ELSEVIER

Systems chemical biology and the Semantic Web: what they mean for the future of drug discovery research

David J. Wild[1], Ying Ding[2], Amit P. Sheth[3], Lee Harland[4], Eric M. Gifford[5] and Michael S. Lajiness[6]

[1] Indiana University School of Informatics and Computing, Bloomington, IN, USA
[2] Indiana University School of Library and Information Science, Bloomington, IN, USA
[3] Wright State University Department of Computer Science and Engineering, Dayton, OH, USA
[4] ConnectedDiscovery, Deal, Kent, UK
[5] Pfizer Global Research and Development, Groton, CT, USA
[6] Eli Lilly, Indianapolis, IN, USA

Systems chemical biology, the integration of chemistry, biology and computation to generate understanding about the way small molecules affect biological systems as a whole, as well as related fields such as chemogenomics, are central to emerging new paradigms of drug discovery such as drug repurposing and personalized medicine. Recent Semantic Web technologies such as RDF and SPARQL are technical enablers of systems chemical biology, facilitating the deployment of advanced algorithms for searching and mining large integrated datasets. In this paper, we aim to demonstrate how these technologies together can change the way that drug discovery is accomplished.

Introduction

Traditionally, drug discovery paradigms involve identifying a protein target that is implicated in disease processes, and then identifying one or more chemical compounds that can safely interfere with these targets, either by activation (agonism) or inhibition (antagonism), and that are then prioritized and tested further for safety and inclusion into clinical trials. Recent failures to bring the projected numbers of new drugs to market, along with increasing postmarket drug withdrawals, have resulted in the questioning of this methodology, in particular the beliefs that the 'reductionist' approach is too simplistic, and cannot properly assess risk of *in vivo* efficacy and safety problems.

Rather than reducing a complex system to simplistic models, the emerging field of chemogenomics seeks to build holistic models around the effects of compounds on multiple biological targets and pathways. Recent work in this area has mostly focused on identifying and predicting different aspects of small-molecule–protein interactions, such as: the use of chemical similarity as a probe of protein function [1]; the prediction of off-target effects of drugs using network methods [2,3]; repurposing of known drugs for new targets [4]; drug–target interaction networks for exploring the kinome [5]; mapping assay networks onto biological networks to relate compounds and targets [6]; and using drug side-effect profiles to predict new biological targets [7]. Although it is sometimes possible to have the luxury of a full matrix of experimental results for compounds against protein targets [5], most work has focused on computational prediction based on available data. Although in its early stages as a research discipline, chemogenomics has demonstrated some early successes, including successful prediction of new targets for known drugs that are later experimentally verified [1,4]. Chemogenomics is limited in that it only considers the relationships of chemical compounds and genes (along with their target proteins). A wider approach has been proposed that involves analyzing networks of many kinds of data including compounds, targets, genes, diseases, side effects, metabolic pathways and so on, with the purpose of investigating the complex systematic effects of drugs and other chemical compounds on biological systems. This field is tentatively termed systems chemical biology [8], although the term chemical systems biology has also been used [2].

Corresponding author: Wild, D.J. (djwild@indiana.edu)

1359-6446/06/$ - see front matter doi:10.1016/j.drudis.2011.12.019

Please cite this article in press as: D.J.. Wild, et al., Systems chemical biology and the Semantic Web: what they mean for the future of drug discovery research, Drug Discov Today (2012), doi:10.1016/j.drudis.2011.12.019

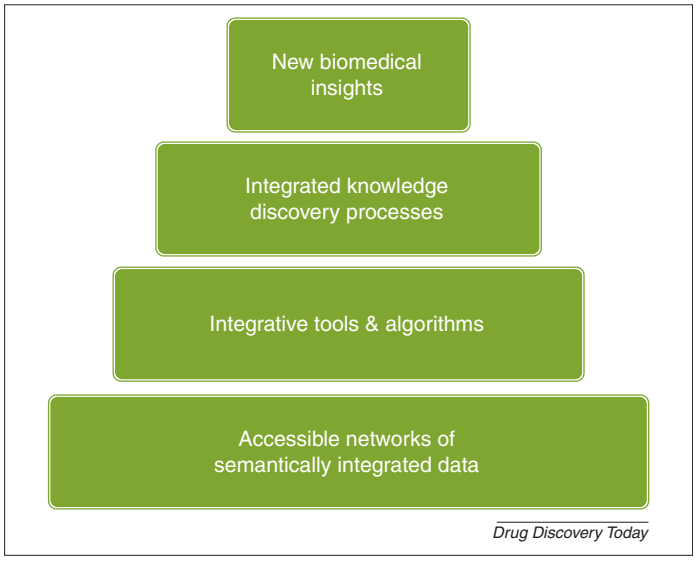

FIGURE 1

Stack of integrative capabilities required to realize the aims of systems chemical biology and chemogenomics.

