Instance-based matching of large life science ontologies

Toralf Kirsten[1], Andreas Thor[2], Erhard Rahm[1,2]

[1] Interdisciplinary Center for Bioinformatics, University of Leipzig, Germany
[2] Dept. of Computer Sciences, University of Leipzig, Germany

tkirsten@izbi.uni-leipzig.de, {thor,rahm}@informatik.uni-leipzig.de

Abstract. Ontologies are heavily used in life sciences so that there is increasing value to match different ontologies in order to determine related conceptual categories. We propose a simple yet powerful methodology for instance-based ontology matching which utilizes the associations between molecular-biological objects and ontologies. The approach can build on many existing ontology associations for instance objects like sequences and proteins and thus makes heavy use of available domain knowledge. Furthermore, the approach is flexible and extensible since each instance source with associations to the ontologies of interest can contribute to the ontology mapping. We study several approaches to determine the instance-based similarity of ontology categories. We perform an extensive experimental evaluation to use protein associations for different species to match between subontologies of the Gene Ontology and OMIM. We also provide a comparison with metadata-based ontology matching.

Keywords: Ontology matching, instance-based matching, match evaluation

1 Introduction

Ontologies become increasingly important in life sciences application domains. Typically, they are used to semantically describe molecular-biological objects, e.g., to annotate genes and proteins with information on the functions and processes they are involved in. Ontologies also provide controlled vocabularies for a uniform naming of concepts to help reduce variations in terminology. Within an ontology, concepts are usually interrelated with is-a and part-of relationships resulting in specialization/ generalization and aggregation hierarchies (trees) or complex graphs of concepts. A very popular ontology is the Gene Ontology (GO) consisting of three (sub-) ontologies on molecular functions, biological processes and cellular components [7]. Genetic disorders are structured in Online Mendelian in Man (OMIM) [17].

The rapid increase in the number of life science data sources is accompanied by a similar growth in the number of ontologies and mappings between data sources and ontologies. This makes it increasingly valuable to match or align ontologies with each other to determine which of their concepts are semantically related. The resulting ontology mappings can be useful in many ways, in particular for enhanced analysis and annotation of genes, proteins or other objects of interest. For example, such objects may only be assigned to one particular ontology, say GO functions. An ontology mapping between GO functions and GO processes can then be useful to newly assign the objects to the second (process) ontology. Curators could thus use ontology mappings to find missing ontology annotations and get recommendations for possible ontology associations. Conversely, existing ontology associations could be validated against a newly determined ontology mapping in order to locate potential mis-associations reducing data quality. Ontology mappings are also helpful for explorative data analysis, e.g., to find objects with similar ontological properties as interesting targets for a comparative analysis.

Ontology matching is a general problem not limited to life sciences and has become an active research area (see Related Work section). Most previously proposed approaches to determine ontology mappings are metadata-based, i.e., they use the ontology representations themselves to find related concepts, in particular the names of concepts and contextual information like the names of the predecessor and successor concepts within the ontologies. Typically, name similarity is determined using generic (syntactical) string similarity functions on the names. However, in the absence of a globally standardized naming scheme such metadata-based approaches are of potentially little usefulness, especially for life science applications. This is because the same names may refer to completely different concepts while different names may describe the same concept. Furthermore, the concept granularities of different ontologies may widely differ so that comparing names may easily lead to correlations between incomparable concepts.

Figure 1 illustrates some of the problems for sample entries of the GO subontologies on molecular functions (MF) and biological processes (BP). We observe that in both subontologies there are highly similar concept names with partially opposite semantics, e.g., *Ion transporter activity* and *Anion transport* or *Organic anion transporter activity* and *Inorganic anion transport*. A name-based matching between molecular functions with biological processes would probably match these concepts despite potentially opposite semantics, e.g., *Ion* vs. *Anion* and *Organic* vs. *Inorganic.* This fact is also supported by [16] showing that nearly 65% of all concepts found in GO subontologies contain another GO concept as a proper substring. While more sophisticated matchers using helper ontologies like thesauri may somewhat reduce these problems there is no general solution due to the inherent difficulty to agree on common terms and constant creation of new terms.

We therefore advocate for instance-based match approaches which utilize existing associations between ontology concepts and instances, i.e., molecular–biological objects like proteins or genes that are described or annotated by the ontology concepts. This assumes that the real semantics of a concept is often better expressed by such existing object associations rather than metadata like the concept name. The example of Figure 1 shows such associations between species-specific proteins of the Ensembl data source [5] and describing concepts of the GO subontologies MF and BP and genetic disorders (GD) of OMIM. Intuitively, we assume that two concepts of different ontologies are related if their associated instances overlap, i.e., when the same instances are associated to them. The degree of concept similarity should take into account the number of shared associated objects or the relative size of the instance overlap.

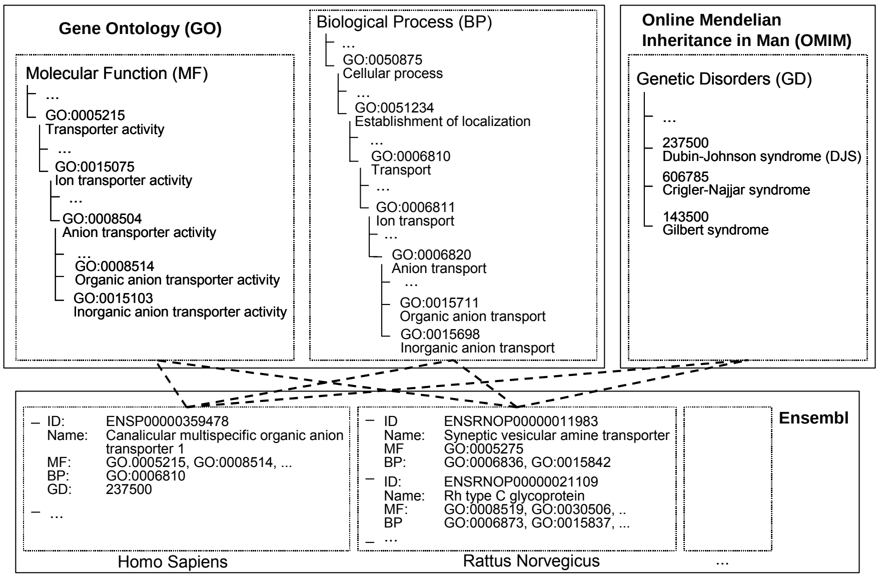

Figure 1: Sample ontology entries and protein associations

We make the following contributions:

