QSAR modelling of biological activity by descriptors calculated with simplified molecular input line entry system (SMILES)

A. Chana[1], A. A. Toropov[1], A. P. Toropova[1], E. Benfenati[1], X.-K Hu[2], H.-M. Hwang[2], B. F. Rasulev[2], T. Puzyn[2] and J. Leszczynski[2]

1) Istituto di Ricerche Farmacologiche "Mario Negri", Via La Masa 19, 20156, Milano, Italy

2) Computational Center for Molecular Structure and Interactions. Department of Chemistry, Jackson State University, 1400 J. R. Lynch Str. P.O. Box 17910, Jackson, MS 39217, USA

Introduction

The prediction by Quantitative Structure-Activity Relationships (QSARs) of biochemical parameters associated with chemical substances is becoming an important tool for risk assessment and is explicitly mentioned within the REACH legislation. Indeed the QSAR approach has several advantages:

- ❑ Low cost of its development
- ❑ No animal tests
- ❑ No chemical waste
- ❑ Easiness of use
- ❑ Fast and accurate prediction
- ❑ Reproducible results and models

On the other hand QSAR methods are highly dependent on the input data, not only on their quality but on the form they are fed to the model. Typically the chemical information is given to the QSAR in the form of molecular descriptors. Here we present some QSAR models based on the search of strings invariants within the SMILES code used to store large sets of molecules which are used as molecular descriptors able to be related with the molecular toxicity.

Method

Optimal SMILES-based descriptors provide an one-variable model for the prediction of endpoints values for substances which have not been examined in direct experiment. The model is EndPoint = C0 + C1*DCW(SMILES). The DCW (descriptor of correlation weights) may be defined as

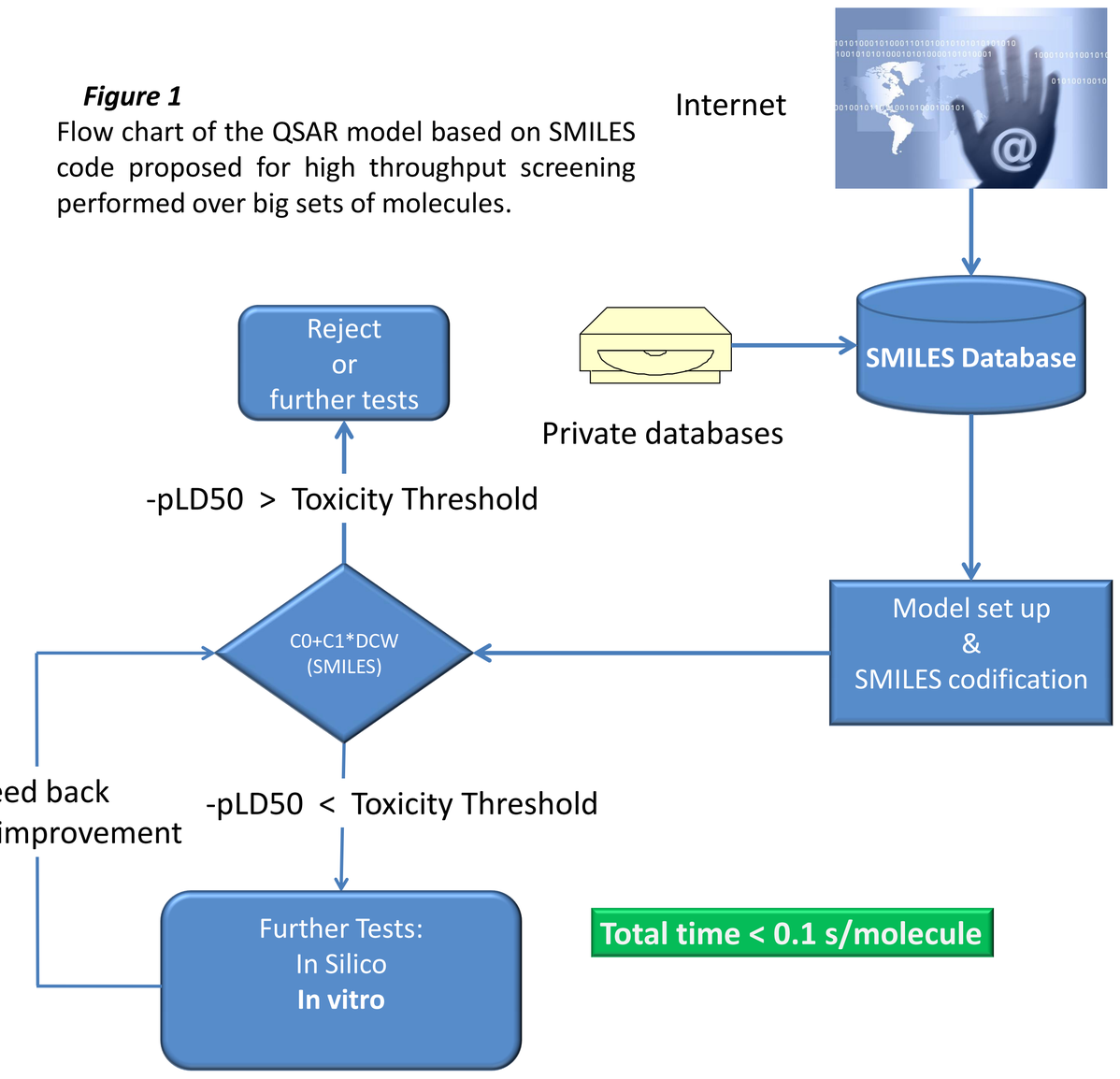
DCW(SMILES) = Σ CW(k-th SMILES attribute) (1)

or

DCW(SMILES) = ∏ CW(k-th SMILES attribute) (2)

The k-th local SMILES attribute may be a symbol of the SMILES notation and/or an combine of the symbols [1-3]. The global SMILES attributes (e.g., number of oxygen, nitrogen atoms, or number of double bonds, etc) may be used in the scheme. The correlation weights of SMILES attributes are calculated by the Monte Carlo method optimization that provide as large as possible correlation coefficient between the DCW(SMILES) and endpoint for the training set. The predictive potential of the model can be validated with an external test set.

