ELSEVIER

Scientific competency questions as the basis for semantically enriched open pharmacological space development

Kamal Azzaoui[1], Edgar Jacoby[14], Stefan Senger[2], Emiliano Cuadrado Rodríguez[3], Mabel Loza[3], Barbara Zdrazil[4], Marta Pinto[4], Antony J. Williams[5], Victor de la Torre[6], Jordi Mestres[7], Manuel Pastor[7], Olivier Taboureau[8], Matthias Rarey[9], Christine Chichester[10], Steve Pettifer[11], Niklas Blomberg[12,a], Lee Harland[13], Bryn Williams-Jones[13] and Gerhard F. Ecker[4]

[1] Novartis Institutes for BioMedical Research, Novartis Pharma AG, Forum 1 Novartis Campus, CH-4056 Basel, Switzerland
[2] GlaxoSmithKline, Medicines Research Centre, Stevenage SG1 2NY, UK
[3] Grupo BioFarma-USEF, Departamento de Farmacología, Facultad de Farmacia, Campus Universitario Sur s/n, 15782 Santiago de Compostela, Spain
[4] University of Vienna, Department of Medicinal Chemistry, Pharmacoinformatics Research Group, Althanstrasse 14, 1090 Wien, Austria
[5] Royal Society of Chemistry, 904 Tamaras Circle, Wake Forest, NC 27587, USA
[6] Structural Computational Biology and National Bioinformatic Institute Unit, Structural Biology and Biocomputing Programme, Spanish National Cancer Research Centre (CNIO), C/ Melchor Fernández Almagro 3, Madrid E-28029, Spain
[7] Chemogenomics Laboratory, Research Programme on Biomedical Informatics, IMIM—Hospital del Mar Research Institute and Universitat Pompeu Fabra, Parc de Recerca Biomèdica, Doctor Aiguader 88, 08003 Barcelona, Catalonia, Spain
[8] Technical University of Denmark, Department of Systems Biology, Kemitorvet, Building 208, 2800 Lyngby, Denmark
[9] Center for Bioinformatics, University of Hamburg, Bundesstraße 43, 20146 Hamburg, Germany
[10] Swiss Institute of Bioinformatics, CALIPHO Group, CMU – Rue Michel-Servet 1, 1211 Geneva 4, Switzerland
[11] School of Computer Science, The University of Manchester, Oxford Road, Manchester M13 9PL, UK
[12] AstraZeneca R&D Mölndal, SE-431 83 Mölndal, Sweden
[13] Connected Discovery Ltd, 27 Old Gloucester Street, London WC1N 3AX, UK
[14] Janssen Research & Development, Turnhoutseweg 30, B-2340 Beerse, Belgium

Molecular information systems play an important part in modern data-driven drug discovery. They do not only support decision making but also enable new discoveries via association and inference. In this review, we outline the scientific requirements identified by the Innovative Medicines Initiative (IMI) Open PHACTS consortium for the design of an open pharmacological space (OPS) information system. The focus of this work is the integration of compound–target–pathway–disease/phenotype data for public and industrial drug discovery research. Typical scientific competency questions provided by the consortium members will be analyzed based on the underlying data concepts and associations needed to answer the questions. Publicly available data sources used to target these questions as well as the need for and potential of semantic web-based technology will be presented.

Introduction

Drug discovery is a data-driven process [1]. The amount and diversity of drug discovery data in the omics- and high-throughput-driven paradigms has significantly grown to the point where current relational data models have reached their performance limits in terms of technical and scientific capabilities [2]. In addition to the need for data integration, it is recognized that providing capabilities for semantic inference is a key challenge and offers a wealth of opportunities. Such a semantic molecular information system was pioneered by the Wild group at Indiana University

Corresponding author:. Ecker, G.F. (gerhard.f.ecker@univie.ac.at)
[a] Present address: ELIXIR, EMBL-European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridgeshire, CB10 1SD, UK.

1359-6446/06/$ - see front matter http://dx.doi.org/10.1016/j.drudis.2013.05.008

with the Chem2Bio2RDF system [3–6], which is based on the Bio2RDF knowledge system provided earlier by Belleau at Laval University, Montreal [7]. The Linking Open Drug Data (LODD) project [8] is a comparable project within the World Wide Web Consortium (W3C) [9] healthcare and life science interest group. Recognizing the challenges and opportunities, the European Union (EU) and European Federation of Pharmaceutical Industries and Associations (EFPIA) decided to develop the Innovative Medicines Initiative (IMI) joint undertaking, and through this [10] the Open Pharmacological Concepts Triple Store (Open PHACTS) consortium [11]. The Open PHACTS project brings together academic and pharmaceutical partners to design and implement a publicly available open pharmacological space. The project is driven by scientific questions and use cases of various complexity that apply to real-world drugs that, for the purpose of the Open PHACTS project, we term 'scientific competency questions'. By focusing on standard use cases for drug discovery, the importance of the competency questions is reflected in their general nature rather than in the specific questions *per se*. The seemingly straightforward questions provide model scenarios that require careful association (mapping) of multiple heterogeneous data across diverse public domain databases. Core underlying data concepts for this medicinal-chemistry-driven platform are 'compound', 'target', 'pathway' and 'disease/phenotype', all of them are relevant for the new fields of chemogenomics [12] and systems chemical biology [13,14].

As will become apparent from the analysis of the scientific competency questions, the Open PHACTS system will build upon the ideas of the Chem2Bio2RDF, Bio2RDF and LODD systems, to address drug discovery research questions specifically. A key feature of the Open PHACTS discovery platform is the openness for new data additions that could include data from text mining of scientific publications as well as opportunities for integration with proprietary or commercial data sources. Another key aspect is the development of novel visualization tools that facilitate the navigation and knowledge extraction from all integrated data made available. It is intended for the end-user tools not only to show how it is possible to build relevant end-user applications on top of the Open PHACTS platform but also to provide the bench scientist with immediate value. Of course, all electronic data should be used with caution and scientists need to be aware of its origin, provenance and reliability. Thus, the goal in the Open PHACTS project is not simply to integrate and query multiple databases but to provide a mechanism to understand how these results were obtained with attribution and provenance of individual data points [15].

