Ontology Consolidation in Bioinformatics

Sven Hartmann[1] **Henning Köhler**[2] **Jing Wang**[3]

[1] Clausthal University of Technology, Germany
Email: sven.hartmann@tu-clausthal.de

[2] The University of Queensland, Brisbane, Australia
Email: henning@itee.uq.edu.au

[3] Massey University, Palmerston North, New Zealand
Email: j.w.wang@massey.ac.nz

Abstract

Ontologies enjoy increasing popularity among bioinformatics researchers who seek assistance in coping with the rapid upgrowth of biological data to be handled and related information to be considered. While communities of committed ontology pioneers drive the development and optimisation of ontologies for a wide range of relevant domains, the amble adoption of ontologies in bioinformatics research is still decelerated by technological, managerial and communication barriers. This paper contains a brief review of current trends, challenges and perspectives in ontology research and practice in bioinformatics.

Keywords: biomedical ontology, ontology uptake

1 Introduction

The advent of high-throughput technology in analytical research laboratories calls for high-throughput tool support that enables bioinformatics researchers to keep pace with processing and analysing ever-increasing amounts of new biological data, and to draw new insights from them. Ontologies guide this process as knowledge bases that comprise existing knowledge about some subject of interest in a systematic and unequivocal way. This helps the bioinformatics community to create a common understanding of the subject and supports bioinformatics researchers in performing a variety of tasks.

In bioinformatics the notion 'ontology' is used to refer to a variety of modelling artifacts, ranging from controlled vocabularies, thesauri and taxonomies to conceptual schemas for particular kinds of biological data. Common to them is the aim to provide a set of concepts that can serve as abstractions for biological entities that are of interest for some application. Typically concepts have a name, a definition and a list of synonyms. Concept names are frequently called terms. The distinction between concepts and terms in not always explicit. Often multiple synonymous terms (including abbreviations and acronyms) are used to refer to the same concept. For convenience, concepts are often equipped with a unique id (or accession number) that can be used for identification and for cross-referencing. Definitions can be textual descriptions or by listing properties that characterise the concept. For an example, see Fig. 1.

Concepts are often organised in hierarchies, as known from taxonomies. Subsumption is most frequent, but meronomy is common, too. More advanced ontologies capture other kinds of relationships, too, that serve as abstractions for further associations between biological entities. For a survey of relationships used by ontologies in bioinformatics we refer to (Smith & al. 2005), and for a brief historical outline to (Bodenreider & Stevens 2006).

AccessionNo	GO:0000001
Name	mitochondrion inheritance
Definition	The distribution of mitochondria, including the mitochondrial genome, into daughter cells after mitosis or meiosis, mediated by interactions between mitochondria and the cytoskeleton.
Synonyms	exact: mitochondrial inheritance

Figure 1: An example of a concept from the Gene Ontology (biological process).

2 Ontologies for Bioinformatics Researchers

Primary sources for bioinformatics researchers are the National Center for Biomedical Ontology (NCBO) and the European Bioinformatics Institute (EBI). As of November 2009 the NCBO BioPortal[1] lists 179 ontologies with a total of 1,438,792 concepts, while the Ontology Lookup Service (OLS)[2] hosted by EBI lists 74 ontologies with a total of 923,606 terms. Both include the about 60 ontologies of the Open Biomedical Ontologies (OBO) foundry[3]. The OBO foundry is a special interest group bringing together ontology providers from the life science disciplines. Major objectives are to establish common design principles for and to foster inter-operability between ontologies (Smith & al. 2007).

Ontologies used in bioinformatics differ with respect to their scope, size and granularity. Some ontologies are devoted to anatomy and physiology of specific organisms, some provide the means for describing biomedical resources and experiments in analytical research labs. Other ontologies are devoted to related areas like health care and medical applications, e.g., disease ontologies. In addition to open ontologies there are proprietary ontologies. Many of them are just used by a single research group, project or tool. Some are publicly available, but may be subject to intellectual property restrictions.

In the next section we will briefly outline a few established OBO ontologies for illustration.

[1] bioportal.bioontology.org/
[2] www.ebi.ac.uk/ontology-lookup/
[3] www.obofoundry.org/

2.1 Examples of Ontologies in Use

Gene Ontology (GO) is probably the most popular ontology in bioinformatics. Its origin dates back to 1998 when several model-organism database projects (initially yeast, fruit fly and mouse) noticed that a common vocabulary will improve inter-operability across databases and simplify data integration (Ashburner & al. 2000). Gene Ontology assembles concepts that serve for classifying gene products according to what they do, where they act and how they perform these activities. It actually comprises three separate ontologies, one for molecular functions, one for cellular components, and one for biological processes. In addition to subsumption (is-a) and meronomy (part-of) it captures a third kind of relationships (regulates) for describing interactions between biological processes and other biological processes, molecular functions or biological qualities.

Most biological databases used by bioinformatics researchers for their investigations contain data at the molecular level. *Sequence Ontology (SO)* provides concepts for describing the features and attributes of biological sequences (Eilbeck & al. 2005). This includes biological, biomaterial and experimental features. SO captures further kinds of spatial relationships, relationships for locating features on sequences.

Molecular-level data studied in research labs is increasingly linked to concepts reflecting the global structure and anatomy of organisms. Suppose we want to know which genes influence the development of organs and tissue, or which genetic mutations cause deviations from the standard phenotype. Then it helps to find out which genes are expressed at which development stages and in which parts of an organism. The *Foundational Model of Anatomy Ontology (FMA)* is a comprehensive ontology of human anatomy, with more than 75,000 concepts for describing elements of human morphology, anatomic structures and the organisation of the human body (Rosse & Mejino 2003). It uses about 170 different kinds of relationships, e.g. to describe spatial associations between anatomic structures like organs and tissue.

