On Semantic Similarity and Relatedness for Knowledge-Driven Discovery in Biomedical Data

Daniela Rosu

Abstract

A great variety of tasks, from word sense disambiguation and document retrieval to assessing the functional similarity of gene products and validating protein-protein interaction networks, depend on the ability to measure the semantic similarity between concepts organized in ontologies. This report is a comprehensive study of classic and recent computational methods measuring semantic relatedness. Motivational arguments set the stage for a survey of the methods, applications and a critical assessment of the methods and of the various evaluation strategies adopted in the literature. Proposals for future directions in the areas of measuring semantic similarity and relatedness, as well as suggestions for improvement, curently under investigation, are offered.

Contents

Glossary

1 Introduction

As various hight-troughput technologies mature and become more cost effective, the major challenge in bioinformatics is no longer how to generate vast quantities of genomic data, but rather how to best collect, manage, and analyze the data. I will attempt to explain in this section how ontologies as "formal, explicit specification[s] of a shared conceptualization"[47], and associated reasoning techniques, in particular the concept of *similarity*, apply to the current biological data analysis methods as well as to the emerging discovery systems. I will then cover existing techniques for measuring concept similarity as well as evaluation strategies for these techniques (section2), a side by side comparison (section 3). In the last section I will discuss several issues important to both improving the measurement techniques and assessing the quality of similarity measures as well as my current research and directions for future research.

The most basic reason bio-ontologies have been receiving an increased amount of attention is their potential to help solve the semantic mismatch, which is a major impediment that data analysis strategies must overcome even for very simple research scenarios. It is widely acknowledged that heterogeneity is inherent in biological data, but there is perhaps less awareness of its extent and pervasiveness.

Heterogeneity occurs not only in the schemas used to store data, but also in the actual data values themselves. For example, comparisons between microarray data are difficult not only because of the biological, technical, and analytical diferences between studies but also because the results may be reported in different gene nomenclatures such as those used by Genbank[1], Entrez Gene[2], EMBL Nucleotide Sequence Database[3], Unigene[4], Affymetrix, etc. The use of ambiguous terms, is another wide spread and difficult to resolve issue. A prominent example is the concept of *gene*. For the Human Genome Database[5], a gene is a "DNA fragment that

[1] Available at: http://www.ncbi.nlm.nih.gov/Genbank/

[2] Available at: http://www.ncbi.nlm.nih.gov/sites/entrez?db=gene

[3] Available at: http://www.ebi.ac.uk/embl/

[4] Available at: http://www.ncbi.nlm.nih.gov/sites/entrez?db=unigene

[5] Available at: http://www.hugo-international.org/

can be transcribed and translated into a protein" but for Genbank[6] a gene is a "DNA region of biological interest with a name and that carries a genetic trait or phenotype". Since the second definition includes nonstructural coding DNA regions like introns, promoters and enhancers, there is a clear semantic distinction between those two notions of *gene* but both continue to be used by different communities. Another commonly used term with multiple meanings is *protein function*. Depending on the context, *function* can refer to a *biochemical function*, e.g. enzyme catalysis, a *genetic function*, e.g. transcription repressor, a *physiological function*, e.g. signal transducer, etc.

The fundamental reason that makes resolving semantic heterogeneity so difficult is that the data sets are developed independently, and therefore varying structures and naming strategies are used to represent the same or overlapping concepts. In many cases the data systems to be integrated were developed for very different business and research needs. Hence, even if they model overlapping domains, they may model them in distinct ways as different agents/actors have varying conceptualizations of their domain of interest. Semantic standardization would impose a certain view of a domain, but in many situations this is not feasible because the domain is changing very fast and/or competing players cannot agree due to costs or diverging interests or views of the domain.

Although ontologies, as computer readable formulations of concepts and relationships among them, are hailed as the potential solution to the semantic interoperability problem, there is no clear consensus on what an ontology really is and depending on the context, an ontology can refer to, for example, a , a with an informal representation, typically consisting of *is-a* relationships, a *conceptual model of a domain*, including rules to infer new knowledge.

From the many ontology formalisms we adopt here the definition given in the Karsrue Ontology Model[34]. This framework does not handle constraints and axioms but it is its simplicity that makes it better suited to bio-ontologies, most of which have very informal representations.

Definition 1.1 *An ontology with datatypes is a structure* $O := (C, T, R, A, I, V, \ \sigma_R, \sigma_A, \leq_C$

[6] Available at: http://www.ncbi.nlm.nih.gov/Genbank/

$, \leq_R, \leq_A, i_C, i_T, i_R, i_A)$ *consisting of*

- *6 disjoint sets C, T, R, A, I and V called* concepts, datatypes, relations, attirbutes, instances *and* datavalues
- *partial orders* concept hierarchy, $\leq_C$ *on C, and* type hierarchy, $\leq_T$ *on T*
- *functions* relation signature, $\sigma_R : R \to C \times C$, *and* attribute signature, $\sigma_A : A \to C \times T$
- *partial orders* $\leq_R$ *on R and* $\leq_A$ *on A*
- *instantiation functions* $i_C : C \to 2^I$, $i_T : T \to 2^V$, $i_R : R \to 2^{I \times I}$, $i_A : A \to 2^{I \times V}$

In this formalism the hierarchical relations $\leq_.$ represent the *is-a* relations from other formal and informal definitions.

I illustrate the ontology definition with an example which relates genes and pathways.

