ELSEVIER

feature

Drug discovery FAQs: workflows for answering multidomain drug discovery questions

Christine Chichester[1,5], christine.chichester@isb-sib.ch, **Daniela Digles**[2,5], **Ronald Siebes**[3], **Antonis Loizou**[3], **Paul Groth**[3] and **Lee Harland**[4]

Modern data-driven drug discovery requires integrated resources to support decision-making and enable new discoveries. The Open PHACTS Discovery Platform (http://dev.openphacts.org) was built to address this requirement by focusing on drug discovery questions that are of high priority to the pharmaceutical industry. Although complex, most of these frequently asked questions (FAQs) revolve around the combination of data concerning compounds, targets, pathways and diseases. Computational drug discovery using workflow tools and the integrated resources of Open PHACTS can deliver answers to most of these questions. Here, we report on a selection of workflows used for solving these use cases and discuss some of the research challenges. The workflows are accessible online from myExperiment (http://www.myexperiment.org) and are available for reuse by the scientific community.

Introduction

In recent years, there has been an explosion in the amount of chemical and biological information available. Various databases are doubling in size every 18 months [1] and new databases are continually being created. If these databases are to remain as information silos, it appears that our ability to generate vast quantities of data, often of varied quality, could surpass our ability to use these data meaningfully. For the pharmaceutical industry, coping with this data load is crucial for the accurate validation of potential drug targets. Therefore, data integration and computational techniques that facilitate searching through multiple heterogeneous data sources and filtering for specific criteria are key technologies in the drug discovery pipeline [2].

The decline in the number of new drugs being brought to the market by pharmaceutical companies has led to paradigm shifts in the strategies of the industry toward research and development [3]. These shifts have occurred throughout the drug discovery pipeline, from research and development to data integration. In the case of data integration, the model of accumulating all public domain data in-house to complement proprietary data independently of other companies is progressively being replaced by precompetitive initiatives that outsource the public data integration task. The Open Pharmacological Concepts Triple Store (Open PHACTS) Discovery Platform [4] (http://www.openphacts.org) originally sponsored by the Innovative Medicines Initiative (http://www.imi.europa.eu) is one of these data integration initiatives. It is distinguished by its use of semantic technologies and its focus on solving widely applicable drug discovery use cases.

Previously, Open PHACTS consortium members published a list of 20 core pharmacology-centered questions [5], which focused on use cases needed for specific research activities as well as for drug discovery in general. The '20 queries' approach is a method that keeps the focus on the most important features that systems should support [6]. These core questions comprise primarily four concepts that are important in pharmacological research: compound, target, pathway and disease. The questions provided model scenarios that drove the data selection and development processes required for integration of a diverse set of public domain databases. After the public release of the Open PHACTS Discovery Platform [4], we were interested in revisiting these

http://dx.doi.org/10.1016/j.drudis.2014.11.006

scientific competency questions to demonstrate how the system can be used to answer the prioritized questions efficiently. To do so, we used the platform in conjunction with the KNIME [7] computational workflow environment, which is commonly used in drug discovery.

A workflow approach

Analysis and hypothesis generation for drug discovery projects requires assembly, overlay and comparison of data from many sources [8], which is ultimately made possible by the use of shared identifiers and common semantics. Scientific workflows are typically used to automate the processing, analysis and management of the scientific data used by these projects. Most scientific workflow programs, such as KNIME [7], Taverna [9], or Pipeline Pilot [10], provide a user-friendly graphical workbench that enables scientists to create and visualize complex workflows easily that might comprise dozens of processing and analytical steps. Furthermore, many workflows provide mechanisms for tracing provenance and other methodologies that foster reproducible science [11,12]. Typically, the functionality of the workflow components falls into various categories, such as data transformation, data preparation and data analysis [13], and can be both part of the workbench or provided via web services (as application programming interfaces or APIs). As the number and variety of scientific open data APIs steadily grows, it is becoming easier for the scientific workflow community to integrate these APIs in their workflows [14].

Without the use of these workflow tools, manual data retrieval methods are typically used that require considerably more time and effort because of the complexities of data access. In most instances, data access and reformatting of different resources is a nontrivial exercise for most bench scientists, who are unskilled in programming languages. In the Open PHACTS project, we developed KNIME utility nodes that are set to leverage automatically the desired content from Open PHACTS web services and simplify the construction of workflow processes by realizing faster and more efficient data retrieval for answering use case questions.