Plant Ontology (PO) assembles concepts that serve for classifying gene products according to morphological and anatomical criteria (Jaiswal & al. 2005). It comprises two ontologies, one for plant structure (anatomy and morphology) and one for growth and development (growth stages in a plant's life cycle and developmental stages of plant structures). Besides subsumption (is-a) and meronomy (part-of) PO also captures a kind of temporal relationships (develops-from) for describing associations between derived structures and progenitor structures.

2.2 Ontology Formats

As Gene Ontology is the longest established ontology in there, most of the OBO ontologies are represented in the native format of GO, also known as the OBO flat file format (OBOF). The OBO repository also contains a translation of FMA into OBOF. This translation, however, omits all kinds of relationships other than is-a, part-of and has-part, as they do not yet conform to the guidelines set by the OBO foundry for relationship specification (Smith & al. 2005). The primary format of FMA is frame-based and can be manipulated with Protégé which is probably the leading general-purpose ontology editor available today.

As a matter of fact, tool support is crucial for the uptake of ontologies by application developers and bioinformatics researchers. A range of tools is available for processing OBO flat files, e.g., ontology browsers like AmiGO (Carbon & al. 2009) and ontology editors like DAG-Edit and OBO-Edit (Day-Richter & al. 2007). These tools were designed particularly for use in bioinformatics, having the needs of bioinformatics researchers in mind. Today there is a clear tendency to use the Web Ontology Language (OWL) for representing ontologies. It enables the use of a variety of tools developed for a wider user group, provides better support for capturing advanced ontology elements like axioms and constraints, and unlocks new application areas such as ontology reasoning and the semantic web, cf. (Golbreich et al. 2008).

In the last few years many of the OBO ontologies have been translated into OWL and are now available in both formats. The NCBO has recently agreed on a mapping between OBOF and OWL (Moreira & al. 2009) and released a conversion tool that can be embedded into OBO-Edit and Protégé for import/export (Moreira & Musen 2007). OBO Explorer (Aitken et al. 2008) is a recent graphical ontology editor based on Protégé and the NCBO conversion tool that allows native access to both formats. As OBOF does not have a formal grammar, the intended semantics of concepts and relationship types had to be mapped. On that occasion, a number of modelling sins were observed that called for further clarification, cf. (Aranguren & al. 2007), some of them are still to be addressed. For a discussion of naming conventions in OBO and OWL see (Schober & al. 2009). Even worse, many current OBO concepts and relationships lack information essential for formal definition in OWL. Other common formats that are in use for exchanging ontologies are plain XML and RDF/S.

2.3 Access to Ontologies

Though most established ontologies are equipped with dedicated web sites, bioinformatics researchers often prefer web portals with an integrated web interface for browsing and querying all ontologies of interest. Examples of such web portal are the NCBO BioPortal (Noy & al. 2009) and the Ontology Lookup Service (OLS) at EBI (Cote & al. 2008). They can be used to search for particular terms in any of the participating ontologies, to retrieve descriptions and additional meta-data stored with the concepts, and to navigate through the ontology along relationships. Both use a relational DBMS as backend to store the concepts and relationships of the participating ontologies. To keep the database up-to-date the ontology providers are polled on a regular basis, updated files are downloaded and scanned for changes.

A major advantage of web portals like OLS and BioPortal is that they greatly reduce the time needed to familiarise oneself with a query interface and to actually execute queries against multiple ontologies. This allows users to consider a larger set of ontologies at no extra costs. Query results are typically presented in a unified format, thus simplifying further processing. While standard tasks like browsing ontologies and detecting ontology-enriched databases, there is currently only limited support for other routine tasks. More usable web applications would enable bioinformatics researchers to make better use of online information. This includes the implementation of ideas from social software (e.g. community feedback, community-based ontology evaluation, collaborative maintenance of ontology mappings), software customisation, service and grid computing, cf. (Li & al. 2007). Both, OLS and BioPortal can be accessed by client applications as web services. They come with standard WSDL descriptions, thus making it easy to invoke them in scientific workflows.

For most routine tasks like data analysis bioinformatics researchers access ontologies through web-based or desktop tools that are integrated with one or more ontologies. Often such tools come along with a

backend database that stores the ontology, or they can be connected to a user-defined database. For a recent survey of 68 ontology-based analysis tools, see (Huang et al. 2009). Many of them still rely on Gene Ontology only, but considerable efforts are underway to better integrate other ontologies, too. For examples, cf. (Khatri & al. 2007, Sherman & al. 2007, Draghici et al. 2006, Kirov & al. 2005).

2.4 Storing Ontologies in Local Databases

If ontologies are frequently used it is worth considering to store them locally. Potential benefits are faster response times to queries, and the option to flexibly use an ontology with available software tools, to keep ontologies together with the actual annotations of biological data, to store several versions of an ontology, to extend ontologies by own concepts, descriptions or other useful meta-data. On the downside, however, local storage increases IT support costs and requires the implementation of a suitable update strategy.

Many providers of OBO ontologies offer users to download their ontology in OBOF or OWL, but also as an SQL dump file. This includes GO, PO and SO. For details see (Harris & al. 2005). OLS offers SQL export of their entire internal term database.[4]. FMA supports SQL export, too. Its authors, however, recommend to use the SQL dump with Protégé. For a detailed discussion of how ontologies are kept in relational databases, we refer to (Keet 2006). Alternatively, if ontologies are offered in XML format for download then XML-enhanced database servers like DB2, Oracle or Tamino can be used to store the ontology, too.