- We propose a simple yet powerful methodology for instance-based matching for life science ontologies utilizing existing associations between object data sources and ontologies. We outline several alternatives to determine the instance-based similarity between ontology concepts based on which the ontology mappings are determined. Each data source with associations to the ontologies to match can be used to derive a new ontology mapping. This way the domain-specific knowledge represented by the associations can be utilized to determine semantically meaningful ontology mappings.
- Our approach is flexible and extensible as several mappings between the same ontologies can be combined, e.g., mappings obtained for different data sources, species or similarity metrics. A combination with metadata-based match results is also feasible in order to improve recall and/or precision. Different ways for combining ontology mappings can be employed, e.g., based on intersection or union.
- We provide an extensive experimental evaluation for matching real ontologies, namely the three GO subontologies and OMIM, based on instance data for three species (human, mouse, rat). We consider direct associations between instances and concepts as well as indirect associations which take intra-ontology relationships into account. We also provide a comparison with metadata-based ontology matching. The evaluation utilizes new approximate recall and precision metrics in order to deal with the problem that the perfect ontology mappings are generally unknown.

The rest of the paper is organized as follows. Section 2 introduces the ontologies and instance associations used for our match evaluation. Section 3 presents the similarity metrics we use to derive and evaluate ontology mappings. In Section 4 we discuss the experimental results for instance-based ontology matching while Section 5 provides an experimental comparison with metadata-based ontology matching. Section 6 overviews related work and Section 7 concludes.

2 Match Scenario: Ontologies and Instance Associations

For our study, we assume that ontologies form a directed acyclic graph of concept nodes. The directed edges between concept nodes represent either *is-a* or *part-of* relationships. Concepts can have multiple associated instances, i.e., objects that are described or classified by the concept. An instance can be associated with multiple concepts, both leaf-level concepts but also to inner concepts of the ontology graph. Hence, the associations between objects (instances) and ontology concepts are of cardinality n:m.

Our experimental evaluation covers four popular life science ontologies: the three Gene Ontology (GO) subontologies on molecular functions, biological processes and cellular components, and genetic disorders of OMIM[1]. To match these ontologies with each other we use protein associations for three species: Homo Sapiens (human), Mus Musculus (mouse) and Rattus Norvegicus (rat). The protein data and ontology associations are obtained from the Ensembl data source (www.ensembl.org).

Table 1 provides base statistics on the considered ontologies, species-specific instance data sources and protein-concept associations. The number of concepts per ontology is shown on top, the number of proteins per species on the left. For instance,

Table 1: Quantity structure of utilized ontologies and instance sources*

#concepts				Gene Ontology									OMIM**		
#proteins	#assoc.			Molecular Functions			Biological Processes			Cellular Components			Genetic Disorders		
				7,514			12,555			1,848			6,535		
Ensembl	(direct assoc.)	Homo Sapiens	43,605	34%	52%	58,539	24%	45%	52,536	34%	44%	37,640	25%	4%	2,618
		Mus Musculus	32,078	31%	61%	57,997	22%	53%	47,646	32%	54%	36,288	0%	0%	0
		Rattus Norvegicus	33,745	29%	38%	29,665	22%	33%	25,703	29%	31%	18,519	0%	0%	0
	(indirect assoc.)	Homo Sapiens	43,605	39%	52%	164,014	35%	45%	209,283	43%	44%	149,548	25%	4%	2,618
		Mus Musculus	32,078	36%	61%	145,646	33%	53%	181,583	40%	54%	139,841	0%	0%	0
		Rattus Norvegicus	33,745	34%	38%	85,429	32%	33%	107,022	37%	31%	75,919	0%	0%	0

* Release states: GO 01/20/2007, OMIM 01/28/2007, Ensembl Release 42 Dec. 2006

** We focus on phenotype descriptions, i.e., entries marked with #, % and without a mark. Please see http://www.ncbi.nlm.nih.gov/Omim/mimstats.html for more details.

[1] OMIM was not originally developed as an ontology but provides a comprehensive set of terms (including term definitions, comments and associated literature) describing genetic disorders which are frequently associated with objects of other data sources. Therefore, OMIM plays an ontology-like role in our evaluation study.

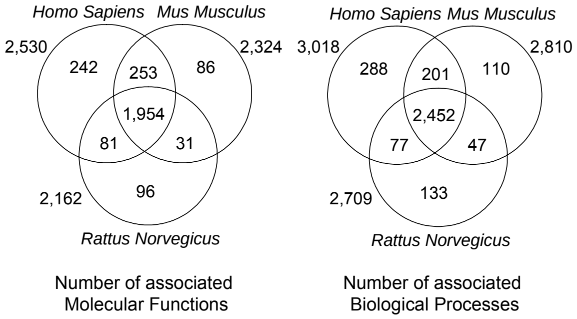

Figure 2: Quantity structure of ontology concepts with at least one associated protein in three selected species

there are 7,514 molecular function concepts in GO and 43,605 human proteins in Ensembl. Furthermore, Table 1 contains the number of associations between proteins and ontology concepts which we separate in direct and indirect associations. Direct associations refer to the original associations recorded in Ensembl and assign objects to the most specific concept of an ontology. For example, there are 58,539 direct associations between human proteins and molecular functions and covering 52% of the human proteins and 34% of the functions. Hence, human protein associations support instance-based matching for up to 34% of the MF concepts. To increase the number of concepts that may be matched we also consider indirect associations which take into account the intra-ontology relationships between concepts. For this we simply assign the direct instances of a concept *c* also to its parents and grandparents within the ontology graph. In the example this provides human protein instances to 39% of the function concepts, however at the expense of a massive increase in the number of object associations (164,014).

