Marie Curie Initial Training Network
Environmental Chemoinformatics (ECO)

Project report
12 October 2011

The ECO Methods for selection of structural features that influence substance toxicities

Duration of Short Term fellowship:
June 2011 – March 2012

Early stage researcher:
Monika Gajewska

Project supervisor:
Prof. Roberto Todeschini

Research Institution:
University of Milano-Bicocca

INTRODUCTION

There are many commercial chemicals found in aquatic systems for which still either no information on toxicity exists or studies are quite limited. Recent legislation requires short-time assessment for their toxicity to aquatic organisms in order to determine which of these chemicals need to be further studied. As a result the new European Union chemical control system adoption, called Registration, Evaluation, and Authorization of Chemicals (REACH), Quantitative structure–activity relationship (QSAR) models are expected to play a crucial role in reducing a number of animals to be used for toxicity testing.Therefore the objective of the study is to collect and analyze available toxicity data for aquatic systems in order to choose the chemicals of interest, then to develop a QSAR models to predict acute in silico toxicity of these chemicals and finally, if possible, to compare the resulting models with the literature ones. However, the main interest of the project lays in the very strategy to construct these models. In order to generate robust, clear and simple predictive models having huge data matrix at one's disposal (up to by several thousands of molecular descriptors for a numerous observations) only the most relevant molecular descriptors should be selected. For this reason reliable and effective variable selection algorithm is needed, which as such, is not yet introduced in literature and there is still much of a controversy among modelers whether a mathematical technique, if any, can be beneficial for model construction.

In this project, the focus is on short term toxicity to aquatic invertebrate organisms (Daphnia, Algae).

MATERIALS AND METHODS

The initial task of the project was a careful collection of reliable experimental values presented in literature or in available on-line databases for acute aquatic toxicity. Only invertebrates (selected species of Algae and Daphnia Magna) were considered. These data would be later used in development and validation of feature selection algorithms.

After a profound insight into available feature selection (FS) methodologies final choice for a method suitable for the investigated problem, to analyze, implement and extend, if possible, has been done.

R software was used, as a tool for FS algorithms implementation. These were presented in a form of a script, incorporating functionalities of several packages and novel functions.

In the final stage QSAR models were to be developed and provided with their sufficient statistical and predictive characteristics, models validation, comparison with existing literature ones, applicability domain; conclusions, suggestions.

Lastly, an effort has been put in preparation of adequate documentation with a simple, straightforward and clear explanation of applied methods.

The software and services used in the project:

- Dragon 6 (http://www.talete.mi.it/products/dragon_description.htm)
- OCHEM portal (http://ochem.eu; online database with modeling environment)
- QSAR toolbox (http://www.qsartoolbox.org/)
- Mobydigs (http://michem.disat.unimib.it/chm/)
- R (http://www.r-project.org/) with the following packages: randomForest, Boruta, caret, party, clValid, cluster, subselect
- Tinn-R, R code editor (http://sourceforge.net/projects/tinn-r/)

Two algorithms, particularly famous in recent biostatistics have been chosen in this study: Random Forest (RF) and Genetic Algorithm (GA).

RESULTS AND DISCUSSION

Database on acute aquatic toxicity for selected invertebrates

Description:

Thorough review of REACH guidelines addressing safe use of chemicals, and in their accordance, careful choice made for the endpoint and species of interest. Subsequently, QSAR models development for a predicting aquatic toxicity. Collection of available experimental data; focus toward variety of organic compounds (industrial organic chemicals, pharmaceuticals, pesticides, surfactants).

Results:

Database on acute aquatic toxicity for Algae and Daphnia Magna with the endpoints: effective concentration (EC50) and lethal concentration (LC50)

SPECIES	ENDPOINT	TYPE AND NUMBER OF COMPOUNDS	DATA SOURCE
Daphnia Magna	48-h LC50	300 Various organics	U.S. EPA AQUIRE (2002)
	48-h LC50	222 Pharmaceuticals	Toxicology Letters 187 (2009) 84–93
	96-h LC50	262 Pesticides	Bioorganic & Medicinal Chemistry 14 (2006) 2779–2788
	48-h EC50 Immobilization	130 Various organics	Journal of Toxicology and Environmental Health, Part A, 72: 1181–1190, 2009
	48-h EC50 Immobilization	17 Substituted benzaldehydes	Chemosphere Vol. 37, No. 1, pp. 79-85, 1998
	48-h EC50 Immobilization	644 Various organics and inorganics	Ministry of the Environment in Japan: Eco-toxicity tests of chemicals (March 2011)
	48-h EC50 Immobilization	22 Benzoic acids	Chemosphere 59 (2005) 255–261
	48-h EC50 Immobilization	6 Anionic surfactants linear alkylbenzene sulphonates (LAS) and 21 ester sulphonates (ES)	Chemosphere 63 (2006)1443–1450
	48-h EC50 Immobilization	74 Organic, inorganic esters	Chemosphere 58 (2005) 559–570
	48-h EC50 Immobilization	40 Various organics	U.S. EPA database ECOTOX (+2000);
	48-h EC50 Immobilization	125 organic chemicals (derived from the European priority list in compliance with Directive 76/464/EEC)	Ecotoxicology and Environmental Safety 49, 206}220 (2001)
Chlorella Vulgaris	15 min EC50; Inhibition of enzyme activity (Fluorescein diacetate)	91 Diverse organic industrial compounds (aliphatic, aromatic)	Chem. Res. Toxicol. 2004, 17, 545-554
	96-h EC50; Inhibition of the activity of acetyl-CoA carboxylase	40 Herbicides	Ecotoxicology and Environmental Safety 51, 128 - 132 (2002)
	96-h EC50; Growth inhibition	14 Pesticide adjuvants	Ecotoxicology and Environmental Safety 58 (2004) 61–67
Pseudokirchneriella Subcapitata	48-h EC50; Biopopulation (Biomass-type based on the cell density)	108 Various organic compounds	Ecotoxicology and Environmental Safety 72 (2009) 1514–1522

