Existing in silico toxicology software for the OpenTox project

C. Helma <helma@in-silico.de>

August 22, 2008

1 Introduction

During the last year we have transfered most of our developments into a framework (tentatively called OpenTox), that can be used as the basis for the OpenTox project.

1.1 Features:

- Modular plugin architecture
- Access to the most important open source bio/chemoinformatics, statistical and data mining packages
  - Openbabel
  - CDK
  - R (and BioConductor)
  - Toxtree
- FUGE (MAGE, ...) compliant database scheme for biological and computational experiments
- GUIs for lazar, Sens-it-iv and a few in-house applications
- Easy integration of additional packages and plugins
- Support of other programming languages (e.g. Java, Python, C, ...)
- Public repositories for source code and public data
- Private repositories for confidential data
- Automated installation procedures

- Tested on various Linux flavours (Debian, Ubuntu, RH Enterprise)
- Should run on other Unix systems (e.g. OS X, BSD, Cygwin, ...)
- Windows port possible

1.2 Technical details:

Programming Language Ruby

Framework Ruby on Rails

Database database independent (we use sqlite3)

Version control Git

In the list below you can find our (already available) contributions to the delivarables, maybe we can use this list also as a template for the contributions from other partners.

2 Work package 1: Framework design (WP Leader: IST)

2.1 Objectives:

To define the requirements and specifications of the OpenTox framework, to evaluate the implemented prototypes and to contribute to standards that are relevant for (Q)SAR model development.

2.2 Deliverables:

2.2.1 Month 6: Initial requirements, standards and APIs (Responsible: IST)

1. Evaluation of common use-cases for toxicological end users, data providers, (Q)SAR model developers and algorithm developers
2. Evaluation of current standards that are relevant for the OpenTox framework
   - FUGE (MAGE,...) as standard for -omics experiments
   - OECD criterea
   - QMRF
   - to be completed
3. Initial specification of requirements and standards for the OpenTox framework
   - Plugin architecture for the integration of external algorithms
4. Definition of APIs for the database interface

- FUGE (MAGE, ...) compliant database scheme available

5. Definition of APIs for the algorithm interface
   - Plugin architecture for the integration of external algorithms available
6. Definition of APIs for the validation interface
   - Exchange format for validation results available

2.2.2 Month 12: Initial GUI design (Responsible: DG)

7. Design of a user interface for toxicological risk assessors
   - Several example GUIs available
8. Design of a user interface for model developers
   - Automated lazar model development and validation

2.2.3 Month 24: Prototype evaluation, improved API and interface designs (Responsible: IST)

9. Evaluation of the prototype implementation
10. Improved design of APIs and interfaces

2.2.4 Month 36: Evaluation of the final implementation (Responsible: DG)

11. Evaluation of the final implementation

3 Work package 2: Framework implementation (WP Leader: IDEA)

3.1 Objectives:

To provide the basic infrastructure for the project and common functionality for the other parts of the project. This will include an easy-to-use graphical user interface (GUI) for toxicological experts, that accesses (Q)SAR models provided by the consortium, a GUI for (Q)SAR model developers with facilities for data import, facilities to retrieve rationales and supporting information for (Q)SAR predictions and a plug-in system for the integration of third party programs and external model developments.

3.2 Deliverables:

3.2.1 Month 6: Project repository and website (Responsible: IDEA)

1. Establishment of a common project repository with version control and project management tools (e.g. mailing lists for users and developers, bug and feature request trackers)
   - Project management site (www.opentox.org) available
   - Version control system available

3.2.2 Month 18: Prototype framework (Responsible: ALU-FR)

2. Implementation of a prototype framework with APIs for work packages 3 (databases), 4 (algorithms) and 5 (validation)
   - Prototype framework available
3. Implementation of a prototype GUI for toxicological risk assessors
   - Several GUIs available
4. Implementation of a prototype GUI for model developers
   - Automated model development and validation from the command line (do we need a GUI?)

3.2.3 Month 24: Prototype server (Responsible: TUM)

5. Implementation of a prototype public access server
   - Public servers for lazar and Sens-it-iv available

3.2.4 Month 33: Final framework implementation (Responsible: IDEA)

6. Final implementation of the framework according to WP 1 specifications after prototype evaluation
7. Final implementation of GUIs according to WP 1 specifications after prototype evaluation with installers for the major operating systems

4 Work package 3: Toxicity databases (WP Leader: ISS)

4.1 Objectives:

To provide a database with data for the training and validation of toxicity (Q)SAR models. The initial database will be built upon the AMBIT database (provided by IDEA). Within this project we will populate it with data that is provided by consortium

members (e.g. ISS ISSCAN, ITEM REPROTOX, ITEM REPDOSE, IDEA AMBIT, IBMC TERA, EPA DSSTOX, FDA GENREPCAR) and enrich them systematically with data from other sources. We will additionally seek to collaborate with other toxicity-related projects for unifying data storage and maintenance. The final version will have facilities to import confidential and commercial data, quality assurance procedures and algorithms for data aggregation. All public data incorporated into the OpenTox database will be available to the public.

