Context-Specific Ontology Integration: A Bayesian Approach

Kshitij Marwah[1,6], Dustin Katzin[1,2,3], Amin Zollanvari, PhD[1,4], Natalya F. Noy, PhD[5], Marco Ramoni, PhD[1,*], Gil Alterovitz, PhD[1,4,6]

[1] Children's Hospital Informatics Program at Harvard-MIT Division of Health Science, Boston, MA; [2]Department of Physics, Massachusetts Institute of Technology, Cambridge, MA; [3] Department of Mathematics, Massachusetts Institute of Technology, Cambridge, MA; [4] Center for Biomedical Informatics, Harvard Medical School, Boston, MA; [5] Stanford Center For Biomedical Informatics Research, Stanford CA; [6]Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, MA; [*] Deceased.

Abstract

We introduce a principled computational framework and methodology for automated discovery of context-specific functional links between ontologies. Our model leverages over disparate free-text literature resources to score the model of dependency linking two terms under a context against their model of independence. We identify linked terms as those having a significant bayes factor ($p < 0.01$). To scale our algorithm over massive ontologies, we propose a heuristic pruning technique as an efficient algorithm for inferring such links.

We have applied this method to translationalize Gene Ontology to all other ontologies available at National Center of Biomedical Ontology (NCBO) BioPortal under the context of Human Disease ontology. Our results show that in addition to broadening the scope of hypothesis for researchers, our work can potentially be used to explore continuum of relationships among ontologies to guide various biological experiments.

1. Introduction

Every year, over 400,000 new articles reportedly enter biomedical literature [1]. This staggering growth of biomedical findings has created an unprecedented corpus of knowledge that is impossible to explore with traditional means of literature consultation and database searches. This information overload has motivated the development of structured information repositories that organize biomedical findings according to hierarchical ontologies.

Ontologies find themselves at the heart of two major complementary activities in biomedical research. Communities of researches create and maintain these ontologies to represent different types of entities and relations in different domains of biomedicine. On the other hand, biomedical experimentalists use ontologies to annotate data in order to facilitate data integration and translational discoveries. This activity is greatly intensified by the development of high-throughput experimental platforms such as gene expression microarrays [2], SNP microarrays [3] and next generation sequencing platforms [4].

The rise of such ontological organization has created a new problem, the proliferation of disparate and seemingly unrelated biomedical ontologies. For example, the National Center of Biomedical Ontology's (NCBO) BioPortal [5] provides over 200 such ontologies to researchers. These ontologies are generally used by scientists to annotate their data, but which ontologies to use and how they relate to each other is generally unclear. What is needed is the integration of these conceptualizations in a principled fashion, a "grand unification" of biological terms. It has been established [6] that the integration of these available ontologies will have a tremendous impact on the advancement of biomedical sciences. These integrated ontologies will provide a complete basis of biomedical knowledge representation and act as a foundation for inference on new biomedical data. Furthermore, a quantitative approach for integration would make the navigation of the complex space of ontologies more amenable to researchers by offering them guidance to numerous links among ontologies, ranking them according to a principled metric, thus making the discovery process faster and efficient.

To date, the mapping and integrating of ontologies in the biomedical domain has relied on discovering links between syntactically and semantically similar terms across ontologies [7]. Such an approach can relate terms with similar meanings but would not deduce any relationships between seemingly disparate functional spaces such as diseases, drugs and anatomy. Approaches in the data integration community for

ontology integration use methods ranging from machine learning [8] to graph matching [9] to natural language processing [10]. These methods again inherently focus on mapping synonyms across ontologies. Recently, Ontology Alignment Evaluation Initiative [11] has been launched as a competition between alignment algorithms on a given standardized dataset. These methods generally cater to the definition of traditional ontology alignment considering synonyms. Even instance-based methods in these initiatives for mappings have the goal of converging two ontologies that represent the same knowledge base. For domains as disparate as biomedical ontologies, such methods do not work and moreover, the computational complexity of these algorithms makes them infeasible for massive scales of such vocabularies. Other approaches to infer these links use standard means of manual curation, which is again a tedious and labor intensive task with extremely bad scaling properties.