A scientific competency question approach

When setting up an open, innovative, data integration and knowledge extraction platform, the first question arising is which out of the more than 1000 open access databases [16] need to be integrated. This obviously depends on the type of queries the system should allow the user to perform. Driven by the fact that in the first instance the target audience will be bench scientists working in drug discovery and development, the Open PHACTS consortium develops a set of core use-cases to guide the project and to assist in prioritizing the data sources selected for integration. Thus, the original 22 Open PHACTS partners (eight EFPIA companies, 12 academic institutions, two SMEs; the project is continuously growing and currently comprises 28 partners now already) were asked to provide 'business' questions that they believed would enable progress in their specific research activities and drug discovery in general. Although most of the questions provided are not challenging *per se*, answering them requires input from multiple data sources hence needs in-depth knowledge of the data models for a large set of systems. Thus, these represent a challenge to the current information systems in use. Analysis of these questions provides valuable information for the design of the graphical user interface and guides the selection of data sources. In total 83 questions were collected in this approach, representing an effective survey of user needs and information priorities for preclinical drug discovery research in pharmaceutical companies and academic institutions. The analysis of the questions followed a structured approach with input and critique from project partner representatives. Because the results represent a clear and prioritized set of requirements and use cases for drug discovery research projects, we believe they will have significant impact and use with regard to knowledge management and systems design. The 83 questions were then grouped and prioritized using a point-based voting system where each partner had one vote to rank the importance of each question as high, medium or low. It is worth noting that there were considerable differences in the rankings between academic institutions from different domains (e.g. University of Vienna and Leiden University Medical Center), but almost perfect correlation between EFPIA companies and academic institutions from the medicinal chemistry domain (e.g. University of Vienna and AstraZeneca).

Prioritizing the 83 collected research questions and subsequent analysis of the top 20 of these led to a deeper insight into the actual information needs of researchers (in the pharma industry as well as in academia and biotech) regarding data associations. This analysis was carried out by extracting the key concepts (compound, target, pathway, disease), as well as crucial mappings between concepts implied in each question needed to start a conventional data search. The application of this procedure resulted in three main groups of 29 target/protein-related, 21 compound-related and 21 either disease- or pathway-related questions. Minor groups from this analysis comprise gene/gene family (six questions), substructure (five), protein family (three), RNA (one) and assay (one).

To complete the above analysis and have an overview of what requesters expect in highly ranked questions, we included for each question the keywords compound, target, pathway and disease, in each category of prioritization (Fig. 1). It immediately becomes obvious that the concept 'compound–target' is predominantly found in the highly prioritized questions. All questions contain the concept 'compound', and 16 out of the top 20 questions also refer to 'target'. The top 20 questions were then grouped in two clusters according to the complexity of information requested about compound–target or compound–target–disease/pathway. The questions and their clusters are summarized in Table 1, and are herein analyzed in terms of required data concepts, required data associations and potential public data sources needed.

Cluster I: compound–target

This first cluster contains 11 questions centered on basic pharmacology. These types of questions are usually asked in early drug

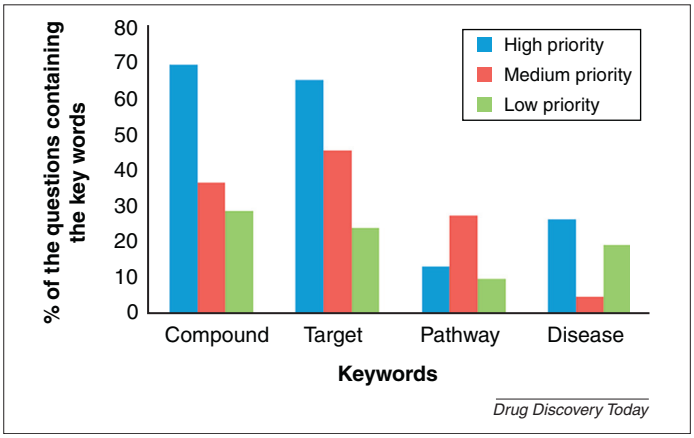

FIGURE 1

The percentage of the 83 questions with respect to the key words: compound, target, pathway, disease; categorized by priorities. It clearly demonstrates that the concepts 'compound' and 'target' are the most dominant.

discovery phases at the hit- or lead-finding stages. Generally, the user wishes to find out more about known interactions between a compound or a set of compounds and a defined primary target and/or other targets. The request can be expanded toward a target family and/or the same target in different species. By defining an activity threshold, the user expects a list of compounds that can be useful for direct screening or to provide input for compound library design targeting a new enzyme, for instance an oxidoreductase (Q1, Table 1).

Although looking deceptively simple, this question is challenging because it requires checking each compound for activity against the >3000 oxidoreductases in both species concerned. The question also illustrates the common approach of 'target families' and hence highlights the need for a well maintained target classification system. A manual search took two scientists three days to retrieve the respective list of compounds. The prototype released internally six months after the start of the project was able to perform this query within a few seconds. Regarding information on multiple targets or ADMET effects (Q2 and Q3, Table 1), questions are motivated by the need to understand the mechanism of possible side-effects of drug candidates and drugs [17]. These are typically asked when lead and drug candidates are assessed to decide whether or not to progress them for further development. The metabolism/toxicity-related issues could at least in part be answered by providing predicted secondary pharmacology data for a given compound of interest. The predicted data could be from an interrogation of existing bioactivity data or based on predictive *in silico* models as applied, for example, in the Chemotargets software for predicting the off-target pharmacology of small molecules [18], the SEA-approach [19] or the PASS algorithm [20].

In some cases the user might be more interested in a profile of activity rather than a single activity or interested in similar compounds with a similar activity profile (Q4 and Q9, Table 1). To answer this type of question one has to provide a defined bioactivity interaction profile, and then search for compounds that share similarity in terms of their bioactivity profile. This use case addresses a typical lead-finding strategy searching for compounds with different chemotypes but similar bioactivity profiles that are also expected to share activity on new targets [21]. In comparison with previous questions the complexity increases remarkably when the definition of the query requires substructure matching capabilities and similarity searching (Q5 and Q10, Table 1). To query and answer the specific questions above might well go beyond the capabilities of a simple, easy-to-use graphical user interface (GUI). Thus, a set of end-user applications, so called example applications (eApps), are developed within the consortium on top of a robust Open PHACTS services application programming interface (API) and will be discussed later. These eApps are proof-of-concept studies to demonstrate the capability of the Open PHACTS discovery platform and API to enable effective services built on top of it. They comprise tools with advanced querying capabilities as well as scientific applications addressing specialized needs.

To answer questions such as Q6 and Q8 (Table 1), a certain level of granularity in the gene classification systems is needed. In fact, the answer to Q6 returns a list of chemical compounds with structures that are active against protein kinase C (PKC)α or all other members of the PKC subfamily of kinases. The question is a typical homology-based hit-finding strategy applicable for projects where a large knowledge base exists [22]. The answer to Q8, where specific interactions are targeted such protein–protein interactions (PPIs), needs a specific database for such request. Finally, an answer to Q11 requires late-phase development clinical data. The question, motivated by the desire to assess clinical compounds, can be part of both clusters because the clinical data can be linked to a disease study. For the specified list of clinical compounds, the system should provide the available clinical data in addition to the bioactivity data. However, although incredibly useful, this is beyond the reach of the Open PHACTS project in its current definition. Nevertheless, the integration capabilities offered by the Open PHACTS discovery platform definitely will allow extensions toward translational data.

