Editor: Carole Goble • carole.goble@manchester.ac.uk

Linked Data in Drug Discovery

Michel Dumontier • *Carleton University*

David J. Wild • *Indiana University*

Drug discovery presents many challenges, but several linked data initiatives are under way to address the huge increase in the amount of data available from chemistry, biology, and drug discovery in the past two decades.

"Information is cheap. Understanding is expensive." This simple but astute insight was recently made by Karl Fast in regard to the general problem of how we can best use the vast amounts of data now available in the world. However, in drug discovery, both information and understanding are expensive.

Drug discovery involves finding therapies (usually chemical compounds) that elicit certain desirable responses in the body without creating unacceptable, adverse side effects. Until the 1960s, this process was entirely empirical. It often involved examining plants and other natural substances with reported beneficial effects to identify in them the chemical compounds responsible for their effects. The most widely used and arguably most valuable drugs today are derived from this process: painkillers, such as aspirin, and antibiotics are two clear examples. From the 1960s onward, increased understanding of the body's molecular mechanisms, as well as the diseases that affect it, enabled a more mechanistic understanding of how drugs act, and began the era of "rational" drug discovery. This is still the prevailing paradigm and generally involves identifying a protein (*target*) involved in a disease state (for example, a protein involved in replicating cells might be implicated in cancer and thus form a potential drug target). This approach has had success stories (such as HIV drugs), but some apparently successful drugs have crashed out of the market due to unforeseen, rare side effects.

The 1990s saw an explosion, primarily in the pharmaceutical industry, in innovative (and expensive) experimental techniques that produce large amounts of data about chemical compounds, protein targets, genes, biological pathways and cells, and their role in how the body functions and in disease states. In the past decade, initiatives such as the Molecular Libraries Initiative,[1] the US Environmental Protection Agency's ToxCast program (www.epa.gov/ncct/toxcast/), and the Human Genome Project have brought this technology (and the resulting data deluge) into the public sphere. This effort represents a third phase of drug discovery that's still ongoing and constitutes a rational investigation scaled up by orders of magnitude. The promise is that producing this data will result in new breakthroughs in drug therapies, particularly for significant diseases such as cardiovascular disease, cancer, and diabetes. Whereas previously a few hundred compounds might be tested for activity against a protein target, now hundreds of thousands can be tested.

However, the rational approach's limitations remain — namely, that it focuses only narrowly on how a compound acts on a particular protein target and doesn't consider the compound's wider impact on the body, or the unexpected cascade effects of interfering with these targets. Thus, the recurring problems are those of efficacy (making drugs that work in the test tube work in the body) and safety (anticipating undesirable side effects).

These experimental techniques have resulted in large, siloed data repositories — both in the public sphere and internal within companies — in which the silos map to the particular experiment types. For example, large public datasets such as

 1089-7801/12/$31.00 © 2012 IEEE

PubChem Bioassay and ChEMBL pertain to how chemical compounds act on protein targets. Others, such as UniProt, pertain to the function and biological pathways of genes (and their protein targets); yet more pertain to drug side effects, gene regulation, clinical findings, and so on. (Table A in the Web appendix at http://doi.ieeecomputersociety.org/10.1109/MIC.2012.122 gives more information about the datasets and repositories mentioned in this article.)

In aggregate, these datasets offer a much more sophisticated understanding of the complex network of actions drugs have on the body. The Semantic Web is a critical enabling technology for this field. By semantically annotating and linking data, researchers can search, explore, and mine the large complex relationships of entities important to drug discovery (drugs, chemicals, proteins, genes, pathways, cells, diseases, and side effects) in an integrated fashion. Many researchers are exploring the possibilities inherent to such integration, including new scientific areas such as systems chemical biology.[2,3]

Current Initiatives

Various initiatives are under way that use linked data in drug discovery, both in academia and in the pharmaceutical industry, and sometimes crossing both (such as the EU Innovative Medicines Initiative's Open-PHACTS project; www.openphacts.org). Uptake of linked data in the pharmaceutical industry is ongoing but is currently at an early stage. Most companies' information systems currently center on large relational databases, and companies routinely link public datasets into these databases (for instance, by creating separate tables for public datasets, sometimes cross-linked with internal data). However, these aren't generally semantically annotated, which is required to integrate and make efficient use of the data.[4]

Still, such integration has fueled numerous tools that can find data paths across datasets.[5] Several pharmaceutical companies are exploring true linked data and semantic methods, mainly using prototype RDF triple stores and semantic searching using commercial tools such as TopBraid (www.topquadrant.com), IO Informatics Sentient (www.io-informatics.com), and Franz Allegrograph (www.franz.com). However, such methods aren't yet mainstream, and centralized repositories remain in relational format, generally not well linked to the "outside world."