We observe that the available object associations cover significant portions of the ontologies (25-39%) so that instance-based matching promises to provide many correspondences between concepts. While OMIM has associations only for human proteins, the GO ontologies are well connected to all three species. There is a similar number of object associations for human and mouse proteins while the coverage for rat is somewhat reduced. On average, an Ensembl protein is directly assigned to 1.5-3.0 concepts of the GO subontologies. The average number of directly associated proteins per GO concept varies between 9 and 62 per species.

Figure 2 illustrates the species-specific distribution of object associations for the three GO subontologies. For example, we observe that 1,954 molecular functions have protein associations to all three species, whereas merely 86 functions are exclusively associated with mouse proteins. On average over 80% of the matched concepts, i.e., functions, processes, or components, are assigned to all three species. Considering species-specific associations is also helpful to determine species-specific ontology mappings. Furthermore, analysis tasks can benefit from focusing on species-specific concepts, e.g., to analyze an ontology mapping for the 86 mouse-specific GO functions with respect to the 110 mouse-specific processes.

3 Similarity and Evaluation Metrics

In order to match two ontologies O_1 and O_2 we need metrics to determine the similarity between O_1 concepts and O_2 concepts. All pairs of concepts from O_1 and O_2 for which the similarity exceeds a certain minimal threshold are called *correspondences* and included in the match result (ontology mapping). The key idea of our instance-based approach to ontology matching is to derive the similarity between concepts from the number of shared instances, i.e., the number of instances associated to both concepts. An important advantage for instance-based ontology matching is that the number of instance associations is typically higher than the number of concepts. This way the match accuracy of the approach can become rather robust against some wrong instance classifications. As discussed, another key advantage is that the instance-based approach is independent from concept names and other metadata.

In the following we first present the used instance-based similarity metrics. We then discuss how to assess the quality of a match result in the absence of a perfect mapping.

3.1 Instance-based similarity metrics

In this paper we study four metrics for determining the instance-based similarity between concepts $c_1 \in C_{O1}$ and $c_2 \in C_{O2}$ of different ontologies O_1 and O_2, namely *baseline*, *minimum*, *dice*, and *kappa* similarity. Most of these metrics are well-know and have already been used in previous match studies (e.g., [8, 21]) however, not yet for an instance-based matching of life science ontologies. To define the similarity of two concepts c_1 and c_2 we use the number of instances that are (or are not) associated to c_1 and c_2. Figure 3 illustrates all relevant combinations for the instance cardinalities.

For example, $N_{c_1\overline{c_2}}$ is the number of instances which are associated to c_1 but not associated to c_2. Furthermore, $N_{c_1}\left(N_{\overline{c_1}}\right)$ is the total number of instances that are (not) associated to c_1. Note that these numbers may be used either for directly associated instances as well as for indirectly associated instances.

The *baseline similarity* metric already matches two concepts c_1 and c_2 if they share at least one object.

$$Sim_{Base}(c_1, c_2) = \begin{cases} 1 \text{ , if } N_{c_1c_2} > 0 \\ 0 \text{ , if } N_{c_1c_2} = 0 \end{cases} \in [0...1], \forall c_1 \in C_{O1}, c_2 \in C_{O2}$$

The baseline approach poses minimal requirements to match two concepts so that it

	$i \in c_2$	$i \notin c_2$	Σ
$i \in c_1$	$N_{c_1c_2}$	$N_{c_1\overline{c_2}}$	N_{c_1}
$i \notin c_1$	$N_{\overline{c_1}c_2}$	$N_{\overline{c_1}\,\overline{c_2}}$	$N_{\overline{c_1}}$
Σ	N_{c_2}	$N_{\overline{c_2}}$	N

Figure 3: Matrix of all possible combinations for the number of shared instances i for two concepts $c_1 \in C_{O1}$ and $c_2 \in C_{O2}$.

can be expected to provide the maximal number of correspondences for instance-based matching. To focus on concept combinations with a higher instance overlap it is necessary to take into account the number of instances per concept.

The *dice similarity metric* [19] considers the concept cardinalities and the number of shared instances:

$$Sim_{Dice}(c_1, c_2) = \frac{2 \cdot N_{c_1 c_2}}{N_{c_1} + N_{c_2}} \in [0...1], \forall c_1 \in C_{O1}, c_2 \in C_{O2}$$

A high dice value indicates a significant instance overlap w.r.t. to both concepts.

A potential limitation of the dice metric is that it can become quite small in case of larger cardinality differences, even if all instances of the smaller concept match to another concept. This aspect is taken care of by the *minimum similarity* metric which determines the relative instance overlap with respect to the smaller-sized concept:

$$Sim_{Min}(c_1, c_2) = \frac{N_{c_1 c_2}}{\min(N_{c_1}, N_{c_2})} \in [0...1], \forall c_1 \in C_{O1}, c_2 \in C_{O2}$$

Our last metric – the *kappa similarity* – is somewhat more complex and adopted from Cohen's kappa coefficient [6]; it has also been adopted in [8] for an e-commerce application. The kappa coefficient measures the agreement of two raters classifying items (e.g., instances) into categories (e.g., concepts). We adopt the kappa coefficient to calculate two probabilities *P* and *P'*. *P* is the agreement among both concepts, i.e., the relative number of shared instances combined with the number of instances that do not appear in any of the two concepts. *P'* is the probability that the agreement that one instance is assigned to both concepts is due to chance. Therefore P and P' are defined as follows:

$$P = \frac{N_{c_1 c_2} + N_{\overline{c_1 c_2}}}{N} \qquad P' = \frac{N_{c_1} \cdot N_{c_2} + N_{\overline{c_1}} \cdot N_{\overline{c_2}}}{N^2}$$

