

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

Campo Grande, 1749-016 Lisboa, Portugal.

NOTICE: This is the author's version of a work accepted for publication by Idea Group Publishing. Changes resulting from the publishing process, including peer review, editing, corrections, structural formatting and other quality control mechanisms, may not be reflected in this document. Changes may have been made to this work since it was submitted for publication. A definitive version was subsequently published as a chapter in the book *Handbook of Research on Text and Web Mining Technologies* (M. Song and Y. Wu, eds.), Idea Group Inc., 2009 doi: 10.4018/978-1-59904-990-8

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,

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 annotations¹ 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 annotations². 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

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.

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 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.

BIOONTOLOGIES

Since Ancient Greece, philosophy has dealt with the need to define and structure reality. Aristotle proposed a system to organize the objects of human perception in well-defined *Categories*, beginning with an explanation of synonyms, homonyms and paronyms. He recognized the importance of having clear unequivocal concepts to identify each object. In the 18th century, Linnaeus applied these same concepts to the natural world and developed a taxonomy for classification of living things. These early ideas have evolved into the current definition of Ontology in philosophy as a systematic account of Existence, and as such much more complex than Classification. Although the concept of Ontology has been in use by philosophy for a long time, it was only with the emergence of Artificial Intelligence that computer science borrowed the term to establish content-specific agreements for the sharing and reuse of knowledge among software systems. In this context, Gruber defines an ontology as *a specification of conceptualisations, used to help programs and humans share knowledge. Conceptualisations* refer to the entities: the terms, the relationships between them, and also the constraints of those relationships (Gruber, 1991). On the other hand, *specification* refers to the explicit representation of the conceptualisations. Using this general description, controlled vocabularies, taxonomies and thesaurus can be considered ontologies (Bodenreider and Stevens, 2006). A controlled vocabulary is a list of terms that have been explicitly enumerated. A taxonomy is a collection of controlled vocabulary terms organised into a hierarchical structure. A thesaurus is a networked collection of controlled vocabulary terms. Ideally, an ontology should contain formal explicit descriptions of the concepts (often called classes) in a given domain, which should be organized and structured according to the relationships

between them. They also make the relationship between concepts explicit, which allows further reasoning and enables a fuller representation of the information by including such aspects as interacting partners, specific roles, and functions in specific contexts or locations.

Ontologies have been classified into three types (Stevens *et al.*, 2000):

1. Domain-oriented: either domain specific (e.g. ontology dedicated to a single species) or domain generalisations (e.g. dedicated to gene function or cellular components);
2. Task-oriented: e.g. for annotation analysis;
3. Generic: defining high-level categories that are maintained across several domains (also called top-level or upper-level ontologies).

A well structured ontology will reuse ontologies of the three types, but in a clearly-defined modular way to allow structural modification and concept reusability.

The role of BioOntologies has changed in recent years: from limited in scope and scarcely used by the community, to a main focus of interest and investment. Although clinical terminologies have been in use for several decades, different terminologies were used for several purposes, hampering the sharing of knowledge and its reliability. This has led to the creation of BioOntologies to answer the need to merge and organize the knowledge, and overcome the semantic heterogeneities observed in this domain. While the first attempts at developing them focused on a global schema for resource integration, real success and acceptance was only achieved later by ontologies for annotating bioentities (Bodenreider and Stevens, 2006). Since then, BioOntologies have been used successfully for other goals, such as description of experimental protocols and medical procedures. The examples that follow represent

some of the most widely-used BioOntologies for some of these goals, and also recent efforts for integrated development of BioOntologies.

Gene Ontology

The Gene Ontology³ (GO) was created for functional annotation of gene products⁴ in a cellular context (Ashburner *et al*, 2000). It is divided in three aspects (or GO types): *molecular function*, *biological process* and *cellular component*; which constitute three orthogonal ontologies.

Each of these ontologies is structured as a Directed Acyclic Graph (DAG), which is identical to a tree with the exception that terms can have multiple parents (see Figure 1). The terms are linked to each other by two types of relationships: *is a* and *part of*, the former expressing a simple class-subclass relationship and the latter expressing a part-whole relationship with the particularity that the existence of the *whole* does not imply the existence of the *part*. Each DAG has a root term homonymous to the corresponding GO type, and all three are linked to the global root term *all*.

GO aims at being species-independent. However, as some functional aspects are not common to all life forms, some terms apply only to a given taxonomical group. In such cases, the terms in question specify the taxonomical group to which they apply preceded by the word *sensu*, as in the term *chromosome organization and biogenesis (sensu Eukaryota)*.

³ <http://www.geneontology.org>

⁴ A gene product is the product of a given gene at any level (DNA, RNA or protein).

