Towards a Semantic Web for Bioinformatics

Rolf Backofen[1], Mike Badea 2, Pedro Barahona[3], Liviu Badea[4] François Bry[5], Gihan Dawelbait [6], Andreas Doms [6], François Fages[7], Carole Goble[2], Andreas Henschel [6], Anca Hotaran [4], Bingding Huang [6], Ludwig Krippahl [3], Patrick Lambrix [8], Werner Nutt [9], Michael Schroeder [6], Sylvain Soliman [7], Sebastian Will [1]

[1] *Friedrich-Schiller-Universität Jena, Germany,* [2] *Victoria University of Manchester, UK,* [3] *Universidade Nova de Lisboa, Portugal,* [4] *National Institute for Research and Development in Informatics, Bucarest, Romania,* [5] *Ludwig-Maximilians-Universität München, Germany,* [6] *Technical University of Dresden, Germany,* [7] *INRIA Rocquencourt, France,* [8] *Linköpings Universitet, Sweden,* [9] *Heriot-Watt University, Edinburgh, UK*

Corresponding author: Michael Schroeder, ms@mpi-cbg.de

Abstract

With the explosion of online accessible bioinformatics data and tools, systems integration has become very important for further progress. Currently, bioinformatics relies heavily on the Web. But the Web is geared towards human interaction rather than automated processing. The vision of a Semantic Web facilitates this automation by annotating web contact and by providing adequate resaoning languages.

In this abstract, we motivate why bioinformatics needs a semantic web and why the semantic web needs bioinformatics. We briefly sketch the bioinformatics effort of the REWERSE project (www.rewerse.net), which is concerned with developing automated reasoning methods and tools for the Web and its application to bioinformatics.

1 Introduction

With recent technological advances, biology has changed dramatically to a data-driven science. Projects like the human genome project, which resulted in a publicly accessible database containing the whole human DNA, are only single examples of an explosion in biological information online accessible. There are hundreds of databases and bioinformatics tools online with ten thousands of 3D structures of proteins, with hundred thousands of protein sequences and with millions of literature abstracts [4]. The current bottleneck upon which future progress in biology depends is the coherent integration of all these public, online resources.

Currently biologists heavily rely on the web to analyse their data.

1.1 The Web

Tim Berners-Lee's seminal vision of the Web, i.e. a platform for exchanging between computers documents related to each other by hypertext links, is immensely successful. Information of any kind is usually better published on the Web than on any other medium because, on the Web, information systems can be accessed from everywhere at any time. This makes updates immediately visible everywhere and thus makes the Web an ideal platform for information systems. Furthermore, the Web is a very flexible framework for data modeling because its premier data modeling languages, HTML and XML, offer (possibly nested) record-like data structures, the so-called HTML or XML "documents", without requiring compliance to any data schema. Furthermore, with XML tags can be freely chosen what makes self-explaining data items easy to specify. In addition, the Web does support the coexistence of very heterogeneous data models, which greatly contributes to its

success. Thus, the Web is convenient for exchanging all kinds of data and it is intensively used for that purpose.

1.2 The Semantic Web

Although biologists heavily use the web, it is inadequate, as it does not allow users to easily integrate different data sources, to incorporate additional analysis tools and to include the user's background in the analysis. This problem is address by the vision of a Semantic Web. The Semantic Web labels contents with "semantic annotations" and uses these annotations for an automated retrieval and composition of Web contents. Keeping with the tolerance to heterogeneity which has been an essential factor in the success of the traditional Web, the Semantic Web vision does not impose any particular form for semantic annotations. In fact, very different formalisms have already emerged like RDF (cf. http://www.w3.org/RDF/) – a variation on the Entity-Relationship database schema language –, OWL (cf. http://www.w3.org/TR/owl-features/), an ontology language stemming from description logics, and thesauri initially developed for information retrieval and/or natural language processing. Current research still needs to integrate such approaches with rules and reasoning capabilities to create true flexibility.

