Life sciences data defragmentation: a case study

6th International Symposium on Computational Methods in Toxicology and Pharmacology Integrating Internet Resources

CMTPi -2011, 3-7 Sep, Maribor, SI

Dr. Nina Jeliazkova
Ideaconsult Ltd., Bulgaria
www.ideaconsult.net
We live in a standardized world

• We take it for granted:
  – Clothes size
  – Phones
  – CDs
  – Computers and computer networks, Internet
• National Institute of Standards and Technology: ~ 800,000 global standards;
• Without standardization, there would be no mass production or mass communication;
• Without standardization there wouldn't be a modern economy;

It all started ~150 years ago with the fight to **standardize simple things** - fasteners - screws, nuts and bolts - that enable making more complex tools:
Standards in Life Sciences

- **Cheminformatics:**
  - Historically, the cheminformatics world has been driven by *de facto* standards, developed and proposed by different vendors;
  - A number of initiatives (relatively recent), have adopted open standardisation procedures (most notably InChI, CML, BlueObelisk initiatives), but there are no requirements for independent interoperable implementations so far;

- **Bioinformatics:**
  - Many (relatively recent) open data initiatives / standardisation efforts.
  - Recent paper in Nature Reviews Drug Discovery: **MIABE - Minimum Information about a Bioactive Entity**
  - EC Cordis news Sep 1, 2011: “Industry and academic actors from chemistry world agree on new bioactive molecule standard”

- **Summary**
  - Chemistry/Biology software and databases can live in their own worlds, unless we want data shared and tools interoperable
  - Interoperability / standards may affect business models
ECHA web pages
What end users really need:

The user profile: organic chemistry background, working in industry, uses computational/modelling tools, but not a developer/programmer

- I can do web search in the following databases and look for a compound (and perhaps later for some toxicity endpoint)
  - Google - http://www.google.com/

(Disclaimer: the list is not comprehensive)

- But how can I retrieve results for multiple compounds and endpoints automatically, without going manually to all the web pages?
  - And if technically possible, is it legal?
Motivation

• Predictive Toxicology applications need **common components**:
  – Access to datasets
  – Algorithms for descriptor calculation and model building
  – Validation routines

• The state-of-the-art involves re-implementation of these components in every new application
  – highly fragmented data

• If we had these components readily available we could
  – Quickly build new applications for specific purposes
  – Experiment with new combinations of algorithms
  – Speed up method development and testing
  – ...
OpenTox Components

- **Compounds**: Structures, names, ...
- **Features**: Chemical and biological (toxicological) properties, substructures, ...
- **Datasets**: Relationships between compounds and features
- **Algorithms**: Instructions for solving problems
- **Models**: Algorithms applied to data yield models, which can be used for predictions
- **Validation**: Methods for estimating the accuracy of model predictions
- **Reports**: Report predictions and models, e.g. to regulatory authorities
- **Tasks**: Handle long running calculations
- **Authentication and Authorization**: Protect confidential data
- **Service registration and querying**: Finding services of specific type
Requirements & Technological choices

• Platform independence
• Interoperability with external programs and data sources
• Transparency for scientific and regulatory credibility
• Open for future extensions

• Web services (REST)
• Communication through well defined interfaces
• Ontologies for the exchange of knowledge and data
• Use and promote open standards
• Open source components
OpenTox API (Application Programming Interface)

- How applications talk to each other
- How developers implement applications

http://opentox.org/dev/apis/api-1.2

<table>
<thead>
<tr>
<th>Feature</th>
<th>GET</th>
<th>POST</th>
<th>PUT</th>
<th>DELETE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dataset</td>
<td>GET</td>
<td>POST</td>
<td>PUT</td>
<td>DELETE</td>
</tr>
</tbody>
</table>

| Compound | GET | POST | PUT | DELETE |

| Algorithm | GET | POST | PUT | DELETE |

| Task      | GET | POST | PUT | DELETE |

| Validation| GET | POST | PUT | DELETE |

| Ontology  | GET | POST | PUT | DELETE |

| AA        | GET | POST | PUT | DELETE |
Algorithms

- Algorithms for **descriptor calculation**: generation and **selection** of **features** for the representation of chemicals (structure based features, chemical and biological properties);
- **Classification** and **regression** algorithms for creation of (Q)SAR models;
- **Rule based** algorithms;
- Algorithms for the **aggregation** of predictions from multiple (Q)SAR models and endpoints, and aggregation of predictions;
- **General purpose algorithms** (e.g. for visualization, similarity and substructure queries, applicability domain, read across, ...)

