An Approach for the Alignment of Biomedical Ontologies based on Foundational Ontologies

Vivian S. Silva[1] , Maria Luiza M. Campos[2] , João Carlos P. Silva[2] and Maria Cláudia Cavalcanti[3]

[1] Isban - Grupo Santander, Brazil
{vivian.ss}@gmail.com
[2] Universidade Federal do Rio de Janeiro, Brazil
{mluiza, jcps}@ufrj.br
[3] Instituto Militar de Engenharia, Brazil
{yoko}@ime.eb.br

Abstract. Genome annotation, the task of assigning a description to each discovered genome sequence, is an important activity within the process of sequencing. It relies on the use of ontologies to maintain a uniform vocabulary and to support interoperability of different information resources. Often, only the Gene Ontology (GO) is used in the annotation process, but the exploration of other ontologies, along with GO, could enrich the vocabulary used in the annotation, complementing it with more details. To facilitate this task, it is necessary the identification of equivalences between terms from GO and terms from other ontologies. This work presents an approach for aligning biomedical ontologies within the genome annotation process that helps to identify equivalent terms between GO and other biomedical ontologies, enabling the annotator to choose which among them is more suitable as a descriptor. Two main points have guided its development: (i) the choice of a subset of similarity measures suited to the characteristics of biomedical ontologies; and (ii) the use of foundational ontologies, allowing the analysis of the conceptual nature of each term, which serves as a fundamental parameter for similarity calculation, reducing the possibility of associations between terms derived from different categories. Initial experiments showed an improvement on the alignment quality, represented by a 14% increase in the number of correct alignments and a 5% decrease in the number of incorrect associations, reinforcing the usefulness of the proposed approach in supporting the annotator's work.

Categories and Subject Descriptors: H.1.m [**Information Systems**]: Models and Principles; I.2.4 [**Artificial Inteligence**]: Knowledge Representation Formalisms and Methods

Keywords: Foundational Ontologies, Genome annotation, Ontology Alignment

1. INTRODUCTION

With the evolution of genome research and frequent studies associated to different new discovered organisms, the importance of bioinformatics has grown steeply. Researchers are exploring technologies that, among other things, support the manipulation of large genomic databases and facilitate data interoperability.

The genome annotation process involves activities to register structural and functional information associated to sequences, including coding regions, biological function, gene regulation, interactions and expression [Belloze 2007]. In order to maintain a uniform vocabulary throughout this process, the use of ontologies is crucial. The Gene Ontology is the most widely used ontology for genome annotation, providing a systematic language that enables consistent descriptions in three key biological domains: cellular component, biological process and molecular function [The Gene Ontology Consortium 2008].

Despite the great popularity of the Gene Ontology, the rapid evolution of biological research requires

---

This work was partially funded by CNPq grants number 309307/2009-0 and 309454/2009-3.

the use of other ontologies. The Open Biological and Biomedical Ontologies (OBO) Foundry, which supports GO, is a collaborative effort to create and maintain a suite of orthogonal ontologies to serve as a reference for the biomedical domain [OBO Foundry 2009]. The use of the Gene Ontology in conjunction with other OBO ontologies can enrich genome annotation, generating more complete and consistent descriptions. Nevertheless, there is still great difficulty in using other ontologies during the annotation process. The preference for GO is mainly due to the availability of mechanisms to support annotation based on similarity. Because existing databases have been previously annotated based on GO, the terms used to describe a sequence are automatically transferred to the annotation of the new similar sequence. Although GO covers many aspects needed by biologists when describing sequences [Bodenreider and Stevens 2006], it is possible that one or a diversity of the other OBO ontologies offer specific concepts more suitable to the particular organism being sequenced, which can directly influence the quality of the annotation.

Using more than one ontology during the annotation process requires the identification of equivalences between the ontologies terms. If the scientist initially uses the GO term x, it is possible to look for terms in other ontologies that have the same meaning of x, and, by comparing their definitions, choose the one that is more suitable for the annotation.

The identification of equivalences between terms of two or more ontologies is possible through the alignment of these ontologies. [Ehrig 2007] defines the task of aligning two ontologies as: "for each entity (concept, relation or instance) in the first ontology, we try to find a corresponding entity, which has the same intended meaning, in the second ontology." As the annotator has initially available the associated GO term, automatically identified in the first phase of the process, the alignment would retrieve similar terms from other ontologies, leaving to the annotator the task of validating the suggested results. Thus, new terms are made available without the need to search the ontologies manually: the concepts equivalent to those expressed in GO can be found even if they have received different names.

The objective of this work is to present an approach for ontology alignment in the Bioinformatics context, performing an analysis of similarity measures best suited to the characteristics of the OBO ontologies, and also using Foundational Ontologies for concepts differentiation based on their nature at the ontological level [Guarino 1994]. It is not intended to be another alignment tool, but a procedure that takes advantage from existing alignment approaches, adding an important semantic feature to them. More reliable alignments can then be obtained, allowing the annotator to quickly identify the terms of interest, enriching the vocabulary to be used, and, therefore, the generated annotation.

The article is structured as follows. Section 2 briefly introduces the ontology-based genome annotation process. Section 3 describes the principles of ontology alignment and its techniques, focusing on the top-level ontology based technique. Section 4 presents an analysis of alignment tools and Section 5 discusses the proposed approach in detail. Section 6 presents the results obtained by an initial experiment. Section 7 reviews related works, followed by the conclusions in the Section 8.

2. GENOME ANNOTATION BASED ON BIOMEDICAL ONTOLOGIES

The interpretation of data generated in the genome sequencing process is one of the most important phases of genome research. The genome annotation corresponds to the description of identified sequences by assigning them biological characteristics. Thus, an annotation is the recording of the biological meaning of each identified gene. Associating annotations to gene fragments provides the context for interpreting genomic data [Frishman and Valencia 2008]. The annotation process involves filtering, transforming and computationally manipulating data, but also often requires some human effort on repairing and curing information. We can identify two types of annotations: automatic, which are generated by analysis programs or imported from public databases, and manual, which are created directly by the researcher. As the annotations will be accessed by many researchers from anywhere in the world, it is important that the language used is of common understanding, based on a standard vocabulary that is shared by everyone in this area. To achieve that, the ontology based annotation is an important mechanism, as ontologies provide a uniform vocabulary, allowing genes to be identified by using similar terms to describe them, regardless of the research group responsible for the annotation.

