

Achieving interoperability in Systems Biology: New informatics methods for user-centric, lightweight integration of heterogeneous data

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

4th Annual International Symposium on Computational Challenges in Systems Biology, April 24 – 25, 2005, Seattle, WA

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

Short title: Interoperability in Systems Biology: New Methods

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

Summary

Conventional informatics methods have yet to meet requirements for integration of systems biology data. Generalizable methods for analyzing functional relationships between heterogeneous data, based on multidimensional analysis according to direct linking and querying of normalized content, have not been achieved. Given these limits, applications such as detection and reproducible analysis of multiple methods bio markers, real-time individualized molecular diagnosis and individualized patient care based on all available information remain out of reach. Innovative data structures are required to enable interoperability and flexibility of data definitions beyond conflicting "open" and "federated" web and XML-based standards. The potentially sensitive and proprietary nature of systems biology research demands more finely-grained security than offered by the most advanced databases available. Current methods for data integration based on software "wrappers" (translation layers serving as interfaces between different applications and resources) require ongoing programming efforts and expensive computing hardware for global data analysis. The poster outlines challenges to interoperability in systems biology and discusses emerging lightweight, user-centric, agent-based, "intelligent data" methods for functional integration of systems biology data.

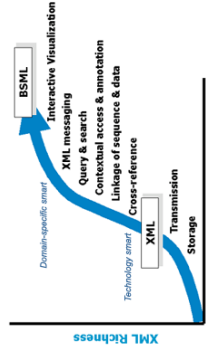
Introduction

Due to its nature, systems biology approaches require interaction between disparate research environments. Resources, research collaboration, reporting and publication involve multiple methods, areas of expertise, proprietary instrumentation, different computer systems and networks. Systems biology-oriented informatics must address disparate data, applications, and selective database access and query methods across networks. Many resources currently are excluded from efficient use and the consequences are 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 systems biology picture. To meet basic requirements for effective systems biology, emerging informatics methods and applications must provide programming-free, low cost methods to bring together the disparate data points required for systems-oriented applications. User-centric data definition, analysis and application frameworks need to be easy to deploy, reproducible and convenient. Systems-oriented informatics applications must provide selective integration of efficient access, normalization and use of disparate data resources. Most importantly, these systems must provide verifiable, validated and reproducible results, and they must be available for audits in a finely-grained security environment with state and processing history.

Methods

Currently, several methods for data integration in systems biology are applied. Such methods include data federation through middleware, data integration through standards and data warehousing, and more recently, data integration through use of data agents. One of the most recent methods, pioneered by companies such as Stanford University; University of Seattle, Washington; Rosetta Inpharmatics; LabBook and the Whitehead Institute, espouses open or federated data ontologies to assist with integration. Although well-formed data standards have proven useful to informatics integration, major challenges remain unaddressed.

Fig. 1: BSM (LabBook) provides an XML-based data integration method using standards supported by several standard converters. It also provides an XML-based browser.



Standards initiatives in molecular biology include AGAVE, BSM, DAS, GEM, GO, and MIAME. Standards may apply at different levels (e.g. interfaces, data definitions, data relationships) and can be based on various coding conventions such as, for example, *ML, RDF or OWL. They require shared, commensurable data ontologies and interoperable DTD schemas. Hardware, programming and processing overhead, conflicting definitions, restricted research flexibility and IP issues associated with shared ontologies characterize common disadvantages. **Middleware-based data federation methods**

evolved from older (TAMBIS, KLEISL) designs. They often involve a combination of object-oriented middleware "request brokers" (e.g. CORBA) supported by Java-based interfaces and "wrappers" translating between various data types and interface requirements. Companies such as Lion (DiscoveryLink) and Invitrogen (Informax/VSL) have applied these methods to relational databases with limited success.

Fig. 2: A classic middleware architecture, depicting a database layer, a server / federation layer and clients for the end-user.