Easy data retrieval

The Open PHACTS API [15] has been constructed to assure compatibility between the data retrieved and inputs required, which facilitates workflow construction. Results from API calls can contain two different types of value that can be used as queries: (i) resource identifiers denoting Web resources and (ii) literals denoting, for example, chemical depictions, such as SMILES and InChIs. Most API calls require Web resource identifiers, known as Universal Resource Identifiers (URIs), which are typically written in the form of a URL (e.g. http://purl.uniprot.org/uniprot/P35968 for VEGFR2 protein) as input. Given that 13 out of 20 use cases (Table 1) needed a sequence of at least two different API calls, the compatibility between results and input greatly eases the assembly of workflows by reducing the need to reformat queries. Moreover, the API has been developed in such a way to return results for which the chaining of calls is not always necessary. Seven of the 20 use cases could be answered directly using only one call to the API.

A brief overview of the Open PHACTS API[a] (Table 2) calls and the data returned from each call is the only background necessary to understand the approach used to answer the use case questions. We mostly relied on two of the main types of API call: (i) information calls and (ii) pharmacology calls.

Information calls (target information, compound information, disease information, pathway information and tissue information) return results that are specific to a query concept (Table 2). For example, the compound information call is used for use cases (Q2 and Q4 in Table 1), where there is a need to retrieve the SMILES strings. The SMILES then can be used sequentially in the chemical structure search using the similarity call to find similar compounds. If desired, there are some common utility methods available from the Open PHACTS API that enable conversion of chemical depictions into URIs (chemical structure conversion: SMILES to URL, InChI to URL, or InChIKey to URL) for use in subsequent pharmacology calls.

In contrast to the 'information' calls, 'pharmacology' calls return results based on an activity relation between a target and a compound, which in turn serves to link them (Fig. 1). Pathway and disease concepts are directly related to targets in API calls (pathway information: get targets, pathways for targets, disease for targets and targets for disease). This enables them to be related, in turn, to compounds via the targets and the pharmacology call. Results from both the 'information' and 'pharmacology' calls can be used directly in subsequent API calls, although the pharmacology calls are the most frequently exploited in the workflows to answer the use cases (18 of the 20 workflows).

[a] The API definition in full can be found at http://dev.openphacts.org.

Answering prioritized use cases

As evaluated by Azzaoui *et al.* [5], the top 20 pharmaceutical use cases fall into two clusters based on the concepts present in the questions (Table 2). Cluster 1 use cases deal primarily with compound–target concepts, whereas, in addition to the basic pharmacology, Cluster 2 questions include the added complexity of disease and pathway concepts as well as text-mined information from patents and literature. Correspondingly, to answer the range of use cases, the Open PHACTS API calls are divided into groups along the same concept lines (target, compound, pathway and disease) for data retrieval. This configuration facilitates the resolving of the previously proposed use cases, as well as enabling any new questions revolving around the same themes to be answered. The methodology for workflow construction (i.e. the chaining together of several calls) requires some understanding of the results returned from the specific API calls. However, the output of each call has been designed to return data specifically formatted to work as input for subsequent API calls.

Building the workflows for the use cases required that the concept central to the use case (target, compound, disease, or pathway) be identified. Then, to resolve its relations to the other concepts, it was often necessary to decompose the natural language of the use case to determine the correct flow for the API calls. This meant that the first concept mentioned in the original use case text was not necessarily the starting point for the workflow. To demonstrate the flow of API calls needed for answering the use cases, the text of the use cases were slightly rewritten (Table 1) from the original published text to align the order of the concepts in the question with the order of the API calls. The succession of API calls that were used for each question is indicated in bold in Table 1. Certain use cases also needed application of some of the filtering options, which are available from the API, to execute the different constraints set out in each use case (italics in Table 1). These filters can be used to restrict the results of an API call, for example to return only pharmacology data for a certain activity type, such as IC_{50}, to introduce activity cutoffs, or to return data for a specific organism only. The filter parameters can accept several values at once. For example, in Q1, by using the organism filter with the parameters of '*Mus musculus*|*Homo sapiens*', both the mouse and human data were retrieved simultaneously using a single API call.