Example 1.1 *Let* $O_{pathway-example}$ *be the structure* $(C, T, R, A, I, V, \sigma_R, \sigma_A, \leq_C, \leq_R, \leq_A, i_C, i_T, i_R, i_A)$, *where*

- $C = \{$ *ROOT, PATHWAY, GENE*$\}$,
- $T = \{$*string*$\}$,
- $R = \{$*is_involved_in, changes_expression_level*$\}$,
- $V = \{$ *"signaling", "metabolic"*$\}$,
- $I = \{$*Insulin_Signaling_Pathway, Glycolysis_Pathway, INS, GAPDH*$\}$,
- $A = \{$*CATHEGORY*$\}$,
- $\leq_C = \{$*(ROOT, PATHWAY), (ROOT, GENE)*$\}$,
- $\leq_R = \{\}$,
- $\leq_A = \{\}$,

- i_C = {*(PATHWAY,*{*Insulin_Signaling_Pathway*}, *(PATHWAY,* {*Glycolysis_Pathway*}*),* *(GENE,* {*INS*}*), (GENE,* {*GAPDH*}*)* },

- i_A = {*(CATEGORY, (Insulin_Signaling_Pathway, ”signaling”)*}*), (CATEGORY,* {*(Glycolysis_Pathway, “metabolic”)*}*)*},

- i_T = {*(string, “signaling”, “metabolic”)*},

- i_R = {*(is_involved_in,* {*(GAPDH, Glycolysis_Pathway)*}*), (changes_expression_level ,*{*(Insulin_Signaling_P*
INS)}*)*}

Among the many ontologies developed for the biomedical domain, Systematized Nomenclature of Medicine - Clinical Terms(SNOMED-CT) and the Gene Ontology(GO)[26] are the most widely used. Several dozen others are maintained by the Open Biomedical Ontology[7] and many more are being developed independently by various research groups around the world. SNOMED-CT has more than 370 000 unique concepts, covering most areas of clinical information such as diseases and microorganisms, which related through semantic relations such as *is-a, treats, prevents, has ingredient*, etc. The concepts are organized into 13 hierarchies united by a root concept.

The Gene Ontology, developed by the Gene Ontology Consortium, is one of the most widely used systems for semantic annotation. Although it has been often criticized for inconsistencies and for not adhering to formal principles[112, 111], it is nevertheless aquiring the status of a standard ontology across various biological domains. The Gene Ontology is structured into three domain ontologies (*molecular function*(MF), *biological process*(BP) and *cellular component*(CC)) with the terms organized in a directed acyclic graph. The relationships between terms are of several types: *is-a, part-of, regulates, positively-regulates*, and *negatively-regulates.* According to the GO documentation, a biological process is “a recognized series of events or molecular functions”, but currently there are no associative relationships in GO indicating whether a molecular function is involved in a biological process. In May 2008 the Gene Ontology consortium announced that it will introduce *regulates* relationships whithin the Molecular Func-

[7] http://www.obofoundry.org/

tion ontology and between the MF and BP ontologies at the end of 2008. This decision comes as a recognition of the necessity to make explicit some relatedness relationships in addition to similarity (*is-a*) or compositionality (*part-of*) relationships. However, although relationships such as the one between *regulation of kinase activity*(BP) and *kinase activity*(MF) will be made explicit with the introduction of the new links, others, such as between *transcription*(BP) and *aryl hydrocarbon receptor binding*(MF), will not. In addition, other relatedness relationships, such as *localization*, for example between *nucleus*(CC) and *DNA binding*(MF), or between *chromosome*(CC) and *sister chromatid biocondensation*(BP) need to be discovered automatically. In some cases it is possible to detect by a simple lexical analysis the localization relationships between CC and MF terms such as *Golgi aparatus*(CC), *Golgi organization and biogenesis*(MF), or between CC and BP terms such as *vacuole*(CC) and *vacuolar protein processing*(BP), but most such relationships, such as between *nucleus*(CC) and *mRNA transcription*(BP), are not immediately evident and more sophisticated techniques are needed.

As amply illustrated by the literature, the importance of the biomedical ontologies, GO especially, goes beyond simply that of simple annotation vocabularies. They are central to a multitude of tasks, from predictive functional genomics to information retrieval and mediating between data sources in data integration engines.

Automatic Annotation of Gene Products

One of the most exciting applications of annotation terms is their use as a predictive instruments for tasks such as assigning functions or localization information to unannotated genes and proteins identified by genome sequencing and other methods. This is an area that has received significant attention in the past years, as many organisms have now been completely sequenced, but establishing the function(s) of various genes is lagging behind. For example, the *Arabidopsis thaliana* (thale cress) genome is completely sequenced, but functional annotation of the genes remains a key challenge as approximately 50% of the 28,000 genes have not been assigned any function.

Another frequent computational task is the analysis of high-throughput experimental data in order to identify genes which are differentially expressed between normal and pathological tissues. This analysis includes associating the significant genes with descriptors that may help explain the biological meaning of the experimental results. The process of finding/predicting the most relevant descriptors can take advantage of the annotations attached to the similar or related gene products in the medical/bio-chemical literature and/or various public and proprietary databases.

Predicting Gene Function

Typically, investigators use computational sequence analysis tools to assign functions to newly found gene products. To date, the most commonly used techniques are based on physical association, genetic interaction, sequence relationships, patterns of gene expression and *enrichment analysis*. Much of the work in enrichment analysis uses statistical methods[53, 115, 64, 17, 1, 120], primarily based on the frequency of terms associated with a list of genes, without taking into account the semantic relationships that may exist between the terms. Ignoring this information, however, may result in failure to identify the similar genes that are annotated with distinct but semantically similar or related terms.

In semantic similarity approaches the functional similarity between gene products is calculated by matching the functional domains that they contain, which addresses the main problem of sequence-based similarity, i.e., when the region of a gene product that is matched by a query sequence is not related to the function of that gene product.

