Exploring GO term definitions to enhance automatic annotation of molecular function

Linair Maria Campos[2], Miguel Gabriel Prazeres de Carvalho[1],
Maria Luiza de Almeida Campos[2], Vanessa Braganholo[1], Maria Luiza Machado Campos[1]
1 - Universidade Federal do Rio de Janeiro –PPGI-IM/NCE
2 - Universidade Federal Fluminense- PPGCI/UFF
marialuizalmeida@gmail.com.br, braganholo@dcc.ufrj.br. mluiza@ufrj.br,
miguelgabriel@ufrj.br, linair@cisi.coppe.ufrj.br

Ontologies have been increasingly used in Bioinformatics, providing a dynamic controlled vocabulary for resources interoperation and annotation. The Open Biomedical Ontologies (OBO) Foundry [1] coordinates the efforts on disseminating a set of shared principles for ontology development, helping to maintain and expand a family of ontologies, expected to be logically well formed and to incorporate accurate representations of the biomedical reality. In this scenario, the Gene Ontology (GO) [2] is clearly the most popular and commonly used OBO ontology. GO has valuable information encapsulated in its terms definitions and hierarchy organization rationale. Unfortunately, not all of its terms are systematically defined and little is known about the strategy adopted for their classification, making their meaning ambiguous to humans and the use of automatic reasoning mechanisms inconclusive to machines [3] [4]. Ambiguity can be minimized if, somehow, the nature of terms could be identified, especially when dealing with different ontologies, when term nomenclature sometimes can be the same, although the intended meaning of terms are different. For example, the term *excretion* can be adopted to refer to the **process** of excretion or to the product of this process. If we can assign its nature (process and object, respectively) to each term, it should be easier to distinguish one from another.

In this context, it is possible to explore GO definitions and documentation in order to make explicit some aspects related to the nature of the represented terms and relations between them (their *ontological commitment*). These aspects can help complement the information already represented on the hierarchy, and therefore help to enhance automatic annotation in general [5]. In the present work we have focused on investigating how molecular function has been represented in different ontologies of the OBO consortium.

One good example is the term *binding*, which, according to GO online documentation has a proposed standard definition for its subclasses: "**x** binding - Interacting selectively and non-covalently with **x**, [brief description of **x**]". Tools can benefit from this standard, since it enables algorithms to be more precise, once one can state beforehand the approach of pattern finding. However, according to the definition of function [6], functions have bearers and are manifested in processes. Additionally, such processes occur in a context, whose elements seem to be reflected in GO subclasses. Considering this, it would be useful if such elements (bearer, process, context) were contained in the definition of functions and existed as individual terms (and not only in the terms textual definition) in ontologies, in order to be used in annotations. For instance, a close examination of *binding* subclasses reveal a multitude of kinds of molecules which interact with bearers, and that could be used in annotations with GO itself and other ontologies if available, such as: chemical elements (*boron binding*), organic compounds (*lipid binding*), inorganic compounds (*nitric oxide binding*). Also, processes occur in places, such as cellular components (*cell surface binding*), among others. This last one, is a clear example of using explicit knowledge about the kind (nature) of **x** to enhance automatic annotation of molecular function: if we state the definition of the branch *binding* as: **x** *binding* [element of context] **x**, [of_kind] **kind of x**, we can easily relate the molecular function *surface binding* (GO:0043498) with its **element of context** of kind **cellular component**: *cell surface* (GO:0009986). Making explicit the notion of **element**

of context and its **kind**, is, as far as we can see, a more useful information than "a brief description of x"; at least from the computational perspective.
Mining tools [7] can be guided to take advantage of such semi-structured information. Besides, if one can make previous associations between kinds and ontologies, it should be possible for tools to suggest candidate terms for automatic annotation.
In order to enhance the exploration of ontologies term definitions, we have proposed a methodology for textual preprocessing combining four text mining approaches: (1) an approach based on machine learning, (2) a dictionary-based approach, (3) a rule-based approach and (4) an approach based on statistics [8]. The challenges of this work lie in dealing with a large amount of information expressed in natural language. Once we are able to state patterns of definitions with correspondent kinds as above, we can obtain more reliable relations which can be used as subsides to ontology enhancement.

References

[1] OBO. **Open Biomedical Ontologies**, 2005. Available at: <http://obo.sourceforge.net.>.

[2] Gene Ontology Consortium.**Creating the gene ontology resource: design and implementation**.Genome Research, vol. 11, n. 8, pp. 1425--1433, (2001).

[3] Köhler, J.; Munn, K.; Rüegg, A.; Skusa, A.; Smith,B. **Quality Control for Terms and Definitions in Ontologies and Taxonomies**, *BMC Bioinformatics,* vol. 7: 212, 2006.

[4] Ogren P, BretonnelCK , Hunter L:**Implications of compositionality in the Gene Ontology for its curation and usage.** Proceedings of the Pacific Symposium on Biocomputing 2005, pp. 174-185, 2005.

[5] OgrenPV, Cohen KB, Acquaah-MensahGK, EberleinJ , Hunter L: **The compositional structure of Gene Ontology terms.** Proceedings of the Pacific Symposium on Biocomputing 2004 , pp. 214-225, 2004.

[6] Arp, R.; Smith, B.; Function, Role, and Disposition in Basic Formal Ontology. In Nature Proceedings, pp 1-4, 2008.

[7] **GATE:General Architecture for Text Engineering**[http://gate.ac.uk]

[8]Ananiadou S,M´Naught J: Text Mining for Biology and BiomedicineNorwood,MA:Artech House ;2006