XML-based approaches for the integration of heterogeneous bio-molecular data

M. Mesiti[1], E. J. Ruiz[2], I. Sanz[2], R. Berlanga[2], G. Valentini[1], P Perlasca[1], D. Manset[3]

[1]*University of Milano, Italy*
Email:{mesiti@dico,perlasca@dico,valentini@dsi}.unimi.it
[2] *Universitat Jaume I, Castello, Spain*
Email: Ernest.Jimenez@alumail.uji.es, isanz@icc.uji.es, berlanga@lsi.uji.es
[3] *Maat Gknowledge, Valencia, Spain*
Email: dmanset@maat-g.com

Abstract: *Experimentation with biological systems produces sources of huge amounts of biological data. In order to make these heterogeneous data sources easy to use, several efforts on data representation and integration have been considered mainly based on XML. In this paper we present the main approaches proposed in literature and we discuss future research directions.*

Keywords: Bio-molecular data types, XML, data representation, data integration.

1. Introduction

Several research institutions produce huge amounts of biological data coming from the experimentation with biological systems. The proper exploitation of these data sources requires great efforts in integrating disparate data structures, protocols and tools.

A standard unified model for the description of data and, consequently, a format for their exchange and representation that is machine readable may face the heterogeneous data integration issues. Several data formats intended to represent biological entities and systems have been reviewed [8] that can be categorized in: textual-based, XML-based, and ontology based. The main problem of the textual-based proposals is the lack of structure consistency, whereas the use of XML overcomes this problem by using a standard data format with a precise structure defined by a DTD or an XML Schema (XSD). Finally, ontology based formats are arising as a solution to increase the content semantics and to formally represent the knowledge to be exchanged.

In this paper we present a survey of some XML-based approaches for bio-molecular data representation and we discuss their main issues and so far proposed approaches for their integration. Finally, we present future research directions.

2. XML representation of bio-molecular data types

XML-like representations have been so far proposed for the (1) principal bio-molecular entities (DNA, RNA and proteins) and their structural properties, (2) biological expression (microarray), and (3) system biology. Initial proposals have been developed within small groups of institutes with the main aim of having a common representation of data structures and language to model their own set of bio-molecular data types, whereas nowadays there are more initiatives to have a wider general agreement by specifying the minimal requirements that such kinds of data structures and languages should have.

XML representation of bio-molecular entities. The Bioinformatic Sequence Markup Language (BSML) [5] describes biological sequences (DNA, RNA, protein sequences) at different granularity levels via sequence data, and sequence annotation. A BSML document usually contains information about how genomes and sequences are encoded, retrieved and displayed. ProXML [3] is used to represent protein sequences, structures and families. A ProXML document consists of an identity section, containing the description of proteins, and the data section, containing properties of such proteins. RNAML [4] has been proposed for the representation and exchange of information about RNA sequences, and their secondary and tertiary structures. A RNAML document can represent RNA molecules as a sequence along with a set of structures that describe the RNA under various conditions or modelling experiments.

XML representation of biological expressions. The MAGE project [7] provides a standard for the representation of microarray expression data to facilitate their exchange among different data systems. MAGE mainly consists of: a data exchange model MAGE-OM (Object Model) and a data exchange format MAGE-ML (Markup Language) according to the standardization project groups responsible of the MIAME (Minimum Information About a Microarray Experiment) and MGED Ontology projects.

XML representation of system biology. The need to capture the structure and content of bio-molecular and physiological systems lead to develop SBML (the System Biology Markup Language), CellML (the Cell Markup Language), BioPAX (the Biological Pathways Exchange Language) and the set of HUPO-PSI (Proteomics Standards Initiative) formats [1]. SBML is used to encode models consisting of biochemical entities (species) linked by reactions to form biochemical networks, whereas, CellML encodes models consisting of a number of more generic components, each described in their own component elements. BioPAX and HUPO-PSI formats are examples of standards used to represent both structure and semantics of biological data based on the use of ontologies as controlled vocabularies providing a non-ambiguous meaning of the domain.

3. Integration of bio-molecular data

Despite the possibility to use standard approaches for the integration of data [2], specific approaches based on the employment of XML in bio-informatics have been proposed:

- Automed [12] targets the problem of multiple and incompatible data types and representation formats by using XML and a simplified version of schema (named

XMLDSS) as a common representation language and a schema type, respectively, supporting the annotations for each source by suitable ontologies.

