# Taking GO where we need it to go: focused automated enrichment of the Gene Ontology

Catia Pesquita[1], Francisco Couto[2]
[1]
LaSIGE, Faculdade de Ciências da Universidade de Lisboa, Campo Grande, Lisboa, Portugal

**Background**

Bio-ontologies are manually developed by experts who join their specialized knowledge with extensive literature analysis to reach a consensus on how to model a specific area of biological knowledge [1]. To minimize this burden, we propose the development of automated enrichment methods, based on text mining and ontology alignment techniques, that are able to propose new terms to bio-ontologies, namely the Gene Ontology [2]. There has been some recent work in this area [3], however, it uses a very distinct strategy, generating new terms using the syntactic relations between existing terms.

**Focused automated enrichment of the Gene Ontology**

We report on ongoing work concerning the automated enrichment of the Gene Ontology (GO) . The automated enrichment process will be composed of three tasks: 1) identification of ontology areas that would benefit from enrichment ('hotspots'); 2) enrichment of 'hotspots' through literature analysis and ontology alignment techniques; 3) validation of enrichment by comparing different GO versions.

We are currently in stage 1, where we identify 'hotspots' as GO areas where there is a divergence between the community's needs (represented by the ontology usage for annotation) and the GO developers efforts. In a preliminary study, we identified 17 'hotspots' [4]. The majority of these 'hotspots' correspond to areas of high electronic annotation activity, illustrating that the manual usage of GO is followed by its developers, but not the automated usage.

These 17 'hotspots' correspond to mid-level GO terms, so we are now delving inside these areas to uncover the specific terms in need of extension (see Figure 1). We expect to identify several candidate terms within these 17 hotspots, that will be used as the starting point for the text mining. We will use 3 features to identify them: lower number of children than their siblings, lower number of annotations that their siblings and/or their children. We will also apply this strategy to terms outside 'hotspots' to recover candidate terms that might benefit from enrichment at a deeper level.

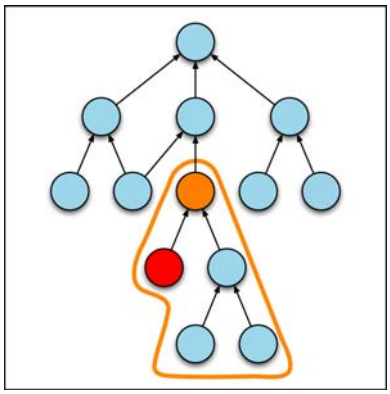

**Figure 1**

Example of a candidate term inside a 'hotspot'. Graphical representation of a portion of the Gene Ontology where circles correspond to GO terms and the arrows correspond to the relationships between them.

The hotspot (area in orange) is characterized by its parent term (in orange). The candidate term (in red) has no children while its sibling has two.

**Conclusion**

The goal of this work is twofold: one one side, to provide GO developers with a starting point for further enrichment of GO; on the other to provide biomedical researchers working in areas currently not under the attention of GO developers with an extended version of GO. By focusing the enrichment on underdeveloped areas we hope to maximize the utility of our work, as well as the automated enrichment process performance.

**Reference**

1.Bodenreider O, Stevens R: **Bio-ontologies: current trends and future directions.** *Brief Bioinform, 7(3),* 2006.
2. GO-Consortium: **The Gene Ontology (GO) database and informatics resource.** *Nucleic Acids Research, 32(Database issue):D258–D261,* 2004.
3. Lee JB, Kim JJ, Park JC: **Automatic extension of gene ontology with flexible identification of candidate terms.** *Bioinformatics, 22(6),* 2006.
4. Pesquita C, Grego T, Couto FM: **Identifying Gene Ontology Areas for Automated Enrichment** *In Proc. 3rd International Workshop on Practical Applications of Computational Biology & Bioinformatics (IWPACBB'09) (in press)*, 2009