The functional relationship is usually estimated by comparing the shared annotation of gene products. The annotation terms most often belong to a controlled vocabulary system, such as GO and several methods exist to assess the similarity of sets of such terms. However, simply identifying shared GO annotations may not be adequate for the estimation of semantic similarity as even if two annotations are different, they can be closely related via their common ancestors in the taxonomy. On the other hand, the shared terms may be too general to be used as evidence for the functional association of annotated gene products and the GO graph structure can be used to improve the sensitivity of semantic measures.

Evaluation of Domain-Domain and Protein-Protein interaction Networks

In addition to enabling the identification of functionally related gene products, similarity measures can also be used to predict and validate high-throughput protein interaction data. The prediction of protein-protein interactions is mainly based on the homology of protein sequences, but the experimental coverage of the interactomes for many organisms is still low and other methods are needed to help validate the posited interactions. In recent years, several techniques[75, 72, 103] have been proposed, with very promising results, for the ab initio prediction of protein-protein interactions and for assessing the quality of extant predictions. The initial evaluation studies all corroborated the conclusion that functional similarity based on the Gene Ontology annotations improve the accuracy of the interaction predictions.

Ontology-based Data Integration

As we approach the post-genomic era, it is estimated that the focus will move from "models-of-analysis" of the existing data, such as algorithms for functional gene clustering, to "models-of-process", which aim at explaining the relationship between genomic data and the biological pathways underlying physiologic processes. The next logical step, and ultimately the goal of genomic research, is relating these processes to clinical outcomes and achieving this goal will rely on methods that perform the semantic integration of various data sources from different levels of biology.

The development of ontologies is seen as a key to succesful semantic data integration [48], but having domain ontologies will not solve the data integration problem right away as even whithin a single domain there are many competing ontologies. For example the *C. elegans development* and *C. elegans anatomy* ontologies from the Open Biomedical Ontology repository [8] and *C. elegans cell and anatomy* ontology developed for WormBase[9] were all developed to describe concepts related to the worm anatomy, but, with slightly different research goals in mind.

As most of the existing resources contain annotations from only one ontology, any researcher interested in performing a cross-species analysis would need a method to combine the annotations contained in all the data sources. As an example, one of the tasks currently receiving a lot of attention is linking genome sequence information to organism function, which is commonly

[8] http://www.obofoundry.org/

[9] http://www.wormbase.org/

accomplished by characterizing phenotypes resulting from mutations. The required bridging between genotype and phenotype information is generally achieved through the integration of knowledge sources such as EntrezGene(EG) and Onlime Mendelian Inheritance in Man (OMIM) [10]. The ontologies used by EG and OMIM, as by most biomedical systems, have been developed independently, and since they do not adhere to a common vocabulary their integration is performed manually or by highly customized software [82]. An automatic mapping system will greatly speed up the integration process, and a significant amount a research is being conducted in this area. Much of the work is aimed at leveraging the results accumulated in the similar area of database schema matching, but new techniques are needed for ontology mapping and integration as this area presents challenges and opportunities not existing in databases.

Both ontologies and schemas (i) provide a vocabulary of terms that describes a domain of interest and (ii) constrain the meaning of terms used in the vocabulary. However, database schemas often do not provide explicit semantics for their data as the semantics is, usually, specified explicitly only at modelling time and it is not a part of a database specification and therefore not available. Formal ontologies, on the other hand, are logical systems that obey some formal semantics so that we can interpret ontology definitions as sets of logical axioms. The mapping strategies also differ in the way they perfom the core operation of assessing the *similarity* between the items being matched. In database schema matching the similarity is evaluated with the help of techniques that “guess” the meaning encoded in the schemas, while the ontology matching systems (primarily) try to exploit the knowledge explicitly encoded in the ontologies.

A comprehensive discussion of the differences and commonalities between ontologies database schemas and other knowledge representation technologies is outside the scope of this paper and I refer the reader to [123] for a good overview.

The focuss of this review is *semantic similarity*, an issue central to data processing algorithms such as functional gene clustering and validation of interaction networks as well as integrative data discovery systems.

[10] Available at: http://www.ncbi.nlm.nih.gov/sites/entrez?db=omim

References

[1] F. Al-Shahrour, R. Diaz-Uriarte, and J. Dopazo. Fatigo: A Web Tool for Finding Significant Associations of Gene Ontology Terms with Groups of Genes. *Bioinformatics*, 20(4):578 – 580, 2004.

[2] H. Alani and C. Brewster. Metrics for Ranking Ontologies. In *International World Wide Web Conference*, Edinburgh, UK, 2006.

[3] M. Altinel and M. J. Franklin. Efficient Filtering of XML Documents for Selective Dissemination of Information. In *VLDB*, Cairo, Egypt, 2000.

[4] F.G Ashby and N.A. Perrin. Toward a Unified Theory of Similarity. *Psychological Review*, 95(1):124 – 150, 1988.

[5] F. Azuaje, H. Wang, and O. Bodenreider. Ontology-driven Similarity Approaches to Supporting Gene Functional Assesment. In *International Conference on Intelligence Systems for Molecular Biology(ISMB)*, Detroit, MI, USA, 2005.

[6] F. Azuaje, H. Wang, H. Zheng, O. Bodenreider, and A. Chesneau. Predictive Integration of Gene Ontology-Driven Similarity and Functional Interactions. In *IEEE International Conference on Data Mining(ICDM)*, Hong Kong, China, 2006.

