# Using Web Technologies for Integrative Drug Discovery

Qian Zhu[1] Sashikiran Challa[1] Yuying Sun[3]
Michael S. Lajiness[2] David J. Wild[1] Ying Ding[3]
*[1]School of Informatics and Computing, Indiana University, Bloomington, IN, USA*
*[2]Eli Lilly and Company, Indianapolis, IN, USA*
*[3]School of Library and Information Science, Indiana University, Bloomington, IN, USA*
*{qianzhu|schalla|pppurohi yuysun|djwild|dingying}@indiana.edu*
*LAJINESS_MICHAEL_S@lilly.com*

## Abstract

*Recent years have seen a huge increase in the amount of publicly-available information relevant to drug discovery, including online databases of compound and bioassay information; scholarly publications linking compounds with genes, targets and diseases; and predictive models that can suggest new links between compounds, genes, targets and diseases. However, there is a lack of tools and methods to integrate this information, and in particular to look for pertinent knowledge and relationships across multiple sources. At Indiana University we are tackling this problem by applying aggregative data mining tools and semantic web technologies including using an extensive web service infrastructure, RDF networks and inference engines, ontologies, and automated extraction of information from scholarly literature.*

## 1. Introduction

There is now an incredibly rich resource of public and proprietary information pertinent to drug discovery: for example, at the time of writing there were 64 million compounds and 1,683 bioassays from PubChem [1], 9 million protein sequences from SwissProt [2] and 58,000 3D structures from PDB [3], etc. One of the greatest challenges is how to use all of this information together to aid in the discovery of the next generation of treatments. To do this requires a framework for integration that includes semantics, cross-domain data mining and advanced determination of relevance. At Indiana University we are tackling this from a compound-centric approach using aggregative web service frameworks, semantic data mining, and automated mining of journal articles.

We have built a prototype aggregate data mining tool called WENDI (Web Engine for Non-obvious Drug Information) which aggregates information about a compound from several different web services and finds non-obvious associations between compound and biological activities.

## 2. Background

At Indiana University, we recently developed a Cyberinfrastructure [4] for cheminformatics, called *ChemBioGrid*, which has made a multitude of databases and computational tools freely available for the first time to the academic community in a web service framework [5]. Based on that, we are employing two key technologies: aggregate web services which call multiple individual web services and aggregate the results, and Semantic Web languages for the representation of integrated data.

In this paper, we would like to introduce an aggregate data mining tool, WENDI extended *ChemBioGrid*, the Web Service infrastructure and employed it as the primary data source.

## 3. Implementation

### 3.1. Databases service

Our compounds related databases are housed on a Linux server running the PostgreSQL database system, with gNova CHORD [6] installed to allow chemical structure searching and 2D similarity searching through the generation of fingerprints. The Databases, like PubChem_Compound [1], PubChem BioAssay [1], Pub3D [7], Drugbank [8], MRTD [9], Medline [10], CTD [11], HuGEpedia [12], ChEMBL [13] are being used in WENDI system.

### 3.2. Chemical-Disease-Gene relationships predictive service

PhenoPred [14], a matrix of predictions of gene-disease relationships based on known relationships mined from the literature and machine learning predictions. We have implemented this matrix as a database in our web service infrastructure, and employed it in WENDI to include prediction of disease-gene relationships any time a gene/disease is identified in the descriptions of PubChem Bioassay, usages of Drug from Drugbank, and abstracts/titles of Journal. Thus we can establish a link, say, between a/an assay/drug/paper, a gene, and/or a disease.

Also we have some other chemogenomics data, like Comparative Toxicogenomics Database (CTD) [11]; HuGEpedia [12]; ChEMBL [13]. These datasets have relationships between compounds and diseases/genes. We employ Tanimoto Coefficient based Swamidass heuristic [15] to reduce the subset of molecules that need to be searched in similarity calculations to speed up the similarity search, to find similar compounds. The comparing performance is shown in Table 1 for the SELECT SQL clauses, one with Swamidass heuristic, and another without. From the Table, we would like to say Swamidass heuristic plays a critical role in the compounds similarity search process. The framework of building chemical compound-disease-gene relationship is shown in Figure 1.

