Ontology based analysis of experimental data

Andrea Splendiani
Institute Pasteur, Unité de Biologie Systémique, rue du dr. Roux 25-28, 75015 Paris, France
University of Milano-Bicocca, DISCO, via Bicocca degli Arcimboldi 8, 20126 Milano, Italy
andrea.splendiani@unimib.it

Abstract

We address the problem of linking observations from reality to a semantic web based knowledge base. Concepts in the biological domain are increasingly being formalized through ontologies, with an increasing adoption of semantic web standards. At the same time biology is becoming a data-centric science, since the increasing availability of high throughput technologies yields a humanly intractable amount of data describing the behavior of biological systems at the molecular level. This creates the need for automated support to interpret biological data given the pre-existing knowledge about the biological systems under study. While this is currently addressed through the analysis of attributes associated to biological entities, the availability of ontologies that represent biological systems makes it possible to improve the extent to which pre-existing knowledge can be used. The semantic web, in particular, provides a framework to integrate and create a formalized biological knowledge base. Linking ontological knowledge to observed data is inherently approximate, because of the quality of observations, the relation between observed data and entities and the classification of entities. We present an overall framework project and its current development status.

1 Introduction

Scientific exploration constantly involves relating experimental evidence to existing knowledge. In the Life Science fields existing knowledge is commonly encoded in a corpus of scientific publications. This knowledge is used by scientists to design and interpret experiments that in turn lead to new discoveries. Traditionally this meant to relate a limited amount of experimental evidence to pre-defined hypotheses.

Recently, the availability of high-throughput technologies, such as DNA sequencing, mRNA and proteomic profiling is challangingchallenging this existing paradigm. Such technologies allow the observation of the behavior of biological systems at the molecular level on a system-wide basis. This means that a humanly intractable amount of data is available, most of which does not relate to previous hypotheses. Thus relating such a large scale experimental evidence to the existing biological knowledge is essential in order to understand the phenomenon under study.

Given the vast amount of data generated by high throughput technologies, this necessitates automated support.

At the same time there is an increasing availability of structured biological information, in the semantic web framework in particular. The Gene Ontology[Ashburner et al., 2000] initiative provides ontologies describing functions of gene products. It currently encompasses more than 17000 terms linked by relations of inheritance and containment and it is available in RDF. MGED-ontology[1] provides an ontology to describe attributes relevant to mRNA experiments in OWL, and the BioPAX[The BioPAX workgroup] initiative is defining a common standard to represent biological pathways and interaction networks in OWL. This last initiative is of particular interest since it provides a common ontological framework for the unification of different resources. The availability of such resources makes it possible to partially automatize the association between experimental evidence and existing knowledge to effectively lead data analysis.

2 Ontologies and data analysis

Focusing on ontologies that describe the behavior of biological systems at the molecular level, there is a range of ontologies that vary in scope and depth. While available knowledge of some biological systems is enough to build causal models, in general such knowledge is limited and most of ontologies have a low ontological commitment. When dealing with system-wide observations, this second class of ontologies is most relevant.

For instance, in the case of mRNA profiling, the behavior of thousands of genes in a cell in response to some sort of stimuli is observed. For each gene, measures of its activity are provided. These data are usually related to Gene Ontology to derive a functional characterization of the cell response.

Associations of genes to specific classes in Gene Ontology are determined based on available knowledge. By its semantics, association of a gene to a class implies association of a gene to its super classes too. Thus a gene is annotated with a set of classes that act as attributes describing specific biological functions.

[1] www.mged.org

It is common practice to define a subset of relevant genes from experimental data and to study the incidence of these attributes derived from Gene Ontology through statistical tests[Beissbarth et al., 2004; Maere et al., 2005].

Sometimes relations of inheritance and independence are used to measure “conceptual distances” among genes[Joslyin et al., 2004].

3 Uncertainty

Uncertainty plays a key role in the task of associating experimental evidence to ontological knowledge, at several levels.

Uncertainty in the definition and relations between classes plays a limited role. There is not a specific support for uncertainty in OWL, and the definition of ontologies is an ongoing task where crisp definitions are valuable.

Association between instances and classes is one point where uncertainty plays a critical role. Almost every ontology encodes a confidence in the association through “evidence codes” (describing the kind of supporting evidence) and eventually a p-value or citations of relevant scientific literature. See [Karp et al., 2004] for an example of an ontology for experimental evidence.

Association between data and ontologies is then inherently uncertain. Uncertainty may come not only from the experimental setup and measurements, but also from the biological source of variability, and from misconceptions or omissions in the available knowledge.

4 Our project

The way experimental data are associated to existing ontologies now does not take into account all the information encoded in ontologies and does not provide a way to reason over related uncertainty. We plan to overcome these limitations by providing a framework for approximate reasoning based in ontologies.

In particular, we focus on OWL ontologies describing biological pathways and on mRNA data. Given an ontology representing a collection of pathways and related concepts (including evidence support), and a set of experimental data, we define a new ontology as the union of the two, representing observed evidence and the previous knowledge.

Thus we plan to use the structure of previous knowledge to compute plausibility of concepts being pertinent to observed conditions. This can be done through a rule based approach, where inherent structure of pathways ontologies would ensure convergence of plausibility distributions.

5 Current development

We have developed an infrastructure where ontologies can be merged and represented. This is based on the Cytoscape[Shannon et al., 2003] software for molecular interaction analysis which is used as a link to experimental data and an interactive visualizer for RDF ontologies. Rule systems for unifying the ontologies and graph transformations to represent views of ontologies are also provided.

Based on this, we plan to provide a Bayesian network along with the ontology, or possible derivation of that, and to update the plausibility of nodes associated to concepts given evidence encoded in roots nodes. This updates involves considering both the experimental evidence, and uncertainty assessment related to it.

6 Conclusion

The Life Science community is one of the early adopters of semantic web technologies. The need to represent and integrate a vast amount of different information is pushing the development of this technology. The analysis of high-throughput data poses naturally the need of approximate reasoning and uncertainty representation.

References

[Ashburner et al., 2000] Ashburner M, et al. *Gene ontology: tool for the unification of biology.* The Gene Ontology Consortium. Nat Genetics 2000 May;25(1):25-9

[The BioPAX workgroup] The BioPAX workgroup. *BioPAX: Biological Pathways Exchange.* www.biopax.org

[Beissbarth et al., 2004] Beissbarth T, Speed TP. *GOstat: find statistically overrepresented Gene Ontologies within a group of genes.* Bioinformatics 2004 Jun12;20(9):1464-5

[Maere et al., 2005] Maere S, Heymans K, Kuiper M. *BiNGO: a Cytoscape plugin to assess overrepresentation of Gene Ontology categories in biological networks.* Bioinformatics 2005 Aug 21; 3448-3449

[Karp et al., 2004] Karp PD, Paley S, Krieger CJ, Zhang P. *An evidence ontology for use in pathway/genome databases.* Pac Symp Biocomput. 2004; 190-201

[Joslyin et al., 2004] Joslyn CA, Mniszewski SM, Fulmer A, Heaton G. *The gene ontology categorizer.* Bioinformatics 2004 Aug 4;20 Suppl 1:I169-I177

[Shannon et al., 2003] Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker Y. Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res. 2003 Nov;13(11):2498-504