[7] M. Baitaluk, X. Quian abd S. Godbole, A. Raval, A. Ray, and A. Gupta. Pathsys: Integrating Molecular Interaction Graphs for Systems Biology. *Bioinformatics*, 5(55), 2006.

[8] S. Banarjee and T. Pedersen. Extended Gloss Overlap as a Measure of Semantic Relatedness. In *Eighteen International Conference on Artificial Intelligenece (IJCAI)*, Acapulco, Mexico, 2003.

[9] D. Barnard, G. Clarke, and N. Duncan. Tree-to-tree Correction for Document Trees. Technical report, Department of Computing and Information Science, Queen's University, Kingston, Ontario, Canada, January 1995.

[10] A. Bernstein, E. Kaufmann, and C. Burki. How Similar is it? Towards Personalized Similarity Measures in Ontologies. In *the 7th Internationale Tagung Wirtschaftsinformatik*, Bamberg, Germany, 2005.

[11] A. Bilke and F. Naumann. Schema Matching using Duplicates. In *International Conference on Data Engineering (ICDE)*, Tokyo, Japan, 2005.

[12] A. Boukottaya and C. Vanoibeek. Schema Matching for Transforming Structured Documents. In *ACM Symposium on Document Engineering, DocEng*, Bristol, UK, 2005.

[13] P. Bouquet and S. Zanobini. Semantic Coordination: a New Approach and an Application. In *International Semantic Web Conference (ISWC)*, Sanibel Islands, Florida, USA, 2003.

[14] R. Brachman. What IS-A is and isn't: an Analysis of Taxonomic Links in Semantic Networks. *Computer*, 16(10):30 – 36, 1983.

[15] J. Brank, D. Mladenic, and M. Grobelnik. Gold Standard Based Ontology Evaluation Using Instance Assignment. In *Workshop on Evaluation of Ontologies for the Web, EON*, Edinburgh, UK, 2006.

[16] A. Brazma, P. Hingamp, J. Quackenbush, G. Sherlock, P. Spellman, C. Stoeckert, J. Aach, W. Ansorge, C.A. Ball, H.C. Causton, T. Gaasterland, P. Glenisson11, F. C.P. Holstege, I.F. Kim, V. Markowitz, J.C. Matese, Helen Parkinson, A. Robinson, U. Sarkans, S. Schulze-Kremer, J. Stewart, R. Taylor, J. Vilo, and M. Vingron. Minimum Information about a Microarray Experiment (MIAME): Towards Standards for Microarray Data. *Nature Genetics*, 29:365 – 371, 2001.

[17] O. Brodenreider, M. Aubry, and A. Burgun. No-lexical Approaches to Identifying Associative Relations in the Gene Ontology. In *Pacific Symposium on Biocomputing*, Big Island, Hawaii, USA, 2005.

[18] K.R Brown and I. Jurisica. Unequal Evolutionary Conservation of Human Protein Interactions in Interologous Networks. *Genome Biology*, 8, 2007. R95.

[19] A. Budanistky and G. Hirst. Semantic Distance in WordNet: an Experimental Application-oriented Evaluation of Five Measures. In *Second meeting of the Nort American Chapter of the Association for Computational Linguistics(NAACL)*, Pittsburg, PA, USA, 2001.

[20] A. Budanitsky and G. Hirst. Evaluating WordNet-based Measures of Lexical Semantic Relatedness. *Computational Linguistics*, 32(1):13 – 47, 2006.

[21] C.J. Bult. Data Integration Standards in Model Organisms: from Genotype to Phenotype in the Laboratory Mouse. *Drug Discov Today*, 1:163 – 168, 2002.

[22] A. Burger, D. Davidson, Y. Yang, and R. Baldock. Integrating Partonomic Hierarchies in Anatomy Ontologies. *Bioinformatics*, 5(184), 2004.

[23] P.A. Champin and C. Solnon. Measuring Similarity of Labeled Graphs. In *International Conference on Case-based Reasoning, ICCBR*, Thondheim, Norway, 2003.

[24] J.M. Cherry, C. Ball, S. Weng, G. Juvik, R. Schmidt, C. Adler, B. Dunn, S. Dwight, L. Riles, R.K. Mortimer, and D. Botstein. Genetic and physical maps of saccharomyces cerevisiae. *Nature*, 387:67–73, 1997.

[25] P. Cimiano, A. Hotho, and S. Staab. Learning Concept Hierarchies from Text Corpora Using Formal Concept Analysis. *Journal of AI Research*, 24:305 – 339, 2005.

[26] The Gene Ontology Consortium. Gene Ontology: Tool for the Unification of Biology. *Nature Genetics*, 25:25 – 29, 2000.

[27] F.M. Couto, M. J. Silva, and P. Coutinho. Implementation of a Functional Semantic Similarity Measure between Gene-products. Technical report, Departamento di Informatica, Lisbon University, Lisbon, Portugal, 2003.

[28] F.M. Couto, M. J. Silva, and P. Coutinho. Semantic Similarity over Gene Ontology: Family Correlation and Selecting Disjuntive Ancestors. In *ACM Fourteenth Conference on Information and Knowledge Management (CIKM)*, Bremen, Germany, 2005.

[29] S.B. Davidson, J. Crabtree, B. P. Brunk, J. Schug, V. Tannen, G. C. Overton, and C. J. Stoeckert. K2/Kleisli and GUS: Experiments in Integrated Access to Genomic Data Sources. *IBM Systems Journal*, 40(2):512 – 531, 2001.

[30] R. Dhamankar, Y. Lee, and A. Doan. iMap: Discovering Complex Semantic Matches between Databse Schemas. In *International Conference on Management of Data (SIGMOD)*, Paris, France, 2004.