**Table 1. Comparasion of SELECT clauses with or without Swamidass heuristics**[a]

| SELECT SQL | Performance |
|---|---|
| FROM pubchem_compound<br>WHERE<br>tanimoto(fingerprint_database[b],<br>fingerprint_query[c]) > 0.8; | Total runtime:<br>83,322.706 ms<br><br>Total rows: 42 |
| FROM pubchem_compound<br>WHERE<br>gfpbcnt_database[d] BETWEEN<br>(gfpbcnt_query[e] * 0.8) AND<br>(gfpbcnt_query / 0.8) AND<br>tanimoto(fingerprint_database,<br>fingerprint_query) > 0.8; | Total runtime:<br>27,168.052 ms<br><br>Total rows: 42 |

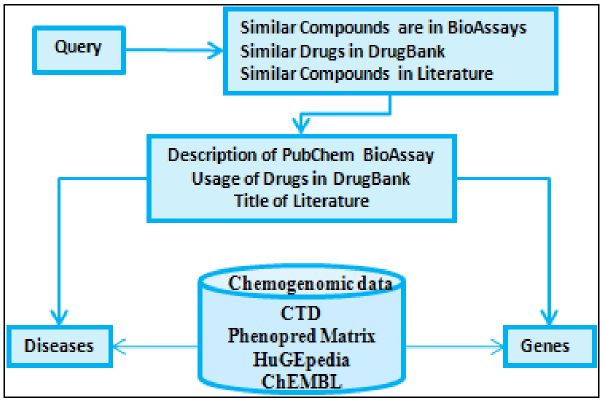

Query
Similar Compounds are in BioAssays
Similar Drugs in DrugBank
Similar Compounds in Literature
Description of PubChem BioAssay
Usage of Drugs in DrugBank
Title of Literature
Chemogenomic data
CTD
Phenopred Matrix
HuGEpedia
ChEMBL
Diseases
Genes

**Figure 1. Framework of Chem-Disease-Gene Relations**

### 3.3. Aggregate web service and client

WENDI is a web service and client that takes a compound as input, and aggregates information from multiple data sources, predictive models, and several algorithms developed at Indiana University. WENDI client will take a SMILES string representing a compound of interest as input, send it to the WENDI web service, and outputs an XML file of information about the compound aggregated by calling multiple web services. This XML file can then be parsed by an intelligent client to integrate information pertinent to compound properties. See Figure 2.

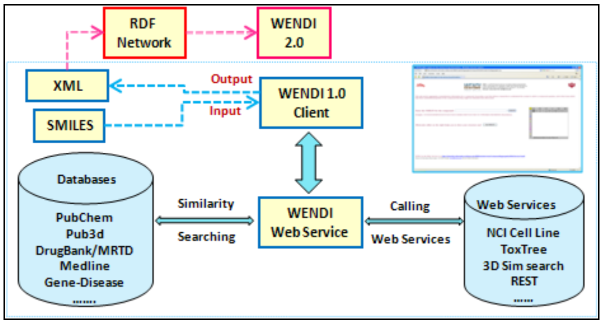


**Figure 2. WENDI 1.0 Workflow**

We have released the WENDI Web Service and WENDI public version through the following link: https:cheminfov.informatics.indiana.edu:8443/WENDI_PUBLIC/WENDI.WSDL https:cheminfov.informatics.indiana.edu:8443/WENDI_PUBLIC/WENDI.jsp

## 4. Corroborating Evidence – Case Study

WENDI automatically aggregates all the information in XML, also it will gather corroborating and conflicting evidence, and will cluster evidence by biological theme automatically.

[a] SMILES for query compound used for searching: 'CC(C)NCC(O)COc1ccccc1CC=C'; 166 MDL MACSS keys calculated as fingerprint.

[b] fingerprint_database: fingerprints from database.

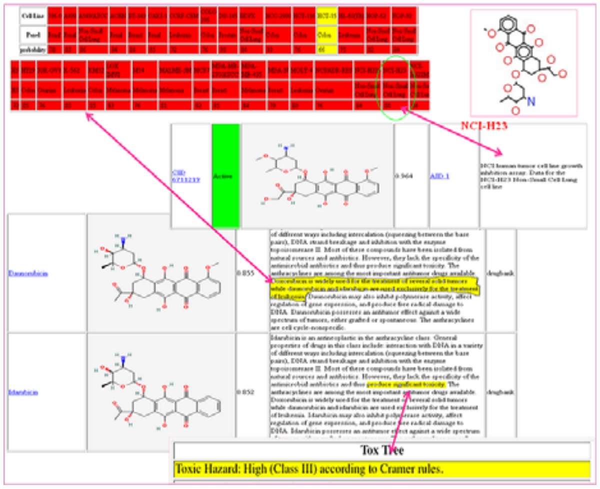
[c] fingerprint_query: fingerprint of the query compound.

[d] gfbcnt_database: total number of 1 in each fingerprint stored in database.

