Managing Biological Information Using Biological Knowledge

Carole A. Goble and Robert D. Stevens
Department of Computer Science
University of Manchester
Oxford Road
Manchester UK
M13 9PL
carole@cs.man.ac.uk

Abstract

The Transparent Access to Multiple Bioinformatics Information Sources (TAMBIS) system uses an ontology to give the illusion of a common query interface to multiple, diverse, heterogeneous bioinformatics resources. The knowledge in the TAMBIS ontology (TaO) allows biologists to form complex, multi-source queries without having to know which source to use, the location of the source, the meaning of terms within the source and how to transfer information between resources. The TaO can be seen to add a semantic layer over these bioinformatics resources. Ontologies are finding a wider use in the bioinformatics arena, emphasising the usefulness of knowledge in bioinformatics. Just as the vision for the Semantic Web envisages the use of knowledge to give a machine processable web, so semantic bioinformatics resources will enable machine processable, rather than only machine readable, bioinformatics resources.

1 Introduction

Many bioinformatics analyses and applications require several resources, both data repositories and analysis tools, to work together to achieve the result. Not only are these resources widely distributed, but they also suffer from a large degree of heterogeneity. This heterogeneity can be broadly broken down into:

i *Syntactic heterogeneity*, where resources lie on different hardware platforms, have different storage paradigms (flat-file and database management system) and different APIs; and

ii *Semantic heterogeneity* where different resources use different conceptualisations to model their data and analysis techniques. In two resources, for instance, a term can have different meanings. Conversely, two different terms can have the same meaning in two resources. An extreme example of heterogeneity is the use of the term *gene* in bioinformatics resources [14].

Both these heterogeneities pose problems for a bioinformatician needing to use multiple resources. The first makes it difficult to engineer resources to co-operate. Once this inter-operation has been achieved, information stored with conflicting meanings or labels needs to be reconciled semantically. Usually, this task relies on the skill and knowledge of the human user.

Traditional techniques used to resolve syntactic heterogeneity are the development of resource wrappers and exchange formats. Wrappers can transform the appearance of a resource to the external world and dictate what services are available. The Common Object Request and Brokerage Architecture (CORBA) offers a standard mechanism by which object views of resources may be developed [16]. Just as CORBA defines the syntactic view of a resource's services, the eXtensible Markup Language (XML) (see `http://www.w3.org/XML`) can be used to define the structure of a resource's data. Again, such a technology works by the adoption of standards, so that similar resources use the same data format.

These technologies, however, only offer a mechanism for plumbing resources together. A common structural view of a resource's data does not necessarily mean a common semantic view of the same information. At best, these mechanisms offer only an intuitive semantics. For example, an XML tag called <`sequence/`>, within a sequence database entry might have a common understanding, to humans if not machines, but it is unlikely that a tag <`gene/`> would have such an intuitive understanding. These interoperation technologies do not resolve any of the problems of heterogeneous conceptualisations or term usage, because they do not have a mechanism for describing the meaning or knowledge associated with a term.

Knowledge is needed to overcome this second type of heterogeneity. Ontologies are used to capture a community's knowledge about a domain and make it accessible to that community and its applications [15]. An ontology can cap-

ture the meaning of terms used within a domain in terms of concepts and that concept's properties. These ontologies can then be used to either give a common terminology within different resources or to map between differing terms used between resources.

At present, this mediation task has to be carried out by the human users of resources, which means that it is often either not carried out at all, or performed incorrectly or inconsistently. The addition of knowledge to resources via an ontology can not only help humans, but also help machines to mediate successfully between resources. To do this, applications need to capture the knowledge that people implicitly bring to their usage of bioinformatics tools or encode in the programs themselves. The following section describes one use of an ontology to manage the heterogeneity existing between bioinformatics resources. The final section attempts to generalise the use of ontologies within the bioinformatics arena.

