OpenTox Overview

CADASTER virtual meeting
10 Nov 2010

presented by Nina Jeliazkova
(Ideaconsult Ltd., Bulgaria)

Uses slides by Barry Hardy (DC), Christoph Helma (IST), Nina Jeliazkova (IDEA), Olga Tcheremenskaya (ISS), Stefan Kramer (TUM), Andreas Karwath (ALU), Haralambos Sarimveis (NTUA)
Motivation

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

• These components have to be re-implemented for every new application

• 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

Ideaconsult Ltd.
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

• Platform independency
• Interoperability for communication with external programs and data sources
• Transparency for scientific and regulatory credibility
• Open for future extensions
Technological choices

• Web services
• Communication through well defined interfaces
• Ontologies for the exchange of knowledge and data
• Use and promote open standards
• Open source components
REpresentational State Transfer (REST)

• What?
  – Architectural style for distributed information systems on the Web
  – Simple interfaces, data transfer via hypertext transfer protocol (HTTP), stateless client/server protocol
    • GET, POST, PUT, DELETE
  – Each resource is addressed by its own web address

• Why?
  – Lightweight approach to web services
  – Simplifies/enables development of distributed and local systems
  – Language independent
## Resources identification

All resources are identified via unique web address, assigned according to the URL templates

<table>
<thead>
<tr>
<th>Component</th>
<th>Description</th>
<th>URL Template (example)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Compound</td>
<td>Representations of chemical compounds</td>
<td><code>http://host:port/compound/{compoundid}</code></td>
</tr>
<tr>
<td>Feature</td>
<td>Properties and identifiers</td>
<td><code>http://host:port/feature/{featureid}</code></td>
</tr>
<tr>
<td>Dataset</td>
<td>Encapsulates set of chemical compounds and their property values</td>
<td><code>http://host:port/dataset/{datasetid}</code></td>
</tr>
<tr>
<td>Model</td>
<td>OpenTox model services</td>
<td><code>http://host:port/model/{modelid}</code></td>
</tr>
<tr>
<td>Algorithm</td>
<td>OpenTox algorithm services</td>
<td><code>http://host:port/algorithm/{algorithmid}</code></td>
</tr>
</tbody>
</table>
| Validation, Report | A validation corresponds to the validation of a model on a test dataset.    | `http://host:port/validation/{validationid}`  
                      |                                                                             | `http://host:port/report/{reportid}` |
| Task               | Asynchronous jobs are handled via an intermediate Task resource. A resource, submitting an asynchronous job should return the URI of the task. | `http://host:port/task/{taskid}` |
| Ontology service   | Provides storage and SPARQL search functionality for objects, defined in OpenTox services and relevant ontologies | `http://host:port/ontology` |
| Authentication and authorisation | Granting access to protected resources for authorised users | `http://host:port/opensso`  
                      |                                                                             | `http://host:port/opensso-pol` |
OpenTox API (Application Programming Interface)

• The way applications talk to each other
• The way developers talk to applications
• http://opentox.org/dev/apis/api-1.1
All components are implemented as REST web services. There could be multiple implementations of same type of components. (Subset of) services could be hosted by the same provider, or by multiple providers on separate locations.

<table>
<thead>
<tr>
<th>Partner No.</th>
<th>Compound</th>
<th>Dataset</th>
<th>Feature</th>
<th>Algorithm (processing)</th>
<th>Algorithm (model)</th>
<th>Model</th>
<th>Validation</th>
<th>Report</th>
<th>Task</th>
<th>Authentication and Authorisation service</th>
<th>Ontology service</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>Y</td>
<td>Y</td>
<td></td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td></td>
<td></td>
<td>Y</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>Y</td>
<td>Y</td>
<td></td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td></td>
<td></td>
<td>Y</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Y</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Y</td>
<td></td>
<td></td>
<td>Y</td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Y</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
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, ...)