Implementation requirements include shared application programming interfaces (APIs) and interface definition languages (IDLs), as well as new computing hardware. Known shortcomings include high hardware expense and processing overhead, inflexible data definitions, missed data dependencies, and requirements for ongoing wrapping "services". Historically, limited systems biology research has used **data warehousing** as a traditional option. This method requires exhaustive schema definition, and results in static data definitions requiring information technology services as analytical needs and data sources change. Importing data often requires extensive curation and import administration. If distributed data locations are involved, porting of data to a proprietary data warehouse will often be blocked by ownership or privacy issues. Emerging "active data" methods using data objects as agents have been developed by institutions such as IO Informatics and Oak Ridge National Labs (ORNL). Such methods are designed to selectively use existing resources while taking advantage of the benefits from standards, middleware and data warehousing without their inherent constraints. ORNL's "data integration agents" integrate XML-based data ontologies, which do not currently provide the full insights required for complete systems biology. IO Informatics' **Intelligent Multidimensional Object (IMO)** data agents are designed to access and normalize analyzed and un-analyzed data from multiple methods, including database cells and records, image data, quantitative applications output, spreadsheets and standards-annotated data.

Fig. 3a (left): IMOs linked within a "Pool" extract a lightweight, audited multidimensional research environment to normalize and analyze otherwise incompatible, massive distributed resources - ideal for systems biology applications

Fig. 3b (right): Selective IMO access to various remote data

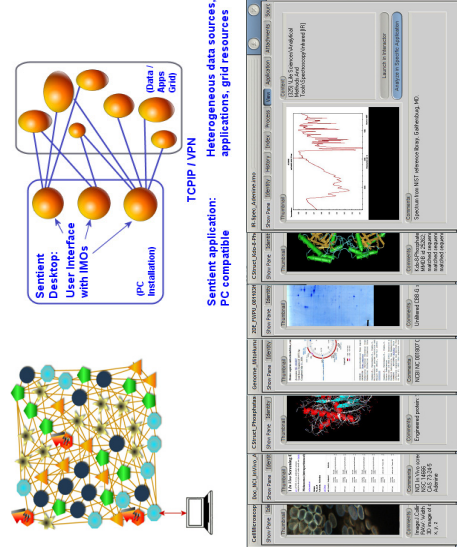


Fig. 3c (bottom): IMOs address virtually any data type - including files from diverse applications, creating an accessible, uniform and secure environment for data exchange.

Central to the IO Informatics' Sentient Desktop is an active, object-oriented data structure. Each IMO interacts directly with users and queries in a no-programming "What you see is what you get" (WYSIWYG) environment. This method provides uniform data access and annotation, multi-dimensional workspace definition and analytical linking within a validated reference database mirror, called **Intelligent Object Pool**. By pointing the software to the data sources to be integrated, a lightweight, unified virtual data pool is created in batch. Source data management and audit across multiple methods are supported via integrated e-lab notebook functions. Within the Sentient Desktop, "point-and-click" relationship definition, normalization and access to systems-oriented analysis and query tools are provided within a uniform framework. This allows users to define applications without programming, such as for multidimensional biomarkers or assays for compound performance. In a high-data density environment efficiency is key. IO informatics objects reduce data size and dimensions to specific "workspace" subsets relevant to a given query or application. Subset definition and processing methods support normalization and analysis of only those parts of raw data needed for processing. This is required for cost-effective high-throughput systems screening and significantly enhances performance.

Results

Best results are obtained from recent methods using the object oriented "data agent" method based on IO Informatics Sentient Desktop. In this case, direct, a dynamic content description based on WYSIWYG selection and definition of multiple data subsets as workspaces provides the framework for multidimensional pathway description and linking and fine-grained field definition for queries. Multiple content and relationship definitions for each data point allow a unified research using all available data resources.

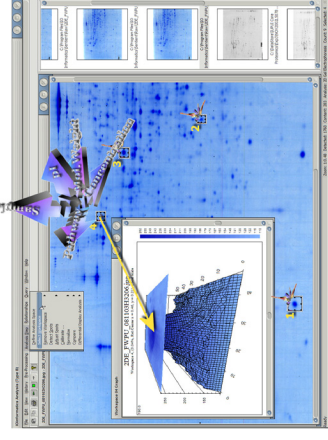


Fig. 4a (left): IMOs within the Sentient Desktop provide methods to extract, analyze and define elements and workspaces according to multiple parameters and pathway linkages.