Following this approach, we were able to answer 16 of the 20 identified use case

TABLE 1

Use case questions and the API calls needed to answer them[a]

ID	Use case question	Sequence of API calls with filters used and link to the workflow
Cluster 1 use cases (Q1–Q11): answers require mainly compound–target pharmacology data		
Q1	Give me all oxidoreductase inhibitors active <100 nM in human and mouse	**Target class pharmacology** (*target_organism=Homo sapiens|Mus musculus; minEx-pChembl=7*); http://www.myexperiment.org/workflows/4504.html
Q2	For a given compound, what is its predicted secondary pharmacology?	**Compound information**>**chemical structure search: similarity**>***compound adverse events***
Q3	Given a target find me all actives against that target, and find and/or predict the polypharmacology of actives	**Target pharmacology** (*minEx-pChembl=5*)>**compound pharmacology** (*minEx-pChembl=0*); http://www.myexperiment.org/workflows/4505.html
Q4	For a given interaction profile, give me similar compounds	**Compound information**>**compound information (Batch)**>**chemical structure search: similarity** (*searchOptions.Threshold=0.85*)>**compound information**; http://www.myexperiment.org/workflows/4516.html
Q5	For molecules that contain substructure X, retrieve all bioactivity data in serine protease assays	**Chemical structure search: substructure**>**compound pharmacology**, **target class members**; http://www.myexperiment.org/workflows/4478.html
Q6	For a specific target family, retrieve all compounds in specific assays	**Target class pharmacology**; http://www.myexperiment.org/workflows/4506.html
Q7	For a target, give me all active compounds with the relevant assay data	**Target pharmacology** (*minEx-pChembl=5*); http://www.myexperiment.org/workflows/4507.html
Q8	Identify all known protein–protein interaction inhibitors	**Target class pharmacology** (*target_type=ppi, minEx-pChembl=5*); http://www.myexperiment.org/workflows/4508.html
Q9	For a given compound, give me the interaction profile with targets	**Compound pharmacology** (*activity_type=IC50|EC50|AC50|Ki|Kd|Potency*); http://www.myexperiment.org/workflows/4509.html
Q10	For a given compound, summarize all similar compounds and their activities	**Chemical structure search: similarity** (*searchOptions.SimilarityType=0; searchOptions.Threshold=0.80*)>**compound pharmacology** (*activity_type=IC50|EC50|AC50|Ki|Kd|Potency*); http://www.myexperiment.org/workflows/4510.html
Q11	Retrieve all data for a given list of compounds depicted by their chemical structure (SMILES) with options to match stereochemistry	**Chemical structure search: exact** (*searchOptions.MatchType=2*)>**compound pharmacology, compound information**, **compound classifications** (*tree=chebi*); http://www.myexperiment.org/workflows/4511.html
Cluster 2 use cases (Q12–Q20): answers requiring pharmacology plus disease, pathway and text-mining data		
Q12	For a given compound, which of its targets have been patented in the context of a disease?	**Compound pharmacology**>***patents calls***>**disease for target**
Q13	For disease X, which targets have ligands in different stages of the development process with publications and/or patents describing these compounds?	**Targets for disease**>**target pharmacology** (*minEx-pChembl=5*), **target information**>***patents calls***
Q14	Target druggability: compounds directed against target X have what indications? Which new targets have appeared recently in the patent literature for a disease?	**Target pharmacology** (*minEx-pChembl=5*)>***indications for compounds***>***patent calls***>**disease for targets**
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?	**Classification of compounds for target** (*minEx-pChembl=5*) **Associations for disease** Competitive Intelligence data not available; http://www.myexperiment.org/workflows/4512.html
Q16	Targets in Parkinson's disease or Alzheimer's disease are activated by which compounds?	**Target for disease**>**target pharmacology** (*minEx-pChembl=5*); http://www.myexperiment.org/workflows/4513.html
Q17	For my specific target, which active compounds have been reported in the literature?	**Target pharmacology** (*minEx-pChembl=5*); http://www.myexperiment.org/workflows/4507.html
Q18	For pathway X, find compounds that agonize targets assayed in only functional assays with potency <1 MM	**Pathway information: get targets**>**target pharmacology** (*activity_type=Potency, max-activity_value=1000, activity_unit=nanomolar*); http://www.myexperiment.org/workflows/4514.html
Q19	For the targets in a given pathway, retrieve the compounds that are active with more than one target	**Pathway information: get targets**>**target pharmacology** (*minEx-pChembl=5*); http://www.myexperiment.org/workflows/4515.html
Q20	For a given disease, retrieve all targets in the pathway and all active compounds hitting those targets	**Targets for disease**>**target pharmacology** (*minEx-pChembl=5*); http://www.myexperiment.org/workflows/4513.html

