Bioinformation by Biomedical Informatics Publishing Group *open access*

www.bioinformation.net **Views & Challenges**

Techniques for integrating -omics data

Siva Prasad Akula[1, 2], Raghava Naidu Miriyala[2,4], Hanuman Thota[2, 4], Allam Appa Rao[3] and Srinubabu Gedela[1, 3, *]

[1]Institute of Glycoproteomics and Systems Biology, Tarnaka, Hyderbad-500017, India; [2]Department of Computer Sciences and Engineering, Acharya Nagarjuna University, Guntur-522510, India; [3] International Centre for Bioinformatics, Andhra University, Visakhapatnam-530003, India; [4]D.M.S S.V.H. College of Engineering, Department of Computer Science, Machilipatnam - 521002, India;
Srinubabu Gedela* - Email: srinubabuau6@gmail.com**;** Phone: 91 40 27006539; Fax: 91 40 40131662; * Corresponding author

received December 23, 2008; accepted December 29, 2008; published January 12, 2009

Abstract:
The challenge for -omics research is to tackle the problem of fragmentation of knowledge by integrating several sources of heterogeneous information into a coherent entity. It is widely recognized that successful data integration is one of the keys to improve productivity for stored data. Through proper data integration tools and algorithms, researchers may correlate relationships that enable them to make better and faster decisions. The need for data integration is essential for present -omics community, because -omics data is currently spread world wide in wide variety of formats. These formats can be integrated and migrated across platforms through different techniques and one of the important techniques often used is XML. XML is used to provide a document markup language that is easier to learn, retrieve, store and transmit. It is semantically richer than HTML. Here, we describe bio warehousing, database federation, controlled vocabularies and highlighting the XML application to store, migrate and validate -omics data.

Keywords: data warehousing; bio warehouse; XML; genomics unified schema; database federation; controlled vocabularies

Background:
The biological science has gone to great heights in the form of genome, proteome, glycome and metablome providing vast quantities of information. Such information resources are of censorious importance to modern biology, medical science and in drug discovery **[1]**. Data obtained from completed and ongoing projects of 'omes' is now widely available over the internet. The World Wide Web (WWW) provides a globally distributed communication framework that is essential for all –omics disciplines **[2]**. It is widely recognized that successful data integration is one of the keys to improve productivity for stored data. There are different data integration tools and algorithms in discovering relationships that enable them to make better and faster decisions about therapeutic targets and thus data integration is a perennial issue in –omics **[3]**. Another challenge is due to high level of biological complexity and the fragmented nature of biological research.

Possible solutions:
Data integration is a need in the area of -omics community. Omics data is currently spread across the internet and throughout organizations in a wide variety of formats. This needs to be concentrated to bridge the gaps between formats which require an integrated view of all relevant data requiring research and development in field of data integration.

Bio warehousing:
Researchers in different life sciences extensively collaborate for sharing genomic, proteomics and glycome data. This requires the integration of different databases in a dynamic manner into a single repository or a warehouse, extracted from various heterogeneous sources, having different degrees of quality and trust. This may not always carefully control data quality **[3]**. In this process, data obtained from various sources is often structured differently at the source. A complete understanding of the database and its format to translate among the schemas is required to exchange information. Advanced computational techniques are needed for integration, mining, and functional interpretation of source data **[7]**. Omics databases are heterogeneous and different in their representation as well as in their query capabilities across diverse sources **[8]**. Even though they cover similar semantics such as genes, glycans, proteins, sequence annotations, data integration is required in order to achieve better results. Different tools were available and one among them is Bio Warehouse, which integrates its component databases into a common representational framework within a single database management system **[9]**. This enables multi-database queries using the Structured Query Language (SQL), through which data mining can be done for data analysis.

In human health and life sciences, researchers extensively collaborate with each other, sharing genomic, biomedical and experimental results. The extensively available knowledge that require dynamically integrating different databases into a single repository or a warehouse. The data integrated in these warehouses are extracted from various heterogeneous sources, having different degrees of quality and trust **[10]**. Bio Warehousing helps in supporting the following -omics data types: chemical compounds, biochemical reactions, metabolic pathways, proteins, genes, nucleic acid sequences, and features on protein and nucleic-acid sequences, organism taxonomies.