Realizing this approach requires a high level of integration of chemical and biological databases, and new kinds of computational tools to enable the use of these integrated databases. Several publicly accessible databases address the integration of chemical and biological data (including the curation and quality control needed for integration); examples include ChEMBL (https://www.ebi.ac.uk/chembldb), a curated set of compound–target interactions, and ChemProt [9], a dataset of over 700,000 unique chemicals and their interactions with over 30,000 proteins (including disease-associated protein–protein interactions). However, the lack of a generalized framework for integrating data sources hampers research in chemogenomics and systems chemical biology, and makes it difficult to replicate published research on other datasets [10].

The case has been made previously for the large-scale integration of heterogeneous datasets, and that this integration must be semantic (i.e. there must be a shared understanding of meaning of and accessibility to tools across the datasets) [10–12]. Such integration is a necessary precursor to systems chemical biology, particularly given the diversity of large public datasets now available describing chemical and biological entities and the relationships between them (e.g. PubChem, ChemSpider, UniProt, ChEMBL, KEGG, to name a few). However, such integration does not affect systems chemical biology: new kinds of algorithms and tools are needed that use these integrated sets, and new methodologies are needed to map these algorithms and tools to real drug discovery problems. In fact, therefore, a stack of capabilities, based on integrated data, is needed (Fig. 1).

Semantically integrated networks of data

Why semantic networks are useful

There are two key advantages of using integrated networks of data from multiple datasets over searching those datasets individually: first, the ability to search multiple datasets through a single framework, with a common understanding of terminology (ontology); second, the ability to search relationships and paths of relationships that go across different datasets. Below, we describe a variety of recent languages and tools designed to affect these types of searching, and which are generally considered to constitute the 'Semantic Web'. Some of the terms frequently found in this field, and in particular in its relation to life sciences, are described in Box 1, and an example of this use is shown in Fig. 2. In Fig. 2 (derived using a tool described later), paths of association are shown between a drug (rosiglitazone) and a side effect (myocardial infarction). These paths cross multiple datasets – in this case the DrugBank drug database, the PharmGKB pharmacology database, the UniProt gene database and the SIDER side-effect database. Datasets that describe a direct relationship are shown in gray rectangles. Using a pathfinding algorithm, literature-supported paths have been found that link the drug and side effect via these datasets. In Fig. 2 the algorithm has found a set of genes and/or targets that rosiglitazone interacts with, and that also interact with other drugs that have the known side effect of myocardial infarction. This kind of search could be used to identify potential risk factors for new drugs, as well as suggesting potential mechanisms of action for side effects [in the case shown here, the interaction of rosiglitazone with apolipoprotein E (APOE) could be

BOX 1

Terminology used in the Semantic Web and life sciences

RDF: Originally used for describing resources on the Web, RDF is now the most fundamental way of representing data points and relationships between them in the Semantic Web. An RDF triple is a noun–verb–noun construct that describes the relationship between entities, often using an 'ontology' to define valid types of entities and the relationships between them. RDF can be represented in a variety of formats, including XML.
OWL: An XML format language for describing ontologies. Commonly used in conjunction with RDF, although an ontology is not required for RDF use. There are three variants of OWL: OWL-Full, OWL-DL (a subset of OWL-Full) and OWL-Lite (a subset of OWL-DL).
Ontology: A formal representation of the terms, relationships, concepts and entities within a particular domain. It provides a set of constraints and generalizations for RDF. For example, it can define that 'compound' and 'target' are valid entities, and 'inhibits' is a valid relation between a compound and target.
SPARQL: A language for describing search queries on RDF. A SPARQL Endpoint will provide a SPARQL-queryable interface to a set of RDF stored in a triple-store or on top of a relational database. SPARQL is in some ways an equivalent of the SQL database query language.
Triple store: A database management system for storage and retrieval of RDF data, usually providing a SPARQL Endpoint.
XML: A simple web format for the description of metadata, designed for easy transmission through HTTP links (and in HTML pages). OWL and RDF can be expressed in XML format.
Linked open data (LOD): A website that integrates a large number of publicly accessible databases in semantic form (http://linkeddata.org/).
OpenPHACTS: An EU-funded initiative to create an open pharmacological space for drug discovery using semantic technologies (http://www.openphacts.org/).
Semantic Web in Health Care and Life Sciences: A W3C consortium interest group whose mission is to develop, advocate for and support the use of Semantic Web technologies across healthcare, life sciences, clinical research and translational medicine.

