# Annotation Concept Synthesis and Enrichment Analysis

## A Logic-Based Approach to the Interpretation of High-Throughput Biological Experiments

Mikhail Jiline[1,*], Stan Matwin[1,2] and Marcel Turcotte[1]
{mjiline, stan, turcotte}@site.uottawa.ca
[1]School of Information Technology and Engineering, University of Ottawa,
800 King Edward Avenue, Ottawa, Ontario, K1N 6N5 Canada.
[2]Institute of Computer Science, Polish Academy of Sciences, Warsaw, Poland
[*] Presenting Author

### Abstract

Annotation Enrichment Analysis (AEA) is a widely used analytical approach to process data generated by high-throughput genomic and proteomic experiments such as gene expression microarrays. The analysis uncovers and summarizes discriminating background information (e.g. GO, Chromosome mapping, KEGG annotations) for sets of genes identified by experiments (e.g. a set of differentially expressed genes, a cluster). The discovered information is utilized by human experts to find biological interpretations of the experiments.

The main drawback of AEA is that it isolates and tests for overrepresentation individual annotation terms or groups of similar terms. As a result, AEA is limited in its ability to uncover complex phenomena involving relationships between multiple annotation terms from various knowledge bases. Also, AEA assumes that annotations describe the whole object of interest, which makes it difficult to apply it to sets of compound objects (e.g., sets of protein-protein interactions) and to sets of objects having an internal structure (e.g., protein complexes).

To overcome the drawbacks shortcoming, we propose a novel logic-based Annotation Concept Synthesis and Enrichment Analysis (ACSEA) approach. In this approach, the annotation information, experimental data and uncovered enriched annotations are represented as First-Order Logic (FOL) statements. ACSEA uses the fusion of inductive logic reasoning with statistical inference to uncover more complex phenomena captured by the experiments. The proposed paradigm allows for a synthesis of enriched annotation concepts that better describe the observed biological processes.

The methodological advantage of Annotation Concept Synthesis and Enrichment Analysis is six-fold. Firstly, it is easier to represent complex, structural annotation information. Information already captured and formalized in OWL and RDF knowledge bases can be directly utilized. Secondly, it is possible to synthesize and analyze complex annotation concepts. Thirdly, it is possible to perform the enrichment analysis for sets of aggregate

objects (such as sets of genetic interactions, physical protein-protein interactions or sets of protein complexes). Fourthly, annotation concepts are straightforward to interpret by a human expert. Fifthly, the logic data model and logic induction are a common platform that can integrate specialized analytical tools (e.g. tools for numerical, structural and sequential analysis). Sixthly, used statistical inference methods are robust on noisy and incomplete data, scalable and trusted by human experts in the field.

In our approach we also develop a novel method for constructing explanatory theories from very large sets of highly synonymical annotation concepts. Such sets are normally encountered in ACSEA due to noisy experimental and annotational data, natural redundancy of annotational data and strong expressive power of First-Order Logic representation model.

We evaluate our approach on large-scale datasets from several microarray experiments and on a clustered genome-wide genetic interaction network using different biological knowledge bases. The discovered interpretations have lower P-values than the interpretations found by AEA, are highly integrative in nature, and include analysis of quantitative and structured information present in the knowledge bases. The results suggest that ACSEA can boost effectiveness of the processing of high-throughput experiments.