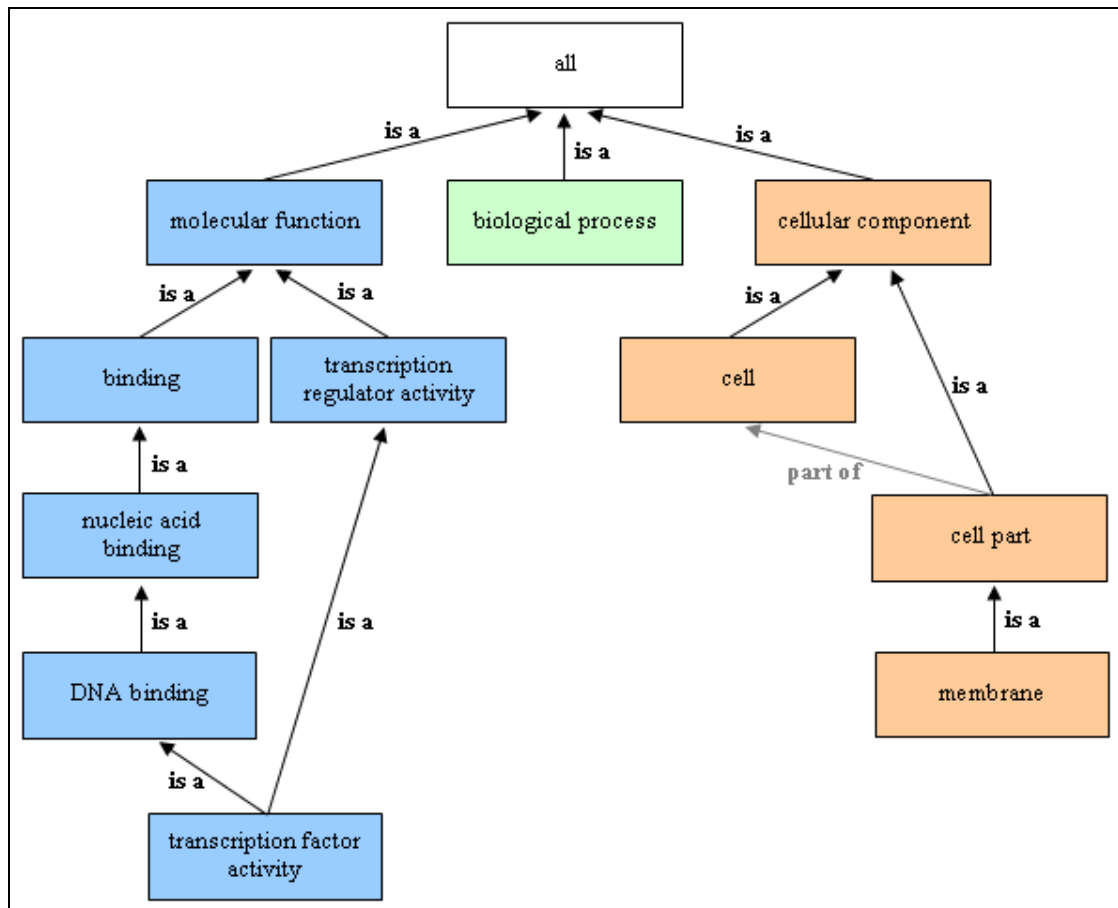


Figure 1: Section of the GO graph showing the three aspects (*molecular function*, *biological process* and *cellular component*) and some of their descendent terms. The fact that GO is a DAG rather than a tree is illustrated by the term *transcription factor activity* which has two parents. An example of a part of relationship is also shown between the terms *cell part* and *cell*.

GO was developed by the GO Consortium, initially a collaboration between three model organism databases (FlyBase, SGD and MGD) to address the need for a common and consistent vocabulary to annotate gene products from different organisms. Since its origin, the GO Consortium has grown to 15 members, which cooperate in maintaining and updating the ontology. GO itself has become widely accepted by most gene and protein databases (both general and species specific) as the main vocabulary for annotation.

One measure of GO's success is that it not only has been extensively adopted by the community for its designed purpose, but has also been used for other purposes beyond

it, such as functional comparison and function prediction of gene products.

The success of GO is due to two key factors. First, GO had a clear and practical goal, and a limited but useful scope. This helped in keeping it focused throughout its development and, above all, ensured its simplicity and usefulness. Second, GO is developed with the involvement of the community, openly addressing its needs. This contributes to make it accessible for the community it wishes to serve, and ensures that it is kept updated.

Despite this success, there is still room for improvement. Some authors suggest that a different model for representing the concepts may be required to deal with the growing compositionality of GO term names, while others have found dependence relationships between terms which are not accounted for or not possible⁵ in the GO structure.

Having been the first ontology of its kind, GO's success has led to a blossoming of the field of BioOntologies. The relative simplicity of GO is what makes it both useful and accessible to the community. Profound changes to GO should be considered with care, since having a perfect ontology is useless if its complexity is beyond the grasp of most of its user community.

Sequence Ontology

The Sequence Ontology⁶ (SO) was developed for annotating biological sequences in a genetic context (Eilbeck *et al*, 2005). It encompasses one main aspect, *sequence feature*, plus three others: *sequence attribute*, *consequences of mutation* and *chromosome variation*; these aspects describe properties of the main aspect at several levels. Like in GO, these three aspects constitute separate ontologies with a DAG

⁵ The GO structure does not allow relationships between terms of different GO types.

⁶ <http://www.sequenceontology.org>

structure, with terms linked by *is a* and *part of* relationships (see Figure 2). However, sequence features can also be linked to sequence attributes with the *has quality* relationship, and sequence features can be linked non-hierarchically with the *adjacent to* or *member of* relationship. Also, as GO, SO is mostly species-independent, although some terms can be limited to certain taxonomic groups (e.g. intron-related terms only occur in *Eukaryota*).

