FLORAL: Semantic and Lexical Matching of Biomedical Ontologies

Catia Pesquita, João D. Ferreira, Francisco M. Couto
Faculdade de Ciências, University of Lisbon

The investigation of complex systems, such as the human body, benefits largely from the integration and interoperability of independent resources within the scope of interest. In this context, ontologies are valuable resources, since they model a portion of biological knowledge, a domain, by providing a formal representation of the concepts in that domain and the relations between those concepts. The common agreement over this representation provides a shared ground for communication both between researchers and computers. However, large and diverse domains such as human physiology, anatomy or molecular biology are usually covered by many different ontologies, with different degrees of overlap. Ensuring the interoperability of these resources is a crucial step towards the understanding of the human body as single complex system, since it provides a basis for integrating data annotated with these ontologies, and allows reasoning across them.

Several existing ontologies are relevant for describing VPH models or data, such as Gene Ontology (GO), Foundational Model of Anatomy (FMA), Chemical entities of biological interest (ChEBI), Phenotypic quality (PATO), Human Phenotype Ontology (HPO), etc. Although there are efforts to guarantee biomedical ontology interoperability, such as the OBO initiative or UMLS, these are ad hoc in nature, so ontologies need to be developed according to them. However, many ontologies have been, and still are, developed outside of these efforts, creating the need for post hoc solutions that are able to find correspondences between ontology concepts and thus enable their interoperability.

Ontology matching aims at identifying these correspondences through the application of automated techniques. These can be classified according to the type of data they use (e.g. labels, internal structure, properties, instances, etc) or the strategy they employ (e.g. linguistics, machine learning, statistics, etc) [1].

So far, the alignment of biomedical ontologies has focused mainly on the alignment of anatomy ontologies, as these have been adopted as a standard for the biomedical track of OAEI [2], an international evaluation of ontology matching systems. The first edition of anatomy alignment in OAEI 2006 aligned the Foundational Model of Anatomy (FMA) and GALEN, a clinical terminology, whereas following editions aligned the Mouse anatomy (MA) and the Human anatomy of NCI thesaurus (HA). The existence of a gold standard reference mapping for the mouse-human alignment supported a more thorough evaluation of alignment systems.

An important finding of this initiative is that many alignments are rather trivial and can be found by simple string comparison techniques. Based on this notion, the work in [3] has applied a simple string matching algorithm, LOOM, to several ontologies available in the NCBO BioPortal, and reported high levels of precision in most cases. The authors mention several possible explanations for this, including the simple structure of most biomedical ontologies, the high number of synonyms they contain and the low language variability. When compared to the top performers in OAEI 2008 [4], SAMBO, SAMBOdtf, and RiMOM, LOOM has a higher precision but a lower recall, probably due to the other algorithm's exploitation of other more complex methods than simple lexical matching. SAMBO and SAMBOdtf employ the UMLS (Unified Medical Language System) as the domain knowledge source to support lexical matching of concepts. RiMOM on the other hand, does not use external knowledge, relying on label and structural similarities. In OAEI 2009 [2], the best systems, SOBOM and AgreementMaker also did not use external knowledge, but both relied on global similarity computation techniques. These techniques represent ontologies as graphs, where concepts are nodes, and the relations between them, edges, and propagate lexical similarities between ontology concepts throughout the ontology graphs. This is based on the assumption that a match between two concepts should contribute to the match of their adjacent concepts, according to a propagation factor.

In here we present a system for ontology alignment designed to leverage on the success of simple lexical matching methods, while still finding alignments where lexical similarity is low, by using global computation techniques. Our system, FLORAL, couples FLOR [5], an algorithm for finding relations between ontology concepts based on the textual components of ontologies, with a novel global similarity computation approach that takes into account the semantics of the edges in ontology graphs.

FLOR measures the similarity between the textual descriptions of ontology concepts. It relies on two information theory notions: the evidence content of a word, and the information content of an ontology concept.

The evidence content (EC) of a word [6] measures the relevance of a word within the vocabulary of an ontology based on its frequency. FLOR performs a syntactic processing step before the computation of word frequencies, that eliminates common English words and uses the TreeTagger package [7] to tokenise, lemmatise and perform part-of-speech tagging, and the Porter Stemmer algorithm [8] to reduce non-nouns to their root word. The final frequency of a word corresponds to the number of terms that contain it in their descriptors. A word that appears multiple times in the name, definition or synonyms of a term is only counted once, preventing bias towards terms that have many synonyms containing very similar word sets.

The information content (IC) of a concept [9] is a measure of how likely the concept is to occur in a given corpus, which can be quantified as the negative log likelihood of occurrence of a term in a specific corpus. For ontologies with an instance corpus, such as GO, we calculate IC using instance data. For ontologies without instance data, the IC of a term is based on the number of children a term has in the ontology. The IC is then transformed into a

relative measure by taking into consideration the size of the corpus [10].