Fig. 4b (right): The panel shows a biomarker query

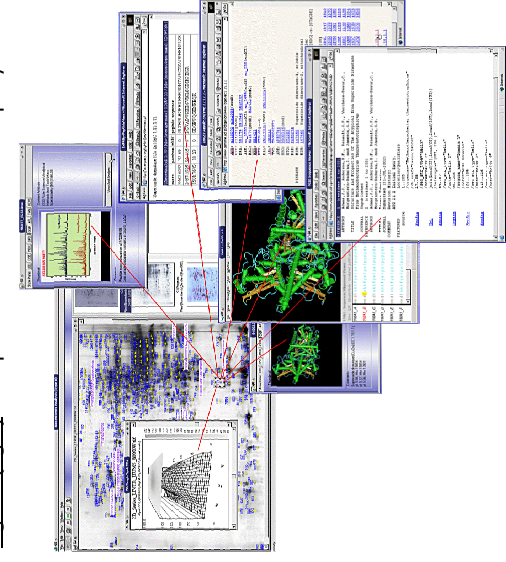


Fig. 5: IMOs can be used to define and link multiple workspaces - manually or through queries

IMOs interact with query tools to external sources to bring content back for integration and local mining. A "Drag-&Drop" version of the "Expert Query" supports parameter transfer directly from quantitatively calibrated scientific images, annotations and other information to a query parsed for submission to multiple external web-based data sources. The "Form Query" delivers structured queries without requiring SQL knowledge to search multiple relational databases and for data mining within the normalized object environment. The architecture supports automated querying, biomarker definition and automated screening for systems-oriented profiles containing data from multiple sources and methods.

Conclusions

The presented user-centric approach to interoperability in Systems Biology provides a innovative way towards faster and easier knowledge building at reduced costs and by using existing IT infrastructure. An analytically integrated interface to content from multiple sources, normalized within the object environment, is provided. "Point and Click" definition of workspaces may be thought of 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, ot multiple method compound activity screening to automate a systems-oriented drug discovery management application. In its common view, all data are considered together within a secure, audited and controlled access environment - thus, increasing the value of analyses, reducing effort duplication and stimulating research innovation through more effective collaborative data sharing across disciplines - all requirements for a true systems biology approach.

References

- Goble, C.A. et al. Transparent access to multiple bioinformatics information sources. *IBM Systems Journal* 40(2): 532-551 (2001).
- Bioinformatics Sequence Markup Language working group (BSML.org). October, 2002. <<http://www.bsml.org/>>
- Gardner, S. "Ontologies, Taxonomies, and Other Buzz Words." *American Genomic / Proteomic Technology* 2(3):10-12 (2002).
- Stanley, R., Gombocz, E., Stevens, R., "Intelligent Molecular Objects: A New Paradigm in Bioinformatics", *Am. Genomic Proteomic Techn.* 2(4) : 22 ff (2002).
- Gombocz, E, Stanley R.: "Intelligent Molecular Objects: Integrated, Data-enabled, Global, Multidimensional Real-Time Analysis in BI and CT", 18th Am. Electrophoresis Society (AES)/Am. Institute Chem. Eng. (AIChE), Proteomics I: State-of-the-art Technologies, Reno, NV: 108e (2001).
- Stanley, R.A., Stevens, R.L., Gombocz, E.A., Heartsong, K.E.: "Intelligent Molecular Object Accelerates Proteomics", *Genet. Eng. News (GEN)* 22 (16): 30-32 (2002).
- Hancock, W.S., Wu, S.L., Stanley, R.A., Gombocz, E.A.: "Publishing Large Proteome Datasets: Scientific Policy Meets Emerging Technologies", *Trends in Biotechnol.* 20 (12): 539-44 (2002) (Review)
- Wong, Limsoon. "Kleisli, a Functional Query System", *Journal of Functional Programming*, 10 (1):19-56 (2000).
- Sheldon, Frederick T., Potok, Thomas E., Elmore, Mark T.: "Dynamic Data Fusion Using an Ontology-Based Software Agent System". Applied Software Engineering Group, Computational Sciences and Engineering Division, ORNL; December 27, 2002.
- Gombocz, E.A.: "Beyond LIMS: Next Level of Scientific Information Integration and 2DE Analysis", 21st Am. Electrophoresis Society (AES)/Am. Institute Chem. Eng. (AIChE), Proteomics I: Applied Bioinformatics, Austin, TX (2004).