**Mathematical structural descriptors and mutagenicity assessment: A study with congeneric and diverse data sets**

Subhabrata Majumdar, Subhash C. Basak, Gregory D. Grunwald, Mircea Diudea and Claudiu Lungu

Abstract: TBD later

Keywords: TBD later

**1. INTRODUCTION**

Hazard assessment of chemicals is often carried out in data poor situations [l]. The Toxic Substances Control Act (TSCA) Inventory, maintained by the United States Environmental Protection Agency (USEPA), currently has about 85,000 entries [2]. A large fraction of these chemicals has very little or no data needed for their hazard estimation [3]. The assessment of chemical mutagenicity is important both for environmental protection and drug discovery. Identification of potential mutagenicity for industrial chemicals and environmental pollutants is prerequisite to the protection of human and ecological health. For drug discovery, early mutagenicity detection for drug candidates can help in the effective allocation of resources in drug design protocol which costs on the average over US $2 billion [4].

Laboratory testing of mutagenicity for all possible candidate chemicals, can be very expensive. Therefore, assessment of potential mutagenicity of chemicals from Quantitative Structure-Activity Relationship (QSAR) models has been accepted for evaluation of chemicals in lieu of experimental mutagenicity data [5].

This paper has a two-fold objective: 1) Apply computed molecular descriptors in the formulation of QSARs for the prediction of mutagenicity of two data sets, viz., a homogeneous set of 95 aromatic and heteroaromatic amines and a large as well as structurally diverse set of 508 chemicals, and 2) Use a battery of various statistical and machine learning approaches in model building for mutagenicity assessment.

**2. MATERIALS AND METHODS**

**2.1. Data**

The two datasets used in this paper represent two different use cases that practitioners are likely to enter while doing QSAR analysis. The first data consists of the mutagenic activities of 95 congeneric amines on two *S. typhimurium* strains (TA98 and TA100) [31]. The data from the original study by Debnath *et al* consisted of 275 descriptors for each compound, and the response variable measures the log number of revertants per nmol when that compound is applied to *S. typhimurium* test cultures of TA98 strain. While the compounds in this dataset are very similar to each other in chemical structure, our second dataset consists of data on 508 chemical compounds from several different chemical classes. **Table 1** summarizes this classification of the chemical compounds (note that a compound can belong to two or more classes). Collected from the CRC Handbook of Identified Carcinogens and Non-carcinogens [32], the response variable in this dataset is the 0/1 Ames mutagenicity status of the chemical compounds. In total the data contains 256 mutagens and 252 non-mutagens. For each of the datasets, we augmented the set of descriptors used in the original study with the descriptors calculated by Diudea *et al*, as used previously in [refs of some Diudea group papers that used their indices].

Table 1: Chemical classes of samples in the 508 compound diverse dataset

|  |  |
| --- | --- |
| Chemical class | Number of compounds |
| Aliphatic alkanes, alkenes, alkynes | 124 |
| Monocyclic compounds | 260 |
| Monocyclic carbocycles | 186 |
| Monocyclic heterocycles | 74 |
| Polycyclic compounds | 192 |
| Polycyclic carbocycles | 119 |
| Polycyclic heterocycles | 73 |
| Nitro compounds | 47 |
| Nitroso compounds | 30 |
| Alkyl halides | 55 |
| Alcohols, thiols | 93 |
| Ethers, sulfides | 38 |
| Ketones, ketenes, imines, quinones | 39 |
| Carboxylic acids, peroxy acids | 34 |
| Esters, lactones | 34 |
| Amides, imides, lactams | 36 |
| Carbamates, ureas, thioureas, guanidines | 41 |
| Amines, hydroxylamines | 143 |
| Hydrazines, hydrazides, hydrazones, traizines | 55 |
| Oxygenated sulfur and phosphorus | 53 |
| Epoxides, peroxides, aziridines | 25 |

**2.2. Descriptors**

For this study we have used two collections of molecular descriptors. One set of descriptors, used frequently by the Cluj team of Diudea and collaborators, were calculated by the programs Schrodinger [6] and TopoCluj [7,8]. More detailed references about these descriptors are given in Supplementary Tables 1 and 4. For the 95 and 508 data sets, ?? and ?? descriptors were calculated by Diudea et al.

The second set of molecular descriptors, used frequently by Basak *et al*, were calculated by the software POLLY [8], MolConnZ (9), Triplet [10], and MOPAC [11]. For the 95 and 508 chemical sets, ?? and ?? descriptors were calculated for this paper by this software.

