In-silico structure activity relationship study of toxicity endpoints by QSAR modeling

Abstract ID: 2273u

www.epa.gov

Kamel Mansouri[1,2], Nisha Sipes[1,2] and Richard Judson[2]

[1]Oak Ridge Institute for Science and Education, Oak Ridge, TN, USA

[2]National Center for Computational Toxicology, Office of Research and Development, U.S. Environmental Protection Agency, RTP, NC, USA

EPA @ SOT

Kamel Mansouri | mansouri.kamel@epa.gov | 919-541-0545

Introduction

Several thousand chemicals were tested in hundreds of toxicity-related *in-vitro* high-throughput screening (HTS) bioassays through the EPA's ToxCast and Tox21 projects. However, this chemical set only covers a portion of the chemical space of interest for environmental risk assessment, leading to a need to fill data gaps with other methods. A cost effective and reliable approach to fullfill this task is to build quantitative structure-activity relationships (QSARs).

In this work, a subset of 1877 chemicals from ToxCast were used to build QSAR models. These models will be applied to predict values for multiple ToxCast assays in a larger environmental database of ~30K chemical structures.

Based on a clustering study by Sipes et al. (2013), the initial molecular targets of this effort consisted of a set of 18 NovaScreen G-protein coupled receptor (GPCR) assays. These assays are part of the aminergic category that showed the highest number of actives within the ToxCast portfolio. Classification methods including SOM, SVM, PLSDA and kNN, were tested. These methods were coupled to variable selection techniques such as genetic algorithms that were applied in order to select the best representative molecular descriptors based on statistical fitness functions. The obtained models were validated and their prediction ability measured. The models that showed good results will be applied within the limits of their established chemical space defined by the applicability domain.

Data preparation

Chemical structure curation:

The initial dataset considered for this study consisted of 1877 chemicals. In order to prepare a consistent QSAR ready dataset, the chemical structures were curated using a KNIME workflow developed for this purpose.

The main steps of this workflow were:

- Check the validity of the molecular file format and retrieve any missing structures
- Remove the inorganic and metallo-organic structures
- Remove salts and counter ions and fulfill valence
- Convert stereo-isomers and tautomers into a unique form to reduce redundancy
- Remove any duplicates
- Calculate descriptors

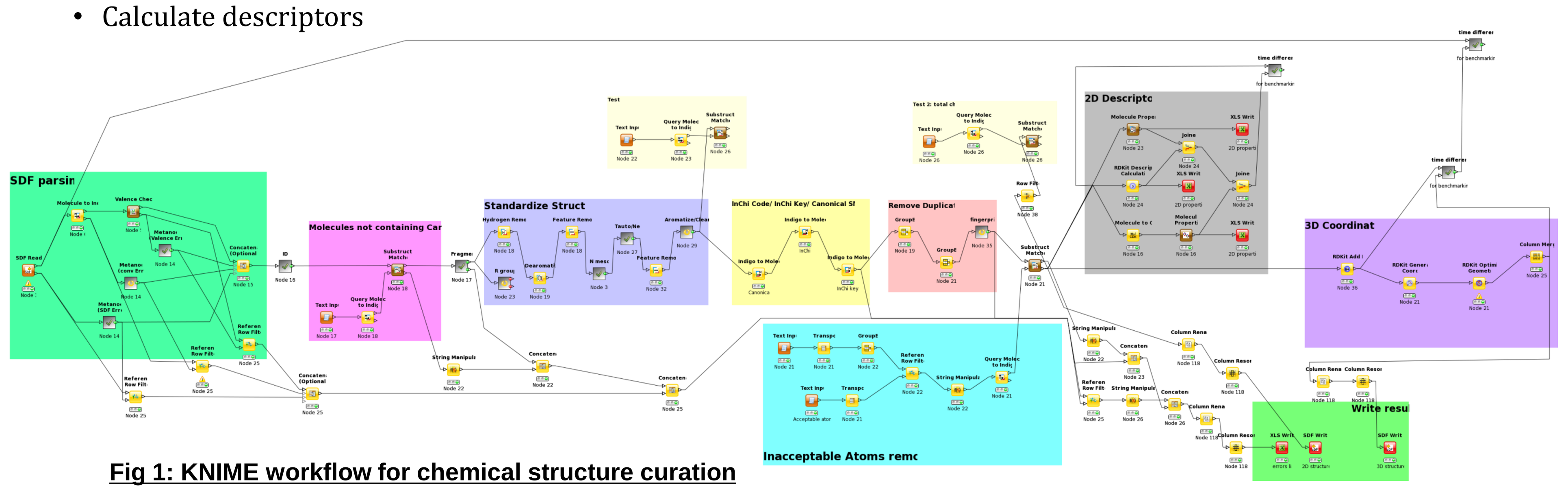

Fig 1: KNIME workflow for chemical structure curation

Molecular descriptors calculation:

A total of 1022 molecular descriptors were calculated from the 2D chemical structures.
The tools used for descriptor calculation were: Indigo, RDKit, CDK and MOE.
In order to reduce collinearity, a correlation filter with a threshold of 0.96 was applied and descriptors with constant, near constant or at least one missing value, were removed. The remaining reduced set consisted of **470** descriptors.

