Untangling BioOntologies for Mining

Biomedical Information

Catia Pesquita, Daniel Faria, Tiago Grego, Francisco M. Couto, Mário J. Silva

University of Lisbon, Faculty of Sciences, LASIGE
Phone: +351 217500253
Fax: +351 217500084
Email: {cpesquita, dfaria, tgrego, fcouto, mjs}@di.fc.ul.pt
Untangling BioOntologies for Mining

Biomedical Information

ABSTRACT
Biomedical research generates a vast amount of information that is ultimately stored in scientific publications or in databases. The information in scientific texts is unstructured and thus hard to access, whereas the information in databases, although more accessible, often lacks in contextualization. The integration of information from these two kinds of sources is crucial for managing and extracting knowledge. By structuring and defining the concepts and relationships within a biomedical domain, BioOntologies have taken a key role in this integration. This chapter describes the role of BioOntologies in sharing, integrating and mining biological information, discusses some of the most relevant BioOntologies and illustrates how they are being used by automatic tools to improve our understanding of life.

Keywords: BioLiterature, Biomedical Databases, BioOntology, Data Mining, Molecular Biology, Ontology, Text Mining
INTRODUCTION

The development of high-throughput techniques, such as DNA sequencing, microarrays and automated gene-function studies, is turning biology into an information-based science. This is reflected in the ever growing amount of biological data stored in databases and articles in scientific publications.

Biomedical databases contain mostly sequence data and annotations on entities, such as genes and proteins. However, sequence data is growing at a far greater rate than the manual annotation of the entities, mainly due to curated annotations requiring experimental results to back them up. These are mostly recorded in the scientific literature. As a result, the annotation of databases falls upon expert curators, which have the difficult and time-consuming task of continuously tracking the literature. This has prompted the development of data and text mining approaches for automated annotation, which are now responsible for the vast majority of current annotations. However, extracting knowledge from the literature is far from trivial, due to the inherent complexity of natural language used in scientific texts, preventing automated annotations from achieving the quality attained by expert curators.

In fact, early automated approaches have produced a significant number of misannotations, which are now being propagated due to extrapolation of new annotations derived from them (Devos and Valencia, 2001). Given that the vast majority of annotations is derived by extrapolation from previous annotations and most annotation efforts do not distinguish between extrapolated and curated annotations, this problem is even more serious (Valencia 2005).

One way of improving the knowledge extraction process is by integration of the

1 An annotation consists of a bioentity (e.g.: gene or protein) linked to a statement describing it (e.g.:in terms of molecular function, or location).
2 Taking the UniProt knowledge base as an example, less than 10% of its protein entries are manually annotated.
concepts and context of the field (a.k.a the domain knowledge) into the computational methods for annotation, so that they can achieve the same levels of performance of expert curators (Spasic et al., 2005). Evidently, this requires the translation of the domain knowledge from natural language into a clear, structured and unequivocal form to enable computational reasoning.

The above reasoning leads to the consideration of creating **ontologies**, which can be defined as data models for representing concepts and their relationships within a given domain, enabling reasoning about the objects in that domain. In addition to their role as a source of domain knowledge in the annotation process, ontologies can also be used directly for annotation: biomedical databases can contain ontology terms annotating their entities instead of containing natural language annotation statements. This makes annotations more precise and consistent, and opens the way for computational reasoning over the annotations.

The use of ontologies is also advantageous in other data management activities, such as data integration, data cleansing and data mining (Gardner, 2005). **Data integration** greatly benefits from the unified view provided by ontologies. If two or more databases share the same ontology for annotating their entities, exchanging and integrating information among them becomes much more efficient. The use of ontologies is also important as a guide for solving semantic conflicts between discrepant data sources. Given these factors, the growing use of ontologies has been a key factor in data integration, shifting the emphasis from knowledge management to knowledge representation.

Data cleansing also benefits from the use of ontologies in that having a structured and precise meaning for concepts in a domain enhances the identification of inconsistent or erroneous database entries and the process of their correction.
Data mining can profit from both data cleansing and data integration, so it benefits indirectly from the use of ontologies. In addition, it also benefits from the use of ontologies as a source of domain knowledge to guide the discovery process and as a semantic setting for expressing discovered patterns in concise terms.