**2.3. Statistical and machine learning methods**

We use three types of methods to build our predictive models.

*2.3.1. Dimension reduction*

The hundreds of descriptors generally used in chemometric analysis generally have a high degree of correlation among them [ref]. For this reason, dimension reduction techniques, such as Principal Component Analysis (PCA) or Partial Least Squares (PLS) have seen widespread use in QSAR model building [refs]. In this paper, we build predictive models using the following two dimension reduction methods:

*Principal Component Regression (PCR)*: We transform descriptor matrix **X**, we transform it by multiplying with a principal component loading matrix:

Where the number of columns in **Γ** denote the minimum number of principal components (PCs) that explain 95% of the total underlying variation. We follow the analysis of Majumdar *et al* [ref] and apply a robust PCA procedure [ref] to obtain the PC loadings. Following this, we use the transformed data matrix as the matrix of predictors in linear and logistic regression models to predict activities in the 95 and 508 compound datasets, respectively.

*Partial Least Squares (PLS)*: Another popular method in QSAR literature, PLS uses latent variables to model the correlation between predictors and the response variables. Mainly used to build models used in prediction purposes, PLS obtains a sequence of linear regression coefficients by successively regressing orthogonal components in the data matrix on those in the response vector.

*2.3.2. Variable selection*

Since the datasets we are dealing with are inherently high-dimensional (i.e. large number of predictors that can potentially be more than the number of samples), we use sparse regression methods for variable selection.

*Least Absolute Shrinkage and Selection Operator (LASSO)*: In a linear or generalized linear model, the lasso method [ref] obtains *sparse* estimates of the coefficient estimates by setting some entries to exactly zero. In the QSAR context, this means some predictors will have zero effect on the response variable. Thus, the lasso method is able to perform simultaneous variable selection and model building.

*Smoothly Clipped Absolute Deviation penalty (SCAD)*: Proposed by Fan and Li [ref], SCAD is another penalization method that selects sparser models than lasso, i.e. models where more entries in the coefficient vector are set at 0, without compromising on the predictive capability of the model.

*2.3.3. Machine learning*

Our goal in this paper is to assess and compare the predictive capabilities of different descriptor sets. Machine learning methods are known to produce models with high predictive performance, even though interpreting them is often difficult [refs]. For this reason, we use the following two methods in our study.

*Random Forest (RF)*: This method trains multiple decision trees on a dataset, each based on a randomly selected subset of total features. The final prediction in a regression problem is taken as the average of individual predictions from all the trees, while in classification problem the final class prediction is done by majority voting. Previous examples of the use of RF models in QSAR include [refs].

*Gradient Boosting Machine (GBM)*: Gradient boosting attempts to fit the data using multiple ‘weak learners’, which are simple models that work slightly better than random guessing. At first a weak learner is trained on the data, residuals are obtained from that model and those are again fit using weak learners. Boosting methods have proven to be very useful in predictive model building since their proposal. Examples of boosting in the QSAR scenario include [refs].

**2.4. Validation**

We use a ‘two-deep’ multi-split cross validation scheme to evaluate our predictive methods. Multi-split means we consider multiple random train-test splits of the data, build a model on the train partition, evaluate them on the test partition, and compare different methods using the average values of a metric (e.g. Root Mean Squared Error, Area Under Curve etc.) across all such test sets. This has been referred in the QSAR literature as Monte-Carlo Cross Validation, and ensures that the true underlying components in a model (e.g. important predictors or principal components) are more and more likely to be recovered accurately as sample size increases [ref]. The phrase ‘two-deep’ means we repeat the dimension reduction/ tuning parameter selection steps of the method being implemented. This ensures that information from the test samples are not used while training the model, and gives a more accurate picture of the predictive capability of the technique being analyzed [refs].

**3. RESULTS**

In this section, we state and discuss the outputs from our analysis. Section 3.1 is concerned about the predictive models and the comparison of outputs across different methods and predictor sets, while in Section 3.2 we list the top principal components for evaluating the effects of the new Cluj descriptors with respect to previous findings on the same datasets. All data analyses were done using the statistical software R v3.3.2 [ref].

