Proceedings of the 2006 Winter Simulation Conference
L. F. Perrone, F. P. Wieland, J. Liu, B. G. Lawson, D. M. Nicol, and R. M. Fujimoto, eds.

A CORPUS-DRIVEN APPROACH FOR DESIGN, EVOLUTION AND ALIGNMENT OF ONTOLOGIES

Thomas Wächter
André Wobst
Michael Schroeder

Biotechnologisches Zentrum
TU Dresden
01062 Dresden, GERMANY

He Tan
Patrick Lambrix

Linköpings universitet
S-581 83 Linköping, SWEDEN

ABSTRACT

Bio-ontologies are hierarchical vocabularies, which are used to annotate other data sources such as sequence and structure databases. With the wide use of ontologies their integration, design, and evolution becomes an important problem. We show how textmining on relevant text corpora can be used to identify matching ontology terms of two separate ontologies and to propose new ontology terms for a given term. We evaluate these approaches on the GeneOntology.

1 INTRODUCTION

Currently, much research is devoted to the design of bio-ontologies, which are used to annotate biomedical data. Intuitively, ontologies can be seen as defining the basic terms and relations of a domain of interest, as well as the rules for combining these terms and relations (Neches et al. 1991). Ontologies are used in many areas, including bioinformatics and systems biology (Lambrix 2004, Lambrix et al. 2006). They are considered to be an important technology for the Semantic Web (e.g., Lambrix 2005; REWERSE). They are used for communication between people and organizations by providing a common terminology over a domain. They provide the basis for interoperability between systems. They can be used for making the content in information sources explicit and serve as an index to a repository of information. Further, they can be used as a basis for integration of information sources and as a query model for information sources. They also support clearly separating domain knowledge from application-based knowledge as well as validation of data sources. The benefits of using ontologies (e.g., Stevens et al. 2000, Lacy and Gerber 2004) include reuse, sharing and portability of knowledge across platforms, and improved maintainability, documentation, maintenance, and reliability. Overall, ontologies lead to a better understanding of a field and to more effective and efficient handling of information in that field.

A prominent example is the Gene Ontology (GO) (The Gene Ontology Consortium 2000), which comprises some 20.000 terms related to cellular components, biological process, and molecular function. The terms of the GO are then used to annotate protein sequences and structures in databases such as UniProt and PDB. While GO covers molecular biology, MeSH, the medical subject headings, focuses more on medicine and includes, e.g., diseases and chemical compounds. Both MeSH and GO overlap. Other ontologies include TAMBIS, GALEN, SNOMED, UMLS, FSTA (food science), FMA (human anatomy) and the collection of ontologies part of the open biomedical ontologies OBO.

In systems biology ontologies are currently being developed in connection to the development of standards for the representation of molecular interaction data. These standards (see, e.g., overviews in (Strömbäck and Lambrix 2005; Strömbäck et al. 2006a,b)) aim to provide the ability to supply information on molecular pathways in a format that supports efficient exchange and integration. This is seen as an important prerequisite for advances in the area. Several standards are being proposed, and use or develop ontologies for the definition of the important terms in the area. For instance, the Systems Biology Ontology, connected to the Systems Biology Markup Language (SBML, Hucka et al. 2003), defines terms used in quantitative biochemistry in four controlled vocabularies: roles of reaction participants, quantitative parameters, rate laws, and simulation frameworks. The Protein-protein interaction ontology, connected to the Proteomics Standards Initiative - Molecular Interaction (Hermjakob et al. 2004, Orchard et al. 2005), defines terms related to protein-protein interactions such as interaction detection methods, experimental roles and biological roles. The Systems Biology Ontology and the Protein-protein interaction ontology are available via OBO. The Biological Pathway Exchange (BioPAX) standard aims to provide an OWL-based data exchange format for pathway data and is developed as an ontology.