The focus of this chapter is explaining what is a BioOntology, describing some successful examples being used by automatic tools to perform important tasks. The rest of this chapter is organized as follows: the next section, *BioOntologies*, will start by presenting a generic definition of the ontology concept and then gives some examples of currently available BioOntologies. It will be followed by the section *Towards Automatic Annotation*, which explains the multiple uses of BioOntologies by automatic annotation tools and gives a brief overview of state-of-the-art tools already using BioOntologies. Finally, the *Future Prospects* section will discuss open questions on this subject, current expectations and possible future directions.
FUTURE PROSPECTS

Due to the quantity and diversity of information it generates, the biomedical sciences are one of the most promising fields for application of ontologies and text mining. The growth of both domains has been mostly the result of investments from large research consortia, which conduct expensive projects that have generated and maintain most of the publicly available biomedical data. Nevertheless, small institutions with limited resources play an important role, complementing the available data and developing innovative approaches that could grow into important trends. For instance, the management of well-founded and broad BioOntologies is clearly an issue to be addressed by large research institutes, but smaller institutions are making important contributions on the development of useful tools to explore that information.

Because of the diversity and evolving nature of biomedical information, designing BioOntologies is a complex task. It requires agreement among the members of a community to define the concepts within its scope, and constant involvement from that community to correct and complete those definitions, since the concepts can change with time or become obsolete, and new concepts can arise. As the success of a BioOntology is directly related to involvement of the community, ontology developers should always consider their expectations and limitations, both when designing and updating a BioOntology.

While BioOntologies are traditionally used mainly for annotation purposes, their ultimate goal should be to accurately represent the domain knowledge so as to allow automated reasoning and support knowledge extraction. The establishment of guiding principles, as in OBO, to guide the development of new BioOntologies is a step in this direction, by promoting formality, enforcing orthogonality, and proposing a common syntax that facilitates mapping between BioOntologies. This not only improves the
quality of individual BioOntologies, but also enables the concerted use of several BioOntologies by computational methods.

However, from the point of view of TM applications, current BioOntologies are still too incomplete, too inconsistent and/or too morpho-syntactically inflexible to efficiently support them. To overcome these limitations, BioOntologies could be designed with TM in mind, for instance by taking advantage of more complex NLP techniques rather than simple text statistics, or even by applying TM techniques in their construction to expand their coverage through automated population and improve their interoperability through automated mapping and integration.

While Bioinformatics has been essential to deal with the growing amount of data and knowledge in Biomedical sciences, its whole potential is still unrealised and it will doubtlessly play a major role in their ultimate goal: understanding how living systems function, and understanding life as a whole (Ideker et al., 2001). Many relevant biological discoveries in the future will result from an efficient exploitation of the existing and newly generated data, which will require innovative and efficient data management and integration approaches. Prominent among these will certainly be the development and use of BioOntologies.

AKNOWLEDGMENTS

This work was partially supported by the Portuguese ‘Fundação para a Ciência e Tecnologia’ with the grants ref. SFRH/BD/29797/2006 and “FIRMS - Future Integrated Research Management Systems” (http://lasige.di.fc.ul.pt/index.php?id=5) ref. POSI/ISFL/13/408.
REFERENCES


KEYWORDS

BioLiterature: The collection of scientific publications in biomedicine.

Biomedical Databases: Databases that store and maintain biomedical data such as gene and protein sequences.

Molecular Biology: Concerns itself with understanding the molecular interactions between the various systems of a cell.

Ontology: Is defined as a specification of a conceptualisation that describes concepts and relationships used within a community.

BioOntology: A BioOntology is an ontology for the biomedical knowledge domain.

Data Mining: The process of discovering meaningful correlations, patterns, and trends by sifting through large amounts of data stored in repositories, using pattern recognition technologies as well as statistical and mathematical techniques.

Text Mining: The process of extracting relevant and non-trivial information and knowledge from unstructured text, usually a collection of documents.