XML in Motion from Genome to Drug

C. Gopi Mohan

Centre for Pharmacoinformatics

National Institute of Pharmaceutical Education and Research (NIPER)

Sector 67, S.A.S. Nagar- 160 062,

Punjab, INDIA.

*Address correspondence to this author at the

Centre for Pharmacoinformatics,

National Institute of Pharmaceutical Education and Research (NIPER),

Sector 67, S.A.S. Nagar, Punjab-160 062,

INDIA.

Phone: 0091-172-2214682-2019,

Fax: 0091-172-2214692

E-mail: cmohan@niper.ac.in and cgopimohan@yahoo.com

Abstract

Information technology (IT) has emerged as a central to the solution of contemporary genomics and drug discovery problems. Researchers involved in genomics, proteomics, transcriptional profiling, high throughput structure determination, and in other sub-disciplines of bioinformatics have direct impact on this IT revolution. As the full genome sequences of many species, data from structural genomics, micro-arrays, and proteomics became available, integration of these data to a common platform require sophisticated bioinformatics tools. Organizing these data into knowledgeable databases and developing appropriate software tools for analyzing the same are going to be major challenges. XML (eXtensible Markup Language) forms the backbone of biological data representation and exchange over the internet, enabling researchers to aggregate data from various heterogeneous data resources. The present article covers a comprehensive idea of the integration of XML on particular type of biological databases mainly dealing with sequence-structure-function relationship and its application towards drug discovery. This e-medical science approach should be applied to other scientific domains and the latest trend in semantic web applications is also highlighted.

Introduction

The term informatics is widely playing an important role in different disciplines of scientific arena. Some of the well known informatics area in biomedical science include bioinformatics, structural bioinformatics, chemoinformatics, pharmacoinformatics, medical informatics, immunoinformatics, genome informatics, microarray informatics, neuroinformatics, biodiversity

informatics, biomolecular informatics, clinical informatics, and pharmacy informatics. It is well known that the size of the human genome has an estimated 30,000 to 40,000 genes, and it would be interesting to consider just how many structural targets this tsunami of data represents. A complete assessment of the number of structural genomics targets for therapeutic intervention is important for the development of post-genomic research era within the pharmaceutical industry. It is well known that an understanding of the molecular structure leads to knowledge of molecular mechanism of action and its function. Molecular structural knowledge is therefore vital if we have to proceed to a complete understanding of life at the atomic level. Molecular information in biological systems are mainly restricted to four type of macromolecules i.e. proteins, nucleic acids, polysaccharides and lipids for interacting among themselves and with small molecules or drugs. The affinity, specificity, toxicity and inability to obtain lead molecules for nucleic acids, polysaccharides and lipids suggest that proteins are the main source for drug targets.

Genomic sequence to structure prediction is one of the most challenging tasks in computational structural biology program. Some well known protein structure prediction techniques are *de-novo* or *ab-initio* modeling and comparative protein modeling which again is divided into homology modeling and protein threading. On the basis of primary sequence information of a given protein different methods have been developed in predicting structure: Prediction of secondary structure (beta sheets, α-helix, coils) from primary sequence, *ab-initio* structure prediction, homology modeling – up to X-ray crystallography model accuracy i.e. 2-3 Å and protein threading (fold family recognition).

Homology modeling uses the fact that if sequence of the unknown protein with $\geq 30\%$ sequence similarity is found with the target protein whose three dimensional structure is known, it is appropriate to assume that the two proteins have similar tertiary folds. This method is useful to a) hypothetically predict functional regions of proteins, b) design or interpret site-directed mutagenesis experiments, c) spectral properties, d) protein structure-based drug design and d) docking analysis e) functional assignment for genomic data. Its limitations are a) structural differences occurring in surface loops of the protein and b) theoretically predicted low resolution structures. Protein threading i.e fold family recognition method classifies known structures into families with similar foldings. Given a sequence of amino acids, we can select the family to which the given sequence most likely belongs.

Discussion and Applications

Bio-Chemo-Pharmaco-informatics disciplines holds out strong expectations of reducing the time and cost of development in pharmaceutical industries from genome to drug by proper integration of *in silico* data with *in vivo* and *in vitro* data for design of new chemical entities, vaccines etc. Moreover, the design and implementation of IT in drug discovery must match well with the development of new scientific methods, algorithms for managing large amounts of sequence and structural data. This IT workflow should complement well with the different scientific programs involved in genome informatics, structural bioinformatics and chemoinformatics suites.