1-4244-0501-7/06/$20.00 ©2006 IEEE

Ontologies may overlap, so that an integration of two ontologies requires the identification of the overlap. This is a difficult problem, as ontology terms may be known under synonyms in the two ontologies and as a term of the same name may refer to different meanings in two ontologies. Thus, alignments based on the names of the terms and the structure of the ontology may not be sufficient. A second problem is the design and evolution of an ontology. The design process varies widely. While GO just requires informally that all paths from a term to the root must be "consistent", SNOMED formally defines all concepts using a simple description logic. No matter whether the ontologies are defined formally or informally, it is still open whether users intuitively understand term names in the ontology. In this article, we propose to approach ontology alignment, design and evolution by analysing a suitable text corpus. Following Wittgenstein's view that the meaning of a word is its use in language, we want to align and extend ontology terms by examining their use in literature. As literature source we use PubMed (PubMed Central), the biomedical literature database, which contains 16.000.000 paper abstracts. For the alignment of ontologies we test two methods. First, a Bayes classifier is generated which associates ontology terms to documents from a corpus. Two terms are aligned if the corresponding classifiers share sufficient documents. Second, PubMed is queried for co-occurrences of two terms from the two ontologies. Regarding the design and evolution of ontologies, we also pursue two approaches. First, we identify common prefixes of existing ontology terms, as often child terms (such as early endosome) are extensions of their parents (endosome). In a second approach, we identify word groups which appear significantly often for a given term. For example, plasma membrane appears frequently with endosome, so it is a good candidate term.

Thus, we test overall four approaches to align and extend ontologies by analysing a suitable text corpus. We first present the two approaches for alignment and then the two for ontology extension. We use test cases based on GO, as this is one of the most mature efforts in the field and has reached the status of de facto standard.

2 ALIGNMENT

Many ontologies have already been developed and many of these ontologies contain overlapping information. Often we would therefore want to be able to use multiple ontologies. For instance, companies may want to use community standard ontologies and use them together with company-specific ontologies. Applications may need to use ontologies from different areas or from different views on one area. Ontology builders may want to use already existing ontologies as the basis for the creation of new ontologies by extending the existing ontologies or by combining knowledge from different smaller ontologies. In each of these cases it is important to know the relationships between the terms (concepts and relations) in the different ontologies. It has been realized that ontology alignment, i.e., finding relationships between terms in the different ontologies, is a major issue and some organisations (e.g., the organisation for Standards and Ontologies for Functional Genomics) have started to deal with it.

We present two approaches for finding inter-ontology relationships based on life science literature from PubMed. In the first approach we define a similarity measure between concepts based on the PubMed literature related to the concepts. Concepts that are closely related according to this similarity measure are candidates for aligning. In the second approach we cluster the concepts of two ontologies according to a distance measure based on PubMed document counts. Concepts from different ontologies in the same cluster are candidates for alignment. We discuss the feasibility of these approaches on parts of GO and Signal-Ontology (SigO) (Takai-Igarashi et al. 1998) as well as GO and the Enzyme Nomenclature (EC).

2.1 Text Classification Approach

We can define a similarity measure between concepts in different ontologies based on the probability that documents about one concept are also about the other concept and vice versa. We implemented a strategy containing the following basic steps (Lambrix and Tan 2006). (i) For each ontology that we want to align we generate a corpus of PubMed abstracts by using the concepts as query terms. In our implementation we generated a corpus of maximally 100 PubMed abstracts per concept. (ii) For each ontology a document classifier is generated. This classifier returns for a given document the concept that is most closely related to the document. To generate a classifier the corpus of abstracts associated to the classifier's ontology is used. In our algorithm we use a naive Bayes classification algorithm. (iii) Documents of one ontology are classified by the document classifier of the other ontology and vice versa. (iv) A similarity measure between concepts in the different ontologies is computed by using the results of step (iii). The similarity is computed as

$$sim(C_1,C_2) = \frac{n_{NBC2}(C_1,C_2)+n_{NBC1}(C_2,C_1)}{n_D(C_1)+n_D(C_2)}$$

where $n_D(C)$ is the number of documents originally associated with C, and $n_{NBCx}(C_p,C_q)$ is the number of documents associated with C_p that are also related to C_q as found by classifier $NBCx$ related to ontology x. (v) Pairs of concepts with a similarity measure higher than or equal to a given threshold are suggested as candidates for aligning.