	48-h EC50; Growth rate inhibition	20 Benzoic acids	Journal of Hazardous Materials 165 (2009) 156–161
	48-h EC50; Growth rate inhibition	13 Substituted anilines	Environmental Toxicology and Chemistry, Vol. 26, No. 6, pp. 1158–1164, 2007
Scenedesmus obliquus	48-h EC50; Growth rate inhibition	40 Substituted benzenes	Chemosphere 44 (2001) 437-440
	48-h EC50; Growth rate inhibition	25 Nitroaromatics	Chemosphere 59 (2005) 467–471

Problems and limitations:

There are limited resources including available scientific literature and online databases with experimental data on endpoints in question for aquatic toxicity. It is difficult to find a single or several comparable studies with numerous compounds investigated in experiment what would be satisfactory and necessary for reliable QSAR models construction.

Review of feature selection methodologies. Regression by Random Forest.

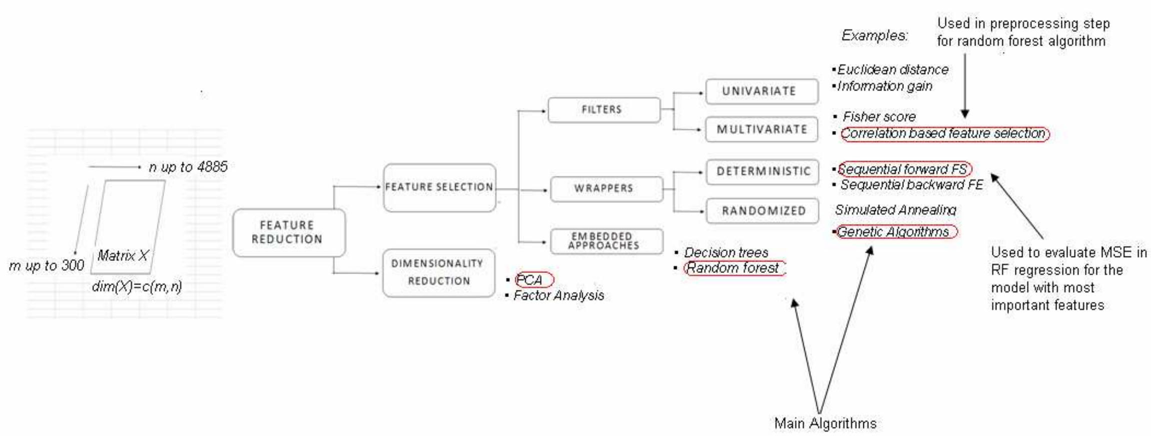
Description:

Development, elaboration and implementation of random forest based R scripts for supervised and unsupervised feature selection; application to collected datasets, performance evaluation, comparison with existing methodologies in OCHEM website.
Initial work toward in silico QSAR modeling to predict environmental toxicity of collected chemicals for Algae and Daphnia. In the final stage, their proper validation, sufficient statistical properties and comparison with the literature models.

Results:

a) Two separate functions implemented in R for dimensionality reduction and information visualization for exploring similarities or dissimilarities in data: principal component analysis and multidimensional scaling

b) Three algorithms for feature selection based on Random Forest: collection of functions for data handling, visualization, storage, generation of clear, comprehensible results.
Short summary of all three algorithms is presented below.



Supervised Random Forest regression

1. A matrix with molecular descriptors (up to 18 descriptors blocks: 0,1,2 D descriptors) for number of observations (compounds) considered as training set created was by means of Dragon 6.
2. Data input into R script, check for correlations, near to zero variance (pre-processing), statistical tests to remove the most irrelevant molecular descriptors.
3. Random Forest variable importance measure (two types) , stepwise addition of variables from the most to the least important one in a model to evaluate mean square error (MSE), choice for a model with the smallest MSE.
4. In search for a model with lower number of variables and lower or comparable MSE error, calculate this error for a model with all possible combinations of up to 10 most important variables (due to high computational burden).
5. Return of a new matrix with all statistically important descriptors, final decision on their acceptance and possible relevance with a measured activity is made by a modeler prior to model build-up.
6. Random Forest model construction, prediction evaluation, application to a previously designated test set, statistics