The kappa similarity for two concepts c_1 and c_2 is then defined as:

$$Sim_{Kappa}(c_1, c_2) = \frac{P - P'}{1 - P'} \in [0...1], \forall c_1 \in C_{O1}, c_2 \in C_{O2}$$

To test the significance of a match between the two concepts $c_1 \in C_{O1}$ and $c_2 \in C_{O2}$, we can utilize a test distribution *Z* as proposed in [8]. *Z* is defined as follows:

$$Z = Sim_{Kappa}(c_1, c_2) \cdot \sqrt{\frac{(N_{c1c2} + N_{c1\overline{c2}} + N_{\overline{c1}c2} + N_{\overline{c1c2}})(1 - P')}{P'}}$$

Z follows a normal distribution so that it can be compared with the standard normal distribution. A significant match correspondence can be assumed if *Z* exceeds the percentile of the standard distribution for a given significance level.

It can easily been shown that for all correspondences between concepts c_1 and $c_{2,}$ it holds:

$$Sim_{DICE}(c_1, c_2) \le Sim_{MIN}(c_1, c_2) \le Sim_{Base}(c_1, c_2) \text{ and}$$
$$Sim_{Kappa}(c_1, c_2) \le Sim_{Base}(c_1, c_2)$$

3.2 Evaluation metrics

To evaluate the quality of a match result and thus the effectiveness of a match approach it is necessary to determine whether all real correspondences have been determined (completeness, high recall) and whether all determined correspondences are real correspondences (correctness, high precision). Exactly determining recall and precision thus requires the perfect match result to be known. Unfortunately, the perfect match result is generally unknown for large real-life match problems, especially for life science ontologies. A manual construction of a perfect match is also too laborious and extremely difficult for broad ontologies such as the Gene Ontology. For our evaluation we therefore focus on the relative quality of the differently obtained match results and use rough approximations for recall and precision.

With respect to recall or completeness we consider the so-called *match coverage,* i.e., the share of concepts that is covered by an ontology mapping, i.e., for which there is at least one correspondence in the match result. Let $C_{O1\text{-}Match}$ ($C_{O2\text{-}Match}$) be the set of matched concepts of ontology O_1 (O_2) and C_{O1} (C_{O2}) the set of all concepts of ontology O_1 (O_2). We then define the match coverage of ontology O_1 (O_2) as follows:

$$MatchCoverage_{O_1} = \frac{|C_{O_1-Match}|}{|C_{O_1}|} \qquad MatchCoverage_{O_2} = \frac{|C_{O_2-Match}|}{|C_{O_2}|}$$

Match coverage can be determined for any match approach, in particular both metadata-based and instance-based schemes. For instance-based approaches the maximal coverage is limited by the number of concepts which have at least one associated instance (w.r.t. the considered instance data source). To take this into account we additionally determine the *instance match coverage* which is defined as the ratio of the matched concepts w.r.t. to the concepts having at least one associated instance. Let $C_{O1\text{-}Inst}$ ($C_{O2\text{-}Inst}$) be the set of concepts of ontology O_1 (O_2) having at least one associated instance. We then define the O_1-specific and O_2-specific instance match coverage as follows:

$$InstMatchCoverage_{O_1} = \frac{|C_{O_1-Match}|}{|C_{O_1-Inst}|} \qquad InstMatchCoverage_{O_2} = \frac{|C_{O_2-Match}|}{|C_{O_2-Inst}|}$$

In addition, we can define the combined instance match coverage for a match result:

$$InstMatchCoverage = \frac{|C_{O_1-Match}| + |C_{O2-Match}|}{|C_{O_1-Inst}| + |C_{O_2-Inst}|}$$

For estimating the precision of a match approach we determine the so-called *match ratio*, i.e., the ratio between the number of found correspondences and the number of matched concepts:

$$MatchRatio_{O1} = \frac{|Corr_{O1-O2}|}{|C_{O1-Match}|} \qquad MatchRatio_{O2} = \frac{|Corr_{O1-O2}|}{|C_{O2-Match}|}$$

Analogously we define the combined *match ratio*.

$$MatchRatio = \frac{2 \cdot |Corr_{O1-O2}|}{|C_{O1-Match}| + |C_{O2-Match}|}$$

In the above formulas, $Corr_{O1\text{-}O2}$ denotes the set of found correspondences in a match result. The intuition is that the precision (and thus value) of a match result is better if a concept is not loosely matched to many other concepts but only to fewer (preferably the most similar) ones. The match ratio for the baseline matcher is expected to provide a worst-case value for instance-based matching and can thus be used as a yardstick.

4 Instance-based Match Results

We first analyze different instance-based match results using direct association. We then study the impact of combining different match results (mappings) and the use of indirect associations.

4.1 Match results using direct association

We applied the introduced instance-based similarity metrics to determine ontology mappings between the three GO ontologies on molecular functions (MF), biological processes (BP), cellular components (CC) and genetic disorders (GD) of OMIM. We thus solved six match tasks: three to match between the GO subontologies (MF-BP, MF-CC, BP-CC) and three GO-OMIM match tasks (MF-GD, BP-GD, CC-GD). As discussed in Section 2, we utilize the Ensembl protein associations for the three species Homo Sapiens, Mus Musculus and Rattus Norvegicus and first focus on direct associations. The three similarity metrics Sim_{Base}, Sim_{Min}, and Sim_{Dice} are evaluated with a high similarity threshold of 1.0; for Sim_{Kappa} we applied a significance level of 95%.