A promising approach to enable the use of multiple ontologies is Chado (Mungall & Emmert 2007) which was developed by the Generic Model Organism Database (GMOD) project[5]. It offers a large relational database schema that is centred around ontologies. It predefines generic tables for flexible storage of ontology concepts and relationships, and can be extended by tables for storing biological data linked to them. For example, there are is a special table collection for storing sequence features. Biological databases that conform to this schema can easily inter-operate with one another, and can be used with software from the GMOD toolkit, e.g., the sequence viewer GBrowse (Donlin 2009). For some major ontologies like GO, SO and PO, Perl scripts are available for translating them into the Chado format.

3 Use of Ontologies

In analytical research labs a good deal of time is spent with acquiring, storing, retrieving and analysing biological data obtained by running experiments or by querying biological databases. Bioinformatics researchers can use of ontologies for all these tasks. For a recent survey of such activities we refer to (Bodenreider 2008, Rubin et al. 2007).

3.1 Annotating Data

Terms (or the ids of the corresponding concepts) can be used to annotate biological data in a consistent manner. Semantic annotations are assertions that link biological entities in the application domain to concepts in the ontology. Annotations are useful for a variety of routine tasks in analytical research labs, such as querying databases or analysing data sets. Often annotations are created on the basis of the data contained in experimental reports, primarily those published in the literature and available from digital libraries for the life sciences, such as PubMed[6]. The annotations may be stored directly in the biological database, or in a separate annotation database together with the ontology. A recent example of annotating organism-specific data is given in (Beck & al. 2009). For a brief summary of where annotations come from and what they mean, see (Hill et al. 2008).

A range of model-organism and multi-species databases have their data enriched by annotations. This includes all major biological databases, such as UniProt or Ensembl. To assure the correctness of annotations, publicly available biological databases often have expert curators in place for this task. The annotation of biological data using ontologies is a time-consuming exercise as it is performed manually for the most part. Empirical studies show that manual curation is far too slow (Baumgartner & al. 2007). Some approaches have been proposed for automated annotation, cf. (Doms & Schroeder 2005, Vinayagam & al. 2006, Couto & al. 2006, Daraselia & al. 2007). So far none of the underlying annotation algorithms works accurately enough to fully compensate human curators. This is partly due to the lack of benchmarks and reliable training data. High-throughput data analysis, however, will eventually require automated annotation to cope with the amount of data to be processed. For the related task of biological text mining, we refer to (Chapman & Cohen 2009). This can help to automatically detect mentions of relevant concepts or interesting biological entities in scientific publications.

3.2 Querying and Analysing Data

Once the data in a biological database is annotated the annotations can be used for queries. For example one can ask for all biological entities linked to a particular concept, or for all concepts a biological entity is linked to, or for biological entities that are linked to the same concepts. Using the appropriate terms a bioinformatics researcher can search, for example, for all genes that are involved in a particular biological process or are expressed in a particular plant structure. Search criteria may be combined, filters may be set, and aggregation functions like counting may be used to form more involved queries. Search results can be ranked by the extend to which they match non-boolean search criteria. Queries can also be combined with similarity search tools like BLAST. The best BLAST hit method (Jones et al. 2005), for example, finds the concepts that the BLAST top hit of a search sequence is linked to. A popular web-based query tool is AmiGO (Carbon & al. 2009).

Experiments undertaken in analytical research labs produce data sets to be analysed. Studying the annotations available for them may offer new insights into the potential biological meaning of an experiment. If an experiment generates a list of genes as output, one may want to identify those concepts that are most characteristic for the genes in the list. For that, over-representation of concepts is determined in the context of some background superset, such as all the genes on a microarray chip, or all the genes in a genome. Such an analysis takes into account annotations that are inherited by transitivity of certain kinds relationships (is-a, part-of), cf. (Grossmann et al. 2007). Generally speaking, a concept may help to differentiate the genes in the list from the rest if the observed over-representation is statistically significant. Similarly one may look for characteristic relationships between the genes in such lists and other biological

[4] www.ebi.ac.uk/ontology-lookup/implementationOverview.do
[5] www.gmod.org
[6] www.ncbi.nlm.nih.gov/pubmed/

entities. Ontology-based data analysis can help to validate hypotheses about the domain of study and make implicit knowledge explicit. In the literature a wide range of analysis tools has been suggested to support bioinformatics researchers in answering such questions. We refer the reader to (Huang et al. 2009, Khatri & Draghici 2005) for a detailed discussion of the scope and capabilities of ontology-based analysis tools, including the statistical models, ontologies, and data mining algorithms deployed.

More recently, scientific workflows have emerged in analytical research labs to streamline and automate data analysis, cf. (Cure & Jablonski 2007, Ram et al. 2008). Popular tools for designing and executing such workflows are Kepler and Taverna (Hull & al. 2006).

3.3 Exchanging and Integrating Data

An important aspect of developing and using ontologies is that concepts need to be identified, named, and enhanced by descriptions. In life sciences there are often several terms in use that refer to the same or similar concepts. Descriptions should help to decide whether terms refer to the same, related or different concepts. As bioinformatics researchers exchange annotated data and retrieve annotated data from different biological databases it is advantageous to use terms consistently for annotations or at least to keep track of terminological associations like synonymy, abbreviations or acronyms. Commonly agreed ontologies help achieving that goal. Currently a large portion of biological data in publicly available databases is annotated with free-text fields that are not yet associated with ontology concepts (Shah & al. 2009). Empirical studies demonstrate benefits of ontology-based over free-text search (Moskovitch & al. 2007). Automated mappings of free-text terms to concepts have studied in (Dai & al. 2008).