The Open Biological and Biomedical Ontologies (OBO) are recognized as a fundamental effort to support the standardization of the annotations resulting from genome sequencing research [Smith et al. 2007]. The GO is certainly the most popular ontology among them, as the Gene Ontology Annotation (GOA) database offers high quality GO-based annotations, generated both manually and automatically [Barrell et al. 2009]. With over 32 million annotations, this database is one of the most used for automatic and semi-automatic annotation of newly investigated organisms, where sequences that are similar to those searched are retrieved and used as a reference for annotation, with the associated GO term being copied as a description for the discovered gene. The wide use of this feature as a starting point for further annotations, consolidates the GO as the primary (and often only) ontology used in the genome annotation process.

3. ONTOLOGY ALIGNMENT

The growing number of publicly available ontologies, as well as the increasing amount of applications that use them, reflect some of the existing problems in semantic interoperability. Despite being developed to provide a common vocabulary within a given domain, there are actually several ontologies in use by different groups and applications. These groups may use different names for the same concept, and the same name to represent different concepts. The way concepts are related and classified vary according to the group's view on the domain, and also according to the purpose of the ontology creation.

Aligning ontologies is a necessary condition to support semantic interoperability between systems, identifying relationships between individual elements of multiple ontologies [Ehrig 2007]. Besides allowing communication and reuse, the alignment process enables the combination of information and knowledge contained in different ontologies, enriching the descriptions of data annotated with these resources.

The term alignment has many interpretations and there is still no consensus on its meaning. [Ehrig 2007] defines the alignment of two ontologies as "[...] for each entity (concept, relation or instance) in the first ontology, trying to find a corresponding entity with the same intended meaning, in the second ontology." [Euzenat and Shvaiko 2007] consider alignment as the result of ontologies matching, defined as the task of finding correspondences between semantically related entities of different ontologies, where the relationships can be of equivalence, disjunction, among others. According to this definition, the alignment is seen as a product, not as a process, as emphasized by Ehrig. [De Bruijn et al. 2006] also reinforce the perspective of alignment as a process, defining alignment as the discovery of correspondences between ontologies, which are represented by their mapping.

For the purposes of this article, we have adopted the definition of [Ehrig 2007], where the alignment is considered the process of establishing one-to-one equality relations between the terms of two ontologies from the same domain, which have some intersection.

3.1 Ontology Alignment Techniques

Several alignment tools have been developed to support the identification of equivalent entities in different ontologies. Behind these tools, there is also a variety of techniques, which are usually combined to calculate the degree of similarity between concepts, relationships and instances. [Euzenat and Shvaiko 2007] present a detailed classification of these approaches, represented in Figure 1. The

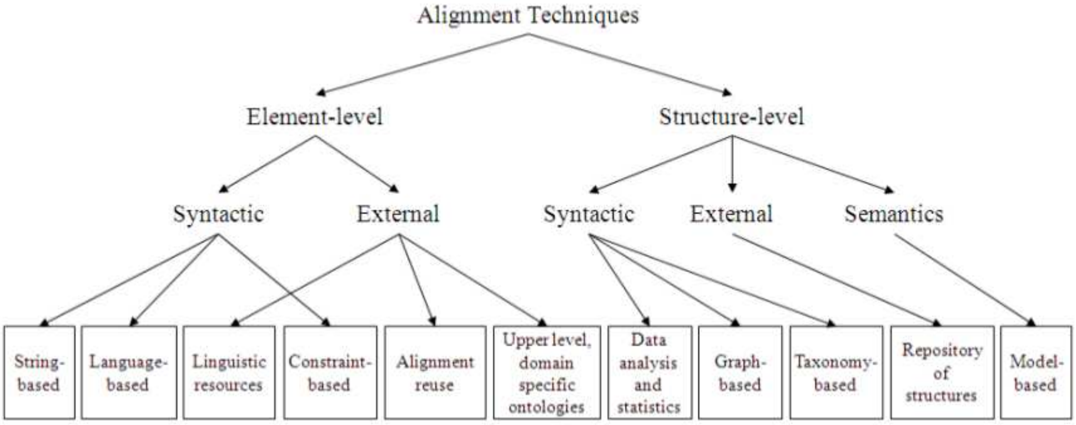

Fig. 1. Classification of elementary alignment approaches. Adapted from EuzenatandShvaiko2007

basic techniques are divided into two categories: element-level and structure-level techniques. The techniques at the element level identify correspondences analyzing entities in isolation, ignoring their relations with other entities. The techniques at the structure level consider how the entities are presented together within the ontology structure to indicate potential matchings.

At the second level of the alignment techniques classification, the syntactic techniques include those where the input is interpreted according to its structure, following a clearly defined algorithm. The external techniques are those that exploit auxiliary external resources of a domain to interpret the input, such as thesauri or additional input from users. Semantic techniques use some kind of formal semantics, as for example, model theory, to interpret the input and justify their results. The last level of the classification tree corresponds to the basic alignment techniques.

3.2 Top-level Ontology based Technique

Top-level, or foundational, ontologies correspond to a set of high-level, domain independent categories, which include notions such as objects, events, attributes, spatial-temporal connections, dependencies and other concepts common to all areas of knowledge. These ontologies provide rigorous formal semantics for these high-level categories, and serve as a conceptual basis for domain ontologies, which will, in turn, model a particular part of the world [Probst 2006]. [Guizzardi 2009] reinforces this definition by stating that a foundational ontology "is a formal framework of generic concepts (i.e. domain independent) of the real world that can be used to talk about material domains”. In general, this kind of ontology is used as a reference model to define allowed concepts in a well- founded conceptual modeling language, enabling it to capture the semantics of the real world.