Please cite this article in press as: D.J.. Wild, et al., Systems chemical biology and the Semantic Web: what they mean for the future of drug discovery research, Drug Discov Today (2012), doi:10.1016/j.drudis.2011.12.019

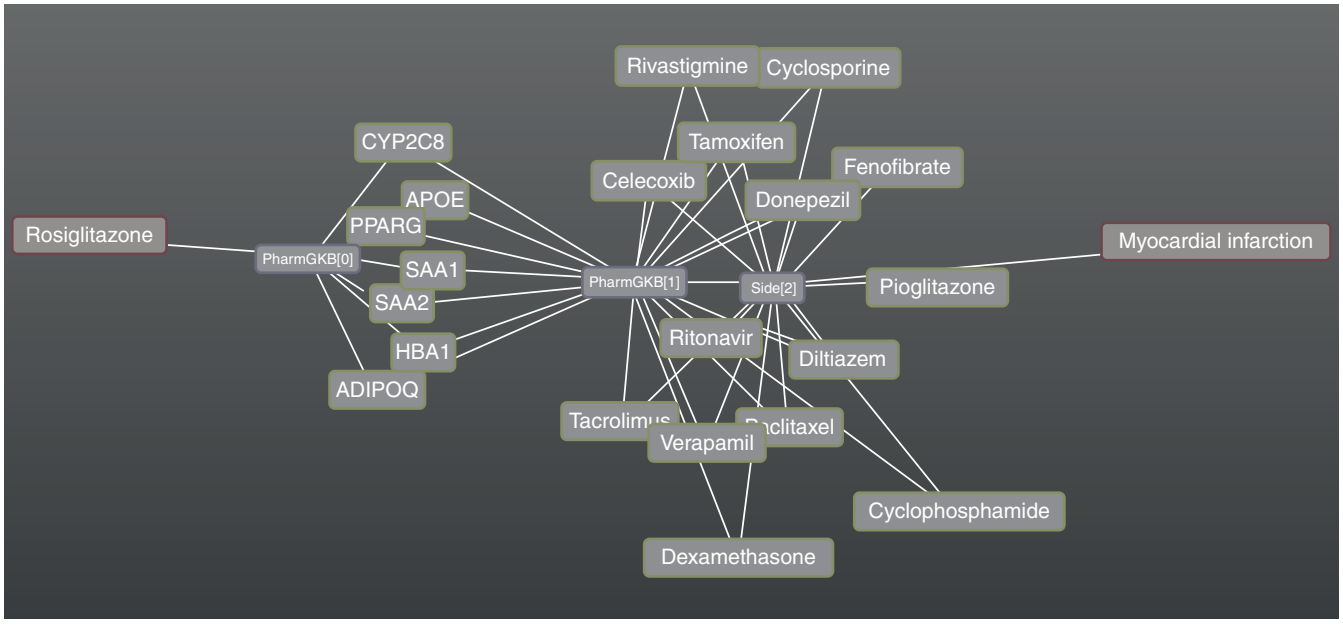

FIGURE 2

Constrained association search between myocardial infarction and rosiglitazone. Showing ranked paths up to three edges in length that (i) contain a gene and (ii) are ranked highly by KL-divergence showing literature support.

significant because interaction with this gene has been shown clinically to be associated with raised low-density lipoprotein (LDL) levels, and thus with myocardial infarction].