Given that ontology concepts usually have several textual descriptors (e.g. name, synonyms, definitions), the relatedness between two ontology concepts is calculated as the maximum similarity between all possible combinations of descriptors. The similarity between two descriptors of ontology concepts depends on whether they can be considered an exact or a partial match. An exact match is found when one descriptor is contained as a substring of the other. In this case the final score is influenced by the information content of the matched concept, which means that matches to more specific terms have a higher score. For partial matches, the final score is based on the EC of the words shared by the descriptors: the more specific the shared words are, the higher is the score. When using FLOR for ontology matching, we have two vocabularies, one for each ontology, so shared words have two EC values. The evidence content of a shared word is given by a weighted sum of each vocabulary EC.

Regarding the global similarity computation, FLORAL uses a novel approach that takes into account the semantics of edges in the ontology graph. The similarity propagated along an ontology edge corresponds to the information content shared by the concepts connected through the edge. Several semantic similarity measures (measures designed to measure the similarity between ontology concepts) explore this notion of shared information content [11], so FLORAL incorporates several semantic similarity measures to compute propagation factors, which are applied to three existing global similarity techniques: similarity flooding [12], DSI and SSC [13]. The main distinction between these three methods is the path of propagation: in similarity flooding, parents and children of ontology concepts can contribute to the propagation of similarity, while in DSI, only parents contribute, and in SSC only siblings contribute.

While the full fledged system is still under development, we present an illustration of the application of FLORAL to the alignment of the mouse (MA) and human anatomy (HA) ontologies, using similarity flooding as the global computation technique. In this example we focus on the mapping of the concept 'Digestive system fluid/secretion' of MA to the concept 'Gastrointestinal fluid or secretion' of HA.

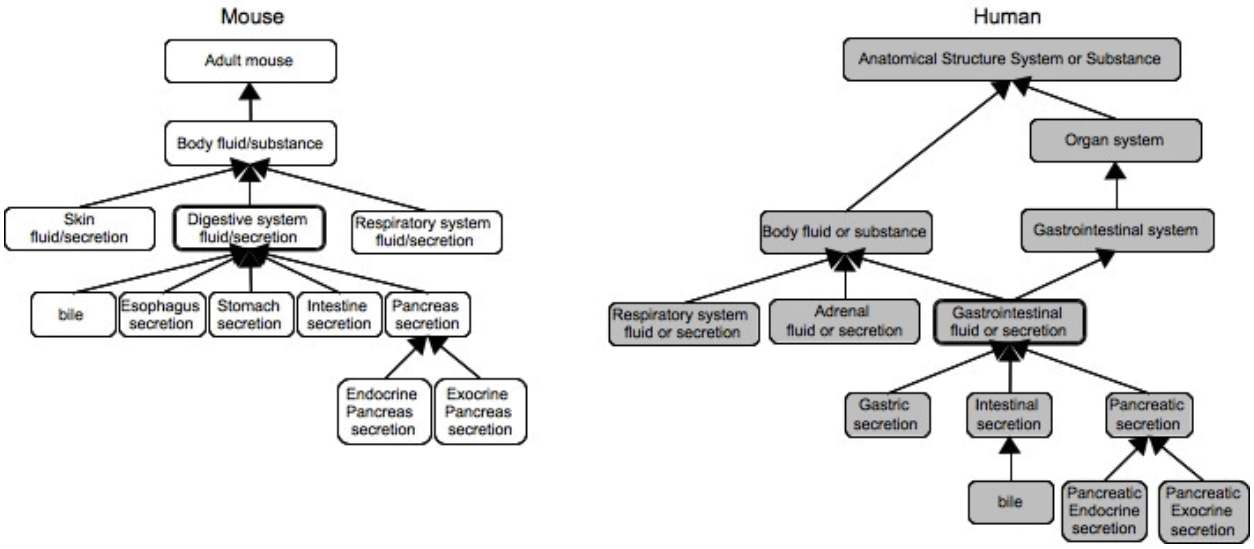

Fig.1 Ontology subgraphs for the mouse and human anatomies, pertaining to the terms 'Digestive system fluid/secretion' and 'Gastrointestinal fluid or secretion'.