The modeling procedure

Assay	Symbol	Name
Hm1-5	CHRM1-5	cholinergic receptor, muscarinic 1-5
gMPeripheral_ NonSelective	M1	Muscarinic receptor peripheral
hAdrb2	ADRB2	adrenergic, beta-2-, receptor, surface
bDR_ NonSelective	DRD1	Dopamine receptor D1
h5HT2A	HTR2A	5-hydroxytryptamine (serotonin) receptor 2A
rAdra1A,B	Adra1a,b	adrenergic, alpha-1A-B, receptor
rAdra1_ NonSelective	Adra1a	adrenergic, alpha-1A-, receptor
hH1	HRH1	Histamine receptor H1
gH2	Hrh2	Guineapig histamine receptor H2
rAdra2_ NonSelective	Adra2a	adrenergic, alpha-2A-, receptor
hAdra2A	ADRA2A	adrenergic, alpha-2A-, receptor
rmAdra2B	Adra2b	adrenergic, alpha-2B-, receptor

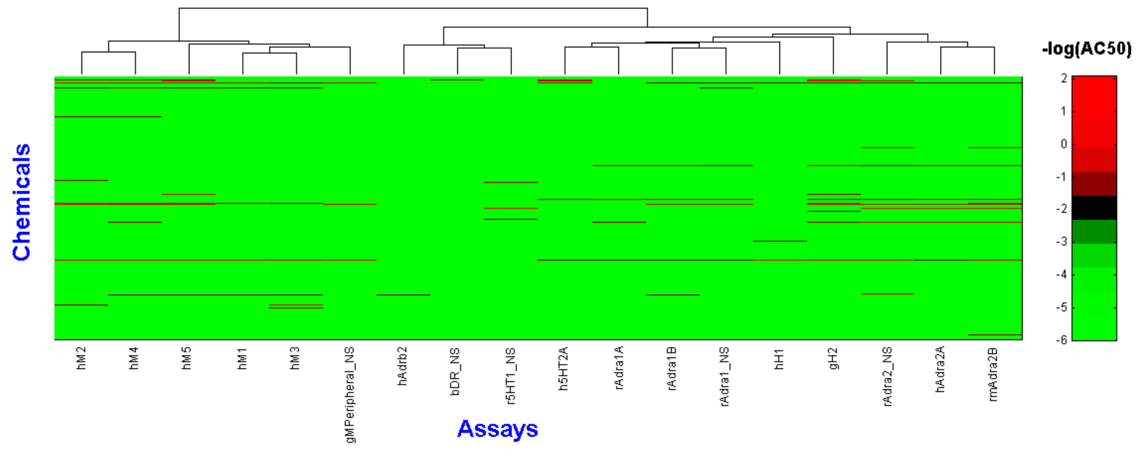

Fig 2: Heatmap of the clustered 18 assays

The resulting SDF file contained **1783** curated 2D chemical structures. Only **1005** compounds tested for all **18** endpoints were considered in the training set. After selecting the best QSAR model to be applied, the 778 entries with missing values will be filled with predictions. A new vector with molecules that are active for at least on of 18 assays was created to model the whole group.

Techniques used :

Kohonen self-organizing maps (SOM):

A machine learning technique that employs artificial neural networks (ANN) to collapse the data samples in a number of clusters (winning neurons) in a two dimensional space.

Partial least squares discriminant analysis (PLSDA):

PLSDA is a partial least squares regression (PLS2-based) with the discrimination power of a classification method. It finds fundamental relations between the matrix of descriptors and the class vector by calculating latent variables (LVs), which are orthogonal linear combinations of the original variables.

K-nearest neighbors (*k*NN):

A molecule is classified according to the classes of the k closest molecules, according to the majority of its k nearest neighbors. The Euclidean metric was used to measure distances between molecules in the descriptors space.

Support vector machines (SVM):

SVM define a decision boundary that optimally separates two classes by maximizing the distance between them. The decision boundary can be described as an hyper-plane that is expressed in terms of a linear combination of functions parameterized by support vectors, which consist in a subset of training molecules.

