Translational Informatics for Systems Biology

by Robert Stanley and Erich Gombocz

Almost since the beginning of the 21st century, life science commentators have been calling it the century of the cell. Having described most of a cell's fundamental machinery through the genomics and proteomics studies that made headlines late in the last century, scientists are now focusing on exploiting this knowledge in vivo. Other 20th-century innovations, such as high-throughput technologies and high-performance computing, are helping scientists measure and assess the impact of different inputs and outputs of biological processes. The resulting combination is systems biology, an approach that treats drug discovery with an engineering approach, where gene function, protein expression, and biochemical processes are harnessed to accomplish specific therapeutic aims.[1,2]

Not all 20th-century innovations, however, are assisting in biology's transformation from science to engineering. In particular, the software used to represent and store information on gene function, protein expression, metabolic fate, biomarkers, and therapeutic pathways fails to support the relationship-centered, integrated approach to data that is the hallmark of systems biology. More often than not, data critical to systems biology are stored in disconnected databases—silos from which data must be extracted and processed before they can be associated with other related data. The most common silos are relational databases, the earliest type of database model, and federated systems, which abstract data into a "metalayer" that can be queried from a top-level interface. Both methods severely limit how data can be accessed, used, and shared. At best, the systems offer shallow query/browse interfaces that must be updated to accommodate new types of data. At worst, they require scientists to convert all the data that they might want to consult into one monolithic database format or another.[3]

Systems biology requires a new approach to managing data, one that is led by the question or relationship that scientists need to explore rather than the format the data take. Research organizations must be capable of thinking outside of the boxes into which they are placing data in order to better focus on the data themselves (i.e., the specific attributes and associations that make one piece of datum unique and connect it to other types of data). Ultimately, these connections are made not just by mining the data directly but also through the continued work of domain experts as they work with and refine results. Based on iterative literature research, data analysis, and peer review, scientists weigh the relevance of individual data values generated by representative experiments and create the associations and relationships that explain how these values (and the biological functions they represent) participate in a system. An effective software infrastructure will enable these associations to be made across applications and shared throughout an organization while preserving the integrity and security of underlying data structures. It will turn an organization's disparate boxes of data into a fluid, interconnected data pool.

Clinical Data, Inc. (Newton, MA) is one organization interested in getting its data "out of the box." A wholly owned part of the company's **Cogenics** division, **Icoria** (Research Triangle Park, NC) has since its 1997 founding developed a systems-based approach to drug discovery. By analyzing relationships among data in three areas—gene function, biochemical profiles, and quantitative tissue studies—**Icoria** identifies biomarkers, specific features that describe or predict how a system responds to a disease, therapy, or other perturbation. With more and more complex data rapidly accumulating in its focus areas, **Icoria** sees efficient integration, management, association, and representation of the data as increasingly important to its business strategy.

In June 2002, the company was awarded an $11.7 million, five-year Advanced Technology Program (ATP) grant—the largest ever awarded in the field of bioinformatics—from the National Institute of Science and Technology. To complete the project, **Icoria** is partnering with **IO Informatics** (Emeryville, CA), a software company that offers a novel approach to describing, accessing, and using data. The intelligent multidimensional object (IMO) is a portable database record that stands to transform data management in the same way that PDFs have transformed file-sharing.

Making semantics work

To better understand how an IMO differs from traditional data handling methods, one can look to an analogous example in Adobe® Acrobat's® (**Adobe Systems Inc.**, San Jose, CA) PDF format. The portable document format takes the product of any application and, in a simple, postprocessing step, converts the file into an "open" format that can be shared free of its original home application. Today, PDFs are ubiquitous, but when the format was first introduced in the early 1990s, many users were skeptical. Why would someone need to convert a file into a portable document format? What would be the advantage of viewing a document outside of its home application? Would users want the full functionality of the home application when working with a file? These questions seem shortsighted today. In fact, the ability to preserve a document's formatting and appearance outside of a single application has revolutionized the way information is published and disseminated, and will continue to change the way information is gathered and used.[4]

