Towards a Semantic Web for Bioinformatics using Ontology-based Annotation

Patrick Lambrix
Linköpings universitet
SE-581 83 Linköping, Sweden
patla@ida.liu.se

Abstract

Nowadays, biologists use biological data sources and tools to find relevant information for their research. However, with the explosion of the amount of online accessible data and tools, finding the relevant sources and retrieving the relevant information is not an easy task. Further, often information from different sources needs to be integrated. The vision of a Semantic Web alleviates these difficulties. The Semantic Web is an extension of the current Web in which information is given a well-defined meaning by annotating Web content with ontology terms. In this paper we discuss the Semantic Web vision and focus on an important technology, ontologies, that is needed to make this vision happen.

1 Introduction

Researchers in various areas, e.g. medicine, agriculture and environmental sciences, use biological data sources and tools to answer different research questions or to solve various tasks [3]. One of the main goals is to understand how various organisms function as biological systems. To achieve this goal, it is important to explore functions and interactions of genome-encoded components. This type of knowledge may be used for different purposes. For instance, it is used to identify genes responsible for a disease, to develop drugs enabling treatment of diseases and to predict organisms' responses to a drug. Also, research is conducted on how the genomes vary between species, how mutations affect functioning of different components in organisms and what differences they cause between organisms. Also, the influence of environmental factors on human health and diseases is investigated.

During recent years an enormous amount of biological data, such as DNA and protein sequences, gene regulatory and protein interaction networks, and secondary together with tertiary structures of molecules, has been generated. This data is spread in a large number of autonomous data sources that are often publicly available on the Web. There are also numerous tools available on the Web such as BLAST, a sequence alignment tool.

Researchers that need to use these databases and tools experience a number of difficulties. A first difficulty is to *locate the relevant data sources and tools*. There are many data sources and users need to have good knowledge about which data sources exist and what information they contain. Often data sources contain overlapping information and the required information can be obtained in a number of ways. Depending on which way is chosen, there may be a difference in the time it takes to obtain results as well as in the quality of the obtained results. Further, it is not easy to stay up to date as the environment is changing frequently. New data is added to the data sources on daily basis. For instance, in 1986 SWISS-PROT, a protein database, contained a few thousand data entries, while in 2005 the database contains over 160 000 entries. Further, new data sources appear frequently. For instance, the yearly database issue of the Nucleic Acids Research [15] journal reported on 386 data sources in 2003, 548 in 2004 and 719 in 2005. Data sources may also disappear. A specific property of biological data sources is further that their structure is frequently modified. This happens, for instance, when new types of data are generated by novel tools and approaches.

A second difficulty is to, once the relevant data sources are identified, *retrieve the relevant information*. Current retrieval approaches are often syntax based and do not provide good precision and recall. This means that the query results often contain information that is not relevant for the user's query. One reason is that terms are often ambiguous. For instance, when looking for information about jaguars, the result will include documents about the animal as well as documents about the car. It is also the case that much relevant information may not be found. For instance, when looking for information about signal transducers also documents about receptors are interesting as receptors are a special kind of signal transducers. However, syntax-based retrieval systems will not return these documents unless signal transducer also occurs explicitly in these documents.

Proceedings of the 14th IEEE International Workshops on Enabling Technologies: Infrastructure for Collaborative Enterprise (WETICE'05)
0-7695-2362-5/05 $20.00 © 2005 IEEE

A third difficulty is that for most tasks data from different sources needs to be *integrated*. For instance, to find publications describing a given disease that relates to a certain type of sequences may require analysis of data sources for publications, diseases and sequences together with some other data sources combining these types of information [8]. To predict properties of a new protein sequence, data sources containing information about protein sequences, protein families and protein structures may be needed. Because the data sources are developed and supported independently by different groups and organizations, the data sources are highly heterogeneous in various aspects [12]. They differ in content, i.e. the type of information that they store, although data sources may also contain overlapping information. The quality of the data may differ. For instance, some data sources contain experimentally verified data while other data sources contain predicted data (e.g. generated by data mining programs). Different kinds of data models are used for the representation of the data such as the relational model, the object-oriented model, semi-structured data and flat files. The sources are also heterogeneous regarding their query languages and query capabilities. Further, there is a terminology discrepancy problem. Data sources can use different terminology to represent the same data or the same term may be used by different sources to refer to different data items.