Following important points are presented in Fig. 1, which should be taken into consideration in the drug discovery program.

- Primary structure (sequence) of a medically important protein (target).

- Structural bioinformatics: Modeled structure of the target protein or Structural molecular biology of the target protein (protein crystallography).

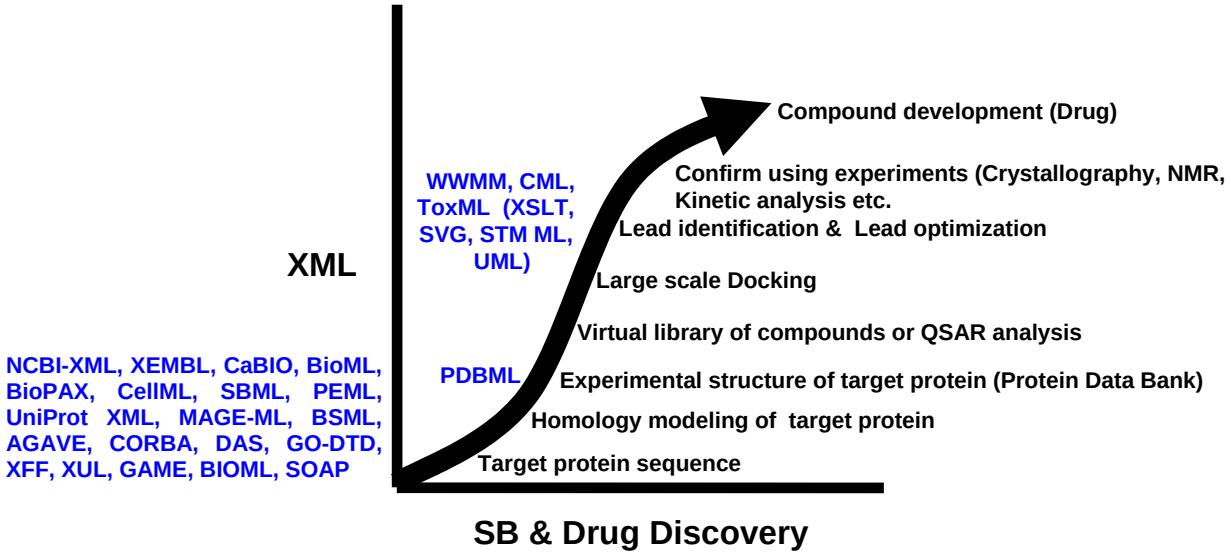

Fig. 1: Schematic outline of the application of SB and its integration with XML in drug discovery process.

- Virtual screening or quantitative structure activity relationship (QSAR) study of drug like compounds to probe the active sites of protein and its mechanism of action.

- High throughput docking analysis of most active (*in silico*) drug like compounds.

- Experimental confirmation of docking analysis (docked compounds) using X-ray crystallography. NMR, Kinetic experiments- binding affinity of docked compounds to target protein.

- Prediction of selective potent drug like compounds. ADMET predictions, drug-drug interactions, *in vitro* test assay of the designed compounds. Pharmacological testing in animal models.

- Clinical trials and US Food and Drug Administration (FDA) approval of novel compounds identified as drug.

Nowdays, in structural genomics, XML representation of sequence and structure is developing, to accelerate the data representation, storage, retrieval and exchange through internet, and is presented in Table 1. Main challenging and interesting task is to deal with this diverse set of data

generated from genome to drug. The key to bioinformatics, chemoinformatics and pharmacoinformatics field is to accelerate discovery and solve critical problems in drug

Table 1: XML formats, DTDs and Schemas prominently used in bioinformatics, chemoinformatics and pharmacoinformatics domain.

