# Context-Based Ontology Integration Over Massive Datasets

Kshitij Marwah[1], Natalya F. Noy[2], Paea LePendu[2], Marco F. Ramoni[1*], and Gil Alterovitz[1]

[1] Children's Hospital Informatics Program at Harvard-MIT Division of Health Sciences and Technology [2] Stanford Center for Biomedical Informatics Research, Stanford University [*] Deceased

## Abstract

• Ontologies are critical tools of biomedical research, providing efficient frameworks for structuring and organizing scientific information.

• Currently, these conceptualizations are developed as disparate isolated silos of biological information with no significant relations amongst them.

• Integration of many different biomedical ontologies into a comprehensive landscape of biomedical knowledge can enable researchers identify novel avenues of investigation and generate new hypotheses.

• We present a computational framework for context-specific and functional integration of ontologies, where context is modeled by the introduction of a third ontology .

• We believe that such a methodological approach would help turn available machine process able ontologies into a single landscape of integrated biomedical concepts and annotations.

## Introduction

• Ontologies[1], currently are at the heart of two complementary activities: for representation of varied biomedical entities, and for experimental data annotations[2].

• We present a novel context-specific integration of these various ontologies in a principled fashion, a "grand unification" of biological terms.

• This quantitative approach strives to provide a complete basis of biomedical knowledge representation, and as a foundation for inference of new biomedical data.

## Methods

• We consider all available ontologies from NCBO's Bioportal[4] interface, and gather raw free-text literature from numerous sources.

• We develop a high-throughput pipeline (Figure 1) to cache sufficient statistics by considering ontology term matches in these free-text sources[5].

• Using the above data-structure we compute the penalized likelihood of context-specific model of dependency of terms against the model of independence (Figure 2).

• To circumvent the complexity of search space, we use a depth first branch and bound heuristic technique to prune insignificant links.

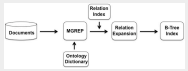


Figure 1. Pipeline used for caching sufficient statistics for model scoring.

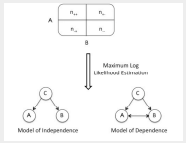


Figure 2. 2 x 2 contingency table to test relationship between ontology terms.

## Results

• We consider 200 ontologies, containing about 3 million terms, and about 1 million free-text abstracts from sources such as Adverse Event Reporting System, Array Express, Gene Expression Omnibus, PubMed and many more.

• We then apply our algorithm to compute integrate Gene Ontology (24,987 concepts) to all other ontologies (Figure 3) under the context of Human Disease (12,033 concepts).

• To validate our links, we take a random sample of about a hundred high information content links[3], and use published literature with a domain expert to compute the efficacy of the algorithm (Figure 4).

• Our preliminary results indicate a high recall value of about 0.88, and a precision value of about 0.76, corresponding to a f-measure of 0.81.

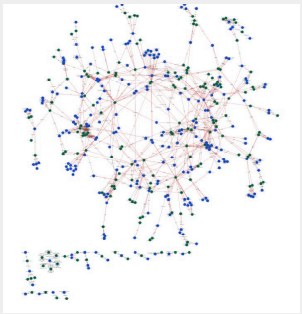

Figure 3. Mapping network showing links between Gene Ontology (blue circles) and Minimal Anatomical terminology (green circles) under Human Disease.

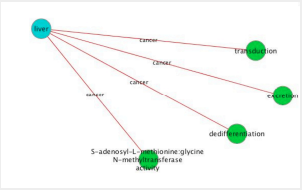


Figure 4. Snapshot showing translation of Gene Ontology to Anatomy under Human Disease.

## Conclusions

• 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, which is not possible if only syntactically relevant links are considered.

• Our work provides a new approach for translating diverse functional spaces in biomedical domain, and making this huge knowledge space amenable to researchers.

• Our integrative method can enable researchers to bear on each single finding, the entire power of established biomedical knowledge.

• For more information please see: http://bcl.med.harvard.edu

## References

1. B. Smith, Ontology (science), Nature Proceedings, 2008.
2. J. Blake, Bio-ontologies-fast and furious, Nature Biotechnology, 2004.
3. G. Alterovitz et. al., Ontology Engineering, Nature Biotech, 2010.
4. N. Noy et. al., Bioportal : ontologies and integrated data resources at the click of a mouse, Nucleic Acids Research, 2009.
5. U. Hahn et. al., Text mining: powering the database revolution, Nature, 2007.

## Acknowledgements

This work was supported in part by the National Library of Medicine (NLM/NIH) under grants 1K99LM009826 and 5T15LM007092 and by the National Human Genome Research Institute (NHGRI/NIH) under grants 2P41HG02273, 1R01HG003354 and 1R01HG004836.

PARTNERS HEALTHCARE BWH

**Biomedical Cybernetics Laboratory**

bcl.med.harvard.edu