Here we propose a novel computational and methodological framework for context-specific integration of biomedical ontologies using free-text literature analysis. We model context specificity using another ontology and derive context-dependent functional links between ontological concepts occurring as phrases in free-text literature. We cache massive amounts of literature data to enable efficient counts of co-occurring ontology terms. Based on these statistics, the penalized likelihood of the model of dependency and independency is computed by applying the well-known bayesian information criterion [12] over a context-sensitive model scoring function. We account for scalability via a depth-first branch and bound heuristic technique, to prune sub-graphs that do not yield significant links.

We believe that such a methodological approach would turn machine-processable ontologies into a single landscape of integrated biomedical concepts and annotations. This would enable researchers to bear on each single finding the entire power of established biomedical knowledge.

2. Methods

A. Caching Sufficient Statistics

We gather raw free-text literature from disparate sources and drive our concept search by finding exact matches of ontology terms. We use the MGREP [13], concept recognition tool that also powers the NCBO Annotator [14] to efficiently find occurrence of concepts in published literature and thus annotate the documents with those concepts. This allows us to leverage on a consolidate vocabulary (of about 4 million ontology concepts) to temper the problem of missing synonyms and term permutations.

We also used a pre-computed index containing the transitive closure of ontology terms for semantically expanding the annotations, propagating them up the hierarchy of the ontology. The document annotations and the concepts are reverse indexed using a disk based b-tree structure an approach commonly used in information retrieval systems.

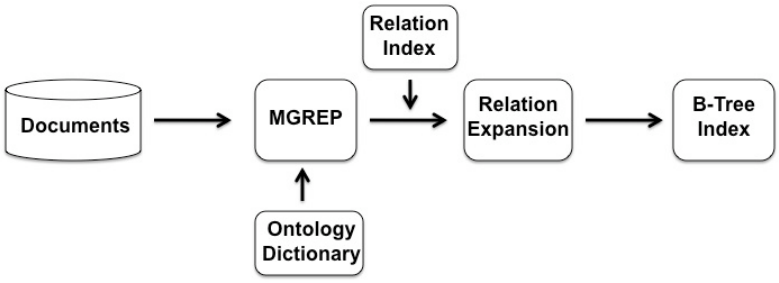

Figure 1: Pipeline used for caching sufficient statistics for model scoring.

We use Lucene [15], an open source high-powered information retrieval engine to create and store the b-tree structure. To answer conjunctive queries for efficient counting we use a bitmap hash-based filter over the stored index. Our integrated pipeline is shown Figure 1 above.

B. Alignment Algorithm

For computing context dependent links between ontology terms, we have developed a novel technique relying on statistical analysis of literature. Our algorithm uses the observed co-occurrence of terms in the literature to infer the relationship between two terms A and B in the context of the ontology term C. As an example the term A can be the ontology concept, 5-fluorouracil, which we want to align with the term B, cell-cycle under the context of term C, say colon cancer.

To do so it builds a contingency table like the one in Figure 3, collecting the frequencies of co-occurrence of the two terms in the literature, a 2 x 2 table where n_{++} is the number of papers in which two

terms appear together, n_{+-} is the number of papers in which A appears but B does not, n_{-+} is the number of papers in which B appears and A does not, and n_{--} is the number of papers in which neither appear all in the context of term C.

