

"Intelligent data": Towards systems biology for proteomics in a single system

Erich A. Gombocz, Robert A. Stanley

IO Informatics, Inc., 2000 Powell Street, Suite 520, Emeryville, CA 94608, USA

Correspondence: Dr. Erich A. Gombocz
(+1) 510.420.8400 Fax: (+1) 510.420.8440 egombocz@io-informatics.com

Keywords: Systems biology, proteomics, interoperability, ontologies, multi-methods, multidimensional analysis, biomarkers, intelligent data, agent-based software

Summary

Conventional informatics methods fail to meet requirements for functional integration of data and effective analysis tools for functional relationships between heterogeneous, but related data, in a normalized environment do not exist. Basic systems biology applications, such as detection and reproducible analysis of multiple-methods biomarkers are currently not supported. To enable interoperability while maintaining propriety and flexibility of data ontologies and annotations beyond conflicting "open" and "federated" web and XML-based standards, is needed. Innovative data structures are required which remedy the common limitations. This poster describes a new approach using multi-dimensional data objects to provide a single system capable of tracking, linking and relating data of any type in a true systems biology manner. This involves a metabolic view of proteomics and an interdisciplinary, secure and regulatory-compliant data sharing and query environment across distributed global resources. The sensitive and proprietary nature of systems biology research demands more finely-grained security than offered by the most advanced databases. Current methods for data integration based on software "wrappers" (translation layers as interfaces between different applications and resources) require programming and expensive hardware for global analysis. Applying intelligence at the data rather than the application level is a significant step which involves workflow, methodological ranking of experimental result relevancy and interdependency between different seemingly independent data in the knowledge building process. "Sentient" object architecture provides a user-centric approach to define and analyze complex, multi-dimensional relationships. This allows differentiating subsets within large datasets based on function.

Using a single system in a systems biology scenario on queries across normalized data subsets from 2DE gels, MS, bioassays, gene arrays and cell imaging, allows the building of knowledge which can provide answers to complex problems in the future.

Introduction

Systems biology requires interaction between disparate research environments. Resources, collaboration, reporting and publication involve multiple methods, areas of expertise, proprietary instrumentation, different computer systems and networks. A systems biology-oriented informatics must therefore address disparate data, applications, and selective database access and query methods across networks. Currently, many resources are excluded from efficient use, leading to lost data, lack of reproducibility and research redundancy. In addition to the difficulties in getting basic access to data from all relevant methodologies, incommensurable definitions from each accessible data source - even though related to a common case - result in missed dependencies, and, ultimately in an incomplete picture. To meet basic requirements for effective systems biology, emerging informatics methods must provide programming-free user-centric methods for data definition and analysis within an easy to deploy, reproducible and scalable framework.. Selective integration of access, normalization, evaluation and ranking of disparate data resources are mandatory. Most importantly, such systems must provide verifiable, validated and reproducible results, and they must be available for audits and signoff in a security environment with state and processing history.

Methods

In contrast to traditional data warehousing requiring exhaustive database schema definition and resulting in static data organization at the time of implementation, emerging "active data" methods which use data objects as agents have been developed. Such approaches as the applied "sentient" technology from IO Informatics are designed to selectively use existing resources while taking advantage of the benefits from standards, middleware and data warehousing without their inherent constraints. The introduced 'Intelligent Multidimensional Object' (IMO) is designed to access, relate and normalize analyzed and un-analyzed data from multiple ontologies and methods, including standards-annotated data of any conceivable type.

Results

Using a traditional proteomics approach (2DE separation, MS identification and bioassays), metabolic expression of protein sets was related to genomic profiles, cell imaging and toxicology providing functional clues on molecular interactions otherwise undetected.

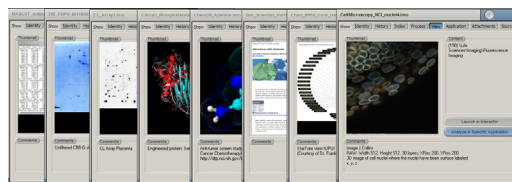


Fig.1: "Intelligent objects" represent any data type with all their attributes in a common view: text, instrument output, charts, graphs, spreadsheets, images, proprietary applications content, and database records.