Cluster II: compound–target–disease/pathway

The second cluster contains nine questions that deal with the previous concepts of compound–target relationships but, in addition, information about more-complex pharmacology in context of pathways and diseases is needed. The information requested in most of the cases needs references associated with it (patents, journal articles). These types of questions are usually asked during the lead optimization phase or proof-of-concept studies. The first question in the cluster (Q12, Table 1) is motivated by the fact that patents constitute an indispensable information source for chemical compounds and biological targets. The complexity of the question is twofold: first, all patents have to be retrieved that relate to the compound and the disease of interest; in a second step, the targets need to be extracted from the patent claims. However, the retrieval and/or tagging of text [23], as well as the recognition of targets within patents [24], is extremely challenging and subject to active research.

Question 13 (Table 1) is an aggregate of different questions seen already; owing to its composite nature the complexity is extremely high. First, as in Q3, the bioactivity of a compound for a specific target needs to be established. As in Q11, the compound is in preclinical or clinical phase and is an advanced compound that should have public data available. The link to the relevant

TABLE 1

The top 20 research questions

Question number	Question
Cluster I	
Q1	Give me all oxidoreductase inhibitors active <100 nM in human and mouse
Q2	Given compound X, what is its predicted secondary pharmacology? What are the on- and off-target safety concerns for a compound? What is the evidence and how reliable is that evidence (journal impact factor, KOL) for findings associated with a compound?
Q3	Given a target, find me all actives against that target. Find/predict polypharmacology of actives. Determine ADMET profile of actives
Q4	For a given interaction profile – give me similar compounds
Q5	The current Factor Xa lead series is characterized by substructure X. Retrieve all bioactivity data in serine protease assays for molecules that contain substructure X
Q6	A project is considering protein kinase C alpha (PRKCA) as a target. What are all the compounds known to modulate the target directly? What are the compounds that could modulate the target directly? I.e. return all compounds active in assays where the resolution is at least at the level of the target family (i.e. PKC) from structured assay databases and the literature
Q7	Give me all active compounds on a given target with the relevant assay data
Q8	Identify all known protein–protein interaction inhibitors
Q9	For a given compound, give me the interaction profile with targets
Q10	For a given compound, summarize all 'similar compounds' and their activities
Q11	Retrieve all experimental and clinical data for a given list of compounds defined by their chemical structure (with options to match stereochemistry or not)
Cluster II	
Q12	For my given compound, which targets have been patented in the context of Alzheimer's disease?
Q13	Which ligands have been described for a particular target associated with transthyretin-related amyloidosis, what is their affinity for that target and how far are they advanced into preclinical/clinical phases, with links to publications/patents describing these interactions?
Q14	Target druggability: compounds directed against target X have been tested in which indications? Which new targets have appeared recently in the patent literature for a disease? Has the target been screened against in AZ before? What information on *in vitro* or *in vivo* screens has already been performed on a compound?
Q15	Which chemical series have been shown to be active against target X? Which new targets have been associated with disease Y? Which companies are working on target X or disease Y?
Q16	Which compounds are known to be activators of targets that relate to Parkinson's disease or Alzheimer's disease
Q17	For my specific target, which active compounds have been reported in the literature? What is also known about upstream and downstream targets?
Q18	Compounds that agonize targets in pathway X assayed in only functional assays with a potency <1 μM
Q19	Give me the compound(s) that hit most specifically the multiple targets in a given pathway (disease)
Q20	For a given disease/indication, give me all targets in the pathway and all active compounds hitting them

publications and patents renders it similar to Q12 and Q17. The compound–publication and target–publication associations are needed. In some questions such as Q14 there is a need to compare the resulting data with proprietary in-house data, which poses a technical challenge because the publicly available Open PHACTS discovery platform needs to be able to integrate proprietary data. This imposes another level of complexity because it requires secure access. In addition, the whole issue of licensing needs to be addressed, which is extremely demanding when it comes to mixing public and private data into one platform.

Question 15 can be related to Q5 and Q10. However, what is new in regard to the related questions is the specification of the concept of a chemical series; hit compounds would need to be clustered around the parent series scaffold. Although various computational definitions for this task exist, there still remains the challenge of agreeing on the method to be used in the first instance. Open PHACTS is thus also actively working on providing standards agreed and widely adopted by the community, also including the pharmaceutical industry. Finally, the question of new targets associated with a disease and the competitive landscape around a target or disease are factors interrogated in Q15. Questions 17 and 18 are similar to Q3. More specifically, they require the knowledge of the pathway(s) where the specific target is representing a node. The capability to extract the upstream and downstream interaction partners from the pathway map adds an extra degree of complexity. Answering Q19 and Q20 imposes two levels of complexity. First, the compounds that hit the known targets in a given canonical disease pathway would need to be identified. In a second step, the most-specific compounds would need to be extracted. Both questions are relevant in the context of target and lead identification in the newer biology-driven drug discovery paradigm [25,26].

Data association and/or data sources

The analysis of the scientific competency questions demonstrates that the data and associations between the concepts 'chemical compound', 'molecular target', 'biological pathway' and 'disease' are essential to address the questions. One can visualize the complexity of the data associations in a network fashion (Fig. 2). With this network, we intend to summarize the data and associations needed to answer just the top 20 prioritized questions and thus to lay the foundation for selecting public databases that provide the respective information. In this section we will go through the details of the data needed and the public databases that provide such data (details and references about the data sources are listed in Table 2).

