HelmholtzZentrum münchen
German Research Center for Environmental Health

Marie Curie Initial Training Network

Environmental Chemoinformatics (ECO)

Final project report /2011

2 February 2012

Reproducibility of linear QSAR models in the JRC database

Duration of Short Term fellowship:
30 August 2011 – 30 November 2011

Early stage researcher:
Elisa D´Onofrio

Project supervisor:
Dr. Igor V. Tetko

Research Institution:
The Helmholtz Zentrum München

Introduction

The JRC database (http://qsardb.jrc.it/qmrf - 1) contains a growing number of descriptions of validated QSAR/QSPR models as QMRF. This database has its major goal to collect descriptions of all validated models to be used for assessment of environmental endpoints with REACH.

Up to now 68 are the published QMRF documents available for download and can be searched by several predefined fields as: Author of the work, Endpoint or Algorithm of the model.

All substances, available in the JRC database, can be searched by exact or similar structure and downloaded.

The QMRF (QSAR Model Reporting Format) is a particular format that provides useful and precise information of each model according to the OECD principles. (Principle 1: defining the endpoint. Principle 2: defining the algorithm. Principle 3: defining the applicability domain. Principle 4: description of the internal and external validation. Principle 5: Providing a mechanistic interpretation where it is possible).

Thanks to JRC database, under the frame of "Environmental ChemOinformatic Marie Curie Initial Training Network", my work was:

i) to upload all the compounds and their relatives ecotoxicological properties in OCHEM database (http://ochem.eu 2)

ii) to re-implement linear models available in the JRC database in order to verify whether they can be reproduced and whether there are any difficulties (if any) with it.

The REACH regulation advocates the use of non-animal testing methods optimizing use of in silico and in vitro information on related compounds, methods such as (Q)SARs models. The predictions of models can be used for hazard and risk assessment, when experimental data are lacking. Because of this aim all the properties compounds found and all the possible models reproduced can be made publicly available at http://ochem.eu and than can be used for the REACH regulation.

Material and Method

OCHEM Database

The Online Chemical Modeling Environment is an online database of experimental measurements integrated with the modeling environment. Submitting experimental data or use the data uploaded by other users is possible build predictive QSAR models for physical-chemical or biological properties.

OCHEM Database is a unique platform on the Web that aims to automate the typical steps required for QSAR modeling. The platform consists of two major subsystems: the database of experimental measurements and the modeling framework. The database is user-contributed and contains a set of tools for easy input, search and simultaneous modification of thousands of records. The OCHEM database is based on the wiki principle and focuses on data quality and verification. The database is tightly integrated with the modeling framework, which supports all the steps required to create a predictive model: data search, calculation and selection of a vast variety of molecular descriptors, application of machine learning methods, validation, analysis of the model and assessment of the applicability domain.

Results and Discussion

Uploading of all the compounds and their relatives eco-toxicological properties in OCHEM database

Thanks to the "Batch Upload Browser" presented in OCHEM, 11863 public properties records now are uploaded in the database. These data are from the linear and non–linear models records available in the JRC database. For each compound I have uploaded the name, the SMILES string, the property related and its unit. If it was available, I have also included experimental conditions, namely temperature, test duration, species/organism/cell culture used. All the properties and the number of the compounds per property are shown in Table 1.

Table 1. The properties added in the OCHEM database

Property	***Number of compounds***
LogP	3919
LogKoc	1993
LC50 aquatic	1780
EC50 aquatic	657
Chromosomal Aberration Index	601
Polyploidy	564
TD50	385
GHLI	374
BCF	317
Mutagenicity	273
Half-life air	250
Abiotic degradation in water	249
Skin sensitization	238
LogK	235
Human health effects	221
LogR	131
LD50bee	124

