

Translational Informatics Applied To Systems Biology: From Data Integration and Process Management to Diagnostics Models

Erich A. Gombocz, Robert A. Stanley, Michael Travers
IO Informatics, Inc., 2000 Powell Street, Suite 520, Emeryville, CA 94608, USA

QB3 Biomedical Engineering Symposium, Genentech Hall, UCSF Mission Bay Campus, April 26, 2006, San Francisco, CA

Correspondence: Dr. Erich A. Gombocz

(+1) 510.420.8400 Fax: (+1) 510.420.8440

egombocz@io-informatics.com

Short title: Translational Informatics: Data Integration, Process Management, Diagnostics Models

Keywords: Systems biology, data integration, multi-methods, semantic models, disease profiling, RDF, XML.

Summary

This study highlights how distributed experimental data is integrated, managed and transformed into semantic models that describe diseases in a systems biology environment.

Integration, management, query and representation of data and knowledge from internal and external sources, will be presented across levels of granularity - from molecular level data (e.g. genomics, proteomics, metabolomics); through organelles; (e.g. quantitative tissue analysis, fMRI); and organisms (e.g. case, clinical data); to systems models suitable for multi-method screening. Biotechnology topics include creating, analyzing and integrating proprietary gene functions and metabolite-based data with related internal, public and subscription data, to develop and validate a biomarker-based disease profiling database and related diagnostics. Informatics topics include practical data integration methods, user-driven (non-programming) data structuring and pathway definition, and traversing the use-case from database and file silos to a coherent, semantic web database using a service and pathway-oriented RDF XML data model.

The project is taking place under a NIST Advanced Technology Program (ATP) grant involving a partnership between IO Informatics and Icoria (division of Clinical Data, Inc.). The project involves a multi-party collaboration including federal, academic, corporate biotechnology and informatics institutions, and applies biotechnology and information technology to data integration, process management and disease modeling for diagnostics.

Introduction

Despite significant advances in science and the ability to measure and assess the impact of different inputs and outputs of biological processes, systems biology approaches still are challenging. This is the case as most software commonly used to represent and store information on gene function, protein expression, metabolic reactions, biomarkers, and therapeutic pathways fails to support the relationship-centered, integrated approach to data which is the foundation of systems biology. Data integration across diverse sources and types, process management, data coherence and querying and representation of data need to address the inherent complexity in a way which provides for management of distributed experimental data and which allows to use semantic multi-dimensional models to describe disease profiles as system. In many cases data critical to systems biology are stored in disconnected databases requiring data extraction and processing prior to their association with other related data. Federated systems, which abstract data into a "meta-layer" - although query-able from a top-level interface - still limit data access, usage and sharing. While most of the work until now has been focused only on certain levels of granularity, the presented approach intends to integrate data from molecular level, cell and organism level into a comprehensive semantic knowledge base. This requires an approach led by the question or relationship scientists need to explore rather than the data format. An effective software infrastructure will enable such associations across applications and allows sharing them throughout an organization while preserving the integrity and security of underlying data. Developing and validating biomarker-based disease profiling and applying semantic models requires such approaches.

Materials and Methods

Disease profiling presents a significant challenge in any -omics category (such as functional genomics or proteomics). Due to multiple simultaneous and complex biomolecular activities, such diseases cannot be adequately characterized by changes in single components nor can pathological changes be understood by observing solely gene expression profiles. Instead, a pattern of subtle changes across multiple tissues and organs with complex associations between corresponding gene, protein and metabolite levels needs to be analyzed in context. In this case study, metabolite data sets generated by LC-MS and tissue analysis data were represented as objects using the Sentient Suite (IO Informatics, Emeryville, CA) as data integration, process management and collaboration platform. In addition to experimental data, public pathway and molecular interaction data sources such as KEGG and BIND were used in queries and to identify, validate and annotate the data set.

Results

The Sentient software represents molecular data (genomics, proteomics, metabolomics), cell data (quantitative tissue analysis, fMRI) and organisms level data (case, clinical data) via a common object, the Intelligent Multidimensional Object (IMO). IMOs enable users to build and represent data relationships, or associations, at multiple levels of detail. Associations allow biologists to link data at project and document levels, drill down to explore the fine-grained details and relationships within and across data, and define and model interactions and pathway functions indicated by specific data. Associations are built in the following ways:

- **Attributes**, which associate data according to their position within a predefined ontology
- **Attachments**, which associate data directly to other IMO level data and queries
- **Annotations**, which associate data via links to analytical content subsets created by users
- **Subset links**, which associate content subsets within IMOs to queries and to other content subsets within objects
- **Queries**, which identify and retrieve similar, covariant, manually, or otherwise semantically associated objects

Using this technique, internal experimental data can be meaningful associated with public-domain or subscription-based data sources, allowing reviewing, representing and mining of all data for their relationships within a common shareable and audited framework. Cross-institutional and interdisciplinary projects are managed throughout the process and can be viewed, analyzed and queried for their relationship. The presented information technology allows scientists to structure and describe complex data focusing on their needs without requiring programming - thus, providing the basis for multi-parametric disease modeling using semantics to describe data relationships in a true systems biology approach.