[a] Priority use case questions and workflow sequence of API calls for retrieving answers. Bold text indicates the Open PHACTS API calls and the sequence of calls that were used to retrieve the data to answer the questions. Filtering parameters (in italics) indicate the values used with the API calls to answer the question. Bold italic text indicates an API call that is not yet realized. All workflows with example queries have been deposited to myExperiment [21] and the link to the workflow is given.

questions. All workflows (http://www.myexperiment.org/groups/1125.html) and KNIME nodes (https://github.com/openphacts/OPS-Knime) are available for reuse by others, from the Open PHACTS group on myExperiment (http://www.myexperiment.org; tagged with 'drug discovery faq' and 'open phacts') and the GitHub OPS-KNIME repository (https://github.com/openphacts/OPS-Knime), respectively.

Workflow construction challenges

From a close reading of the use cases, it is obvious that there are expressions that can be interpreted in more than one way, as well as uncertainty in the meaning of terms. Ambiguity

TABLE 2

A list of Open PHACTS API calls frequently used in workflows[a]

Open PHACTS API call	Data types	Provenance
Target information	Target depictions: amino acid sequence, number of residues, theoretical pI, mass, textual labels	SwissProt, DrugBank
	Target annotations: function, interacting proteins, target components, cellular localization, GO terms, UniProt keywords	SwissProt, Chembl, DrugBank
	Drugs specific for target	DrugBank
	Links to other datasets: GO annotations, Protein Databank	SwissProt
Compound information	Compound depictions: **SMILES, InChI, InChIKey**, molecular formula, textual labels	OPS Chemical Registry
	Compound properties: hydrogen bond donors/acceptors, molecular weight, rule of five violations, rotatable bonds, logP, polar surface area, melting point	OPS Chemical Registry
	Compound annotations: biotransformation, toxicity, protein binding, description, classification as per drug approval process, drug–drug interactions	DrugBank
Tissue information	Tissue description	neXtProt
	Links to other data sets (cross-references): foundational model of anatomy, Brenda tissue ontology, UBERON, medical subject headings	neXtProt
Disease information	Disease depictions: **textual label**	DisGeNet
	Annotations: disease class	DisGeNet
Pathway information	Pathway depictions: **textual label, pathway organism**, pathway elements	WikiPathways
	Annotations: pathway description, pathway ontology terms	WikiPathways
Target pharmacology	Pharmacology components: **target**, target textual name, **compound**	ChEMBL
	Compound–target activity annotations: published activity type, published activity value, activity units, pChEMBL value, **assay**, assay comment, assay organism, DOI, target type	ChEMBL
Compound pharmacology	Pharmacology components: **compound**, **target**	ChEMBL
	Compound–target activity annotations: published activity type, published activity value, activity units, pChEMBL value, **assay**, assay description, assay comment, assay organism, DOI, drug type, drug generic name	ChEMBL, DrugBank

[a] The Open PHACTS Information and Pharmacology API calls can return several data types from various different sources. The table lists the API call, the data returned (Data types) and the sources of the data (Provenance). Although most data returned for a query are optional (i.e. can be missing from the output if no data are available), some information (indicated in bold) must be available for any result to be returned from the call.