[31] H.H. Do and E. Rahm. Flexible Integration of Molecular-Biological Annotation Data: the GeneMapper Approach. In *Advances in Database Technology - EDBT 2004/ Lecture Notes in Computer Science*, volume 2992, pages 811 – 822. Springer Berlin / Heidelberg, 2004.

[32] A. Doan, J. Madhavan, R. Dhamankar, P. Domingos, and A.Y. Halevy. Learning to Match Ontologies on the Semantic Web. *Journal of Very Large Databases*, 12(4):303 – 319, 2003.

[33] A. Doan, J. Madhavan, P. Domingos, and A. Halevy. Learning to Map between Ontologies on the Semantic Web. In *International World Wide Web Conference*, Honolulu, Hawaii, USA, 2002.

[34] M. Ehrig and P. Haase. Similarity for Ontologies - a Comprehensive Framework. In *European Conference on Information Systems (ECIS)*, Regensburg, Germany, 2005.

[35] M. Ehrig and S. Staab. QOM - Quick Ontology Mapping. In *International Semantic Web Conference*, Hiroshima, Japan, 2004.

[36] M. Ehrig and Y. Sure. *Ontology Mapping - an Integrated Approach*, pages 76 – 91. Springer Berlin, 2004.

[37] D. Embley, L. Xu, and Y. Ding. Automatic Direct and Indirect Schema Mapping: Experiences and Lessons Learned. *SIGMOD Record*, 33(4), 2004.

[38] D. Empley, D. Jackman, and L. Xu. Attribute Match Discovery in Information Integration: Exploiting Multiple Facets of Metadata. *Journal of the Brazilian Computer Society*, 8(2), 2002.

[39] C. Fellbaum, editor. *WordNet. An electronic lexical database.* MIT Press, Cambridge, MA, USA, 1998.

[40] F.Giunchiglia, P. Shvaiko, and M. Yatskevich. Semantic schema matching. Technical report, Department of Information and Communication Technology, University of Trento, Trento, Italy, March 2005.

[41] A.J. Finkel, editor. *Current Medical Information and Terminology.* American Medical Association, Chicago, IL, USA, 1981.

[42] H. Frohlich, N. Speer, A. Poustka, and T. Beissbarth. GOSim - an R-package for Computation of Information Theoretic GO Similarities between Terms and Gene Products. *Bioinformatics*, 8(166), 2007.

[43] A. Gangemi, C. Catenacci, M. Ciaramita, and J. Lehman. A Theoretical Framework for Ontology Evaluation and Validation. In *Workshop on Semantic Web Applications and Perspectives (SWAP)*, Trento, Italy, 2005.

[44] F. Giunchiglia and P. Shvaiko. Semantic Matching. Technical report, Department of Information and Communication Technology, University of Trento, Trento, Italy, April 2003.

[45] F. Giunchiglia, P. Shvaiko, and M. Yatskevich. S-match: an Algorithm and an Implementation of Semantic Matching. Technical report, Department of Information and Communication Technology, University of Trento, Trento, Italy, February 2004.

[46] F. Giunchiglia and M. Yatskevich. Element Level Semantic Matching. Technical report, Department of Information and Communication Technology, University of Trento, Trento, Italy, June 2004.

[47] T. Gruber. A Translation Approach to Portable Ontology Specification. *Knowledge Aquisition*, 5:199 – 220, 1993.

[48] M. Gruninger. Model-theoretic Approaches to Semantic Integration. In *Dagstuhl Seminar: Semantic Interoperability and Integration*, Dagstuhl, Germany, 2005.

[49] X. Guo, C.D. Shriver, H. Hu, and M.N. Liebman. Semantic Similarity-based Validation of Human Protein-Protein Interactions. In *IEEE Omputational Systems Bioinformatics Conference Workshops(CSBW)*, Stanford, CA, USA, 2005.

[50] B. Bagheri Hariri, H. Abolhassani, and A. Khodaei. A New Structural Similarity Measure for Ontology Alignment. In *International conference on Semantic Web & Web Services, SWWS*, Las Vegas, NV, USA, 2006.

[51] E. Heit. Features of Similarity and Category-based Induction. In *An Interdisciplinary Workshop On Similarity And Categorisation (SimCat)*, Edinburgh, UK, 1997.

[52] T. Hernandez and S. Kambhampati. Integration of Biological Sources: Current Systems and Challenges Ahead. *Sigmod Record*, 9, 2004.

[53] P. Hu, G. Bader, D.A. Wigle, and A. Emili. Computational Prediction of Cancer-Gene Function. *Nature Reviews Cancer*, 7:23 – 34, 2007.

[54] Rosetta Inpharmatics. Gene expression markup language, 2007.

[55] P. Jaccard. Distribution de la flore alpine dans le bassin des dranses et dans quelques regions voisine. *Bulletin del la Socit Vaudoisdes Sciences Naturelles*, 37:241 – 272, 1901.

[56] K. Janowicz. Kinds of Contexts and their Impact on Semantic Similarity Measurement. In *the 5th IEEE Workshop on Context Modeling and Reasoning (CoMoRea) at the 6th IEEE International Conference on Pervasive Computing and Communication (PerCom0*, Hong Kong, China, 2008.

[57] M. Jarmasz and S. Szpakowicz. Roget's Thesaurus and Semantic Similarity. In *International Conference on Recent Advances in Natural Language Processing(RANLP)*, Borovets, Bulgaria, 2003.

[58] J.J Jiang and D.W. Conrath. Semantic Similarity Based on Corpus Statistics and Lexical Taxonomy. In *International Conference on Computational Linguistics(ROCLING X)*, Taiwan, 1997.