Figure 4 illustrates the obtained values for combined *instance match coverage* for the three GO match tasks and the three considered species. Table 2 shows the corresponding *match ratios* for Homo Sapiens; the match ratios for the other species are similar and omitted due to space constraints. We observe that there are big differences between the considered similarity metrics while the match coverage results are very similar for the three species. The latter is because the species-specific proteins match the same concepts to a large degree (as noted in Section 2) so that the derived ontology mappings are also very similar for a given similarity metric and match task. As expected the baseline similarity metric Sim_{Base} achieved the best coverage (recall) and worst match ratios (precision) for all match tasks. Its instance match coverage is up to 99% (for Homo Sapiens and the MF – BP match) so that almost every concept with an associated instance is matched. On the other hand, match ratios achieve values between about 8 and 46, i.e., concepts are mapped to many other concepts indicating a low precision. On the other hand, Sim_{Dice} and Sim_{Kappa} turn out to be very restrictive with match ratios close to 1.0. This is they focus on the best matching concepts. Unfortunately this is only achieved for very few correspondences so that the match cov-

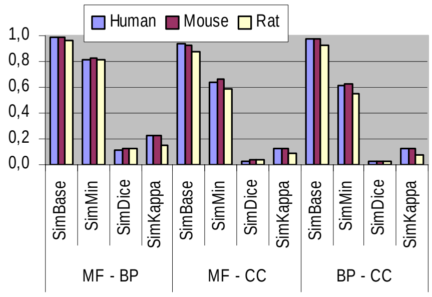

Figure 4: Combined Instance Match Coverage of GO ontology mappings (direct associations)

Table 2: Match Ratios of GO ontology mappings (direct associations; Homo Sapiens)

	MF – BP		MF – CC		BP – CC	
	MF	BP	MF	CC	BP	CC
Base	20.4	17.0	7.6	28.6	9.8	46.3
Min	4.4	4.0	2.2	7.8	2.4	8.6
Dice	1.3	1.2	1.0	1.3	1.0	1.3
Kappa	2.0	2.0	1.9	2.7	1.7	2.6

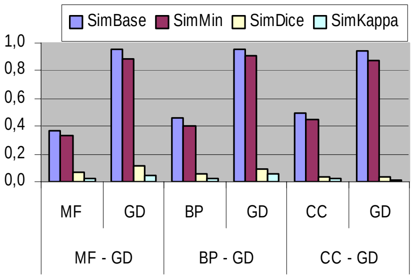

Figure 5: Instance Match Coverage of GO-GD mappings (direct associations; Homo Sapiens)

Table 3: Match Ratios of GO-GD mappings (direct associations; Homo Sapiens)

	MF – GD		BP – GD		CC – GD	
	MF	GD	BP	GD	CC	GD
Base	7.1	4.3	2.5	6.3	2.5	3.4
Min	5.9	3.5	2.5	4.6	1.7	3.4
Dice	1.6	1.5	1.1	1.5	1.4	1.4
Kappa	1.4	1.2	1.1	1.2	1.1	1.2

erage remains rather low (around 5-10% for Sim_{Dice} and 10-20% for $\text{Sim}_{\text{Kappa}}$). For all match tasks the metric Sim_{Min} achieves very promising precision/recall values which lie between the extreme cases discussed so far. In particular instance match coverage is as good as between 60-80% while match ratios are significantly lower than for $\text{Sim}_{\text{Base.}}$ On average, a concept is matched with 2–9 concepts of another ontology which is still a reasonably low number, e.g., to be checked by a biologist.

The GO-OMIM match tasks are only performed for the species Homo Sapiens since there are no protein associations to OMIM for the other two species. Figure 5 shows the instance match coverage for the three match tasks; Table 3 illustrates the corresponding match ratios. For these experiments (and in contrast to the previous match tasks) we observed substantial coverage differences for the individual ontologies so that we indicate the ontology-specific coverage values in Figure 5. We observe that for both Sim_{Base} and Sim_{Min} the instance match coverage of the GO ontologies is only about half of the instance coverage of GD (40-50% vs. more than 88%). The reasons are twofold. On the one hand, many proteins are associated with concepts of the GO ontologies but have no correspondence to OMIM. On the other hand, if a protein is associated with OMIM then it is mostly also connected with a concept of the GO ontologies. For instance, there are 20,936 proteins of the Homo Sapiens that have at least one molecular function, but only 1,581 of these proteins are associated with a genetic disorder. Conversely, only 110 human proteins are described by a genetic disorder but not by a molecular function.

The relative outcome for the different similarity metrics is in agreement with the observations made for the previous match tasks. While Sim_{Base} and Sim_{Min} have a relatively high recall (instance match coverage), the metrics Sim_{Dice} and $\text{Sim}_{\text{Kappa}}$ are very restrictive but precise (only about 1 to 2 correspondences per matched concept on average).

4.2 Combining ontology mappings

The match results discussed so far were each derived for a certain similarity metric and a species-specific set of instances. Combining several such ontology mappings for a given match task is a promising way to obtain an improved ontology mapping, e.g., with improved recall and/or precision. For example, taking the union of two independently derived ontology mappings is likely to improve recall (coverage) while

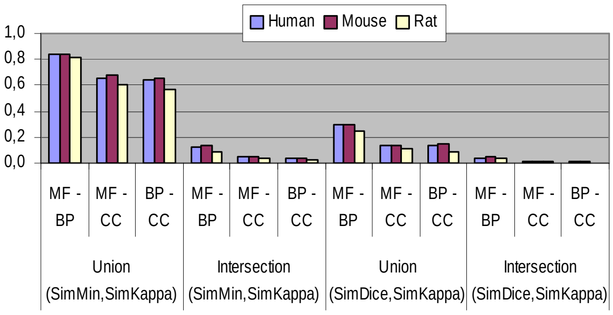

Figure 6: Combined Instance Match Coverage of combined GO ontology mappings

Table 4: Match Ratios of combined GO ontology mappings (Homo Sapiens)

	Match		∪	∩
Min – Kappa	MF-BP	MF	4.6	1.2
		BP	4.2	1.3
	MF-CC	MF	2.4	1.0
		CC	8.0	1.3
	BP-CC	BP	2.5	1.0
		CC	8.8	1.2
Dice – Kappa	MF-BP	MF	1.8	1.1
		BP	1.8	1.1
	MF-CC	MF	1.8	1.0
		CC	2.6	1.3
	BP-CC	BP	1.6	1.0
		CC	2.4	1.5

building the intersection can improve precision. Other combination strategies are also conceivable (e.g., weighted or majority-based selection of correspondences) but are not further considered in this paper.

