

CADASTER

Case studies on the Development and Application of in-Silico Techniques for Environmental hazard and Risk assessment

Grant agreement no.: 212668

Collaborative Project

Sub-Priority ENV2007 3.3.1.1: In-silico techniques for hazard-, safety-, and environmental risk-assessment

WP5 QSPR-THESAURUS: Web site and standalone tools for dissemination of project results

Task 5.4 Public QSPR-THESAURUS site
--

Due date of deliverable: December 31, 2010

Actual submission date: December 28, 2010

Start date of project: 1 January 2009

Duration: 4 years

Corresponding authors of document: *Igor V. Tetko*¹, *Stefan Brandmaier*¹, *Wolfram Teetz*¹

1. Institute of Bioinformatics and Systems Biology, Helmholtz Zentrum Muenchen, German Research Center for Environmental Health, Ingolstaedter Landstrasse 1, D-85764 Neuherberg, Germany (i.tetko@helmholtz-muenchen.de)

CADASTER

Project co-funded by the EU Commission within the Seventh Framework Programme		
Dissemination Level		
PU	Public	X
RE	Restricted to a group specified by the consortium (including the Commission Services)	
CO	Confidential, only for members of the consortium (including the Commission Services)	

WP 5: QSPR-THESAURUS: Web site and standalone tools for dissemination of project Results

Work Package Leader: Igor Tetko (HMGU, partner 6)

Task 5.3 - Methods to Estimate the Applicability Domain and Experimental Design (WWW-based tools + report) (Deliverable 5.3)

Overview

The goal of this task is to make the database and tools developed within the CADASTER project publicly available to web users. The developed web site <http://www.qspr-thesaurus.eu> is based on the OCHEM <http://ochem.eu> platform and includes a number of features that were specifically developed for the CADASTER project. This report contains a brief general overview of the main features of the database and the additional tools that were developed and made available to outside-users. The overview is supplemented with displays of specific screens encountered by users when using the publicly available database and additional functionalities.

Figure 1. The entry page of the QSPR-THESAURUS database.

Revision 2010-12-20 16:54:47 by 146.107.19.12 checked in on null. Built from null on null

Safari 5 on Mac - Not tested
Welcome, Guest! Logout

CADASTER
Case studies on the Development and Application of in-Silico Techniques for Environmental hazard and Risk assessment

Home Database Models A+ a-

Models applier browser

Step 1. Select a model from the list

Filter by model name: and property name: or by article id: Models visibility: [refresh]

1 - 10 of 10

MP_TAZs_remap , published by simona	predicts Melting Point using TAZ_MP_Training set (56)	MLRA	2010-11-12			
LogVP_TAZs_remap , published by simona	predicts Vapor Pressure using TAZ_LogVP_Training set (33)	MLRA	2010-11-12			
LogWS_TAZ , published by simona	predicts Aqueous Solubility using TAZ_LogWS_Training set (49)	MLRA	2010-11-11			
LogWS_PFCs , published by simona	predicts Aqueous Solubility using PFC_LogWS_Training set (20)	MLRA	2010-11-11			
LogVP_PFCs_remap(p4-2C) , published by simona	predicts Vapor Pressure using PFC_LogVP_Training set (35)	MLRA	2010-11-12			
Papa_PDE_Henry new , published by wolfram	predicts Henry using Ester_Papa_LogH (7)	MLRA	2010-07-05			
Papa_PDE_TM new , published by wolfram	predicts Melting Point using Ester_Papa_TM (25)	MLRA	2010-07-05			
Papa_PDE_Log_PL new , published by wolfram	predicts Vapor Pressure using Ester_Papa_Log_PL (34)	MLRA	2010-07-05			
Papa_PDE_Log_KOW new , published by wolfram	predicts logPow using Ester_Papa_LogKOW (20)	MLRA	2010-07-05			
Papa_PDE_LogKOA New , published by wolfram	predicts LogKoa using Ester_Papa_LogKOA (30)	MLRA	2010-07-05			

1 - 10 of 10

Next>>

Figure 2. Publicly available models developed within the CADASTER FP7 project.

The OCHEM platform was provided as background information to the CADASTER project and it was customized to the project requirements within the framework of the project. The OCHEM itself comprises more than 100,000 lines of Java, C++, and shell script code and it was developed by a team of 5 people over three years. The majority of data collected within CADASTER project were hidden and password protected. Starting as of end of December 2010, these data are released to the public at large.