1.1 The TAMBIS Approach to Information Management

The TAMBIS (Transparent Access to Multiple Bioinformatics Information Sources) system uses an ontology of molecular biology and bioinformatics tasks to manage the two forms of heterogeneity existing in bioinformatics resources. It provides the illusion of a common query interface to a variety of distributed, autonomous, heterogeneous bioinformatics resources [7]. A TAMBIS applet may be found at `http://img.cs.man.ac.uk/tambis`. The ontology allows users to formulate complex, yet precise, queries over several resources. The transparency removes the need for the user to choose the resources, know the resources' location, know how to manipulate that resource, and know how to manage the semantic heterogeneity between those resources. Queries are formed against the global schema provided by the TaO; the conceptual query is passed to the query processor, which interacts with the TaO and a catalogue of wrapper functions and concept mappings, indexed by the TaO to rewrite the query to a concrete form [12]. The wrapper layer gathers results and passes them back to the user for viewing. At this stage the query concept is still available for iterative refinement, the cycle being repeated until appropriate results are obtained.

The TAMBIS ontology (TaO) [2] describes a wide range of bioinformatics tasks and molecular biology. Figure 1 illustrates the TAMBIS architecture and emphasises the central role of the TaO. The roles the TaO plays within TAMBIS are:

- To describe the molecular biology in bioinformatics sources and the tasks performed on these data;

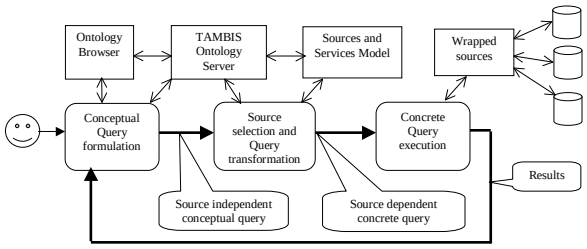

Figure 1: The flow of information through the TAMBIS architecture.

- To drive the query formulation interface;
- To act as a semantic index to the wrapper and mapping catalogue; and
- To facilitate the query transformation process.

The TaO covers a wide range of biological concepts such as macromolecules, their sequence components, their function, cellular location and the processes in which they take part. Secondary and tertiary structure are also described. These broad categories are specialised to the level of entry classes, features and keywords found in the entries of resources such as SWISS-PROT, PROSITE, CATH AND EMBL. Conceptualisations pertinent to bioinformatics, such as AccessionNumber and isHomologousTo, also appear. These concepts are not necessary to the description of biological knowledge, but are vital for performing bioinformatics tasks. Other conceptualisations contradict domain knowledge, but are also needed for bioinformatics tasks. The translation of DNA to protein is an example of this kind of conceptualisation.

The TaO is encoded in the Description logic (DL) GRAIL [13] and is delivered to TAMBIS through a terminology server (TeS) [3], that offers reasoning services to TAMBIS. The reasoning services allow new concepts to be formed dynamically by users and automatically classified within the TaO. So, the concept Protein can be joined to the function Receptor via the relationship hasFunction, to form the new concept ReceptorProtein. This new concept is automatically classified as a kind of Protein. The knowledge in the TaO only allows users to create biologically sensible concepts.

Concepts in the TaO, such as Protein, and those dynamically created by users, such as Receptor Protein can act as queries. A concept represents a set of instances. Retrieving those instances described by a concept is answering a query. The TaO holds knowledge about what questions may be asked of bioinformatics resources and TAMBIS uses the TaO to drive a query formulation interface [4]. In the query builder, a user chooses a concept, such as Protein, upon

which to base a query. TAMBIS can then ask the TaO what other concepts can legitimately be joined to or replace the current concepts. The query builder is asking the TaO 'what can be said about this biological concept?'. The knowledge held in the TaO means that what can be said about the growing concept changes as more knowledge is added to the query. This knowledge, for instance, stops a user asking for DNA motifs that are part of a protein sequence [2].

This knowledge driven query builder delivers the conceptual query to the query processing part of TAMBIS [12]. Conceptual queries are source-independent, declarative and need to be transformed into source-dependent, ordered, concrete query plans. The Sources and Services Model (SSM) in Figure 1 uses the TaO to associate concepts to wrapper functions and arguments that form part of the query plan. At this point some of the semantic heterogeneity between the resources can be addressed: For instance, one concept can appear in the SSM several times with different mappings depending on the resource involved.

