Complex Query Formulation Over Diverse Information Sources Using an Ontology

Robert Stevens[1,2], Carole Goble[1], Norman Paton[1],
Sean Bechhofer[1], Gary Ng[1], Patricia Baker[2] and Andy Brass[2]
Department of Computer Science[1],
School of Biological Sciences[2]
University of Manchester
Oxford Road, Manchester M13 9PL, UK
tambis@cs.man.ac.uk

Abstract

Biologists increasingly need to ask complex questions over the large quantity of data and analysis tools that now exist. To do this, the individual sources need to be made to work together. The knowledge needed to accomplish this places a barrier between the bench biologist and the question he or she wishes to ask. The TAMBIS project (Transparent Access to Multiple Bioinformatics Information Sources) has sought to remove these barriers – thereby making the process of asking questions against multiple sources transparent. Central to the TAMBIS system is an ontology of biological terms. This allows TAMBIS to be used to formulate rich, complex queries over multiple sources. The ontology is constructed in a manner that ensures only biologically meaningful queries can be posed. TAMBIS then uses the ontology to transform the declarative, source-independent query into an optimised, ordered sequence of source dependent requests, that is then executed against the individual sources.

1 Introduction

Molecular biology is a data rich discipline, holding a vast quantity of sequence and other data. Most of these are held in different databanks and separate analysis

tools. These information sources are automonous and distributed across heterogeneous platforms and have differing call interfaces and query interfaces (where they exist). Added to this is the semantic heterogeneity between the many data-sources and analysis tools [5].

Many bioinformatics tasks may be supported by the individual sources, but increasingly, biologists wish to ask rich, complex questions, that span a range of the sources available. This places a barrier between a biologist and the task he or she wishes to accomplish; namely having to know what source to use, the location of that source, how to use that source (both syntactically and its semantics) and how to transfer data between the sources. In addition, the biologist has to use the sources in the appropriate order and manage the semantic heterogeneity between the sources during transfer of data.

TAMBIS attempts to avoid these pitfalls by using an ontology of molecular biology and bioinformatics to manage the presentation and usage of the sources. This ontology allows TAMBIS to have the following attributes: A homogenising layer over the numerous databases and analysis tools; an opportunity to manage the semantic heterogeneity between the datasources; and a common, consistent query-forming user interface that allows queries across sources to be precisely expressed and progressively refined.

TAMBIS represents knowledge of what concepts exist in these domains and the relationships that exist between those concepts. It is this knowledge that TAMBIS uses to give transparent access to a wide range of bioinformatics databanks and tools. The TAMBIS ontology is used for retrieving instances represented by concepts in the model. A concept is a description of a set of instances, so a concept or description can also be viewed as a query.

This approach allows a biologist to ask questions such as 'find all antigenic human apoptosis receptor protein homologues and their phosphorylation site motifs'. This query is answered by the three sources Swiss-prot, Blast and Prosite. The user does not have to choose the sources, the key-words with which to filter the proteins etc.

Other systems, such as SRS [4], Bio-Kleisli/CPL (Collection Programming Language) [3] and OPM [5], have attempted to integrate multiple bioinformatics sources, but have done so less transparently. All have the integration and reconciliation of some aspects of heterogeneity in common. They all, however, still leave the user to choose sources, the attributes to be used and order of query subcomponents in complex tasks. All can also allow biologically meaningless queries to be formed.

The remainder of this paper is organised as follows. Section 2 describes the TAMBIS ontology, its underlying representation and how the ontology is central

to the TAMBIS architecture. A worked example of TAMBIS in use is given in Section 3 and concluding remarks appear in Section 4.

2 The TAMBIS Ontology and System Architecture

An ontology is a system that describes concepts and their relationships. The TAMBIS ontology is expressed in a Description Logic (DL) [2], which are a type of knowledge representation language. One of the key features for TAMBIS is that these representations offer reasoning services about concepts and their relationships to their clients. As well as the traditional 'is a kind of' relationships (a `Protein` is a kind of `Biopolymer`), there are partitive, locative and nominative relationships etc. This means the TAMBIS ontology can describe relationships such as: 'Motifs are parts of proteins'; 'Organelles are located inside cells' and a 'gene has a name gene name'. The ontology initially holds only asserted concepts, but these can be joined together, via relationships to form new, compositional concepts. For example, `Motif` can be joined to `Protein` using the relationship `is component of` to form a new concept `Protein motif`. The ontology is a dynamic model, what is present in the model is the description or potential of what concepts can be formed in the domain of molecular biology and bioinformatics. As these new, compositional concepts are described they are automatically placed within the lattice of existing concepts by the DL reasoning services. For example, the compositional concept `Protein motif` is automatically classified as a kind of `Motif`. This new concept is then available to be re-used in further compositional concepts.