In the remainder of the paper we introduce the vision of the Semantic Web and show how a first step towards this vision can be taken (section 2). In section 3 we discuss ontologies, a key technology for this step towards a Semantic Web. We introduce the notion of ontologies, discuss different kinds of representations as well as ontology tools. In section 4 we show how this step alleviates the three difficulties discussed above. The paper concludes in section 5.

2 Semantic Web

The current Web is essentially a collection of documents that are interconnected by links and it is used as a portal to applications. For instance, the biological data sources available on the Web provide access through Web pages. To query the data sources often the users fill out forms. The results are again presented to the users as Web pages. The Web pages are presented to the users based on mark-up. This mark-up mainly represents rendering information, such as the font and color of the text, and links to other Web pages. Therefore, the current Web is mostly a medium of documents for people rather than for information that can be processed automatically by computers [2].

The Semantic Web is a vision of a further development of the World Wide Web "in which information is given well-defined meaning, better enabling computers and people to work in cooperation" [2]. The World Wide Web Consortium states it as follows [21]: "The Web can reach its full potential only if it becomes a place where data can be shared and processed by automated tools as well as by people. ... The Semantic Web is a vision: the idea of having data on the Web defined and linked in a way that it can be used by machines not just for display purposes, but for automation, integration and reuse of data across various applications."

As a first step towards this vision of making the content of Web pages machine-understandable, people have started to use semantic annotation. One way to do this is to annotate the Web pages with 'meaningful' tags. In this case we annotate the Web pages with XML mark-up to distinguish the meaningful parts of the document. For instance, in a Web page about a protein we may distinguish between its name, coding DNA, three dimensional structure, family, function, source organism, etc. Then we use programs that recognize the mark-up and the different parts of the document can then be used in other programs based on the meaning represented by the mark-up. However, for this approach to be successful, there is a need for agreement on the annotation. A solution to this is to use ontologies to specify the meaning of the annotations. The ontologies define a vocabulary, specify the meaning of the terms and define how new terms can be formed by combining existing terms. We introduce ontologies in the next section.

3 Ontologies

Intuitively, ontologies (e.g. [9, 5]) can be seen as defining the basic terms and relations of a domain of interest, as well as the rules for combining these terms and relations. Ontologies are used for communication between people and organizations by providing a common terminology over a domain. They provide the basis for interoperability between systems. They can be used for making the content in information sources explicit and serve as an index to a repository of information. Further, they can be used as a basis for integration of information sources and as a query model for information sources. They also support clearly separating domain knowledge from application-based knowledge as well as validation of data sources. The benefits of using ontologies include reuse, sharing and portability of knowledge across platforms, and improved maintainability, documentation, maintenance, and reliability. Overall, ontologies lead to a better understanding of a field and to more effective and efficient handling of information in that field.

3.1 Bio-ontologies

Within bioinformatics many ontologies exist. Many of the model organism databases such as Flybase and Mouse Genome database can be seen as simple ontologies. Further, there are ontologies focusing on things such as protein

functions, organism development, anatomy and pathways. (For examples we refer to e.g. [9, 16, 19].)

The use of ontologies in bioinformatics has grown drastically since database builders concerned with developing systems for different (model) organisms joined to create the Gene Ontology Consortium (GO, [4]) in 1998. The goal of GO is to produce a structured, precisely defined, common and dynamic controlled vocabulary that describes the roles of genes and proteins in all organisms. Currently, there are three independent ontologies publicly available: biological process, molecular function and cellular component. The GO ontologies have become a de facto standard and are used by many biological data sources for annotation.