A subset of SO consisting only of *sequence features* is available under the name SOFA⁷, which can be used for automated sequence annotation whereas the full SO is intended to be used only by curated genome annotation projects.

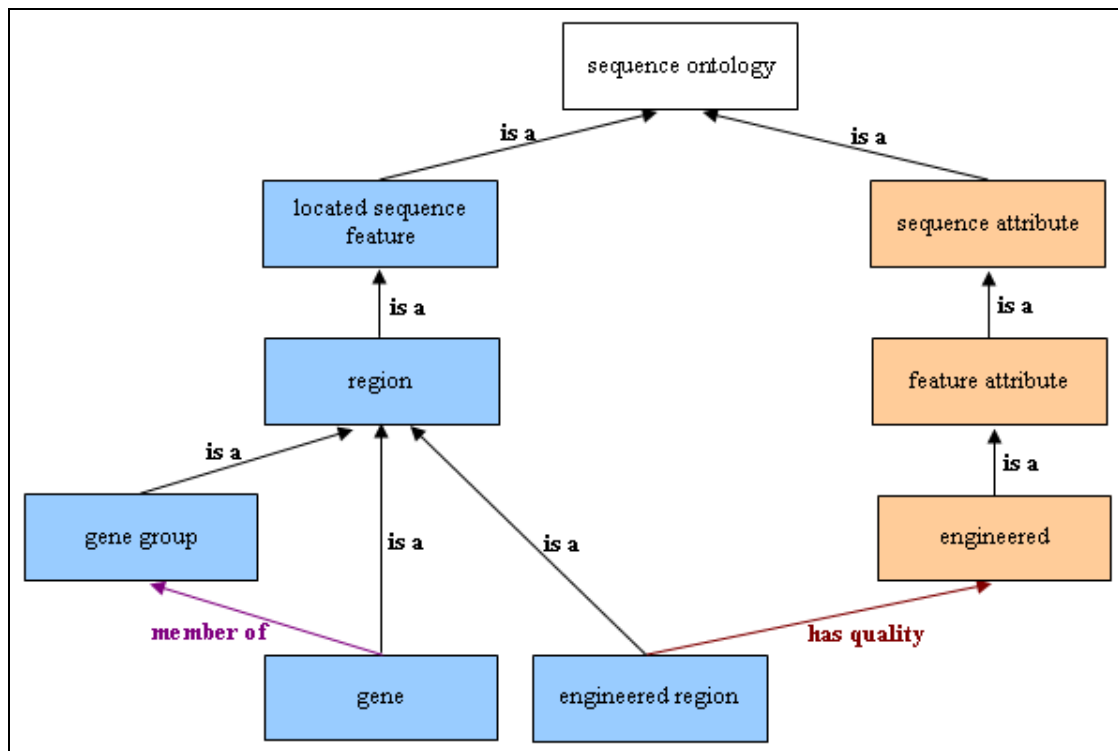


Figure 2: Section of the SO graph, showing the aspects *located sequence feature* and *sequence attribute*, and some of their descendent terms. The *has quality* relationship that links these two aspects is illustrated with the term *engineered region* which has the quality *engineered*.

⁷ Sequence Ontology Feature Annotation

SO was also developed by the GO consortium with the goal of unifying the vocabulary used to describe sequence features, facilitating information exchange and retrieval and enabling computational reasoning over sequence annotations. It is a natural complement to GO, with the two together accounting for a large portion of the biological aspects for which there is a need for annotations in a large scale.

However, when SO was developed, the main sequence databases (GenBank, EMBL and DDBJ) already had a well established terminology for sequence annotation (the *Feature Table*). This hampers SO's acceptance by the community. The main argument in favour of SO is that it provides an underlying structure for the annotations, whereas the *Feature Table* is a controlled vocabulary with no formalized structure. The underlying structure of SO greatly facilitates the use of computational tools for mining sequence data, and may lead to its increasing adoption.

MGED Ontology

The Microarray Gene Expression Data (MGED)⁸ Ontology (MO) was created for describing microarray experiments, encompassing all aspects from the methodology and experimental design to biological samples (Christian *et al*, 2003). MGED is divided in two parts, a core ontology (MCO) and an extended ontology, the former providing a stable basic structure to ensure continuous compatibility with software applications and the latter an extension that enables content evolution (see Figure 3).

MO has a simple structure consisting of several orthogonal trees corresponding to its various aspects, mostly linked by *is a* relationships that are relatively short in length (i.e. the number of levels between the root node and the leaf nodes is small). It

⁸ <http://www.mged.org>

includes three categories of descriptors: classes, like *Organism* and *BioMaterial*, which define the types of data required for describing the experiment; properties, like *has_disease_state* and *has_additive*, which relate classes to descriptors characterizing them; and individuals, like *cell_type* and *exon*, which instantiate the classes. Some classes, namely *Organism* and *Compound*, are instantiated from identified external resources. For describing biological sequences, the Sequence Ontology is directly used.