4.2 Deliverables:

4.2.1 Month 6: Initial vocabularies and ontologies for toxicological data (Responsible: ISS)

1. Definition of vocabularies and ontologies for toxicological and chemical data
   - partially available

4.2.2 Month 12: Prototype database with initial data (Responsible: IDEA)

2. Definition of requirements for data inclusion
3. Identification of suitable data sources
4. Implementation of a prototype database according to the requirements from WP1
   - available
5. Import of initial data into the prototype database
   - DSSTox data available

4.2.3 Month 21: Tools for the integration of confidential data (Responsible: IST)

6. Implementation of tools for the integration of confidential data
   - available

4.2.4 Month 33: Redesigned database with additional content (Responsible: ISS)

7. Redesign and implementation of the database according to WP1 specifications after prototype evaluation
8. Modification of the database content according to WP 1 specifications after prototype evaluation

5 Work package 4: (Q)SAR algorithms (WP Leader: TUM)

5.1 Objectives:

This work package will implement a framework for the integration of state-of-the-art statistical, data mining and chemoinformatics algorithms into the OpenTox project. New algorithms will be developed and implemented according to the requests of WP1 (Framework design), WP3 (Toxicity databases) and after a weak-point analysis of currently available techniques.

5.2 Deliverables:

5.2.1 Month 6: Selection of algorithms for the prototype (Responsible: TUM)

5.2.2 Month 18: Initial prototype of (Q)SAR algorithms (Responsible: NTUA)

1. Integration and implementation of algorithms for the generation of structural features (e.g. paths, trees, subgraphs, multiple neighborhood of atoms, pharmacophore descriptors)
   - access to Openbabel and CDK available
   - path generation available
   - tree and subgraph generation available soon
2. Integration and implementation of algorithms for the calculation of chemical properties (e.g. logP, surface parameters, reactivity indices)
   - access to Openbabel, CDK and Dragon (requires license) available
3. Implementation of algorithms for the retrieval of bioassay data from sources like PubChem
   - partially available
4. Integration and implementation of algorithms for feature selection (e.g. statistical filters, closed sets, principal component analysis)
   - chi-square and KS filter
   - easy implementation of additional algorithms with R
5. Integration and implementation of algorithms for classification and regression (e.g. k-nearest neighbors, linear regression, neural nets, support vector machines, decision and regression trees)
   - access to all R packages
   - lazar program
   - Weka access possible

6. Integration and implementation of algorithms for the aggregation of predictions (rule based and data driven)
    - example in-house application
7. Integration and implementation of supporting algorithms (e.g. applicability domain estimation, various measures of chemical similarity, structure and property based searches)
    - AD estimation for lazar
8. Implementation and a plugin system for external (commercial) programs These tasks will run in parallel and the progress will be monitored by WP1 (Framework design). Priorities for the implementation and development of new algorithms will be set in collaboration with WP1.
    - example plugin for Dragon

5.2.3 Month 33: Final version of (Q)SAR algorithms (Responsible: TUM)

6 Work package 5: (Q)SAR model validation (WP Leader: ALU-FR)

6.1 Objectives:

To provide tools for the unbiased evaluation of (Q)SAR models, regardless of the underlying algorithms. The automated creation of validation reports that are compliant with international standards (e.g. OECD guidelines, ECB QSAR model reporting format) and facilities for the toxicological interpretation of validation results will be provided for an independent external review of (Q)SAR models. Facilities for validation against confidential data will be provided for the same purpose. We will start with existing validation routines from project members (e.g. IST lazar, IDEA AMBIT, NTUA Y-scrambling) and sequentially add features that are requested by WP1 (Framework design).

6.2 Deliverables:

6.2.1 Month 18: Prototype validation routines (Responsible: ALU-FR)

1. Implementation of validation methods based on artificial test sets (e.g. crossvalidation, leave-one-out, simple training/test set splits)
    - Automated loo and external validation for lazar models

6.2.2 Month 24: Report generation facilities (Responsible: IBMC)

2. Implementation of standards-compliant validation report generation facilities
    - Automated lazar validation reports according to the OECD Guidelines

6.2.3 Month 30: Validation facilities for confidential data (Responsible: NTUA)

3. Implementation of validation techniques for confidential data

6.2.4 Month 33: Final implementation of validation routines (Responsible: ALUFR)

4. Revision of the implemented techniques according to WP1 evaluation