Recently, Open Biomedical Ontologies (OBO, [16]) was started as an umbrella Web address for ontologies for use within the genomics and proteomics domains. The member ontologies are required to be open, to be written in a common syntax, to be orthogonal to each other, to share a unique identifier space and to include textual definitions. Many bio-ontologies are already available via OBO.

The field has matured enough to start talking about standards. An example of this is the organization of the first conference on Standards and Ontologies for Functional Genomics (SOFG) in 2002 and the development of the SOFG resource on ontologies [19]. The work on ontologies is also recognized as essential in some of the grand challenges of genomics research [3] and there is much international research cooperation for the development of ontologies (e.g. OBO) and the use of ontologies for the Semantic Web (e.g. the EU Network of Excellence REWERSE [18]).

3.2 Knowledge representation for ontologies

Ontologies differ regarding the kind of information they can represent. From a knowledge representation point of view ontologies can have the following components (e.g. [20]). *Concepts* represent sets or classes of entities in a domain. The concepts may be organized in taxonomies, often based on the is-a relation or the part-of relation. *Instances* represent the actual entities. They are, however, often not represented in ontologies. Further, there are many types of *relations*. Finally, *axioms* represent facts that are always true in the topic area of the ontology. These can be such things as domain restrictions, cardinality restrictions or disjointness restrictions.

Depending on which of the components are represented and the kind of information that can be represented, we distinguish between different kinds of ontologies. A simple type of ontology is the *controlled vocabulary*. These are essentially lists of concepts. When these concepts are organized in an is-a hierarchy, we obtain a *taxonomy*. Many of the current bio-ontologies are taxonomies. A slightly more complex kind of ontology is the *thesaurus*. In this case the concepts are organized in a graph. The arcs in the graph represent a fixed set of relations, such as synonym, narrower term, broader term, similar term. The *data models* allow for defining a hierarchy of classes (concepts), attributes (properties of the entities belonging to the classes, functional relations), relations and a limited form of axioms. There are also the *knowledge bases* which are often based on a logic. They can represent all types of components and provide reasoning services such as checking the consequences of the statements in the ontology and building the is-a hierarchy.

An ontology and its components can be represented in a spectrum of representation formalisms ranging from very informal to strictly formal. This would include natural language, limited and structured forms of natural language, formally defined languages and logics with formal semantics [6]. The choice of which formalism to use depends on the characteristics of the ontology as well as on its intended use. An interesting family of logics for representing ontologies are description logics (e.g. [1]). One of the languages within this family, DAML+OIL, was previously recommended by the BioOntology Consortium as its choice of ontology representation language. Its successor, OWL, is now one of the languages that may be used for representing OBO ontologies.

In practice, in bioinformatics ontologies such as the GO ontologies, have started out as controlled vocabularies. This allowed the ontology builders, which were domain experts, but not necessarily experts in knowledge representation, to focus on the gathering of knowledge and the agreeing upon definitions. More advanced representation and functionality was a secondary requirement and was left as future work. However, some of the bio-ontologies have reached a high level of maturity and stability regarding the ontology engineering process and their developers have now started investigating how the usefulness of the ontologies can be augmented using more advanced representation formalisms and added functionality.

3.3 Ontology tools

In the same way as there exist many tools in software engineering that provide support for the different software development phases, there are now also tools that provide support for the different phases of the ontology engineering process. Based on the tasks and processes that are supported, the ontology tools can be grouped in the following clusters [17]. *Ontology development tools* are used for building new ontologies. These tools usually support editing, browsing, documentation, export and import from different formats, views, libraries and they may have attached inference engines. *Ontology alignment, merge and integration tools* support users in merging or integrating ontologies in the same domain. Within an area there are always a num-

ber of ontologies, each with their own focus. For instance, in bioinformatics ontologies may cover different aspects in molecular biology such as molecular function and cell signaling. Many of these ontologies contain overlapping information. For instance, a protein can be involved in both cell signaling and other biological processes. In applications using ontologies it is therefore of interest to be able to use multiple ontologies. However, to obtain the best results, we need to know how the ontologies overlap and align them (define the relationships between the ontologies) or merge them (create a new ontology containing the knowledge included in the original ontologies). Another reason for merging ontologies is that it allows for the creation of ontologies that can later be composed into larger ontologies. Also, companies may want to use de facto standard ontologies and merge them with company-specific ontologies. *Ontology evaluation tools* support ensuring a certain level of quality for the ontologies. A kind of tools that may become more important for bioinformatics, are the *ontology-based annotation tools*, which allow users to insert ontology-based mark-up in Web pages. Further, there are also *ontology learning tools* that derive ontologies from natural language texts and *ontology storage and querying tools*.