Languages for semantic integration

Traditionally, data integration in pharmaceutical research has been achieved by developing schema that map relational database tables together within a single database management system (a tortuous manual process), by *ad hoc* merging of data files to meet a particular immediate integration need or by employing external vendor solutions often for organization-wide data integration. However, no widely accessible, noncommercial technology has existed, until recently, for relatively straightforward integration of heterogeneous datasets between organizations and data silos. Three foundational components of what we now recognize as the Semantic Web, all recommendations developed by the World Wide Web Consortium, do now constitute a common core of technologies for such integration: RDF (resource description framework), OWL (Web ontology language) and SPARQL. RDF is a simple language, implementable in a variety of formats (e.g. XML), that enables the representation of pairs of entities and the relationships between them (RDF triples). Because of their simple nature, RDF triples are extremely flexible in representing any kind of relationship between chemical or biological entities. This direct representation of relationships is crucial for capturing semantics of data, which has been missing in the popular relational model. Furthermore, each RDF triple can be considered as two nodes of a network connected by an edge, and so in aggregate. A set of RDF triples describes a network of entities and relationships between them. OWL is used to represent ontologies, providing shared nomenclature or core vocabulary, and capture a richer model of the domain using subclass relationships and constraints. SPARQL is a language for querying RDF triples, similar conceptually to the relational SQL language but enabling powerful integrative searching (i.e. involving multiple, heterogeneous sets linked by OWL ontologies). Recently, and key to the practical implementation of Semantic Web based resources, triple stores are also available for fast access and searching of reasonably large sizes of data in RDF. These technologies are only now reaching a point of maturity where they are practically effective, as demonstrated below, leading to the Semantic Web unfortunately being rejected prematurely in some quarters as 'not up to the job' of practical integration.

Implementation of semantically integrated networks

There are now many successful demonstrations and deployments of Semantic Web technologies biological applications in the public sphere and in industry [10,13–16], although there are clearly significant research challenges ahead [17]. A 2007 *BMC Bioinformatics* paper made the case for the use of the Semantic Web in translational medical research, giving examples of its successful use [14]: this has since become the all-time most viewed article in the journal. Since 2005, linked open data (LOD) has become a significant force in open sharing of data, where large corpuses of Semantic Web data are published as bubbles in the LOD cloud, and integrated with other datasets using a set of standard protocols. Escalating importance is being given to the use of such methods in pharmaceutical research, as exemplified by the recent large EU grant given to the OpenPHACTS initiative specifically for the development of Semantic Web methods for drug discovery (http://www.openphacts.org) and the active and growing membership of the W3C Semantic Web in Health Care and Life Sciences (SWHCLS) Interest Group (http://www.w3.org/2001/sw/hcls/). Bio2RDF [18] and a rapidly growing biological component of the LOD cloud index approximately five billion triples of biological data. A subset of the LOD cloud relevant to drug discovery is the Linked Open Drug Data project [19]. A recent special issue of

Please cite this article in press as: D.J.. Wild, et al., Systems chemical biology and the Semantic Web: what they mean for the future of drug discovery research, Drug Discov Today (2012), doi:10.1016/j.drudis.2011.12.019

the *Journal of Cheminformatics* included papers describing the current uses of RDF in chemistry and cheminformatics [20] and demonstrated its use in a linked open drug data cloud [19], for providing open toxicology data [21], creating an open QSAR framework [22], in semantic text mining of journal articles [23] and in describing chemical structure and reaction data [24].

A variety of triple-store technologies are now available for practical implementation. Examples of demonstrated scalabilities of current systems are maintained on the W3C Consortium (http://www.w3.org/wiki/LargeTripleStores). Experiments demonstrate the ability to store and search tens of billions of RDF triples in real time, with current implementations easily able to store and search several hundred million triples on a small server implementation; however, it remains to be seen how well current systems scale in production environments.