Name	URL
AGAVE: Architecture for Genomic Annotation Visualization and exchange	http://www.agavexml.org/
BioML: BIOpolymer Markup Language	http://proteometrics.com/BIOML/
BioPAX: Biological Pathways Exchange	http://www.biopax.org
BSML: Bioinformatic Sequence Markup Language	http://www.bsml.org
CellML	http://www.cellml.org
DAS: Distributed Annotation System	http://www.biodas.org
Gene Ontology (GO) DTD	http://www.geneontology.org/xml-dtd/go.dtd
MAGE-ML: MicroArray Gene Expression Markup Language	http://www.mged.org/mage
NCBI DTDs: Numerous DTDs, including GBSeq, TinySeq, and NCBI Blast	http://www.ncbi.nlm.nih.gov/dtd
CML: Chemical Markup Language	http://xml.coverpages.org/cml.html
PSI-MI: Proteomics Standards Initiative Molecular Interaction	http://www.psidev.info/
SBML: The System Biology Markup Language	http://www.sbw-sbml.org/
UniProt XML	http://www.pir.uniprot.org/

development by representing and integrating these tsunami data sets using semantic web approach based on XML. In the end, this world wide semantic web applications enables

individual researchers, software engineers, database managers, to easily exchange and share their relevant and aggregate data from multiple sources as well as mine this data for making important scientific clues [1-6].

National Center for Biotechnology Information (NCBI) and European Bioinformatics Institute (EBI) provide world-wide free public services by developing, organizing, networking and distributing different scientific databases. Main classification of these databases are, nucleotide databases, protein databases, structure databases, proteome databases, microarray databases, and literature databases etc. Various bioinformatics toolboxes for protein function analysis, proteomic analysis, sequence analysis, structure analysis, similarity and homology modeling etc. are incorporated by appropriate submission facilities for sequence, structure and other biological data. In short, NCBI and EBI mainly serve as an important platform in performing structural bioinformatics work.

XML for Structural Bioinformatics:

Structural bioinformatics (SB) devotes mainly to answer questions about molecular way of life by using computational resources. One of the main goals of SB community is to create intelligent database systems and associated softwares which is capable of storing, retrieving, analyzing and exchanging large sets of genomic and structural data. An exponential growth in biological databases is challenging to SB community for interrelating and sharing data for open scientific exchange. This bottleneck is overcome by the proper usage of XML for data representation and its exchange over the internet. XML is officially specified by the World Wide Web Consortium (W3C), started in 1996, and is open standard, independent of computer operating systems and

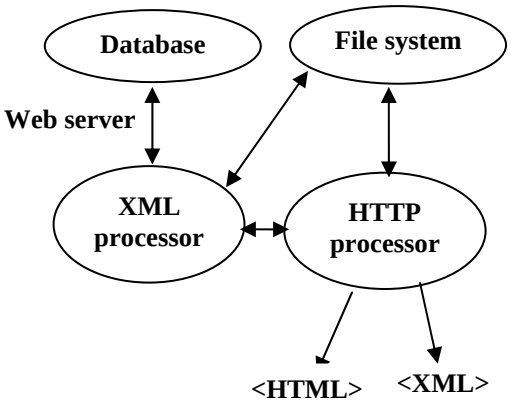

Fig. 2: Backbone representation of XML schema

programming languages. The application of XML in bioinformatics is essential to bridge the gap, such as, (i) No standard nomenclature exists for genomics, proteomics, chemoinformatics and other biological data, (ii) No standard data format exists to exchange biological data, (iii) No standard data model exists, and (iv) Difficulties in using and exchanging data.

XML is highly flexible, internet-oriented and has very rich capabilities for linking databases represented in simplest form in Fig. 2. XML and document type definition (DTD) files are human readable and can be easily edited by people with only few computer skills. XML provides an open framework for defining standard specifications, because SB clearly lacks standardization. This language is being extensively used nowadays to manage documents and informations such as in macromolecular sequence, macromolecular structure, spectra, organic molecules, crystallography, quantum chemistry, hypertext (HTML), databases, regulatory processes, molecular databases, publishing and many others. Development of protein data bank in XML format (PDBML) is good example in this direction [7]. All of the released PDB entries are now available in PDBML/XML format**.** Some other well known bioinformatics contents such as NCBI-XML, XEMBL contribute its application at genome level (Tables 1 and 2).