ISSN 0973-2063 (online) 0973-8894 (print)
Bioinformation 3(6): 284-286 (2009)

Bioinformation by Biomedical Informatics Publishing Group open access
www.bioinformation.net

Views & Challenges

Extensible Markup Language

The Extensible Markup Language (XML) is a general-purpose markup language that helps in sharing data across heterogeneous systems. XML is a format of choice for storing information with an inherent hierarchical structure and has been widely accepted in the -omics industry as a means of data exchange. Today, there is a growing need in -omics to combine available software tools into chains, thus building complex applications from existing single-task tools. To create such workflows, the tools involved have to be able to work with each other's data - therefore, a common set of well-defined data formats is needed **[4]**. Information resources on the World Wide Web play a central role in modern biology. But integrating data from multiple sources is often confine by the need to reformat data files, convert between naming systems, or perform ongoing maintenance of local copies of public databases. Opportunities for new ways of combining and re-using data are arising as a result of the increasing use of web protocols to transmit structured data **[5, 6]** and where XML is coming into the picture, this facilitates data integration and application interoperability through the adoption of standards for representing certain types of data, e.g., genome annotations or microarray experiments. It provides a common format for expressing both data structures and contents making it a standard for data representation and migration. Hence, it can help in integrating structured, semistructured, and unstructured data over the web repositories **[11]**.

Recent prospectus using XML

Recently Gupta and colleagues **[12]** developed Neuroscience Information Framework (NIF) a one-stop-shop for neuroscience. This system is targeted to find information from web pages, publications, databases, web sites built upon databases, XML documents and any other modality in which such information may be published. This system is currently in beta test and may be accessed through http://nif.nih.gov. In the same way, Sho and colleagues **[13]** developed XML-based genomic information and provided scalable vector graphics output named Blujay for genome annotation. They incorporated the gene expression information, comparative analysis with an unlimited number of other genomes in the same view, numerous display customization features, GPS-style waypoints for marking multiple points of interest on a genome and user relevance feedback of annotated genes or textual items. **[13]** They have also developed a user friendly system based on XML that enables end-users who use a quality labeled software package, to automatically produce all the legally necessary documents concerning the registry of physiotherapy electronic healthcare records. Gil-Redondo R **[14]** developed the user-configurable XML virtual screening data management on an integrated platform that stores ligand information, modular and pluggable architecture. Hunter and colleagues **[15]** developed the InterPro database (http://www.ebi.ac.uk/interpro/) that integrated predictive models or 'signatures' representing protein domains, families and functional sites from multiple, diverse source databases: Gene3D, PANTHER, Pfam, PIRSF, PRINTS, ProDom, PROSITE, SMART, SUPERFAMILY and TIGRFAMs. They have incorporated other developments which include the provision of non-signature data, such as structural data, in new XML files on our FTP site, as well as the inclusion of matchless UniProtKB proteins in the existing match XML files. Recently Schaefer and coworkers developed a XML based pathway interaction database, a freely available collection of curated and peer-reviewed pathways composed of human molecular signaling and regulatory events and key cellular processes. It offers a range of search features to facilitate pathway exploration.

Database federation

A federated database is a logical association of independent databases that provides a single, integrated, coherent view of all resources in the federation. There are many proven technologies in the scientific usage, such as CORBA (Common Object Request Broker Architecture), J2EE (Java 2 Platform, Enterprise Edition), that access sources of interest directly and combine the data retrieved from those sources with the application itself. This approach always works, but it is expensive. Database federation is cost effective for private and government organizations and mainly it does not require modification of the primary data stores as most are large number of heterogeneous databases, especially databases like PDB, PROMALS, PSIPRID, NCBI, Swissprot, etc are in the public domain and thus not directly modifiable by researchers. This technology has have been into for further development by introducing join scheduling algorithms by Xiaodan and coworkers **[16]** which balances network utilization metric to optimize the use of all network paths in a global-scale database federation. This metric allows algorithms to exploit excess capacity in the network, while avoiding narrow and long-haul paths.

Controlled vocabularies

In -omics context, controlled vocabularies offer a form of data integration by enforcing naming conventions for data elements that ultimately appear in -omics databases. In this type of techniques integrating heterogeneous -omics data sources are based on one of a common field, ontology or cross-reference like, Plant Ontology Consortium (POC, http://www.plantontology.org) The POC builds upon groundbreaking work by the Gene Ontology Consortium (GOC) by adopting and extending the GOC's principals. POC is a collaborative effort among model plant genome databases and plant researchers that aim to create maintain and facilitate the use of a controlled vocabulary (ontology) for plants. The ontology allows users to describe attributes of plant structure (anatomy and morphology) and developmental stages to data types, such as genes and phenotypes, to provide a semantic framework to make meaningful cross-species and database comparisons **[17]**. In the same way protein ontology is designed to provide a structured protein data specification for protein data representation. Protein ontology is a standard for representing protein data in a way that helps in defining data integration and data mining models for protein structural and functional properties **[18]** whose instantiations, which are

ISSN 0973-2063 (online) 0973-8894 (print)
Bioinformation 3(6): 284-286 (2009)

undergoing evolution, need a good management and maintenance system. Protein ontology instantiations signify relevant information about proteins that is shared and has evolved to reflect development in protein ontology project and proteomics domain itself. Protein ontology needs to be populated with data and information from data and information sources proteomics domain **[18]**.