RDF triple stores are made significantly more useful by the employment of ontologies, usually in the OWL language, which structure the allowed content of the RDF triple statements. Without an ontology, links between heterogeneous sets are mostly limited to 'same-as' statements (e.g. 'compound in set X is the same as drug in set Y') but with an ontology individual data fields can be mapped to higher level classes that could be described differently between sets (for example, to distinguish an IC_{50} from a percent inhibition). Attempts to create grand-scale ontologies (for instance to cover the whole of chemistry) have generally failed in the life sciences owing to problems of complexity and fuzzy boundaries with other disciplines. However, there are several well-used ontologies available that have a wide scope, including the Gene Ontology, and of particular note the recent Translational Medicine Ontology [25] provides a 'bridge' between diverse areas of medical research. Indeed, the most successful approaches to ontologies now seem to be to use existing ontologies where possible, and to build new ontologies for specific purposes to 'fill the gaps' with proper linking to existing ontologies. This is facilitated by open ontology portals, most notably for the life sciences, OBO (http://obofoundry.org/) and NCBO BioPortal (http://bioportal.bioontology.org/). The latter includes well over 250 ontologies at the time of writing. For industry use, it is also valid to develop internal ontologies closely mapped to internal data sources, but externally linked to other public ontologies to promote integration between internal and external data.

Integrative tools and algorithms

SPARQL searching

When drug discovery data is represented in RDF format in a triple store, the most basic kind of searching is to use a SPARQL Endpoint (i.e. an access point for searching the RDF data using the SPARQL language). This approximately maps using SQL to search a relational database, but it is much more powerful (especially if an OWL ontology is employed), because it permits searches that span heterogeneous sets in a single query. Inference is supported that can, for example, enable the use of a single general class of drug to be mapped into all its subclasses and variants. Demonstrated examples of this integrated searching include finding compounds with similar polypharmacology profiles to a known drug, suggesting multiple target inhibitors of mitogen-activated protein (MAP) kinase, and the identification of metabolic pathways with multiple gene associations that map to a given side effect [26]. SPARQL has significant limits, however; in particular, it is primarily a searching language, and thus does not provide access to advanced data mining algorithms. It is also a complex language for humans to learn, relegating its use to computing specialists rather than end-user scientists. Use of ontology-supported graphical query formulation tools such as Cuebee [27] now make it significantly easier to give a scientist access to the more powerful capabilities of the Semantic Web without the need to learn new languages.

End user tools

The first generation of generic Semantic Web tools have been designed primarily for browsing, visualizing and searching RDF data and are now being developed with more-powerful tools such as hypothesis testing. Topbraid (http://www.topquadrant.com) is a series of tools used for the integration of existing internal data sources into RDF-based formats, and for operating these integrated data. IO Informatics (http://www.io-informatics.com/) produces a suite of software designed specifically for life science users for integrating heterogenous data into RDF format, and then visualizing and searching the data in a variety of ways. Franz Allegrograph (http://www.franz.com/) combines a cloud-enabled RDF triple store (which it is claimed can handle over 300 billion triples) with tools for SPARQL searching of the data, visualization and limited reasoning capabilities. The RDFScape project (http://www.bioinformatics.org/rdfscape/wiki/) adds Semantic Web features to the free Cytoscape network visualization tool, enabling it to query, visualize and reason on ontologies represented in OWL or RDF within Cytoscape. Several free generic RDF graph visualization tools are available including RDF Gravity (http://semweb.salzburgresearch.at/apps/rdf-gravity/), SIMILE (http://simile.mit.edu/) and Triple Map (http://www.triplemap.com/).

Graph and network algorithms

When dealing with networks of data, it is useful to be able to apply graph theoretic algorithms such as breadth/depth first search and shortest path finding. Methods for handling graph theoretic querying [28] and semantic association finding [29] have been previously described and an algorithm for computing semantic associations has been recently applied to RDF drug discovery data [30]. This enables multiple shortest or otherwise meaningful paths between any two entities in a network to be identified. This has been implemented, along with the BioLDA algorithm, for literature mining [31] into a prototype association search tool that shows, for all pairs of entities, the network paths between them that have the highest level of literature support (as measured by KL-divergence). This has shown promise for suggesting gene associations that can account for the side effects of a drug or interactions with a disease. An example of this is given in Fig. 2, which shows the gene-based associations between one of the drugs from the thiazolinedione class, rosiglitazone (Avandia), and the side effect of myocardial infarction. This is significant because rosiglitazone has been found to have rare but serious cardiac side effects, and thus this provides a mechanism for suggesting potential gene actors in the process. This association-finding tool is now being implemented in a variety of systems at Pfizer.