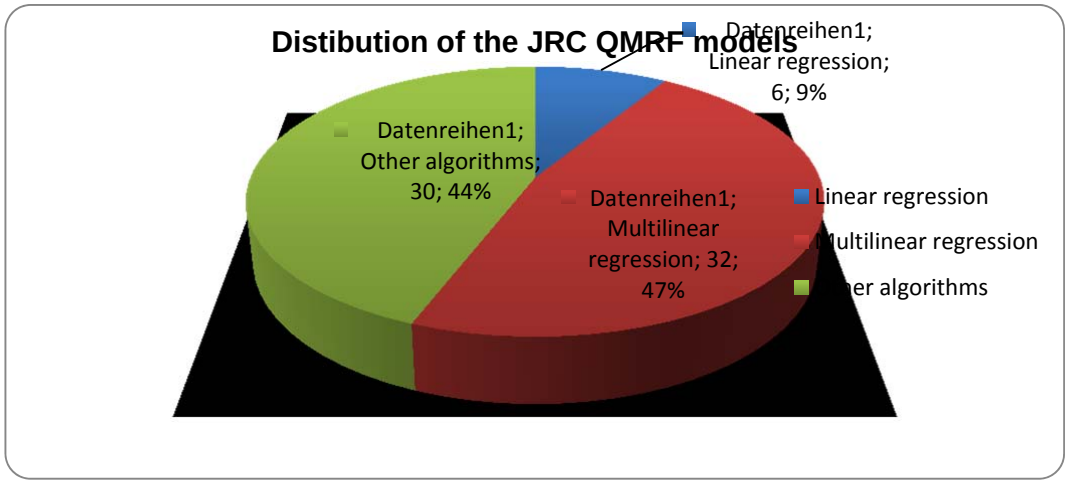
Property	*Number of compounds*
LD50	111
IC50 Acute oral toxicity	100
EC50 bioluminescence	98
logLOAEL	94
pEC50	93
logK' hsa	85
Eye irritation/corrosion	72
logP(BBB)	60

Re-implement linear models available in the JRC database in order to verify whether they can be reproduced and whether there are any difficulties (if any) with it.

68 QSPR/QSAR validated models available in the JRC database can be classified in according to the algorithm used for their development (Table 2):

Table 2. Classification of the JRC models

Algorithm used	**Number of models**
Linear regression	6
Multilinear regression	32
Non linear (other)	30



LINEAR MODELS

Below are listed the six linear models available in the JRC database in which the only descriptor used for their development is the partition coefficient LogKow. It is a ratio of concentrations of compound in the two phases of a mixture of the two immiscible solvents at the equilibrium: water and octanol (Table 3).

Table 3. Six linear models

QMRF Number	QMRF Link JRC database	QMRF Title
154	Q15-28-8-162	TOPKAT NTP Rodent Carcinogenicity Model (Female Mouse).
162	Q15-28-8-162	QSAR for narcosis to fathead minnow, including non-polar and polar narcosis.
317	Q19-39-8-317	Non-polar narcosis QSAR for fathead minnow acute toxicity.
318	Q19-39-8-318	Polar narcosis QSAR for fathead minnow acute toxicity
319	Q27-39-8-319	Polar narcosis QSAR for tetrahymena pyriformis acute toxicity.
320	Q27-40-8-320	Non-polar narcosis QSAR for tetrahymena pyriformis acute toxicity.

For the first linear model, QMRF number 154, it was not possible to verify the reproducibility in OCHEM database since the dataset of the compounds used for its development was not provided.

For the second model, QMRF number 162, it was not possible to verify the reproducibility in OCHEM database since the descriptor data used for its development contained both experimental and calculated values. However, it was still possible to create new models using ALogPS_logP descriptor (octanol/water partition coefficient calculated using ALOGP 2.1 program) available in OCHEM and then comparing the published statistical performances of each model (Table 4). The algorithm was developed with 12908 molecules from the PHYSPROP database using 75 E-state indices. 64 neural networks were trained using 50% of molecules selected by chance from the whole set. The logP prediction accuracy is root mean squared error rms=0.35 and standard mean error s=0.26 (4;5).

The descriptor used for the development of the other four models (QMRF number 317; 318; 319 and 320) was calculated using the program KOWWIN, that is a part of EPIsuite software. The EPI (Estimation Programs Interface) Suite™ is a Windows®-based suite of physical/chemical property and environmental fate estimation programs developed by the EPA's Office of Pollution Prevention Toxics and Syracuse Research Corporation (SRC). KOWWIN™ estimates the log octanol-water partition coefficient, log Kow, of chemicals using an atom/fragment contribution method.

This descriptor is not yet implemented in OCHEM. However, it was possible to create new models using ALogPS_logP descriptor (octanol/water partition coefficient calculated using ALOGP 2.1 program) available in OCHEM as the previous case.