More details about this algorithm as well as some extensions can be found in (Tan et al. 2006).

2.2 Cluster Approach

This method uses the normalized information distance (Li et al. 2003) between two concepts. It is based on counting the number of hits when querying PubMed using the concepts in the ontologies. The algorithm contains the following basic steps. (i) For each concept the number of documents retrieved from PubMed when using the concept as a query term is retrieved (by querying PubMed). For each pair of concepts the number of documents retrieved from PubMed when using the conjunction of the concepts as a query term is retrieved. (ii) A distance measure between the concepts is computed by using the results of step (i). The distance is computed as

$$NPD(C_1,C_2) = \frac{max(logf(C_1),logf(C_2)) - logf(C_1,C_2))}{logM - min(logf(C_1),logf(C_2))}$$

where M is the total number of documents in PubMed (in our case 16.000.000), $f(C)$ is the number of documents retrieved from PubMed when using C as a query term, and $f(C_1,C_2)$ is the number of documents retrieved from PubMed when using C_1 *and* C_2 as query term. (This distance measure is similar to the normalized google distance in Cilibrasi and Vitanyi 2004.) (iii) The concepts are clustered based on their distance and a given threshold using complete-link hierarchical clustering. The maximal distance between elements in the same cluster is lower than or equal to the threshold. We used an implementation available from LingPipe (`<http://www.alias-i.com/lingpipe/>`). (iv) Concepts from different ontologies in the same cluster are suggested as candidates for alignment.

2.3 Evaluation

Test cases: Two test cases are based on parts of GO and SigO. The first case, *B* (behavior), contains 57 terms from GO and 10 terms from SigO. The second case, *ID* (immune defense), contains 73 terms from GO and 17 terms from SigO. Domain experts were asked to analyse the cases and provide alignment relationships based on equivalence and is-a relations. We used the ontologies and the alignment relationships from the experts as they were provided to us. For B this resulted in 4 expected alignments (*EA* in Tables 1 and 2) and 8 for ID. The other test cases are based on GO and EC. The case *EC1.3* contains 141 terms from GO and 129 terms from EC. The case *EC1.1.3* contains 29 terms from GO and 31 terms from EC. We used the ec2go mapping (2006/03/27) available on the GO homepage as the expected result set (115 expected alignments for EC1.3 and 27 for EC1.1.3).

Table 1: Quality of the Suggestions – Classification; ts/cs: Total Number of Suggestions (ts), Number of Correct Suggestions (cs).

Case	B	ID	EC1.3	EC1.1.3
Th/EA	4	8	115	27
0.4	4/2	9/6	23/15	8/6
0.5	2/2	5/5	18/13	7/5
0.6	2/2	2/2	13/11	4/4
0.7	2/2	1/1	9/8	4/4
0.8	1/1	0/0	7/6	4/4

Table 2: Quality of the Suggestions – Cluster; tc/cc/cs: Total Number of Clusters (tc), Number of Clusters Containing Correct Suggestions (cc), Number of Correct Alignment Suggestions (cs).

Case	B	ID	EC1.3	EC1.1.3
Th/EA	4	8	115	27
0.6	3/3/4	5/3/5	35/28/32	11/9/10
0.5	4/2/3	5/3/4	36/28/32	12/10/10
0.4	4/2/2	5/3/4	36/29/30	12/10/10
0.3	2/2/2	4/3/4	37/29/30	11/9/9
0.2	2/2/2	3/3/3	26/21/21	6/6/6

Regarding the querying of PubMed, we use the concept name as query term for GO and SigO concepts and search in the titles and abstracts of PubMed documents. For EC concepts we use the EC number as query term and search in the EC number field for PubMed documents.