Graph-theoretic analysis can also be used to predict new associations based on an existing graph. Eli Lilly has employed a tool called Chemogenomic Explorer, based on a previous profiling tool

Please cite this article in press as: D.J.. Wild, et al., Systems chemical biology and the Semantic Web: what they mean for the future of drug discovery research, Drug Discov Today (2012), doi:10.1016/j.drudis.2011.12.019

[32] that uses a rule-based inference engine to suggest potential disease associations for a new compound [33]. On the basis of manually curated rules, 'evidence paths' (chains of linked RDF statements) linking compounds and genes are created that then in aggregate represent a cluster of independent or semi-independent evidence linking a compound to a disease. A similar approach is taken in the HyQue tool [34], which enables diverse hypotheses to be evaluated through SPARQL queries and query evaluation rules. Such 'evidence clustering' could be important as a way of mitigating the risks of errors in data, as well as the known propensity for individual pieces of published medical research to be proved incorrect later on [35]. Probabilistic methods can also be applied to networks to provide a quantitative measure of association between any two entities based on the semantics and topology of the network. An ongoing project at Indiana University is investigating the use of such methods for the prediction of 'missing links' in networks, and also as a virtual screening method. Published methods, such as the SEA analysis [1], can also be used for this purpose.

RDF also offers the possibility of encoding data in scholarly publications, and then applying algorithms to mine the data. In recent work [31], a database of recent PubMed abstracts (those published during the past four years) was analyzed to identify Bioterms (i.e. terms that can be associated with chemical and biological entities that already exist in OWL ontologies: compounds, drugs, genes, among others). These Bioterms constitute an RDF association that can be mined. A latent Dirichlet allocation algorithm was used to identify latent topics in the PubMed literature based on these terms, which are then used to create a measure of distance between entities (via topics) known as KL-divergence.

Knowledge discovery processes and biomedical insights

'The proof of the pudding is in the eating' and, thus, a significant research endeavor must be applied to the evaluation of these new integrative tools and processes in real drug discovery efforts. Thus, as the 'horizontal' effort is needed to develop a wide-range of tools and algorithms, 'vertical' efforts are needed to discover how these new approaches can complement existing computational approaches (such as docking, QSAR, sequence similarity searching and ligand-based virtual screening) to accelerate the discovery of new drugs for specific therapeutic purposes, and to identify the key pieces of knowledge (biomedical insights) necessary for understanding disease processes. One can imagine a convergence of tools into 'integrative virtual screens' that fuse and balance a variety of virtual screening methods (including network-based methods) but also specific and perhaps even unique combinations of tools being applied for individual drug discovery problems. At the time of writing, little research has been carried out at this tier of the stack (i.e. how the tools can be mapped most effectively to real drug discovery problems), although work in related areas, such as data fusion and virtual screening, should help.

What this means, and what needs to be done

We believe the work described here constitutes a first step in realizing systems chemical biology (i.e. in providing a progressive framework for the development of integrated data resources, algorithms and tools, and knowledge discovery processes that combine systems chemical biology with more-traditional approaches). Data integration efforts using RDF are well underway in the public sector and internally in pharmaceutical organizations, but care does need to be taken so that these efforts are at least sufficiently aligned and that mapping entities between repositories is straightforward (for example, by maintaining PubChem identifiers for internal repository compounds that are also externally available). Further work needs to be done on technical details such as the representation of quantitative relationships (e.g. IC_{50} or similarity values) and provenance (i.e. source and history of data). Key to this are collaborative efforts such as W3 SWHCLS, the Pistoia Alliance and OpenPHACTS, along with publicly available open resources such as Chem2Bio2RDF and Bio2RDF. Also crucial is the separation of tools from data. Historically many tools and algorithms have been developed to work on specific datasets or repositories, and are not easily extensible to other sets. This must be addressed; in particular, tools should not be dependent on a specific ontological mapping in a set.

Addressing quality is an essential step, and one that is fraught with numerous complexities. Example questions that demonstrate this challenge are: is a PubChem BioAssay IC_{50} result comparable with one in ChEMBL or from an internal assay? Is an experimental result always more significant than a predicted result or an association extracted from a journal article? What happens when we get so many links between things so that we cannot separate the signal from the noise?