Besides the benefits for building conceptual models of a domain, top-level ontologies may also be useful during the ontology alignment process. By identifying the meta-categories from which the concepts are derived, it is possible to establish their nature, making it explicit the differences between an object and a process, types of things from their roles, among others. This distinction may help to prevent incorrect associations in the alignment process, restricting the indication of equivalent terms to those derived from the same meta-category, i.e. those having the same conceptual nature. Figure 2 illustrates the use of top-level ontologies in ontology alignment. The presented fragment of a foundational ontology was introduced by [Guizzardi 2007], where only objects are modeled (endurants). In this ontology, the class "Kind”, which is characterized by a rigid sortal (which provides a principle of identity), is disjoint from the class "Role", which is not a rigid sortal. Simply put, a rigid sortal is one that does not vary over time (e.g., a "Person" is always a "Person"), unlike a non-rigid sortal, which establishes a condition that is only valid during a given moment in

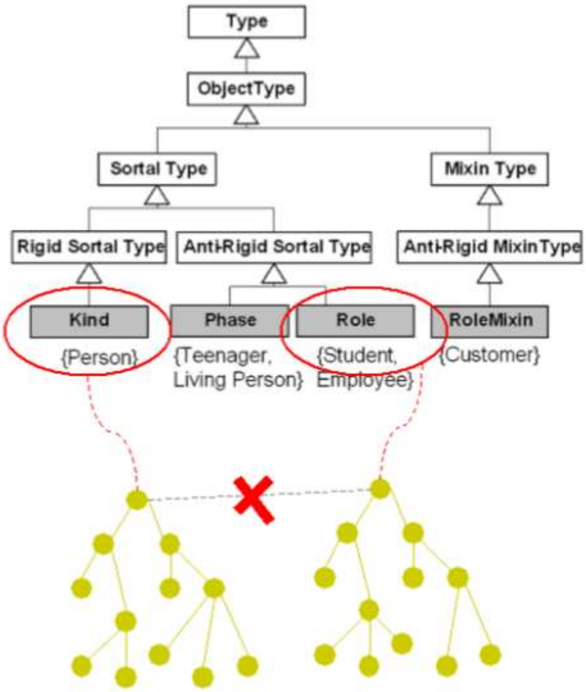

Fig. 2. Use of a foundational ontology in ontology alignment

time (for example, a "Student" at one time may no longer be a "Student" at another).

Figure 2 shows the alignment of two ontologies, and the matching of classes where one of them describes a kind and the other one, a role. By identifying the origin of these classes with the help of a top-level ontology, this type of association can be avoided, even if some similarity measure returns a high value, because concepts with a different nature can never be regarded as equivalent.

3.3 Similarity between Ontologies

The identification of similarities between two ontologies is a fundamental step in the alignment process. The similarity, which corresponds to a numeric value that indicates how two elements are similar or different, may be obtained from the comparison of whole ontologies or only of their sub-elements [Ehrig 2007] The similarity relation can be established between concepts, relationships or instances. By considering both representation and meaning of an entity in the computation of similarity, many measures can be used, each contemplating a distinct characteristic of the compared concepts. A list of these measures, including those used in this work, can be found in Ehrig's work.

4. ALIGNMENT TOOLS

In the last years, motivated by an increase on ontology use by several applications, the research on ontology alignment has flourished. The development of a variety of ontologies on the same domain, or with a significant intersection among them, and the need for communication and interoperability between the systems that use them, have led to the creation of dozens of tools for ontology alignment, merge and integration. Differently from the first developed tools, like Chimaera [McGuinness et al. 2000] and PROMPT [Noy and Musen 2000], which involved just simple techniques, such as string or structure comparison, more recent developed tools also use sophisticated techniques (Table I), such as linguistic external resources and taxonomy based techniques.

Table I. Ontology alignment tools and their techniques

Tools	Techniques										
	Element-based Techniques						Structure-based Techniques				
	String Based	Language Based	Linguistic resources	Restriction based	Alignment Based	Top-level Ontologies	Data analysis	Graph based	Taxonomy based	Structure repository	Model based
ASMOV [Jean-Mary et al. 2009]	x		x	x				x	x		x
DSSim [Nagy et al. 2006]	x		x					x			
OntoDNA [Kiu and Lee 2006]	x						x	x	x		
Falcon [Jian et al. 2005]	x	x						x			
FOAM [Ehrig and Sure 2005]	x		x	x				x	x		
SAMBO [Lambrix and Tan 2006]	x	x	x					x	x		
RIMOM [Li et al. 2007]	x	x						x			
Lily [Wang and Xu 2008]	x		x	x				x			
CIDER [Gracia and Mena 2008]	x		x					x	x		
Aroma [David et al. 2007]	x	x		x				x	x		x

In order to identify tools and their respective approaches, to elaborate the proposal of this work, a selected set of tools were analyzed. This selection was based on the best performance results reported at OAEI 2007/2008 and the availability of a well-documented API. Table I summarizes the selected set of tools characterization with respect to the alignment techniques discussed in the last section.

Among all these tools, FOAM was chosen to validate the present proposal due to its technical superiority. It is clearly one of the most complete approaches in terms of supported techniques. At the time of this work development, it was available as an open source code in JAVA, together with a detailed description of classes and methods. In addition, FOAM code presents a good level of modularity and flexibility, which facilitates possible extensions, such as the inclusion of new similarity calculation techniques. Therefore, FOAM was adapted to perform OBO ontologies alignment, and consequently, support the genomic annotation process.

5. OBOAEA ALIGNMENT APPROACH

Most of the alignment tools available nowadays are generic, i.e., they do not address specific domain ontologies. However, within some domains, these ontologies may present some particular characteristics that require special treatment, such as to carry deep or shallow hierarchies, ontologies without instances, focus on constraint representation, among others. This is the case of the ontologies under the supervision of the OBO consortium, which are focused on providing a standard vocabulary and hierarchy, with no instances and just a few object properties.

In this context, the main goal of this work is to elaborate an alignment approach, named OBOAEA (from OBO Alignment for Enriched Annotation), which takes into account the specific characteristics of the OBO ontologies. The idea is to select the most appropriate similarity measures, discarding those that focus on resources that are not present in such ontologies. Therefore, it is possible to take advantage of the OBO ontologies resources, without the impact of unnecessary calculations on the whole alignment process. Besides the usual similarity measures, the identification of similar classes also counts on the support of top-level ontologies. These ontologies enable to identify the nature of the concepts that are under comparison, and help to improve the quality of the alignments by adding more information to the similarity calculation. Therefore, this work develops a mechanism that can be incorporated to a genome annotation environment, enabling the annotator to use multiple ontologies simultaneously. The annotator can easily identify equivalent concepts on different ontologies, and select the one that presents the most detailed description, thus producing an enriched final annotation.

5.1 OBOAEA approach steps

The main goal of the OBOAEA approach is to identify GO terms, which were already used in the genomic annotation process, in other OBO ontologies, and provide the annotator a (possibly) more detailed term, which comes from one of the other ontologies, to enrich his/her annotation. Before the alignment itself, a set of preparation procedures are necessary, not only to select and prepare the resources to be used, but also to facilitate the manipulation of the ontologies and reduce the execution time of the system. These steps are described as follows.