Currently, only few evaluations of ontology tools using bio-ontologies have been performed and almost no support or benchmarks exist yet for performing evaluations. In [11] Protégé-2000, Chimaera, OilEd and DAG-Edit were evaluated as ontology development tools using GO ontologies as test ontologies. Protégé-2000 with PROMPT, Chimaera and SAMBO were also evaluated as ontology merging tools [10, 13]. A general framework for aligning ontologies where different alignment strategies can be combined is presented in [14]. It can be used to experiment with combinations of strategies and is a first step towards defining a framework that can be used for comparative evaluations of alignment strategies. In [23] Protégé-2000 was assessed as a tool for maintaining and developing the GO ontologies. For a survey of ontology tool evaluations and proposals (including other areas) we refer to [7].

4 Ontology-based locating, retrieval and integration

The use of ontologies and semantic annotation can alleviate the difficulties for users of Web-based data sources and tools as described in section 1.

Locating the relevant data sources and tools. An approach that would alleviate the difficulty of finding the relevant data sources and tools is to use Semantic Web services. In the current Web service [22] approach, data sources and tools can be seen as service providers and announce their services. Data sources, for instance, can announce their content and query capabilities. Users can be seen as consumers that request services based on their task. User requests and services are matched by service matchers. When we semantically enable the Web service approach, service providers are able to use ontologies to describe their services and users can use ontologies when formulating their requests. The service matchers will then more easily find relevant services.

Retrieving the relevant information. By using ontologies during information retrieval, it is possible to reduce the amount of non-relevant information in the returned results. For instance, when looking for information about jaguars, the user may use an ontology to state that she is interested in the animal. The result will then only include documents about the animal. It is also possible to find more relevant information. For instance, when looking for information about signal transducers, we may take into account information from an ontology that states that receptors are a special kind of signal transducers. Therefore, also documents about receptors will be returned.

Integrating data sources. Semantic annotations can enhance the integration process. Entities in different data sources that are annotated with the same or related ontology terms are likely related. Relations between data items could be derived from relations (e.g. equivalent, is-a, part-of) between the ontology terms they are annotated with.

5 Conclusion

In this paper we discussed the fact that there are a number of difficulties that researchers face when they use biological data sources and tools to find relevant information for their research. The vision of a Semantic Web alleviates these difficulties. We discussed a first step towards the Semantic Web. Further, we focused on an important technology, ontologies, that is needed to make this vision happen. The focus in bioinformatics has been on the construction of a community reference and on creating a vocabulary for annotation. Ontologies will still be used in this way in the future, while the other uses such as ontology-based search and ontology-based integration of information sources will gain in importance. Regarding representation the focus has been on controlled vocabularies. However, to be able to use the ontologies in more advanced ways, more expressive representation formalisms with reasoning capabilities will need to be used. We also mentioned some ontology engineering tools. The current ontology engineering tools are a good start, but research is still needed to be able to develop high-quality industrial strength ontology engineering tools. More research is also needed to develop approaches for comparative evaluations of ontology tools.

Acknowledgements

Much of the research at Linköpings universitet concerning the topics in this paper has been performed together with Vaida Jakonienė, He Tan and Lena Strömbäck. We also thank Bo Servenius, Jan Maluszynski and Nahid Shahmehri for discussions, and Bassam Abdullahad, Anna Edberg, Manal Habbouche, Georgios Lounis, Carolyn Manis and Marta Pérez for help with implementations and evaluations. We acknowledge the financial support of the Center for Industrial Information Technology (CENIIT), the Swedish national graduate school of computer science (CUGS) and the EU Network of Excellence REWERSE (Sixth Framework Programme project 506779).