Vocabulary in –omics databases which deals with genetic related information also requires controlled vocabulary by following the rules of Gene Ontology Consortium. Ontologies is the specifications of the entities, their attributes and relationships among the entities in a domain of discourse - are increasingly enabling biomedical researchers to accomplish these tasks. In fact, bio-ontologies are beginning to proliferate in step with accruing biological data. The myriad of ontologies being created enables researchers not only to solve some of the problems in handling the data explosion but also introduces new challenges **[19]**. In controlled vocabulary, Gene Ontology (GO) supports data integration for biomedical researchers by enabling them to store results and generate reports using a common terminology in annotating genes and gene products. Functional analysis using the GO is importent for array analysis, but it is often difficult for researchers to assess the amount and quality of GO annotations associated with different sets of gene products **[20]**. Here vocabularies are used in subject indexing schemes, subject headings, thesaurus and taxonomies. This schemes mandate the uses of predefined, authorized terms that have been preselected by the designer. This is used to improve the effectiveness of information storage and retrieval systems via some sort of description using language. Many bioinformatics applications also rely on controlled vocabularies or ontologies to consistently interpret and seamlessly integrate information scattered across public resources. Experimental data sets from metabolomics studies need to be integrated with one another, but also with data produced by other types of -omics studies in the spirit of systems biology, hence the pressing need for vocabularies and ontologies in metabolomics has high impact for better understanding of integrated data **[21]**. The primary purpose of vocabulary control is to achieve consistency in the description of content objects that helps in accurate and fast retrieval.

Conclusion:

Data refers to a collection of results, including the results of experience, observation or experiment, or a set of premises. Data can be utilized at the maximum when made available to all in a common format. Different organizations around the world have started storing the data in their own formats; this diversity of data sources is caused by many factors including lack of coordination among the organizations and research laboratories. These intellectual gaps can be bridged by adopting new technology, mergers, acquisitions, and geographic coordination of collaborating groups. In conclusion, we showed an over view of existing solutions for overcoming problems with heterogeneous and distributed data. We have identified four different styles of integrations solutions that may be implemented at different contexts and requirements, in order to reuse the data across applications for research collaboration.

References:

[01] K. McGarry *et al.*, *Expert Systems*, 23 : 330 (2006)
[02] T. Clark *et al.*, *Brief. Bioinformatics*, 5: 59 (2004) [PMID: 15153306]
[03] C. Goble *et al.*, *J Biomed Inform*, 41: 687 (2008) [PMID: 18358788]
[04] P. N. Seibel *et al.*, *BMC Bioinformatics*, 7: 490 (2006) [PMID: 17087823]
[05] J. C. Bare *et al.*, *BMC Bioinformatics*, 8: 456 (2007) [PMID: 18021453]
[06] F. Moussouni *et al.*, *Lecture Notes in Computer Science*, 4832: 5 (2007)
[07] H. Huang *et al.*, *Frontiers in bioscience* 12: 5071 (2007) [PMID: 17569631]
[08] B. Al-Daihani *et al.*, *Lecture Notes in Computer Science*, 4316: 78 (2006)
[09] T. J. Lee, *BMC Bioinformatics*, (2006) 7: 170 [PMID: 16556315]
[10] F. Moussouni, *Lecture Notes in Computer Science*, 4832: 5 (2007)
[11] E. Bertino *et al.*, *IEEE Internet Computing*, 5: 75 (2001)
[12] A. Gupta *et al.*, *Neuroinformatics*, 6: 205 (2008) [PMID: 18958629]
[13] J. Soh *et al.*, *BMC Bioinformatics*, (2008) 9: 450 [PMID: 18940007]
[14] R. Gil-Redondo *et al.*, *J Comput Aided Mol Des.*, (2008) [PMID: 18941902]
[15] S. Hunter *et al.*, *Nucleic Acids Res.*, 37: D211 (2008) [PMID: 18940856]
[16] W. Xiaodan, *International Conference on Data Engineering*, 4497467, 586 (2008)
[17] S. Avraham *et al.*, *Nucl. Acids Res.*, 36: D449 (2008) [PMID: 18194960]
[18] A. S. Sidhu *et al.*, *IEEE Symposium on Computer-Based Medical Systems*, 4262715, 607 (2007)
[19] D. L. Rubin *et al.*, *Brief Bioinformatics*, 9: 75 (2008) [PMID: 18077472]
[20] T. J. Buza *et al.*, *Nucl. Acids Res.*, 36: e12 (2008) [PMID: 18187504]
[21] I. Spasic *et al.*, *BMC Bioinformatics*, 9: S5 (2008) [PMID: 18460187]

Edited by P. Kangueane

Citation: Akula *et al.*, Bioinformation 3(6): 284-286 (2009)

ISSN 0973-2063 (online) 0973-8894 (print)
Bioinformation 3(6): 284-286 (2009)