Figure 1 presents the ontology subgraphs that contain these terms. A simple word matching algorithm, such as LOOM would be unable to match these concepts, and even a more complex algorithm able to match the words 'fluid' and 'secretion', would result in the mapping of these concepts to many others, such as 'Adrenal fluid or secretion' or 'Respiratory system fluid or secretion'. The application of global similarity techniques directly over these simple matches would not significantly improve the matching score, since the matching of parents (e.g. 'Body fluid/substance' to 'Body fluid or substance') affects equally all of their children. The only matching of descendants found through simple matching corresponds to 'Bile'. However, since 'Bile' is not a direct child of 'Gastrointestinal fluid or secretion', the contribution of similarity is to the match between 'Digestive system fluid/secretion' and 'Intestinal secretion'.

By using FLORAL, many more lexical matchings are found (see Table 1) due to the higher degree of syntactical analysis performed by FLOR. The quality of these matchings is ensured by the application of IC and EC notions. This results in a more successful application of similarity flooding that benefits from the higher quality of the initial mappings.

Mouse anatomy	Human anatomy	FLOR	FLOR+similarity

			flooding
Digestive system fluid/secretion	Gastrointestinal fluid or secretion	0.39	0.92
Digestive system fluid/secretion	Respiratory system fluid or secretion	0.79	0.82
Digestive system fluid/secretion	Adrenal fluid or secretion	0.22	0.25
Body fluid/substance	Body fluid or substance	0.90	1.00
Respiratory system fluid/secretion	Respiratory system fluid or secretion	0.97	1.00
Pancreas secretion	Pancreatic secretion	0.98	1.00
Intestine secretion	Intestinal secretion	1.00	1.00
Endocrine pancreas secretion	Pancreatic endocrine secretion	1.00	1.00
Exocrine pancreas secretion	Pancreatic exocrine secretion	1.00	1.00
Endocrine pancreas secretion	Pancreatic exocrine secretion	0.47	0.73
Exocrine pancreas secretion	Pancreatic endocrine secretion	0.47	0.73
Bile	Bile	1.00	1.00
Esophagus secretion	Gastric secretion	0.34	0.46
Esophagus secretion	Intestinal secretion	0.34	0.46
Stomach secretion	Pancreatic secretion	0.33	0.45

Table 1. Examples of matches for the ontology concepts related to the concept 'Digestive system fluid/secretion' of the mouse anatomy, and to the concept 'Gastrointestinal fluid or secretion' of the human anatomy using FLORAL.

The final version of FLORAL will take advantage of the full integration between global similarity computation techniques and semantic similarity measures, while being able to remain independent of external sources of information.

We believe that ontology matching systems such as FLORAL, are crucial to improve the interoperability of ontological resources, and these in turn are cornerstone to the understanding of human physiology and pathology.

References

[1] J Euzenat and P Shvaiko. Ontology Matching. *Springer*. 2007

[2] J Euzenat, A Ferarara, L Hollnick, A Isaac, C Joslyn, V Malaise. Results of the Ontology Alignment Evaluation Initiative 2009. *ISWC workshop on Ontology Matching (OM-2009)*. 2009

[3] A Ghazvinian, NF Noy, MA Musen. Creating mappings for ontologies in biomedicine: simple methods work. *AMIA Annu Symp Proc. 2009*:198-202. 2009

[4] C Caracciolo, J Euzenat, L Hollnick, R Ichise, A Isaac, V Malaise. Results of the Ontology Alignment Evaluation Initiative 2008. *ISWC workshop on Ontology Matching (OM-2008)*. 2008

[5] C Pesquita, FM Couto. Retrival of related Bio-ontology Concepts. (to appear) 2010

[6] FM Couto, MJ Silva, and PM Coutinho. Finding genomic ontology terms in text using evidence content. *BMC Bioinformatics*, 6 Suppl 1:S21, Jan 2005.

[7] H Schmid. Improvements in part-of-speech tagging with an application to German. In Natural Language Processing of very large corpora. *Springer*, 1999.

[8] MF Porter. An algorithm for suffix stripping. *Program* 14(3) 130-137. 1980

[9] P Resnik. Semantic Similarity in a taxonomy: an information-based measure and its application to problems of ambiguity in natural language. *Journal of Artificial Intelligence Research* 11, 95-130. 1999

[10] C Pesquita, D Faria, H Bastos, A Ferreira, AO Falcao, FM Couto. Metrics for GO based protein semantic similarity: a systematic evaluation. *BMC Bioinformatics* 9 Suppl 5:S4 2008

[11] C Pesquita, D Faria, O Falcao, P Lord, FM Couto. Semantic similarity in biomedical ontologies. *PLoS Comput Biol* 5(7), 2009

[12] S Melnik, H Garcia-Molina, E Rahm. Similarity Flooding: A versatile graph matching algorithm and its application to schema matching. *Proc. 18th ICDE*, 2002

[13] IF Cruz, W Sunna. Structural alignment methods with applications to geospatial ontologies. *Transactions in GIS*. 2008