Automatic Construction of Extensional Ontology from Genome Databases

Takuya Yagyu

School of Knowledge Science,
Japan Advanced Institute of Science and Technology
March 2002

Keywords: genome databases, extensional ontology, occurrence-based meaning

Abstract

Biological ontology is one of the most important and interesting subjects in today's bioinformatics. Efforts have already been focused on the need to construct bio-ontology (Gene Ontology[1], EcoCyc[2], etc.) and to develop tools (GKB-Editor[3]). These efforts are bearing fruits, however, the basic philosophy of bio-ontology is oriented toward the construction of a reliable and carefully screened hierarchy of biological concepts, which is mostly done by domain experts, admitting that intelligent softwares assist them. In other word, these efforts can be regarded as the efforts to establish a hierarchy of the "intentional" meaning of technical terms, similar to a comprehensive and peer reviewed biomedical dictionary. For this reason, it is hard to build a large amount of bio-ontology.

On the contrary, one can infer the "extensional" meaning of a technical term from all the occurrence of it. This approach might be supplementary to current majority of bio-ontologies, and might have advantages in automatization and quantity. In this paper, our approach toward automatic construction of "an extensional, occurrence-based bio-ontology" and a result are described.

Before forming a concept hierarchy, it is needed to obtain a long and correct list of technical terms. As an opening gambit, we tried to extract large amount of "probably correct" technical terms from various genome databases integrated in GenomeNet[4]. After rough estimation of

about 370 different fields of 20 databases, 77 fields were selected, which were expected to consist mainly of technical terms (keywords, organisms, reacting chemicals, etc.). Then, all the terms were extracted with the occurrence information (database name, field name, and entry ID). Starting from a list containing 1,911,950 terms, our approach to construct an extensional bio-ontology proceeds as follows:

1. Exclude probably incorrect terms from the list using heuristics and simple pattern matching.
2. Transform the remaining terms into some normalized form, in order to merge terms with essentially the same meaning.
3. Structure terms into lattice(s), using set inclusion. In a lattice, if a normalized form A includes B, then A expresses the more specific concept than B.

As to the transformation into normalized form, the initial list containing 1,911,950 terms shrinked into 1,850,933 terms by case transformation (downcase). Furthermore, the initial list shrinked to 1,770,562 terms by transforming all the special characters (with some exceptions, e.g. a period between two numbers) into white spaces, chunking successive ones, and deleting beginning and ending ones. Finally, by ignoring the order of words in a term, the initial list shrinked to 1,745,877 terms. This result shows that, using the simple transformations mentioned above, nearly 150,000 polymorphic terms could be merged into abstract forms.

Next, lattices of terms were constructed based on the relation called child-parent relationship representing generality of terms. When all the words in a normalized form A are contained in a normalized form B, A is a parent of B. For example, if apple is compared with green apple, the concept of green apple will be more specific. Furthermore, We discriminated two types of relationships called direct parents (or children) and indirect parents (or children).

Based on the idea that terms can be classified from the viewpoint of similarity of their occurrence pattern, the extracted and normalized terms were classified by using LBG algorithm[5]. From this experiment, it was turned out that the terms could be appropriately classified in some extent.

Finally, a system for utilizing the extensional ontology was developed. This system is roughly divided into the following three CGIs, and these are in connected to each other.

(1) CGI for matching technical terms from an input text
(2) CGI for displaying child-parent relationship
(3) CGI for displaying related terms by using the result of clustering

The first CGI compares the extensional ontology and an input text and highlights the matching. Second CGI displays all the direct/indirect parents/children of a term. Third CGI displays other terms classified into the same cluster as an input term, by using the result of clustering. So, a user of this system can navigate the space of this extensional ontology.

As a future work, it is needed to measure terminological importance and the terminological degree of relation from the created hierarchy and the collected occurrence information.

Reference

[1] The Gene Ontology Consortium. "*Gene Ontology: tool for the unification of biology*", nature genetics volume 25 may 2000, pp.25-29,2000

[2] P. D. Krap. et al. "*The Ecocyc Database*", Nucleic Acids Research, 30(1):56 2002

[3] "*Generic Knowledge-Base Editor*", http://www.ai.sri.com/~gkb/

[4] "*GenomeNet WWW server*", http://www.genome.ad.jp/

[5] Y. Linde, A. Buzo, and R. Gray. "*An algorithm for vector quantizer design*", IEEE Trans. Commun. vol. 28, pp. 84--95, Jan. 1980