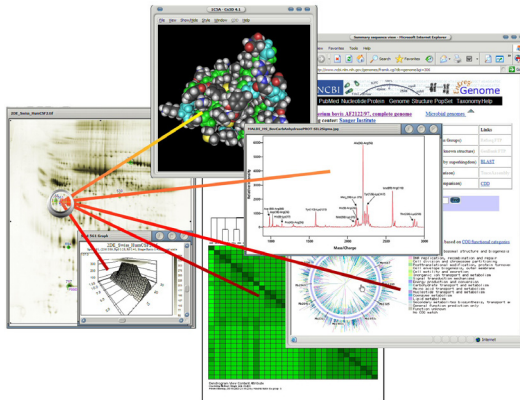


Fig.2: Example of complex data relationships across methods: a protein spot, its mass spectrum, genetic map, structure and biological activity.

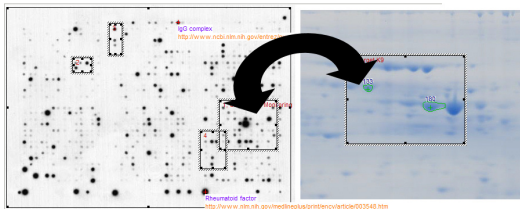


Fig.3: User-defined data subsets provide query-able normalized workspaces within large data sets: dynamically create "fields" as needed.

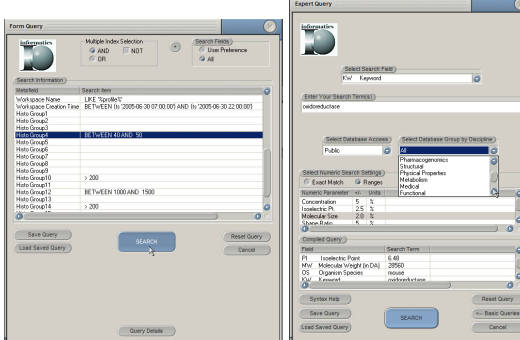


Fig.4: Knowledge-building through query and linking: relate internal results to public domain and propagate similarities to attachments, annotations and indexed metadata. Audited, regulatory-compliant collaboration across disciplines is provided to assist comprehensive discovery based on all relevant information in a secure e-notebook style framework.

Discussion

The presented user-centric approach using "intelligent data" towards a unified proteomics systems biology provides secure interoperability required for knowledge building and is fully based on existing IT infrastructure, data sources and applications. "Point and Click" definition of workspaces can be understood as user-defined creation of multiple database fields in real-time. This function allows users to define multiple fields from multiple sources and to put them to a joined query - crucial to commercializing systems biology applications such as biomarker detection and definition, biomarker screening, and multiple method compound activity screening to automate a systems-oriented drug discovery. In its common view, all data are considered together within a secure, audited and controlled access environment - thus, stimulating research innovation through more effective collaborative data sharing across disciplines - all requirements for a true systems biology approach which is particularly needed in the arena of a metabolic proteomics.

References

- 1) W.S. Hancock, S.L. Wu, R.A. Stanley, E.A. Gombocz: "Publishing Large Proteome Datasets: Scientific Policy Meets Emerging Technologies", Trends in Biotechnol. 20 (12): 539-44 (Review, 2002).
- 2) E.A. Gombocz, R.A. Stanley: "Achieving interoperability in Systems Biology: New informatics methods for user-centric, lightweight integration of heterogeneous data " 4th Int'l Symposium on Challenges in Systems Biology, ISB, Seattle, WA, 4/24-25, 2005 (2005).
- 3) E. Gombocz, R. Stanley: "Bringing Life Sciences together: Lightweight, regulatory-compliant knowledge management across heterogeneous data sources" 229th ACS, San Diego, CA, 3/13-17, 2005 (2005).
- 4) E.A. Gombocz, R.A. Stanley, R.L. Stevenson: "From Acquisition to Identification Using Global Data Resources: Fast, Comprehensive Knowledge Building in 2-D Electrophoresis" LabAutomation 2004, San Jose, CA, 1/31-2/5, 2004 (2004).