Table 4. Statistical parameters and equations of the linear QSAR models

QMRF Number	Number of compounds		Published Parameters (using KOWWIN descriptor)				Parameters (using ALogPS_logP descriptor)			
			R^2	Q^2	RMSE	MAE	R^2	Q^2	RMSE	MAE
317	66		0.895	/	0.386	/	0.92	0.92	0.33	0.26
		Equation	log 1/LC50= - 4.90 + 0.979 logP				log 1/LC50= - 4.748 + 0.896 ALogPS_logP			
318	43		0.713	/	0.48	/	0.76	0.76	0.44	0.33
		Equation	log 1/LC50=-3.73 + 0.694 logP				log 1/LC50 = -3.679 + 0.679 ALogPS_logP			
319	138		0.763	/	0.397	/	0.69	0.69	0.45	0.35
			log (1/IGC50) = - 0.997+ 0.619 logP				log (1/IGC50) = - 0.954 + 0.604 ALogPS_logP			
320	87		0.957	/	0.27	/	0.96	0.96	0.27	0.2
		Equation	log (1/IGC50)= - 2.07 + 0.825 logP				log (1/IGC50) = -1.931 + 0.716 ALogPS_logP			

As the results shown in the table above, the equations of each model are almost identical: the values of the bias and of the descriptor coefficient are very similar ones.

Regarding the performance of the models, in three out of four cases the models developed using ALogPS_logP instead of KOWWIN descriptor provided slightly better results.

From this comparison we can say that ALogPS_logP can be used instead of KOWWIN to provide models with similar or even better accuracy.

MULTI LINEAR MODELS

Below I have listed 32 multi-linear validated QSAR models available in the JRC database (Table 5).

Table 5. 32 multi-models

QMRF Number	QMRF Link JRC database	QMRF Title
Koc	The link for this model is not provided because it does not have QRMF ID yet.	Model for organic carbon-sorption partition coefficient (logKoc) prediction
53	Q8-29-23-53	Catalogic base surface narcotic model for aquatic toxicity to Daphnia Magna
83	Q2-10-6-83	QSAR for mutagenicity (Salmonella typhimurium TA98 strain)
108	Q2-15-8-108	QSAR for skin sensitization via Schiff base formation
112	Q7-17-11-112	QSAR for the Global Half-Life Index (GHLI) of Persistent Organic Pollutants (POPs) (8)
119	Q2-10-14-119	QSAR for honey bee acute contact toxicity (ester derivatives)
121	Q8-10-13-121	QSAR for honey bee acute contact toxicity (amine derivatives)
126	Q2-17-11-126	QSAR for acute toxicity to Pimephales promelas (Fathead Minnow) (9)
135	Q2-22-1-135	QSAR for eye irritation (Draize test)
136	Q2-10-1-136	QSAR for acute toxicity to fish (Danio rerio)
140	Q2-17-16-140	QSAR for bioconcentration factor in fish (7)
144	Q2-18-14-144	QSAR for honey bee acute contact toxicity (amine derivatives)
150	Q8-10-14-150	QSAR model for acute toxicity to rainbow trout
153	Q8-10-14-153	QSAR for female rat carcinogenicity (TD50) of nitro compounds

QMRF Number	QMRF Link JRC database	QMRF Title
155	Q8-26-8-155	QSAR for haloacetic acid mutagenicity
169	Q8-10-14-169	QSAR for honey bee acute contact toxicity (ether derivatives not containing amide groups)
171	Q8-10-14-171	QSAR for Relative Binding Affinity to Estrogen Receptor
173	Q8-10-24-173	QSAR for bioconcentration (flow-through fish test) of polychlorinated biphenyls
174	Q2-10-14-174	QSAR for acute toxicity to fathead minnow
175	Q8-10-14-175	QSAR for bioconcentration (flow through fish test) of pesticides
176	Q8-10-14-176	QSAR for acute oral toxicity (in vitro)
177	Q2-10-25-177	QSAR for human serum albumin binding
179	Q2-10-26-179	QSAR for soil adsorption coefficient Koc
184	Q2-10-25-184	QSAR for blood-brain barrier (BBB) partitioning
207	Q8-10-14-207	QSAR for the bioconcentration factor of non-ionic organic compounds
208	Q8-10-28-208	QSAR for algae toxicity of benzene derivatives
209	Q8-10-27-209	QSAR for acute toxicity to algae
220	Q8-10-29-220	QSAR model for acute oral toxicity of benzene derivatives- Acute Toxic Class Method
288	Q8-10-30-288	QSAR for rat chronic LOAEL
299	Q19-10-30-299	QSAR for acute toxicity to Daphnia magna (LC50)
300	Q19-10-30-300	QSAR for octanol-water partition coefficient (logP) for pesticides
309	Q8-10-1-309	QSAR for toxicity to activated sludge

For the QMRFs Number: 53; 108; 171; 299 and 300 is not possible to verify the reproducibility in OCHEM database since the datasets of the compounds used for their development are not provided.

