Chapter XIX
Untangling BioOntologies for Mining Biomedical Information

Catia Pesquita
University of Lisbon, Portugal

Daniel Faria
University of Lisbon, Portugal

Tiago Grego
University of Lisbon, Portugal

Francisco M. Couto
University of Lisbon, Portugal

Mário J. Silva
University of Lisbon, Portugal

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 Bio Ontologies 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.
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 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
Related Content

Incremental Algorithm for Discovering Frequent Subsequences in Multiple Data Streams
[www.igi-global.com/article/incremental-algorithm-discovering-frequent-subsequences/58635?camid=4v1a](www.igi-global.com/article/incremental-algorithm-discovering-frequent-subsequences/58635?camid=4v1a)

Analytical Processing Over XML and XLink
[www.igi-global.com/article/analytical-processing-over-xml-xlink/61424?camid=4v1a](www.igi-global.com/article/analytical-processing-over-xml-xlink/61424?camid=4v1a)

TtoO: Mining a Thesaurus and Texts to Build and Update a Domain Ontology
[www.igi-global.com/chapter/ttoo-mining-thesaurus-texts-build/7575?camid=4v1a](www.igi-global.com/chapter/ttoo-mining-thesaurus-texts-build/7575?camid=4v1a)

Knowledge Management Portals for Empowering Citizens and Societies
[www.igi-global.com/chapter/knowledge-management-portals-empowering-citizens/29064?camid=4v1a](www.igi-global.com/chapter/knowledge-management-portals-empowering-citizens/29064?camid=4v1a)