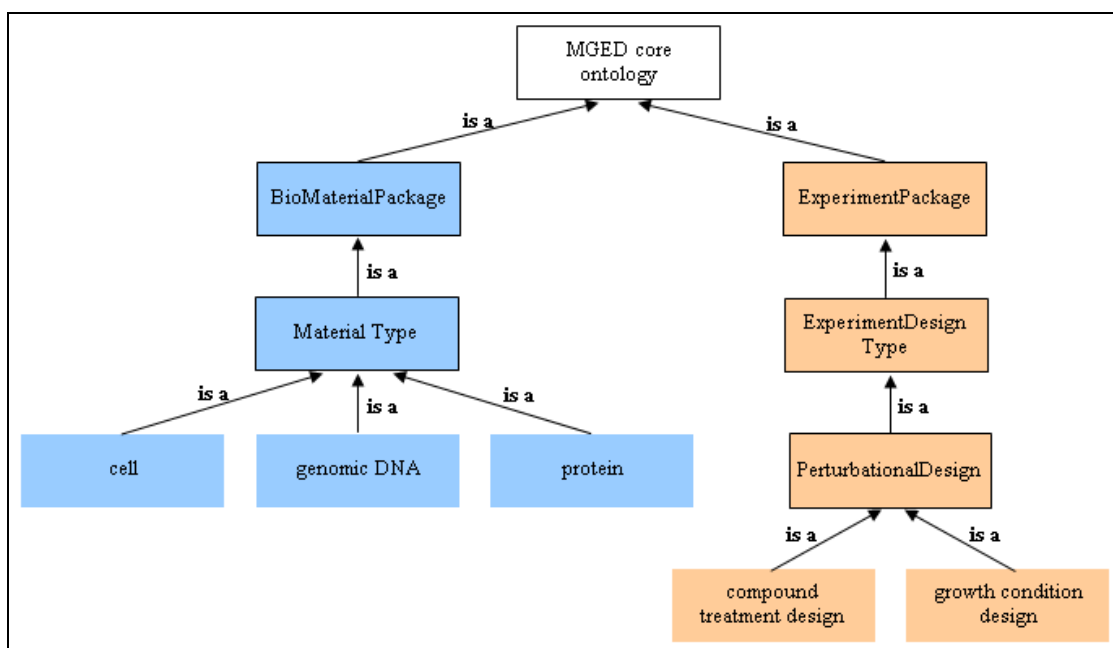


Figure 3: Section of the MCO tree, showing two of the main aspects (or packages), *BioMaterial* and *Experiment*. Boxes with frames represent classes, whereas boxes without frames, which are also leaf nodes, represent individuals. Only two or three class levels usually separate individuals from the root of the ontology, showing the relatively small height of the MCO tree.

MO was developed by the MGED Society with the goal of providing the required semantics to support the existing MAGE-OM⁹ data model, which already provided a standardized format for representing and exchanging microarray experiment data. In addition, it was designed to serve as a resource for the development of computational tools for mining microarray data.

⁹ MicroArray Gene Expression Object Model

Besides its wide use for describing microarrays, MO is also being used to describe other types of functional genomics experiments such as proteomics experiments. Being the first ontology describing a biological experiment, MO has paved the way for other related efforts, which lead to the creation of the integrative Ontology for Biomedical Investigations project (formerly FuGO¹⁰).

Unified Medical Language System

The Unified Medical Language System¹¹ (UMLS) is a compendium (or an integrated ontology) of text mining-oriented biomedical terminology encompassing all aspects of medicine (Bodenreider, 2004). It comprises three distinct knowledge sources: the *Metathesaurus*, the *Semantic Network*, and the *SPECIALIST lexicon*.

The *Metathesaurus* is an extensive, multi-purpose vocabulary database that integrates information from over one hundred clinical and biomedical databases and information systems, such as ICD, MeSH, SNOMED and GO. It defines biomedical concepts, listing their various names and relationships and mapping synonyms from different sources, thus providing a common knowledge basis for information exchange. It can be used autonomously for a variety of applications, namely linking between different clinical or biomedical information systems, and linking patient records to literature sources and factual databases. However, its utility is enhanced when used with the other UMLS knowledge sources. Since the *Metathesaurus* contains concepts and terms from diverse sources for diverse purposes, many specific applications require a customized reduced version of it, where only the areas of interest are included.

The *Semantic Network* is an ontology of biomedical subject categories (*semantic types*) and relationships between them (*semantic relations*) with the purpose of

¹⁰ Functional Genomics Ontology

¹¹ www.nlm.nih.gov/research/umls/

semantically categorizing the concepts from the *Metathesaurus* (each term in the *Metathesaurus* is linked to at least one *semantic type*). *Semantic types* are organized in a tree-structure with major types including *organism*, *anatomical structure*, *biologic function*, *chemical*, and *event*. The tree edges are labeled with the main *semantic relation*, *is-a*, although several other non-hierarchical semantic relations also exist, grouped in five major categories: *physically related to*, *spatially related to*, *temporally related to*, *functionally related to*, and *conceptually related to*.

The *SPECIALIST Lexicon* is an English language lexicon focused on biomedical vocabulary, but also including common English words. Each entry in the lexicon, or lexical item, includes syntactic, morphological and orthographic information, essential for natural language processing (NLP). This lexicon was developed to support an NLP system, also called *SPECIALIST*, which is available with the UMLS as a set of *lexical tools*.