2 Bioinformaics and the Semantic Web

Bioinformatics is an ideal field for testing Semantic Web technologies for three reasons: First, Web-based systems and Web databases have been applied very early in bioinformatics [4], second the dramatic increase of data produced in the field calls for novel processing methods, third, the high heterogeneity of bioinformatics data require semantic-based integration methods.

2.1 Scenario

Consider the following scenario: a biologist obtains a novel DNA sequences nothing is known about. He or she wants to run an alignment, but has specific requirements for the alignment. These requirements are captured as rules and constraints, which are taken into account by the online accessible Semantic Web enabled sequence comparison service.

The researcher found a number of significantly similar sequences in yeast for which there is gene expression data available. The scientist requests from the Semantic Web enabled gene expression database and tool expression data for the relevant genes. He or she defines rules, which capture which expression profiles are interesting, e.g. all genes which are highly expressed at the beginning and end of the experiment are of interest.

The genes are part of a larger process and the researcher is interested in their gene products. A query to the protein database SWISSPROT determines these. Do these proteins interact with each other? To answer this question a Semantic Web service is queried, which computationally determines protein interactions. A user-defined rule formulating what constitutes a protein domain interaction, is applied on the fly to SCOP, the structural classification of proteins, and PDB, a large protein structure database. The rule-based sequence similarity tool mentioned above is used to determine whether the scientists proteins of interest are similar to any interacting proteins computed from SCOP and the PDB.

Finally, the scientist wishes to relate the protein interaction network to metabolic pathways. As all the tools used refer to the same ontologies and terminology defined through the gene ontology, the researcher can easily investigate a mapping from the interaction network to a relevant metabolic pathway obtained from a Semantic Web enabled pathway server.

During the above information foraging, the scientist constantly used literature databases to read relevant articles. Despite the tremendous growth of 8000 articles a week, the biologist still manages to quickly find the relevant articles as he or she uses an ontology-based search facility, which guides the search, automatically specializing querying, where too many hits are obtained, and generalizing, where too few articles can be found.

2.2 Rules for systems integration and modeling

One approach to tackle the challenges in bioinformatics in general and the scenario above in particular, will rest upon the use of rules, reasoning and ontologies. Broadly, these can be applied in two contexts:

1. Systems integration: Rules for mediation, including consistency, and to formulate complex queries.
2. Modeling: Rules to model biological systems.

Currently much interaction with online data sources is manually through HTML pages [4]. However, many online Bioinformatics tools already provide XML output, so that more complex querying and interaction is possible. These could be simple queries to a single XML document retrieving some of its attributes (e.g. given an XML document representing a protein structure, query the protein document for a domain of that protein). The above structures could be complemented by similar sequences for which the structure has been predicted using rules and constraints.

Queries can be complex, involving different sources and ontologies. E.g. use a rule that specifies if a search for literature in the medical literature database PubMed retrieves too many (or too few) results, then find keywords in gene ontology and specialize (or generalize) them and re-issue the query to PubMed. As another example for complex queries, consider the Biomolecular interaction network (BIND), which already provides some output in XML format. The problem of defining interactions is very complex, and enhanced expressivity would be required to define the different classes of interactions. Interactions can then be deduced from several sources, such as PDB, metabolic/regulative pathways or networks, etc. Rules can be used again to model these networks and to query them.

The problems above have already begun to be addressed, in particular regarding transparent access to bioinformatics databases [9] and integration protein annotations [11, 12, 10]. In depth experience in merging bioinformatics ontologies [8], integration of an ontology with gene expression data [2, 3], and consistent annotation of proteins [11, 12, 10] have already been gathered. Relevant work concerning protein structure prediction with constraints [1, 7, 13] and concerning the ability to flexibly query data in the form of networks [5] and in the light of mismatches [6] between concepts have already been carried out.