The thorough analysis of new experimental data requires bioinformatics researchers to consult multiple data sources such as local databases with outcomes of earlier experiments and studies, and publicly available databases with reference data for comparisons. Relevant biological data is spread across and disseminated through an ever-increasing number of databases. Ontologies provide common languages to overcome semantic heterogeneity among independently designed and maintained biological databases, enable data exchange between different research groups, foster inter-operability between diverse databases, and detect distributed data that is interrelated by associations between the corresponding biological entities. Ontologies can guide data transformations for data warehouses, but also query transformations for mediation-based integration, cf. (Hernandez & Kambhampati 2004, Bodenreider 2008). OntoFusion (Perez-Rey & al. 2006) is an example of a mediation tool that exploits Gene Ontology. For a recent case study on ontology-driven data integration, see (Smedley & al. 2008).

When data from diverse sources is integrated data inconsistencies are likely to occur. These inconsistencies can be due to inaccurate and missing data, or due to wrong decisions made during integration. Data analysis based on inconsistent biological databases may lead to uninteresting and wrong results. Ontologies can help to discover such inconsistencies in advance (Chen et al. 2007).

3.4 Representing and Sharing Knowledge

Ontologies offer a conceptualisation of a scientific domain in a formal and unambiguous way. Their formal specification makes the available knowledge about the domain explicit in a form that is accessible to both humans and machines. Ontologies enable researchers to describe the entities in the domain consistently. This forms the basis for a shared and commonly accepted understanding of that domain. Ontologies are consistent but not necessarily complete. So they are challenged each time when new information about the underlying domain becomes available, and updated if necessary. Due to the rapid growth of information this happens frequently in bioinformatics.

4 Development of Ontologies

4.1 Ontology Design

Most of the established ontologies in bioinformatics were originally designed by hand. Since then, however, a range of methodologies and tools have emerged that aim to support ontology designers in developing and maintaining ontologies. In the literature a few design methodologies have been proposed, cf. (Cristani & Cuel 2005, Kamel et al. 2007, Darlington & Culley 2008) but none of them is widely known nor used. In practice new ontologies are usually created in an iterative refinement process, similar to other modelling artifacts like database schemas. For an interesting discussion of differences between ontology and database schema design, we refer to (Noy & Klein 2004). The generic design principles (Smith & al. 2007) of the OBO foundry provide arguably the most influential guidance for the design of ontologies in bioinformatics, even though not all of the principles are well-justified, e.g. the call for a unique root element.

Recurring steps in the design process are the creation of new concepts (including names, description, and synonyms) and the specification of relationships to other concepts (including the placement of the concept in is-a and other hierarchies). Design decisions need to be checked against already existing parts of the ontology, e.g., name clashes should be avoided and descriptions should be consistent with specified relationships. Ontologies can be built bottom-up with the most specific concepts first, or top-down staring with the most general ones, or both approaches may be mixed. Some ontologies were built in a modular fashion that allows the cooperation of several designers or the integration of already existing ontologies or parts thereof, cf. (Thomas & al. 2006, Pathak et al. 2009). Design patterns were exploited in (Aranguren & al. 2008) for designing the *Cell Cycle Ontology (CCO)*. Some ontologies were built from database schemas of databases to be annotated, cf. (Lubyte & Tessaris 2009, Zhao & Chang 2007).

4.2 Ontology Evolution and Maintenance

High-throughput analysis of biological data results in new observations and insights that contribute to our understanding of the respective application domain. Ontologies need to be updated to keep pace with the new knowledge. Typical update operations are the addition of new concepts and relationships to meet new requirements or to reflect new insights, the modification of exiting concepts and relationships if they need to be clarified or adapted to stay consistent with new insights, or the deletion of outdated concepts and relationships that are no longer needed or in line with our understanding. For example, one might want to add a new intermediate concept into the is-a hierarchy to support a new abstraction level. To implement such an update one needs to name and describe the new concept and to decide where to insert it.

Updates in ontologies may be motivated by the desire to coordinate the development of ontologies as promoted by the OBO foundry. When ontologies

cover related domains it is likely that they possess common or similar concepts. There are plenty of reasons to make such implicit connections explicit. This can be done, for example, by unifying concepts or by defining mappings between matching concepts. Discovering such connections is by far not trivial and often hampered by semantic heterogeneity or different levels of granularity across ontologies and the scientific communities using them. There are tools available that compute mappings automatically, cf. (Ghazvinian et al. 2009, Jean-Marya et al. 2009) but in most cases human expertise is needed.

Once ontologies are in use the evolution process needs to be adequately documented and managed. Updates of ontologies are likely to affect exiting annotations and exiting mappings to alternative formats or to other ontologies. Updates need to be propagated effectively and potential problems need to be resolved, e.g. in case of modifications and deletions that may impair previously correct annotations. This results in a trade-off between robustness and up-to-dateness. While new knowledge should be incorporated as soon as possible to be ready-to-use, existing tools and applications require a certain level of stability to be useful and constrain maintenance costs. Due to the dynamic nature of the represented knowledge, many ontology providers release separate versions of their ontology at regular intervals. Updates are accumulated in a new release of the ontology while previously released versions stay unchanged. For a systematic study ontology evolution, including examples and consequences for annotations and ontology mappings, we refer to (Hartung et al. 2008).