Table 2: Some of the well known Bioinformatics/ Chemoinformatics/ Toxicoinformatics databases/languages in XML

XML Databases/ Languages	Importance	URL
PDB/PDBML (Macromolecular Structure database in XML)	The Protein Data Bank Markup Language (PDBML) provides a representation of archival macromolecular structure data in XML format.	http://pdbml.rcsb.org/
EMBL/XEMBL (Bioinformatics database in BSML and AGAVE)	XEMBL has been a service to display data from the EMBL Nucleotide Sequence Database in XML formats, BSML (Bioinformatics Sequence Markup Language) and AGAVE (Architecture for Genomic Annotation, Visualization and Exchange) developed by Double Twist. for managing, visualizing and sharing annotations of genomic sequences.	http://www.ebi.ac.uk/xembl/index.html http://www.bsml.org http://www.agavexml.org/
caBIO (Bioinformatics cancer database)	The cancer Bioinformatics Infrastructure Objects (caBIO) model is an open source project and all caBIO objects can be transformed into XML, and XSL/XSLT is used to present data in documents, web pages or other interfaces.	http://ncicb.nci.nih.gov/NCICB/infrastructure/cacore_overview/caBIO
NCBI-XML (Bioinformatics	NCBI software tools can now automatically produce data as XML. This provides developers access to the	http://www.ncbi.nlm.nih.gov/IEB/ToolBox/XML/

database in XML)	full internal NCBI data set using a variety of Open Source tools.	
CML (Chemoinformatics database in XML)	Chemical Markup Language (CML) An approach for representing chemical structures in XML format.	http://xml.coverpages.org/cml.html
ToxML (Toxicoinformatics database in XML)	ToxML is an XML database system developed for toxicology information.	https://www.leadscope.com/toxml.php

XML for Chemoinformatics

Development of XML in chemoinformatics i.e. Chemical Markup Language (CML) is another approach to manage molecular data. CML files are extensible in that new concepts can be added through its unique dictionary-based system. Namespaces allow many different disciplines to be combined, such as graphics, mathematics, biology, text etc. CML supports and interfaces with developments such as Java, C++ and Corba. CML has support for atom- and bond-based stereochemistry. CML can hold 2-D molecular information in a variety of ways: a) Connection tables, b) SMILES and c) 2-D coordinates. CML can hold 3-D molecular information in a variety of ways: a) 3-D Cartesian coordinates, b) fractional coordinates, c) Z-matrix, d) crystallographic unit cell, e) molecular symmetry and f) space group symmetry. Jumbo- an XML browser is designed to work with CML and XML. Jumbo browser can display XML style sheets and can use CML to draw molecules [8,9].

Chemical coordinates of molecules in 2D and 3D represented in CML can be visualized graphically using Scalable Vector Graphics (SVG), a text-based graphics language describing

images with vector shapes, text, and embedded raster graphics. SVG files are compact and provide high-quality graphics on the web, in print, and on resource-limited handheld devices. In addition, SVG supports scripting and animation, so is ideal for interactive, data-driven, personalized graphics. The molecular data is originally stored in CML, then an EXtensible Stylesheet Language Transformations (XSLT) document automatically converts each CML file to an SVG file.

World Wide Molecular Matrix (WWMM) is a molecular repository that manages chemical information of molecules entirely in XML and CML/ or CCML (computational chemical markup language) [10]. WWMM Wiki - an online database for molecular informatics, can edit it at any time, from anywhere. Its main application is turning the web into a chemical knowledge base, getting robots to do routine tasks automating CCML, Chemistry software toolkit development and research, chemical databases (design and implementation), peer-to-peer ("Napster") Chemistry, GRID and eScience technologies.

WWMM enable scientists and chemists to publish at source, and provide them with a mechanism that ensures their molecular data is freely available for everyone around the globe. The open-source nature of the matrix technology guarantees accessibility of scientific data for the community. Currently there aren't any mechanisms that would allow to globally store molecular data along with measured, calculated or computed properties. The WWMM only holds XML data, but converters to and from many common formats are provided. The matrix toolkit is able to convert the most common molecular file formats such as MOL/SMILES and PDB for submission to the WWMM. This group also generated IUPAC International Chemical Identifier

(InChIs) for 250,000 molecules from the NCI database and around 10,000 molecules from the KEGG database respectively. These solutions reduce the high costs of research and accelerate revenue by using intelligent documents to transform manual, paper-based processes into automated digital workflows. Further, it will support *in silico* prediction of molecular and reaction properties for use in pharmaceutical design, enzymatic, clinical and toxicological processes.