OpenTox

Ideaconsult Ltd.
Uniform approach to data processing

Read data from a web address - process - write to a web address

Dataset + Algorithm = Dataset

- Feature
  - GET
  - POST
  - PUT
  - DELETE

- Compound
  - GET
  - POST
  - PUT
  - DELETE

- Dataset
  - GET
  - POST
  - PUT
  - DELETE

- Feature
  - GET
  - POST
  - PUT
  - DELETE

- Compound
  - GET
  - POST
  - PUT
  - DELETE

http://myhost.com/dataset/trainingset1

http://myhost.com/dataset/results

http://myhost.com/algorithm/{myalgorithm}
Models

• **Models**: Models are generated by respective algorithms, given specific parameters and data
  – Statistical models are generated by applying statistical/machine learning algorithms to specific dataset and parameters
  – Models can be other than statistical, e.g.
    • expert defined rules,
    • quantum mechanical calculations,
    • metabolite generation, etc.
• The intention of the framework is to be generic enough to accommodate varieties of predictive models.
• Model services provide facilities to inspect, store and delete models. Every model is identified by unique web address.
Uniform approach to models building

Read data from a web address - process - write to a web address

Dataset
GET
POST
PUT
DELETE

Feature
GET
POST
PUT
DELETE

Compound
GET
POST
PUT
DELETE

Algorithm
GET
POST
PUT
DELETE

Model
GET
POST
PUT
DELETE

http://myhost.com/dataset/trainingset1

http://myhost.com/algorithm/neuralnetwork

http://myhost.com/model/predictivemodel1

OpenTox

Ideaconsult Ltd.
# OpenTox datasets: Uniform data access

## Table analogy (but defined as OWL)

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><a href="http://myhost/compound/413">http://myhost/compound/413</a></td>
<td>N,N-dimethyl-4-aminoazobenzene</td>
<td>CN(C1=CC=C(C=C1)N=N/C2=CC=CC=C2)C</td>
<td>3.123</td>
</tr>
<tr>
<td><a href="http://myhost/compound/4497">http://myhost/compound/4497</a></td>
<td>4-acetamidofluorene</td>
<td>O=C(Nc3c2c1ccccc1Cc2cc3)c</td>
<td>2.085</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

```owl
http://myhost.com/feature/22114
a ot:Feature , ot:NumericFeature ;
dc:title "XLogP" ;
ot:hasSource <http://myhost/algorithm/XLogPDescriptor> ;
= otee:Octanol-water_partition_coefficient_Kow .
```

All columns have explicit and machine readable links to originating algorithms, models or data.
Toxicological data: needs for standards

- Needs for data standards for automatic data integration
  - Example: Carcinogenic Activity

CPDBAS: Carcinogenic Potency Database
http://www.epa.gov/ncct/dsstox/sdf_cpdbas.html#SD

ISSCAN: Chemical Carcinogens Database
http://www.iss.it/ampp/dati/cont.php?id=233&lang=1&tipo=7

FFields

ActivityOutcome
- active
- unspecified/blank
- inactive

Integration

Canc
- 3 = carcinogen;
- 2 = equivocal;
- 1 = noncarcinogen

OpenTox datasets represent endpoint data as features. Features can have arbitrary names (e.g. “Canc”), but are also associated with entries from relevant ontologies.

E.g. (simplified example)
http://opentox.org/echaEndpoints.owl#Carcinogenicity
Our open source AMBIT REST software package implements a large subset of the OpenTox API (data and processing), and is available both as online services and as a downloadable archive.

