ABSTRACT

There is a proliferation of research and industrial organizations that produce sources of huge amounts of biological data issuing from experimentation with biological systems. In order to make these heterogeneous data sources easy to use, several efforts at data integration are currently being undertaken based mainly on XML. Starting from a discussion of the main biological data types and system interactions that need to be represented, the authors deal with the main approaches proposed for their modelling through XML. Then, they show the current efforts in biological data integration and how an increasing amount of Semantic information is required in terms of vocabulary control and ontologies. Finally, future research directions in biological data integration are discussed.
**INTRODUCTION**

Bioinformatics is the science of storing, extracting, organizing, analyzing, interpreting, and utilizing information from biological sequences and molecules. It has been fuelled mainly by advances in DNA sequencing and genome mapping techniques. Great opportunities arise for developing novel data analysis methods. Some of the great challenges in bioinformatics include protein structure prediction, homology search, multiple alignment and phylogeny construction, genomic sequence analysis and gene finding, as well as applications in gene expression data analysis, drug discovery in the pharmaceutical industry, etc. Nowadays, there is a proliferation of research institutions that produce sources of huge amounts of biological data derived from experimentation with biological systems. These data sources can be fully exploited only if a great effort is made to integrate disparate data formats, protocols and tools. Data integration and system interoperability are currently being undertaken in order to overcome the high level of heterogeneity currently present in the available resources.

One way to expand the utility and interpretability of the individual resources would be to create a standard unified model for the description of data and, consequently, a format for their exchange and representation that is machine readable. In the literature, we can find several data formats intended to represent biological entities and systems: non XML-based, XML-based and ontology-based files. FASTA (Pearson, 1994) is an example of a non XML-based data format for the representation of sequence data. The main problem with this type of format is the lack of structure consistency, thereby leading to a possibly different interpretation of a correct file. The second group tries to overcome the problem of a consistent structure definition by using XML as the data format. Within this group, two approaches are distinguished depending on how the structure is validated, that is, whether they are using XML document type definitions (DTDs) or XML schema definitions (XSDs). The use of XSDs can be richer than DTDs since users can specify not only the structure but also the Semantics of the XML tags by defining conditions and constraints. However, the potential of XSDs is addressed in only a few proposals. Finally, ontology-based formats have emerged as a solution to the lack of Semantics and will allow the formal representation of the knowledge to be exchanged. The ontology Web language (OWL) and the open biomedical ontology (OBO) are the main languages used to represent ontologies. Unlike XML Schemas, the use of well-defined ontologies will guarantee the correct representation of the content Semantics. It must be taken into account that in the ontology-based group, we also consider the XML formats that link the content with ontologies or controlled vocabularies.

It is worth mentioning that some of the previous efforts have also defined a specific XML-like language (i.e. SBML, BioPax, PSI-MI) for the representation of biological data. The aim of these efforts is the creation of public and well-known standards, so that data source providers are able to format and externalize their biological data according to the schemas and restrictions given by the standard. This is quite an interesting approach since data source providers know how to format the data to be externalized and shared, and this data can be easily used and integrated in applications such as Taverna or Pegasys to compose complex workflows. However, as will be discussed in this chapter, other efforts are still necessary in several directions.

Another important open issue will be the selection of relevant information from XML files. Since biological data will come from heterogeneous sources, it requires the identification of approximate retrieval systems, specifically tailored for Bioinformatics, in order to extract interesting portions from XML files. The pos-
Related Content

Using a Semiotic Framework to Evaluate UML for the Development of Models of High Quality
www.igi-global.com/chapter/using-semiotic-framework-evaluate-uml/30573?camid=4v1a

Labeling XML Documents
www.igi-global.com/chapter/labeling-xml-documents/41502?camid=4v1a

The Debugging of Rule Bases
www.igi-global.com/chapter/debugging-rule-bases/35864?camid=4v1a

Use of UML Stereotypes in Business Models
www.igi-global.com/chapter/use-uml-stereotypes-business-models/30546?camid=4v1a