**3.1. Output of predictive models**

Table 2: Average and standard deviations (in brackets) of Area Under Curve (AUC) for different methods applied on the 508 compounds heterogeneous dataset

|  |  |  |  |
| --- | --- | --- | --- |
| **Method** | **Descriptor set used** | | |
| **Combined** | **Basak lab** | **Diudea lab** |
| **PCR** | 0.59 (0.055) | **0.78 (0.038)** | 0.58 (0.057) |
| **PLS** | **0.86 (0.035)** | 0.85 (0.033) | 0.79 (0.038) |
| **Lasso** | 0.72 (0.048) | **0.75 (0.045)** | 0.63 (0.06) |
| **SCAD** | 0.57 (0.061) | 0.58 (0.059) | **0.62 (0.063)** |
| **RF** | **0.81 (0.036)** | 0.80 (0.042) | 0.79 (0.040) |
| **GBM** | 0.80 (0.04) | **0.82 (0.04)** | 0.75 (0.042) |

Table 3: Median and mean absolute deviations (in brackets) of Mean Square Prediction Error (MSPE) for different methods applied on the 95 amines dataset

|  |  |  |  |
| --- | --- | --- | --- |
| **Method** | **Descriptor set used** | | |
| **Combined** | **Basak lab** | **Diudea lab** |
| **PCR** | **29.11 (13.79)** | 57.08 (93.829) | 76.02 (24.72) |
| **PLS** | **18.86 (6.03)** | 19.86 (7.464) | 75.70 (24.689) |
| **Lasso** | **26.85 (9.049)** | 28.72 (8.825) | 72.75 (17.998) |
| **SCAD** | **25.81 (8.962)** | 31.77 (21.442) | 74.94 (18.322) |
| **RF** | **17.25 (6.498)** | 18.98 (6.587) | 84.59 (21.735) |
| **GBM** | **14.79 (5.836)** | 18.03 (6.296) | 74.78 (17.426) |

For all descriptor sets, PLS has the best performance among all methods, while Boosting performs the best for the 95 amines data.

Methods that depend directly on sparse linear combinations of predictors: Lasso, SCAD do not perform well in either case. This means there is high degree of nonlinearity among the relationship between the responses and predictors, and activities of compounds are more dependent on lower-dimensional subspaces in the predictor space than individual predictors.

PLS performs well in both cases. This implies there are low-dimensional substructures in the predictor spaces that are predictive of the responses.

For the 508 compounds dataset, more predictors do not always equate to better prediction. A reason for this can be the fact that this dataset is composed of chemical compounds from diverse classes. In comparison, the homogeneous 95 compound dataset always gives better prediction with the combined set of predictors than either group of predictors alone.

3.2. Principal Component Analysis of descriptor sets.

4 Discussion (To be completed later after results are written)

5) CONFLICT OF INTEREST

We confirm that there is no conflict of interest on the content of this paper.

6) ACKNOWLEDGEMENTS

7) SUPPLEMENTARY MATERIAL

Supplementary material (Supplementary tables 1-4) is available on the publisher’s web site along with the published article.

8) REFERENCES:

1. National Research Council. Toxicity Testing Strategies to Determine Needs and Priorities, National Academy Press: Washington, DC, 1984.
2. Toxic Substances Control Act (TSCA) Inventory: <https://19january2017snapshot.epa.gov/tsca-inventory/about-tsca-chemical-substance-inventory_.html>; Accessed on April 11, 2018.
3. Auer, C.M.; Nabholz, J.V. and Baetcke, K.P. Mode of action and the assessment of chemical hazards in the presence of limited data: use of structure-activity relationships (SAR) under TSCA, Section 5. Environ. Health Perspect. 1990, 87, 183-197.
4. Innovation in the pharmaceutical industry: New estimates of R&DcostsJoseph A. DiMasia, Henry G. Grabowski, Ronald W. Hansen, (2016) Innovation in the pharmaceutical industry: new estimates of R&D costs. J. Health Econ. 47, 20–33
5. Quantitative structure-activity relationship (QSAR) Models for Mutagens and Carcinogens; Romaualdo Benigni, Ed.; CRC Press, Boca Raton, FL, 2003.
6. Schrodinger ref (Claudiu to provide)
7. TopCluj ref (Claudiu to provide)
8. S. C. Basak, D. K. Harriss and V. R. Magnuson, "POLLY v2.3," Copyright of the University of Minnesota, 1988.
9. MolconnZ v4.05, Quincy, MA: Hall Ass. Consult., 2003.
10. S. C. Basak, G. Grunwald and A. Balaban, "TRIPLET," Copyright of the Regents of the University of Minnesota, 1993.
11. J. Stewart, MOPAC Version 6.00, QCPE #455, Frank J. Seiler Research Laboratory: US Air Force Academy, CO, 1990.