Like a PDF, an IMO involves a conversion. Rather than a postprocessing step, however, in which an application product is converted into an open format, IMOs are created in a preprocessing step that transforms a specific data type into a freeform relational object. By pointing and clicking, users can select subsets of data contained within IMOs, which then become values available for analysis. Subset selections may be propagated across similar data sets to create new data fields where none existed before. The data themselves remain intact and housed securely in the box into which they have been placed. IMOs simply point to these data much like a Windows "shortcut" (only much more sophisticated). In other words, IMOs define and characterize the data they describe. As a result, they enable users to describe, utilize, relate, and share data in any number of ways. Most importantly, an IMO can reside anywhere, can be opened anywhere, and can define anything (from a single data value captured from a larger data set or an entire spreadsheet to images and unstructured data or the results of a search query).

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. Biologists build associations through:

- Attributes, which associate data according to their relationship or position within a predefined ontology, or area of interest, i.e., "These data look like this and belong in this area"
- Attachments, which associate data directly to other IMO level data and queries, i.e., "These data go along with these data"
- Annotations, which associate data via links to analytical content subsets created by users, i.e., "I analyzed these data, and I think this particular feature is interesting or important"
- Subset linkages, which associate content subsets within IMOs to queries and to other content subsets within objects, i.e., "I analyzed these data, and I have found that the particular subset feature that I noted here is explained by or influences something in these data over here"
- Queries, which identify and retrieve similar, covariant, manually, or otherwise semantically associated objects, i.e., "I am interested in all of the data related to this particular feature."

If the fundamental capabilities of the IMO sound familiar, it is because they are based on the same principles that underlie the "Semantic Web," which utilizes an innovative data description method for representing data subjects, properties, and relationships in a coherent, meaningful way that has been championed by Web inventor Tim Berners-Lee, and others.[4–6] The Semantic Web, however, has proven difficult to implement. Users, rather than IT staffs, must be able to interpret the connections between data, and without standards to support these interpretations, the meaning, or semantics, is lost.

continued

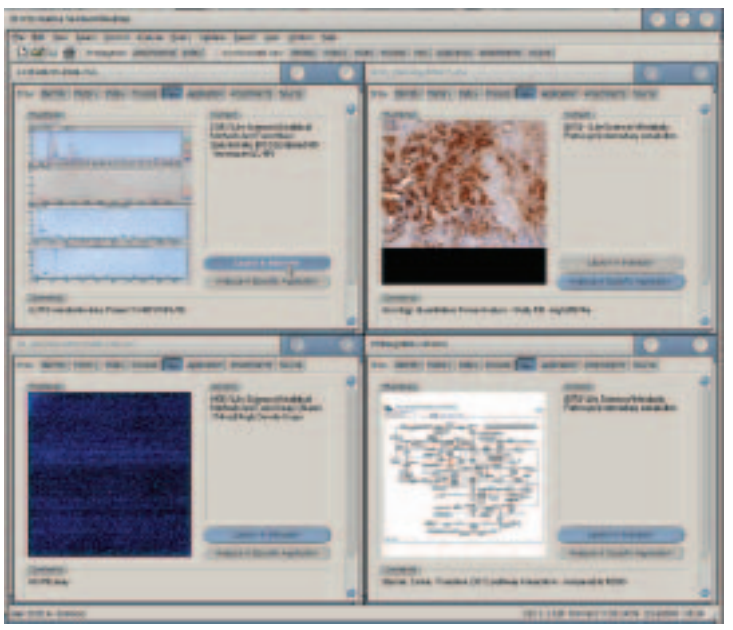

Figure 1 Heterogeneous data sources in biomarker research and discovery. The figure shows several data sources key to the Gene to Cell to System approach to drug discovery displayed within the unified Sentient environment. The data include LC-MS metabolites, gene expression, quantitative tissue analysis, and pathway diagrams. The software filters for profiles and presents these different data in context, revealing otherwise undetected data relationships.