Text classification: Table 1 presents the results (ts/cs) of the classification approach. Terms in different ontologies with a similarity value higher than or equal to the threshold *Th* are suggested alignment candidates (ts). For instance, for ID and threshold 0.4 the approach suggests 9 alignment candidates of which 6 (cs) are correct. The other 3 may be wrong or redundant.

Cluster: Table 2 presents the results (tc/cc/cs) of the cluster-based approach. The table presents for each case and threshold the number of clusters with elements from both ontologies (tc), the number of clusters containing both terms in at least one expected alignment (cc) and the number of correct alignments (cs) found using this method. For instance, for ID and threshold 0.6 the approach creates 5 clusters with terms from both ontologies, 3 of these contain terms in correct alignments and in total 5 correct alignments can be identified in these 3 clusters.

2.4 Discussion

The quality of the suggestions for both approaches varies in the different cases in this evaluation. For instance, when the threshold is set to 0.4 the recall of the classification approach is 0.5 for the case B but only 0.13 for the case EC1.3. Similarly, for the threshold 0.6 the recall for the cluster approach is 1 for B but only 0.27 for EC1.3. The precision for the classification approach with threshold 0.4 is, however, higher for EC1.3 (0.65) than for B (0.5). All clusters with terms from both ontologies in the cluster approach with threshold 0.6 are relevant for B, but only 80% of the clusters are relevant for EC1.3.

For all cases, the recall of the classification approach goes down when the threshold becomes higher, e.g., in the EC1.3 case the recall goes down from 0.13 to 0.05 when the threshold goes up from 0.4 to 0.8. The recall for the cluster approach goes down when the threshold becomes lower, but the change is not as large as in the classification approach.

The quality of the suggestions depends heavily on the related literature in PubMed. In the case B there are no related documents for 10% of the concepts and less than 10 related documents for an additional 4% of the concepts. In the case EC1.3 there are no related documents for 40% of the concepts and less than 10 related documents for an additional 26% of the concepts.

Although the approaches find correct alignments, they also may give wrong results, even when many documents are available. For instance, in the case EC1.3 there are 699 documents related to EC:1.3.99.3, 523 documents related to GO:acyl-CoA dehydrogenase activity and among these documents 123 contain both concepts. This is a correct alignment. However, the classification approach assigns a low similarity value (0.335) and in the cluster approach they do not appear in the same cluster for the tested thresholds.

In most cases of this evaluation, the cluster approach outperforms the classification approach and the alignments found by the classification approach are also found by the cluster approach. One factor could be that the computation of the distance in the cluster approach is normalized. For instance, in the case EC1.3 there are 593 documents related to EC:1.3.1.2, two related documents for GO:dihydropyrimidine dehydrogenase (NADP+) activity and these two documents contain both concepts. In the cluster approach the distance is computed as 0.358 which means that the concepts are relatively similar. In the classification approach the similarity between the two concepts using the formula in Section 2.1 can be at most 0.0067.

3 ONTOLOGY LEARNING

Learning from text corpora is based on methods which try to extend ontologies by applying natural language processing techniques to text (Gomez-Perez and Manzano-Macho 2003). Early publications focus on pattern-based concept and relation extraction, where a concept or relation will be added to the ontology if it is found to match a predefined pattern (Morin 1999). As in classical shopping cart analysis, association rules can also be used for corpus-based learning of ontologies (Maedche and Staab 2000). Association rules evaluate the co-occurrence of items within an item set and use the likelihood of an item *A* being member of a set, if *B* is already a member. A different technique called conceptual clustering was proposed in (Faure and Poibeau 2000). After the acquisition of syntactic frames in a text, the learning method relies on the observation of syntactic regularities in the context of words. Concepts found are grouped according to their semantic distance and become this way ordered in a hierarchy. For this, no annotation is needed beforehand, but the validation of the result is performed manually and is therefore time-consuming. A pattern-based learning approach instead would use labeled examples for extracting instances from texts. While the annotation of the learning examples is time-consuming, the quality of the learning results would be predictable and could be validated automatically.