The UMLS was developed and is maintained by the US National Library of Medicine, with its main goal being the improvement of accessibility to biomedical information by facilitating its interpretation by computer systems. It successfully addresses the problem of coping with the multiplicity of vocabularies and terminologies in use in medicine through an integrative approach (the *Metathesaurus*) and complements it with a semantic structure that facilitates computer reasoning (the *Semantic Network*) and lexical information. This enables NLP-based text mining tools to explore the biomedical literature (the *SPECIALIST Lexicon*). These three factors, together with the all-encompassing scope of UMLS, make it an invaluable tool for mining medical data in any of its aspects.

Open Biomedical Ontologies

The Open Biomedical Ontologies¹² (OBO) Foundry is a project dedicated to coordinating the development of new BioOntologies (Smith *et al*, 2007). It was established to deal with the growing number of efforts in the field that followed after the success of the Gene Ontology.

The OBO Foundry defines a set of principles to which new ontologies should adhere in order to be accepted as members of the project. These are set to ensure high quality and formal rigor, and also interoperability between OBO member ontologies. As of December 2007, the key principles enforced are:

- Open access: the OBO is intended to be a shared community resource, with all member ontologies openly available.
- Delineated content: every new ontology must be orthogonal to other OBO ontologies to avoid the problems of redundant information and conflicting definitions,
- Textual definitions: ontology concepts should include a precise textual definition to avoid ambiguity and convey the meaning within the context of the ontology.
- Defined relationships: the OBO Foundry includes the *Relation Ontology*, which defines the possible relationship types between the terms of member ontologies; OBO ontologies should use these relationships or others defined in a similar way.

¹² <http://obofoundry.org/>

- Shared syntax: OBO ontologies must be expressed or expressible in a common shared syntax, which can be either the OBO syntax developed by the project, or OWL, the web ontology language defined by the W3C.

The result of complying with these principles is that OBO ontologies have similar structural aspects and a shared syntax, which facilitates the integration of their information and makes the use of common software tools possible. Furthermore, by ensuring that the ontologies are orthogonal, redundancy is minimized and the problem of having concurrent definitions for the same concepts is avoided.

In this manner, the OBO Foundry project aims at preventing the problem that UMLS was designed to solve. By ensuring that new BioOntologies grow in concert with each other, no *a posteriori* integrative solution should be required.

TOWARDS AUTOMATIC ANNOTATION

One of the main applications of text mining and data mining in biomedical research is the automatic annotation of biological entities.

The main source of annotation data is the scientific literature, since text is still the preferred medium of communication among biomedical researchers. The main link between text and BioOntologies is a terminology where textual terms are associated to concepts in the BioOntology (Spasic *et al.*, 2005). However, two main issues arise when linking textual terms to ontologies: the imprecise and inconsistent use of terminology in text and the incompleteness of ontologies. There is a high degree of term ambiguity and variation in the biomedical field, often preventing a direct mapping between ontology concepts and terms in text. Term variation arises from the many synonyms that exist for gene products, diseases, etc, whereas term ambiguity

originates from the various sub-domains and niches inherent to the field, where terms can have different meanings depending on the context.

Text Mining (TM) can be used to extract relevant information from the scientific literature to aid in bioentity annotation. The most widespread use of TM in biomedical applications is on the retrieval of small chunks of relevant information from large collections of unstructured text (Couto and Silva, 2006). Typically, TM makes use of: Information Retrieval (IR), for retrieving and filtering of relevant texts; Information Extraction (IE), to select specific information about predefined entities; Natural Language Processing (NLP), to process natural language into a machine-readable form; and Machine Learning (ML), to classify, cluster and extract relations. All these approaches can benefit from the use of BioOntologies to assist in the semantic interpretation and integration of text.

IR tools are frequently used by the biomedical community (e.g. PubMed). However, it is important to take into consideration synonyms and polysemes, and not restrict IR to exact term matching, in order to achieve a balance between loss of information and loss of relevance. BioOntologies provide not only a semantic layer to define such cases, but also a hierarchical organization which allows expanded querying (e.g. retrieving documents that do not have the query term but one of its descendents or ancestors).

Biomedical IE can range from simple Named Entity Recognition (NER) to the more complex extraction of relations, networks, etc. NER is the identification and mapping of a term detected in text to a concept; since many term occurrences are variants, it is possible to use the list of terms present in a BioOntology to derive a training set to detect new terms. To extract more complex types of information, BioOntologies should be used beyond their lexical properties, guiding and constraining the semantic

analysis with their structural and relational properties.

The application of NLP to BioLiterature can profit from the integrated use of BioOntologies at different levels: tokenization (e.g.: recognizing “androgen receptor” as an entity, rather than two separate ones), syntactic processing (parsing the syntactic structure often implies the semantic relations between the concepts), sense disambiguation (referring to the definition of a term in a BioOntology can elucidate the correct meaning of the term in the text). In addition, the simultaneous use of NLP and BioOntologies allows higher quality inferences to be made, by translating the linguistic structures generated by NLP into an ontology-based schema with its finer-grained representation of knowledge (Friedman *et al.*, 2006).

