Integration of bioactive substances data for preclinical testing with Cheminformatics and Bioinformatics resources

Branko Arsic[1], Marija Djokic[1], Vladimir Cvjetkovic[1], Petar Spalevic[2], Marko Zivanovic[3], Milan Mladenovic[4]

[1]Faculty of Science, University of Kragujevac, Department of Mathematics and Informatics
[2]Faculty of Technical Sciences, University of Kosovska Mitrovica
[3]Faculty of Science, University of Kragujevac, Department of Biology and Ecology
[4]Faculty of Science, University of Kragujevac, Department of Chemistry

Email: brankoarsic@kg.ac.rs, m.djokic@kg.ac.rs, vladimir@kg.ac.rs, petar.spalevic@pr.ac.rs, zivanovicm@kg.ac.rs, mmladenovic@kg.ac.rs

Abstract

Finding and comparing information published by institutions with similar goals can be a real challenge due to the fact that it is often necessary to interpret large amounts of data with different nomenclature and structure presentation having equivalent meaning. Using of Semantic Web technologies for publishing data accessible as Linked Open Data (LOD) encourages the integration of these datasets. In this paper, we aim to integrate and extend earlier developed ontology based information system with different datasets from PubChem, ChEMBL, DrugBank, ChemProt, etc. The information system supports the Research Center for Preclinical Testing (RC) which performs monitoring of in vitro effects of active substances on cell lines of different origin, primarily cancer cell lines and primary cells isolated from different tissues. In this way the researchers can be better focused on tested drugs with small IC_{50} factor for planning new experiments, saving the resources and time. Available data and Semantic Web technologies significantly improve synthesizing new substances, QSAR (Quantitative/Qualitative Structure Activity Relationships) analyses, the development of advanced algorithms for searching and establish future cancer bank for personalized medicine.

1 Introduction

The subject of various analysis that are carried out at the RC [1] includes monitoring of *in vitro* effects of active substances on cell lines of different origin, primarily cancer cell lines and primary cells isolated from different tissues. Experiments include cytotoxic active substances in human cancer cell lines, while monitoring includes the type of cell death, the mechanisms of apoptosis, migration and angiogenesis and prooxidant-antioxidant mechanisms which are important for regulation of these processes. Experiments are based on protocols such as MTT cytotoxicity test, AO/EtBr staining of cells for examination of the type of cell death, Western blot technique for examining proteins, Multiplex PCR, etc. Complete testing procedures consist of specific and complex relationships among various terms and concepts from the RC work area. As we predicted, the structure of experiment is expected to be further expanded as a consequence of complex research tasks that require flexible modeling and representation that can be easily updated [2].

Nowadays, biomedical researchers frequently need to use datasets derived from other systems. Data integration becomes an important precondition for successful performance of biomedical research. Over the past decade in the field of cheminformatics and bioinformatics research, the huge accumulation of various data (compounds, target, cell line, experiment, etc.) has generated a significant amount of knowledge. Similar institutions with the same purpose and goals are independent in their work, so their associated information models, protocols, cancer cell-lines, compounds, areas of interest and naming systems are different. This reflects to heterogeneity of data models, integration and retrieval data methods.

Semantic Web [3] technologies have mechanisms to solve these shortcomings. Ontology [4] as a main component of Semantic Web gives us a mechanism to connect similar data sources, published within different structures and for specific needs, providing the semantic context by adding semantic information to models. The capability to integrate heterogeneous datasets using a common terminology defined in ontology and search heterogeneous datasets in a single SPARQL [5] query represent a powerful tool for solving these problems.

In this paper we extended existing ontologies in RC and integrated it with other datasets from PubChem [6], ChEMBL [7], ChemProt [8] and DrugBank [9]. For example, this integration gives us a complete set of data for compounds used in an experiment: related compounds, chemical and physical properties, identifiers, IC_{50} value for targets in cancer cell line, and many more. Similarly, the same holds for cancer cell lines, the targets in cell-lines and applied protocols. Now, the researchers can find tested drugs with small IC_{50} factor for planning experiments with different conditions and protocols. The chemists can synthesize

new, untested substances following similar results in other laboratories which is an introduction for QSAR. Creating a system for intelligent and automatic saving of experimental data according to defined standards and publishing as LOD is one of the upcoming challenges.

This paper is organized in the following way: The second section gives an overview of the literature in existing field of work. The third section describes data integration mechanism between local ontologies and other related data sources. One part of this section is dedicated to SPARQL queries used for integration. Conclusion contains short survey of paper key points and directions for future work.