One of the earliest public collaborative efforts to develop ideas, standards, and projects around the use of linked data for pharmaceutical and clinical research arose from the W3C's Semantic Web for Health Care and Life Sciences Interest Group (HCLSIG). The HCLSIG is an open forum that puts executives, scientists, researchers, programmers, and policymakers together to work toward developing standards and demonstrate Semantic-Web-enabled biomedical solutions that support translational research. The HCLSIG has worked on several problems, including those pertaining to capturing scientific discourse, integrating life science and clinical data, providing guidelines for publishing proteomics and genomics data, and undertaking biomedical research.

Understanding drugs' impact on a dynamical network is increasingly important.

One early HCLSIG demonstration involved exploring hypotheses related to Alzheimer's disease over an integrated store of knowledge.[6] This effort focused on pathological information from SenseLab, neuronal circuitry from CoCoDa, receptor-ligand data from the PDBSP K_i database, biomolecular interactions from the Biomolecular Interaction Network Database (BIND), gene information from NCBI Gene, antibodies from the Antibody Directory, pathways from the Kyoto Encyclopedia of Genes and Genomes (KEGG) and Reactome, and terminology from NCI Metathesaurus and a growing collection of open biomedical ontologies (OBOs). A big part of this effort was delineating the methodology to marshal a heterogeneous collection of source data (flat-files, XML files, SQL databases, and so on) into RDF and dealing with the complexity that arises from mashing together disparate data sources with different types and relations. In the end, researchers can query the knowledge base for information that relates to hypotheses and clinical guidelines, molecular targets, antibodies, and mouse models, all of which contribute to advancing science and improving healthcare.

Delineating on and off drug targets and understanding drugs' impact on a dynamical network is increasingly important in drug discovery. Recent work shows how researchers developed and used an HIV-focused mashup of biological resources (Affymetrix array, RefSeq, Gene, Online Mendelian Inheritance in Man [OMIM], the HIV-1 Human Protein Interaction Database [HHPID], PubMed, Medical Subject Headings [MeSH], and Gene Ontology) to identify a protein interaction network that emerges from significantly expressed genes during a time-course microarray in the first hours of an HIV infection of primary human macrophages that had or had not been treated with interferon, an antiviral product.[7] The paper

demonstrates that the difference between the interferon-treated and non-treated networks identifies interferon-responsive elements, while an analysis of MeSH-enriched terms from associated publications suggests molecular and disease associations that could be important in better understanding off-target-induced side effects. A key biological data resource that powered this and many other studies is Bio2RDF,[8] which provides more than 40 datasets, including chemicals, genes, gene expression, proteins, pathways, molecular interactions, diseases, and scientific literature.

Some significant demonstrations have occurred as regards the possibilities of large-scale semantic integration in drug discovery. Chem2Bio2RDF integrates a wide range of drug-discovery-related datasets covering compounds, drugs, targets, genes, assays, diseases, side effects, and pathways in a single, integrated format, annotated with a chemogenomic ontology called Chem2Bio2OWL.[9] By embedding bioinformatics and cheminformatics functionality into SPARQL queries, researchers can issue simple, integrative queries to extract new and meaningful relationships among multiple datasets. In the original Chem2Bio2RDF paper, the authors presented three proof-of-concept case studies showing how the networks can help researchers identify

- compounds that share the same multi-target poly-pharmacological profile as known drugs by linking PubChem Bioassay compounds with Drugbank compounds via targets;
- 58 potential multi-target inhibitors of the mitogen-activated protein kinase (MAPK) signaling pathway by linking KEGG and Reactome pathways with PubChem Bioassay via targets; and
- possible involvement of pathways in drug hepatotoxicity by linking KEGG pathways with adverse effects recorded in Drugbank.