The dataset web service provides generic means to query, retrieve and upload chemical compounds and aggregate various data.

Built-in heuristics for automatic discovery of 2D chemical structure inconsistencies

Uploading a file with chemical structures and properties makes it automatically available in several formats.

Algorithm & model service: Descriptor calculation (CDK, MOPAC, more), machine learning methods (Weka), expert rules (Toxtree), applicability domain, wrapper for remote services, etc.)

1) HTTP POST a file with chemical structures and properties to an OpenTox dataset service.

   The structures and data are assigned a dataset URI and become available by multiple formats (RDF, Chemical MIME, CSV, Weka ARFF)

2) Assign metadata

   PUT /dataset/{id}/metadata

3) Annotate any of dataset features /dataset/{id}/feature by assigning links to relevant ontologies

   PUT /feature/{id}
OpenTox dataset service (content)

- **ECHA list of pre-registered substances (PRS)**: 143835 entries - names, CAS and EINECS numbers. Using the identifiers, structures were retrieved from
  - Chemical Identifier Resolver (structures retrieved by CAS)
  - ChemIDplus (structures retrieved by CAS)
  - OPSIN (Name to structure converter)
  - ChemDraw (Name to structure converter + partially manual expert inspection)
  - JRC PRS list (subset of ECHA PRS, generated by ACD/Name to structure converter)

- and imported into the dataset service via the OpenTox API

- The following datasets have been also imported (structures + data)
  - EPA DSSTox, ECETOC skin irritation, LLNA skin sensitisation, Bioconcentration factor (BCF) Gold Standard Database, EPA ToxCast, Benchmark Data Set for pKa Prediction of Monoprotic Small Molecules the SMARTS Way, Benchmark Data Set for In Silico Prediction of Ames Mutagenicity, Bursi AMES Toxicity Dataset, EpiSuite data, PubChem, Leadscope* data and Pharmatrope* data
OpenTox dataset service
http://ambit.sourceforge.net/dataset_compare.html
### OpenTox dataset service