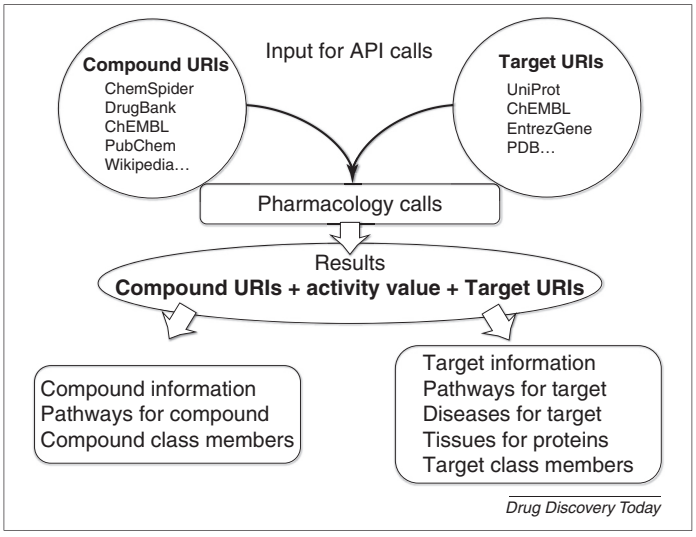

FIGURE 1

Schematic of the input and main results from the Open PHACTS Pharmacology API calls and the available subsequent application programming interface (API) calls. The Open PHACTS API calls allow several different identifiers (in the format of an URI) as input. The figure shows examples of different compound and target identifiers that can be used for input in the pharmacology calls. The main results from the pharmacology API calls are again compound and target URIs, which are connected by their reported bioactivity. These URIs can be used directly for several different API calls, as shown.

is an intrinsic phenomenon in scientific language and a fundamental property of linguistic expression, which can cause difficulties when developing universal workflows. For instance, the notion of 'active' compound depends heavily on the interpretation of a threshold concentration at which a compound exerts an effect. There were six (Q3, Q7, Q16, Q17, Q19 and Q20) use cases that contained the notion of 'active' compounds without specifying a specific activity threshold, and two use cases that did give threshold values (Q1 and Q18 with >100 nM and 1 nM, respectively). The pharmacology APIs have several filtering options, which can be set to retrieve data corresponding to specific thresholds. For the use cases with unspecified activity thresholds, in our deposited workflows, we filtered for a pChEMBL value >5 as our determination of active. pChEMBL is defined as the −log molar IC_{50}, XC_{50}, EC_{50}, AC_{50}, *Ki*, *Kd*, or potency. This value permits roughly comparable measures of half-maximal response concentration, potency and/or affinity to be compared on a negative logarithmic scale. For example, an IC_{50} measurement of 10 MM would have a pChEMBL value of five [16].

Another difficulty is the ambiguity of phrases, such as 'interaction profile' in Q4. 'Interaction profile' could refer to the interaction profile with targets as specified in Q9, making Q4 and Q9 very similar. Instead, 'interaction profile' was interpreted as the drug–drug interaction profile, which can be retrieved from the compound information API call, to highlight other available data.

Data quality is always an important concern in solving domain use cases [17]. Here, certain questions emphasize the requirement of retrieving 'all' results (Q6, Q8, Q10, Q11 and Q20). Clearly, it is possible that not every data point is available in the integrated data resources and, therefore, depending on the requirements of the user, might not produce the expected result. Data quality ultimately has many dimensions and, thus, is a complex problem when it comes to meeting user expectations. For example, in Q17, retrieving activity results from the ChEMBL data set implies that they were published, because the data are primarily assembled from the literature, although curation issues might cause discrepancies between the returned result and user expectations.

Finally, some use cases comprised several questions, of which not all could be resolved. Often these questions required data that are out of the scope of the project or not currently available from open access data providers. For example, Q15 has one component that fits the range of the currently integrated data (targets for disease), whereas the second part of the use case is focused on competitive intelligence data, which is not currently in the system. Answering Q13 and Q14 requires text mining of literature and patent data. Although not available currently, in subsequent releases of the platform data from SureChEMBL, an open database of the