Chem2Bio2RDF is now part of the Linked Open Data cloud. Recent work has shown how researchers can use this large-scale linked data in path finding[10] and drug-target prediction.[11]

The idea of an open, integrated, semantic repository of drug discovery data has been taken up by the large OpenPHACTS project, which is building an extensive, precompetitive Open Pharmacological Space (OPS) of linked data for drug discovery that involves multiple academic institutions and pharmaceutical industry partners. Keeping such repositories up to date is a significant challenge, and one that's greatly assisted when data providers make available RDF versions (annotated with appropriate ontologies) of their most up-to-date data at source. For example, the European Bioinformatics Institute (EBI) is now making an RDF version of their ChEMBL dataset available.

Integration with clinical and patient data is considered key to translational research. By combining linked open drug data (LODD) with a database of clinical trials and patient data, HCLSIG participants showed how the Translational Medicine Ontology (TMO) could be crafted to provide a conceptual schema by which queries could more easily be constructed across drugs, genes, proteins, pathways, diseases, and pharmacogenomic knowledge.[12] The paper demonstrates how such a semantically integrated knowledge base can provide insight into drug repurposing through common target modulation and suggest alternative therapeutic avenues by comparing the mechanism of actions.

Although ontologies play a dominant role in providing types for semantic annotation, researchers are increasingly using them as a means of enhanced knowledge representation for drug discovery. Anika Oellrich and colleagues report on identifying causal genes for an underlying disease by comparing the similarity of phenotypes arising from mouse models with those of the human condition.[13] Another study demonstrates how researchers can use mappings in biomedical ontologies to integrate pharmacogenomics databases to enable novel analyses.[14] The authors identified disease-prone pathways using a novel, multi-ontology enrichment analysis that employed the Human Disease Ontology, PharmGKB, the Comparative Toxicogenomics Database, Drugbank, the Anatomical Therapeutic Chemical Classification System (ATC), and the MeSH thesaurus. Moreover, significant associations between drugs and pathways suggest interesting avenues for computational drug repurposing.

Perspective

Integration is still a huge challenge in drug discovery, and ontology-based mapping of datasets from different experimental and other sources remains nontrivial. One major consideration that lies at the heart of Semantic-Web-based data integration is using a common syntax to name data resources. Identifiers.org aims to provide this common naming service[15] and is being backed by a growing consortium, including the EBI, the W3C HCLSIG, the BioSharing consortium (www.biosharing.org), Bio2RDF, the Systems Biology Markup Language (SBML; http://sbml.org), Reactome, and even more organizations. The EBI itself is pursuing trial investigation to provide RDF-based data access for some of its more prized datasets, including UniProt for proteins, ChEMBL for bioactive chemicals, Array Express for gene expression data, and Biomodels for computable models of biology. Persistent naming also acts as the foundation for nanopublications,[16]

which can capture meaningful biological statements and link their sources.

As Semantic Web search engines such as Sindice crawl websites, blogs, twitter feeds, published RDF/OWL files, and SPARQL end points for interlinked data, Semantic Web technologies will demonstrate their power to federate and integrate the world's knowledge in a manner that existing technologies simply can't.

As simpler and more effective methodologies emerge to facilitate data integration and publishing, opportunities abound in the innovative development and application of data mining algorithms to the resulting knowledge graph (and subsets of it). Approaches such as edge and path weighting, along with advanced graph prediction, enrichment and quantitative analyses, data mining algorithms, and enhanced visualizations could allow for intelligent evidence gathering, clustering, consensus building, and exploration. Companies will be able to ask, "What is the most significant proprietary data that we own and why?" and extract and mine comprehensive knowledge graphs related to particular therapeutic areas.