The query processor consults both the SSM and the TaO to transform a query to a concrete query plan. This plan is then delivered to the wrapper layer, which executes the query and collects the results. It can be seen that the TaO impinges at all stages of retrieval, except in query execution, within TAMBIS. The TaO enables TAMBIS to give the illusion of a common query interface to diverse resources and allows semantic heterogeneity to be managed. The knowledge within the TaO allows complex, yet precise queries to be asked, without biologically non-sensical queries being allowed [2].

Some examples of TAMBIS queries are: Find the motifs on protein homologues of a named protein; find phosphorylation sites on proteins with a given function; find all human proteins with a known seven propeller domain architecture; find homologues of human, apoptosis receptor proteins; retrieve the receptor proteins involved in lactation and disease processes; find motifs on enzymes with a thymine substrate and an iron cofactor. This small sample represents a range of queries from simple, standard bioinformatics tasks, to complex queries over a range of sources. There are two prominent characteristics of TAMBIS queries: First, the potential specificity of the query, for example, only finding a particular class of motifs on a particular class of proteins; second, the common, conceptual view of the wider domain, that allows queries to be refined until the task is achieved.

1.2 Lessons Learnt from TAMBIS

The TAMBIS approach to bioinformatics resource interoperation has been seen to work. Users are able to generate complex, multi-source queries without knowledge of which resource to use, its location, how to use that resource, etc. This approach is, however, a high-price, high-gain one, which has costs as well as benefits. The high-price comes from the investment of effort in building and maintaining the TaO and integrating external resources into the system. The benefits include the TaO itself, a model that captures and makes explicit knowledge of molecular biology and bioinformatics, that is available to the wider community. Many users find the transparency offered by TAMBIS an advantage, allowing focus on the query, rather than the query mechanics. This allows the rich, multi-source queries characteristic of TAMBIS to be formed. This transparency can, however, be a barrier to some expert users. TAMBIS has a single source assumption – so Protein maps to only SWISS-PROT and no other sequence repository. Many users will wish to choose which resources are used to answer queries and otherwise manipulate all or part of the query independent of TAMBIS.

The TaO is broad and shallow in its scope. Detail or specialisation of concepts can be generated through the concepts formed by users. The point at which users currently start constructing queries lacks detail – for instance, protein functions are no more specialised than Receptor.

1.3 Current Developments in TAMBIS

A new version of TAMBIS is under development. The prototype version suffered from its own heterogeneity, work being distributed over several platforms and using several programming forms. The new version of TAMBIS is a Java only form, with a new ontology and re-worked query processing as its most significant changes. Only the changes to the ontology will be discussed here.

The TaO is being re-encoded and partially re-conceptualised. More detail is being added to the TaO, making it less shallow. The new TaO has extended the model further, to accommodate the points at which users wish to initiate their queries. In addition, some of the TaO's concepts currently map directly to terms used within some of the resources. An increased number of resources in TAMBIS makes this untenable. Increased generality in the TaO avoids any bias in the conceptualisation.

The TaO is currently encoded in GRAIL [13], a relatively inexpressive DL. Whilst this made the query processing easier, the conceptualisation, at some points,was awkward and contrived. The new TaO is encoded in the Ontology Inference Layer (OIL) [6], supported by the reasoning supplied by the DL FaCT [9]. GRAIL only had conjunctive concept forming operations, whereas OIL can use conjunction, disjunction and negation to form concepts. A Cofactor can be either a MetalIon *or* a OrganicMolecule, a simple disjunction. There are two kinds of Cofactor: Coenzyme and prosthetic-group, distinguished by the former binding loosely to a protein and the latter binding tightly to a protein. In addition, a Prosthetic-Group is a kind of

Cofactor, but *not* a Metal-Ion. OIL can express both the disjunction and negation used here, as well as the conjunctive form, the only concept forming operation available in GRAIL. OIL's expressivity allows the new TaO to be clearer and capture domain knowledge with high-fidelity.

The drawbacks of the TAMBIS approach, mentioned above will be addressed. The single source assumption will be relaxed, thus allowing some users to express a preference for certain resources. This will necessitate the maintenance of data provenance or audit trails within TAMBIS. Currently, the integration of resources into TAMBIS is hand-crafted. The development of resource wrappers and the mapping of those resources into the TAMBIS ontology will be supported by TAMBIS tools. Currently, results are an end-point in TAMBIS. Results should be storable, as well as being part of the input to future queries. For example, Receptor Proteins can be gathered and form the input to a query about human receptor proteins.