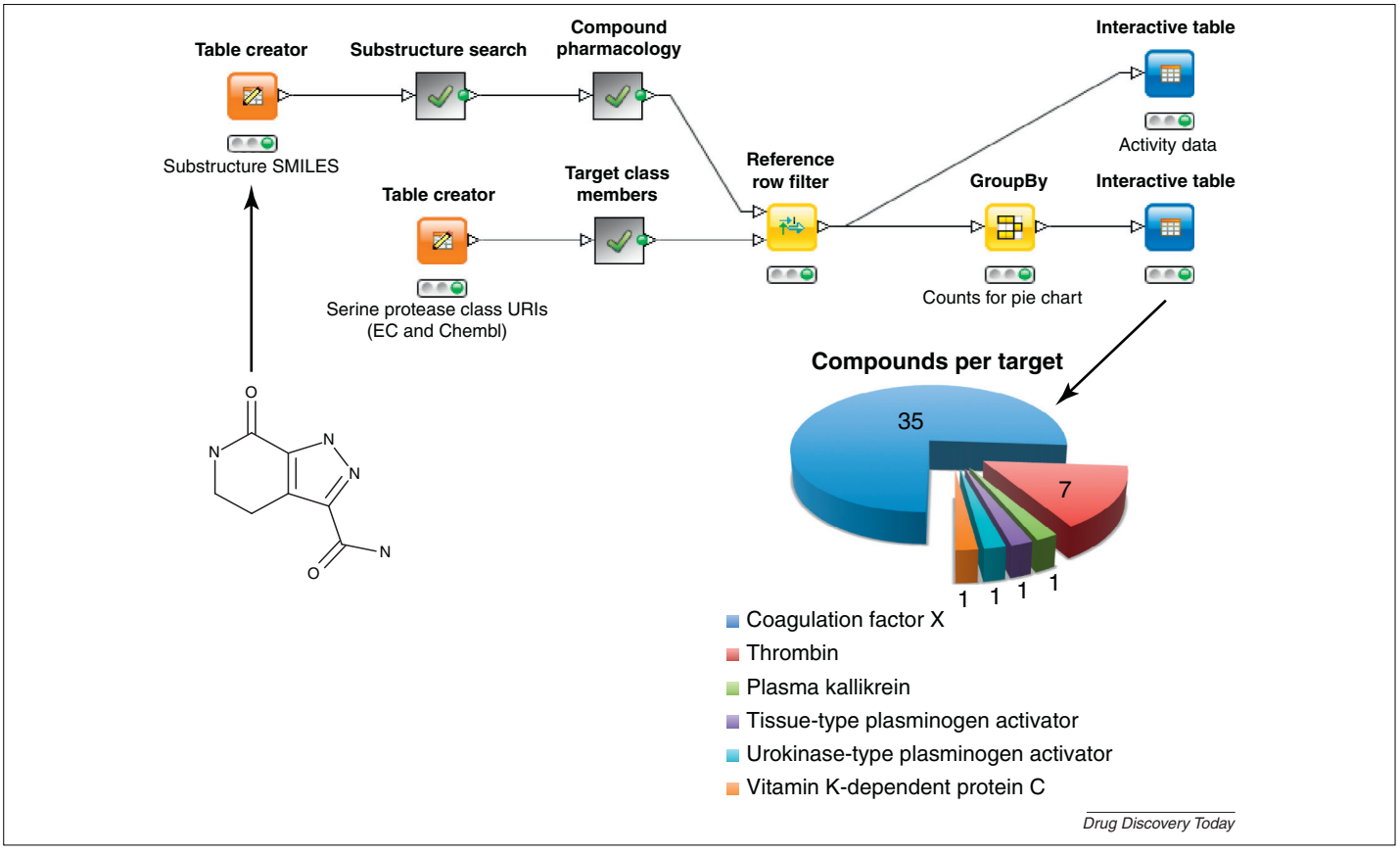

FIGURE 2

Example KNIME workflow for Q5. The workflow for Q5 uses three different Open PHACTS application programming interface (API) calls. A SMILES string is starting input for the chemical structure search:substructure call. The resulting compounds from this call are passed to the compound pharmacology call. Concurrently, the serine protease class URIs from the ENZYME and ChEMBL target classification hierarchies are the input for the target class members call. Subsequently, the pharmacology results are filtered according to the class members to determine the compounds for the serine protease type targets. The input structure and the resulting number of compounds per target are shown.

biopharmaceutical and associated therapeutic agents found in patents will be added, which will enable these use cases to be addressed. Moreover, when the data sources are updated, the user's expectations can change and some of the use cases could have alternate solutions using new data sources and API calls.