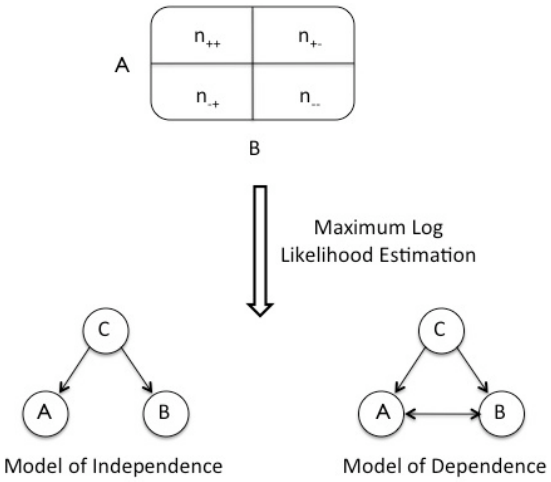

Figure 2: 2 x 2 contingency table to test relationship between two ontology terms A and B under C.

Our method uses the Bayesian information criterion to compute the penalized likelihood of dependence A ⇔ B | C (where two terms are related) and the model of independence A ⇑ B | C (where the two terms are unrelated) as

$$BIC = -2\,MLL + k\log(N), \tag{1}$$

where N is the number of observations, k is the number of parameters of the model, and MLL is the marginal log likelihood of the model. We assume that both the models of dependence and independence are equally likely in which case maximizing the posterior probability converges with maximizing the marginal likelihood as shown in Equation 1.

The marginal log likelihood for the model of dependency is:

$$\begin{aligned} MLL(A \Leftrightarrow B \mid C) = {} & [\ln(\Gamma(\alpha)) - \ln(\Gamma(\alpha + n))] \\ & + [\ln(\Gamma(\alpha_k + n_{++})) - \ln(\Gamma(\alpha_k))] \\ & + [\ln(\Gamma(\alpha_k + n_{+-})) - \ln(\Gamma(\alpha_k))] \\ & + [\ln(\Gamma(\alpha_k + n_{-+})) - \ln(\Gamma(\alpha_k))] \\ & + [\ln(\Gamma(\alpha_k + n_{--})) - \ln(\Gamma(\alpha_k))] \end{aligned} \tag{2}$$

whereas the marginal log likelihood for the model of independence is:

$$\begin{aligned} MLL(A \Uparrow B \mid C) = {} & [\ln(\Gamma(\alpha)) - \ln(\Gamma(\alpha + n))] \\ & + [\ln(\Gamma(\alpha_k + n_{++} + n_{+-})) - \ln(\Gamma(\alpha_k))] \\ & + [\ln(\Gamma(\alpha_k + n_{-+} + n_{--})) - \ln(\Gamma(\alpha_k))] \\ & + [\ln(\Gamma(\alpha)) - \ln(\Gamma(\alpha + n))] \\ & + [\ln(\Gamma(\alpha_k + n_{-+} + n_{++})) - \ln(\Gamma(\alpha_k))] \\ & + [\ln(\Gamma(\alpha_k + n_{--} + n_{+-})) - \ln(\Gamma(\alpha_k))] \end{aligned} \tag{3}$$

where Γ is the gamma function, n_{++}, n_{+-}, n_{-+}, n_{--} are the co-occurrence frequencies as described above, α is the prior precision and, α_k is the prior precision per term, that is, α/|T|, where |T| is the number of terms

in the dependency: in our particular case, |T| = 2. In our case, we use $\alpha = 4$ for 2 x 2 tables, so that for the initial prior precision we put 1 in each cell, maintaining the uniformity of the distribution and the lowest possible precision, so as to minimize bias on the precision.

By plugging the marginal log likelihood into equation (1), we obtain respectively the penalized likelihood of dependency BIC ($A \Leftrightarrow B \mid C$), where the two terms are linked, and the model of independence BIC ($A \Uparrow B \mid C$), where the two terms are not linked. The final score is the bayes factor

$$Score = BIC(A \Leftrightarrow B \mid C) - BIC(A \Uparrow B \mid C) \quad (4)$$

that estimates how many times the model linking term A and B in the context of C is more likely than the model in which the terms are not related.

We use the pipeline explained in the previous section to efficiently count the co-occurrence frequencies, for computing the bayes factor. Context-dependent functional links are then selected as the ones having a bayes factor greater than 20 ($p < 0.01$).