Variable selection using genetic algorithms (GA):

GA is a nature inspired technique applied to find the optimal subset of molecular descriptors. It starts from an initial random population of chromosomes (present molecular descriptors). An evolutionary process is simulated to optimize a defined fitness function. New chromosomes are obtained by coupling those of the initial population with genetic operations (crossover and mutation). This operation was repeated for *n* runs. Subsequently, a forward selection was performed based on the most frequently included descriptors during the *n* runs (Fig. 6) .

In this work, the feature selection procedure was performed in two steps. First, GA was launched for 50 runs, 100 evolutions for each, on the total number of 470 descriptors. Then, another 50 runs were performed on the most selected subset during the first step which improves the results (Fig. 5).

Goodness of fit measure and validation methods:

5-fold cross-validation was coupled to GA in order to validate the models and avoid overfitting (by-chance correlation). It consists of, repeatedly, fitting a model using 80% of the data and predicting the 20% left out.

The used fitness function was the non-error rate (NER) called also balanced accuracy: $NER\% = (Sn + Sp)/2$

and $$Accuracy = \frac{TP + TN}{TP + TN + NP + FN}$$

where $Sp \equiv TNR = \dfrac{TN}{TN + FP}$ and $Sn \equiv TPR = \dfrac{TP}{TP + FN}$

Results

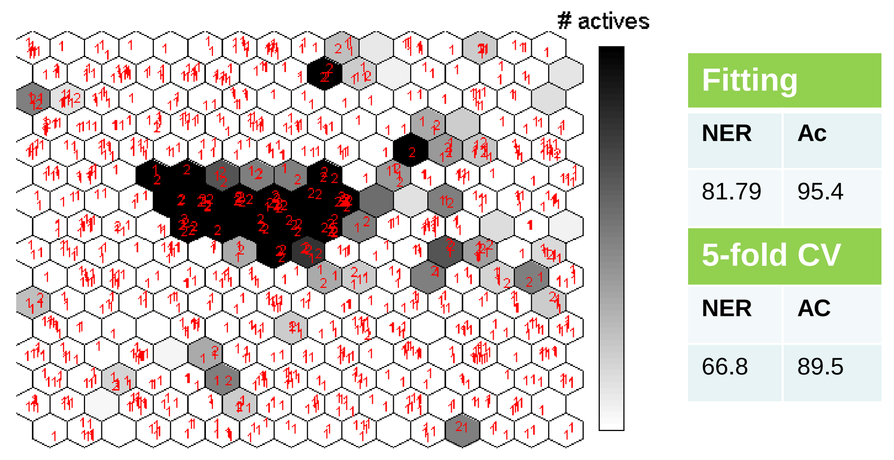

Fig 3:Supervised-learning SOM map of all active compounds for the 18 assays with the full set of 470 descriptors.

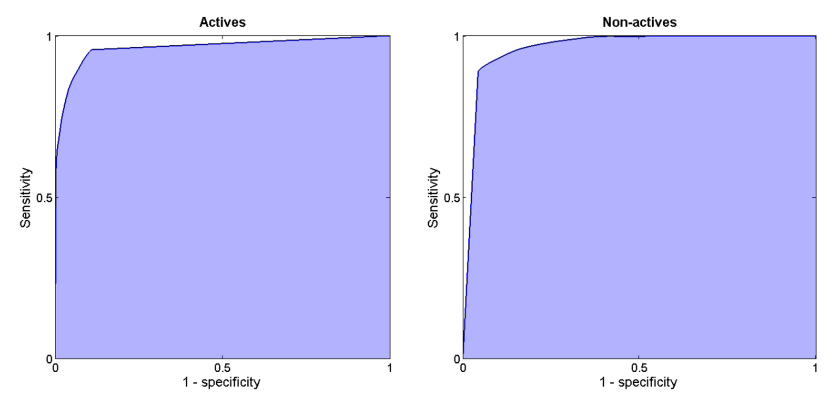

Fig 4: ROC curves of the SOM map in fitting for active and non-active compounds.