Preparation: In this step, all the ontologies involved in the alignment process are identified: source, target, and top-level ontologies. The source ontology is the one that is in use for semi-automatic annotation, which in our example is the GO. A target ontology is an ontology in which there may be equivalent terms to the annotated source ontology terms, and therefore, may be aligned to the source ontology. Finally, the top-level ontology is the one that will support the alignment between the first two ontologies. After this identification, the association between the top-level and the domain (source and target) ontologies are previously and manually done, as a previous step for the source-target alignment. A detailed description of this top-level association is provided in Section 5.2. It is worth to mention that more than one target ontology can be used and previously prepared for the alignment. Later on, the target ontology is a user's choice for the source-target alignment. It is also important to emphasize that the top-level ontology is a user's choice as well: any foundational ontology can be used within the approach.

Semi-automatic annotation: In this step, a target (or set of target) genome sequence(s) is(are) previously and automatically annotated, through the identification of similar sequences, which are retrieved from sequence databases such as the GOA (Gene Ontology Annotation) Database [Barrell et al. 2009]. For each similar sequence found, the corresponding genome annotations are copied to the local database, and later on complemented or curated manually by the annotator. This step corresponds to the way genome annotation is performed nowadays. It provides the main input to the proposed alignment approach, which is a selected set of source ontology terms.

The OBOAEA alignment approach is organized in the following steps:

(1) **Identification of source ontology terms:** Once the annotations recorded in the local annotation database are available, a direct access to this database is sufficient to retrieve the set of terms of the source ontology which are of interest for the annotator. These terms (with no duplications) are the main input to the next step.

(2) **Source ontology fragment extraction:** Aiming at alignment processing time reduction, a set of source ontology fragments is extracted. OBO ontologies are usually large ontologies. The GO, for instance, has more than 29.000 terms. An alignment process involving it implies in high costs, if it is computable. Therefore, in our approach, we recommend the source ontology fragmentation as a way of reducing the cost and enabling the alignment. For each source ontology term identified in the previous step, a source ontology fragment is generated, which includes the term itself and the other terms and properties structurally close to it.

(3) **Ontology cleaning and name treatment:** The set of ontology fragments is then submitted to some cleaning and transformation procedures. Alternatively, this step could be performed with the original complete source ontology, during the preparation step. However, it is recommended as part of the alignment proposal, as it is much faster when applied to the ontology fragments. OBO ontologies usually come with extra information, such as metadata (e.g. definitions, source datasets, external resources, etc.), which are not used by the traditional alignment techniques.

The metadata removal may reduce the size of ontologies in 65%. Another necessary procedure is the name treatment. OBO ontologies use numeric identifiers as the name of the ontology classes (terms). For instance, in the GO, the names of the classes are similar to GO_XXXXXXX, where each X corresponds to a 0-9 digit. Usually, in OBO ontologies, alphanumeric terms are registered as labels associated to their corresponding class. Therefore, in order to facilitate the syntactic comparison, ontology class names need to be converted to their corresponding labels.

(4) **Ontology alignment:** Once source and target ontologies are prepared, the alignment is then applied. In this step, the OBOAEA adopts NOM (Naïve Ontology Mapping), which is used by the FOAM tool. Originally, FOAM uses many similarity measures, as described in [Ehrig 2007]. However, as OBO ontologies present specific characteristics, such as a reduced number of properties and no instances (individuals), some of these measures were excluded in OBOAEA approach: extensional similarity, domain and scope similarity, and concept similarity. This way, the following measures are used: equality, syntactic similarity, object equality, multi similarity, label similarity and taxonomic similarity. In addition, this step also takes into account previous alignments, which serve as reference to validate correct alignments, and also to discard incorrect alignments, avoiding that these ones are repeatedly presented to user validation in the next step. It is worth to mention that in this step, the alignment counts on the top-level ontologies associated to the source fragment ontologies and to the target ontologies.

(5) **Equivalent terms presentation:** Alignment results are then presented for the user examination. Only a selected set of pairs (source ontology term, target ontology term) are presented to the user, discarding those pairs in which the source ontology term is not among the annotated terms (selected in the terms identification step).

The following step is the final step, which completes the OBOAEA approach. It provides suggestions of new annotations to the annotation environment. Then, the annotator may proceed with the annotation enrichment based on the results of the alignment process.

(6) **Validation:** After examining carefully each pair of terms, the annotator, who is a domain expert, validates each one by tagging it as correct or incorrect. At this point, the annotator may then choose which terms on the target ontology may enrich the original annotation, providing a more detailed description to the final annotation. Thus, the enriched annotation (final annotation) is composed of three items: the semi-automatic annotation description, the source ontology automatic annotated term, and the suggested target ontology term.

Figure 3 shows a schema with the OBOAEA approach, described previously, using the GO as the source ontology and the INOH Event (IEV) ontology as the target ontology.

5.2 Top-level and domain ontologies association

To establish the relationship among top-level ontologies and domain ontologies, a manual procedure is recommended as the most adequate one [[Mika et al. 2004] ; [Fallahi et al. 2008]; [Damjanović et al. 2007]; [Probst 2006]], because a high level of expertise on the domain is required in order to correctly interpret each domain element within the context of the top-level ontology. There is not a known way of doing that without human intervention. Therefore, in the preparation step we recommend a manual association to be performed between the domain ontologies and the top-level selected ontology.

According to [Probst 2006], the best approach to do this kind of association is to align the most specific concepts of the top-level ontology with the most general concepts of the domain ontology. For each first level concept at the domain ontology, a top-level specific concept is associated. Consequently, the domain concept subclasses (subconcepts) inherit from the top-level concept associated to it. The cost of doing such manual alignment approach is rather small, as there are usually a reduced number of concepts at the first level of a domain ontology.

