TAMBIS: Transparent Access to Multiple bioinformatics Information Sources

R. D. Stevens, P. G. Baker, S. Bechhofer, G. Ng,
A. Jacoby, N. Paton, C. A. Goble and A. Brass
Department of Computer Science and School of biology
the University of Manchester
Oxford Road
Manchester
M13 9PL

Abstract

Summary: TAMBIS is an application that allows biologists to ask rich and complex questions over a range of bioinformatics resources. It is based on a model of the knowledge of the concepts and their relationships in molecular biology and bioinformatics. Availability: TAMBIS is available as an applet from `http://img.cs.man.ac.uk/tambis`. Supplementary Information: A full user manual, tutorial and videos can be found at `http://img.cs.man.ac.uk/tambis`. Contact: `tambis@cs.man.ac.uk`.

Keywords: Transparent access; multiple sources; complex queries.

TAMBIS is an application for performing bioinformatics tasks over multiple information sources, such as databases and analysis tools, with a single, interactive user interface. TAMBIS is, however, unlike other tools, in that a user does not have to choose sources, analysis tools and attributes to retrieve information, and does not need to decide the order in which to perform sub-tasks of complex queries. This prototype version of TAMBIS, which only accesses five sources (Swiss-Prot, Enzyme, Cath, blast and Prosite), can ask questions that cannot be asked by other systems. For example, retrieve 'Apoptosis receptor protein Phosphorylation sites' or find all the 'human seven propeller domain proteins'.

The key feature of TAMBIS is its use of a model of knowledge in the world of bioinformatics and molecular biology. This is the knowledge of what 'things' exist in bioinformatics and molecular biology and how these 'things' are related to each other. TAMBIS allows a biologist to browse through this model to see what it is possible to say about 'things' or *concepts* in molecular biology. As well as exploring what can be said about concepts, a biologist can join concepts together by their relationships to build new concepts. These new concepts are descriptions of 'things' in the world of bioinformatics. TAMBIS takes this description, goes to the relevant resources, collects the results, and displays them in a web browser. TAMBIS chooses the resources to use to answer the question and the order in which to use them. The user does not need this knowledge – the system is *transparent*. One of the features of this domain knowledge is that at each stage in the building of a description or query, TAMBIS will show only what it is biologically sensible to ask.

There are two versions of TAMBIS available:

1. An *unlinked* applet that allows the user to browse around a large model, containing about 1 800 concepts of bioinformatics and molecular biology, and form queries covering proteins and nucleic acids, their motifs, functions and the processes in which they take part. It also covers enzymes, reactions, protein structure, and allows a biologist to ask about homology, parts of sequences, the standard features of sequence database entries, etc. Whilst queries may be formed, they cannot be submitted to external sources for evaluation.

2. An applet, *linked* to external resources, that can be used to browse the model and answer questions. This applet uses a smaller model covering only proteins. This model can answer any query that can be made using the 250 concepts and relationships in this model.

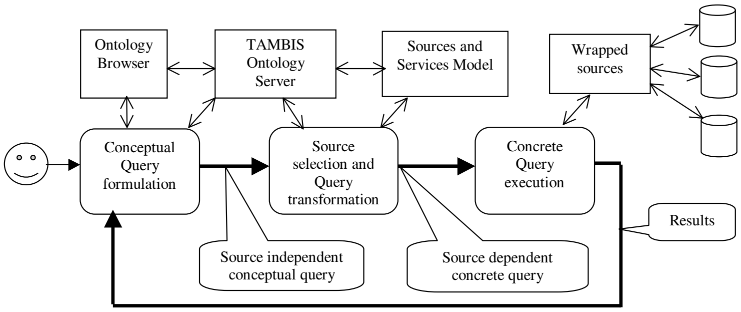

Figure 1: The TAMBIS system architecture, showing the processing of a conceptual query into a query plan.

The TAMBIS model has relatively few concepts and relationships asserted, enough to describe the basic knowledge in the domain. The knowledge representation techniques used, however, mean that many more concepts can be generated. Concepts can be joined, via their relationships, to form new concepts. For example, the concepts `Motif` and `Protein` may be joined by the relationship `is component of` to form a new concept `Protein motif`. This is a new *kind of* motif, and will be automatically placed at the correct location in the existing 'is a kind of' classification that forms the TAMBIS model. Part of the knowledge in the model is when it is biologically sensible for relationships to hold between concepts. This means that the model can describe that bio-polymers can have motifs, but not all motifs occur in all types of bio-polymer; some are specific to nucleic acids and some to proteins. So, only biologically sensible concepts can be described. Once made, new concepts are placed in the model and hold the appropriate relationships and properties as determined by their location.

Figure 1 shows the TAMBIS system architecture, with the model server component forming the backbone of the system, offering knowledge services to the remainder of the system. This enables TAMBIS to generate its user interface; map concepts to functions that retrieve instances of those concepts and rewrite the conceptual, source independent query to the concrete, source dependent query plan. To ask a question, a user describes a new concept. the user can browse the model, choosing the concepts to add to the description of what he or she wishes to retrieve. For any concept the user interface will only display those concepts and relationships which can be held by that concept in that context. To do this, the user interface component has to ask the model what are the parents, children and siblings of the concept, as well as what relationships to which other concepts are valid.

TAMBIS is one of several systems that attempt to allow queries to be formed over multiple information sources, such as SRS [3], OPM [4] and CPL [2]. The principal difference between TAMBIS and these systems is the level of transparency. TAMBIS is completely transparent, the user need only express their biological goal in order to form a query. TAMBIS uses the conceptual model to give the illusion of a common user interface and manage the processing of the query.

The TAMBIS WWW release thus allows access to two distinct bioinformatics resources:

1. A substantial ontology of bioinformatics concepts [1]. This illustrates how a wide range of biological concepts can be modelled using a knowledge representation system.

2. A novel query system that builds upon the domain model, to allow a wide range of queries to be addressed to 5 popular bioinformatics sources.

Acknowledgements: This work is funded by Zeneca 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] 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.

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

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