Ontology use cases

The prioritized use cases are mainly focused on the use of the pharmacology API calls. However, there are additional API calls (compound classifications, compound class members, classification of targets for compound, compound class pharmacology, target classifications, target class members, classification of compounds for target, target class pharmacology and hierarchy calls), used in only four of the 20 uses cases, that take advantage of searching in organized hierarchies or ontologies to find relevant information. Having a variety of ontologies available for exploration can be useful for drug discovery use cases because they offer a formal description of the relations between concepts in a specific domain. Gene Ontology (GO) [18] is arguably the most widely used ontology in life sciences. It structures information about biological processes, molecular functions and cellular components in a loosely hierarchical fashion. Structurally similar to GO, ChEBI [19] provides an ontology of chemical compounds of biological interest based on relations between chemical structural and functional features. The ChEMBL Target Classification scheme and the ENZYME classification (EC) [20] are also available in API calls and organize protein concepts in terms of parent–child hierarchies. The use of these ontologies to retrieve knowledge concerning sets of concepts within a domain is demonstrated in Q5 (Fig. 2). Here, we used two different hierarchies, the ChEMBL Target Classification and the EC, to retrieve the class of serine proteases as organized by two different authorities. The Q5 workflow is executed as follows; first, all pharmacology data corresponding to the specific substructure are retrieved using the sequence of chemical structure search:substructure followed by the compound pharmacology call. These data are then filtered for targets that are members of the serine protease class of the ChEMBL classification and the EC as retrieved with the target class members API call. The outcome is activities for six different serine protease targets. In a hypothetical use case, GO could be used in a similar workflow, for instance to obtain all compounds that are pharmacologically active against proteins known to be involved in processes such as cell replication or immune system response. These ontology resources provide a knowledge representation of a domain from a specific viewpoint in a machine usable format. Of course, what is specified in any ontology is not the absolute truth, but is only reflective of the available state of knowledge as documented by the data provider. As such, the availability of several different ontologies can offer another approach for information analysis, knowledge and intellectual property creation.

Concluding remarks

From the perspective of a researcher, the ideal data infrastructure should make it easy to search across different data sources containing data about drugs, clinical trials, diseases, pharmaceutical companies and so on, to identify novel and meaningful correlations. The Open PHACTS Discovery Platform provides this capability; leveraging public domain data, a full corpus of computational workflows has been developed that answers most of the high-priority use case questions from the pharmaceutical industry. Even though there are many ways to gather the data necessary to answer the use cases, the aim of the Open PHACTS Discovery Platform is to make these complex analyses simpler and faster to perform. For instance, identifier schemes do not need to be reconciled because the platform supports all the main inputs (SMILES, InChI, UniProt, EntrezGene, among others) and the data structures of the different integrated data sets are homogenized into a single API output. This highly available and reliable API generates output that also encourages application developers to build real-world applications (https://dev.openphacts.org/apps). Finally, the availability of workflow components (KNIME and Pipeline Pilot) lessens the load for busy scientists, because they do not need to worry about (i) installing local copies of the various databases; (ii) learning to write code or (iii) performing hefty Excel manipulations to address the questions they most frequently ask.

With the addition of new data sets, such as the information related to adverse events and patents, the relations between the concepts already available will be expanded to provide new paths for traversing the knowledge network. Concretely, the integration of these new data sets will enable all 20 use cases to be completely addressed. Furthermore, new and more elaborate workflows for computational drug discovery can be implemented. The use of ontologies is also a key step forward for structuring drug discovery data in a way that helps scientists to understand the relations that exist between concepts in specialized areas of interest, as well as to provide different perspectives on the relevant domains that should be more fully exploited.

Acknowledgements

The authors would like to acknowledge the contribution of the many Open PHACTS Consortium members for input in all aspects of the platform development. A list of Consortium members can be found at https://www.openphacts.org/partners/consortium. The research leading to these results has received support from the Innovative Medicines Initiative Joint Undertaking under grant agreement no. [115191], resources of which comprise financial contribution from the European Union's Seventh Framework Programme (FP7/2007–2013) and in-kind contribution of EFPIA companies.