2 Related work

Ontologies for applications in biology and chemistry are becoming increasingly important especially for standardization of annotations essential for experimental work. Ontologies are finding their way in many branches of life sciences [10]. In [11] ontologies have been recommended for representation of molecular structures. A number of ontologies have already been developed in the disciplines of biomedicine. The National Center for Biomedical Ontology (NCBO) [12] builds a library of biomedical ontologies known as the Open Biomedical Ontologies (OBO) [13] which is now comprised of more than 70 biomedical ontologies (such as National Cancer Institute Thesaurus [14]). The Ontology for Biomedical Investigations (OBI) [15] project has developed an integrated ontology for the description of life-science and clinical investigations.

For the improvement of experimental work in any laboratory such as RC, there exists a need to involve and analyze several data sources which contain the information for compounds, cancer cell-lines, experiment protocols and behaviors in specific circumstances that depends on particular purpose. This approach requires a high level of integration of chemical and biological databases, the knowledge structure for comparing and predicting results, making decisions. For example, the obtained data will be more valuable if the information for drug effects is combined with cancer cell lines, various conditions and substance concentrations, different protocols, tumor markers results during control tests and so on.

Some examples of successful exploitation and integration of Semantic Web technologies with biological applications are [16] and [17]. In a paper [18] the authors indicated the importance of data integration in cheminformatics and bioinformatics. The OpenPHACTS initiative implements Semantic Web methods for drug discovery. Simon Jupp et al. [19] have developed the EBI RDF platform to provide a new entry point to querying and exploring integrated resources available at the EBI. Willighagen et al. [20] presented integration with other web resources including Bio2RDF, Chem2Bio2RDF, ChemSpider and ChEMBL database. Semantic Automated Discovery and Integration (SADI) Framework exposes the QSAR descriptor functionality of the Chemistry Development Kit and integration of computational functionality into formal ontologies [21].

3 Data integration process

The main activities of the RC are aimed at testing of the importance of physiological, genetic, molecular and biological tumor markers in assessing the effects of active substances and the prediction of pathological conditions in humans. In cooperation with the chemists, the activities start with the chemical synthesis of new varieties with improved properties of active substances. After this process, experiments include monitoring of in vitro effects of obtained substances in the cancer cell and primary cells isolated from different tissues [1]. This field of work results in complex experiment structure [2].

Earlier developed ontology based information system for center support is built as a union of separate ontologies. Separate ontologies enable integration with different datasets, easier for handling and presenting, easier for update and less error prune.

The core ontology is *Experiment*. Roughly speaking, each experiment is characterized by three main parts: active substance that is examined, model system (cell line) used for testing and protocol which is standard experimental procedure applied in the experiment. Accordingly, three external ontologies (*ActiveSubstance*, *ModelSystem* and *Protocol*) for storing data were constructed. The relationships between individuals (*Experiment-ActiveSubstance*, *Experiment-ModelSystem* and *Experiment-Protocol*) are implemented with object properties (Figure 1).

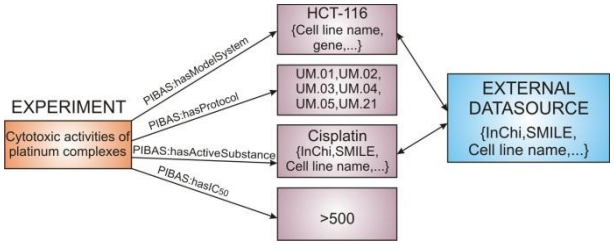

Figure 1. PIBAS ontologies integration

The same methodology was used for data integration between current local information system with external open services. *ActiveSubstance* ontology store the information about substance which is used in particular experiment. If some data is missing, researcher can use other systems to complement the knowledge. In LOD cloud there are a huge number of performed experiments and tested substances, with complete process results. In order to integrate resources we used InChi and SMILE substance notation, cell line name and gene. In this way the information about tested cell lines, affected targets, IC_{50}, and other experimental parameters can be available. Also, the researchers can obtain the list of all applied substances in different institutions, for which cell line is applied, avoiding

replications. An opposite direction will be possible too in upcoming time. The similar can be applied to *ModelSystem* ontology.

3.1 Data integration framework

The main idea for "integrating" application is the connection between several biochemical databases via LOD and searching heterogeneous sets in a single query via SPARQL. Web application has been developed using PHP and deployed on a Windows Web 2008 server.