Most established ontologies in bioinformatics have seen continuous change since their creation. In Gene Ontology, for example, the number of concepts has reduplicated since 2002. OnEX (Hartung et al. 2009) is a simple web-based tool for exploring the evolutionary aspect of major ontologies in bioinformatics. A more generic approach towards monitoring ontology evolution is presented in (Park et al. 2008). Traditionally, many OBO ontologies had a strong focus on concepts while relationships were often neglected. Considerable efforts are underway to overcome this situation, cf. (Smith & al. 2007). Following the recommendation of the OBO foundry, logical definitions are provided for relationships, and relationships are defined at both concept- and instance-level when appropriate. We also refer to (Eilbeck & Mungall 2009) for the particular example of Sequence Ontology.

The maintenance of established ontologies is in most cases done by dedicated teams of ontology curators who retrieve update proposals from the literature and from the scientific community, reach consensus in case of conflicting proposals, verify and approve updates, and generate new releases of the respective ontologies. This is time-consuming and requires highly specialised experts, however, many ontology providers consider this necessary due assure a reasonable level of quality and scientific rigour. Automated extension has also been suggested and first experiences recorded, e.g. for extending Gene Ontology using text mining and ontology matching techniques to extract new terms, concepts and relationships (Pesquita et al. 2009). For a generic approach towards automated ontology integration, alignment and extension, see (Novacek et al. 2009).

5 Trends, Challenges and Directions

5.1 User Involvement

Despite the recent growth in the number, size and coverage of ontologies in bioinformatics they are still far from unfolding their full potential in bioinformatics research. This is mainly due to communication barriers between different communities and research groups. Still today many bioinformatics researchers come in touch with ontologies only as part of analysis tools. However, there is an increasing interest and awareness of ontologies among researchers. Yet only few of them know how precisely ontologies can accelerate their work. As web portals and tools for querying annotation databases mature their users will become increasingly curious about the scope and granularity of the underlying ontologies.

Most bioinformatics researchers have their focus on a small, individualised portion of the huge multi-faceted life science domain, however, in their sub-domain they are interested in highly specialised knowledge. Browsing the integrated term databases underlying web portals like OLS or the NCBO BioPortal can be confounding due to their sheer complexity. On the other hand, selecting particular ontologies that better meet individual needs is not an easy exercise either. Often there is not enough information available to make an informed choice. Even when ontology providers offer detailed documentation and online tutorials this will only be perceived useful later on. For efficient decision-making most users will prefer more condensed information.

As argued in (Tan & Lambrix 2009) ontology selection relies on evaluation and comparison of candidate ontologies. Ontologies can be evaluated against a catalogue of human-made criteria (Lozano-Tello & Gomez-Perez 2004) or exploit statistics about the ontology (Gangemi et al. 2006) and its usage (Maiga & Williams 2008, Porzel & Malaka 2004). None of these approaches is perfect. Community-based assessment, collaborative filtering and recommender services as known from social networks and Web 2.0 applications can help to tackle that problem. Social functionality like community feedback and peer-review as recently added to the BioPortal is likely to improve the uptake of ontologies and ontology-based tools. Ontology pioneers can act as multipliers in these user communities.

When using ontologies bioinformatics researchers will find it helpful if they are tailored to their particular needs. It might be desirable to have user- or application-specific ontologies at hand that cover just the sub-domain of interest and are sufficiently fine-grained. An application of such personalised ontologies for information extraction is discussed in (Tao & Embley 2007). Specialised ontologies may be extracted and further evolved from existing ontologies that continue to serve as reference ontologies. Users are domain experts and will play an active role in the generation of such specialised ontologies. Since bioinformatics researchers have only limited experience with ontology evolution and maintenance a new generation of ontology editors will be required to keep the learning curve flat. Eventually this may lead to the creation of a collection of coupled ontologies centred around a master ontology that need to be harmonised on a regular basis.

The development of specialised ontologies may again be a collaborative effort of research groups or small communities with a common interest in certain sub-domains. Some first experience with collaborative ontology evolution using an extension of the ontology editor Protégé have been reported in (Tudorache et al. 2008), and using a wiki-based tool in (Hoehndorf & al. 2009). To foster reuse of specialised ontologies it would be helpful to set up a dedicated ontology registry that allows other researchers to discover existing ontologies that best meet their needs.

5.2 Enhancing Ontologies for Reasoning

Ontologies as formal specifications of knowledge are amenable to automated reasoning. Reasoning tools can help in developing and maintaining ontologies, but also in making inferences from existing knowledge, and checking hypotheses against ontologies. Ontology design often focusses on concepts, but relationships and axioms deserve more attention as they represent valuable knowledge, too. The popularity of OWL is partly due its expressiveness that allows one to capture more of the semantics of the underlying domain than other ontology formats like OBOF. Thus OWL promotes the creation of more precise and comprehensive artifacts. In particular, it provides powerful means for specifying axioms, even if we restrict ourselves to the OWL-Lite or OWL-DL fragments of OWL that guarantee decidability in reasoning. These fragments are also well-supported by reasoning tools such as FaCT++, Pellet and Racer that all can be integrated with Protégé. Potential applications of reasoning in ontology design were presented in (Lutz & al. 2006), and requirements for reasoning in bioinformatics were discussed in (Keet et al. 2007).