References

1 Mizrachi, I. (2013) *GenBank. The NCBI Handbook* (2nd edn), NCBI
2 Marti-Solano, M. *et al.* (2014) Integrative knowledge management to enhance pharmaceutical R&D. *Nat. Rev. Drug Discov.* 13, 239–240
3 Goldman, M. (2012) Public–private partnerships need honest brokering. *Nat. Med.* 18, 341
4 Williams, A.J. *et al.* (2012) Open PHACTS: semantic interoperability for drug discovery. *Drug Discov. Today* 17, 1188–1198
5 Azzaoui, K. *et al.* (2013) Scientific competency questions as the basis for semantically enriched open pharmacological space development. *Drug Discov. Today* 18, 843–852
6 Szalay, A.S. *et al.* (2009) Gray's laws: database-centric computing in science. In *The Fourth Paradigm* (Hey, T. *et al.* eds), pp. 5–11, Microsoft Research
7 Ratnam, J. *et al.* (2014) The application of the Open Pharmacological Concepts Triple Store (Open PHACTS) to support drug discovery research. *PLOS ONE* (in press)
8 Berthold, M.R. *et al.* (2008) KNIME: the Konstanz Information Miner. In *Data Analysis, Machine Learning and Applications* (Preisach, C. *et al.* eds), pp. 319–326, Springer
9 Wolstencroft, K. *et al.* (2013) The Taverna workflow suite: designing and executing workflows of web services on the desktop, web or in the cloud. *Nucleic Acids Res.* 41, 57–561
10 Warr, W.A. (2012) Scientific workflow systems: pipeline pilot and Knime. *J. Comput. Aid. Mol. Des.* 26, 801–804
11 Littauer, R. *et al.* (2012) Trends in use of scientific workflows: insights from a public repository and recommendations for best practice. *Int. J. Digit. Curation* 7, 92–100
12 Curcin, V. *et al.* (2014) Implementing interoperable provenance in biomedical research. *Fut. Gen. Comp. Symp.* 34, 1–16
13 Garijo, D. *et al.* (2014) Common motifs in scientific workflows: an empirical analysis. *Fut. Gen. Comp. Symp.* 36, 338–351
14 Bhagat, J. *et al.* (2010) BioCatalogue: a universal catalogue of web services for the life sciences. *Nucleic Acids Res.* 38, W689–W694

15 Groth, P. *et al.* (2014) API-centric linked data integration: the Open PHACTS Discovery Platform case study. *Web Semant. J.* 359 (In press)
16 Bento, A. *et al.* (2013) The ChEMBL bioactivity database: an update. *Nucleic Acids Res.* 42, D1083–D1090
17 Missier, P. *et al.* (2006) Quality views: capturing and exploiting the user perspective on data quality. In *Proceedings of the 32nd International Conference on Very Large Data Bases (VLDB '06)* (Dayal, U. *et al*. eds), In pp. 977–988, VLDB Endowment
18 Gene Ontology Consortium (2012) Gene Ontology annotations and resources. *Nucleic Acids Res.* 41, D530–D535
19 Hastings, J. (2013) The ChEBI reference database and ontology for biologically relevant chemistry: enhancements for 2013. *Nucleic Acids Res.* 41, D456–D463
20 Bairoch, A. (2000) The ENZYME database in 2000. *Nucleic Acids Res.* 28, 304–305
21 Goble, C. *et al.* (2010) myExperiment: a repository and social network for the sharing of bioinformatics workflows. *Nucleic Acids Res.* 38, W677–W682

Christine Chichester[1,5], Daniela Digles[2,5], Ronald Siebes[3], Antonis Loizou[3], Paul Groth[3], Lee Harland[4]

[1]Swiss Institute of Bioinformatics, CALIPHO Group, CMU, Rue Michel-Servet 1, 1211 Geneva 4, Switzerland
[2]University of Vienna, Department of Medicinal Chemistry, Pharmacoinformatics Research Group, Althanstrasse 14, 1090 Vienna, Austria
[3]VU University Amsterdam, De Boelelaan 1081a, 1081 HV Amsterdam, The Netherlands
[4]Connected Discovery Ltd, 27 Old Gloucester Street, London WC1N 3AX, UK
[5]These authors contributed equally to this work.