For the QMRFs Number: 126; 140; 112 and Koc it was possible to verify the reproducibility in OCHEM database. The comparison of the models published in JEC and the model reproduced in OCHEM are listed in table 6. For each models I have provided the results using two different energetic optimization of the chemical structures compounds: Corina and Mopac. These 4 models were developed using DRAGON descriptors (6), the versions DRAGON 5.4 and DRAGON 5.5, which are available in OCHEM database. The light differences among the parameters are probably caused by the different minimization used.

Table 6. Results of four reproduced models

QMRF Nr 140: QSAR for bioconcentration factor in fish

		(Gramatica et al. 7)			**Corina**				**Mopac**			
Nr of compounds		R2	Q2	RMSE	R2	Q2	RMSE	MAE	R2	Q2	RMSE	MAE
Training	179	0.81	0.8	0.57	0.76	0.75	0.64	0.52	0.76	0.76	0.63	0.52
Test	59	0.9	0.87	0.57	0.89	0.81	0.69	0.55	0.89	0.81	0.69	0.55

QMRF Nr 112 QSAR for the Global Half-Life Index (GHLI) of Persistent Organic Pollutants

		(Gramatica et al. 8)			**Corina**				**Mopac**			
Nr of compounds		R2	Q2	RMSE	R2	Q2	RMSE	MAE	R2	Q2	RMSE	MAE
Training	125	0.85	0.83	0.76	0.86	0.86	0.68	0.52	0.86	0.86	0.68	0.52
Test	125	0.79	/	0.78	0.79	0.79	0.79	0.6	0.79	0.79	0.79	0.6

QMRF Nr 126 QSAR for acute toxicity to Pimephales promelas (Fathead Minnow)

		(Gramatica et al. 9)			**Corina**				**Mopac**			
Nr of compounds		R2	Q2	RMSE	R2	Q2	RMSE	MAE	R2	Q2	RMSE	MAE
Training	249	0.79	0.78	0.38	0.79	0.79	0.6	0.47	0.79	0.79	0.6	0.47
Test	200	0.71	/	0.64	0.72	0.7	0.64	0.51	0.72	0.7	0.64	0.51

QMRF Koc INSBUBRIA QSPR Model for organic carbon-sorption partition coefficient (logKoc)

		(Gramatica et al.)			Corina				Mopac			
Nr of compounds		R2	Q2	RMSE	R2	Q2	RMSE	MAE	R2	Q2	RMSE	MAE
Training	93	0.82	0.80	0.56	0.77	0.77	0.59	0.43	83	83	0.52	0.39
Test	549	0.77	0.78	/	0.75	0.73	0.62	0.46	0.81	0.8	0.53	0.42

The remaining 23 models were developed by “Molcode model development team”. (10). Molcode is using its own descriptors. It has developed 1000+ original molecular descriptors that are calculated solely from the molecular structure. These are not implemented in OCHEM database yet. It is also difficult to find similar descriptors used for the model developing since the abbreviation name of the descriptors changes in according to the software used for their calculation and these abbreviations are not informative and are very difficult to be interpreted. So far, for these models it was impossible to verify the reproducibility of them in OCHEM.

For 19 of these models (QMRF number: 83; 119; 121; 135; 144; 150; 155; 169; 174; 175; 177; 179; 184; 207; 208; 209; 220; 288 and 309) the values of the descriptors for each compound are provided. So even if the descriptors are not known exactly, they could be used in further studies to find the most similar ones.

For 4 of these (QMRF number: 136; 153; 173 and 176) models the values of the descriptors for each compound are not provided. Since the future remapping of the descriptors is not possible, it was decided to analyze the performance of the 4 four models by building new models. I used ASNN (ASsociative Neural Networks) method (11) instead of MLR (Multiple Linear Regression) with the default parameters available in OCHEM database: E-state and ALogPS as molecular descriptors. SuperSAB was used as the training method, 3 neurons and 1000 learning iterations, 64 Ensemble as parameters of networks. The calculated results are shown in table 7.