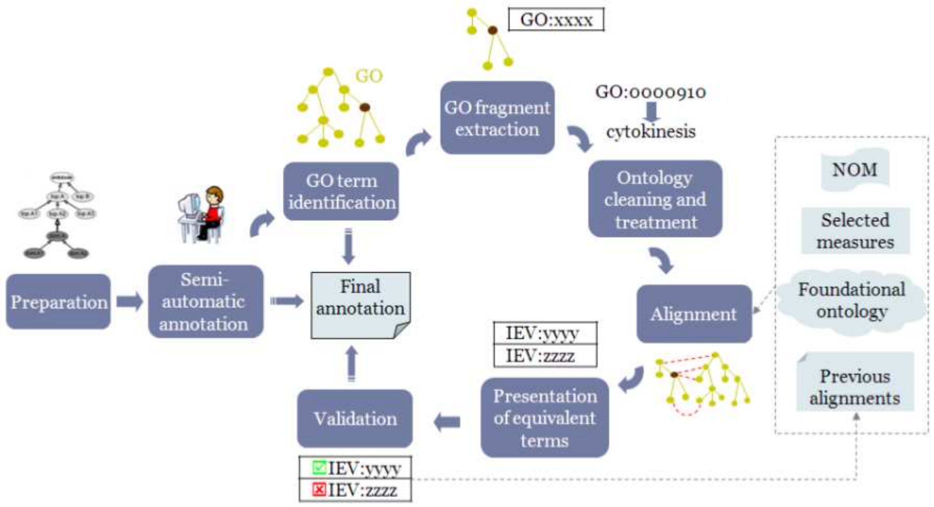

Fig. 3. Proposed approach for biomedical ontology alignment within the genome annotation process

Thus, the result is a unique integrated ontology, composed by the domain ontology and some of the meta-categories of a top-level ontology. The advantage of such association is that the resulting ontology can be used in the alignment process, either as a source or target ontology, automatically benefiting from the additional information brought by the meta-categories. This information is especially useful for the taxonomic similarity measure [Ehrig 2007], as it becomes possible to compare upper level concepts in the hierarchy, when a candidate pair of concepts is under analysis. Briefly, the taxonomic similarity measure recursively compares all the superclasses of the classes under comparison. Given that the top concepts become superclasses of the domain classes, they also become part of the features set used in those classes comparison. Using FOAM, which already includes such measure, it is possible to perform a richer alignment. However, to reach the best results, both source and target ontologies must be associated to the same top-level ontology.

6. EXPERIMENT

To validate the approach to verify if we can obtain a larger number of alignments considered valid by the annotator, the following experiment was performed. The experiment was divided into two stages to enable the comparison among the results originally obtained with a general purpose alignment tool with those obtained using the process adapted to the characteristics of OBO ontologies. In addition to the foundational ontology, only a subset of selected similarity measures was used, as described in Section 5.1. In the first stage, the alignment between GO and other OBO ontology was carried out using only the FOAM tool, and in the second stage the same tool was used, but it was modified to include only the selected similarity measures, counting also with the aid of a foundation ontology, as described in Section 5.2.

In both stages we used the following parameters: fully automatic alignment, number of iterations equals to 10 and cutoff value (i.e., minimum value of similarity for the pair of entities to be considered equivalent) equals to 0.95. In both cases, GO was divided into slices, each one aligned with the second ontology, thus generating subsets of partial results. These subsets were then consolidated, eliminating duplicates and generating for each step a single list of pairs of terms considered equivalent. Therefore, we obtained two sets of results, each one concerning to a step of the experiment. The results of each part were then submitted to the validation of an expert able to assess whether the returned pairs of terms were indeed equivalent. This resulted in a quantitative comparison which allowed assessing objectively whether the features introduced in the alignment process improved the results. The main steps followed in the experiment are described below:

Preparation: The first step was choosing the ontologies to be aligned. The GO was chosen as the source ontology. That was a natural choice because most of the genomic research that adopts ontology-based annotation uses GO for this purpose. As target ontology, INOH Event proved to be the best option since there was already a manual alignment between this ontology and GO. This alignment is derived from a list of 800 GO terms used in the sequencing of the organism T. rangeli, conducted by a research group at Instituto Oswaldo Cruz [Wagner 2006]. Among these 800 terms, we noticed that 26 had verbal coincidence with terms in INOH Event. The identification of this overlapping area between the source and target ontologies helped to delimit the context of the experiment. With respect to the foundational ontology, the most popular ones were taken into account: Cyc [Matuszek et al. 2006], SUMO [Pease et al. 2002], BFO [Grenon et al. 2004], DOLCE [Gangemi et al. 2002], Sowa's Ontology [Sowa 1999] and UFO [Guizzardi and Wagner 2004]. Cyc and SUMO were discarded because they are very large ontologies, which would difficult the manual association between them and the domain ontologies. Sowa's Ontology and UFO, although being well founded conceptual models, don't have an OWL representation, which is required for the physical integration between domain and top-ontologies, as described in Section 5.2. DOLCE and BFO were good candidates, and BFO was chosen because it has already been developed within the biomedical field to guide the development of new ontologies. This fact could simplify the association between top and domain concepts task, since the former has already been developed specifically to categorize biological concepts. Once we have defined the foundational ontology to be used, we could establish the association among the most specific concepts of this ontology with the most general terms of Biological Process (from the GO) and INOH Event, as described in Section 5.2. Discarding the root class ("biological process" in GO and "Event" in INOH Event) at the first level of the two ontologies, there are 21 classes at the second level of Biological Process, which correspond to the most general terms of this ontology. In INOH Event, there are only two classes at the second level, so we chose to use the classes at the third level, totalizing 6 most general concepts for this ontology. This association was made with the aid of a biologist, and the results are presented in Tables II and III. Once such association was established, the Biological Process and INOH Event ontologies were edited to incorporate the foundational ontology as part of their hierarchies.

Selection of terms, extraction of fragments, cleaning and treatment: The fragmentation was based on the list of 26 terms mentioned above. Among them, 25 terms belong to the Biological Process branch of GO, and only one was part of the Molecular Function branch. As the INOH Event also refers to biological processes, we decided to work only with the Biological Process branch, and then 25 small ontologies were generated, all of them having an overlapping area with INOH Event. For the stages of cleaning and treatment we developed small applications to automate these tasks.

Alignment: After the cleaning and treatment phase, the first part of the experiment was done using only FOAM, with no further resources. Alignments between INOH Event and each of the 25 fragments extracted from GO were carried out. In these alignments, 57 pairs of terms were considered equivalents. In the second part, the adapted version of FOAM (considering the selected set of measures) and the foundational ontology BFO were used to perform alignments between GO and INOH Event. FOAM calculates the similarity between a pair of terms using a rule for each similarity measure, so, for taking into account only the chosen measures, the tool was modified to apply only the rules of interest, disabling the other ones. For adding the foundational ontology to the alignment process, no modifications were needed in FOAM, but rather in the domain ontologies, as described in Section 5.2. In this part of the experiment, we obtained 64 pairs of terms considered equivalent, discarding the associations involving terms of the BFO ontology.