- SWAMI [9] defines a rich middleware architecture for the integration of different databases, formats and computational resources, based on two layers: the *presentation layer* which receives user requests, and the *workbench core*, which processes the request and returns the result. XML is used as interface among the layers.
- The index-driven integration approach [6] supports queries on several databases simultaneously, whose answers "can be syntactically and semantically heterogeneous with each other" so that "some of them can be exact while others are approximate". The set of partial results extracted from heterogeneous sources is then merged in order to obtain an integrated answer.
- B-Fabric system [11] provides abstraction for *samples* (input data for experiments), which are catalogued and annotated, and *extracts*, which are prepared from samples in the laboratory and subject to measurements. XML is mainly used as a medium for the specification of each component and its mapping throughout the architecture.
- The Pegasys [10] is a workflow management system that includes and integrates several tools for different purpose (i.e. sequence alignment, gene prediction, etc.). Pegasys allows the creation of sequence analysis workflows, described in XML and represented as a DAGs, and the exportation of computational results in General Feature Format (GFF) and GAME XML format to use them for further analyses.

4. Future Trends

As we have seen, the only approach which is purely for XML integration purposes is the index-driven integration approach. In all the other cases, the systems are focused on computational elements, where XML is used as a sort of interfacing language. Still, XML is generally used as a common data format, but the actual data integration between heterogeneous sources is left to hand-coded implementations of ad-hoc components.

Besides this basic limitation, there are some other important issues in data integration which are not addressed by these systems:

- Data security, as the focus of the reviewed systems seems to be the collaboration between localized groups. In general, the extension of the proposed architectures to support distributed teams poses a number of challenges which are not addressed.
- Evolution of data. Given that, in most cases, the underlying data sources are described by hand-made specifications, any change in the structure or semantics of the sources is inherently problematic.
- Efficiency. No attempt is made to use any technique to improve the efficiency of distributed queries, such as maintaining statistics for tuning query execution.
- "Partially incomplete or partially correct" data can still be vital to researchers. The problem of uncertainty and high source heterogeneity should be further coped with.

It can be noticed that conflicts at physical and syntactic levels are almost solved through Internet and XML technologies. However conflicts at the semantic layer are still an open issue for seamless biological data integration. In [6] we have devised some technologies

and experimental systems that try to address some of the shortcomings identified in the previous section. These include: ontology-based systems, which exploit semantic characteristics of XML sources to facilitate their integration; multi-similarity systems, which provide advanced querying systems in the presence of complex sources; grid-based systems, which allow the collaboration of widely distributed teams and resources.

5. References

[1] A. Brazma et al. Standards for systems biology. Nat. Rev. Genet., 7, 593–605. 2006.

[2] A. Halevy et al. Data Integration: The Teenage Years. VLDB: 9-16. 2006.

[3] D. Hanisch et al. ProML - the Protein Markup Language for specification of protein sequences, structures and families. In Silico Biology 2(3): 313 – 324. 2002.

[4] S. Harvey et al. RNAML. A standard syntax for exchanging RNA information. RNA 8(6):707-717. 2002.

[5] M. Hucka et al. The Systems Biology Markup Language (SBML): A Medium for Representation and Exchange of Biochemical Network Models. Bioinformatics, 19(4): 524-531. 2003.

[6] E. Hunt et al. Index-Driven XML Data Integration to Support Functional Genomics. *LNCS* 2994, 95-109. 2004.

[7] MAGE project: http://www.mged.org/Workgroups/MAGE/mage.html

[8] M. Mesiti et al. Data Integration Issues and Opportunities in Biological XML Data Management. Technical report. 2008. Available at:

[9] Rifaieh et al. SWAMI: Integrating Biological Databases and Analysis Tools Within User Friendly Environment. In Data Integration in the Life Sciences 2007, LNCS 4544, 48-58. 2007.

[10] S.P.Shah et al. Pegasys: software for executing and integrating analyses of biological sequences. BMC Bioinformatics 2004, 5:40. 2005.

[11] C. Türker et al. B-Fabric: A Data and Application Integration Framework for Life Sciences Research. In Data Integration in the Life Sciences, LNCS 4544, 37-47. 2007.

[12] L. Zamboulis et al. Bioinformatics Service Reconciliation By Heterogeneous Schema Transformation. In Data Integration in the Life Sciences, LNCS 4544, 89-104. 2007.