C. Heuristic Pruning Using Depth First Branch and Bound

To apply our algorithm we, in the worst case, would have to compare all possible triples of terms representing the ontologies. Such an approach would work for small ontologies but will not scale up to massive ontologies even with cached statistics. We apply a depth first branch and bound algorithm to prune away ontology sub-graphs where the likelihood of finding functional links is extremely low. We use the bayes factor as a scoring cue to find such sub-graphs.

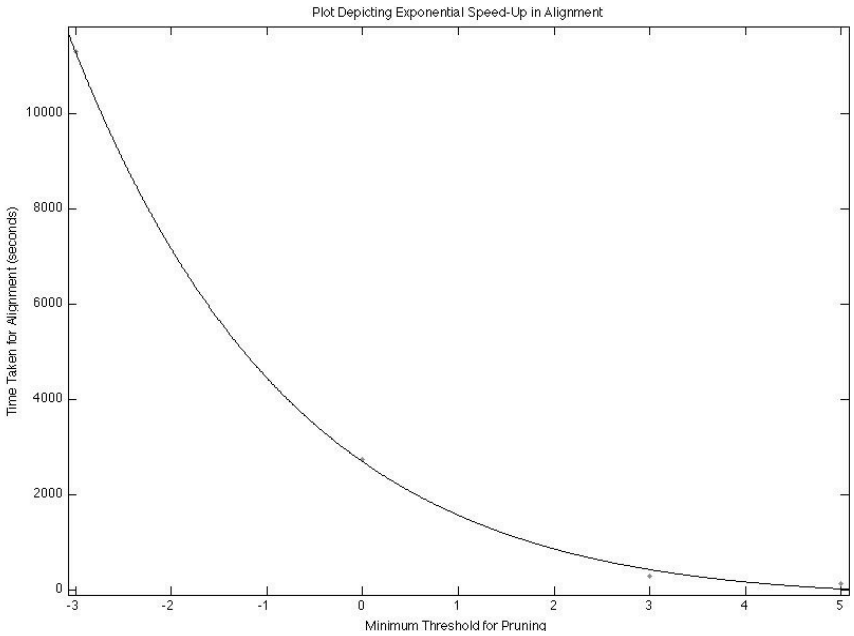

Figure 3: Graph depicting exponential reduction in running time as the minimum threshold for pruning increases.

We build on the empirical observation that if the bayes factor for an ontology concept A mapped to another ontology concept B under the context C is less than given a custom user-set threshold ε, then the bayes factor for mappings amongst majority of A's children with the concept B under C would also be less than ε. An intuition towards such an observation can be gauged from the fact that any instance of a specific concept, say a paper, is also an instance of a more general concept. This follows the subsumption property that the taxonomy structure of an ontology follows. Hence, if not enough evidence is found for linking A to B under C, as demonstrated by the computed bayes factor it follows that a major fraction of A's children would also not have enough evidence of a map to B under C.

Further extending the empirical observation to span sub-graphs under A and B in context of the sub-graph under C helps us to use the metric to prune away insignificant portions in the ontological graph. We rather than giving theoretical bounds on the likelihood of matches, experimentally analyze the effect of the given threshold ε over the running time and the amount of false negatives. Our results show below an expected exponential reduction in computations for inferring functional links.

We also depict below the linear increase in the amount of false negatives if we prune the full graph.