[e] gfbcnt_query: total number of 1 in the fingerprint for query compound.

The corroborating evidence with WENDI 1.0 can be identified by a selected query compound, Doxorubicin [http://en.wikipedia.org/wiki/Doxorubicin], in Figure 3.

1) The nonspecific tumor inhibition properties of Doxorubicin are identified by our NCI tumor Cell Line predictive models (First two tables, background color of most of cells are red, in Figure 1). The predictive probabilities greater than 0.7 (more active) are in Red, the probabilities between 0.7 and 0.5 shown in Yellow, otherwise in Grey. Ontologically marked-up text in the description of Danorubicin, a 0.855 similar drug from Drugbank gives the similar properties as NCI predictive models.
2) Doxorubicin is predicted active in NCI-H23 (lung cancer) (see green circle) and a 0.964 similar compound was shown to be experimentally active in NCI-H23 in a PubChem bioassay.
3) The compound is predicted to have toxicity issues by using our ToxTree predictor and this is noted in the ontologically marked-up text of similar drugs.



CID 6711239 | Active | 0.964 | AID 1 | NCI human tumor cell line growth inhibition assay. Data for the NCI-H23 Non-Small Cell Lung cell line

Daunorubicin | 0.855 | drugbank

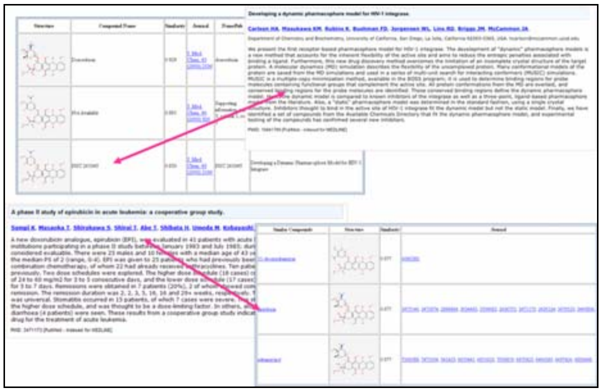
Idarubicin | 0.852 | drugbank

Tox Tree

Toxic Hazard: High (Class III) according to Cramer rules.

**Figure 3. Identification of corroborating evidence with WENDI 1.0**

WENDI will also gather some insights from similar compounds in the literature. By identifying journal articles that contain similar compounds to the query, we highlight potential therapeutic applications of each similar compound. In Figure 4, we identify similar compounds noted in the journal articles which are reported with a well known leukemia usage, but also lesser known application of similar compounds against HIV-1 Integrase.



Developing a dynamic pharmacophore model for HIV-1 integrase.

Carlson HA, Masukawa KM, Rubins K, Bushman FD, Jorgensen WL, Lins RD, Briggs JM, McCammon JA.

A phase II study of epirubicin in acute leukemia: a cooperative group study.

**Figure 4. Gathering insights from similar compounds in the literature in WENDI 1.0.**

## 5. OWL/RDF Representations and Inference

Whilst WENDI 1.0 provides automated aggregation of information, it relies on manual interpretation of the output and identification of corroborating or conflicting information. We are developing a new version, WENDI 2.0 which will provide automated organization of the information by disease area and identification of corroborating and conflicting information, as well as expanding the range of information accessed about a compound. In particular, it will use networks of RDF statements (e.g. the query compound is similar to compound X, compound X is active in assay Y, assay Y is associated with gene BRCA1) along with deductive reasoning tools to infer relationships between the query compound and genes and diseases. This will allow us to cluster insights by disease, and then prioritize the output based on the amount of evidence linking a compound to a disease. The system will automatically highlight positive and negative evidence for a compound-disease hypothesis, giving the rationale for the hypotheses based on the RDF chains.

### 5.1. Generation of RDF’s based on WENDI Ontology

Information is extracted from the XML using XML DOM and it is converted into triples in Turtle format based on an Ontology created in house called WENDI Ontology.

WENDI Ontology is created using OWL (Web Ontology Language). The Classes in the WENDI Ontology are: Chemical Compound, BioAssay, Journal Article, Gene, and Disease. The Object Properties defined in the Ontology with the domain and range classes specified as isSimilarTo, is3DSimilarTo, isActiveIn, isInactiveIn, isContainedIn, hasGenes, hasDisease, isAssociatedWith, hasSimilarity, etc.

Some of the triples in Turtle format generated based on the WENDI Ontology automatically listed in table 2. And the WENDI RDF graph is shown in Figure 5, the inferred relationships in red arrows can be pointed out based on the obvious distinct ones, like query-PubChem ID-PubChem BioAssay ID, etc.