http://ambit.sourceforge.net/dataset_compare.html

<table>
<thead>
<tr>
<th>Dataset</th>
<th>ECHA II</th>
<th>ChemIDp</th>
<th>ChemIDr</th>
<th>ChemDra</th>
<th>CPDBAS</th>
<th>DBPCAN</th>
<th>EPAFHM</th>
<th>FDAMDD</th>
<th>HPVCSI</th>
<th>HPVISD</th>
<th>IRISTR</th>
<th>KIERBI</th>
<th>NCTRER</th>
<th>NTPBSI</th>
<th>NC50</th>
</tr>
</thead>
<tbody>
<tr>
<td>ECHA II</td>
<td>143835</td>
<td>55.9%</td>
<td>50.7%</td>
<td>15.7%</td>
<td>0.7%</td>
<td>0.1%</td>
<td>0.4%</td>
<td>0.6%</td>
<td>1.5%</td>
<td>0.4%</td>
<td>0.3%</td>
<td>0.2%</td>
<td>0.1%</td>
<td>1.3%</td>
<td>61</td>
</tr>
<tr>
<td>TOXICST</td>
<td>90.0%</td>
<td>89.4%</td>
<td>89.0%</td>
<td>12.2%</td>
<td>18.7%</td>
<td>0.0%</td>
<td>7.2%</td>
<td>0.3%</td>
<td>9.4%</td>
<td>1.9%</td>
<td>40.3%</td>
<td>7.7%</td>
<td>5.5%</td>
<td>19.4%</td>
<td>16</td>
</tr>
<tr>
<td>TOXICST</td>
<td>90.0%</td>
<td>89.4%</td>
<td>89.0%</td>
<td>12.2%</td>
<td>18.7%</td>
<td>0.0%</td>
<td>7.2%</td>
<td>0.3%</td>
<td>9.4%</td>
<td>1.9%</td>
<td>40.3%</td>
<td>7.7%</td>
<td>5.5%</td>
<td>19.4%</td>
<td>16</td>
</tr>
<tr>
<td>TOXICST</td>
<td>90.0%</td>
<td>89.4%</td>
<td>89.0%</td>
<td>13.2%</td>
<td>18.7%</td>
<td>0.0%</td>
<td>7.2%</td>
<td>0.3%</td>
<td>9.4%</td>
<td>1.9%</td>
<td>40.3%</td>
<td>7.7%</td>
<td>5.5%</td>
<td>19.4%</td>
<td>16</td>
</tr>
<tr>
<td>TOXICST</td>
<td>90.0%</td>
<td>89.4%</td>
<td>89.0%</td>
<td>13.2%</td>
<td>18.7%</td>
<td>0.0%</td>
<td>7.2%</td>
<td>0.3%</td>
<td>9.4%</td>
<td>1.9%</td>
<td>40.3%</td>
<td>7.7%</td>
<td>5.5%</td>
<td>19.4%</td>
<td>16</td>
</tr>
<tr>
<td>ECBPRS</td>
<td>91.6%</td>
<td>88.2%</td>
<td>81.8%</td>
<td>27.5%</td>
<td>1.3%</td>
<td>0.2%</td>
<td>0.7%</td>
<td>1.1%</td>
<td>2.6%</td>
<td>0.8%</td>
<td>0.6%</td>
<td>0.3%</td>
<td>0.2%</td>
<td>2.4%</td>
<td>18</td>
</tr>
<tr>
<td>NAME2ST</td>
<td>100.0%</td>
<td>87.8%</td>
<td>81.3%</td>
<td>28.9%</td>
<td>1.0%</td>
<td>0.2%</td>
<td>0.7%</td>
<td>0.2%</td>
<td>2.6%</td>
<td>0.7%</td>
<td>0.5%</td>
<td>0.3%</td>
<td>0.2%</td>
<td>2.1%</td>
<td>18</td>
</tr>
<tr>
<td>PubChem</td>
<td>1.3%</td>
<td>1.2%</td>
<td>0.7%</td>
<td>0.7%</td>
<td>0.1%</td>
<td>0.0%</td>
<td>0.1%</td>
<td>0.1%</td>
<td>0.2%</td>
<td>0.0%</td>
<td>0.1%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.3%</td>
<td>0.9%</td>
</tr>
</tbody>
</table>

**OpenTox**

Ideaconsult Ltd.
### OpenTox database Quality Assurance

####Automatic Classification

<table>
<thead>
<tr>
<th>Structure</th>
<th>Consensus Quality Label</th>
<th>Structure Quality Label</th>
</tr>
</thead>
<tbody>
<tr>
<td>55-65-2 guanethidine</td>
<td>Majority[1:3]</td>
<td>OK</td>
</tr>
<tr>
<td>55-65-2 guanethidine</td>
<td>Majority[1:3]</td>
<td>MightOK</td>
</tr>
<tr>
<td>55-65-2 guanethidine</td>
<td>Majority[1:3]</td>
<td>MightOK</td>
</tr>
<tr>
<td>55-65-2 guanethidine</td>
<td>Majority[1:3]</td>
<td>ProbablyERROR</td>
</tr>
</tbody>
</table>

####Initial Quality Label Assigned

<table>
<thead>
<tr>
<th>Consensus</th>
<th>Majority</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>OK</strong></td>
<td>ProbablyOK for the structure that belongs to the majority</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Ambiguous</th>
<th>Unknown (multiple sources)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unconfirmed</td>
<td>Unknown (single source)</td>
</tr>
</tbody>
</table>
OpenTox database
Quality labels distribution

- OK
- ProbablyOK
- Unknown
- Probably ERROR
Dataset comparison

Use OpenTox algorithms and models

http://ambit.sourceforge.net/demo_datacomparison.html
Dataset comparison using OpenTox algorithms and models