[59] I. Jurisica. Context-based Similarity Applied to Retrieval of Relevant Cases. Technical report, University of Toronto, Toronto, Canada, 1994.

[60] I. Jurisica. *Theory, Implementation and Applications of Similarity-based Retrieval for Case-based Reasoning*. PhD thesis, University of Toronto, Toronto, Canada, 1998.

[61] Yannis Kalfoglou and Marco Schorlemmer. Ontology Mapping: the State of the Art. In Y. Kalfoglou, M. Schorlemmer, A. Sheth, S. Staab, and M. Uschold, editors, *Semantic Interoperability and Integration*, number 04391 in Dagstuhl Seminar Proceedings, Dagstuhl, Germany, 2005. Internationales Begegnungs- und Forschungszentrum fuer Informatik (IBFI).

[62] V. Kashyap and A. Sheth. Semantic and Schematic Similarities between Database Objects: a Context-based Approach. *Very Large Database Journal*, 5(4):276 – 304, 1996.

[63] C. Kessler. Similarity Measurement in Context. In *the 6th International Conference and Interdisciplinary Conference, CONTEXT*, Roskilde, Denmark, 2007.

[64] P. Khatri and S. Draghici. Ontological Analysis of Gene Expression Data: Current Tools, Limitations, and Open Problems. *Bioinformatics*, 21(18):3587 – 3595, 2005.

[65] W. Kim, S. Park, S. Bang, and S. Lee. An Ontology Mapping Algorithm between Heterogeneous Product Classification Taxonomies. In *International Workshop on Ontology Matching collocated with the 5th International Semantic Web Conference (ISWC)*, Athens, Georgia, USA, 2006.

[66] Y.W. Kim and J.H Kim. A Model of Knowledge Based Information Retrieval with Hierarchical Concept Graph. *Jourbnal of Documentation*, 46(2):113 – 136, 1990.

[67] L.B. Larkey and A.B. Markman. Processes of similarity judgement. *Cognitive Science*, 29:1061 – 1076, 2004.

[68] J.H. Lee, M.H. Kim, and Y.J. Lee. Information Retrieval Based on Conceptual Distance in Is-a Hierarchies. *Journal of Documentation*, 49(2):188 –207, 1993.

[69] L.J. Lee. *Similarity-based Approaces to Natural Language Processing.* PhD thesis, Harvard University, Cambridge, MA, USA, 1997.

[70] C. Lemer, H. Anerhour, J.M Maniraja, O. Sand, J. Richelle, and S.J Wodak. The aMAZE Database Goes Public. In *International Conference on Intelligent Systems for Molecular Biology (ISMB)/ European Conference on Computational Biology (ECCB)*, Glasgow, UK, 2004.

[71] D. Lin. An Information-Theoretic Definition of Similarity. In *International Conference on Machine Learning*, Madison, Wisconsin, USA, 1998.

[72] N. Lin, B. Wu, R. Jansen, M. Gerstein, and H. Zhao. Information assessment on predicting protein-protein interactions. *BMC Bioinformatics*, 5(154), 2004.

[73] P.W. Lord, R.D. Stevens, A. Brass, and C.A. Goble. Investigating Semantic Similarity Measures Across the Gene Ontology: the Relationship between Sequence and Annotation. *Bioinformatics*, 19(10):1275 – 1283, 2003.

[74] P.W. Lord, R.D. Stevens, A. Brass, and C.A. Goble. Semantic Similarity Measures as Tools for Exploring the Gene Ontology. In *Pacific Symposium on Biocomputing*, Lihue, Hawaii, USA, 2003.

[75] L.J. Lu, A. Paccanaro, H. Yu, and M. Gerstein. Assessing the limits of genomic data integration for predicting protein networks. *Genome Research*, 15:945 – 953, 2005.

[76] A. Maedche and S. Staab. Comparing Ontologies - Similarity Measures and a Comparison Study. Technical report, Institute AIFB, University of Karlsrue, Karlsrue, Germany, 2001.

[77] A. Maedche and S. Staab. Measuring Similarity between Ontologies. In *International Conference on Knowledge Engineering and Knowledge Management: Ontologies and the Semantic Web (EKAW)*, Siguenza, Spain, 2002.

[78] Y. Matar, E. Egyed-Zsigmog, and S. Lajmi. Kwsim : Concepts Similarity Measure. In *Conference an Recherche d'Information et Applications*, Tregastel, France, 2008.

[79] D. Maynard, W. Peters, and Y. Li. Metrics for Evaluation of Ontology-based Information Extraction. In *International World Wide Web Conference*, Edinburgh, UK, 2006.

[80] D. Medin, J. Son, and D. Gentner. Respects for Similarity. *Psychological Review*, 100(2):254 – 278, 1993.

[81] G.A. Miller and W.G. Charles. Contextual Correlates of Semantic Similarity. *Language and Cognitive Processes*, 1(6):1 – 28, 1991.

[82] J.A Mitchell, A.T. McCray, and O. Bodenreider. From Phenotype to Genotype: Issues in Navigating the Available Information Resources. *Methods of Information in Medicine*, 42(5):557 – 563, 2003.

[83] P. Mitra, N. Noy, and A. Jaiswal. Omen: a Probabilistic Ontology Mapping Tool. In *International Semantic Web Conference*, Galway, Ireland, 2005.

[84] H.A. Nguyen and H. Al-Mubaid. New Ontology-based Semantic Similarity Measure for the Biomedical Domain. In *IEEE International Conference on Granular Computing*, Atlanta, GA, USA, 2006.