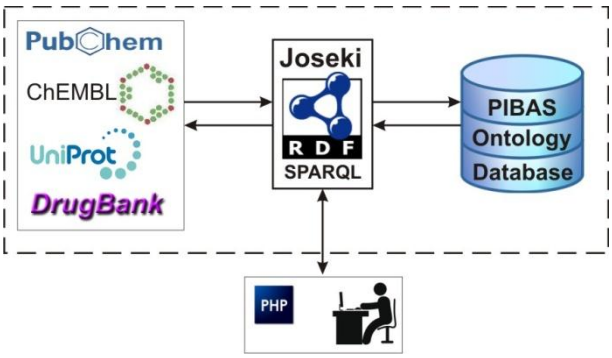

Figure 2. System architecture

Users can create individual customized queries that can retrieve any data from SPARQL endpoints given in advance. We extended SPARQL using ARQ [22] in Jena for services availability and data linking. Created Federated SPARQL query is forwarded to JOSEKI layer of architecture. Created SPARQL query is forwarded to JOSEKI layer of architecture. JOSEKI is SPARQL server which contains local ontology knowledge base. When the query is forwarded to the server and executed, the results from our ontology system and external data sources are obtained (Figure 2). As a results are shown combined data from different sources.

In following examples we demonstrated the mechanism of data integration system using Federated SPARQL queries.

Find IC_{50} value for target which reacted on a given active substance:

```
PREFIX compound: <http://chem2bio2rdf.org/pubchem/resource/>
PREFIX bindingdb: <http://chem2bio2rdf.org/bindingdb/resource/>
PREFIX pibas: <http://cpctas-lcmb.pmf.kg.ac.rs/2012/3/PIBAS#>

SELECT ?compound ?target ?IC50
from <http://cpctas-lcmb.pmf.kg.ac.rs/2012/3/PIBAS/expMethod.owl>
WHERE{
    ?compound_inchi pibas:inChI ?inchi.
    FILTER(?inchi="InChI=1S/C18H12BrN3/c19-14-6-3-7-15(10-14)22-
18-16-8-12-4-1-2-5-13(12)9- 17(16)20-11-21-18/h1-11H,(H,20,21,22)").
    SERVICE <http://cheminfov.informatics.indiana.edu:8890/sparql>
    {
        ?compound compound:std_inchi ?inchi.
        ?chemical bindingdb:cid ?compound.
        ?target bindingdb:Monomerid ?chemical;
        bindingdb:ic50_value ?IC50.
    }
}
```

Find the assay details of some compounds (see results in Figure 3):

```
PREFIX compound: <http://chem2bio2rdf.org/pubchem/resource/>
PREFIX chembl: <http://chem2bio2rdf.org/chembl/resource/>
PREFIX uniprot: <http://chem2bio2rdf.org/uniprot/resource/>
PREFIX rdfs: http://www.w3.org/2000/01/rdf-schema#
PREFIX pibas: <http://cpctas-lcmb.pmf.kg.ac.rs/2012/3/PIBAS#>

select ?pubchem_compound ?target ?activities
where
{
    SERVICE <http://cheminfov.informatics.indiana.edu:8890/sparql>
    {
      ?pubchem_compound compound:std_inchi ?inchi.
      ?compound chembl:cid ?pubchem_compound .
      ?activities chembl:molregno ?compound ;
      chembl:standard_value ?standard_value ;
      chembl:standard_units ?standard_units ;
      chembl:assay_id ?assay_id .
      ?assay2target chembl:assay_id ?assay_id.
      ?assay2target chembl:tid ?target.
      ?target chembl:pref_name ?pre_name.
      ?activities chembl:doc_id [chembl:pubmed_id ?pubmed_id].
    }
    FILTER regex(?pre_name, "Peroxisome proliferator", "i") .
}
LIMIT 6
```

SPARQL endpoints:

CPCTAS-LCMB

Get SPARQL result

SPARQL result

Number of rows: 6 results.

pubchem_compound	target	activities
http://chem2bio2rdf.org/pubchem/resource/pubchem_compound/3339	http://chem2bio2rdf.org/chembl/resource/chembl_targets/133	http://chem2bio2rdf.org/chembl/resource/chembl_activities/2649192
http://chem2bio2rdf.org/pubchem/resource/pubchem_compound/10021239	http://chem2bio2rdf.org/chembl/resource/chembl_targets/133	http://chem2bio2rdf.org/chembl/resource/chembl_activities/2649193
http://chem2bio2rdf.org/pubchem/resource/pubchem_compound/10021239	http://chem2bio2rdf.org/chembl/resource/chembl_targets/10958	http://chem2bio2rdf.org/chembl/resource/chembl_activities/2649163
http://chem2bio2rdf.org/pubchem/resource/pubchem_compound/3339	http://chem2bio2rdf.org/chembl/resource/chembl_targets/163	http://chem2bio2rdf.org/chembl/resource/chembl_activities/2649207
http://chem2bio2rdf.org/pubchem/resource/pubchem_compound/10021239	http://chem2bio2rdf.org/chembl/resource/chembl_targets/163	http://chem2bio2rdf.org/chembl/resource/chembl_activities/2649208
http://chem2bio2rdf.org/pubchem/resource/pubchem_compound/3339	http://chem2bio2rdf.org/chembl/resource/chembl_targets/163	http://chem2bio2rdf.org/chembl/resource/chembl_activities/2649177

Figure 3. Results for previous query

Conclusion

The mission of the Center is improvement of existing and development of new knowledge about the effects of different active substances and their potential application to living systems (cells, tissues, organs, organisms) using various contemporary scientific research methods and techniques, with strong emphasis on permanent improvement of research work. Required information support uses developed database with data about all examined substances, i.e. with data of the structures, chemical properties and results of biological tests with given conditions.