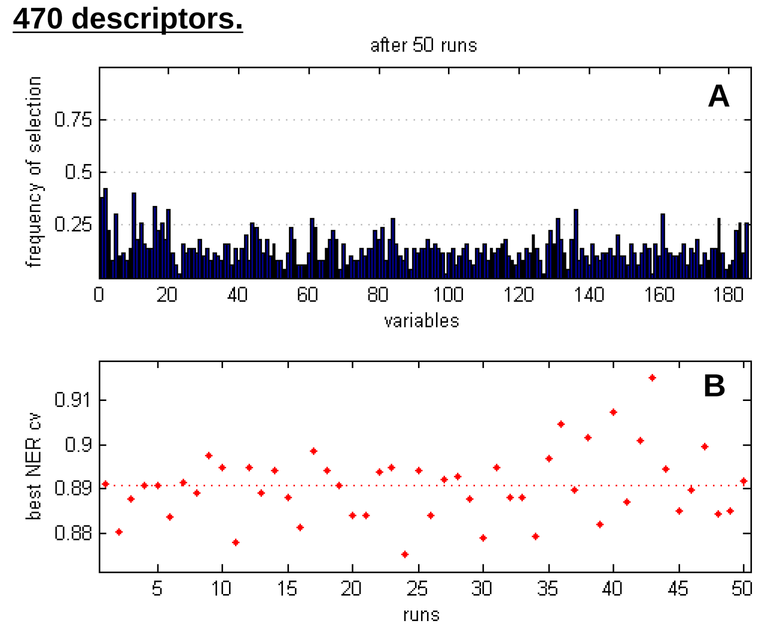

Fig 5: The frequency of selection for each descriptor after 50 GA runs (A) and the NER in 5-fold CV for each GA run (B) for hH1.

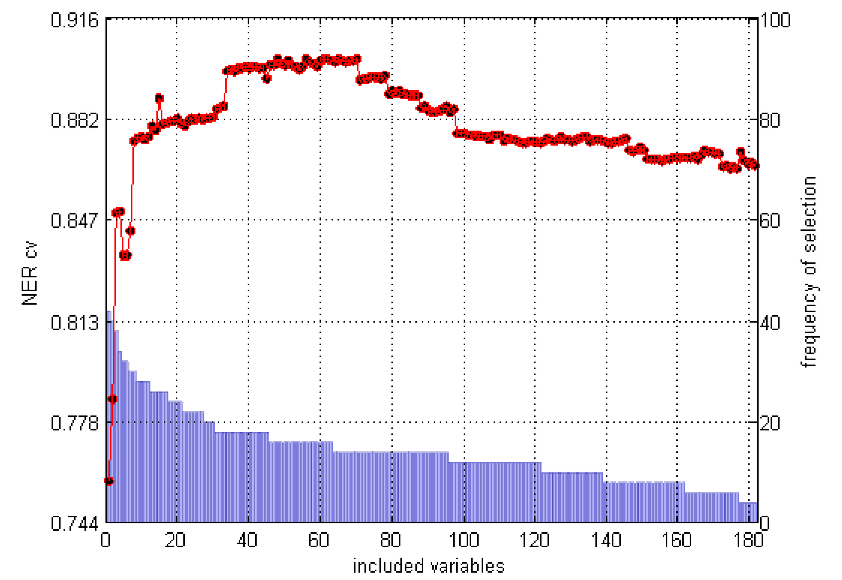

Fig 6: The final forward selection based on the frequency of selection for each descriptor after 50 GA runs for hH1.

Method	PLSDA						KNN						SVM					
Endpoint (Positives)	Dsc	LVs	Fitting		5-fold CV		Dsc	k	Fitting		5-fold CV		Dsc	C	Fitting		5-fold CV	
			NER	Ac	NER	Ac			NER	Ac	NER	Ac			NER	Ac	NER	Ac
hH1 (37)	26	5	91.9	89.4	91.5	88.6	27	1	81.1	96.2	81.2	96.3	30	5	87.8	99.1	68.5	96.9
hM1-5 (76)	15	3	84.1	82.9	85.7	83.6	19	4	76.7	93.8	78.8	94.3	12	10	94.7	99.2	76.6	94.7
All (115)	20	5	84.0	85.8	82.3	84.1	22	1	76.8	90.3	78.1	90.1	16	10	94.3	98.7	75.9	92.1

Dsc: Number of descriptors; LVs: Latent variables; Ac: Accuracy; K: Number of nearest neighbors; C: Cost, SVM's degree of fitting; CV: Cross-validation.

Conclusions and future work

- The QSAR models showed good results especially using the PLSDA method.
- Except for SVM, the statistics of the models between fitting and 5-fold CV are balanced indicating low overfitting.
- Modeling all aminergic category assays together showed acceptable accuracy. This is a good approximation for such similar assays. The accuracy increased with SVM compared to the single model for hH1.
- Future work:
  - Apply the same procedure for all the 18 assays and build regression models to predict AC50 values.
  - Link the molecular descriptors selected in each model to the predicted biological activity.
  - Develop consensus models using the predictions of all the tested methods to increase accuracy.
  - Employ muti-criteria decision making (MCDM) techniques to improve and simplify the models.

U.S. Environmental Protection Agency
Office of Research and Development