Table II. Association between GO (Biological Process) and BFO concepts

Gene Ontology (Biological Process)	BFO Concepts																								
	Generically Dependent Continuant	Quality	Disposition	Function	Role	FiatObjectPart	Object	ObjectAggregate	ObjectBoundary	Site	ZeroDimensionalRegion	OneDimensionalRegion	TwoDimensionalRegion	ThreeDimensionalRegion	FiatProcessPart	Process	ProcessAggregate	ProcessBoundary	ProcessualContext	SpatiotemporalInstant	SpatiotemporalInterval	ScatteredSpatiotemporalRegion	TemporalInstant	TemporalInterval	ScatteredTemporalRegion
Biological process																									
anatomical structure formation																x									
biological adhesion																		x							
biological regulation																			x						
cell killing															x										
cellular process																			x						
death																		x							
developmental process															x										
establishment of localization																x									
growth															x										
immune system process																			x						
localization																		x							
locomotion																x									
metabolic process																			x						
multi-organism process																x									
multicellular organismal process																x									
pigmentation																x									
reproduction																		x							
reproductive process																x									
response to stimulus																			x						
rhythmic process																			x						
viral reproduction																x									

Table III. Association between INOH Event and BFO concepts

INOH Event	BFO Concepts																								
	Generically Dependent Continuant	Quality	Disposition	Function	Role	FiatObjectPart	Object	ObjectAggregate	ObjectBoundary	Site	ZeroDimensionalRegion	OneDimensionalRegion	TwoDimensionalRegion	ThreeDimensionalRegion	FiatProcessPart	Process	ProcessAggregate	ProcessBoundary	ProcessualContext	SpatiotemporalInstant	SpatiotemporalInterval	ScatteredSpatiotemporalRegion	TemporalInstant	TemporalInterval	ScatteredTemporalRegion
Event																									
Biological event																									
Cellular event																			x						
Molecular event																			x						
Organism event																			x						
Physiological event																			x						
Environmental event																									
Medium condition										x															
Treatment																			x						

Validation and analysis of results: The two results sets obtained in the previous step were analyzed and validated by a biologist with expertise in the genome sequencing and annotation area.

Table IV presents the validation results for the two parts of the experiment: on the left side of the column "Results" (Part 1), we present the results with respect to the 57 pairs of terms and on the right one (Part 2), the results with respect to the 64 pairs obtained with the application of our approach.

If we just focus on correct (rating of 5) and incorrect (rating of 0) results, we note that for the 57 pairs of terms considered equivalent by FOAM without additional resources, 47are incorrect. In

Table IV. Experiment validation results

Classification		**Results**			
		Part 1		*Part 2*	
Rating	*Meaning*	*Amount*	*Percentage*	*Amount*	*Percentage*
5	correct	27	47%	27	42%
4	strong relationship	8	14%	12	19%
3	average relationship	7	12%	7	11%
2	weak relationship	2	4%	1	2%
1	insignificant relationship	8	14%	9	14%
0	incorrect	5	9%	8	13%
	Total	**57**	**100%**	**64**	**100%**

Table V. Experiment validation results considering a cutoff value equal to 0.97

Classification		**Results**			
		Part 1		*Part 2*	
Rating	*Meaning*	*Amount*	*Percentage*	*Amount*	*Percentage*
5	correct	27	50%	27	64%
4	strong relationship	8	15%	8	19%
3	average relationship	6	11%	2	5%
2	weak relationship	2	4%	0	0%
1	insignificant relationship	7	13%	4	10%
0	incorrect	4	7%	1	2%
	Total	**54**	**100%**	**42**	**100%**

our approach, for the 64 pairs of terms returned, 42% are correct and 13% are incorrect, showing a decrease of 5% accuracy rate and an increase of 4% in error rate. These results correspond to the set of alignments whose cutoff was established as 0.95, as previously described. This means that all pairs of terms returned have the similarity value higher than 0.95.

Analyzing these similarity values, we note that there is a greater variation in the values of the second part of the experiment compared to values obtained in the first part. If we increase the cutoff value and consider pairs of terms with a similarity value greater than or equal to 0.97, in the first part we discard 3 of 57 pairs and in the second we discard 22 of 64 pairs. The validation results for this subset of pairs obtained from the use of a higher cutoff value are listed in Table V.

In Table V we can see that, using a cutoff value of 0.97, using FOAM we obtained 50% of correct pairs and 7% incorrect pairs. Using our approach, 64% of the aligned pairs were correct, while 2% were totally incorrect. As we can see in Table V, with the new cutoff value we have an increase (about 14%) of the percentage of correct alignment and a decrease (about 5%) of incorrect alignment. In addition, we also note a 4% increase in pairs where the terms have strong relationship, which may also help the annotation process, and a decrease in the percentage of couples who have weaker relationships, which could represent only noise data.

We can conclude from this analysis that although our approach returns a greater number of errors, these incorrect alignments have a lower similarity value and can be filtered using an appropriate cutoff value, which can be obtained experimentally. Moreover, without the resources offered by our approach, FOAM returns alignments (correct, incorrect and related) with very close similarity values, making it difficult to filter out the results.

Another point to be emphasized is the gain on new results obtained by the application of our

Table VI. Gains on new results applying the OBOAEA approach

Classification		**OBOAEA approach results**			
		Old (overlap)		*New*	
Rating	*Meaning*	*Amount*	*Percentage*	*Amount*	*Percentage*
5	correct	17	85%	10	45%
4	strong relationship	2	10%	6	27%
3	average relationship	0	0%	2	9%
2	weak relationship	0	0%	0	0%
1	insignificant relationship	1	5%	3	14%
0	incorrect	0	0%	1	5%
	Total	**20**	**100%**	**22**	**100%**

approach. In Table VI, we notice that, in the second part of the experiment, 27 pairs were validated as correct. Among them, 10 pairs are new results, i.e., pairs that were not found by FOAM in the first part of the trial, but were returned when our approach was applied. With respect to the 8 pairs where the terms have a strong relationship in the second part of the experiment, 6 are new results found by our approach. Table VI shows the gain on new results, splitting the results obtained in the second stage of the experiment into two groups: old results (which have overlapped with the first stage) and the new pairs discovered by our approach.

We can observe in Table VI that 20 of the 42 pairs of terms returned in the second part of the experiment have been recovered in the first part by FOAM. Among them, 17 were assessed as correct, i.e., 85% of the alignments found by both FOAM and our approach are correct. The remaining pairs (22) correspond to the results found exclusively by our approach and, among them, 10 are totally correct, which represents 45% of useful results for the scorer among the new results.