3 REWERSE

The above problems and scenarios are investigted as part of REWERSE, a research "Network of Excellence" of the 6th Framework Programme of the EU Commission. REWERSE stands for "REasoning on the WEb with Rules and SEmantics". REWERSE will develop a miniaml set of rule-based languages catering for reasoning on the web. Besides teh bioinformatics application working group, the project has technology working groups dedicated to query languages, reactivity and evolution of data, composition of rules, policies and standards.

The overall objective of REWERSE is to strengthen Europe in the area of reasoning languages for Web systems and applications, especially Semantic Web systems and applications aiming at enriching the current Web with reasoning capabilities as described above. REWERSE will also develop university education and training as well as technology transfer and awareness activities so as to spread excellence within its research field in Europe. REWERSE involves 27 European research and industry organizations and about 100 computer science researchers and professionals and runs tunil 2008.

More information on REWERSE can be found at `http://www.rewerse.net`.

Acknowledgement

This research has been funded by the European Commission within the 6th Framework Programme project REWERSE number 506779 (cf. `http://www.rewerse.net`).

The authors of the poster are the organisations' representatives of the working group. Additional members of REWERSE's bioinformatics working group are

References

[1] Rolf Backofen, Sebastian Will, and Erich Bornberg-Bauer. Application of Constraint Programming Techniques for Structure Prediction of Lattice Proteins with Extended Alphabets. *Journal of Bioinformatics*, 15(3):234–242, 1999.

[2] Liviu Badea. Functional Discrimination of Gene Expression Patterns in Terms of the Gene Ontology. In *Proceedings of the Pacific Symposium on Biocomputing (PSB)*, 2003.

[3] Liviu Badea and Doina Tilivea. Integrating Biological Process Modelling With Gene Expression Data and Ontologies for Functional Genomics (Position Paper). In *Proceedings of the International Workshop on Computational Methods in Systems Biology*, University of Trento, 2003. Springer-Verlag.

[4] François Bry and Peer Kröger. A Computational Biology Database Digest: Data, Data Analysis, and Data Management. *Distributed and Parallel Databases*, 13(1):7–42, 2002.

[5] Nathalie Chabrier and François Fages. Symbolic Model Checking of Biochemical Networks. In *Proceedings of the 1st International Workshop on Computational Methods in Systems Biology (CMSB)*, LNCS, Riverto, Italy, March 2003. Springer-Verlag.

[6] David Gilbert and Michael Schroeder. FURY: Fuzzy Unification And Resolution Based on Edit Distance. In *Proceedings of BIBE2000 - IEEE International Symposium on Bio-Informatics and Biomedical Engineering*. IEEE Press, 2000.

[7] L. Krippahl and P. Barahona. PSICO: Solving Protein Structures with Constraint Programming and Optimisation. *Constraints*, 7(3/4):317–331, July/October 2002.

[8] P. Lambrix and A. Edberg. Evaluation of Ontology Merging Tools in Bioinformatics. In *Proceedings of the Pacific Symposium on Biocomputing (PSB)*, pages 589–600, 2003.

[9] P. Lambrix and V. Jakoniene. Towards Transparent Access to Multiple Biological Databanks. In *Proceedings of the 1st Asia-Pacific Bioinformatics Conference*, pages 53–60, Adelaide, Australia, 2003.

[10] S. Möller, E. V. Kriventseva, and R. Apweiler. A Collection of Well Characterised Integral Membrane Proteins. *Bioinformatics*, 16(12):1159–1160, 2000.

[11] S. Möller, U. Leser, W. Fleischmann, and R. Apweiler. EDITtoTrEMBL: a Distributed Approach to High-Quality Automated Protein Sequence Annotation. *Bioinformatics*, 15(3):219–227, 1999.

[12] S. Möller, M. Schroeder, and R. Apweiler. Conflict-Resolution for the Automated Annotation of Transmembrane Proteins. *Comput. Chem.*, 26(1):41–46, 2001.

[13] P.N. Palma, L. Krippahl, J.E. Wampler, and J.J.G. Moura. BiGGER: A New (Soft) docking Algorithm for Predicting Protein Interactions. *Proteins: Structure, Function, and Genetics*, 39:372–384, 2000.