Ontology learning is the automatization of the ontology building process with the aim to lower development costs and shorten the development time. For automatic learning, information sources like the ontology itself or a text corpus of relevant documents are needed. We will illustrate our work on the example of the GO and PubMed abstracts.

We discuss two approaches for the automatic prediction of candidate terms, namely the *superstring prediction* and the *term co-occurence analysis*. Both approaches require to make a selection on PubMed abstracts for a given GO term. We identify subsets of documents using the textmining capabilities of the GoPubMed project (Doms and Schroeder 2005). We only regard perfect matches for terms.

3.1 Superstring Prediction

In (Ogren et al. 2004) the compositional structure of GO terms was analyzed. The authors found that many GO terms contain each other and many GO terms are derived from each other. For example, the term *membrane* [GO:0016020] has *inner membrane* [GO:0019866] as a direct subconcept. This knowledge can be used to automatically generate new candidate terms following the observed patterns. We analyzed whether these superstring relations observed in GO can be verified in the text.

* Methodology: By analyzing the GO we identified 3129 out of 20223 terms, where the term is a superstring of its children. Further for 1781 of these terms, they and their children were found in the documents. For each term, we used the PubMed abstracts giving this evidence for our analysis (see also Table 3). Based on these texts we

Table 3: Analysis of Gene Ontology Annotations for PubMed Abstract with Respect to Parent–Child Relationships.

Terms in GO	20223
Terms found in abstracts	14905
Terms having children containing themselves	3129
(parent found in text)	2692
(parent and one child found in text)	2239
(parent and all children found in text)	**1781**
Terms having children	7451
(parent found in abstracts)	5964
(parent and one child found in abstracts)	5185
(parent and all children found in abstracts)	**3757**

identify the words which precede the actual term and rank them by their frequency of occurrence. This leads to a list of newly identified candidate terms to be possibly included in the ontology. Below we carry out this analysis for some example terms and highlight how many of the predicted terms are indeed children in GO.

* Example: **GO:0005773 ‘vacuole’** A vacuole is defined as a closed structure, found only in eukaryotic cells, that is completely surrounded by unit membrane and contains liquid material. The term has the children ‘autophagic’, ‘contractile’, ‘lytic’, ‘parasitophorous’ and ‘storage vacuole’. All are found in the first 50 predicted terms (see Table 4).

* Example: **GO:0005096 ‘GTPase activator activity’** GTPases are molecular switches. A GTPase activator is an enzyme that catalyzes the hydrolysis of GTP. GTPase activator activity has the children ‘ARF’, ‘Rab’, ‘Rac’, ‘Ral’, ‘Ran’, ‘Rap’, ‘Ras’, ‘Rho’ and ‘Sar GTPase activator activity’. Five of the children can be automatically found (Table 5).

* Example: **GO:0016265 ‘Death’** This term has the children ‘aging’, ‘tissue death’ and ‘cell death’. Out of these three terms the superstring prediction method is only capable to find ‘tissue death’ and ‘cell death’ (Table 6). While ‘cell death’ is found first, ‘tissue death’ is not found within the first 50 predicted terms. Nevertheless by carefully investigating the result list one will find, that many terms are from the medical domain rather than molecular biology. Terms like ‘cardiac death’, ‘neuronal death’, ‘infant death’, ‘fetal death’, ‘brain death’ and ‘neonatal death’ make perfectly sense for a medical ontology. Predicted prefixing words like ‘sudden’, ‘early’ and ‘late’ can easily be filtered using knowledge about their frequency of occurrence in the English language.

Table 4: Example – Vacuole.

pos.	count	candidate	GO
1	autophagic	1219	**child**
2	cytoplasmic	1048	unknown
3	parasitophorous	933	**child**
4	large	684	unknown
5	food	496	unknown
6	contractile	387	**child**
7	phagocytic	383	unknown
8	rimmed	383	unknown
9	lipid	378	unknown
10	intracellular	303	unknown
11	intracytoplasmic	295	unknown
12	digestive	265	descendant
13	endocytic	260	unknown
14	small	247	unknown
15	membrane-bound	240	unknown
..	..	...	...
20	storage	175	**child**
..	..	...	...
44	lytic	36	**child**
..	..	...	...