Ultimately, we are constrained by the data sources available: we have a choice which datasets to include or exclude and we have methods (such as provenance tracking) for keeping track of the history of a piece of data, but we are bound by the quality of which data we choose to use. Quality should thus be addressed primarily at the tool level, enabling users to select which datasets they are comfortable using and understanding the caveats in doing so. There is a case in some instances for using only limited datasets of known quality, and at other times using all available data. Ideally, it will be possible to make such quality determinations within tools and environments in a meaningful way. There is also a need for research into the use of multiple semi-independent evidence paths found in networks of data as a way of 'building consensus' that mitigates quality issues in specific data sources, which in turn makes a case for improved provenance tracking in Semantic Web implementations [36,37].

Once we can apply validated integrative tools and algorithms freely on data of our choosing from the full breadth of available information, the problem becomes one of what are the correct questions to ask of the data, and how to interpret and follow up on the results. This can only be done by the practical application of these methods in real drug discovery problems. Ideally, research efforts will occur in academia (and perhaps in precompetitive collaboration between industry and academia) so that effective integrative methodologies for drug repurposing can be publicly validated.

In the near future, emerging patient-level datasets, including those derived from electronic medical records (EMRs), next-generation sequencing and/or genome-wide association search (GWAS) tools and metagenomics will massively increase the available data and will thus introduce issues of scale that will need to be addressed at the triple-store and algorithm levels. However, these

Please cite this article in press as: D.J.. Wild, et al., Systems chemical biology and the Semantic Web: what they mean for the future of drug discovery research, Drug Discov Today (2012), doi:10.1016/j.drudis.2011.12.019

sets will also provide the opportunity to gain understanding of how individuals will respond to drugs, rather than the body as a generic entity. Research is needed at the interface of these datasets with existing chemical, biological and pharmacological sets, to provide a public corpus of data that in aggregate will form a biomedical map that bridges the molecular and clinical spectra – 'from molecule to human'.

If we assume that successful discovery on new, safe, effective drugs is going to require that we step beyond the 'lock and key' model of drug and protein targets to understand the much greater complexities of how drugs interact with the body, realizing the emerging disciplines of chemogenomics and systems chemical biology through enabling integrative technologies (such as the Semantic Web) is going to be a crucial foundation to success in 21st century drug discovery. Promising efforts are already underway, but there is much more basic research and industry–academia collaboration required to accelerate progress in these fields.