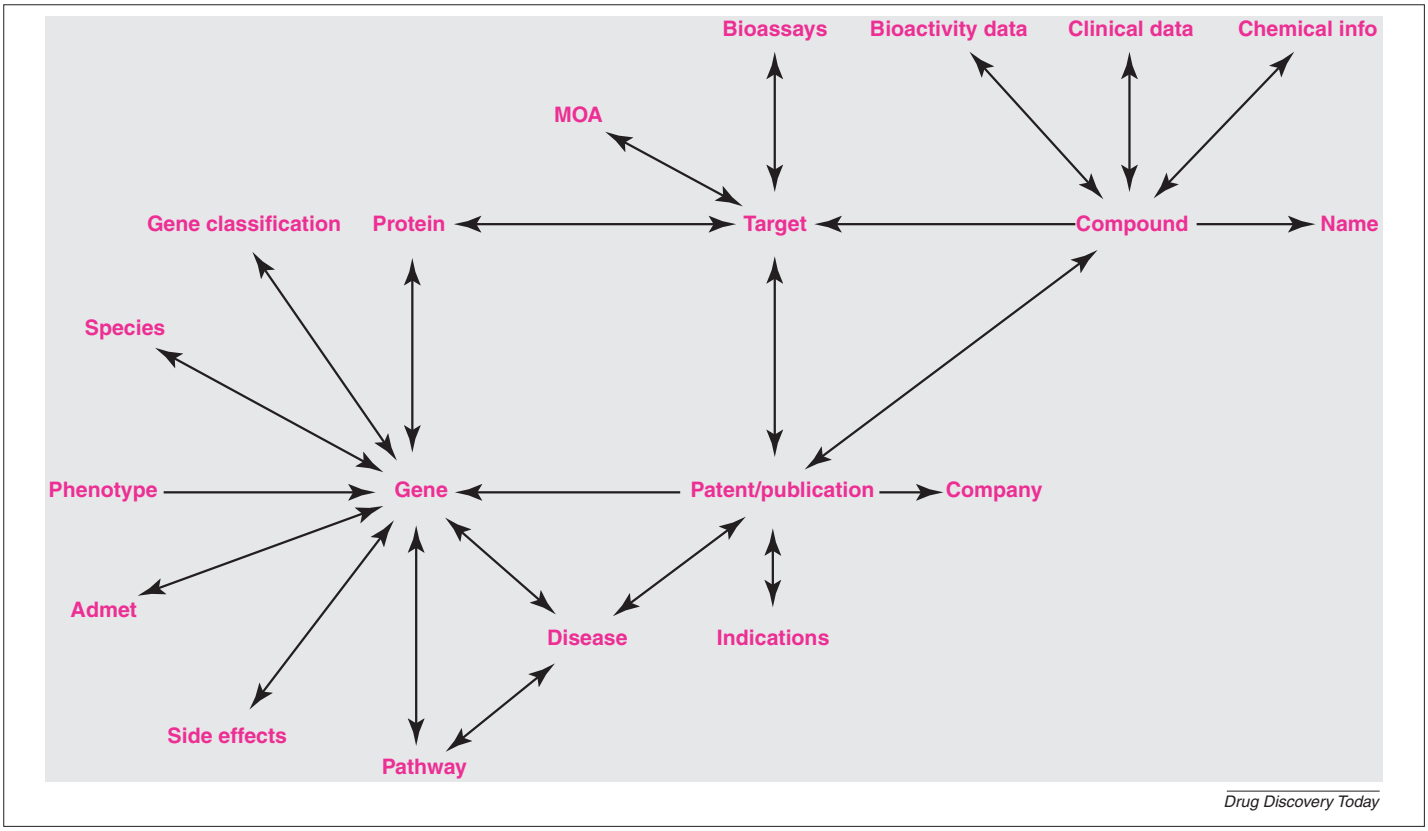

FIGURE 2

Network of data associations needed to answer the top-ranked scientific competency questions. The network reflects a cartoon that summarizes the data associations that are needed to target the top 20 research questions.

Chemistry node

The information about drug names and the required search mode depends strongly on the use of correct chemical names such as systematic names, trade names or synonyms. Such dictionaries of chemical name–structure associations are available from a number of chemical databases including ChEBI, DrugBank and ChemSpider [27]. In addition, chemical structure information is needed for defining queries such as substructure matching or similarity searches. This point stresses the need for data quality in general and especially on the exactness of representation of chemical structures, including stereochemistry, tautomers and protonation states. The quality of data in public domain databases has been discussed in a number of publications and highlights the need for a consistent and publicly described framework for normalization [28,29]. Within the Open PHACTS project those chemical-related queries will be provided via an interface to the ChemSpider molecular information system. For the compound–bioactivity association several highly popular databases will be integrated in the Open PHACTS discovery platform. In the first instance, those are ChEMBL, Chebi and DrugBank, which comprise large-scale public sources for the compound and bioactivity data. The bioactivity databases provide data at the level of primary activity at a single concentration of compound or data from dose–response-based experiments such as IC_{50} and/or EC_{50}, K_i or K_b. Hence, there is a need for the system to handle quantitative data within the semantic interoperability framework.

Biology nodes

It is important to recognize that a target in a given assay can be of heterogeneous nature, including proteins, cells or even whole organisms. To illustrate this, a search performed in ChEMBL for propafenone, a class 1C antiarrhythmic agent, revealed target name instances such as *Ratus norwegicus*, *Plasmodium falciparum*, CYP 450 2D6 and CCRF CEM 1000. It is thus necessary to represent the target–protein–gene associations in full detail. The gene–phenotype association can be found in the OMIM or GO data systems. More specifically, gene–side-effect associations can be found in specific literature on safety profiling [17]. The IUPHAR database provides structured pharmacological data on ion channels, G-protein-coupled receptors (GPCRs) and nuclear receptors. The SIDER database at EMBL contains information on marketed medicines and their recorded adverse drug reactions. However, it should be mentioned that the definition of secondary pharmacological data is 'fuzzy' given that a compound can have more than one primary target in the view of its polypharmacology, which finally defines its pharmacodynamic *in vivo* profile [30].

Protein–pathway association data are provided by the WikiPathways and Reactome data repositories. Also the GO classification of genes will be useful in this context. Gene–disease association data are part of the OMIM and Diseaseome datasets. This nicely illustrates the need for granularity in the gene classification systems. Because the previously explored target families such target ontologies have been elaborated [31]. The key difficulty is to have

TABLE 2

Summary view of data sources, content and data concepts that are of interest for an open pharmacological space, recorded in March 2012. Emphasized is the open and free public access of the data sources