3.2 Term Co-occurrence Analysis

The second hypothesis we wanted to verify, is based on the co-occurrence of ontology terms in scientific text.

* Methodology: By analyzing the GO we identified 7451 out of 20223 terms, which have children. Furthermore 3757 of these terms and their children are found in documents. Again, we used the PubMed abstracts giving this evidence for our analysis (see also Table 3) for each term.

Additionally to the questions in the previous section, for the co-occurrence we are now interested in the predicted terms which are themselves terms in the GO.

* Example: **GO:0005768 ‘endosome’** This term has the children ‘early endosome’, ‘endosome lumen’, ‘endosome membrane’ and ‘late endosome’ (Table 7). Further terms deeper in the hierarchy are mainly combinations of the named terms. On the one side all children of ‘endosome’ apart from ‘endosome lumen’ where predicted. On the other side predicted terms like ‘recycling endosome’, ‘transferrin receptor’, ‘epithelial cell’, ‘trans-golgi network’, or ‘receptor-mediated’ endocytosis are valuable candidates to be included in the GO.

* Example: **GO:0001739 ‘sex chromatin’** This GO term only has two children: ‘Barr body’ and ‘XY body’. Both child terms are found among the top 15 (Table 8). Many of the other predictions such as DNA content, cell nuclei, and electron microscopy are meaningful terms, too.

Table 5: Example – GTPase Activator Activity.

pos.	candidate	count	GO
1	ras	133	**child**
2	rho	106	**child**
3	small	100	similar term
4	intrinsic	88	unknown
5	gap	37	synonym
6	p21ras	34	unknown
7	family	29	unknown
8	arf	23	**child**
9	triphosphatase	19	similar term
10	rac	17	**child**
11	p21	16	unknown
12	rab	12	**child**
..	..	...	...

Table 6: Example – Death.

pos.	candidate	count	GO
1	cell	60678	**child**
2	sudden	11521	unknown
3	cardiac	7179	unknown
4	neuronal	5326	unknown
5	infant	3925	unknown
6	fetal	3636	unknown
7	brain	3468	unknown
8	early	2658	unknown
9	late	2079	unknown
10	neonatal	2038	unknown
..	..	...	...

Table 7: Example – Endosome.

pos.	candidate	count	GO
1	early endosome	675	**child**
2	plasma membrane	487	existing term
3	late endosome	361	**child**
4	cell surface	327	existing term
5	degrees c	205	(can be filtered)
6	endosome fusion	200	existing term
7	endocytic pathway	190	unknown
8	*recycling endosome*	189	unknown
9	cell line	145	unknown
10	*transferrin receptor*	144	unknown
11	membrane protein	137	unknown
12	endosomal compartment	136	unknown
13	growth factor	130	unknown
14	results suggest	125	(can be filtered)
15	*epithelial cell*	121	unknown
16	endosomal membrane	116	unknown
17	*trans-golgi network*	114	unknown
18	*endocytic vesicle*	109	unknown
19	*receptor-mediated endocytosis*	98	unknown
..	..	...	...
29	endosome membrane	75	**child**
..	..	...	...

Table 8: Example – Sex Chromatin.

pos.	count	word	explanation
1	257	sex chromatin	existing term
2	59	sex chromosome	existing term
3	35	chromatin body	not a term
4	32	barr body	**child**
5	18	dna content	not a term
6	18	meiotic prophase	existing term
7	18	male female	(can be filtered)
8	18	cell nuclei	not a term
9	16	sex body	not a term
10	15	male sex	existing term
11	14	klinefelter syndrome	not a term
12	14	sex determination	existing term
13	13	electron microscopy	not a term
14	13	xy body	**child**
15	13	bone marrow	existing term
..	..	...	...

