Improving bio-ontologies matching using types and adaptive weights

Bastien Rance[1] and Christine Froidevaux[1]

LRI, UMR 8623, Univ. of Paris-Sud, CNRS
F-91405 Orsay CEDEX France
firstname.surname@lri.fr

Functional annotation consists in assigning a biological function to a given protein. It is a crucial task in biology and has various impacts on many fields, including understanding cellular processes and drug designing. In order to be able to share and reuse annotations, biologists and bioinformaticians have developed structured controlled vocabularies that were first simple classifications and then more elaborated ontologies such as the Gene Ontology [1].

In our project, biologists and bioinformaticians collaborators are interested in proteins annotated with two distinct ontologies, such that no protein is annotated with both of them. These ontologies are merely functional hierarchies (Subtilist [2] and FunCat [3]) that share common features: (i) a simple structure with no explicit relationships (subsumption relationships can be deduced from concepts identifiers), (ii) high broadness and small depth, and (iii) variable size.

The system O'Browser [4] we have designed to align functional hierarchies, is based on a weighted combination of matchers as many ontology matching systems [5], with two original characteristics. Indeed, we had to face two issues: (a) a high number of candidates pairs of concepts, and (b) a variable quality of the results of the matchers with respect to the gold standard built by the expert.

As the number of candidates pairs of concepts can be unnecessarily huge, we propose to reduce it by exploiting domain knowledge. For it, we have used **types** (groups of concepts sharing the same semantic context). Concepts that are related to the same field (in our case the same functional genomic field) are assigned to the same type. As an example, the concepts *Utilization of Carbon* and *Synthesis of Glucose* are related to the type *Metabolism*. As in [6], concepts of distinct types will never be mapped (e.g. *Germination* in the context of plants and *Germination* in the context of bacteria). In our approach, an expert manually assigns types to the top concepts of the hierarchies, that represent only a small part of the whole set of concepts of both hierarchies. Types are then spread to all concepts using subsumption relationships. In our experiment, the use of types has allowed to divide the number of candidate pairs by 7. The originality of our contribution is to propose a machine learning strategy to assign types to concepts.

The second issue is about the variable quality of the scores of a given matcher. It has been shown that the good results of a matcher may be spoiled by the scores of other matchers [7, 8]. To address this issue, we would like to give a high weight to a matcher in a combination of matchers only when its results are informative. We claim that the weight of a matcher in a combination should partially depend

on its scores (**adaptive weighting**). As an example, let us consider a string-based matcher that compares concepts from two biological ontologies. If the labels of the concepts are close, the two concepts are likely to be equivalent. On the opposite, distant labels do not indicate necessarily that the concepts are distant. Consequently the weight of the string-based matcher should be high for high scores and weak for low scores.
For each matcher, we define a weighting function which associates a weight to each score of the matcher. Let O_1 (resp. O_2) be the set of concepts of the first (resp. second) ontology and let M_i be a matcher: $O_1 \times O_2 \rightarrow Dom_i$, the weighting function W_i is defined on Dom_i and has $[0, 1]$ as a range. For example, assume that the range of the string-based matcher is $Dom_{String-based} = [0, 1]$. Then a weighting function could be the following simple function: $W_{String-based} : [0, 1] \rightarrow [0, 1]$, where $W_{String-based}(\alpha) = 1$ if $\alpha > 0.5$ and $W_{String-based}(\alpha) = 0.25$ otherwise. Unlike in [9], we allow to associate a strong confidence (and thus a high weight) to low results of a matcher in the case where the score of the matcher is a strong indicator of the absence of equivalence between the considered concepts.

We successfully used types and adaptive weighting to align Subtilist and FunCat and compared the results to the gold standard. O'Browser with adaptive weighting found 80 % of the actual correspondences, while O'Browser with the best classical matcher combination found only 70 % of them.

References

1. The Gene Ontology Consortium: Creating the gene ontology resource: design and implementation. Genome Res. **11** (2001) 1425–1433 http://www.geneontology.org.
2. Moszer, I., Jones, L., Moreira, S., Fabry, C., Danchin, A.: Subtilist: the reference database for the Bacillus subtilis genome. Nucleic Acids Res **30** (2002) 62–5
3. Ruepp, A., Zollner, A., Maier, D., Albermann, K., Hani, J., Mokrejs, M., Tetko, I., Gldener, U., Mannhaupt, G., Mnsterktter, M., Mewes, H.: The FunCat, a functional annotation scheme for systematic classification of proteins from whole genomes. Nucleic Acids Res. **14**((32)18) (2004) 5539–5545
4. Rance, B., Gibrat, J.F., Froidevaux, C.: An adaptive combination of matchers: application to the mapping of biological ontologies for genome annotation. In: Proc. of the 5th Data Integration in the Life Sciences workshop DILS'09. LNBI 5647 (2009) 113–126
5. Euzenat, J., Shvaiko, P.: Ontology matching. Springer-Verlag, Heidelberg (DE) (2007)
6. Zhang, S., Mork, P., Bodenreider, O., Bernstein, P.A.: Comparing two approaches for aligning representations of anatomy. Artificial Intelligence in Medicine **39**(3) (2007) 227–236
7. Ghazvinian, A., Noy, N.F., Musen, M.A.: Creating mappings for ontologies in biomedicine: Simple methods work. Technical report, Stanford Center for Biomedical Informatics Research (2009)
8. Ontology Alignment Evaluation Initiative: http://www.oaei.ontologymatching.org
9. Mork, P., Seligman, L., Rosenthal, A., Korb, J., Wolf, C.: The harmony integration workbench. J. Data Semantics **11** (2008) 65–93