Activities performed

The main development within this task included customization of the OCHEM to the demands of the CADASTER project and addition of the required functionality. Below, we briefly summarize the main activities with a focus on the web development. Some of the activities were more deeply covered within the respective CADASTER WPs.

Upload of models developed within the CADASTER project

The majority of models contributed within the project are linear ones and are based on linear regression or partial least squares (PLS) approaches. We have developed a module, which accepts as input an Excel file with descriptors and coefficients of regression equations. Once the model is uploaded, it is used to predict molecules from the training and validation set(s), and several statistical parameters of the model are calculated for these sets of chemicals. The tool was tested by the project participants and a number of models developed by the QSAR Research Unit in Environmental Chemistry and Ecotoxicology of the University of Insubria (partner no 3 of the CADASTER project) were uploaded. Help on how to introduce models is provided.

While models developed using 2D descriptors could be easily reproduced, there are difficulties to do this for 3D-based model, since calculated descriptors dramatically depend on the conformation of molecules. This problem was addressed within the next task.

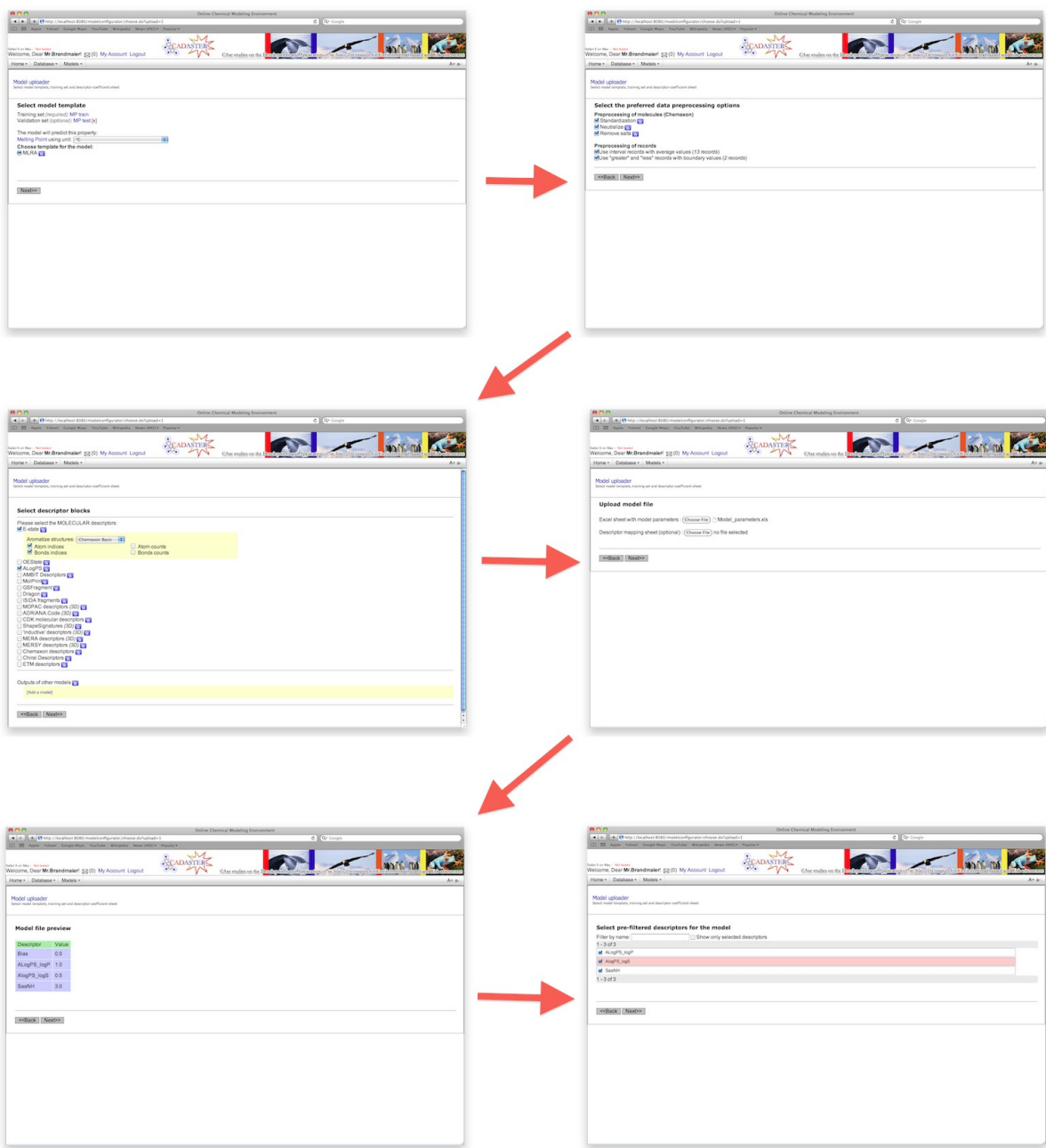


Figure 3. Workflow for model upload, illustrating the Sequential selection of required parameters.