We also notice that, despite the gain of new alignments, some correct pairs were lost. Table VI shows that FOAM returned 27 pairs classified as correct in the first part of the experiment. Among those pairs, 17 were also recovered by our approach. The remaining 10 pairs, despite being assessed as correct, were not found in the second step. This loss may be due to the tool policy comparison, associated with structural changes undertaken in the ontologies, as a result of the integration of the foundational ontology with the hierarchy of domain ontologies. As FOAM compares all classes of the first ontology with all classes of the second one, returning a similarity value for each pair, the comparison among classes of the foundational ontology with the domain ontologies may be impacting negatively on the similarity of neighboring pairs during subsequent iterations.

Some additional adjustments in FOAM, such as comparing classes that belong to the foundational ontology only among themselves, could decrease the number of correct alignments that were not found by our approach, avoiding this way comparisons between upper and domain concepts, that probably return lower similarity values, introducing noise in the results.

7. RELATED WORK

According to [Euzenat and Shvaiko 2007], foundational ontologies are logic-based systems, and therefore the alignment techniques that exploit them are based on semantics. Thus, their use along with other techniques represents an advantage over purely syntactic methods. Nevertheless, there are almost no reports of systems using this approach. So far, the only work we found, developed by [Mascardi et al. 2010], uses a set of algorithms that exploit upper ontology "semantic bridges" in the alignment process. These algorithms perform the alignment between two domain ontologies, and between each of them with an upper ontology, combining later the results obtained in both cases. This approach shows good results, increasing the recall and maintaining a comparable accuracy, with respect to the direct alignment of the two ontologies (without the mediation of the foundational ontology). However, the fact that a fully automatic process was used increases the likelihood of incorrect associations.

The main difference between this work and the proposed approach is that the first one initially executes alignments between foundational and domain ontologies, which returns good results only when the top ontology has also domain specific concepts (this is the case of Cyc and SUMO), and the second one performs alignments only between domain ontologies, integrating the top concepts in their hierarchies, enabling the use of totally domain independent foundational ontologies.

Another type of related work held more often is the alignment of domain ontologies with foundational ontologies, i.e., the combination of concepts from one ontology to higher level categories to solve problems of conceptual ambiguity and semantic heterogeneity. [Mika et al. 2004] and [Fallahi et al. 2008] perform the alignment of Web Services description ontologies with the foundational ontology DOLCE, in order to remove ambiguities, increasing the precision of service discovery. Although the authors call it alignment, the tasks described in these works resemble what we called association between top and domain concepts, performed in the preparation step, before the actual ontology alignment.

8. CONCLUSION

Currently, a large number of biomedical ontologies has been developed, providing a comprehensive vocabulary for the task of annotation. The Gene Ontology is the most important example, being the most (and often the only one) used in this process. The purpose of our work is to facilitate the use of different and multiple biomedical ontologies in the genome annotation process.

By coming up with an ontology alignment approach for the biomedical area, it becomes possible, from a term used in the GO annotation and recovered automatically, to identify equivalent terms in other biomedical ontologies, enabling the annotator to choose which among them has a more detailed description, making it the most appropriate descriptor for annotation.

Performing a comparative experiment, between the results returned by the execution of FOAM (with no additional resources), and those returned by the application of the OBOAEA approach, both assessed by a field professional, it was shown that the approach, in conjunction with a higher cutoff value, provided greater accuracy in the results, leading to a 145% decrease in error rate.

Besides the improvement in the results obtained, another important contribution is the introduction of foundational ontologies in the alignment process, helping to increase the influence of semantic factors in this task, further expanding the universe of information to be explored during the alignment.

Future work includes performing further experiments, involving different sets of domain and foundational ontologies and a larger number of biologists validating the alignments to reinforce the obtained results. Important questions also to be taken into account are improvements in the use of the foundational ontology, such as a wider concept mapping including not only classes in the highest level of the domain ontologies, but also in intermediary levels, and the addition of external resources in the alignment process, like genome textual information from the Web.

ACKNOWLEDGMENT

The authors would like to thank specialists of the Fiocruz Computational Biology and Systems Laboratory for their collaboration with this work.

REFERENCES

Barrell, D., Dimmer, E., Huntley, R. P., Binns, D., O'Donovan, C., and Apweiler, R. The GOA database in 2009 - an integrated Gene Ontology Annotation resource. *Nucleic Acids Research* 37 (Database Issue): 396–403, 2009.

Belloze, K. *Uma Extensão do Processo de Anotação Genômica para Ampliar o Uso e a Evolução Colaborativa de Ontologias no Domínio da Biologia Molecular.* M.S. thesis, Instituto Militar de Engenharia, Rio de Janeiro, Brazil, 2007.

Bodenreider, O. and Stevens, R. Bio-ontologies: current trends and future directions. *Briefings in Bioinformatics* 7 (3): 256–274, 2006.

Damjanović, V., Behrendt, W., Ploßnig, M., and Holzapfel, M. Developing Ontologies for Collaborative Engineering in Mechatronics. In E. Franconi, M. Kifer, and W. May (Eds.), *ESWC 2007.* Lecture Notes in Computer Science, vol. 4519. Springer, pp. 190–204, 2007.

David, J., Guillet, F., and Briand, H. Association Rule Ontology Matching Approach. *International Journal on Semantic Web and Information Systems* 3 (2): 27–49, 2007.

De Bruijn, J., Ehrig, M., Feier, C., Martin-Recuerda, F., Scharffe, F., and Weiten, M. Ontology mediation, merging and aligning. In J. Davies, R. Studer, and P. Warren (Eds.), *Semantic Web Technologies. Trends and Research in Ontology-based Systems.* Wiley and Sons, UK, pp. 95–113, 2006.

Ehrig, M. *Ontology Alignment: Bridging the Semantic Gap.* Springer, 2007.

Ehrig, M. and Sure, Y. FOAM - Framework for Ontology Alignment and Mapping - Results of the Ontology Alignment Evaluation Initiative. In *Proceedings of the K-CAP 2005 Workshop on Integrating Ontologies.* Banff, Canada, pp. 72–76, 2005.

Euzenat, J. and Shvaiko, P. *Ontology Matching.* Springer, 2007.