To illustrate the idea we analyze the combination of mappings obtained for different similarity metrics. This is not useful for all metrics since according to Section 3 all instance-based similarity measures generate subsets of correspondences of the baseline approach and Sim_{Dice} produces a subset of correspondences of Sim_{Min} . Therefore, we comparatively study the intersection and union of the ontology mappings generated by Sim_{Dice} (Sim_{Min}) and $\text{Sim}_{\text{Kappa}}$.

Fig. 6 depicts the instance match coverage of the combined mappings between GO ontologies, while Table 4 shows the corresponding match ratios (for Homo Sapiens). We observe that the union mappings for Sim_{Min} and $\text{Sim}_{\text{Kappa}}$ only slightly improve coverage (84%) compared to Sim_{Min} (81%). The match ratios are also not significantly higher than for Sim_{Min} alone (Table 2). This is because Sim_{Min} alone achieved already a high coverage so that $\text{Sim}_{\text{Kappa}}$ could add only few new correspondences. On the other hand, the union mapping between Sim_{Dice} and $\text{Sim}_{\text{Kappa}}$ is very effective and more than doubles coverage (30%) compared to Sim_{Dice} alone (12%). The match ratio still remains low (1.8–2.6) indicating a high-quality ontology mapping.

4.3 Match results using indirect instance associations

Another way to improve match coverage is to not only consider direct but also indirect object associations. As already discussed in Section 2 (Table 1), this increases the number of concepts for which instance-based matching can be applied (e.g., the number of GO processes with associated instances is increased by 45%). Although we restrict the propagation of object associations to two levels (parents, grandparents) the number of object associations is increased by almost a factor of 3 compared to direct associations.

Figure 7 shows the instance match coverage results for the GO match tasks using indirect associations; Table 5 illustrates the corresponding match ratios (for the species Homo Sapiens). The coverage for Sim_{Base} was already high for direct associations; the use of indirect associations primarily is thus little helpful but leads to ex-

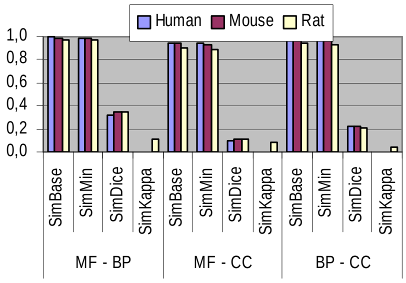

Figure 7: Combined Instance Match Coverage of GO ontology mappings (indirect associations)

Table 5: Match Ratios of GO ontology mappings (indirect associations; Homo Sapiens)

	MF – BP		MF – CC		BP – CC	
	MF	BP	MF	CC	BP	CC
Base	90.2	60.4	27.2	96.2	38.8	210.2
Min	16.6	10.9	7.4	23.3	6.6	33.9
Dice	1.9	1.3	1.2	1.7	1.6	1.8
Kappa	6.7	5.1	5.5	7.6	6.3	11.7

tremely high match ratios (27–210). For $\text{Sim}_{\text{Min,}}$ on the other hand, the instance match coverage improvement is substantial, e.g., from 61% (direct) to 86% (indirect) for the match BP - CC. However, match ratios are also increased, e.g., from 6 (direct, MF) to almost 17 (indirect, MF) for matching MF with BP using Sim_{Min} and human proteins.

The results suggest that the use of indirect associations can be helpful but also be harmful. Hence we see a need for more sophisticated approaches to intelligently make use of intra-ontology relationships in combination with instance-based matching. One idea is to restrict the use of indirect associations to concepts that remain otherwise unmatched. Another option is to use direct associations to determine instance-based concept similarities which are then propagated along intra-ontology relationships by a context matcher [18].

5 Metadata-based Match Results

5.1 Metadata match results using concept names

For comparison purposes we also use a simple metadata-based matcher to determine mappings between the considered ontologies. We apply a name matcher based on trigram similarity for comparing pairwise the concept names of different ontologies. Table 6 shows the name matcher results for the six match tasks by using the trigram similarity and different thresholds (≥ 0.5). Note that the match coverage values refer to all concepts not only to the ones with instances.

We observe a rather low number of correspondences especially for a similarity threshold of 0.7 or higher. This indicates a high diversity in the concept names so that name matching is not very effective. There are no correspondences with a threshold of 0.9 or greater (not shown in Table 6). The match coverage and match ratios grow for smaller similarity thresholds but probably due to many wrong correspondences.