Pharmaceutical and other companies will get the best added value by integrating proprietary, purchased commercial data with public data, but this must happen in a way that intelligently maps the sets together and allows selective data filtering. Further efforts to integrate healthcare data (including electronic medical records) and publication and personnel data will increase the possibilities for this field, including in such important new areas as pharmacogenomics, personalized medicine, and health biomarkers.[17]

References

1. C.P Austin et al., "NIH Molecular Libraries Initiative," *Science*, 2004, vol. 306, no. 5699, 2004, pp. 1138–1139.
2. T.I. Oprea et al., "Systems Chemical Biology," *Nat'l Chemical Biology*, vol. 3, 2007, pp. 447–450.
3. D.J. Wild et al., "Systems Chemical Biology and the Semantic Web: What They Mean for the Future of Drug Discovery Research," *Drug Discovery Today*, vol. 17, 2012, pp. 469–474.
4. T. Slater, C. Bouton, and E.S. Huang, "Beyond Data Integration," *Drug Discovery Today*, vol. 13, nos. 13–14, 2008, pp. 584–589.
5. Q. Zhu et al., "WENDI: A Tool for Finding Nonobvious Relationships between Compounds and Biological Properties, Genes, Diseases and Scholarly Publications," *J. Cheminformatics*, vol. 2, 2010, p. 6.
6. A. Ruttenberg et al., "Advancing Translational Research with the Semantic Web," *BMC Bioinformatics*, vol. 8, supplement 3, 2007, S2.
7. M.A. Nolin et al., "Building an HIV Data Mashup using Bio2RDF," *Brief Bioinformatics*, vol. 13, no. 1, 2012, pp. 98–106.
8. F. Belleaud et al., "Bio2RDF: Towards a Mashup to Build Bioinformatics Knowledge Systems," *J. Biomedical Informatics*, vol. 41, no. 5, 2008, pp. 706–716.
9. B. Chen et al., "Chem2Bio2RDF: A Semantic Framework for Linking and Data Mining Chemogenomic and Systems Chemical Biology Data, *BMC Bioinformatics*, vol. 11, 2010, p. 255.
10. B. He et al., "Mining Association Paths in Relational Biomedical Data," *PloS One*, vol. 6, no. 12, 2011, e27506.
11. B. Chen, Y. Ding, and D.J. Wild, "Assessing Drug Target Association using Semantic Linked Data," *PLoS Computational Biology*, vol. 8, no. 7, 2012, e1002574.
12. J.S. Luciano et al., "The Translational Medicine Ontology and Knowledge Base: Driving Personalized Medicine by Bridging the Gap between Bench and Bedside," *J. Biomedical Semantics*, vol. 2, supplement 2, 2007, S1.
13. A. Oellrich et al., "Improving Disease Gene Prioritization by Comparing the Semantic Similarity of Phenotypes in Mice with Those of Human Diseases, *PLoS One*, vol. 7, no. 6, 2012, e38937.
14. R. Hoehndorf, M. Dumontier, and G.V. Gkoutos, "Identifying Aberrant Pathways through Integrated Analysis of Knowledge in Pharmacogenomics," *Bioinformatics*, vol. 28, no. 16, 2012, pp. 2169–2175.
15. N. Juty, N. Le Novère, and C. Laibe, "Identifiers.org and MIRIAM Registry: Community Resources to Provide Persistent Identification," *Nucleic Acids Research*, vol. 40, Jan. 2012, pp. D580–586.
16. G.P. Patrinos et al., "Microattribution and Nanopublication as Means to Incentivize the Placement of Human Genome Variation Data into the Public Domain," *Human Mutation*, 26 June 2012.
17. M. Samwald et al., "Semantically Enabling Pharmacogenomic Data for the Realization of Personalized Medicine," *Pharmacogenomics*, vol. 13, no. 2, 2012, pp. 201–212.

Michel Dumontier is an associate professor of Bioinformatics in the Department of Biology, Institute of Biochemistry and School of Computer Science at Carleton University, Ottawa, Canada. His research aims to develop semantics-powered computational methods to increase our understanding of how living systems respond to chemical agents. Dumontier is Scientific Director for the open source Bio2RDF linked data for life sciences project and currently serves as a chair for the W3C's Semantic Web in Health Care and Life Sciences Interest Group (HCLSIG). Contact him at michel.dumontier@gmail.com.

David J. Wild is an assistant professor of informatics and computing at Indiana University, where he directs the Cheminformatics and Chemogenomics Research Group (CCRG). This research focuses on large-scale semantic integration, searching, and prediction on chemical and biological information. Wild is co-editor in chief of the *Journal of Cheminformatics*. Contact him at djwild@indiana.edu; http://djwild.info.