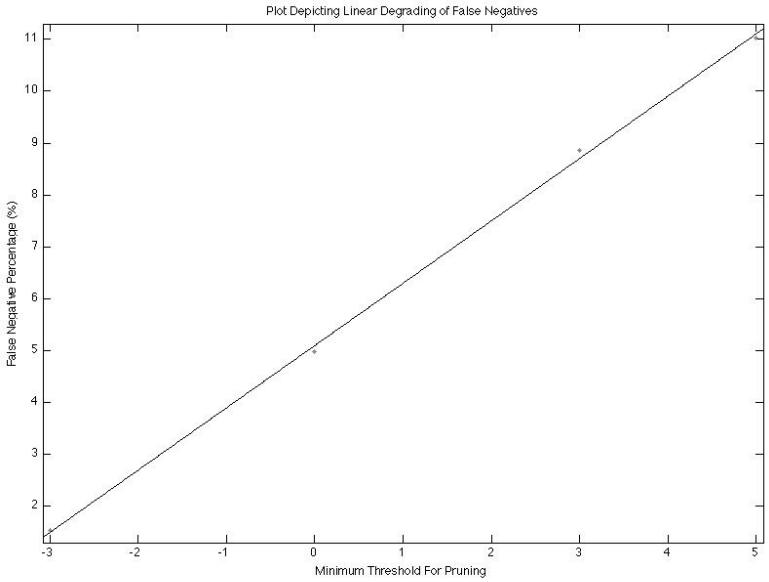

Figure 4: Graph depicting linear degradation in the amount of inferred links as the minimum threshold for pruning increases.

We implement the depth first branch and bound algorithm allowing us to compute functional links with much greater efficiency with a trade-off in loss of some alignments. The minimum threshold can be controlled by the user, depending on the efficacy of the results required. A suitable threshold can be determined empirically, by running the algorithm with different thresholds and observing the occurrence of "false positive" links. Once this threshold is chosen, we say that if the bayes factor is greater than ε (or corresponding desired significance level via corresponding p-value), than a high-confidence link exists between concepts.

3. Results

We obtain in all about 200 ontologies from the National Center for Biomedical Ontology's BioPortal interface. For caching sufficient statistics we obtain the dictionary of all available ontology concepts (4,153,358 terms) for searching in the corpora. We further create our b-tree index on the corpus containing the following:

1. Adverse Event Reporting System [16] database containing about 774,606 records.
2. Array Express [17] containing 9281 records.
3. BioSiteMaps [18] data containing 1013 records.
4. caNanoLab [19] data containing 444 records.
5. Conserved Domain Databases [20] containing 34,735 records.
6. Clinical Trials [21] database containing 75,828 records.
7. Drug Bank [22] containing 4774 records.
8. Database of Phenotypes and Genotypes [23] having 184 records.
9. Gene Expression Omnibus [24] containing 15,968 records.
10. Stanford Microarray Database [25] containing 16,148 records.
11. Published articles in PubMed [26] containing about 100,000 records.

Each element of the corpus contains the full abstract of corresponding published article. We then apply our proposed algorithm over the heuristic pruning technique described earlier to integrate Gene Ontology (containing 24,987 concepts) to all available ontologies in BioPortal under the context of Human Disease Ontology (containing 12,033 concepts). The threshold for a significant link was set to be with a bayes factor greater than twenty ($p < 0.01$), while the threshold for pruning was set to be with bayes factor less than zero.

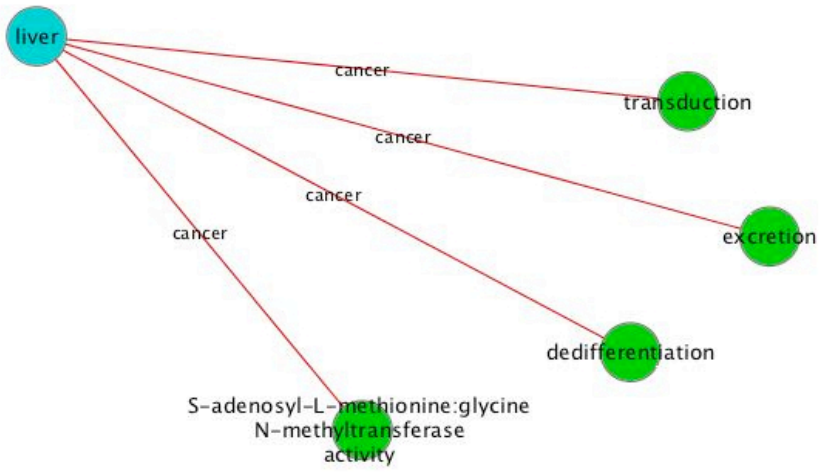

Figure 5: A part of mapping network showing links between Gene Ontology (green circles) and Minimal Anatomical Terminology (blue circles) under the context of Human Disease (red links).