Database name	Internet resource URL/source	Specification of data and information available
ADME-AP	http://bioinf.xmu.edu.cn/databases/pathways/ADMEAP.htm	ADME-AP provides comprehensive information about all classes of ADME-associated proteins described in the literature including physiological function of each protein, pharmacokinetic effect, ADME classification, direction and driving force of disposition, location and tissue distribution, substrates, synonyms, gene name and protein availability in other species. Cross-links to other databases are also provided to facilitate the access of information about the sequence, 3D structure, function, polymorphisms, genetic disorders, nomenclature, ligand binding properties and related literatures of each protein. ADME-AP currently contains entries for 321 proteins and 964 substrates
Cancer Central Clinical Database (C3D)	https://cabig.nci.nih.gov/community/tools/c3d	C3D is a clinical trials data management system that collects clinical trial data using standard case report forms based on common data elements. It utilizes security procedures to protect patient confidentiality and maintain an audit trail as required by FDA regulations
ChEBI	http://www.ebi.ac.uk/chebi/	ChEBI (Chemical Entities of Biological Interest) is a freely available dictionary of chemical compounds, with IUPAC and NC-IUBMB endorsed terminology. Currently three data sources have been incorporated into ChEBI, namely KEGG Ligand, IntEnz and Chemical Ontology
ChemBank	http://chembank.broadinstitute.org/	ChemBank is a freely available chemoinformatics database. The data are derived from small molecules and small-molecule screens and resources for studying these data. It was developed through a collaboration with the Chemical Biology Program and Platform at the Broad Institute of Harvard and MIT
ChEMBL	https://www.ebi.ac.uk/chembl/	ChEMBL is a database of bioactive drug-like small molecules. This database also contains 2D structures, calculated properties (e.g. log *P*, molecular weight, Lipinski parameters) and abstracted bioactivities (e.g. binding constants, pharmacology and ADMET data)
ChemSpider	http://www.chemspider.com/	ChemSpider is a free chemical structure database from the Royal Society of Chemistry providing fast text and structure/substructure search access to over 26 million structures from over 400 data sources
ClinicalTrials.gov	http://clinicaltrials.gov/	ClinicalTrials.gov is a registry and results database of federally and privately supported clinical trials conducted in the USA and around the world. It gives information about a trial's purpose, who may participate, locations and phone numbers, among others
Diseases Database	http://www.diseasesdatabase.com/	The Diseases Database is a database that underlies a free website that provides information about the relationships between medical conditions, symptoms and medications
DISEASOME	http://diseasome.kobic.re.kr/	DISEASOME provides the genes that are associated with diseases and potentially deleterious SNPs among the genes that are strongly associated with specific diseases and clinical phenotypes. Currently, it contains 14,674 records on genetic variation and 109,715 records on genes related to human diseases
DrugBank	http://www.drugbank.ca/	The DrugBank database combines detailed drug (i.e. chemical, pharmacological and pharmaceutical) data with comprehensive drug target (i.e. sequence, structure and pathway) information. The database (version 3.0) contains 6708 drug entries. 4229 nonredundant protein (i.e. drug target/enzyme/transporter/carrier) sequences are linked to these drug entries
GenBank	http://www.ncbi.nlm.nih.gov/genbank/	GenBank is the NIH genetic sequence database, an annotated collection of all publicly available DNA
GO Database	http://www.geneontology.org/	The GO (Gene Ontology) database is a relational database comprising the GO ontologies and the annotations of genes and gene products to terms in the GO. The advantage of housing the ontologies and annotations in a single database is that powerful queries can be performed over annotations using the ontology
HMDB	http://www.hmdb.ca/	The Human Metabolome Database (HMDB) is a freely available electronic database containing detailed information about small molecule metabolites found in the human body. The database (version 2.5) contains over 7900 metabolite entries. Additionally, approximately 7200 protein (and DNA) sequences are linked to these metabolite entries
IntAct	http://www.ebi.ac.uk/intact/	IntAct provides a freely available, open source database system and analysis tools for protein interaction data. All interactions are derived from literature curation or direct user submissions and are freely available
InterPro	http://www.ebi.ac.uk/interpro/	InterPro is an integrated database of predictive protein signatures used for the classification and automatic annotation of proteins and genomes. It classifies sequences at superfamily, family and subfamily levels, predicting the occurrence of functional domains, repeats and important sites. InterPro adds in-depth annotation, including GO terms, to the protein signatures

TABLE 2 (*Continued*)

Database name	Internet resource URL/source	Specification of data and information available
IUPHAR Database	http://www.iuphar-db.org/	The IUPHAR Database is the official database of the IUPHAR Committee on Receptor Nomenclature and Drug Classification. It incorporates detailed pharmacological, functional and pathophysiological information on G-protein-coupled receptors, voltage-gated ion channels, ligand-gated ion channels and nuclear hormone receptors
KDBI	http://bidd.nus.edu.sg/group/kdbi/kdbi.asp	Kinetic data of biomolecular interaction (KDBI) is a collection of experimentally determined kinetic data of protein–protein, protein–RNA, protein–DNA, protein–ligand, RNA–ligand, DNA–ligand-binding or reaction events described in the literature
MetaCyc	http://metacyc.org/	MetaCyc is a database of nonredundant, experimentally elucidated metabolic pathways. MetaCyc contains more than 1100 pathways from more than 1500 different organisms. MetaCyc is curated from the scientific experimental literature and contains pathways involved in primary and secondary metabolism, as well as associated compounds, enzymes and genes
OMIM	http://www.ncbi.nlm.nih.gov/omim	OMIM (online Mendelian inheritance in man) is a comprehensive, authoritative and timely compendium of human genes and genetic phenotypes. The full-text, referenced overviews in OMIM contain information on all known Mendelian disorders and over 12,000 genes. OMIM focuses on the relationship between phenotype and genotype
2P2I	http://2p2idb.cnrs-mrs.fr/	The 2P2I database stores structural information about PPIs with known inhibitors and provides a useful tool for biologists to assess the potential druggability of their interfaces
PDSP	http://pdsp.med.unc.edu/	PDSP (Psychoactive Drug Screening Program) provides screening of novel psychoactive compounds for pharmacological and functional activity at cloned human or rodent CNS receptors, channels and transporters
PharmGKB	http://www.pharmgkb.org/	The PharmGKB database is a central repository for genetic, genomic, molecular and cellular phenotype data and clinical information about people who have participated in pharmacogenomics research studies. The data includes, but is not limited to, clinical and basic pharmacokinetic and pharmacogenomic research in the cardiovascular, pulmonary, cancer, pathways, metabolic and transporter domains
Protein Data Bank (PDB)	http://www.rcsb.org/pdb/home/home.do	The PDB archive contains information about experimentally determined structures of proteins, nucleic acids and complex assemblies. Users can perform simple and advanced searches based on annotations relating to sequence, structure and function
PubChem/PubMed	http://pubchem.ncbi.nlm.nih.gov/ http://www.ncbi.nlm.nih.gov/pubmed/	PubChem is a free database of small molecules and information on their biological activities. The system is maintained by the National Center for Biotechnology Information (NCBI), a component of the National Library of Medicine, which is part of the United States National Institutes of Health (NIH). It is linked to NIH PubMed/Entrez information
REACTOME	http://www.reactome.org/ReactomeGWT/entrypoint.html	REACTOME is an open source, open access, manually curated knowledgebase of biological pathways in humans. Pathway annotations are authored by expert biologists, in collaboration with Reactome editorial staff and cross-referenced to many bioinformatics databases
SIDER	http://sideeffects.embl.de/	SIDER (Side Effect Resource) contains information on marketed medicines and their recorded adverse drug reactions. The information is extracted from public documents and package inserts. The available information includes side-effect frequency, drug and side-effect classifications as well as links to further information, for example drug–target relationships
STITCH	http://stitch.embl.de/	STITCH (search tool for interactions of chemicals) is a searchable database that integrates information about interactions from metabolic pathways, crystal structures, binding experiments and drug–target relationships. This database contains interactions for between 300,000 small molecules and 2.6 million proteins from 1133 organisms
SUPERTARGET	http://bioinf-apache.charite.de/supertarget_v2/	SuperTarget is a database that contains more than 2500 target proteins, which are annotated with about 7300 relationships to 1500 drugs; the vast majority of entries have pointers to the respective literature source
TOXLINE	http://toxnet.nlm.nih.gov/cgi-bin/sis/htmlgen?TOXLINE	TOXLINE records provide bibliographic information covering the biochemical, pharmacological, physiological and toxicological effects of drugs and other chemicals. It contains over 4 million bibliographic citations, most with abstracts and/or indexing terms and CAS registry numbers
UniProt	http://www.ebi.ac.uk/uniprot/	UniProt is a freely accessible database of protein sequence and functional information, many of it derived from genome sequencing projects. It contains a large amount of information about the biological function of proteins derived from the research literature
WikiPathways	http://wikipathways.org/index.php/WikiPathways	WikiPathways is an open, collaborative platform dedicated to the curation of biological pathways. It was built on the MediaWiki software and thus enables a broad usage by the entire community