Algorithm

- GET
- POST
- PUT
- DELETE
Algorithms: Descriptors and feature selection

• **Descriptor calculation**: services based on
  – OpenBabel
  – Joelib2
  – CDK
  – Multi-level neighborhood of atoms (MNA)
  – Substructure/fragment generation
  – MOPAC

• **Feature selection**
  – Services for feature selection based on **information gain**
  – Service for feature selection based on **Chi^2** statistics
  – PCA
  – **Filter** pipeline for **preprocessing**: combining approaches for handling missing values, feature selection, ...
Algorithms

Classification/SAR

- Simple baseline: \textit{k-Nearest neighbor}
- Machine learning algorithms
  - Decision trees (J48)
  - Support Vector machines (SVM)
- Probabilistic / graphical models
  - Bayesian network
  - Gaussian process regression

Regression /QSAR

- Simple baseline: \textit{k-Nearest neighbor}
- Classical statistical algorithms
  - Multiple linear regression (MLR)
  - Partial Least squares (PLS)
- Machine Learning algorithms
  - Model trees (M5)
  - Support vector regression
- Probabilistic/graphical models:

Rule based

- Toxtree
### Datasets

upload, read, modify, delete, search

Uniform access to data: described by W3C RDF (Resource Description framework)

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><a href="http://myhost.com/compound/413">http://myhost.com/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</td>
<td>3.31</td>
<td>225.3</td>
<td>YES</td>
<td>3.123</td>
</tr>
<tr>
<td><a href="http://myhost.com/compound/44497">http://myhost.com/compound/44497</a></td>
<td>4-acetamidofluorene</td>
<td>O=C(Nc3c2c1ccc(c)cc1Cc2ccc3)C</td>
<td>1</td>
<td>NP</td>
<td>223.28</td>
<td>YES</td>
<td></td>
</tr>
</tbody>
</table>

http://myhost.com/feature/21573

```
af:21573 a ot:Feature , ot:NumericFeature ;
  dc:creator "http://www.epa.gov/NCCT/dsstox/sdf_isscan_external.html" ;
  dc:title "Cancer" ;
  ot:hasSource <http://myhost.com/algorithm/Benigni+%2F+Bossa+rulebase+%28for+mutagenicity+and+carcinogenicity%29> ;
    dc:creator "http://www.blueobelisk.org/ontologies/chemoinformatics/algorithms/#xlogP" ;
    dc:title "XLogP" ;
    ot:hasSource <http://myhost.com/algorithm/org.openscience.cdk.qsar.descriptors.molecular.XLogPDescriptor> ;
    otee:Octanol-water_partition_coefficient_Kow .
```
Datasets can be easily merged, compared, and calculations reproduced, regardless of their physical place.

The dataset service offers property, compound, substructure and similarity searches via uniform OpenTox Application Programming Interface (API).
Ontologies

• What?
  – Formal, shared conceptualization of a domain

• Why?
  – Distributed services need to be able to “talk to each other”, e.g. have a common understanding of endpoints, properties, methods, etc.
  – Allows us to integrate existing knowledge from many related domains
Ontologies

• Standards: RDF (OWL-DL) as representation language and SPARQL as query language

• There are many ongoing biological ontology projects

• Our strategy: use existing work and standards wherever possible

• However, there are new ontology needs for OpenTox applications, e.g. for algorithms, toxicological endpoints

OpenTox
Ontology Working Group
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#SDFFields

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

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
Mapping of the ISSSCAN entry - ToxML xsd scheme

CAS 67-66-3
Substance ID 4
ChemName Chloroform
Synonyms Formyl trichloride; methane trichloride; methenyl trichloride; Methyl trichloride; R 20; r 20 (refrigerant); Refrigerant R20; trichloroform; Trichloromethane

SAL negative
Canc positive
Reference CPDB

TD50_Rat 262
TD50_Mouse 90.3
Rat_Male_Canc positive
Rat_Male_NTP ND
Rat_Female_Canc positive
Rat_Female_NTP ND
Mouse_Male_Canc positive
Mouse_Male_NTP ND
Mouse_Female_Canc positive
Mouse_Female_NTP ND

MolWeight 119.38
Formula CHCl3
SDF file Connection table
SMILES CIC(Cl)Cl
OpenTox Toxicological Endpoint Ontology

• Why we need an ontology?
  • Distributed services need to be able to “talk to each other”, i.e. have a common understanding of endpoints, any type of property, methods, etc

• Methodology
  • Starting from 5 toxicological endpoints
  • following OBO Foundry principles