An example of such a network is shown above in Figure 5. This is a part of a full network containing about 2000 relevant links. Figure 6 another network in which we switch the context to Minimal Anatomical Terminology from Human Disease. In such ways, our framework can take any two ontologies and compute scalable mappings under any given context.

To validate the soundness of our context-sensitive mappings we take a random sampling of about a thousand high information content links [27], having a significantly high bayes factor. We repeat the experiment about ten times and use published literature and a domain expert in the field of molecular biology to validate these links. The number of repetitions are constrained by time resource available at our disposal for the domain expert. The precision number for the algorithm using this approach was found to be about 0.78.

We further validate the completeness of our mappings by again taking a random sampling of about a thousand high information content triplets of nodes. We then use published literature and the domain expert to predict links amongst these concepts. These predicted links are then matched against the ones inferred by our algorithm to get recall. We repeat the experiment about ten times to get the recall number for the algorithm, which was found to be about 0.91. This corresponds to f-measure about 0.83. These numbers underscore the robustness and quality of our inferred links.

4. Discussion

This work is based on data that is changing and evolving over time. New data enters the biomedical literature and ontological databases constantly. Thus, conclusions and links can change over time. This framework provides an efficient and scalable algorithm to incorporate big data prevalent in the biomedical domain. A limitation of such analysis is its inability to differentiate between positive and negative correlation. Though nodes may be connected but their type of association is not computed. Incorporating some shallow semantics from natural language processing domain would help such a cause. A sliding window that detects relationships in conjunction with ontology concepts can be implemented to classify these alignments.

A better algorithm to incorporate and update new data would be a nice addition accompanied by a graphical visualization toolkit to succinctly map such links. We only consider textual abstracts for caching statistics ontology terms. Expanding to full-text articles and incorporating varied datasets like images and experimental data would be interesting and challenging. A further extension of such a framework to propagate annotations over these links and perform enrichment analysis on ontologies other than Gene Ontology would be extremely useful. Another exciting analysis for future work would be to look at the evolution of the derived links over time as biological knowledge expands. Such a network can provide insights of how different biological terms relate to each other as advancements and new knowledge is added. They can also be used to detect and predict clusters of influence and propagation. Combining these links into a continuous bridge between different domains can help guide

biological experiments and analyses.

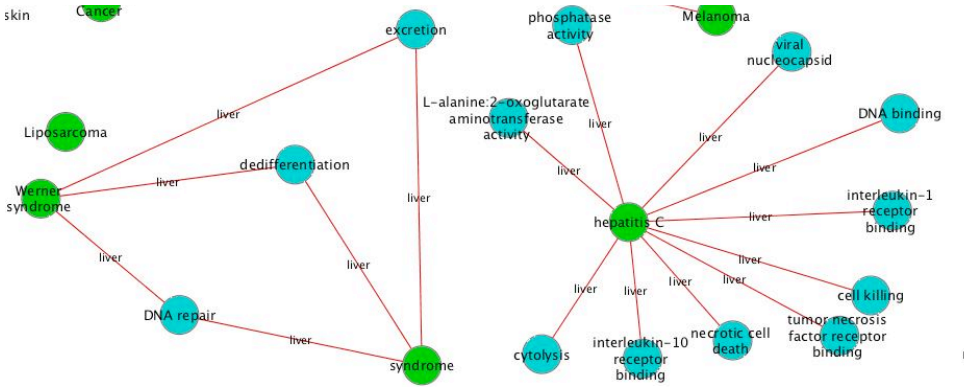

Figure 6: Portion of network showing context-specific links between Gene Ontology (blue circles) and Human Disease (green circles) in context of Minimal Anatomical Terminology (red links).