BioOntologies can be used as training corpora for ML techniques, either as simple lists of classified terms, or making use of the relational and hierarchical information to perform clustering and classification.

Biomedicine is an inherently complex area and, as such, coherent and concise annotations of bioentities are crucial for computational reasoning. Traditional TM techniques have been shown to fall short of the biomedical community’s needs, performing worse than in other domains. To be successful, TM applications need to be supported by an explicit semantic representation of the kind provided by ontologies. Below, we describe four tools than can be used to retrieve and extract relevant information for annotation, all of which make use of BioOntologies.

GoPubMed

GoPubMed¹³ is an ontology-based literature search tool, which extracts GO terms from PubMed abstracts retrieved by keyword search. PubMed is a service of the U.S.

¹³ <http://www.gopubmed.org/>

National Library of Medicine that includes over 16 million citations from MEDLINE and other life science journals for biomedical articles back to the 1950s. PubMed includes links to full text articles and other related resources. The extracted GO terms are then used to induce a relevant and browsable sub-ontology, which allows for a quick navigation from general to more specific terms, due to the hierarchical nature of GO.

GoPubMed makes use of the Gene Ontology for two different tasks: GO term extraction, which uses an algorithm that explores the inherent characteristics of GO (hierarchy and substring relationships between terms), and the construction of the minimal sub-ontology that contains all the extracted terms. These allow enhanced keyword searches, which usually demand a good understanding of the domain to obtain good results. The technique allows detection of relevant keywords derived from GO, even when they are not mentioned in the articles.

The tool also enables exploration of the abstracts at different levels of detail by structuring them according to the induced sub-ontology, making the large amounts of information retrieved more manageable (Delfs *et al*, 2004).

GoPubMed refines traditional PubMed keyword search by incorporating domain knowledge from GO, gearing it towards the molecular biology community. It provides researchers with a more relevant and structured set of results, which could be overlooked when using PubMed queries.

Textpresso

Textpresso¹⁴ is an ontology-based information retrieval and extraction system for biomedical literature first developed for *C. elegans*. Instead of using an established

¹⁴ <http://www.textpresso.org/>

ontology, like GO, Textpresso uses its own ontology, which is organized into a shallow hierarchy with several parent categories of terms, some of which overlap GO and constitute the majority of the Textpresso lexicon. These categories are split into three groups: the first consists of biological entities, such as genes, cells or species; the second group contains terms that characterize a biological entity or establish a relation between two of them (e.g. binding, regulation); the third group contains auxiliary categories involved in semantic analyses of sentences.

Textpresso contains a collection of full-text scientific articles where each word or phrase is labelled according to the Textpresso lexicon, which makes it easier to query.

The search engine of Textpresso allows the user to formulate queries by combining keywords and Textpresso categories, which enables the formulation of semantic queries that impart much more meaning than simple keyword searches. The user can query against whole categories to retrieve all the information pertaining to a broad area, or he can combine keywords, categories and sub-categories to confine the search to a more specific theme. The categories that include entities and relationships enable the semantic contextualization of the query, whereas the auxiliary categories allow for a better retrieval of the relevant information from the texts. So the user queries the literature in the framework of the ontology and obtains sentences to be inspected. A typical result page shows a list of documents with all bibliographical information, abstract and all sentences having a match for an ontology term, links for the full-text available online; PubMed related articles are also provided when available (Müller *et al.*, 2004).

The main advantages of Textpresso lie on the use of the full-text of scientific articles, as well as on the possibility of building meaningful queries by the use of categories,

since each category corresponds in practice to a large set of keywords. However, Textpresso may be hindered by the lack of complete literature coverage, and the use of an ontology or lexicon that are not rich enough.

EBIMed

EBIMed¹⁵ combines document retrieval with co-occurrence based summarization of MedLine abstracts. The tool retrieves abstracts selected through keyword queries and filters them for biomedical terminologies maintained in different public bioinformatics resources: UniProtKb provides the terminology for proteins, Gene Ontology for describing cellular components, biological processes and molecular functions, MedLinePlus for identification of drugs and the NCBI taxonomy as terminology resource for species. EBIMed makes use of these resources as a simple source of terminology, any extra information that can be conveyed, such as relationships in the Gene Ontology and hierarchies in the NCBI taxonomy, is not used.

EBIMed looks for every UniProtKb protein in the text that co-occurs with another UniProtKb protein, GO term, drug or mention of a species because these can be interpreted as protein-protein interactions, functional annotations, drug targets and model organisms. The extracted sentences and terminology are used to generate an overview table of these paired co-occurring terms (Rebholz-Schuhmann *et al*, 2007).

The advantage of EBIMed is the extensive use of biomedical terminology resources to process PubMed abstracts and report associations between terms, thus giving an overview of a multitude of relations and organizing the information.

GOAnnotator

¹⁵ www.ebi.ac.uk/Rebholz-srv/ebimed/

GOAnnotator¹⁶ links the GO terms present in the uncurated annotations with evidence text automatically extracted from the documents linked to UniProt entries. The input to GOAnnotator is a UniProt accession number, which is used to access the bibliographic links in the UniProt database and retrieve the documents. Additional text for mining is retrieved from the GeneRIF database or supplied by the user. GOAnnotator then extracts GO terms from the documents and ranks them according to their similarity to the GO terms present in the uncurated annotations (see Figures 4 and 5).