2 General Use of Bio-Ontologies and the Semantic Web of Biological Resources

TAMBIS uses knowledge to give transparency to the process of inter-operating between diverse, distributed bioinformatics resources. The use of ontologies to capture biological knowledge is increasing within the bioinformatics arena. The uses to which ontologies are put can be divided into two broad categories:

i Ontologies are used as controlled vocabularies within and across data repositories. The ontology defines the meanings of terms used to annotate data items, such as gene products, so that terms have a shared meaning and are used consistently. The structure of the ontology also helps in query refinements, as users can move within the ontology to specialise or generalise query terms used. Such controlled vocabularies also enhance recall and precision – all of the data items for a given term are retrieved, but only data with those terms are retrieved. The Gene Ontology (GO) [17] is an example of this approach. The Schulze-Kremer ontology for molecular biology (MBO) [14] has a related purpose, offering a community wide description of domain knowledge, giving a shared understanding of differing conceptualisations used across resources.

ii Ontologies are also used to provide knowledge services to bioinformatics applications. In addition to TAMBIS, the RiboWeb ontology [1] and the EcoCyc ontology [10, 11] both use ontologies to provide database schema. The expressivity of their encoding surpasses that of conventional schema languages and allows data instances to be described with high-fidelity. The RiboWeb ontology also lends knowledge to the analysis of ribosome component structure, guiding a user as to which methods to use and high-lighting contradictions with current knowledge.

These and other bio-ontologies capture a large amount of domain knowledge. The content and conceptualisation contained in these ontologies is undoubtedly skewed by the use to which the ontology is put [15]. In spite of this, due to the effort needed to create an ontology, there is a desire to re-use ontologies. Regardless of the difficulties of differing conceptualisations, the encoding of ontologies presents something of a heterogeneity barrier of its own. There are many types of knowledge representation (KR) language, that differ in expressivity, formality and rigour. Phrase-based ontologies are hand-crafted taxonomies of natural language terms, with variable numbers of cross-taxonomy relationships. The Gene Ontology is an example of a phrase-based ontology. Such ontologies are expressive, as they use natural language forms, but lack formality and rigour and can, therefore, sometimes be difficult to re-use. Frame-based systems are again hand-crafted, but take an object view of the world, where concepts are classes and properties of concepts are expressed as relationships to other classes. They are more formal, but vary in their expressivity. Their object view of the world is, however, intuitive and attractive to many modellers. Description logics, such as that lying behind OIL, are expressive and have a strong, formally defined semantics. The reasoning available with DLs afford them a rigour not seen in other KR languages. The logical form, however, is often seen as difficult to use and is unattractive to many users.

OIL, however,seeks to combine the best of the frame-based and DL modelling paradigms. OIL uses a frame-based structure, where concepts are described in terms of classes and their slots or properties. Classes and slots can, however, be described with highly expressive concept expressions (such as that seen for cofactor above). An underlying DL reasoning service can be used to check the logical consistency of the definitions and infer a subsumption hierarchy from the concept descriptions. So, OIL combines the modelling perspective of frame-based systems, with the expressivity and rigour of a DL system, together with the offer of reasoning support. This support is optional, so purely hand-crafted, phrase-based ontologies can be encoded, as simulacra of frame-based ontologies such as RiboWeb, as well as DL ontologies such as TAMBIS. Such OIL encoded ontologies can evolve from a less to a more rigourous, descriptive form. As well as suiting a wide range of modelling styles and formalisms, OIL can be mapped to and from several current encodings, making it suitable as an exchange language. This should facilitate the wider use of individual

ontologies.

One of the visions for the web is to transform it from a machine readable, human understandable, to a machine readable and understandable or processable resource [5]. This would mean that applications and agents could use information on the web, without the intimate intervention of human users. Central to this semantic web are the use of ontologies to describe content of web resources. Two languages, OIL and the DARPA Agent Markup Language (DAML) have been combined into a language DAML+OIL [8] (`http://daml.org`), which is proposed as a standard encoding, management and delivery mechanism for ontologies on the web. Just as the W3C recognise knowledge as the key to the formation of the semantic web, knowledge is also the key to semantic bioinformatics resources. Due to their machine processability, knowledge rich bioinformatics resources will inter-operate seamlessly, within a variety of applications to enhance the usefulness of bioinformatics resources.