Most correspondences are found between molecular functions and biological processes which are the largest ontologies considered (Table 1). As already indicated by the examples in Figure 1 many similar terms only differ in pre-/suffixes or an additional word, such as *activity* for naming a function. Moreover, in many cases concepts inherit their name from their parents and use an additional term representing the specialization, such as *transport*, *anion transport* (both BP), *transporter activity* and

Table 6: Name matching results between selected ontologies

Sim$_{\text{Trigram}}$		0.5			0.6			0.7			0.8		
Match		# Corres-pondences	Match Coverage	Match Ratio	# Corres-pondences	Match Coverage	Match Ratio	# Corres-pondences	Match Coverage	Match Ratio	# Corres-pondences	Match Coverage	Match Ratio
MF – BP	MF	15,415	47%	4.4	2,770	15%	2.4	602	6%	1.4	69	<1%	1.1
	BP		18%	6.9		8%	2.9		3%	1.4		<1%	1.1
MF – CC	MF	2,663	14%	2.5	1,274	6%	2.7	225	3%	1.1	31	<1%	1.1
	CC		23%	6.3		15%	4.6		8%	1.5		1%	1.2
BP – CC	BP	2,563	8%	2.5	693	3%	1.7	175	1%	1.4	32	<1%	1.1
	CC		40%	3.4		17%	2.0		7%	1.4		2%	1.2
MF – GD	MF	667	7%	4.4	124	2%	1.1	27	<1%	1.0	1	<1%	1.0
	GD		2%	6.9		1%	2.1		<1%	1.1		<1%	1.0
BP – GD	BP	1,400	9%	1.3	174	1%	1.0	11	<1%	1.0	2	<1%	1.0
	GD		2%	8.8		<1%	5.0		<1%	1.1		<1%	1.0
CC – GD	CC	364	11%	1.8	36	2%	1.2	1	<1%	1.0	0	0%	0.0
	GD		1%	5.9		<1%	2.3		<1%	1.0		0%	0.0

anion transporter activity (both MF). Hence, if the additional word is short enough then concepts from different levels are matched, e.g., *anion transport* with *transport activity*. Of course, a low threshold (e.g., 0.5) can lead to the generation of false correspondences, e.g., between the function *Inorganic anion transporter activity* (MF) and the process *Organic anion transport* (BP) due to a trigram similarity of 0.66.

Most correspondences for OMIM GD are found for the GO subontology on biological processes. The reason is that some genetic disorders refer to biological processes, such that their names only differentiate in modified suffixes or additional words. For instance, the concepts *vitamin A metabolism* (BP) and *vitamin A metabolic defect* (GD) are matched with a trigram similarity of 0.72. Of course, low threshold values also lead to false positives matches, such as *betaine transport* (BP) and *citrulline transport defect* (GD) with a trigram similarity of 0.5.

5.2 Comparison between metadata and instance-based matching

To study the relationship between metadata- and instance-based matchers, we analyze the union and intersection (overlap) of the generated ontology mappings. For this purpose, we combine the name matcher results (threshold ≥ 0.5) with the instance-based results using the similarity metric Sim$_{\text{Min}}$ (similarity threshold = 1) and direct instance associations. Figure 8 shows the match coverage per ontology for the union results (species Homo Sapiens). The highest coverages are achieved for molecular functions (approx. 60%) in the combined MF–BP match result and for cellular components (54%) in the BP–CC result, both when using a trigram similarity of 0.5. These high coverage values are mainly due to the name matcher. According to Table 6, the name-based correspondences for threshold 0.5 cover already 47% of the functions (match MF-BP) and 40% of the components (match BP-CC). For trigram thresholds of 0.6 and higher, match coverage is primarily influenced by instance-based matching using Sim$_{\text{Min}}$. This is also the case for the unified match results between GO subontologies and OMIM; around 22% of the genetic disorders and be-

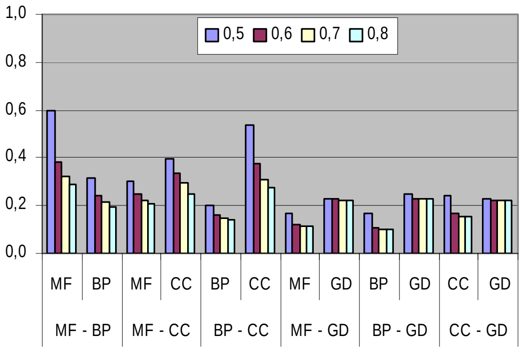

Figure 8: Match Coverage per ontology for unified matches of Sim_{Name} and Sim_{Min} (Homo Sapiens)

Table 7: Match Ratios of combined ontology mappings between Sim_{Name} (0.7) and Sim_{Min} (Homo Sapiens)

Match		∪	∩
MF-BP	MF	4.1	1.0
	BP	3.7	1.0
MF-CC	MF	2.2	1.0
	CC	6.7	1.0
BP-CC	BP	2.4	1.0
	CC	7.6	1.3
MF-GD	MF	5.8	1.0
	GD	3.4	1.0
BP-GD	BP	5.4	0.0
	GD	4.5	0.0
CC-GD	CC	12.9	0.0
	GD	2.5	0.0

tween 11% (MF, BP) and 15% (CC) of GO subontology concepts are covered by using Sim_{Min}.

The match coverage of the intersection results is in most cases only 1% or less (and therefore not shown in an extra plot). This is because the name-based and instance–based match results have only a very low number of correspondences in common. Especially for a lower trigram threshold (0.5) the vast majority of name correspondences has no corresponding instance similarity.

Table 7 illustrates the achieved match ratios for both, the union and intersection of the ontology mappings generated by the name matcher (similarity threshold = 0.7) and Sim_{Min}. We observe a moderate ratio (mostly less then 6) for the union results while the ratios for mapping intersection is seldom larger than 1.0. This is influenced by the fact that the number of correspondences is very low. The intersection of the mappings between genetic disorders and biological processes (cellular components) is even empty, therefore the match ratios also equal zero.

The experiment shows that simple name matching is not very effective and less promising than the proposed instance-based approaches. Still we believe that more sophisticated metadata-based matchers may be helpful to complement instance-based matching and leave the investigation of such combined approaches for future work.

6 Related Work

Overviews of approaches for ontology matching in general are given in [18, 10, 2, 20]. Typically, matching utilizes metadata, associated instances or both. The match approaches [13, 15, 12, 1] are based on metadata, such as concept names, synonyms and descriptions, and applied in different domains. More specific to bioinformatics, [4] utilizes a metadata matcher to link GO with ChEBI, an ontology of chemical entities for biological interest.