5. Conclusion

Our framework and algorithms combine disparate sources of data for discovery of relationships between ontologies. Unlike prior work, our approach tries to find context-specific functional links between ontologies, which is not possible if only semantically-relevant links were considered. By developing a novel algorithm we identified links across ontologies, which can be used for guided expansion of various biomedical experiments. We then augmented this algorithm with heuristic approaches, for scaling up to massive data sizes with marginal loss in functional quality of links. We further validated the utility of our algorithm, by manual verification using a domain expert, increasing confidence in our methodological approach. Our work provides a new approach for translationalizing diverse functional spaces in biomedical domain, making this huge space of knowledge amenable to researchers.

6. Acknowledgements

This work was supported by the NIH grants P01CA89392 from the National Cancer Institute, 5R21DA025168-02 (G. Alterovitz), 1R01HG004836-01(G. Alterovitz), and 4R00LM009826¬03 (G. Alterovitz).

7. References

1. F. Moerchen, D. Fradkin, M. Dejori, B. Wachmann, Emerging Trend Prediction in Biomedical Literature, *AMIA Annual Symposium Proceedings*, 485 (2008).
2. M.K. Kerr, G.A. Churchill, Experimental design for gene expression microarrays, *Biostatistics* 2(2): 183-201. (2001).
3. A.C. Syvänen, Toward genome-wide SNP genotyping, *Nature Genetics* 37: S5-S10. (2005).
4. B. Smith, Ontology (Science), *Nature Proceedings,* 2008.
5. N. Noy, N. Shah, P. Whetzel, B. Dai, M. Dorf, N. Griffith, C. Jonquet, D. Rubin, M. Storey, C. Chute, M. Musen,, BioPortal: ontologies and integrated data resources at the click of a mouse, *Nucleic Acids Research*, 37 (2009).
6. L. Jensen, P. Bork, Ontologies in Quantitative Biology: A Basis for Comparison, Integration, and Discovery, *PLoS Biology* (2010).
7. N. Silva, J. Rocha, Complex semantic web ontology mapping, *Web Intelligence and Agent Systems* 1(3-4): 235-248. (2003).
8. A. Doan, J. Madhavan, P. Domingos and A. Halevy, Ontology Matching: A Machine Learning Approach, *International World Wide Web Conference,* 2002.
9. N. Noy and M. Musen, PROMPT: Algorithm and Tool For Automated Ontology Merging and Alignment, *National Conference on Artificial Intelligence,* 2000.
10. M. Ehrig and S. Staab, QOM – Quick Ontology Mapping, *International Semantic Web Conference,* 2004.

11. C. Caracciolo, J. Euzenat, L. Hollink, R. Ichise, A. Issac, V. Malaise, C. Meilicke, J. Pane, P. Shivko, H. Struckenschmidt, O.S. Zamazal and V. Svatek, Results of the Ontology Alignment Evaluation Initiative, *International Semantic Web Conference Workshop on Ontology Matching* (2008).

12. A. Liddle, Information criteria for astrophysical model selection, *Monthly Notices of the Royal Astronomical Society*, 377 (2007).

13. M. Dai, An Efficient Solution for Mapping Free Text to Ontology Terms, *AMIA Summit on Translational Bioinformatics*, (2008).

14. C. Jonquet, N. Shah, C. Youn, M. Musen, C.Callendar, M. Storey, NCBO Annotator: Semantic Annotation of Biomedical Data, *International Semantic Web Conference* (2009).

15. E. Hatcher, O. Gospodnetic, Lucene in Action, *JavaOne Conference*, (2004).

16. S.D. Ross, M.W. Reynolds, Use of the FDA spontaneous adverse event reporting system (SAERS), or why your MedWatch reports really do matter, *Journal of Clinical Ontology, 2004 ASCO Annual Meeting Proceedings (Post-Meeting Edition)* 22(14S). (2004).