Acknowledgements: Robert Stevens is supported by BBSRC/EPSRC grant 4/B1012090.

References

[1] R. Altman, M. Bada, X.J. Chai, M. Whirl Carillo, R.O. Chen, and N.F. Abernethy. RiboWeb: An Ontology-Based System for Collaborative Molecular Biology. *IEEE Intelligent Systems*, 14(5):68–76, 1999.

[2] P.G. Baker, C.A. Goble, S. Bechhofer, N.W. Paton, R. Stevens, and A Brass. An Ontology for Bioinformatics Applications. *Bioinformatics*, 15(6):510–520, 1999.

[3] S. Bechhofer, A.L. Goble, C.A.and Rector, W.D. Solomon, and W.A. Nowlan. Terminologies and Terminology Servers for Information Environments. In *Proceedings of the IEEE 8th International Conference on Software Technology and Engineering Practice*, pages 35–42, 1997.

[4] Sean Bechhofer, Robert Stevens, Gary Ng, Alex Jacoby, and Carole Goble. Guiding the User: An Ontology Driven Interface. In Norman W. Paton and Tony Griffiths, editors, *Proc. User Interfaces to Data Intensive Systems (UIDIS99)*, pages 158–161. IEEE Press, September 1999.

[5] T. Berners-Lee. *Weaving the Web*. Orion Business Books, 1999.

[6] D. Fensel *et al.* OIL in a nutshell. In *Proc. of EKAW-2000*, LNAI, 2000.

[7] C.A. Goble, R. Stevens, G. Ng, S. Bechhofer, N.W. Paton, P.G. Baker, M. Peim, and A. Brass. Transparent Access to Multiple Bioinformatics Information Sources. in press, 2001.

[8] J. Hendler and D. L. McGuinness. The DARPA agent markup language. *IEEE Intelligent Systems*, jan 2001.

[9] I. Horrocks. Using an Expressive Description Logic: FaCT or Fiction? In A.G.Cohn, L.K. Schubert, and S.C.Shapiro, editors, *Principles of Knowledge Representation and Reasoning: Proceedings of the Sixth International Conference (KR'98)*. Morgan Kaufmann Publishers, San Fransisco, CA, 1998.

[10] P. Karp and S. Paley. Integrated Access to Metabolic and Genomic Data. *Journal of Computational Biology*, 3(1):191–212, 1996.

[11] P.D. Karp, M. Riley, M. Saier, I.T. Paulsen, S.M. Paley, and A. Pellegrini-Toole. The EcoCyc and MetaCyc Databases. *Nucleic Acids Research*, 28:56–59, 2000.

[12] N.W. Paton, R.D. Stevens, P.G. Baker, C.A. Goble, S. Bechhofer, and A. Brass. Query Processing in the TAMBIS Bioinformatics Source Integration System. In *et al*. Z.M. Ozsoyoglo, editor, *Proc. 11th Int. Conf. on Scientific and Statistical Database Management (SSDBM)*, pages 138–147, Los Alamitos, California, July 1999. IEEE Press.

[13] A. Rector *et al.* The GRAIL concept modelling language for medical terminology. *Artificial Intelligence in Medicine*, 9:139–171, 1997.

[14] S. Schulze-Kremer. Ontologies for Molecular Biology. In *Proceedings of the Third Pacific Symposium on Biocomputing*, pages 693–704. AAAI Press, 1998.

[15] R. Stevens, C.A. Goble, and S. Bechhofer. Ontology-based Knowledge Representation for Bioinformatics. *Briefings in Bioinformatics*, 1(4):398–416, November 2000.

[16] R. Stevens and C. Miller. Wrapping and Interoperating Bioinformatics Resources Using CORBA. *Briefings in Bioinformatics*, 1(1):9–21, 2000.

[17] The Gene Ontology Consortium. Gene Ontology: Tool for the Unification of Biology. *Nature Genetics*, 25:25–29, 2000.