Fallahi, G. R., Mesgari, M. S., Rajabifard, A., and Frank, A. U. A Methodology Based on Ontology for Geo-Service Discovery. *World Applied Sciences Journal* 3 (2): 300–311, 2008.

Frishman, D. and Valencia, A. *Modern Genome Annotation: The BioSapiens Network.* Springer-Verlag, 2008.

Gangemi, A., Guarino, N., Masolo, C., Oltramari, A., and Schneider, L. Sweetening Ontologies with DOLCE. In V. Benjamins and A. Gomez-Perez (Eds.), *EKAW-2002.* Lecture Notes in Computer Science, vol. 2473. Springer, pp. 166–181, 2002.

Gracia, J. and Mena, E. Ontology Matching with CIDER: Evaluation Report for the OAEI 2008. In *Proceedings of the 3rd International Workshop on Ontology Matching.* Karlsruhe, Germany, pp. 140–146, 2008.

Grenon, P., Smith, B., and Goldberg, L. Biodynamic Ontology: Applying BFO in the Biomedical Domain. In D. M. Pisanelli (Ed.), *Ontologies in Medicine.* IOS Press, The Netherlands, pp. 20–38, 2004.

Guarino, N. The Ontological Level. In R. Casati, B. Smith, and G. White (Eds.), *Philosophy and the Cognitive Science.* Holder-Pivhler-Tempsky, Austria, pp. 443–456, 1994.

Guizzardi, G. *Ontological Foundations for Structural Conceptual Models.* Ph.D. thesis, University of Twente, Drienerlolaan 5, 7522 NB Enschede, The Netherlands, 2007.

Guizzardi, G. Ontology-Driven Conceptual Modeling - II Seminario de Pesquisa em Ontologia no Brasil. http://ontobra.comp.ime.eb.br/apresentacoes/curso2/, 2009.

Guizzardi, G. and Wagner, G. A Unified Foundational Ontology and some Applications of it in Business Modeling. In *Proceedings of the 16th International Conference on Advances in Information Systems Engineering (CAiSE).* Riga, Latvia, pp. 129–143, 2004.

Jean-Mary, Y., Shironoshita, E., and Kabuka, M. Ontology matching with semantic verification. *Journal of Web Semantics: Science, Services and Agents on the World Wide Web* 7 (3): 235–251, 2009.

Jian, N., Hu, W., Cheng, G., and Qu, Y. FalconAO: Aligning Ontologies with Falcon. In *Proceedings of the K-CAP 2005 Workshop on Integrating Ontologies.* Banff, Canada, pp. 85–90, 2005.

Kiu, C.-C. and Lee, C.-S. Ontology Mapping and Merging through OntoDNA for Learning Object Reusability. *Educational Technology & Society* 9 (3): 27–42, 2006.

Lambrix, P. and Tan, H. A System for Aligning and Merging Biomedical Ontologies. *Web Semantics: Science, Services and Agents on the World Wide Web* 4 (3): 196–206, 2006.

Li, Y., Li, J., and Tang, J. RiMOM : Ontology Alignment with Strategy Selection. In *Proceedings of the 6th International Semantic Web Conference.* Busan, South Korea, pp. 51–52, 2007.

Mascardi, V., Locoro, A., and Rosso, P. Automatic Ontology Matching Via Upper Ontologies: A Systematic Evaluation. *IEEE Transactions on Knowledge and Data Engineering* 22 (5): 609–623, 2010.

Matuszek, C., Cabral, J., Witbrock, M., and Deoliveira, J. An Introduction to the Syntax and Content of Cyc. In *Proceedings of the 2006 AAAI Spring Symposium on Formalizing and Compiling Background Knowledge and Its Applications to Knowledge Representation and Question Answering.* Stanford, USA, pp. 44–49, 2006.

McGuinness, D., Fikes, R., Rice, J., and Wilder, S. An Environment for Merging and Testing Large Ontologies. In *Proceedings of the Seventh International Conference on Principles of Knowledge Representation and Reasoning (KR2000).* Breckenridge, USA, pp. 483–493, 2000.

Mika, P., Sabou, M., Gangemi, A., and Oberle, D. Foundations for OWL-S: Aligning OWL-S to DOLCE. *Papers from 2004 AAAI Spring Symposium -Semantic Web Service* 1 (SS-04-06): 52–60, 2004.

Nagy, M., Vargas-Vera, M., and Motta, E. DSSim-ontology mapping with uncertainty. In *Proceedings of the 1 st International Workshop on Ontology Matching (OM-2006).* Athens, USA, pp. 115–123, 2006.

Noy, N. and Musen, M. PROMPT: Algorithm and Tool for Automated Ontology Merging and Alignment. In *Proceedings of the National Conference on Artificial Intelligence*. Austin, USA, pp. 450–455, 2000.

OBO Foundry. The Open Biomedical Ontologies. http://www.obofoundry.org, 2009.

Pease, A., Niles, I., and Li, J. The Suggested Upper Merged Ontology: A Large Ontology for the Semantic Web and its Applications. In *Working Notes of the AAAI-2002 Workshop on Ontologies and the Semantic Web*. Edmonton, Canada, 2002.

Probst, F. Ontological Analysis of Observations and Measurements. In M. R. et al. (Ed.), *GIScience*. Lecture Notes in Computer Science, vol. 4197. Springer, pp. 304–320, 2006.

Smith, B., Ashburner, M., Rosse, C., Bard, C., Bug, W., Ceusters, W., Goldberg, L. J., Eilbeck, K., Ireland, A., Mungall, C. J., Consirtium, T. O., Leontis, N., Rocca-Serra, P., Ruttenberg, A., Sansone, S.-A., Scheuermann, R. H., Shah, N., Whetzel, P. L., and Lewis, S. The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. *Nature Biotechnology* 25 (11): 1251–1255, 2007.

Sowa, J. F. *Knowledge Representation: Logical, Philosophical and Computational Foundations*. Brooks Cole Publishing Co., 1999.

The Gene Ontology Consortium. The Gene Ontology Project in 2008. *Nucleic Acids Research* 36 (1): 440–444, 2008.

Wagner, G. *Geração e análise comparativa de sequências genômicas de Trypanosoma rangeli*. M.S. thesis, Instituto Osvaldo Cruz, Rio de Janeiro, Brazil, 2006.

Wang, P. and Xu, B. Lily : Ontology Alignment Results for OAEI 2008. In *Proceedings of the 3rd International Workshop on Ontology Matching*. Karlsruhe, Germany, pp. 167–175, 2008.