17. A. Brazma, H. Parkinson, U. Sarkans, M. Shojatalab, J. Vilo, N. Abeygunawardena, E. Holloway, M. Kapushesky, P. Kemmeren, G.G. Lara, A. Oezcimen, P. Rocca-Serra, S.A. Sansone, ArrayExpress—a public repository for microarray gene expression data at the EBI, *Nucleic Acids Research* 31(1):68-71. (2003).

18. L. Marenco, R. Wang, G.M. Shepherd, P.L. Miller, The NIF DISCO Framework: Facilitating Automated Integration of Neuroscience Content on the Web, *Neuroinform* 8:101-112. (2010).

19. V. Maojo, F. Martin-Sanchez, C. Kulikowski, A. Rodriguez-Paton, M. Fritts, Nanoinformatics and DNA-Based Computing: Catalyzing Nanomedicine, *Pediatrics Research* 67(5): 481-489. (2010).

20. A. Marchler-Bauer, J.B. Anderson, F. Chitsaz, M.K. Derbyshire, C. DeWeese-Scott, J.H. Fong, L.Y. Geer, R.C. Geer, N.R. Gonzales, M. Gwadz, S. He, D.I. Hurwitz, J.D. Jackson, Z. Ke, C.J. Lanczycki, C.A. Liebert, C. Liu, F. Lu, S. Lu, G.H. Marchler, M. Mullokandov, J.S. Song, A. Tasneem, N. Thanki, R.A. Yamashita, D. Zhang, N. Zhang, S.H. Bryant, CDD: specific functional annotatioin with the Conserved Domain Database, *Nucleic Acids Research* 37: D205-10. (2009).

21. M. Mi, Clinical Trials Database: Linking Patients to Medical Research http://clinicaltrials.gov, *Journal of Consumer Health On the Internet*, 9(3): 59-67. (2005).

22. D.S. Wishart, C. Knox, A.C. Guo, D. Cheng, S. Shrivastava, D. Tzur, B. Gautam, M. Hassanali, DrugBank: a knowledgebase for drugs, drug actions and drug targets, *Nucleic Acids Research* 36(Database issue):D901-6. (2008).

23. M.D. Mailman, M. Feolo, Y. Jin, M. Kimura, K. Tryka, R. Bagoutdinov, L. Hao, A. Kiang, J. Paschall, L. Phan, N. Popova, S. Pretel, L. Ziyabari, Y. Shao, Z.Y. Wang, K. Sirotkin, M. Ward, M. Kholodov, K. Zbicz, J. Beck, M. Kimelman, S. Shevelev, D. Preuss, E. Yaschenko, A. Graeff, J. Ostell, S.T. Sherry, The NCBI dbGaP database of genotypes and phenotypes, *Nature Genetics* 39(10): 1181-1186. (2007).

24. T. Barrett, D.B. Troup, S.E. Wilhite, P. Ledoux, D. Rudnev, C. Evangelista, I.F. Kim, A. Soboleva, M. Tomashevsky, K.A. Marshall, K.H. Phillipy, P.M. Sherman, R.N. Muertter, R. Edgar, NCBI GEO: archive for high-throughput functional genomic data, *Nucleic Acids Research* 37(Database issue):D5-15. (2009).

25. J. Hubble, J. Demeter, H. Jin, M. Mao, M. Nitzberg, T.B. Reddy, F. Wymore, Z.K. Zachariah, G. Sherlock, C.A. Ball, Implementation of GenePattern within the Stanford Microarray Database. *Nucleic Acids Res*earch 37(Database Issue):D898-901. (2009).

26. A. Plake, T. Schiemann, M. Pankalla, J. Hakenberg, U. Leser, ALIBABA: PubMed as a graph, *Bioinformatics* 22(19):2444-2445. (2006).

27. G. Alterovitz, M. Xiang, D.P. Hill, J. Lomax, J. Liu, M. Cherkassky, J. Dreyfuss, C. Mungall, M.A. Harris, M.E. Dolan, J.A. Blake, M.F. Ramoni, Ontology engineering, *Nature Biotechnology* 28:128-130. (2010).