[85] A. Ouangraoua and P. Ferraro. A New Constrained Edit Distance between Quotiented Ordered Trees. Technical report, Laboratoire Bordelais de Recherche en Informatique, Bordeaux, France, May 2007.

[86] S. Patwardhan and T. Pedersen. Using WordNet Based Context Vectors to Estimate the Semantic Relatedness of Concepts. In *EACL Workshop Making Sense of Sense - Bringing Computational Linguistics and Psycholinguistics Together*, Trento, Italy, 2006.

[87] P.Bille. Tree Edit Distance, Allignment Distance and Inclusion. Technical report, The IT University of Copenhagen, Copenhagen, Denmark, 2003.

[88] T. Pedersen, S.V.S Pakhomov, S. Patwardhan, and C.G. Chute. Measures of Semantic Similarity and Relatedness in the Biomedical Domain. *Journal of Biomedical Informatics*, 40:288 – 299, 2007.

[89] Y. Peng, Z. Ding, R. Pang, Y. Yu, B. Kulvatunyou, N. Ivezic, A. Jones, and H. Cho. In *Industrial Engineering Research Conference*, Nashville, Tennessee, USA, 2007.

[90] C. Pesquita, D. Faria, H. Bastos, A.E.N. Ferreira, A.O. Falcao, and F.M. Couto. Metrics for GO-based Protein Semantic Similarity: Systematic Evaluation. *Bioinformatics*, 9, 2008.

[91] M. Popescu, J. M. Keller, and J. A Mitchell. Gene Ontology Automatic Annotation Using a Domain Based Gene Product Similarity Measure. In *The 14th IEEE International Conference on Fuzzy Systems*, Changsha, China, 2005.

[92] C. Posse, A. Sanfilippo, B. Gopalan, R. Riensche, N. Beagley, and B. Baddeley. Cross-ontological Analytics: Combining Associative and Hierarchical Relations in the Gene Ontologies to Assess Gene Product Similarity. In *International Conference on Computational Science*, Reading, UK, 2006.

[93] R. Rada and C. Coccia. A Knowledge-base for Retrieval Evaluation. In *ACM Annual conference*, Denver, Colorado, USA, 1985.

[94] R. Rada, H. Mili, E. Bicknell, and M. Blettner. Development and Application of a Metric on Semantic Nets. *IEEE Transactions on Systems, Man and Cybernetics*, 19:17 – 30, 1989.

[95] C. Kessle1and M. Raubal and K. Janowicz. The Effect of Context on Semantic Similarity Measurement. In *On the Move to Meaningful Internet Systems: OTM 2007 Workshops*, Vilamoura, Portugal, 2007.

[96] P. Resnik. Using Information Content to Evaluate Semantic Similarity in a Taxonomy. In *the 14th International Joint Conference on Artificial intelligence*, Montreal, Canada, 1995.

[97] P. Resnik. Semantic Similarity in a Taxonomy: an Information-based Measure and its Application to Problems of Ambiguity in Natural Language. *Journal of Artificial Intelligence Research*, 11:95 – 130, 1999.

[98] M.M. Richter. Classification and Learning of Similarity Measures. Technical report, Universitat Kaiserslautern, Kaiserslautern, Germany, 1992.

[99] M.A. Rodriguez and M. J. Egenhofer. Comparing Geospatial Entity Classes: an Asymmetric and Context-Dependent Similarity Measure. *International Journal of Geographical Information Science*, 18:229–256, 2004.

[100] R.Pan, Z. Ding, Y. Yu, and Y. Peng. A Bayesian Network Approach to Ontology Mapping. In *International Semantic Web Conference*, Galway, Ireland, 2005.

[101] H. Rubenstein and J.B. Goodenough. Contextual Correlates of Synonymy. *Communications of the ACM*, 8:627 – 633, 1965.

[102] A. Schlicker and M. Albrecht. Funsimmat: a Comprehensive Functional Similarity Database. *Nucleic Acids Research*, 36:D434 – D439, 2008.

[103] A. Schliker, C. Huthmacher, F. Ramirez, T. Lengauer, and M. Albrecht. Functional Evaluation of Domain-Domain Interactions and Human Protein Interaction Networks. *Bioinformatics*, 23(7), 2007.

[104] S. Schulze-Kremer. Ontologies for Molecular Biology and Bioinformatics. *In Silico Biology*, 2:422 – 433, 2002.

[105] " N. Seco, Tony Veale, and J. Hayes". "An Intrinsic Information Content Metric for Semantic Similarity in WordNet". In *European conference on Artificial Intelligence*, Valencia, Spain, 2004.

[106] P. Shavaiko. A Classification of Schema-based Matching. Technical report, Department of Information and Communication Technology, University of Trento, Trento, Italy, 2004.

[107] P. Shavaiko. Iterative Schema-based Semantic Matching. Technical report, Department of Information and Communication Technology, University of Trento, Trento, Italy, 2004.

[108] P. Shvaiko and J. Euzenat. A Survey of Schema-based Matching Approaches. *Journal on Data Semantics*, IV:146 – 171, 2005.

[109] T. Slimani, B. Ben Yagahlane, and K. Mellouli. A New Similarity Measure Based on Edge Counting. *Proceedings of the World Academy of Science, engineering and Technology*, 17, 2006.

[110] D. Slotta. *EvaluatingBiological Data using Rank Correlation Methods.* PhD thesis, Virginia Polytechnic Institute, 2005.

[111] B. Smith, J. Kohler, and A. Kumar. On the Application of Formal Principles to Life Science Data: a Case Study in the Gene Ontology. In *Data Integration in Life Sciences*, Leipzig, Germany, 2004.