Instance-based ontology matching is investigated in [8, 9, 3, 11]. They follow statistical or machine learning approaches and apply them in different application domains. [8] focuses on integrating internet catalogs, represented by hierarchical collections of web links. Similar to our study, it applies the Kappa similarity metric including a significance test. [9] applies decision trees and Bayesian networks to create matches between GO subontologies that is different to our approach. It uses available annotations (instances) of two species (mouse and human) as training data and for cross validation to test the models. In contrast to our approach using the proposed evaluation metrics, the predicted match result is evaluated by a manual selection of 100 correspondences which are then validated by an expert (41 judged to be true, 42 judged to be plausible). [3] utilizes three non-lexical approaches to create ontology matches, namely a vector space model, a statistical co-occurrence analysis and association rule mining. In contrast to our match application where we are interested in correspondences between GO ontologies, they associate GO concepts without a distinction whether the concept is a function, process or component. Therefore, the result can also contain associations between concepts of the same GO subontology, e.g., between two functions. [11] applies association rule mining and formal ontological concepts to create mappings between the GO subontologies whereas we use simple and comprehensible metrics for ontology matching.

[14] is a mixed match approach, i.e., it follows lexicographic and instance-based approaches, with the goal to create a second ontology layer that maps the GO subontologies. Instead of using complete concepts names as we have applied they create specific patterns for the metadata-based matching such that ontology-specific words (e.g., *activity* for molecular functions) are ignored. Moreover, it applies association rule mining by using available gene annotations and reuses existing associations to metabolic pathways to create ontology matches. In contrast to our match scenario, the generated matches are validated by human experts.

Conclusions

We proposed the use of simple instance-based approaches for matching life science ontologies. The idea is to utilize the domain knowledge expressed in existing object-ontology associations for finding related concepts in different ontologies. The approach is extensible as ontology mappings obtained for different match approaches or different instance sources (e.g., different species) can be combined to improve overall recall or precision. We experimentally evaluated four alternatives for instance-based matching and one metadata-based approach for six match tasks involving the GO subontologies and OMIM. We observed that instance-based matching using the Sim_{Min} metric achieves a high match coverage while limiting the number of correspondences per matched concept.

In future work, we will further study combined approaches for ontology matching and the interplay between instance-based and metadata-based matching in life sciences. We also plan applications that utilize the computed ontology mappings and gather user feedback to help validate the proposed match correspondences.

References

1. D. Aumüller, H.-H. Do, S. Massmann, E. Rahm: Schema and ontology matching with COMA++. Proc. ACM SIGMOD, 2005.
2. A. Avesansi, F. Giunchiglia, M. Y. Yatskevich: A large taxonomy mapping evaluation. Proc. 4th Int. Semantic Web Conference (ISWC), 2005.
3. O. Bodenreider, M. Aubry, A. Bugrun: Non-lexical approaches to identifying associative relations in the Gene Ontology. Proc. Pacific Symposium on Biocomputing, 2005.
4. O. Bodenreider, A. Bugrun: Linking the Gene Ontology to other biological ontologies. Proc. ISMB meeting on Bio-Ontologies, 2005.
5. T. Hubbard, D. Andrews, M. Caccamo et al.: Ensembl 2005. Nucleic Acid Research 33(Database Issue): D447-D453, 2005.
6. J. Cohen: A coefficient of agreement for nominal scales, Educational and Psychological Measurement 20: 37–46, 1960.
7. The Gene Ontology Consortium: The Gene Ontology (GO) database and informatics resource. Nucleic Acids Research, 32: D258-D261, 2004.
8. R. Ichise, H. Takeda, S. Honiden: Integrating multiple internet directories by instance-based learning. Proc. 18th Intl. Joint Conf. on Artificial Intelligence (IJCAI), 2003.
9. O. D. King, R. E: Fougler, S. S. Dwight et al.: Predicting gene function from patterns of annotation. Genome research, 13(5): 896-904, 2003.
10. Y. Kalfoglou, M. Schorlemmer: Ontology mapping: The state of the art. The Knowledge Engineering Review Journal, 18(1): 1-31, 2003.
11. A. Kumar, B. Smith, C. Borgelt: Dependence relationships between Gene Ontology terms based on TIGR gene product annotations. Proc. 3rd Intl. Workshop on Computational Terminology (CompuTerm), 2004.
12. P. Mork, P. Bernstein: Adapting a generic match algorithm to align ontologies of human anatomy. Proc. 20th Intl. Conf. on Data Engineering (ICDE), 2004.
13. A. Maedche, S. Staab: Measuring similarity between ontologies. Proc. 13th Conf. on Knowledge Engineering and Management, 2002.
14. S. Myhre, H. Tveit, T. Mollestad, A. Laegreid: Additional Gene Ontology structure for improved biological reasoning. Bioinformatics. 22(16): 2020-2037, 2006.
15. N. Noy, M. Musen: The PROMPT suite: Interactive tools for ontology merging and mapping. Intl. Journal of Human-Computer Studies, 59(6): 983-1024, 2003.
16. P. Ogren, K. Cohen, G. Acquaah-Mensah et al.: The compositional structure of Gene Ontology terms. Proc. Pacific Symposium on Biocomputing, 2004.
17. Online Mendelian Inheritance in Man, OMIM. McKusick-Nathans Institute for Genetic Medicine, Johns Hopkins University (Baltimore) and National Center for Biotechnology Information, National Library of Medicine (Bethesda), 2000.
18. E. Rahm, P. Bernstein: A survey of approaches to automatic schema matching. The VLDB Journal, 10(4): 334-350, 2001.
19. C. J. van Rijsbergen: Information retrieval. Butterworths, London, 2nd edition, 1979.
20. P. Shvaiko, J. Euzenat: A survey of schema-based matching approaches. Journal on Data Semantics, LNCS 3720 (JoDS IV): 928-943, 2005.
21. A. Thor, T. Kirsten, E. Rahm: Instance-based matching of hierarchical ontologies. Proc. 12th German Database Conf. (BTW), 2007.