References

[1] F. Baader, D. Calvanese, D. McGuinness, D. Nardi, P. Patel-Schneider. *The Description Logic Handbook*. Cambridge University Press, 2003.

[2] T. Berners-Lee, J. Hendler, O. Lassila. The Semantic Web. *Scientific American*, May issue, 2001.

[3] F. Collins, E. Green, A. Guttmacher, M. Guyer. A Vision for the Future of Genomics Research. *Nature* 422:835-847, 2003.

[4] The Gene Ontology Consortium. Gene Ontology: tool for the unification of biology. *Nature Genetics*, 25(1):25-29, 2000. http://www.geneontology.org/

[5] A. Gómez-Pérez. Ontological Engineering: A state of the Art. *Expert Update* 2(3):33-43, 1999.

[6] R. Jasper, M. Uschold. A Framework for Understanding and Classifying Ontology Applications. *Proceedings of the 12th Workshop on Knowledge Acquisition, Modeling and Management*, 1999.

[7] Knowledge Web Network of Excellence. Deliverable D2.1.1 (State of the art on the scalability of ontology-based technology), 2004. http://knowledgeWeb.semanticWeb.org/

[8] Z. Lacroix, H. Murthy, F. Naumann, L. Raschid. Links and Paths through Life Science Data Sources. *Proceedings of International Workshop on Data Integration in the Life Sciences*, LNCS 2994, pp 203-211, 2004.

[9] P. Lambrix. Ontologies in Bioinformatics and Systems Biology. Chapter 8 in Dubitzky, Azuaje (eds) *Artificial Intelligence Methods and Tools for Systems Biology*, pp 129-146, Springer, 2004. ISBN: 1-4020-2859-8.

[10] P. Lambrix, A. Edberg. Evaluation of ontology merging tools in bioinformatics. *Proceedings of the Pacific Symposium on Biocomputing*, 8:589-600, 2003.

[11] P. Lambrix, M. Habbouche, M. Pérez. Evaluation of ontology development tools for bioinformatics. *Bioinformatics* 19(12):1564-1571, 2003. (Also republished in the 2005 edition of the IMIA Yearbook, pp 547-554.)

[12] P. Lambrix, V. Jakonienė. Towards Transparent Access to Multiple Biological Databanks. *Proceedings of the Asia-Pacific Bioinformatics Conference*, pp 53-60, 2003.

[13] P. Lambrix, H. Tan. Merging DAML+OIL Ontologies. Barzdins, Caplinskas (eds) *Databases and Information Systems - Selected Papers from the Sixth International Baltic Conference on Databases and Information Systems*, pp 249-258, IOS Press, 2005.

[14] P. Lambrix, H. Tan. A Framework for Aligning Ontologies. *Third Workshop on Principles and Practice of Semantic Web Reasoning*, 2005.

[15] The online Nucleic Acids Research journal. http://nar.oupjournals.org

[16] Open Biomedical Ontologies. http://obo.sourceforge.net/

[17] OntoWeb Consortium. Deliverables 1.3 (A survey on ontology tools) and 1.4 (A survey on methodologies for developing, maintaining, evaluating and reengineering ontologies), 2002. http://www.ontoWeb.org

[18] REWERSE Network of Excellence. Deliverables A2-D1 (State-of-the-art in Bioinformatics), 2004 and A2-D2 (Usage of bioinformatics tools and identification of information sources), 2005. http://www.rewerse.net

[19] Standards and Ontologies for Functional Genomics. http://www.sofg.org

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

[21] World Wide Web Consortium. Semantic Web. http://www.w3.org/2001/sw/

[22] World Wide Web Consortium. Web Services. http://www.w3.org/2002/ws/

[23] I. Yeh, P. Karp, N.F. Noy, R. Altman. Knowledge acquisition, consistency checking and concurrency control for Gene Ontology (GO). *Bioinformatics* 19(12):241-248, 2003.