Table 7. Comparison of four models using different approach

Nr QMRF	Set	Nr of compounds	Molcode models MLR R2	OCHEM ANN R2	Q2	RMSE	MAE
173	Training	58	0.93	0.89	0.89	0.29	0.2
136	Training	61	0.804	0.44	0.44	1.06	0.79
	Test	6	0.847	0.13	-0.01	1.01	0.84
153	Training	42	0.73	0.46	0.41	0.66	0.5
	Test	3	0.94	0.93	-1.13	0.67	0.54
176	Training	45	0.78	0.66	0.64	1.13	0.79
	Test	5	0.82	0.97	0.88	0.57	0.4

The models developed using ASNN showed considerably lower statistical parameters compared to the original results, probably because the default descriptors used (E-state and ALogPS) are not proper to explain the correct relationship between the structure and the response of the compounds.

The suggestion is to try other molecular descriptors or other statistical methods to redevelop models.

This distribution of the 32 multi-linear models is summarized in figure 1.

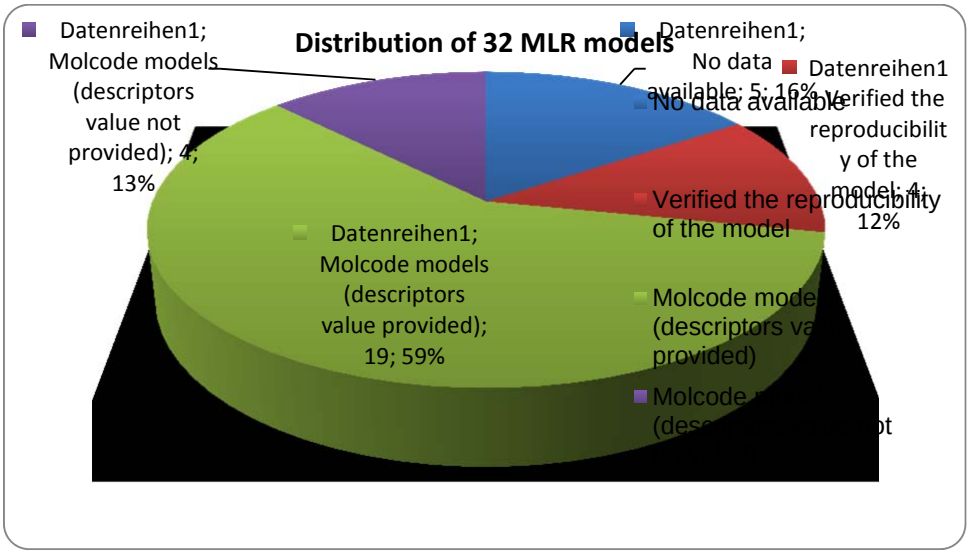


Figure 1. Classification 32 multi-model

NON-LINEAR MODELS

The remaining 30 models are non-linear ones. The dataset with the compounds used for their developments are not provided for 12 of them. 11 models were developed using neural networks, 4 models were multilinear models derived with BMLR method (Best Multi Linear Regression) and three models were developed with Canonical Discriminate Analysis (Fig. 2).