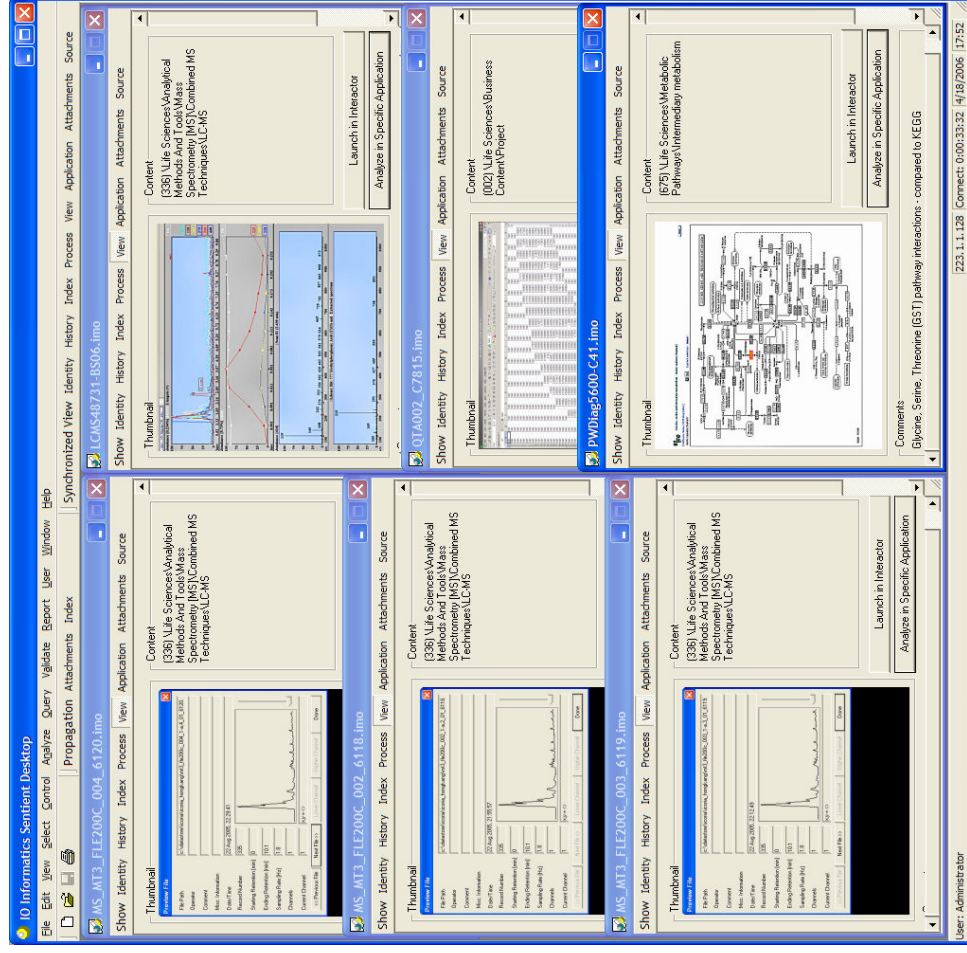


Fig. 1: Sentient Intelligent Multidimensional Object (IMO) Views
LC-MS metabolites, tissue matrices and lookup of data-relevant Glycine-Serine-Threonine pathway in a public reference database (KEGG). Any data type and format can be represented and interacted with in a common collaboration environment without requiring the original application.