Integration of 3D structure optimisation using MOPAC

A database for on-line optimisation of molecules using the MOPAC 7 program was developed. The MOPAC 7 program is available as <http://www.cadaster.eu/mopac>. Users can submit a single molecule or a batch of molecules and they can download structures optimised by MOPAC. If molecules are not yet optimised, they will be automatically submitted for analysis and calculations and the user can retrieve the results later, once the calculations are finished. It is also possible to submit and retrieve molecules using the web services. The user can request the minimum energy conformation (considering or ignoring stereochemistry) as well as request the optimized conformation starting from

the given conformation. The calculations are based on a BOINC server. The structures optimized using this tool are used to calculate 3D descriptors with the CADASTER.

Records Export

1 - 5 of 14

Items per page:

Page of 3

ID:86911
Source:original

ID:86912
Source:balloon

ID:87780
Source:original

TotalEnergy: -10441.63752.....

Figure 4. Results with 3D optimization of molecules within the MOPAC/BOINC.

Similarity search

The integration of AMBIT substructure and similarity search was performed. All molecules from the CADASTER database were uploaded to the AMBIT database. By using REST webservice technologies, the AMBIT functionality for similarity and sub-structure search was wrapped and embedded in the database. Both similarity search and substructure search are based on the SMARTS patterns. Users have the possibility to submit the requested substructures either directly as SMILES, or in several other formats. Furthermore, molecules can be also drawn in the JME Molecule editor.

Integration of QMRF



The (Q)SAR Model Review Format (QMRF) <http://www.cadaster.eu/QMRF> was integrated in the CADASTER database. The users can contribute the description of their models as

required by the OECD principles directly from the web site of the project. In addition to the statistical parameters available on the OCHEM web site, the users can provide several predefined parameters (R^2 , RMSE, Q^2_{Loo} , etc.) to document the output of both the internal and the external validation of their models. Furthermore the possibility to add any kind of graphical representation (graphs, tables, images) to support or explain the validation results as well as to describe the applicability domain of the models is provided. After filling the QMRF form and adding a new model description to the database, users receive an ID that can be used to publish the report on the <http://cadaster.eu/database>. The data introduced to QMRF about the model's validation performed in the publications are displayed as an embedded tab of the analysed model.

The graphical representation, as well as statistical information facilitates evaluation of models and allows better estimation of the quality of prediction of new compounds.

www.CADASTER.eu

Home REACH CHALLENGE Database Partners WP Meetings News Forum Login

Case studies on the Development and Application of in-Silico Techniques for Environmental hazard and Risk assessment

Home

QSAR Model Reporting Format

Displaying a model description

1. QSAR identifier

[-] Expand

1.1 QSAR identifier (title) Help

Polybrominated Diphenyl Ethers, LogKOA

1.2 Other related models Help

Polybrominated Diphenyl Ethers, LogKOW Polybrominated Diphenyl Ethers, Log_PL Polybrominated Diphenyl Ethers, TM Polybrominated Diphenyl Ethers, Henry

1.3 Software coding the model Help

	Name:	URL:	Description:	Contact:
1	MOBY- DIGS (Ver. 1.0)	http://www.talete.mi.i		

2. General information

[+] Expand

3. Defining the endpoint - OECD Principle 1

[+] Expand

Figure 5. CADASTER implementation of the QMRF form.

Applicability domain

Methods to describe the applicability domain as developed by IDEA (partner no. 7 of the CADASTER project) were incorporated and made available within the database: IDEA AD. The implementation of IDEA AD is based on applicability domain algorithms, as indicated in Table 1. They are implemented as Web REST services, compliant to the OpenTox Application Programming Interface 1.1 [4,5].

Table 1.