Typical examples include existence, disjointness, covering, cardinality and universal constraints, transitivity and symmetry. (Aranguren & al. 2007) claims that biologists would not widely use axioms as that would require more knowledge than is usually available. We rather believe that axioms can be beneficial in answering questions that go beyond simple instance or relationship checking. Clearly it will take considerable efforts to add information that is currently missing, e.g. by reinspecting annotations or even rerunning experiments. Relationships and axioms can impose necessary and/or sufficient conditions on entities to be linked to certain concepts. For example, there has been a long discussion about the meaning of the part-of relationships in Gene Ontology and others. Axioms can reflect that entities linked to a particular concept must occur in a certain relationship, or that occurrence is optional. Reasoning helps ontology designers to understand the consequences of design decisions. Careful reconsideration of existing ontologies helps to detect inconsistencies, modelling errors and gaps. Automated reasoning can guide this process and provide informed feedback for ontology repair. Reasoning helps bioinformatics researchers to think about what new observations mean in the context of what is already known. Reasoning with adequately enhanced ontologies offers better support for objectives like computing derived relationships, generating new hypotheses, consistency checking, discovering equivalence or similarity among concepts, reformalising ill-specified domains, reusing ontologies, or extracting and evolving specialised ontologies.

5.3 Automation

Bioinformatics researchers will allocate increasing portions of their time to ontology engineering rather than data engineering. For that, however, they need to be released from tedious tasks like data processing or experimentation. The combination of high-throughput experimental devices and analysis tools is a first step to automatise many routine tasks that results in a very large number of experiments that can be executed simultaneously. Data analysis workflows can be executed in flexible ontology-based workflow tools that integrate experimental data with other data resources to validate hypotheses. So far, however, the design of experiments and workflows is still done by human researchers. Experience tells that these tasks are tedious, too, in particular when they need to be reiterated several times.

A major step towards lab automation are robot scientists as presented in (King & al. 2009). The prototype system has been used to refine hypotheses about yeast and test them by conducting the respective experiments. The operation is fully automatised except for the periodic replacement of lab consumables. The system uses an in-house developed ontology that has been evolved from the *Ontology of Scientific Experiments (EXPO)* (Soldatova et al. 2006). EXPO captures generic knowledge about experiments and is available in OWL, thus enabling automated reasoning. (Epple & Scherf 2009) reports on an ontology-based expert system that implements a strategy for building hypotheses from inspected data. Automated hypothesis generation may guide bioinformatics researchers to experiments that will result in potentially interesting observations. In the near future we will see more examples of hypothesis-led investigations. For that, however, ontologies are required that capture the concepts and principles of human knowledge discovery.

6 Discussion

Ontologies in bioinformatics are still 'work-in-progress'. Over the last few years ontologies have gradually transformed the way bioinformatics researchers approach experiments, analyse biological data and generate new knowledge. Active involvement of users enabling them to adopt ontologies for their own needs, better reasoning support through high-quality ontologies, and ontology-based automation will promote the future uptake of ontologies in analytical research labs.

Acknowledgements

We are grateful to the organisers of the Asia-Pacific Conference on Conceptual Modelling (APCCM) for inviting us to write this survey paper.

References

Aitken, S., Chen, Y. & Bard, J. (2008), 'OBO Explorer: an editor for open biomedical ontologies in OWL', *Bioinformatics* **24**(3), 443–444.

Aranguren, M. E. & al. (2007), 'Understanding and using the meaning of statements in a bio-ontology: recasting the Gene Ontology in OWL', *BMC Bioinformatics* **8**(57), 1–13.

Aranguren, M. E. & al. (2008), 'Ontology design patterns for bio-ontologies: a case study on the cell cycle ontology', *BMC Bioinformatics* **9**(Suppl5), S1.

Ashburner, M. & al. (2000), 'Gene ontology: tool for the unification of biology', *Nature Genet* **25**, 25–29.

Baumgartner, W. A. & al. (2007), 'Manual curation is not sufficient for annotation of genomic databases', *Bioinformatics* **23**, i41–i48.

Beck, T. & al. (2009), 'Practical application of ontologies to annotate and analyse large scale raw mouse phenotype data', *BMC Bioinformatics* **10**(Suppl5), S2.

Bodenreider, O. (2008), Biomedical ontologies in action: Role in knowledge management, data integration and decision support, *in* 'IMIA Yearbook Medical Informatics', pp. 67–79.

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

Carbon, S. & al. (2009), 'AmiGO: online access to ontology and annotation data', *Bioinformatics* **25**(2), 288–289.

Chapman, W. W. & Cohen, K. B. (2009), 'Current issues in biomedical text mining and natural language processing', *J Biomedical Informatics* **42**(5), 757–759.

Chen, Q., Chen, Y.-P. P. & Zhang, C. (2007), 'Detecting inconsistency in biological molecular databases using ontologies', *Data Min Knowl Disc* **15**, 275–296.

Cote, R. G. & al. (2008), 'The ontology lookup service: more data and better tools for controlled vocabulary queries', *Nucleic Acids Research* **36**, 372–376.

Couto, F. M. & al. (2006), 'GOAnnotator: linking protein GO annotations to evidence text', *J Biomed Discov Collab* **1**, 19.

Cristani, M. & Cuel, R. (2005), 'A survey on ontology creation methodologies', *Int J Semantic Web Inf Syst.* **1**, 49–69.

Cure, O. & Jablonski, S. (2007), Ontology-based data integration in data logistics workflows, *in* 'Advances in Conceptual Modeling', Vol. 4802 of *LNCS*, pp. 34–43.

Dai, M. & al. (2008), An efficient solution for mapping free text to ontology terms, *in* 'AMIA Summit on Translational Bioinformatics'.

Daraselia, N. & al. (2007), 'Automatic extraction of gene ontology annotation and its correlation with clusters in protein networks', *BMC Bioinformatics* **8**, 243.

Darlington, M. J. & Culley, S. J. (2008), 'Investigating ontology development for engineering design support', *Advanced Engineering Informatics* **22**(1), 112–134.