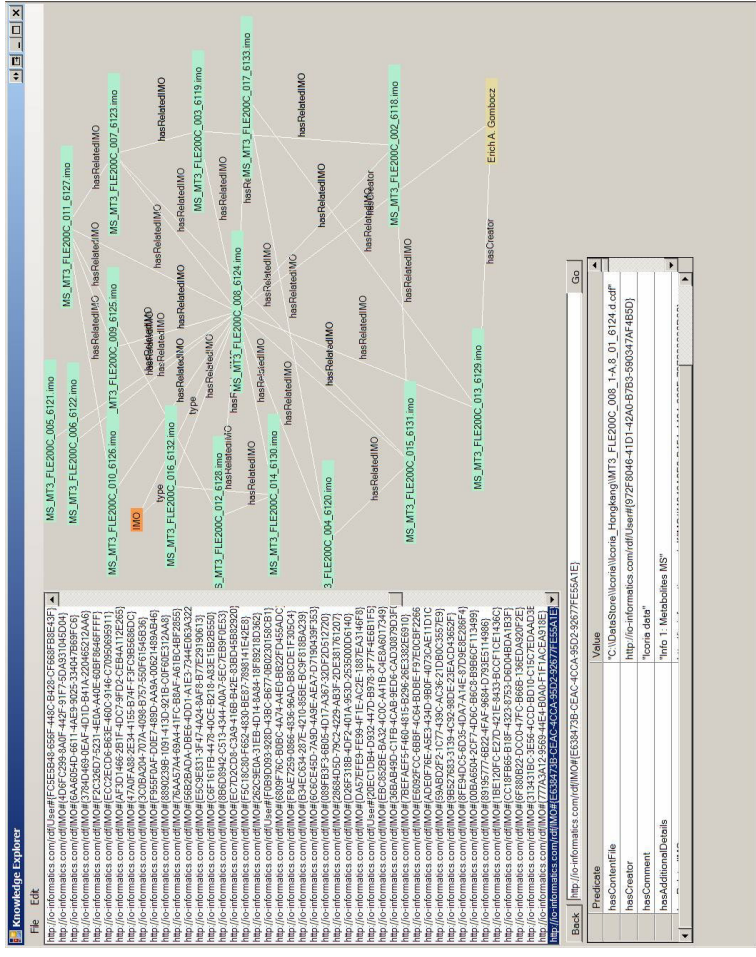


Fig. 2: Sentient Knowledge Explorer
Example of an interactive semantic view of metabolite relationships based on annotations and attributes.

Discussion

The opportunity to engineer new therapeutic pathways for predicting and controlling disease requires new informatics environments capable of creating coherence and preserving connections among complex, disparate, and increasingly large collections of biologically relevant data. The introduced tools, based on multidimensional objects, will allow scientists to build, share, and search relationships among data within a semantic RDF XML data model which supports several aspects of translational informatics:

- Permitting high-level integration of critical data across challenging boundaries (such as the intersection of clinical and laboratory research)
- Maintaining application-independent, open data formats that can be accessed by anyone, anywhere, in an organization
- Creating dynamic, searchable profiles for validating diagnostics, predicting or detecting adverse events, or targeting appropriate patients or markets for a therapy
- Creating an auditable, consistent data trail tracking creation and usage of data within an organization.

Future work will support the ability to import, merge, and create ontologies consisting of formal data definitions, interaction pathways, and other semantically important associations, which can be applied towards a better understanding and representation of disease models, compound activities, and adverse events. Further project goals include continued efforts to improve system usability, particularly in the area of clinical data capture.

References

- 1) Stanley, R.; Hancock, W. Bioinformatics in the clinic: challenges and opportunities for improved trials and clinical care. *Genom. Proteom. Techn.* **2003**, *3*(3), 29-36.
- 2) Glassbrook, N.; Ryzals, J. A systematic approach to biochemical profiling. *Curr. Opin. Plant Biol.* **2001**, *4*(3), 186-90.
- 3) Hancock, W.; Wu, S.; Stanley, R.; Gombocz, E. Publishing large proteome datasets: the meeting of scientific policies and emerging technologies. *Trends in Biotechnology (Suppl.)* **2002**, *20*(12), 39-44.
- 4) Neumann, E. A life science semantic web: are we there yet? *Science Signal Transduction Knowledge Environment (STKE)* **2005**, *283*, pe22.
- 5) Berners-Lee, T. What do HTTP URIs identify? www.w3.org/DesignIssues/HTTP-URI, July 27, 2002.
- 6) Bouquet, P.; Giunchiglia, F.; van Harmelen, F.; Serafini, L.; and Stuckenschmidt, H. C-OWL: Contextualizing *Web Semantics: Science, Services, and Agents on the World Wide Web 2004*, *1*, 325-43.
- 7) Wang X.; Gorlitsky R.; Almeida, J.S. From XML to RDF: how semantic web technologies will change the design of "omic" standards. *Nat. Biotech.* **2005**, *23*(9), 1099-1103.