In this paper we developed software that can match tested substances from other services and ontologies, which enables RC staff to have better and

improved insight to all performed experiments and results.

It also enables the communication in the opposite direction, where other researchers can access RC experimental data and integrate it in their systems.

Acknowledgment

This paper was supported by the Ministry of Education, Science and Technological Development of the Republic of Serbia (scientific projects TR32023, III41010, ON174033 and III44006).

References

[1] CPCTAS-LCMB, Faculty of Science, University of Kragujevac, Serbia, http://cpctas-lcmb.pmf.kg.ac.rs

[2] V. Cvjetković, M. Đokić, B. Arsić and M. Ćurčić, "The ontology supported intelligent system for experiment search in the scientific research center", Kragujevac Journal of Science, Vol. 36, pp. 95-110, 2014.

[3] T. Berners-Lee, J. Hendler and O. Lassila, "The Semantic Web", Scientific American, 284 (5), pp. 29-37, May 2001.

[4] T. Gruber, "What is an Ontology?", Stanford University, Retrieved 2009-11-09.

[5] J. Pérez, M. Arenas and C. Gutierrez, "Semantics and Complexity of SPARQL", in 4th International Semantic Web Conference (ISWC), Athens, GA, USA, pp. 30-43, 2006.

[6] E. E. Bolton, et al. "PubChem: integrated platform of small molecules and biological activities", Annual reports in computational chemistry 4, pp. 217-241, 2008.

[7] A. Gaulton, et al. "ChEMBL: a large-scale bioactivity database for drug discovery", Nucleic acids research, 40 (D1): D1100–D1107, 2011.

[8] O. Taboureau, et al. "ChemProt: a disease chemical biology database", Nucleic acids research 39 (suppl 1): D367-D372, 2011.

[9] V. Law, et al. "DrugBank 4.0: shedding new light on drug metabolism", Nucleic acids research, 42(D1), D1091-D1097, 2014.

[10] M. Konyk, A. De Leon and M. Dumontier, "Data integration in the life sciences", Lecture Notes in Computer Science, Springer Berlin Heidelberg, pp. 169-176, 2008.

[11] H. J. Feldman, M. Dumontier, S. Ling, N. Haider and C. W. Hogue, "CO: A chemical ontology for identification of functional groups and semantic comparison of small molecules", FEBS letters, 579(21), pp. 4685-4691, 2005.

[12] A. M. Musen, "National Center for Biomedical Ontology", Encyclopedia of Systems Biology, Springer New York, pp. 1492-1492, 2013.

[13] B. Smith, M. Ashburner, C. Rosse, J. Bard, W. Bug, W. Ceusters and S. Lewis, "The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration", Nature biotechnology, 25(11), pp. 1251-1255, 2007.

[14] A. M. Musen, "National Cancer Institute Thesaurus", Encyclopedia of Systems Biology, pp. 1492-1492, 2013.

[15] M. Courtot, et al. "The OWL of Biomedical Investigations", OWLED, Vol. 432, 2008.

[16] H. Min, et al. "Integration of prostate cancer clinical data using an ontology", Journal of biomedical informatics, 42(6), pp. 1035-1045, 2009.

[17] D. Salvi, et al. "Merging Person-Specific Bio-Markers for Predicting Oral Cancer Recurrence Through an Ontology", Biomedical Engineering, IEEE Transactions on, 60(1), pp. 216-220, 2013.

[18] J. W. David, et al. "Systems chemical biology and the Semantic Web: what they mean for the future of drug discovery research", Drug discovery today, 17(9), pp. 469-474, 2012.

[19] S. Jupp, et al. "The EBI RDF platform: linked open data for the life sciences", Bioinformatics, 30(9), pp. 1338-1339, 2014.

[20] L. E. Willighagen, et al. "The ChEMBL database as linked open data", Journal of cheminformatics, 5(1), pp. 1-12, 2013.

[21] L. L. Chepelev and M. Dumontier, "Semantic Web integration of Cheminformatics resources with the SADI framework", Journal of cheminformatics, 3(1), pp. 1-12, 2011.

[22] ARQ http://www.openjena.org/ARQ