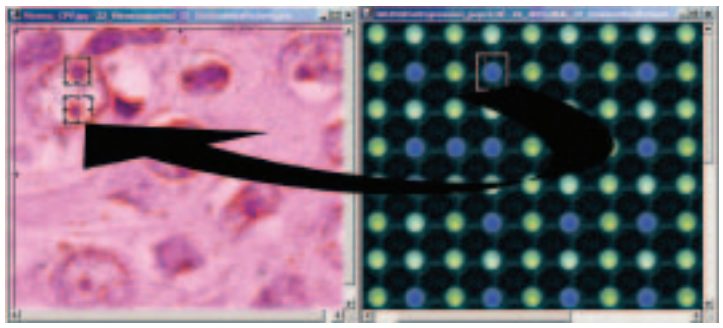

Figure 2 Data values defined and linked as subsets relating tissue analytics and gene expression data. Two example elements on the tissue image at left are shown linked to the single gene expression data marker at the right. The resulting IMO would describe the relationship between these data sets. In addition, as part of the hepatotoxicity iPool, the IMO helps build a research and screening profile for liver toxicity.

To make semantics meaningful outside IT departments, **IO Informatics** has created the Sentient Suite, a software environment that enabled domain experts, rather than software programmers, to create coherent representations of systems using real data from all sources relevant to their research. With the software, biologists can point and click to create or capture semantic properties, such as nodes, edges, and pathways in data, which were once accessible only to computer scientists. They can structure and define data relationships, view these relationships, query them, and capitalize on them—all within a secure, auditable framework that helps organizations comply with regulatory requirements and accumulate valuable intellectual property. The software's interactive tools, search capabilities, and security features help users turn the basic distillation capabilities of the IMO into a usable, actionable service.

Just as the PDF has not replaced but instead adds value to existing applications, Sentient and the fundamental IMO format do not disrupt or alter an organization's existing data infrastructure. By creating IMOs, users can turn any collection of databases and application files into a rich, fluid data pool—the Sentient iPool. An organization's underlying data remain intact, safely stored in whichever box they reside. The software simply enables users to think outside of those boxes.

Supporting systems biology

Clinical Data has built one of the largest independent sources of pharmacogenomics services in the world through the acquisition in 2005 of three companies: **Genaissance Pharmaceuticals** (New Haven, CT), which develops DNA-based diagnostics; **Lark Technologies** (Houston, TX), a contract research organization specializing in custom molecular biology services; and **Icoria**. The combined knowledge of each of these companies feeds into **Clinical Data's** proprietary Gene to Cell to System approach to pharmaceutical and life science discovery. Mapping data from gene expression, biochemical profile, and quantitative tissue studies onto biological pathways helps identify specific biomarkers associated with both drug action and patient response, enabling **Clinical Data** and its clients to triangulate the causes of disease and develop targeted therapeutics (see *Figure 1*).

The challenge for the company, though, has been maintaining coherence among the various types of data accumulated in its Gene to Cell to System methodology. The needs are threefold. Technology must first be able to organize and integrate the vast, complex data streams generated by high-throughput technologies. It must also be able to account for the high degree of variability—in both processes and systems—encountered when developing useful profiles and controls for biological processes. Most importantly, the technology must also be easy for scientists to use. Domain experts must be able to work directly with data from each source within a common environment where they can endow disparate, disconnected bits of data with coherence and relational semantics. System-oriented software capable of both creating and managing these semantic connections will inform discovery decisions today while supporting critical next-generation applications, such as targeted therapeutics and near real-time adverse event reporting.[7]

Systems biology applications ultimately are not about the data and where they are stored. They are about the work to be done with the data—what they mean and how they can be used. The same data might support several different complex screening profiles, even competing models, for example, and informatics should not force those data to inhabit one space or another. In systems biology, informatics must be capable of creating mutable connections among data types that can themselves be efficiently aggregated, semantically integrated, and easily shared and applied throughout an organization and beyond—from the laboratory bench to clinical point-of-care and back again.

