Semantic Web Platforms for Bioinformatics and Life Sciences

Massimiliano Picone
University of Rome Sapienza, Italy

Maurizio Lenzerini
University of Rome Sapienza, Italy

INTRODUCTION

The amount of data being generated in the life sciences has increased exponentially in the past few years, with DNA sequencing beating Moore's law since 2008. New challenges are thus being posed for data integration and analysis in order to cope with this massive amount of information. We review some of the most promising platforms that are leveraging the Semantic Web approach, a powerful paradigm that has the potential to address many of the issues being faced in bioinformatics. In doing so, we introduce the field by evaluating ontologies and middleware, highlighting present and future trends. In the last couple of years the interest in Big Data and NoSQL technologies have outshined the vision of the Semantic Web; here we advocate the need of merging some of the technologies in order to leverage both paradigms. The time is ripe for industry-driven and research-driven architectures to come together in order to deliver usable tools in several interdisciplinary fields.

The scientific discovery process has shifted from the traditional approach of formulating hypothesis, doing experiments and interpreting them in a cycle, towards a more complex and data-driven workflow as many fields have changed dramatically, thanks to the heavy use of computers. Life sciences in particular have become more and more information and data centric in the last decade, most notably thanks to the availability of new sequencing and measurements techniques.

Dealing with this vast amount of information requires splitting the traditional interpretation task into several steps, leaning towards a data-driven methodology comprising of Data Management, Analysis and Mining. This approach encompasses several skills and requires contributions from different expertise in order to properly formulate the experiments, analyze the results and drive new insights and conclusions.

As more and more experiments in the field of biology and life sciences become more and more high throughput, expressing data in ways that can be read by computers and ways that can be shared from one experiment to another and from one data source to another is thus becoming increasingly important. The Semantic Web (Berners-Lee, 2001) is a technology stack backed by the World Wide Web Consortium (W3C) that tackles this problem; ontologies that facilitate semantic search and information integration are a fundamental part of this stack.

Despite the dynamicity of biological information has limited the development of ontologies to support dynamic reasoning for knowledge discovery, we advocate that the time is right for bio-ontologies to be developed and exploited at their full potential. Here we aim at introducing the field and some use cases of bio-ontologies.

BACKGROUND

One of the main problems with natural languages is ambiguity: the same word can have different meanings and the same meaning can be expressed using different words. The ambiguity embedded in human languages was of course targeted by many fields throughout history, beginning with philosophy, linguistics and more recently by information science and then computer science.

A way of dealing with this problem, leveraging an information science approach, is through the use of controlled vocabularies, where to a particular concept is assigned an official term in order to avoid any confu-

DOI: 10.4018/978-1-4666-5888-2.ch655

sion. In these tools simple relationships are defined like synonyms and antonyms. There are variations on the scope of these *thesauri*: some take a broad approach in order to classify all human knowledge (like the Library of Congress Subject Headings), while others go more in depth in a particular field (like the Medical Subject Headings or *MeSH*). Others take a semantic network approach and are called *metathesauri*, like the Unified Medical Language System (UMLS).

In addition, these controlled vocabularies define a hierarchy of terms going from the most general, broader term, to the more specific narrower term. The data structure in use is usually a tree, where the broader term lies on top and the narrower term on the bottom.

Ontologies are more complex than controlled vocabularies especially in the complexity of their relationships and in their architecture, something that allows algorithms to infer knowledge that is not explicitly defined in the ontology itself.

Unlike those alternative hierarchical views of concepts such as taxonomies, ontologies often have a graph structure characterized by complex relationships.

Thus, a growing number of biomedical ontologies are being developed. The most common of these is the *Gene Ontology* (GO), developed as a cross-species platform to describe genes and genes products in terms of their molecular functions, biological processes and cellular components. It has already been used to provide useful Data Analysis results for a large number of genes (Beißbarth, 2004). GO is now part of the *Open Biological and Biomedical Ontologies* (OBO)[1] project, supported by the National Center for Biomedical Ontology (NCBO) in the US and others, where dozens of ontologies are hosted. This 'umbrella' is the de facto standard in biomedical ontologies.

A category somewhat in between databases and ontologies is constituted by pathway databases. A pathway is a grouping of a functionally related set of genes; it can be so complex and its interactions and relationships so intertwined that it's challenging to properly represent and store its information.

The *Pathway Resource List*[2] has hundreds of pathway databases available, going from protein-protein interaction databases to knowledge bases on biological entities. The most notable example of a pathway ontology with a unified semantics is *BioPax* (Demir, 2010). Its goal is to provide a common source of data exchange between providers and users, as different providers traditionally describe pathway information in different ways. BioPax defines the class structure as a standard, allowing information providers to map their information as instances.

For instance, the Biological Process category of the Gene Ontology can be seen as a pathway ontology. Ontologies can also communicate between each other and classes in different ontologies can be mapped, with mappings from the NCBO BioPortal being fully documented[3].

GO and several other bio-ontologies are in the OBO file format. This format originated from the Gene Ontology and, during time, it evolved following the goals of human readability, ease of parsing, extensibility and minimal redundancy. Therefore this format is nowadays the backbone of many annotation and data analysis tools in the biomedical field.

In parallel with these developments, ontologies in general have become more prevalent in information technology, with the most visible push coming from the W3C in the form of the Web Ontology Language (OWL), proposed as a standard for building ontologies. Tools to create, maintain and view ontologies without writing any code have also evolved[4], while some are supporting both OBO and OWL[5].

In the Open Biomedical Ontologies (OBO) file format, a concept can either be a term, defined as a class, or a typedef (relationship). ID spaces are assigned to concepts and relationships are represented as triples. This core representation can be mapped to RDF.

OBO has the ability to annotate concepts with metadata, like names and comments, and supports relationship types like sub-class and sub-property, in addition to domain and range attributes. This set of functionalities can be mapped to RDF Schema.

The OBO Ontology Extensions (OBO-OE) layer defines tags for expressing metadata on the entire ontology, allowing for synonyms, equivalences and deprecation of OBO concepts. It can also express specific properties of OBO terms, like set operations, and typedefs like transitivity and symmetry; OBO-OE require constructs defined in OWL.

The main difference between the two is that OWL has globally unique identifiers (URIs), whereas OBO allows local identifiers. In addition, OBO features the 'subset' construct, which has no direct equivalent in OWL; the OBO subset is a collection composed of terms only, defined as part of an ontology. Each term can be part of multiple subsets and the ontology allows for several subsets.