The TAMBIS ontology is described using a DL called GRAIL [7]. In spite of its inexpressiveness, the GRAIL representation has a useful property in its ability to describe constraints about when relationships are allowed to be formed. For example, it is generally true that a `Motif` is a component of a `Biopolymer`, but not all `Motifs` are, however, components of all biopolymers. An α`-helix` is only a component of `Protein`, but not `Nucleic acid`, both of which are `Biopolymers`. this constraint mechanism allows the TAMBIS model to capture this distinction, and thus only allow the description of biologically meaningful concepts. It prevents `Restriction site` being described as part of `Protein`, but allowing α`-helix`, whilst both are still `Motif`s and thus kinds of `Biopolymer`.

The TAMBIS ontology describes both molecular biology and bioinformatics. Concepts such as `Protein` and `Nucleic acid` are part of the world of molecular biology. An `Accession number` lies outside this domain, but is essential for describing bioinformatics tasks in molecular biology. The TAMBIS ontology

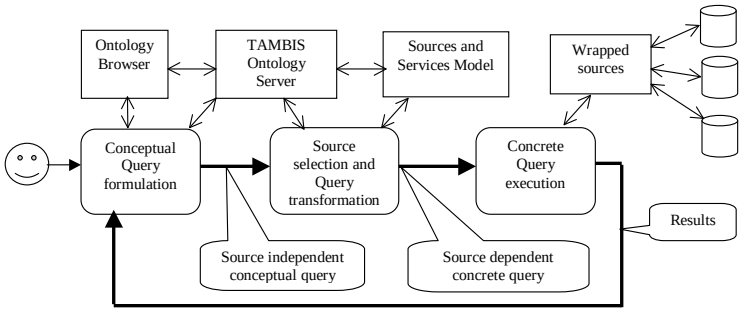

Figure 1: The flow of information through the TAMBIS architecture.

has been designed to cover the standard range of bioinformatics tasks of retrieval and analysis. This means that a broad range of biology has been described. The model is, however, currently shallow as the detail present is enough to allow most tasks to be described. Whilst it is possible to make more specific concepts through composition, it is likely that the model will have to become more detailed.

The model is centred about the biopolymers `Protein` and `Nucleic acid` and their obvious children, such as `Peptide`, `Enzyme`, `DNA`, `RNA`, etc. Biological functions and processes are described to detail such as `Receptor`, `Cytokine`, `Lactation` and the top-level `Enzymic function`s. Sequence components such as protein motifs and structure are described in accordance with standard databases and classifications. Sequence components also contain concepts such as `Gene`, `Ribosome binding site`, `Restriction sites`, etc. and the protein secondary structures.

These basic concepts are present in the 'is a kind of' hierarchy. Other relationships add richness to the model such that a wide range of biological features can be described. For example, `Motif` (and its children) can be components of `Protein` or `Nucleic acid`. `Protein` can have both `Biological function` and `Biological process` via the 'has function' and 'functions in process' relationships. `Protein`, for example, can 'have homology to' both `Protein` and `Nucleic acid`, and 'be expressed by' a `Gene`. Proteins also 'have accession number' and 'have name'. The model is difficult to describe on paper, but is reviewed in more detail in [1] and can be browsed via an applet on the TAMBIS WEB SITE (`http://img.cs.man.ac.uk/tambis`).

The TAMBIS ontology is supplied as a software component that acts as a server. Other components can ask questions of the knowledge in the ontology component. It is the back-bone of the architecture, all other components either

directly or indirectly using the ontology. These other components ask questions such as: 'Is this a concept'; 'what are the parents, children or siblings of this concept'; 'which relationships are held by this concept' and 'what is the English version of this concept'. Figure 1 shows how a conceptual query is processed through the TAMBIS system and how it is transformed to a concrete query plan using the ontology component.

In brief, the processing of a TAMBIS query is as follows:

- A query is formulated in the query builder window on the user interface, which is created from knowledge in the ontology. Concepts can be either specialised or generalised by the user.
- This declarative, source-independent conceptual query is then submitted to the query processor. It is rewritten to an optimised, source-dependent concrete query plan expressed in the middleware language CPL [6].
- Sources and services wrappers, written in CPL, enable external sources to be included into the system and transformed to provide a syntactically consistent interface for the query processor. The sources currently used in TAMBIS are Swiss-Prot, Enzyme, Cath, Blast and Prosite.
- The sources and services model (SSM) acts as a catalogue for the query processor, indexed by the ontology. As well as mappings from concepts to their representations in the source, the model maps concepts to functions in the CPL wrappers. Both the SSM and CPL source wrappers manage some of the heterogeneities existing between the sources.

3 TAMBIS in Use

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.

For each of the queries above, the concepts involved can be replaced by other concepts, e.g. `Lipidation site` for `Phosphorylation site`. In addition, each query can be either further added to in order to specialise, or removed from, to generalise, the scope of the query. The model can also be explored in order to show what it is possible to ask, and to serve as a tutorial, in molecular biology.