ontologies at the genome-wide level, with the need for a sustained development of ontologies for medicinal chemistry targets. This further expands toward ontologies especially dealing with ADME and toxicity. Because of its intrinsic complexity a toxicology ontology represents a particular challenge and is addressed within the IMI eTox Consortium [32]. Finally, to retrieve assay data not only indirectly from databases but also directly from the primary literature poses an immense challenge for chemical and biological text mining [33–35]. Some specific questions were related to PPIs, which again stresses the need for granularity in the gene classification systems. With the 2P2I database a specific data source regarding PPI inhibitors is emerging. The value of PPI datasets such as IntAct would need to be evaluated in the perspective of its relevance to contribute to the answer for Q8. One of the highest-prioritized questions was related to oxidoreductase enzymes. In this case, the Enzyme Commission (EC) classification system in UniProt provides a source for the enzyme classification.

Patent and publication node

Patents constitute a valuable source of information for chemical compounds and biological targets. To answer some of the questions related to patents, all patents that relate to the compound and the disease of interest have to be retrieved. In a second step, the targets need to be extracted from the patent claims. Currently, no public domain database with the required specific data associations exists. The European Bioinformatics Institute (EBI) in collaboration with the European Patent Office (EPO) is working toward this goal [36]. Also, the SureChem family of products [37] offers access to a comprehensive international patent collection, which is normalized and curated. Finally, the IBM patent analytics platform is a pharmaceutical industry consortium focusing on the same area [38,39]. However, accurate extraction of chemical structure information from patent documents constitutes a key challenge [40], which needs considerable attention. With respect to publications, the most popular public database used to access references of publications is PubMed. Although target and chemical names can be directly extracted from abstracts, extracting all chemicals in a publication with its associated data will certainly remain a significant challenge.

Finally, it has to be pointed out that the main goal of Open PHACTS – connecting publicly available databases – inherently puts a 'data bias' onto the whole system. In addition, the selection of data sources is also heavily influenced by their respective license. Thus, although some data sources can be better than others, the final choice is based on a list of criteria rather than on best coverage of the respective domain. These include, among others, the compatibility of the license with the Open PHACTS license, the availability of an RDF version, regular updates and maintenance, and data quality.

OPS example end-user applications (eApps)

The Open PHACTS infrastructure will consist of a series of software components that together form a platform that multiple applications can be built upon. All applications access the underlying data via an API that provides access to optimized queries of the system. Form-based queries within the core Open PHACTS interface (Open PHACTS explorer) will enable the bench scientist to address key use cases like the retrieval of compound, target and pathway information. In addition to the API, several other end-user applications will be co-developed with the Open PHACTS discovery platform, such as a target dossier, a polypharmacology browser, a chembio navigator and an application specialized for linking to the toxicity data store established under the framework of the eTox project. The relevance of these tools is not only to show how it is possible to build relevant end-user applications on top of the core platform but also to provide the bench scientist with immediate value. The target dossier is designed to provide a comprehensive view on pharmacologically relevant targets to answer questions regarding druggability, tissue expression profiles and implications in pathways, disease states and physiological mechanisms. The polypharmacology browser is a tool that aims at enabling scientists to define a target profile of interest and interrogate the Open PHACTS discovery platform for compounds having affinity for some or all of those targets, as well as for additional proteins that might show a degree of cross-pharmacology with any of the targets in the profile. The chembio navigator will enable intuitive browsing of the chemical and biological spaces. With filtering by various structure and physicochemical descriptors, as well as by chemical substructure and bioactivity data, the chembio navigator enables an interactive analysis of datasets. Connected to the Open PHACTS discovery platform, it will enable the user to drill down into the primary data via hyperlinks.

Challenges

Heterogeneity of data and chemical data integration

When integrating databases of different origin and built on different concepts (compound, target, pathway, disease, among others), the heterogeneous nature of the data needs to be emphasized, as does the diverse nature of the quality of data contained in the various databases. This poses a significant challenge to their integration within the framework of more-classical relational databases, especially for compound databases where the mappings between chemical compounds are generally based on an electronic structure format such as molfiles, SMILES and InChIs [41]. Indeed, currently there is no public information system available that is based on the relational database model and that allows the type of more complex competency questions posed here to be addressed. The majority of existing information systems has succeeded in integrating associations between pairs of the above mentioned data concepts. The ChEMBL system for instance provides an excellent integration of the compound–target concepts including an ontology for the represented targets. Also, within pharmaceutical companies, large-scale integration of chemical structure and bioactivity data – the SAR estate – has been successfully completed. AstraZeneca, for instance, reported recently the development of a relational enterprise application containing 45 million unique chemical structures from 18 internal and external sources [42]. The system merges compound-to-assay-to-result-to-target relationships enabling users to search with drug names, synonyms, chemical structures, patent numbers and target protein identifiers, at a scale not previously available. Similarly, the ChemSpider database has mapped together 26 million chemical structures from over 400 distinct data sources, mostly available online, and provides links out to PubChem (where assay information can be accessed), to Google Scholar and Pubmed (where articles can be accessed) and to Google Patents and the SureChem patent service

where patents can be retrieved. ChemSpider also has links out to the ligands on Protein Data Bank, and to the majority of the compound-based databases mentioned in this article including ChEBI, ChEMBL, DrugBank and HMDB. The handling of complex chemical queries such as structure-, substructure- and similarity-based searching for the Open PHACTS discovery platform will be handled by accessing web services provided by the ChemSpider database. ChemSpider holds the role of compound-based data aggregator for the project and brings together compound-based datasets, providing the appropriate database mappings to the triple store, and uses stringent controls for data processing and validation to ensure as high a quality as possible at the compound level. Crowdsourcing capabilities are available on the platform so that new compound data can be deposited and existing data can be annotated and curated. ChemSpider contributes a subset of the entire dataset available to the core platform. The data slice coincides with the data sources identified to be of value to the Open PHACTS project and includes ChEBI, ChEMBL and DrugBank, together with ChemSpider identifier mappings, chemical names and synonyms and structure representations including SMILEs and InChIs. On the basis of feedback from the EFPIA members of the Open PHACTS consortium, and in alignment with chemical compound standardization recommendations from the FDA [43], the ChemSpider database will be re-standardized in the near future to ensure compatibility between FDA standards and the Open PHACTS discovery platform. However, although all these successful efforts demonstrate that it is indeed possible to bring together the majority of the relevant chemical space, significant issues remain, such as synchronized updates and maintaining integrity of links, quality control of the chemical structures and, last but not least, the big issue of normalization and standardization of chemical structures. The latter comprises protonation states, salts and tautomers.