Name	Web service location	Type
Leverage	http://apps.ideaconsult.net:8080/ambit2/algorithm/leverage	Descriptor based
Descriptor ranges in transformed PCA space	http://apps.ideaconsult.net:8080/ambit2/algorithm/pcaRanges	Descriptor based
Euclidean distance	http://apps.ideaconsult.net:8080/ambit2/algorithm/distanceEuclidean	Descriptor based
City block distance	http://apps.ideaconsult.net:8080/ambit2/algorithm/distanceCityBlock	Descriptor based
Mahalanobis distance	http://apps.ideaconsult.net:8080/ambit2/algorithm/distanceMahalanobis	Descriptor based
Non parametric density estimation	http://apps.ideaconsult.net:8080/ambit2/algorithm/nparamdensity	Descriptor based
Hashed fingerprints Tanimoto distance	http://apps.ideaconsult.net:8080/ambit2/algorithm/ftpanimoto	Structure based
Hashed fingerprints number of missing bits	http://apps.ideaconsult.net:8080/ambit2/algorithm/ftpmissingfragments	Structure based

The algorithm AD service is used to create the AD model, specific to the training set. The model then becomes available as an OpenTox Model service, with unique URL, e.g. <http://apps.ideaconsult.net:8080/ambit2/model/2443>. The AD model can be estimated by using HTTP POST operation with dataset_uri parameter, pointing to the dataset to be estimated. The result of the POST operation is an URL of a new Dataset, containing the result.

Implementation of experimental design methods

The approaches for optimal design were implemented on the web site of the project. This included implementation of the traditional D-Optimal design and newly proposed within the project PLS-optimal design. For both approaches, preliminary measured values can be taken into consideration. These preliminary values are obligatory for the PLS-based approach and can be either measurements from a literature research or newly measured values from a lab. To use this feature, users should create a molecule basket with measured values of the property, for which the experimental design should be performed. In case of the D-Optimal design, the maximum diversity set is selected in the descriptor space (represented by principal components) while in case of the PLS-Optimal design, the selection is based on latent variables derived from a PLS model. The results of this experimental design are depicted as a clickable 2D-plot using principal components or the latent variables. Compounds are discriminated into three different groups, represented by different colours. The first group contains all compounds for which a preliminary

measured value was available, the second one contains all molecules the have been selected for testing and the last group contains all other molecules.

The figure displays two screenshots from the QSPR Thesaurus website. The top screenshot shows the 'Experimental design' interface. It includes a navigation menu with 'Home', 'Database', and 'Models'. The main content area is titled 'Experimental design' and contains the following information: 'Molecules to be predicted (required): Cadaster ED2', 'Experimental set (required for PLS only): ExpD', 'Method: D-Optimal design PLS-Optimal design', and 'Number of compounds to select: 10'. A 'Next>>' button is located at the bottom. A red arrow points from this screenshot to the second one below. The second screenshot shows the 'Model editor' interface. It features a navigation menu with 'Home', 'Database', and 'Models'. The main content area is titled 'Model editor' and contains 'Experimental design - results'. This section includes a scatter plot with data points colored in green, red, and blue, and a chemical structure of a molecule: CC(=O)C(C)CS(=O)C.

Figure 6. Experimental design workflow integrated in QSPR Thesaurus site.

Descriptor mapping

The different software tools (including different versions) calculate different descriptors. The absence of descriptors may not allow the reimplementing of models. In order to solve this problem the descriptors mapping tool was developed. It allows uploading the calculated descriptors together with molecules and calculates the best mapping between the uploaded descriptors and the descriptors available in the database. The user can verify the quality of calculated mapping and use it instead of the original descriptors. In case there is 100% mapping of old and new descriptors, the mapping can be immediately

accepted. In other cases, some additional analysis can be used to decide whether such mapping is adequate or not.

Conclusion and activities foreseen

The goals and expected developments within this WP were fully achieved. The site is functional and has already been used to host several models developed within the project by CADASTER partners.

Several suggestions from project partners are under implementation now. This includes:

- 1) calculation of models by providing external descriptors
- 2) storage of predicted values

Both these requirements will allow publishing of models, for which descriptors are absent in the database. The external descriptors can be those calculated by programs absent at the web database (i.e., those that could not be remapped) of the project as well as experimental physico-chemical properties of molecules, in case some models are based on them. The user will be requested to introduce such descriptors when applying his/her model to new data. The second feature will allow avoiding the first step, in case the predicted values for analyzed molecules were already uploaded to the database.

Several other enhancements with respect to display of model results are also planned. This includes better description of model results (i.e., explanation of axes of plots, visualization of the full regression equation, calculation of several other applicability domains) as requested by the project participants.