XML for Toxicoinformatics

The ability to accurately predict toxicity using quantitative structure-activity relationships (QSAR) is becoming more important as toxicoinformatician worldwide incorporate QSAR-based evaluations for hazard identification and risk assessment in safety purposes. Several practical and fundamental challenges need to be tackled in building reliable QSAR models, one being access to high-quality data from which accurate predictive models can be derived. In this regard, FDA has been actively collaborating with the public toxicity database standardization effort, ToxML, to create a set of controlled vocabulary to represent toxicity data (Table 2). Through a cooperative research and development agreement involving the Center for Food Safety and Applied Nutrition and the Center for Drug Evaluation and Research, FDA scientists now have access to data in consolidated databases which can be more easily incorporated into the review process. These databases will greatly expand safety analysis by utilizing analog information and improve efficiency by electronic retrieval. Currently, studies are added to these databases in a separate process following the safety evaluation of the submission. This method is problematic in that it duplicates effort and creates a lag-time between completion of the review and addition of

the data to the QSAR training set. In an effort to resolve this issue, FDA scientists are exploring reviewing submissions with real-time data entry in an effort to expand these databases. So *in silico* toxicology databases for various endpoints need to be properly integrated and XML (i.e. ToxML) should play an important role in organizing and managing this data.

Recent trend is an integrated *in silico* with *in vitro* and *in vivo* approaches to early ADME/T screening, for making effective decisions on lead molecule selection, which will help support and accelerate drug discovery projects. So an integrated *in silico, in vitro* and *in vivo* toxicology data need to be properly integrated and XML should play an important role in organizing and managing this drug discovery process. In addition to improving safety evaluations at the FDA, these databases will increase the global knowledge base and expand the scope of predictive toxicology.

CONCLUSIONS AND FUTURE PROSPECTIVES

The inspiration to write this article came by teaching M.S. (Pharm.) Pharmacoinformatics students at NIPER. This article serves as a comprehensive collection of biomedical informatics, where XML and semantic web is used with scientific intuition. The usage of XML for structural biologists, bioinformatician, chemoinformatician, toxicoinformatician and other allied fields is touched upon with clarity and its importance towards drug discovery program is highlighted. The challenging path of "bench to bedside" in the drug discovery program was considered, and the flow of XML along with the semantic web, for the proper integration of this tsunami data set of *in silico* to *in vivo* and *in vitro* is appreciated and directed towards its proper usage.

References

1) XML, bioinformatics and data integration. Bioinformatics. 2001, 17(2):115-25.

2) From XML to RDF: how semantic web technologies will change the design of 'omic' standards. Nat. Biotechnol. 2005, Sep;23(9):1099-103

3) Creating a bioinformatics nation. Stein, L. 2002, Nature **417**, 119-120.

4) Bioinformatics: bringing it all together. Chicurel, M. 2002, Nature **419**, 751, 753, 755.

5) Scientific Publications in XML - towards a global knowledge base. *Data Science* 2002, 1, 84-98

6) Virtual screening using grid computing: the screensaver project. Richards WG (2002), Nature Rev Drug Discovery, 1, 551-555.

7) PDBML: the representation of archival macromolecular structure data in XML. John Wesbrook, Nobutoshi Ito, Haruki Nakamura, Kim Henrick and Helen M. Berman, Bioinformatics, 21(7), 988-992, 2005.

8) Chemical Markup, XML, and the World Wide Web. 6. CMLReact, an XML Vocabulary for Chemical Reactions**,** Gemma L. Holliday, Peter Murray-Rust, and Henry S. Rzepa, J. Chem. Inf. Model.; 2006; *46*(1) pp 145 - 157

9) Chemical Markup, XML and the Worldwide Web. Part 4. CML Schema, *J. Chem. Inf. Comp. Sci*., 2003, 43, issue 4

10) The World Wide Molecular Matrix - a peer-to-peer XML repository for molecules and properties. *EuroWeb2002*, 2002, The British Computer Society, 163-164. http://wwmm-svc.ch.cam.ac.uk/wwmm/html/