TAMBIS is used to formulate the query "find all antigenic human, apoptosis receptor homologues of the protein called 'lymphocyte associated receptor of death' " in the following manner. Figure 2 shows the conceptual form of this query in the TAMBIS user interface[1].

First, the *topic* of the query is described. In this case, it is `Protein` of a particular type that needs to be retrieved. The description `Protein` is a query that means 'find all known proteins', a query resulting in 80 000 entries being returned. The next task is to *restrict* the `Protein` to be a certain 'kind of' `Protein`. The restrict menu consults the ontology to ask what can be asked about the concept `Protein`. The user interface can use the ontology to display only what it is possible to ask about any concept in any context within the model. One of the restrictions allowed is that a `Protein` can be `homologous to protein`. Choosing this option gives a new kind of conceptual description of proteins that are homologous to protein, a more refined query, but no less huge.

The second protein (the *filler* of the relationship) can then be further *restricted* to be a particular protein, by choosing `has name protein name` from the restriction menu. The user interface allows a value to be set for certain attributes, such as `Protein name` and in this case it is set to 'lymphocyte associated receptor of death'. Now, instead of having all protein homologues described, only homologues of this particular protein are described.

The desired query retrieves all homologues, but only those from a certain `Species`, with a certain `Function` and involvement in a certain `Process` are required. The topic is then specialised, first by *restriction* to come from a certain species and then to have a particular `Biological function` and in a `Biological process`. The latter two concepts are general concepts, with

[1] A video showing this query being formed and answered can be found at `http://img.cs.man.ac.uk/tambis`

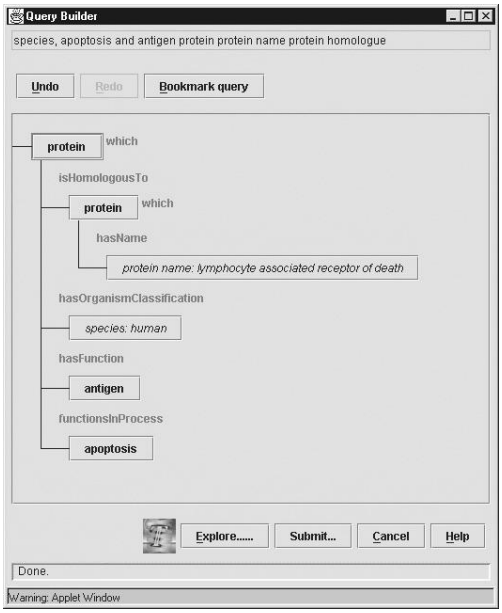

Figure 2: The TAMBIS user interface showing the conceptual query "human, antigenic apoptosis homologues of the protein 'lymphocyte associated receptor of death' ".

many children. These general concepts are then *replaced* with a specific kind of process or function, such as `Apoptosis` or `Antigen` respectively.

The query may be submitted for rewriting and evaluation at any stage, refinements being made after each submission. It is important to note that the building of such a query can take place in almost any order. This ability to progressively refine a query to a precise description, but only being able to make biologically meaningful changes, distinguishes TAMBIS from other source integration systems.

4 Conclusions

TAMBIS has been able to use an ontology of biological concepts to manage the integration of diverse bioinformatics sources. The ontology is the back-bone of the system, driving both the user interface and the query transformation process. It gives the illusion of a common user interface to many sources. The knowledge in the system allows the choice of which source to use, where that source is located,

how to use that source and how to transfer data, to be taken away from the user. This leaves the biologist free to build rich and precise queries to either retrieve the data or perform the analyses he or she wishes. **Acknowledgements:** This work is funded by AstraZeneca Pharmaceuticals and the BBSRC/EPSRC Bioinformatics programme, whose support we are pleased to acknowledge.

References

[1] Patricia G. Baker, Carole A. Goble, Sean Bechhofer, Norman W. Paton, Robert Stevens, and Andy Brass. An Ontology for Bioinformatics Applications. To Appear in: Bioinformatics, 1999.

[2] A. Borgida. Description Logics in Data Management. *IEEE Trans Knowledge and Data Engineering*, 7(5):671–782, 1995.

[3] P. Buneman, S.B. Davidson, K. Hart, C. Overton, and L. Wong. A data transformation system for biological data sources. In *Proceedings of VLDB*, Sept 1995.

[4] T. Etzold, A Ulyanov, and P Argos. SRS: Information Retrieval System for Molecular Biology Data Banks. *Methods in Enzymology*, 266:114–28, 1996.

[5] V.M. Markowitz and O. Ritter. Characterizing Heterogeneous Molecular Biology Database Systems. *Journal of Computational Biology*, 2(4), 1995.

[6] 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. To Appear in Scientific and Statistical Databases, 1999.

[7] A. L. Rector, S. K. Bechhofer, C. A. Goble, I. Horrocks, W. A. Nowlan, and W. D. Solomon. The GRAIL Concept Modelling Language for Medical Terminology. *Artificial Intelligence in Medicine*, 9:139–171, 1996.

Index