Other freely available resources: DSSTOX, GoReni (ITEM), etc

Protégé, free open source OWL (Web Ontology Language) editor

Re-use of pieces or terms defined in neighboring ontologies (OWL and OBO)
Toxicological Ontology: graphical representation
## Linked resources:

### Compound, Algorithm, Model, Dataset, Features

<table>
<thead>
<tr>
<th>Dataset Resource</th>
<th>Descriptor resource</th>
<th>Assay resource</th>
<th>Chemical compound</th>
</tr>
</thead>
</table>

### Regression, Classification, Quantum Chemistry, Descriptors, etc.
- Blue Obelisk algorithms ontology
- OpenTox algorithm types ontology
- Toxicology related ontologies

---

**Ideaconsult Ltd.**
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.
- Models services provide facilities to inspect, store and delete models. Every model is identified by unique web address.
Uniform approach to models creation

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

Feature
GET
POST
PUT
DELETE

Dataset
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
Uniform approach to data processing (e.g. Descriptors calculation)

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

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/{descriptorX}
Build a predictive model

Create a model, Run calculations with dataset http://host1/dataset/id

Dataset service

Structures, descriptors, endpoints

HTTP POST

Returns the model URL http://host1/model/id

Algorithm service

Regression
Classification
Quantum Chemistry
Descriptors, etc.,

Validation service

/valuation/id

Model service

/http POST

/model/{id}

Http POST

Ontology service

Published models, Algorithms, Ontologies, metadata

OpenTox
Uniform approach to model prediction

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

http://myhost.com/dataset/id1

http://myhost.com/model/predictivemodel1

http://myhost.com/dataset/results1
Apply predictive models

/model/{id} HTTP POST

SPARQL

HTTP POST

Ontology service

Published models, Algorithms, Ontologies, metadata

Retrieve available endpoints and model URLs, e.g. http://host1/model/id

HTTP POST

Returns the results dataset URL http://host/dataset/id

Apply the model http://host1/model/id to dataset http://host2/dataset/id

Ontology service

/ model/{id}
Validation

Algorithm Validation
• common best practices such as k-fold cross validation, leave-one-out, scrambling

QSAR Validation (Model Validation)
• OECD Principles
  www.oecd.org/dataoecd/33/37/37849783.pdf
• QSAR Model Reporting Format (QMRF)
  qsardb.jrc.it/qmrf/help.html
• QSAR Prediction Reporting Format (QPRF)
  ecb.jrc.it/qsar/qsar-tools/qrf/QPRF_version_1.1.pdf

Reports

REACH
• Guidance on Information Requirements and Chemical Safety Assessment

Part F
• Chemicals Safety Report
• Appendix Part F
  guidance.echa.europa.eu/guidance_en.html
Goodness-of-fit, robustness and predictivity

- OpenTox is developing unified and objective validation routines for model and algorithm developers and for external (Q)SAR programs, including procedures for validation with artificial test sets
  - (e.g. n-fold cross-validation, leave-one-out, simple training/test set splits, bootstrapping, Y-scrambling).
  - Validation services are completely independent of algorithm and model services

- An important goal is to integrate
  - statistical tests for the comparison of (Q)SAR models under consideration,
  - a versioned database to store validation results and their history,
  - and tools for the inspection of the toxicological plausibility of (Q)SAR predictions.
Implemented validation algorithms

**Classification methods**

- Number of correctly classified instances
- Number of incorrectly classified instances
- `weighted_area_under_roc`
- `f_measure`
- `num_false_positives`, `negatives`
- `num_true_positives`, `negatives`
- `sensitivity`
- `specificity`
- Classification confusion matrix

**Regression methods**

- `root_mean_squared_error`
- `mean_absolute_error`
- `sum_squared_error`
- `r_square`
- `correlation_coefficient`

Uniform approach to models validation and report generation

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

http://myhost.com/dataset/trainingset1

http://myhost.com/dataset/predictedresults1

http://myhost.com/validation

Validation
GET
POST
PUT
DELETE

Dataset
GET
POST
PUT
DELETE

= +

Model generating predictions
Model
GET
POST
PUT
DELETE

Validation report
Report
GET
POST
PUT
DELETE

http://myhost.com/report/1

http://myhost.com/model/predictivemodel1

http://myhost.com/dataset/trainingset1

http://myhost.com/dataset/predictedresults1

http://myhost.com/validation

OpenTox
Ideaconsult Ltd.
A. The predictive model itself provides estimation of applicability domain
   • Lazar