GOAnnotator uses the Gene Ontology for two tasks: recognize terms in the text and, as a framework for calculating the semantic similarity between pairs of terms.

The extraction of GO terms is performed by FiGO, a rule-based method that does not use make use of NLP techniques and does not require manual intervention. FiGO assigns a confidence value to each GO term that represents the terms' likelihood of being mentioned in the text based on the nomenclature of GO.

GOAnnotator uses the Gene Ontology Annotation (GOA) database, which provides GO annotations to proteins in the UniProt Knowledgebase (UniProtKB) and International Protein Index (IPI). GOA is a central dataset for other major multi-species databases, such as Ensembl and NCBI. GOAnnotator ranks the documents based on the extracted GO terms from the text and their similarity to the GO terms present in the uncurated annotations, using the measure proposed by Lin (Lin, 1998). This measure combines GO hierarchy and term usage in the GOA database to achieve a measure of GO term semantic similarity (Couto *et al*, 2006).

GOAnnotator not only provides the evidence to support uncurated annotations, as well as predicts new and more specific annotations. It can also benefit from the

¹⁶ <http://xldb.fc.ul.pt/rebil/tools/goa/>

incorporation of other text-mining methods, since FiGO was not designed specifically for the extraction of annotations.











PubMedId	Title	MostSimilarTermExtracted	Scope	Authors	Year	Extract	AddText
11594756(FullText)	Distinct phosphoinositide binding specificity of the GAP1 family proteins: characterization of the pleckstrin homology domains of MRASAL and KIAA0538.	100% GTPase activator activity (f)	GeneRIF	3	2001		
11448776(FullText)	CAPRI regulates Ca(2+)-dependent inactivation of the Ras-MAPK pathway.	100% GTPase activator activity (f)	SEQUENCE FROM N.A.	3	2001		
9628581(FullText)	Prediction of the coding sequences of unidentified human genes. IX. The complete sequences of 100 new cDNA clones from brain which can code for large proteins in vitro.	28% cell communication (p)	SEQUENCE FROM N.A.	7	1998		
14702039(FullText)	Complete sequencing and characterization of 21,243 full-length human cDNAs.	-	GeneRIF	154	2004		
12853948(FullText)	The DNA sequence of human chromosome 7.	-	SEQUENCE FROM N.A.	107	2003		

Figure 4. GOAnnotator: Some of the documents retrieved for the protein Ras GTPase-activating protein 4. The documents are sorted by the most similar term extracted from their text. The curator can use the Extract option to see the extracted terms together with the evidence text. By default GOAnnotator only uses the abstract, but the curator can use the AddText option to replace or insert text.

Similar GO Terms Extracted	GOA Electronic Term: intracellular signaling cascade (p) [-]
inactivation of MAPK (p) [-]	CAPRI regulates Ca ²⁺ -dependent inactivation of the Ras-MAPK pathway Ca ²⁺ is a universal second messenger that is critical for cell growth and is intimately associated with many Ras-dependent cellular processes such as proliferation and differentiation [1].
protein kinase C activation (p) [-]	A role for intracellular Ca ²⁺ in the activation of Ras has been previously demonstrated, e.g., via the nonreceptor tyrosine kinase PYK2 [3] and by Ca ²⁺ /calmodulin-dependent guanine nucleotide exchange factors (GEFs) such as Ras-GRF [4]; however, there is no Ca ²⁺ -dependent mechanism for direct inactivation .
phosphoinositide-mediated signaling (p) [-]	Previously, we have shown that these C2 domains do not regulate Ca ²⁺ - mediated membrane association; instead, membrane targeting is mediated by phosphoinositide binding PH domains [11, 12 and 13].
Comment: <input type="text"/>	New Terms: <input type="text"/>
Evidence: [-] <input type="button" value="Add"/>	

Figure 5. GOAnnotator: For each uncurated annotation, GOAnnotator shows the similar GO terms extracted from a sentence of the selected document. If any of the sentences provides correct evidence for the uncurated annotation, or if the evidence supports a GO term similar to that present in the uncurated annotation, the curator can use the Add option to store the annotation together with the document reference, the evidence codes and any comments.

The use of BioOntologies

The tools reviewed above make use of BioOntologies in quite diverse manners (see Table 1). The most straightforward approach is implemented by EBIMed, where BioOntologies are used as a source of terminology to match the entities present in the literature. BioOntologies also provide evidence of association between different kinds of bioentities. Textpresso, on the other hand, uses its own built-in BioOntology to allow word meaning to be queried. The possibility of semantic query formulation enables the usage of this tool both as a search engine and as a curation tool. GoPubMed uses both the concepts and the structure of its BioOntology, not only matching the concepts to terms in the literature, but also exploring the hierarchical relationships among the retrieved terms to provide a multi-level navigation interface for accessing the retrieved texts at different resolutions. GOAnnotator goes beyond concept definition and ontology structure, by integrating both to calculate similarities between concepts. This enables this tool propose new annotations in addition to retrieving evidences from text support existing annotations.