In the final phase of the ATP grant, **Clinical Data** will work with **IO Informatics** to create a translational informatics environment based on IMOs and the Sentient technology platform. This environment will encompass all of the relevant data collected throughout the Gene to Cell to System process. The company has already successfully completed three technical milestones in this ATP grant: It has developed, validated, and analyzed two increasingly complex coherent data sets and has produced prototype data coherence tools. The data sets comprise information collected from a study of how acetaminophen affects the liver of rats. In addition to assessing the clinical effects of acetaminophen, the study also discovered biomarkers associated with liver toxicity (hepatotoxicity).

The first step in developing the translational informatics environment will be to create IMOs describing these validated data sets and building an iPool for the hepatotoxicity information. Once data are available in the IMO format, important causal elements can be selected, defined, and linked, to begin creating models reflecting types of toxicity or compound activity (*Figure 2*). Unlike other data management exercises, which require importing and reformatting data, creating an IMO is a simple conversion step that leaves the original data intact. As a result, the company's existing data will be accessible and usable throughout the process, and the conversion process itself will take just a few weeks. Next, to ensure efficient and repeatable processes, core users will define and implement work flows to streamline the creation of IMOs by group members. In many cases, these work flows are already dictated by existing laboratory information management systems and work flow documents. Once the basic collaborative approach has been agreed upon, **Clinical Data** will release the system to additional groups.

With Sentient in place, biologists will be able to import, merge, and create ontologies consisting of formal data definitions, interaction pathways, and other semantically important associations, which they can use to better understand and represent disease models, compound activities, and adverse events. **Clinical Data** scientists will also be able to create targeted, dynamic profiles to iteratively mine data for known and emerging models, activities, and events as the data enter the system. These associations and profiles will be searched and viewable through the Sentient client application and through a Sentient Web portal, which enables researchers to apply complex queries in an ad hoc manner; use automated, dynamic filters to screen for data similar to threat profiles; and click through from results back to the original data. Project goals also include continued efforts to improve system usability, particularly in the area of clinical data capture. Ultimately, creating a shareable data type should be as easy as dragging and dropping the data into a folder where IMO conversion will occur.

IMOs and the Sentient Suite will support any number of business-level needs at **Clinical Data, Inc.** by:

- 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 that tracks not just the creation of data, but how the data are used in an organization.

Out of the box and into the pool

The opportunity to engineer new therapeutic pathways for predicting and controlling disease requires new informatics environments. These environments must be capable of creating coherence and preserving connections among the complex, disparate, and increasingly large collections of biologically relevant data. Together, **Clinical Data, Inc.** and **IO Informatics** are building the next generation of tools for target assessment and life sciences research. These tools, based on an independent multidimensional object, will allow scientists to build, share, and search relationships among data. What the data signify and how they can be used will be defined and dictated by scientists, rather than by the boundaries of a software application or the initiative of IT experts.

References

1. Stanley, R.; Hancock, W. Bioinformatics in the clinic: challenges and opportunities for improved trials and clinical care. *Genom. Proteom. Tech.* **2003**, *3*(3), 29–36.
2. Glassbrook, N.; Ryals, 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. The challenge of publishing large proteome datasets: the meeting of scientific policies and emerging technologies. *Trends Biotechnol. (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.; Stuckenschmidt, H. C-OWL: contextualizing ontologies. *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.

Mr. Stanley is Chief Technology Officer, and Dr. Gombocz is Chief Science Officer, ***IO Informatics, Inc.,*** *2000 Powell St., Ste. 520, Emeryville, CA 94608, U.S.A.; tel.: 510-420-8400; fax: 510-420-8440; e-mail: rstanley@io-informatics.com.*