Figure 1
Flow chart of the QSAR model based on SMILES code proposed for high throughput screening performed over big sets of molecules.



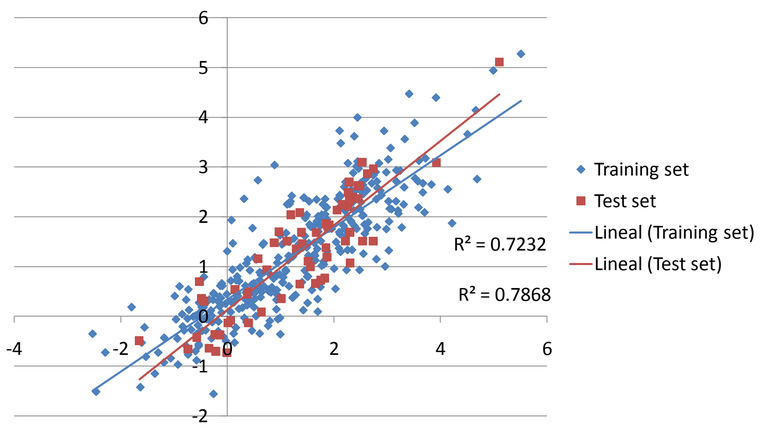


Figure 2
Scattered plot Predicted vs. Calculated of the carcinogenicity model. Training and external test sets are indicated in blue and red respectively

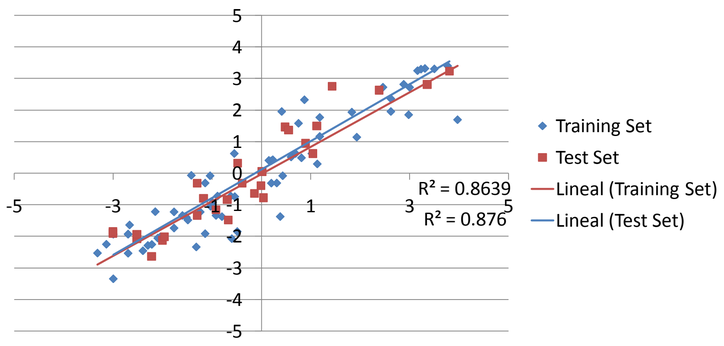


Figure 3
Scattered plot Predicted vs. Calculated of the Mutagenicity model. Training and external test sets are indicated in blue and red respectively

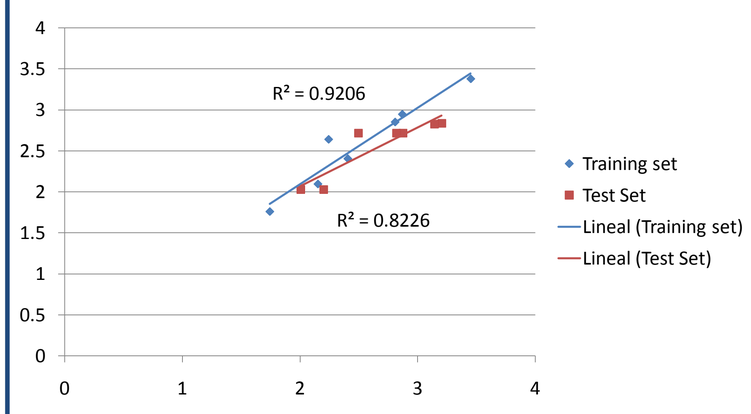


Figure 4
Scattered plot Predicted vs. Calculated of the nanosized particles toxicity model. Training and external test sets are indicated in blue and red respectively

Results

The model results presented here are the application of the molecular invariant approach applied to the SMILES code. It is a rather versatile approach since we are showing the application to two different sort of problems, carcinogenicity and nanosized particles toxicity. For carcinogenicity optimal descriptors have been calculated taking into account, not only common sets of characters but cycles codes which reflect quality and quantity of cycles contained within the molecular structure.

&(5-memb. cycl)(6-memb. cycl.)(heteroatoms)

O=C2Oc1c4C5C=COC5Oc4cc(OC)c1C=3CCC(O)C2=3

The cyclicity code &321

Such invariant is not necessary for nanosized particles instead other kind of invariants such as **Al**, **=**, **[**, **Zn**, etc., has been found relevant for the structure-toxicity relationships.

In any case this is a good measure of the versatility of the methodology able to model biological properties for a wide variety of chemical compounds.

References

[1] A. A. Toropov, E. Benfenati Cur. Drug. Disc. Tech, 4 (2007) 77-116
[2] A. A. Toropov, D. Leszczynska, J. Leszczynski Mat. Lett. 61 (2007) 4777-4780
[3] A. A. Toropov, E. Benfenati Computat.

Conclusions

The molecular descriptors and the QSAR associated are able to:

- ❑ Evaluate big datasets in short time
- ❑ Deal with chemical compounds with diverse chemical nature
- ❑ handle different endpoints
- ❑ provide good predictivity

Acknowledgements

The authors thank the *Marie Curie Fellowships* for financial support through the contract MIF1-CT-2006-039036 - CHEMPREDICT and the EC funded project CAESAR (contract SSPI-022674)