	GOAnnotator	GoPubMed	Textpresso	EBIMed
Text Mining / NLP	Rule-based No NLP	NLP	Rule-based/ low NLP	NLP
Ontology	GO	GO	Textpresso Ontology	GO / NCBI taxonomy / MedLine Plus / UniProt
Usage of Ontologies	Extract terms and compute similarities	Creation of GO subontology based on extracted terms	Enhance keyword query	Retrieval of term co- occurrences
Goal	Retrieval of evidence for uncurated annotations	Construction of browsable relevant sub- ontology	Document and statement retrieval based on relevant ontology categories	Retrieval of MedLine statements

Table 1. Overview of the surveyed tools according to TM/NLP techniques, BioOntologies and goals.

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.

ACKNOWLEDGMENTS

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

Ashburner, M., Ball, C., Blake, J., Botstein, D., Butler, H., Cherry, J., Davis, A., Dolinski, K., Dwight, S., Eppig, J., Harris, M., Hill, D., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J., Richardson, J., Ringwald, M., Rubin, G. & Sherlock, G. (2000). Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature Genetics* 25(1):25-29.

Bodenreider, O. (2004). The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic Acids Research* 32(Database issue):D267-270.

Bodenreider, O. & Stevens, R. (2006). Bio-ontologies: current trends and future directions. *Briefings in Bioinformatics*, 7(3):256-274.

Christian, J., Stoeckert Jr & Parkinson, H. (2003). The MGED ontology: a framework for describing functional genomics experiments. *Comparative and Functional Genomics* 4(1):127-132.

Couto, F., Silva, M., Lee, V. Dimmer, E. Camon, E. Apweiler, R. Kirsch, H. & Rebholz-Schuhmann, D. (2006). GOAnnotator: linking protein GO annotations to evidence text. *Journal of Biomedical Discovery and Collaboration*, 20;1:19.

Couto, F. & Silva, M. (2006). Mining BioLiterature: Towards Automatic Annotation of Genes and Proteins. Hsu, H. (Ed), *Advanced Data Mining Technologies in Bioinformatics* (pp.283-295) Idea Group Inc.

Devos, A. & Valencia, A. (2001). Intrinsic errors in genome annotation. *Trends in Genetics*, 17(8):429-431.

Delfs, R., Doms, A., Kozlenkov, A. & Schroeder, M. (2004). GoPubMed: ontology-based literature search applied to Gene Ontology and PubMed. *Proceedings of German Bioinformatics Conference*. LNBI Springer, Bielefeld, Germany, pp.169-178.

Eilbeck, K., Lewis, S., Mungall, C., Yandell, M., Stein, L., Durbin, R. & Ashburner, M. (2005). The Sequence Ontology: a tool for the unification of genome annotations. *Genome Biology* 6(5):R44.

Friedman, C., Borlawsky, T., Shagina, L., Xing, H. & Lussier, Y. (2006). Bio-Ontology and text: bridging the modeling gap. *Bioinformatics*, 1;22(19):2421-2429.

Gardner, S. (2005). Ontologies and semantic data integration. *Drug Discovery Today*, 15;10(14):1001-1007.

Gruber, T. (1991) The role of common ontology in achieving sharable, reusable knowledge bases. Allen JF, Fikes R, Sandewall E (eds). *Proceedings of KR'1991: Principles of Knowledge Representation and Reasoning*. San Mateo, California: Morgan Kaufmann, pp. 601–602.

Ideker, T., Thorsson, V., Ranish, J., Christmas, R., Buhler, J., Eng, J., Bumgarner, R., Goodlett, D., Aebersold, R. & Hood, L. (2001). Integrated genomic and proteomic analyses of a systematically perturbed metabolic network. *Science*, 4;292(5518):929-934.

Lin, D. (1998). An information-theoretic definition of similarity. *Proceedings of the 15th International Conference on Machine Learning* pp. 296-304.

Müller, H., Kenny, E. & Sternberg, P. (2004). Textpresso: An Ontology-Based Information Retrieval and Extraction System for Biological Literature. *PLoS Biology*, 2(11) e309.

Rebholz-Schuhmann, D., Kirsch, H., Arregui, M., Gaudan, S., Riethoven, M., & Stoehr, P. (2007). EBIMed--text crunching to gather facts for proteins from Medline. *Bioinformatics*, 15;23(2):e237-244.

Smith, B., Ashburner, M., Rosse, C., Bard, J., Bug, W., Ceusters, W., Goldberg, L., Eilbeck, K., Ireland, A., Mungall, C., The OBI Consortium, Leontis, N., Rocca-Serra, P., Ruttenber, A., Sansone, S., Scheuermann, R., Shah, N., Whetzel, P. & Lewis, S. (2007). The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. *Nature Biotechnology* 25(11):1251-1255.

Spasic, I., Ananiadou, S., McNaught, J. & Kumar, A. (2005). Text mining and ontologies in biomedicine: making sense of raw text. *Briefings in Bioinformatics*, 6(3):239-251.

Stevens, R., Goble, C. & Bechhofer, S. (2000). Ontology-based knowledge representation for bioinformatics. *Briefings in Bioinformatics*, 1(4):398-414.

Valencia, A. (2005). Automatic annotation of protein function. *Current Opinion in Structural Biology*, 15(3):267-274.

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.