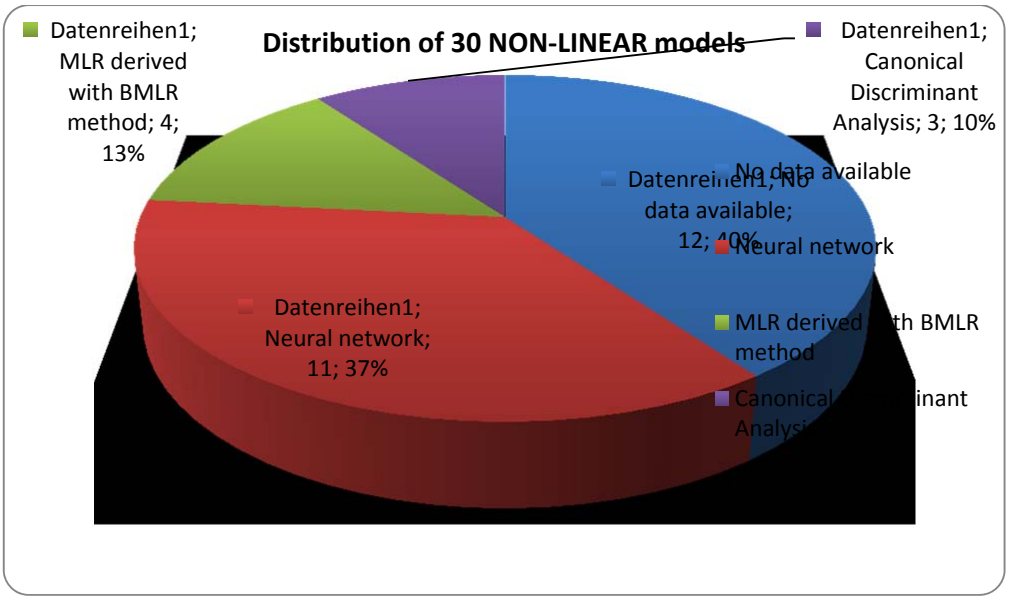


Figure 2. Classification 32 multi-models

Regarding non-linear models is important to underline that for 40% of them is impossible to verify the reproducibility because the data are not provided. For the rest of them the re-implementing of non-linear models available in JRC database can be a future important task.

Summary

In conclusion this work indicates the usefulness of JRC database. It provides in QMRF format information about models according to the OECD principles. Thanks to it I was able to upload 11863 compounds and their relatives ecotoxicological properties in OCHEM database. I was able to re-implement some linear models available in the JRC database. I also indicated which problems did not allow me to reproduce other models as previously reported.

I suggest future tasks that can be interesting to continue this work:

- Add/remap molecular descriptors to find the “unknown Molcode descriptors” of the linear models;
- redevelop models with different approaches (ANN) and to find the best way to redevelop a model;
- re-implement non-linear models available in JRC database;
- add/implement new descriptors:
  - ✓ KOWWIN (EPIsuite);
  - ✓ MOLCODE descriptors

in order to verify the reproducibly of the models that are based on these molecular descriptors.

Is important to underline a lack of data for some models in JRC database: 16% of linear models and 40% of non-linear model do not contain data. This makes impossible to reproduce and re-implement these models.

References

1. http://qsardb.jrc.it/qmrf/
2. Sushko I, Novotarskyi S, Körner R, Pandey AK, Rupp M, Teetz W, Brandmaier S, Abdelaziz A, Prokopenko VV, Tanchuk VY, Todeschini R, Varnek A, Marcou G, Ertl P, Potemkin V, Grishina M, Gasteiger J, Schwab C, Baskin II, Palyulin VA, Radchenko EV, Welsh WJ, Kholodovych V, Chekmarev D, Cherkasov A, Aires-de-Sousa J, Zhang QY, Bender A, Nigsch F, Patiny L, Williams A, Tkachenko V, Tetko IV. Online chemical modeling environment (OCHEM): web platform for data storage, model development and publishing of chemical information. J Comput Aided Mol Des. 2011 Jun;25(6):533-54
3. http://www.epa.gov/oppt/exposure/pubs/episuite.htm
4. Tetko, I. V.; Tanchuk, V. Y. Application of associative neural networks for prediction of lipophilicity in ALOGPS 2.1 program, J. Chem. Inf. Comput. Sci., 2002, 42, 1136-45.
5. Tetko, I. V.; Tanchuk, V. Y.; Villa, A. E. Prediction of n-octanol/water partition coefficients from PHYSPROP database using artificial neural networks and E-state indices, J. Chem. Inf. Comput. Sci., 2001, 41, 1407-21.
6. R.Todeschini and V.Consonni: "Molecular Descriptors for Chemoinformatics", (2 volumes), WILEY-VCH,Weinheim (Germany) 2009, 1257 pp
7. An Update of the BCF QSAR model based on theoretical molecular descripritors. QSAR & Combinatorial Science 24, 953-960
8. Gramatica P and Papa E (2007). Screening and Ranking of POPs for Global Half-Life: QSAR Approaches for Prioritization Based on Molecular Structure. Environmental Science & Technology 41, 2833-2839.
9. Papa E, Villa F and Gramatica P (2005). Statistically Validated QSARs, Based on Theoretical Descriptors, for Modeling Aquatic Toxicity of Organic Chemicals in Pimephales promelas (Fathead Minnow). Journal of Chemical Information and Modeling 45, 1256-1266.
10. http://www.molcode.com
11. Tetko, I. V.; Associative neural network Methods Mol Biol. 2008;458:185-202.