such as semantic restrictions, and illustrate the possibility in the extended version of this paper. We also suggest how to find a suitable direction for the possible extension of an ever-growing ontology such as GO.

Finally, we have developed a toolset for manipulating the extended GO, searching newly introduced GO terms through PubMed, and annotating relevant terms in the retrieved articles. These three components of the toolset are closely related to one another, and the toolset can be utilized to construct an annotated corpus as well as to extend a given ontology. It is thus expected that our toolset will reduce a lot of the burden for both developers and curators of GO.

Availability: http://www.biopathway.org

Full article forthcoming in *Bioinformatics*.

Address for Correspondence:
Jong C. Park
Computer Science Division and AITrc, KAIST
373-1 Guseong-dong, Yuseong-gu, Daejeon 305-701
South KOREA