[112] B. Smith and A. Kumar. Controlled Vocabularies in Bioinformatics: a Case Study in the Gene Ontology. *Biosilico*, 2(6):246 – 252, 2004.

[113] R.R Sokal and P.H.A Sneath. *Principles of Numerical Taxonomy.* Freeman, San Francisco, CA, USA, 1963.

[114] S. Staab and R. Studer, editors. *Handbook on Ontologies*, chapter Ontology Matching: a Machine Learning Approach, pages 385 – 404. Springer Verlag, 2004.

[115] E. Stoica and M. Hearst. Predicting Gene Functions from Text Using a Cross-Species Approach. In *Pacific Symposium on Biocomputing*, Grand Wailea, Maui, Hawaii, 2006.

[116] M. Sussna. Word Sense Disambiguation for Free-text Indexing Using a Massive Semantic Network. In *the Second International Conference on Information and Knowledge Management, CIKM*, Washington, DC, USA, 1993.

[117] O. Svab and V. Svatek. Combining Ontology Mapping Methods Using Bayesian Networks. In *Workshop on Ontology Matching at ISWC*, Athens, Georgia, USA, 2006.

[118] KC. Tai. The Tree-to-Tree Correction Problem. *Journal of the Association for Computing Machinery*, 26(3):422 – 433, 1979.

[119] Z. Tang, S. Phan, Y. Pan, and F. Famili. Prediction of Co-Regulated Gene Groups through Gene Ontology. In *the IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology*, Honolulu, HW, USA, 2007.

[120] Y. Tao, L. Sam, J. Li, C. Freidman, and Y.A. Lussier. Information Theory Applied to the Sparse Gene Ontology Annotation Network to Predict Novel Gene Function. *Bioinformatics*, 23:i529 – i538, 2007.

[121] J. Tuikkala, L. Elo, O.S. Nevalainen, and T. Aittokallio. Improving Missing Value Estimation in Microarray Data with Gene Ontology. *Bioinformatics*, 22:566 –572, 2006.

[122] A. Tversky. Features of similarity. *Psychological Review*, 84:327 – 352, 1977.

[123] M. Uschold and M. Gruninger. Ontologies and Semantics for Seamless Connectivity. *SIGMOD Record*, 33(4), 2004.

[124] W.R. van Hage, S. Katrenko, and G. Schreiber. A Method to Combine Linguistic Ontology-Mapping Techniques. In *International Semantic Web Conference*, Galway, Ireland, 2005.

[125] A. Wagner. Estimating Frequency Counts of Concepts in Multiple-Inheritance Hierarchies. *LDV Forum*, 19:81–91, 2004.

[126] H. Wang, F. Azuaje, O. Bodenreider, and J. Dopazo. Gene Expression Correlation and Gene Ontology-based Similarity: an Assessment of Quantitative Relationships. In *IEEE*

Symposium on Computational Intelligence in Bioinformatics an d Computational Biology(CIBCB), La Jolla, CA, USA, 2004.

[127] J. Wang, Z. Du, R. Payattakool, P.S. Yu, and C-F. Chen. A New Method to Measure the Semantic Similarity of GO Terms. *Bioinformatics*, 23(10):1274 — 1281, 2007.

[128] J. Wei. Markov Edit Distance. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(3):311 – 321, 2004.

[129] P. C. Weinstein and W. P. Birmingham. Comparing Concepts in Differentiated Ontologies. In *12th Workshop on Knowledge Acquisition, Modeling and Management (KAW)*, Banff, Alberta, Canada, 1999.

[130] Dominic Widdows. A Mathematical Model for Context and Word-Meaning. In *Fourth International and Interdisciplinary Conference on Modeling and Using Context*, Stanford, CA, USA, 2003.

[131] Dominic Widdows. *Geometry and Meaning.* CSLI Publications, Stanford University, Stanford, CA, USA, 2004.

[132] H. Wu, Z. Su, F. Mao, V. Olman, and Y. Xu. Prediction of Functional Modules Based on Comparative Genome Analysis and Gene Ontology Application. *Nucleic Acids Research*, 33(9):2822 – 2837, 2005.

[133] X. Wu, L. Zhu, D.Y. Zhang J. Guo, and K. Lin. Predictions of Yeast Protein - Protein Interaction Network: Insights from the Gene Ontology and Annotations. *Nucleic Acids Research*, 34(7):2137 – 2150, 2006.

[134] Z. Wu and M. Palmer. Verb Semantics and Lexical Selection. In *the 32nd Annual Meeting of the Associations for computational Linguistics*, Las Cruces, NM, USA, 1994.

[135] D. Yang, Y. Li, H. Xiao, Q. Liu, M. Zhang, J. Zhu, W. Ma, C. Yao, J. Wang, D. Wang, Z. Guo, and B. Yang. Gaining Confidence in Biological Interpretation of the Microarray

Data : the Functional Consistence of the Significant GO Categories. *Bioinformatics*, 24(2), 2008.

[136] H. Yu, R. Jensen, G. Stolovitzky, and M. Gerstein. Total Ancestry Measure: Quantifying the Similarity in Tree-like Classification, with Geometric Applications. *Bioinformatics*, 23(16):2163 – 2173, 2007.

[137] K. Zhang. A Constrained Edit Distance between Unordered Labeled Trees. *Algorithmica*, 15:205 – 222, 1996.

[138] S. Zhang and O. Bodenreider. NLM Anatomical Ontology Alignment System Results of the 2006 Ontology Alignment Contest. In *International Semantic Web Conference*, Athens, Georgia, USA, 2006.