Day-Richter, J. & al. (2007), 'OBO-Edit - an ontology editor for biologists', *Bioinformatics* **23**(16), 2198–2200.

Doms, A. & Schroeder, M. (2005), 'GoPubMed: exploring pubmed with the gene ontology', *Nucleic Acids Research* **33**, W783–W786.

Donlin, M. J. (2009), 'Using the generic genome browser (GBrowse)', *Curr Protoc Bioinform* **9**.

Draghici, S., Sellamuthu, S. & Khatri, P. (2006), 'Babels tower revisited: a universal resource for cross-referencing across annotation databases', *Bioinformatics* **22**, 2934–2939.

Eilbeck, K. & al. (2005), 'The Sequence Ontology: a tool for the unification of genome annotations', *Genome Biology* **6**(R44), 1–12.

Eilbeck, K. & Mungall, C. J. (2009), Evolution of the Sequence Ontology terms and relationships, *in* 'Nature Precedings'.

Epple, A. & Scherf, M. (2009), Bibliosphere hypothesis generation in regulatory network analysis, *in* 'Bioinformatics for Systems Biology', pp. 401–412.

Gangemi, A., Catenacci, C., Ciaramita, M. & Lehmann, J. (2006), Modelling ontology evaluation and validation, *in* 'European Semantic Web Conference'.

Ghazvinian, A., Noy, N. F. & Musen, M. A. (2009), Creating mappings for ontologies in biomedicine, *in* 'AMIA Annual Symposium'.

Golbreich, C., Horridge, M., Horrocks, I., Motik, B. & Shearer, R. (2008), OBO and OWL: Leveraging semantic web technologies for the life sciences, *in* 'Semantic Web', Vol. 4825 of *LNCS*, pp. 169–182.

Grossmann, S., Bauer, S., Robinson, P. N. & Vingron, M. (2007), 'Improved detection of overrepresentation of gene-ontology annotations with parentchild analysis', *Bioinformatics* **23**(22), 3024–3031.

Harris, M. A. & al. (2005), 'The gene ontology (GO) database and informatics resource', *Nucleic Acids Research* **32**, D258–D261.

Hartung, M., Kirsten, T., Gross, A. & Rahm, E. (2009), 'OnEX: Exploring changes in life science ontologies', *BMC Bioinformatics* **10**, 250.

Hartung, M., Kirsten, T. & Rahm, E. (2008), Analyzing the evolution of life science ontologies and mappings, *in* 'DILS', Vol. 5109 of *LNBI*, pp. 11–27.

Hernandez, T. & Kambhampati, S. (2004), 'Integration of biological sources: Current systems and challenges ahead', *SIGMOD Record* **33**(3), 51–60.

Hill, D. P., Smith, B., McAndrews-Hill, M. S. & Blake, J. A. (2008), 'Gene Ontology annotations: what they mean and where they come from', *BMC Bioinformatics* **9**(Suppl5), S2.

Hoehndorf, R. & al. (2009), 'BOWiki: an ontology-based wiki for annotation of data and integration of knowledge in biology', *BMC Bioinformatics* **10**(Suppl5), S5.

Huang, D. W., Sherman, B. T. & Lempicki, R. A. (2009), 'Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists', *Nucleic Acids Research* **37**, 1–13.

Hull, D. & al. (2006), 'Taverna: a tool for building and running workflows of services', *Nucleic Acids Research* **34**, W729W732.

Jaiswal, P. & al. (2005), 'Plant ontology (PO): a controlled vocabulary of plant structures and growth stages', *Comp Funct Genom* **6**, 388–397.

Jean-Marya, Y. R., Shironoshitaa, E. P. & Kabuka, M. R. (2009), 'Ontology matching with semantic verification', *Web Semantics: Science, Services and Agents on the World Wide Web* **7**(3), 235–251.

Jones, C. E., Baumann, U. & Brown, A. L. (2005), 'Automated methods of predicting the function of biological sequences using GO and BLAST', *BMC Bioinformatics* **6**, 272.

Kamel, M. N., Lee, A. Y. & Powers, E. C. (2007), A methodology for developing ontologies using the Ontology Web Language (OWL), *in* 'ICEIS', pp. 261–268.

Keet, C. M. (2006), Using and improving bio-ontologies stored in relational databases, *in* 'SBI-OLBD'.

Keet, C. M., Roos, M. & Marshall, M. S. (2007), A survey of requirements for automated reasoning services for bio-ontolgies in OWL, *in* 'OWL: Experiences and Directions'.

Khatri, P. & al. (2007), 'Onto-Tools: new additions and improvements', *Nucleic Acids Research* **35**, W206–W211.

Khatri, P. & Draghici, S. (2005), 'Ontological analysis of gene expression data: current tools, limitations, and open problems', *Bioinformatics* **21**, 3587–3595.

King, R. D. & al. (2009), 'The automation of science', *Science* **324**, 85–89.

Kirov, S. A. & al. (2005), 'GeneKeyDB: a lightweight, gene-centric, relational database to support data mining environments', *BMC Bioinformatics* **6**, 72.

Li, K.-C. & al. (2007), BioPortal: A portal for deployment of bioinformatics applications on cluster and grid environments, *in* 'High Performance Computing for Computational Science', Vol. 4395 of *LNCS*, pp. 566–578.

Lozano-Tello, A. & Gomez-Perez, A. (2004), 'Ontometric: A method to choose the appropriate ontology', *J Database Management* **15**(2), 1–18.