Further databases and need for ontologies

The focus aim of the Open PHACTS project is to cover the four high-priority data concepts: chemical compound, molecular target, biological pathway and disease, including the relevant ontologies. Via mappings of identifiers, the semantic approach opens a model-free approach for integration [44]. The data model is included in the data and, in principle, no further specification is needed. An explicit data model, as implied in relational databases, restricts the query capabilities. Thus, Open PHACTS fosters the reutilization and integration of the large public investment in data sources rather than generation of additional databases. Although the concepts 'chemical compound', 'biological target' and 'biological pathway' reference databases are well established, relatively few data sources can be found around the concepts 'disease/phenotype'. The development of public mechanistically determined clinical trial data sources are thus further encouraged and will help to advance application in translational medicine [45]. In addition, several other IMI projects, such as eTRIX, DDMORE and EHR4CR, are focusing on this area. A typical application would be, for instance, the systematic repurposing of marketed and experimental therapeutics by enabling the discovery of new potential therapeutic indications [46,47]. In early drug discovery, the development of databases of microscopic cell and organism phenotype data as generated in high-content screening (HCS) will potentially enable further closing of the gap between molecular and physiological observations. A large number of phenotypic compound screens are 'black box' screens where the identification of active compounds needs to be followed-up by *in silico* and *in vitro* chemogenomics target fishing approaches to elucidate their possible mechanisms of action. Although there is no concrete scientific competency question for this among the top 20 questions, many of them would be part of an HCS work-up workflow. The analysis of the scientific competency questions also demonstrates the need for further development of ontologies. Ontologies are computable formal explicit specifications of a shared concept of a knowledge domain. They define entities and the relationships between these entities and express this knowledge in a formal and computable way. These ontologies and classifications are indeed essential to allow not only data aggregation and navigation but also inference-based modeling.

In the perspective of medicinal-chemistry-driven applications as outlined in the scientific competency questions, comprehensive target and ADME ontologies are of the utmost importance, as is the contribution of ADME data to the public domain [48]. The GO classifications enable interesting aggregations for chemical biology applications such as providing known compounds active through a given biological mechanism like apoptosis or autophagy. The InterPro classification enables the aggregation of data for a given protein domain type. The ChEBI ontology [49,50] provides a chemical information ontology that enables aggregation of compound classes like lipids (or many others), and was established to provide biologists with a more intuitive access to the chemical space. Finally, as outlined above, ChemSpider allows large-scale chemistry integration and also aims at providing high-quality chemical compound standardization.

Concluding remarks

The top 20 priority questions presented herein are prevalently motivated by chemogenomics and chemical biology applications that are essential to the early drug discovery process. They refer to the basic concepts of compound–target and compound–target–disease/pathway relationships. Other questions are motivated by the more general scientific need to enable integrative assessment of information on compounds, targets, pathways and diseases. The first examples of such capabilities were provided by the Wild group using their semantic system to investigate the genetic basis of side-effects of thiazolinedione drugs, resulting in a hypothesis for the recently discovered cardiac side-effects of rosiglitazone (Avandia®) and a prediction for pioglitazone, which is backed up by recent clinical studies [51]. These scientific motivations are also well aligned with the overall goal of the Open PHACTS project to deliver an open, integrated and sustainable chemistry, biology and pharmacology resource for drug discovery. The openness of the system for profit-based and non-profit-based organizations, as well as the long-term sustainability plan for the Open PHACTS discovery platform and API, will be crucial to its success in industrial and academic drug discovery.

The functional Open PHACTS discovery platform will enable the retrieval of existing data and associations. Given the complexity of biological data, critical analysis by the end-user scientist will be needed to interpret the findings appropriately. This critical assessment will be especially important for the inferred knowledge

that is enabled by the semantic rule-set and inference technologies. In addition, with the enormous progress in modern biology, computational scientists are confronted with the need for flexible integration of a wide source of data, which requires new approaches. Finally, the concepts and technological solutions outlined and pursued for the open pharmacological space could be easily expanded to create an ecosystem of interoperable open spaces [open transcriptomic space (OTS), open genomic space (OGS): generally OXS] in the life science area.

Acknowledgements

The authors should like to thank all colleagues participating in the Open PHACTS consortium for discussions and insights leading to this consolidated view. The research leading to these results has received support from the Innovative Medicines Initiative Joint Undertaking under grant agreement no. [115191], resources of which are composed of financial contribution from the European Union's Seventh Framework Programme (FP7/2007-2013) and in-kind contribution of EFPIA companies.