Supervised Random Forest Regression with Conditional Variable Importance

1. A matrix with molecular descriptors (up to 18 descriptors blocks: 0,1,2 D descriptors) for number of observations (compounds) considered as training set created was by means of Dragon 6.
2. Data input into R script, check for correlations, near to zero variance (pre-processing).
3. Unconditional variable importance measure used to define a threshold for variables pre-selection for conditional measure (statistical test implemented in “party” R package).
4. Variables are checked for existing trends and eventual monotonicity to distinguish from random fluctuations. Only descriptors stated: important and with non-random trend are returned.
5. Modeler makes final decision about the variables used for model build-up.
6. Random Forest model construction, prediction evaluation, application to a previously designated test set, statistics

Unsupervised Random Forest Regression with Cluster Analysis

1. A matrix with molecular descriptors (up to 18 descriptors blocks: 0,1,2 D descriptors) for number of observations (compounds) considered as training set created was by means of Dragon 6.
2. Data input into R script, check for correlations, near to zero variance (pre-processing).
3. RF proximity matrix replaced by dissimilarity matrix. Variable importance calculation.
4. Internal Validation for existence of clusters and appropriate clustering method. In case of clustering, their geometrical interpretation, otherwise choice for representatives on a basis of dissimilarity level.
5. Return a matrix of variables stated important and adequately dissimilar
6. Further model construction possible by any available method (not only random forest)

Problems and limitations:

- For each of the algorithms separate R script must be developed.
- Programming is time consuming and needs constant improvements, modifications and check for correctness.
- High dimensionality problem (far too higher number of molecular descriptors than observations), this exclude the application of many FS algorithms, however Random Forest is said to perform well in such difficulty.

- Problem of correlated descriptors. This affects robustness and performance of a model. An effort is done to reduce their number in a final matrix. Therefore, conditional variable importance is used next traditional approach.
- Problem of overfitting. This is almost unavoidable problem in regression, however, an effort in first two approaches is done to limit it. Last, unsupervised method is implemented where no experimental values (thus no regression) are considered so overfitting is not there a limitation.

Future tasks:

- Finalization of these three (supervised and unsupervised) R scripts for feature selection. They should be clear, simple and comprehensible, easily used by anyone. Few more novel additions and detailed check for the scripts functionality have to be done.
- Careful comparison with methodologies implemented in OCHEM services
- Separate attention given to Genetic Algorithm (MobyDigs and R packages) and its comparison with Random Forest approach.
- QSAR Models; their construction and prediction performance is not yet well completed and validated. For each of species and a given endpoint, model is to be created with a satisfactory explanation and literature comparison.
- Proper data directory and documentation on algorithms functionality must be provided.
- Possible work toward applicability domain, further work suggestions

REFERENCES

1. Hartmann W. M. Abstract: **"**Dimension Reduction vs. Variable Selection"; SAS Institute, Inc., Cary NC, USA

2. Dudek A. Z., Arodz T., Gálvez J. (2006) Combinatorial Chemistry & High Throughput Screening, 9, 213-228

3. Guyon I., Elisseeff A. (2003) Journal of Machine Learning Research 3 1157-1182

4. Kursa M.B., Rudnicki W. R.,(2010) Journal of Statistical Software Vol. 36, Issue 11

5. Liaw A. and Wiener M. (2002) "Classification and Regression by Random Forest" Vol. 2/3

6. Svetnik V., Liaw A., (2003) J. Chem. Inf. Comput. Sci.**,** 43, 1947-1958

7. Janecek G.K., Gansterer W.N, JMLR: Workshop and Conference Proceedings 4: 90-105

8. Strobl C., Boulesteix A.-L. (2008); Technical Report Number 23, Department of Statistics University of Munich

9. Polishchuk P.G., Muratov E.N. (2009); J. Chem. Inf. Model.**,** 49, 2481–2488

TRAININGS & SCIENTIFIC MEETINGS & SCHOOLS

- Attendance in seminars organized within the group to present the research results; Two presentations given: " Database setup for QSAR studies" and "Introduction to R"
- COSMOS Workshop ; 16th June 2011; JRC Ispra, Italy; Introduction to molecular (systems biology models), cellular (DEBTox models),organs (2D liver model) and organisms (PBTK models)
- ECO project online training: 1st June, 3rd August 2011
- Participation to doctoral seminar, presentation given by Faizan Sahigara; 6th July 2011 and Kamel Mansouri; 14th September 2011
- August 2011 Internship at Dr Igor Tetko's group at the Institute of Bioinformatics and SystemsBiology, Helmholtz Zentrum München- German Research Center for Environmental Health
- Participation in OpenTox InterAction Meeting: Innovation in Predictive Toxicology, In Vitro and
- In Silico Modelling, Applications, REACH, Risk Assessment **-** 9-12th August 2011
- 19th-30th September 2011, participation in Environmental ChemOinformatics Summer School at Leiden University (LU). http://www.eco-itn.eu/node/86