References

1 Keiser, M.J. *et al.* (2009) Predicting new molecular targets for known drugs. *Nature* 462, 175–181
2 Xie, L. *et al.* (2011) Structure-based systems biology for analyzing off-target binding. *Curr. Opin. Struct. Biol.* 21, 189–199
3 Chang, R.L. *et al.* (2010) Drug off-target effects predicted using structural analysis in the context of a metabolic network model. *PLoS Comput. Biol.* 6, 9
4 Kinnings, S.L. *et al.* (2010) Drug discovery using chemical systems biology: repositioning the safe medicine comtan to treat multi-drug and extensively drug resistant tuberculosis. *PLoS Comput. Biol.* 5, 7
5 Metz, J.T. *et al.* (2011) Navigating the kinome. *Nat. Chem. Biol.* 7, 200–202
6 Chen, B. *et al.* (2009) PubChem as a source of polypharmacology. *J. Chem. Inf. Model.* 49, 2044–2055
7 Campillos, M. *et al.* (2008) Drug target identification using side-effect similarity. *Science* 321, 263–266
8 Oprea, T.I. *et al.* (2007) Systems chemical biology. *Nat. Chem. Biol.* 3, 447–450
9 Taboureau, O. *et al.* (2011) ChemProt: a disease chemical biology database. *Nucleic Acids Res.* 39, 367–372
10 Wild, D.J. (2009) Mining large heterogeneous datasets in drug discovery. *Expert Opin. Drug Discov.* 4, 995–1004
11 Slater, T. *et al.* (2008) Beyond data integration. *Drug Discov. Today* 13, 584–589
12 Guha, R. *et al.* (2010) Advances in cheminformatics methodologies and infrastructure to support the data mining of large, heterogeneous chemical datasets. *Curr. Comput. Aided Drug Des.* 6, 50–67
13 Cheung, K.H. *et al.* (2011) Semantic Web for health care and life sciences: a review of the state of the art. *Brief. Bioinform.* 10, 111–113
14 Ruttenberg, A. *et al.* (2007) Advancing translational research with the Semantic Web. *BMC Bioinformatics* 8, 1–16
15 Chen, H. and Xie, G. (2010) The use of web ontology languages and other semantic web tools in drug discovery. *Expert Opin. Drug Discov.* 5, 413–423
16 Choi, J. *et al.* (2010) A Semantic Web ontology for small molecules and their biological targets. *J. Chem. Inf. Model.* 50, 732–741
17 Dumontier, R. (2010) Building an effective Semantic Web for health care and the life sciences. *Semantic Web – Interoperability Usability Applicability* 1 Special Issue: Vision Statements
18 Belleau, F. *et al.* (2008) Bio2RDF: towards a mashup to build bioinformatics knowledge systems. *J. Biomed. Inform.* 41, 706–716
19 Samwald, M. *et al.* (2011) Linked open drug data for pharmaceutical research and development. *J. Cheminf.* 3, 19
20 Willighagen, E.L. and Brändle, M.P. (2011) Resource description framework technologies in chemistry. *J. Cheminf.* 3, 15
21 Jeliazkova, N. and Jeliazkov, V. (2011) AMBIT RESTful web services: an implementation of the OpenTox application programming interface. *J. Cheminf.* 3, 18
22 Chepelev, L.L. and Dumontier, M. (2011) Semantic Web integration of cheminformatics resources with the SADI framework. *J. Cheminf.* 3, 16
23 Hawizy, L. *et al.* (2011) ChemicalTagger: a tool for semantic text-mining in chemistry. *J. Cheminf.* 3, 17
24 Chepelev, L.L. and Dumontier, M. (2011) Chemical entity semantic specification: knowledge representation for efficient semantic cheminformatics and facile data integration. *J. Cheminf.* 3, 20
25 Dumontier, M. *et al.* (2011) The translational medicine ontology: driving personalized medicine by bridging the gap from bedside to bench. *J. Biomed. Semantics* http://www.ncbi.nlm.nih.gov/pubmed/21624155, 2(Suppl 2):S1.
26 Chen, B. *et al.* (2010) Chem2Bio2RDF: a semantic framework for linking and data mining chemogenomic and systems chemical biology data. *BMC Bioinformatics* 11, 255
27 Mendes, P.N. *et al.* (2008) TcruziKB: enabling complex queries for genomic data exploration. *International Conference on Semantic Computing* pp. 432–439
28 Anyanwu, K. *et al.* (2007) SPARQ2L: towards support for subgraph extraction queries in RDF databases. In *Proceedings of the 16th International Conference on World Wide Web* ACM
29 Anyanwu, K. and Sheth, A. (2003) ρ-Queries: enabling querying for semantic associations on the semantic web. In *Proceedings of the 12th International Conference on World Wide Web* ACM
30 He, B. *et al.* (2011) Mining relational paths in relational biomedical data. *PLoS One* 6, e27506
31 Wang, H. *et al.* (2011) Finding complex biological relationships in recent PubMed articles using Bio-LDA. *PLoS One* 6, e17243
32 Zhu, Q. *et al.* (2010) WENDI: a tool for finding non-obvious relationships between compounds and biological properties, genes, diseases and scholarly publications. *J. Cheminf.* 2, 6
33 Zhu, Q. *et al.* (2011) Semantic inference using chemogenomics data for drug discovery. *BMC Bioinformatics* 12, 256
34 Callahan, A. *et al.* (2011) HyQue: evaluating hypotheses using Semantic Web technologies. *J. Biomed. Semantics* 2(Suppl 2): S3
35 Ioannidis, J. (2005) Why most published research findings are false. *PLoS Med.* 2, e124
36 Sahoo, S.S. *et al.* (2008) Semantic provenance for eScience: managing the deluge of scientific data. *Internet Comput.* 12, 46–54
37 Sahoo, S.S. *et al.* (2010) Provenance Context Entitiy (PaCE): scalable provenance tracking for scientific RDF data. *Lect. Notes Comput. Sci.* 6187, 461–470

Please cite this article in press as: D.J.. Wild, et al., Systems chemical biology and the Semantic Web: what they mean for the future of drug discovery research, Drug Discov Today (2012), doi:10.1016/j.drudis.2011.12.019