4 CONCLUSION

We have presented four approaches which use literature to define the semantics of terms and hence align terms (classification and cluster approach) and predict new child terms for a given term (superstring prediction and co-occurrence analysis).

The evaluation results suggest that using life science literature is a possible approach for aligning and extending ontologies, although there still is much room for improvement of the current algorithms. In (Lambrix and Tan 2006) the classification approach for alignment was compared with other approaches. In one case the classification approach performed best, but it was outperformed by other approaches in other cases. It was also shown that the classification approach could be combined with other approaches to obtain superior results. The cluster approach can also be used in another way. Instead of using the clusters to suggest alignments, the clusters could be used to filter alignment suggestions proposed by other algorithms. Similar ideas were explored and found useful using a structure-based cluster approach in (Chen et al. 2006). Regarding the ontology extension methods, results can be improved by filtering common words. For all approaches the choice of the text corpus and appropriate natural language processing is vital for good results.

ACKNOWLEDGMENTS

We acknowledge the financial support of the Swedish Research Council (Vetenskapsrådet), the Center for Industrial Information Technology (CENIIT), and the EU projects Sealife (FP6-IST-027269) and REWERSE (FP6-IST-506779).

REFERENCES

BioPAX, Biological Pathway Exchange. `<http://www.biopax.org/>`.

Chen, B., H. Tan, and P. Lambrix. 2006. Structure-based filtering for ontology alignment. In *Proceedings of the IEEE WETICE Workshop on Semantic Technologies in Collaborative Applications*.

Cilibrasi, R., and P. Vitanyi. 2004. Automatic meaning discovery using Google. `<http://xxx.lanl.gov/abs/cs.CL/0412098>`.

Doms, A., and M. Schroeder. 2005. GoPubMed: exploring PubMed with the gene ontology. *Nucleic Acids Research* 33:W783–W786.

Enzyme Nomenclature. Nomenclature Committee of the International Union of Biochemistry and Molecular Biology (NC-IUBMB). `<http://www.chem.qmul.ac.uk/iubmb/enzyme/>`.

Faure, D., and T. Poibeau. 2000. First experiments of using semantic knowledge learned by ASIUM for information extraction task using INTEX. In *Proceedings of the ECAI Workshop on Ontology Learning*.

The Gene Ontology Consortium. 2000. Gene ontology: tool for the unification of biology. *Nature Genetics* 25(1):25-29. `<http://www.geneontology.org/>`.

Gomez-Perez, A., and D. Manzano-Macho, editors. 2003. *A survey of ontology learning methods and techniques*. OntoWeb Deliverable 1.5.

Hermjakob, H., L. Montecchi-Palazzi, G. Bader, J. Wojcik, L. Salwinski, A. Ceol, et al. 2004. The HUPO PSI's Molecular Interaction format - a community standard for the representation of protein interaction data. *Nature Biotechnology* 22(2):177-183.

Hucka, M., A. Finney, H. Sauro, H. Bolouri, J. Doyle, H. Kitano, and the rest of the SBML Forum. 2003. The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models. *Bioinformatics* 19(4):524-531.

Lacy, L., and W. Gerber. 2004. Potential modeling and simulation applications of the web ontology language OWL. In *Proceedings of the Winter Simulation Conference*, 265-270.

Lambrix, P. 2004. Ontologies in bioinformatics and systems biology. In *Artificial Intelligence Methods and Tools for Systems Biology*, ed. Dubitzky and Azuaje,129-146. Springer.

Lambrix, P. 2005. Towards a semantic web for bioinformatics using ontology-based annotation. In *Proceedings of the 14th IEEE International Workshops on Enabling Technologies: Infrastructures for Collaborative Enterprises*, 3-7. Invited talk.

Lambrix, P., and H. Tan. 2006. SAMBO – a system for aligning and merging bio-ontologies. *Journal of Web Semantics, Special issue on Semantic Web for the Life Sciences* 4(3).