ECHA PRS

CPDBAS

ToxCast I

PubChem

OpenTox Model REST API

Model name | Algorithm
---|---
Lipinski Rule of Five Text legend

Dataset URI

https://ambit.uni-plovdiv.bg:8443/ambit2/dataset/6489

Predict
Dataset comparison using OpenTox algorithms and models

ECHA PRS

PubChem

OpenTox Model REST API

Model id=22

Model name
MolecularWeight

Algorithm

Dataset
https://ambit.uni-plovdiv.bg:8443/ambit2/dataset/3

Predict
Dataset comparison using OpenTox algorithms and models

**ECHA Preregistration list**

- Alert for SNAr
- Alert for Schiff base...
- Alert for Michael Acceptor
- Alert for Acyl Transfer...
- No DNA binding alerts

**PubChem**

- Alert for SNAr
- Alert for Schiff base...
- Alert for Michael Acceptor
- Alert for Acyl Transfer agent
- No DNA binding alerts

OpenTox Model REST API

<table>
<thead>
<tr>
<th>Model name</th>
<th>Algorithm</th>
<th>Dataset</th>
<th>Independent variables</th>
<th>Dependent</th>
<th>Predicted</th>
</tr>
</thead>
<tbody>
<tr>
<td>ToxTree: Skin sensitisation alerts (M. Cronin)</td>
<td><a href="https://ambit.uni-plovdiv.bg:8443/ambit2/algorith/toxtreeskinsens">https://ambit.uni-plovdiv.bg:8443/ambit2/algorith/toxtreeskinsens</a></td>
<td>Search</td>
<td>🟠 Independent variables</td>
<td>🟠 Dependent</td>
<td>🟠 Predicted</td>
</tr>
</tbody>
</table>

Dataset URI

https://ambit.uni-plovdiv.bg:8443/ambit2/dataset/3

Ideaconsult Ltd.
OpenTox dataset service
(Annotation and query example)

ToxCast Phase I datasets have been uploaded and fields annotated
OpenTox dataset service (an example)

http://ambit.sourceforge.net/api_ontology.html

Retrieve list of features, representing Estrogen receptor gene (http://bio2rdf.org:geneid:2099) related assays from ToxCast Phase 1 datasets:

Once we know which features are representing the Estrogen receptor related studies, the three columns can be collated with the ToxCast dataset and data retrieved in various formats (all via the OpenTox API)
Application, based on OpenTox components [http://toxypredict.org/](http://toxypredict.org/)
All components are implemented as REST web services.

No database schema mentioned so far. The API is database engine and database schema agnostic.

A common API may be implemented by any existing database or web service, without the need to change the underlying implementation.

Anybody can:
- Install and run separate instances of the services, either on Intranet or publicly on the Internet
- Publish own data, own algorithms and models
Conclusions

• It is possible to define an implementation independent protocol (Application Programming Interface), covering the common chemoinformatics functionality:
  – Data access
  – Data upload
  – Chemical structures and data queries
  – Encapsulate various calculation and prediction methods under an uniform interface

• The uniform interface to chemical compounds, data and models:
  – Helps reducing the diversity of processing to the simple paradigm “read data from a web address, perform processing, write to a web address”
  – **Serves as a basis for web mashups, GUI applications, workflow components**
Had the simple components not been standardized, the entire course of the modern economy might look very different.

http://www.wired.com/wired/archive/10.01/standards_pr.html
Acknowledgements

OpenTox project - An Open Source Predictive Toxicology Framework
http://www.opentox.org/
EU FP7 HEALTH-2007-1.3-3 Promotion, development, validation, acceptance and implementation of QSARs for toxicology
Project Reference Number
Health-F5-2008-200787 (2008-2011)

This presentation includes slides made by
Barry Hardy (DC),
Christoph Helma (IST)
Nina Jeliazkova (IDEA)
Olga Tcheremenskaya (ISS),
Stefan Kramer (TUM)
Andreas Karwath (ALU)
Haralambos Sarimveis (NTUA)
Thank you!

Try OpenTox demo applications

Build an application with OpenTox REST Web Services API
http://opentox.org/dev/apis

Download OpenTox software  http://opentox.org/downloads
(services, libraries, applications)

Download AMBIT Implementation of OpenTox API
and launch your OpenTox service
http://ambit.sourceforge.net