B. Applicability domain is estimated by a procedure, separate from the predictive model
   • PCA ranges
   • Euclidean distance
   • Cityblock distance
   • Mahalanobis distance
   • Nonparametric density estimation
   • Leverage
   • Fingerprints, Tanimoto distance
• QMRF and QPRF
  – What are they?
    • Harmonized templates for summarizing and reporting key information on (Q)SAR models and predictions, generated by these models
  – Why it is important in OpenTox?
    • QMRF and QPRF are expected to be the communication tool between industry and the authorities under REACH
User perspective

ToxPredict → QSAR Model A → Compound → QMRF Editor → QMRF report

ToxCreate → QSAR Model B → Compound → QPRF Editor → QPRF report

OpenTox

Ideacconsult Ltd.
Creating reports

- Compound
- QSAR Model
- REACH report Web service
- QPRF report
- QMRF report
- Validation web service
- Model web service
- Dataset web service
Storing and editing reports

1) QMRF editor OpenTox version
By Albert-Ludwigs-Universitat-Freiburg (ALU-FR) Germany

- Automatically populate relevant fields based on information, available in (distributed) algorithm, model, data, validation and reporting services
- Users can edit to add missing information
- Reports can be downloaded, uploaded, deleted from/to OpenTox reporting service

2) QPRF editor (Q-Edit)
http://opentox.ntua.gr/Q-edit/dist/launch.jnlp
Live demo by Pantelis Sopasakis, National Technical University of Athens
What can you do with OpenTox

• Build simple applications, based on existing algorithms, methods, and data
• Distributed applications, integrating wide range of data and methods
• Examples:
  – ToxCreat (web application), ToxPredict (web application), QMRFEditor (Java web start), QPRF Editor (Java web start)
  – More under development
ToxCreate creates models from user supplied datasets. Developed and hosted by IST (Christoph Helma). Uses OpenTox algorithm, model, compound, dataset and validation services.
ToxCreate (behind the scenes)
ToxPredict uses existing OpenTox models to estimate chemical compound properties. Developed and hosted by IdeaConsult.
ToxPredict (behind the scenes)

**ToxPredict Web Application**

Find structure by name, registry number, SMILES, InChI, structure, substructure, similarity...

**OT Dataset API HTTP GET**

text/uri-list, application/rdf+xml, chemical/x-daylight-smiles chemical/x-mdl-sdfile,...

**OT Dataset Service**

Here is the list of structures as URI links, RDF, MOL or SMILES.
What prediction models are available? Is there a model for endpoint X?

HTTP GET SPARQL query

Here is the list of model URIs and related endpoints and algorithms in SPARQL format.
Q-Edit demo

http://opentox.ntua.gr/Q-edit/dist/launch.jnlp

Q-Edit uses existing OpenTox models and data services to create QPRF report. Developed and hosted by NTUA.
What can you do with OpenTox

• Integration into
  – Workflow systems: Taverna, Knime, Pipeline Pilot
  – Applications: Bioclipse

• Run your own instances of (subset) OpenTox services
  – Expose your new predictive algorithm as OpenTox algorithm or model service
  – Publish your data as OpenTox dataset

• Query ontology services to find out
  – datasets or models (possible remote)
  – for particular endpoint, type of algorithm, etc.
Linked resources: Compound, Algorithm, Model, Dataset, Features
Summary

- OpenTox is a framework for predictive toxicology
- Designed for language independence, transparency and extensibility
- Implemented as open source REST web services
- Exchange of data and knowledge with ontologies (RDF, OWL-DL)
- OpenTox components: Compound, Feature, Dataset, Algorithm, Model, Validation, Report, Task, Authentication and Authorisation
- Documentation: www.opentox.org/dev
Thank you!

EC FP7 CADASTER
http://www.cadaster.eu

EC FP7 OpenTox
http://www.opentox.org