Lambrix, P., H. Tan, V. Jakonienė, and L. Strömbäck. 2006. Biological ontologies. In *Semantic Web: Revolutionizing Knowledge Discovery in the Life Sciences*, ed. Baker and Cheung, to appear. Springer.

Li, M., X. Chen, X. Li, B. Ma, and B. Vitanyi. 2003. The similarity metric. In *Proceedings of the 14th Annual ACM-SIAM Symposium on Discrete Algorithms*, 863–872.

Maedche, A., and S. Staab. 2000. Discovering conceptual relations from text. In *Proceedings of the 14th European Conference on Artificial Intelligence*, 21–25.

Morin, E. 1999. Automatic acquisition of semantic relations between terms from technical corpora. In *Proceedings of the Fifth International Congress on Terminology and Knowledge Engineering*, 268–278.

Neches, R., R. Fikes, T. Finin, T. Gruber, T. Senator, and W. Swartout. 1991. Enabling technology for knowledge engineering. *AI Magazine* 12(3):26-56.

Ogren, P. V., K. B. Cohen, G. K. Acquaah-Mensah, J. Eberlein, and L. Hunter. 2004. The compositional structure of gene ontology terms. In *Proceedings of the Pacific Symposium on Biocomputing*, 9:214–225.

Orchard, S., L. Montecchi-Palazzi, H. Hermjakob, and R. Apweiler. 2005. The use of common ontologies and controlled vocabularies to enable data exchange and deposition for complex proteomic experiments. In *Proceedings of the Pacific Symposium on Biocomputing*, 10:186-196.

PubMed Central. `<http://www.pubmedcentral.nih.gov/>`.

REWERSE. EU Network of Excellence on Reasoning on the Web with Rules and Semantics, Working group A2. `<http://rewerse.net/>`.

SBML. Systems Biology Markup Language. `<http://sbml.org>`.

Strömbäck, L., D. Hall, and P. Lambrix. 2006a. A review of standards for data exchange within systems biology. *Proteomics*. Invited contribution, to appear.

Strömbäck, L., V. Jakonienė, H. Tan, and P. Lambrix. 2006b. Representing, storing and accessing molecular interaction data: a review of models and tools. *Briefings in Bioinformatics*. Invited contribution, to appear.

Strömbäck, L., and P. Lambrix. 2005 Representations of molecular pathways: an evaluation of SBML, PSI MI and BioPAX, *Bioinformatics* 21(24):4401-4407.

Tan, H., V. Jakonienė, P. Lambrix, J. Aberg, and N. Shahmehri. 2006. Alignment of biomedical ontologies using life science literature. In *Proceedings of the International Workshop on Knowledge Discovery in Life Science Literature*, LNBI 3886, 1–17.

Takai-Igarashi, T., Y. Nadaoka, and T. Kaminuma. 1998. A database for cell signaling networks. *Journal of Computational Biology* 5(4):747-754.

AUTHOR BIOGRAPHIES

THOMAS WÄCHTER is a Ph.D. student at Dresden University of Technology. His current main research interests are ontology evolution and text mining. His e-mail address is `<thomas.waechter@biotec.tu-dresden.de>`.

HE TAN is a Ph.D. student at Linköpings universitet. Her current main research interests are in Semantic Web and ontologies for the Life Sciences. Her e-mail address is `<hetan@ida.liu.se>`.

ANDRÉ WOBST is a Diploma student at Dresden University of Technology. His current main research interests are statistical methods for ontology learning and natural language processing. His e-mail address is `<andre.wobst@biotec.tu-dresden.de>`.

PATRICK LAMBRIX is an associate professor of computer science at Linköpings universitet. His current main research interests are in Semantic Web, ontologies and databases for the Life Sciences. His e-mail address is `<patla@ida.liu.se>`.

MICHAEL SCHROEDER is a professor for Bioinformatics at Dresden University of Technology. His current main research interests are protein interactions, textmining, and ontologies. His e-mail address is `<ms@biotec.tu-dresden.de>`.