References

1 Barnes, M.R. *et al.* (2009) Lowering industry firewalls: pre-competitive informatics initiatives in drug discovery. *Nat. Rev. Drug Discov.* 8, 701–708
2 Mons, B. *et al.* (2011) The value of data. *Nat. Genet.* 43, 281–283
3 Chen, B. *et al.* (2010) Chem2Bio2RDF: a semantic framework for linking and data mining chemogenomic and systems chemical biology data. *BMC Bioinformatics* 11, 255
4 Zhu, Q. *et al.* (2010) WENDI: a tool for finding non-obvious relationships between compounds and biological properties, genes, diseases and scholarly publications. *J. Cheminform.* 2, 6
5 Zhu, Q. *et al.* (2011) Semantic inference using chemogenomics data for drug discovery. *BMC Bioinformatics* 12, 256
6 Wild, D.J. *et al.* (2011) Systems chemical biology and the Semantic Web: what they mean for the future of drug discovery research. *Drug Discov. Today* 17, 469–474
7 Belleau, F. (2008) Bio2RDF: towards a mashup to build bioinformatics knowledge systems. *J. Biomed. Inform.* 41, 706–716
8 Samwald, M. (2011) Linked open drug data for pharmaceutical research and development. *J. Cheminform.* 3, 19
9 W3C. Available at: http://www.w3.org/Consortium/ (accessed May 2013)
10 Innovative Medicines Initiative. Available at: http://www.imi-europe.org (accessed May 2013)
11 Open PHACTS. Available at: http://www.OpenPHACTS.org (accessed May 2013)
12 Jacoby, E. (2011) Computational chemogenomics. *WIREs Comput. Mol. Sci.* 1, 57–67
13 Oprea, T.I. *et al.* (2011) Computational systems chemical biology. *Methods Mol. Biol.* 672, 459–488
14 Oprea, T.I. *et al.* (2007) Systems chemical biology. *Nat. Chem. Biol.* 3, 447–450
15 Bechhofer, S. *et al.* (2010) Why linked data is not enough for scientists. *Sixth IEEE e-Science Conference* Available at: In: http://eprints.soton.ac.uk/271587/
16 Galperin, M.Y. and Fernandez-Suarez, X.M. (2012) The 2012 nucleic acids research database issue and the online molecular biology database collection. *Nucleic Acids Res.* 40, D1–D8
17 Whitebread, S. *et al.* (2005) *In vitro* safety pharmacology profiling: an essential tool for successful drug development. *Drug Discov. Today* 10, 1421–1433
18 Vidal, D. *et al.* (2011) Ligand-based approaches to in silico pharmacology. *Methods Mol. Biol.* 672, 489–502
19 Hert, J. *et al.* (2008) Quantifying the relationships among drug classes. *J. Chem. Inf. Model.* 48, 755–765
20 Lagunin, A. *et al.* (2000) PASS: prediction of activity spectra for biologically active substances. *Bioinformatics* 16, 747–748
21 Cheng, T. *et al.* (2011) Identifying compound–target associations by combining bioactivity profile similarity search and public databases mining. *J. Chem. Inf. Model.* 51, 2440–2448
22 Jacoby, E. *et al.* (2009) Knowledge-based virtual screening: application to the MDM4/p53 protein–protein interaction. *Methods Mol. Biol.* 575, 173–194
23 Sayle, R. *et al.* (2012) Improved chemical text mining of patents with infinite dictionaries and automatic spelling correction. *J. Chem. Inf. Model.* 52, 51–62
24 Suriyawongkul, I. *et al.* (2010) The Cinderella of biological data integration: addressing some of the challenges of entity and relationship mining from patent sources. In *Data Integration in the Life Sciences* (Lambrix, P. and Kemp, G., eds), pp. 106–121, Springer Verlag
25 Fishman, M.C. and Porter, J.A. (2005) Pharmaceuticals: a new grammar for drug discovery. *Nature* 437, 491–493
26 Baggs, J.E. (2010) The network as the target. *WIREs Syst. Biol. Med.* 2, 127–133
27 Pence, D. and Williams, A.J. (2010) ChemSpider: an online chemical information resource. *J. Chem. Educ.* 87, 1123–1124
28 Williams, A.J. (2008) Public chemical compound databases. *Curr. Opin. Drug Discov. Dev.* 11, 393–404
29 Williams, A.J. and Ekins, S. (2011) A quality alert and call for improved curation of public chemistry databases. *Drug Discov. Today* 16, 747–750
30 Mestres, J. *et al.* (2008) Data completeness – the Achilles heel of drug–target networks. *Nat. Biotechnol.* 26, 983–984
31 Schuffenhauer, A. *et al.* (2002) An ontology for pharmaceutical ligands and its application for in silico screening and library design. *J. Chem. Inf. Comput. Sci.* 42, 947–955
32 The eTOX Website. Available at: http://www.etoxproject.eu/ (accessed May 2013)
33 Altman, R.B. *et al.* (2008) Text mining for biology – the way forward: opinions from leading scientists. *Genome Biol.* 9 (Suppl. 2), 7
34 Zimmermann, M. *et al.* (2005) Information extraction in the life sciences: perspectives for medicinal chemistry, pharmacology and toxicology. *Curr. Top. Med. Chem.* 5, 785–796
35 van Haagen, H. and Mons, B. (2011) *In silico* knowledge and content tracking. *Methods Mol. Biol.* 760, 129–140
36 Industry partnerships. Available at: http://www.ebi.ac.uk/industry/ (accessed May 2013)
37 SureChem – Patent chemistry made easy and accessible. Available at: http://www.surechem.com (accessed May 2013)
38 Griffin, T.D. *et al.* (2010) Annotating patents with Medline MeSH codes via citation mapping. *Adv. Exp. Med. Biol.* 680, 737–744
39 Robson, B. *et al.* (2011) Drug discovery using very large numbers of patents: general strategy with extensive use of match and edit operations. *J. Comput. Aided Mol. Des.* 25, 427–441
40 Downs, G.M. and Barnard, J.M. (2011) Chemical patent information systems. *WIREs Comput. Mol. Sci.* 1, 727–741
41 Williams, A.J. *et al.* (2012) Towards a gold standard: regarding quality in public domain chemistry databases and approaches to improving the situation. *Drug Discov. Today* 17, 685–701
42 Muresan, S. *et al.* (2011) Making every SAR point count: the development of Chemistry Connect for the large-scale integration of structure and bioactivity data. *Drug Discov. Today* 16, 1019–1030
43 Food and Drug Administration Substance Registration System Standard Operating Procedure. Available at: http://www.fda.gov/downloads/ForIndustry/DataStandards/SubstanceRegistrationSystem-UniqueIngredientIdentifierUNII/ucm127743.pdf (accessed May 2013)
44 Allemang, D. and Hendler, J., eds) (2011) *Semantic Web for the Working Ontologist: Effective Modeling in RDFS and OWL*, Morgan Kaufmann
45 Oprea, T.I. *et al.* (2011) Associating drugs, targets and clinical outcomes into an integrated network affords a new platform for computer-aided drug repurposing. *Mol. Inform.* 30, 100–111
46 Loging, W. *et al.* (2011) Cheminformatic/bioinformatic analysis of large corporate databases: application to drug repurposing. *Drug Discov. Today* 8, 109–116
47 Cavalla, D. and Singal, C. (2012) Retrospective clinical analysis for drug rescue: for new indications or stratified patient groups. *Drug Discov. Today* 17, 104–109
48 Ekins, S. and Williams, A.J. (2010) Precompetitive preclinical ADME/Tox data: set it free on the web to facilitate computational model building to assist drug development. *Lab Chip* 10, 13–22
49 de Matos, P. *et al.* (2012) A database for chemical proteomics: ChEBI. *Methods Mol. Biol.* 803, 273–296
50 Hastings, J. *et al.* (2011) The chemical information ontology: provenance and disambiguation for chemical data on the biological semantic web. *PLoS ONE* 6, e25513
51 He, B. *et al.* (2011) Mining relational paths in integrated biomedical data. *PLoS ONE* 6, e27506