**Table 2. Some Triples based on WENDI Ontology**

| |
|---|
| WO: querycmpd WO: isSimilarTo<br>WO: cid24871487. |
| WO: cid24871487 rdf:type WO: ChemicalCompound;<br>WO: isActiveIn WO: aid1469. |
| WO: aid1469 rdf:type WO: BioAssay;<br>WO: hasGenes WO: COL4A4. |
| WO: COL4A4 rdf:type WO: Gene;<br>WO: isAssociatedWith WO: Nephritis. |

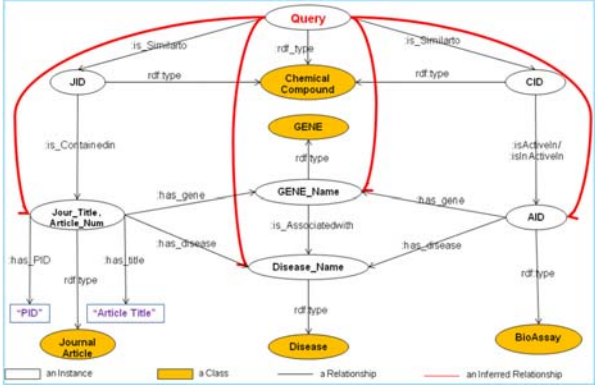


**Figure 5. RDF graph for WENDI**

### 5.2. Framing the Rules:

Once the RDF triples are generated, they are loaded into Ont Model Class in Jena, a Java Platform to build semantic web applications. The rules are written as shown below:

```
[rule1 :(? querycmpd WO: isSimilarTo ?cid)
        (? cid WO: isActiveIn ?aid)
- >(? qerycmpd  WO:mightBeActiveIn ?aid)]
```

The rule 1 means that if there is a triple with ‘isSimilarTo’ as property, query compound as subject and a compound C as object, and if there is a triple with ‘isActiveIn’ as property, Compound C as subject, bioassay id as object, then infer a relationship between query compound and bioassay id by creating a triple with ‘mightBeActiveIn’ as property, query compound as subject and bioassay id as the object. The actual inferencing is done by the Jena generic reasoner. We employ this rule-based inference algorithm to derive new relationships from the RDF network; these rules can be extended further to include genes and diseases using forward inferencing.

WENDI2.0 (accessed by the link: https://cheminfov.informatics.indiana.edu:8443/WENDI_LILLY_2/Compound Information Aggregation.jsp) provides a big picture for the Medicinal Chemists to understand the nature of the compound and its potential in curing a particular disease. Thus by having the information generated by WENDI as RDF triples automatic inferencing could be done.

## 6. Future work

Since WENDI performance mostly depends on the dataset, if no similar compounds can be found from our database, then no pertinent results will return back. So more chemical/biological/chemogenomic data will be added into WENDI, beyond this point, some predictive models will be built on the fly, then the predictive values will be given out in case no results from the database. Also we have an ongoing project named Chem2Bio2RDF, which is allowing a marriage of chemical information systems with network biology systems, we would like to make particular relations between these two, then more information about systems chemical biology can be retrieved based on some specific SPARQL queries.

## 7. References

[1] *http://pubchem.ncbi.nlm.nih.gov/*
[2] http://www.expasy.org/sprot/

[3] http://www.rcsb.org/pdb/home/home.do

[4] http://en.wikipedia.org/wiki/Cyberinfrastructure

[5] Dong, X., Gilbert, K.E., Guha, R., Heiland, R., Kim, J., Pierce, M.E. Pierce, Fox, G.C. and Wild, D.J. "Web service infrastructure for chemoinformatics." *J. Chem. Info. Model.* 2007 47(4):1303-1307.

[6] http://www.gnova.com

[7] Pub3D is a 3D version of PubChem, in which we have generated a single conformer for 99% of PubChem using the smi23d suite of programs.

[8] http://www.drugbank.ca

[9] http://www.fda.gov/CDER/Offices/OPS_IO/MRTD.htm

[10] http://www.nlm.nih.gov/bsd/licensee/2009_stats/baseline_med_filecount.html

[11] http://ctd.mdibl.org

[12] http://hugenavigator.net

[13] http://www.ebi.ac.uk/chembldb/

[14] A project in the Bioinformatics group at IU http://206.176.233.171/

[15] S. J. Swamidass and P. Baldi "Bounds and Algorithms for Fast Exact Searches of Chemical Fingerprints in Linear and Sublinear Time." *J. Chem. Inf. Model.*, 2007, 47:302-317.