Lubyte, L. & Tessaris, S. (2009), Automatic extraction of ontologies wrapping relational data sources, *in* 'DEXA', Vol. 5690 of *LNCS*, pp. 128–142.

Lutz, C. & al. (2006), Reasoning support for ontology design, *in* 'Experiences and Directions'.

Maiga, G. & Williams, D. (2008), 'A user centered approach for evaluating biomedical data integration ontologies', *European J Scientific Research* **24**(1), 55–68.

Moreira, D. A. & al. (2009), The NCBO OBOF to OWL mapping, *in* 'Nature Precedings', pp. 1–6.

Moreira, D. A. & Musen, M. A. (2007), 'OBO to OWL: A Protege OWL tab to read/save OBO ontologies', *Bioinformatics* **23**, 1868–1870.

Moskovitch, R. & al. (2007), 'A comparative evaluation of full-text, concept-based, and context-sensitive search', *J Am Med Inform Assoc* **14**(2), 164–174.

Mungall, C. J. & Emmert, D. B. (2007), 'A Chado case study: an ontology-based modular schema for representing genome-associated biological information', *Bioinformatics* **23**, i337–i346.

Novacek, V., Laera, L., Handschuh, S. & Davis, B. (2009), 'Infrastructure for dynamic knowledge integration - automated biomedical ontology extension using textual resources', *J Biomedical Informatics* **41**(5), 816–828.

Noy, N. F. & al. (2009), 'BioPortal: Ontologies and integrated data resources at the click of a mouse', *Nucleic Acids Research* **37**, W170–W173.

Noy, N. F. & Klein, M. (2004), 'Ontology evolution: Not the same as schema evolution', *Knowledge and Information Systems* **6**(4), 428–440.

Park, J. C., Kim, T. & Park, J. (2008), 'Monitoring the evolutionary aspect of the gene ontology to enhance predictability and usability', *BMC Bioinformatics* **9**(Suppl3), S7.

Pathak, J., Johnson, T. M. & Chute, C. G. (2009), 'Survey of modular ontology techniques and their applications in the biomedical domain', *J Integrated Computer-Aided Engineering* **16**(3), 225–242.

Perez-Rey, D. & al. (2006), 'ONTOFUSION: ontology-based integration of genomic and clinical databases', *Comput Biol Med* **36**(7-8), 712–730.

Pesquita, C., Grego, T. & Couto, F. (2009), Identifying gene ontology areas for automated enrichment, *in* 'Distributed Computing, Artificial Intelligence, Bioinformatics, Soft Computing, and Ambient Assisted Living', Vol. 5518 of *LNCS*, pp. 934–941.

Porzel, R. & Malaka, R. (2004), A task-based approach for ontology evaluation, *in* 'ECAI Workshop Ontology Learning and Population'.

Ram, S., Zhang, K. & Wei, W. (2008), Linking biological databases semantically for knowledge discovery, *in* 'Advances in Conceptual Modeling', Vol. 5232 of *LNCS*, pp. 22–32.

Rosse, C. & Mejino, J. V. L. (2003), 'A reference ontology for biomedical informatics: the Foundational Model of Anatomy', *J Biomed Inform* **36**, 478–500.

Rubin, D. L., Shah, N. H. & Noy, N. F. (2007), 'Biomedical ontologies: a functional perspective', *Briefings in Bioinformatics* **7**, 75–90.

Schober, D. & al. (2009), 'Survey-based naming conventions for use in OBO foundry ontology development', *BMC Bioinformatics* **10**(125), 1–9.

Shah, N. H. & al. (2009), 'Ontology-driven indexing of public datasets for translational bioinformatics', *BMC Bioinformatics* **10**(Suppl2), S1.

Sherman, B. T. & al. (2007), 'DAVID knowledgebase: a gene-centered database integrating heterogeneous gene annotation resources to facilitate high-throughput gene functional analysis', *BMC Bioinformatics* **8**, 426.

Smedley, D. & al. (2008), 'Solutions for data integration in functional genomics: a critical assessment and case study', *Briefings in Bioinformatics* **9**(6), 532–544.

Smith, B. & al. (2005), 'Relations in biomedical ontologies', *Genome Biology* **6**(R46), 1–15.

Smith, B. & al. (2007), 'The OBO foundry: coordinated evolution of ontologies to support biomedical data integration', *Nature Biotechn* **25**, 1251–1255.

Soldatova, L. N., Clare, A., Sparkes, A. & King, R. D. (2006), 'An ontology for a robot scientist', *Bioinformatics* **22**(14), e464–e471.

Tan, H. & Lambrix, P. (2009), Selecting an ontology for biomedical text mining, *in* 'Workshop on BioNLP', pp. 55–62.

Tao, C. & Embley, D. (2007), Seed-based generation of personalized bio-ontologies for information extraction, *in* 'Advances in Conceptual Modeling', Vol. 4802 of *LNCS*, pp. 74–78.

Thomas, C. J. & al. (2006), Modular ontology design using canonical building blocks in the biochemistry domain, *in* 'Formal Ontology in Information Systems', pp. 115–127.

Tudorache, T., Noy, N. F., Tu, S. & Musen, M. A. (2008), Supporting collaborative ontology development in protege, *in* 'The Semantic Web', Vol. 5318 of *LNCS*, pp. 17–32.

Vinayagam, A. & al. (2006), 'GOPET: a tool for automated predictions of Gene Ontology terms', *BMC Bioinformatics* **7**(161), 1–7.

Zhao, S. & Chang, E. (2007), Mediating databases and the semantic web: a methodology for building domain ontology from databases and existing ontologies, *in* 'Semantic Web & Web Services'.