STUDIA UBB CHEMIA, LXII, 3, 2017 (p. 197-204)
(RECOMMENDED CITATION)
DOI:10.24193/subbchem.2017.3.16

In memory of prof. dr. Simion Gocan

DIMENSIONALITY OF BIG DATA SETS EXPLORED BY CLUJ DESCRIPTORS

CLAUDIU LUNGU[a], SARA ERSALI[a], BEATA SZEFLER[b],
ATENA PÎRVAN-MOLDOVAN[a], SUBHASH BASAK[c], MIRCEA V. DIUDEA[a *]

ABSTRACT. Dimensionality of a relatively big data set (95 compounds) observed for toxicity (mutagenicity) was explored in order to compute QSAR models. Distinct molecular descriptors were used. Dimensionality of data, using PCA, correlation plots and clustering, was evaluated. Analyzing data dimensionality allowed model optimization. Docking studies and PCA were used in order to expand data dimensionality. Pearson correlation coefficient (r^2) values, obtained for both perceptive and predictive models, were satisfactory.

Keywords: *topological descriptor, QSAR, data dimensionality, mutagenity, principal component analysis (PCA), Ames test*.

INTRODUCTION

In a data case, involving big data, one faces the curse of dimensionality, reflected by the minimum number of variables necessary to represent the data without any loss of information. A dataset in R_p is said to have Intrinsic Dimensionality (ID) equal to *m* if its elements lie entirely within an *m*-dimensional subspace of R_p (where $m < p$). In a multivariate statistical scenario, using methods like principal components analysis[1] (PCA), first few selected principal

[a] *Babes-Bolyai University, Faculty of Chemistry and Chemical Engineering, 400028 Cluj, Romania*
[b] *Nicolaus Copernicus University, Faculty of Pharmacy, Department of Physical Chemistry, Collegium Medicum, 85-096 Bydgoszcz, Poland*
[c] *University of Minnesota Duluth Natural Resources Research Institute and Department of Chemistry and Biochemistry, 5013 Miller Trunk Highway, Duluth, MN 55811, USA*
[*] *Corresponding author: diudea@chem.ubbcluj.ro*

[1] Jolliffe I.T. Principal Component Analysis, Series: Springer Series in Statistics, 2nd ed., Springer, NY, , XXIX, 487 p. 28 . ISBN 978-0-387-95442-4, **2002.**

components (PCs) explaining a reasonably high (90-95%) fraction of the variance in the original variables may be taken as an approximate measure of *m*. The abundance of data (big data) poses a challenge in many fields of chemometrics[2]. In toxicological research, strategies are manifold: grouping and classifying of data, searching of patterns and searching of correlations to biological activity, related to particular toxic endpoints[3]. One can use a perceptive model to evaluate a phenomenon occurrence. If occurrence is confirmed, a predictive model is used to find a best prediction[4].

RESULTS AND DISCUSSION

First a perceptive model was built using commercial descriptors, generated using 2D, 3D and ADME descriptors (which simulate the behavior of compounds in culture medias – used for toxicity). A selection algorithm led to the results shown in Table 1: the best model was obtained with 15 descriptors. The toxicity model equation is: $y = -0.0196617 + 0.9002748x$; $r^2 = 0.900$; $p = 0.946501$; $q^2 = 0.900$; RMSD = 0.604; it is plotted in Figure 1.

Table 1. Descriptors in the toxicity model (ordered in non-increasing Pearson r^2); (..) = no. descriptors

r^2 (..)	DESCRIPTOR
0.900 (15)	E_nb; E_stb; Gcut_Peoe_2; Gcut_SlogP_0; SlogP_VSA9; vsurf_HL1; vsurf_IW6; SMR_VSA6; logS; opr_nring; opr_nrot; opr_violation; radius; vsurf_CW5; vsurf_DD13.
0.892 (14)	E_nb; E_stb; Gcut_Peoe_2; Gcut_SlogP_0; SMR_VSA6; logS; radius; vsurf_HL1; vsurf_DD13; vsurf_IW6; SlogP_VSA9; opr_nring; opr_nrot; opr_violation.
0.887 (13)	E_nb; E_stb; Gcut_Peoe_2; SMR_VSA6; SlogP_VSA9; logS; opr_nring; opr_nrot; opr_violation; vsurf_DD13; vsurf_HL1; vsurf_IW6; radius.
0.880 (12)	E_stb; Gcut_Peoe_2; SMR_VSA6; SlogP_VSA9; logS; opr_nring; opr_nrot; opr_violation; radius; vsurf_DD13; vsurf_HL1; vsurf_IW6.
0.759 (4)	E_stb; logS; opr_nring; opr_nrot.
0.743 (3)	E_stb; logS; opr_nring
0.723 (2)	logS; opr_nring
0.690 (1)	opr_nring

[2] Hair, J. F. Jr., Anderson, R. E., Tatham, R. L. & Black, W. C. Multivariate Data Analysis (3rd ed). New York: Macmillan, **1995.**

[3] Wallace A.D. *Progress in Molecular Biology and Translational Sciences*, **2012,** *112*, 89.

[4] Basak, S.C.; Vraćko, M.; Witzmann, F.A. *Current Computer* Aided Drug Design, **2016,** *12*(4), 259**.**

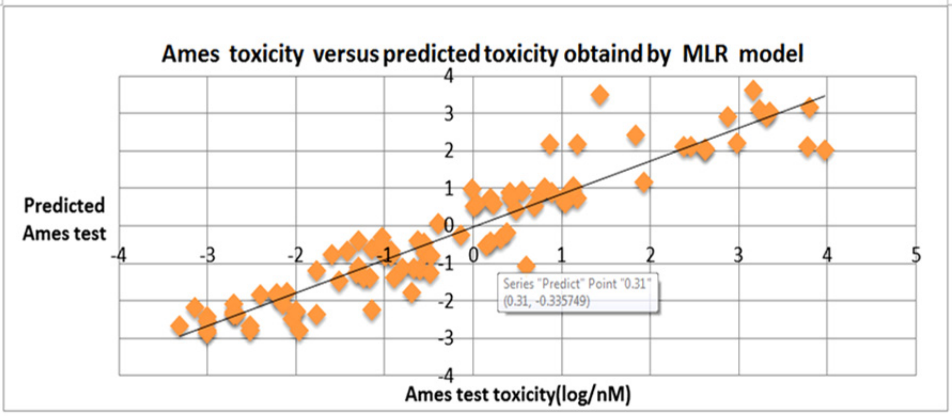

Figure 1. Correlation between observed Ames test values and predicted values.

Note that the descriptor opr_nring alone explains 69 % of the toxicity variance; thereby the aromatic nature of the compounds was further investigated, manly in a docking study (see below). As expected, this descriptor correlates with all other 14 descriptors, having a very low tolerance and increased inflation VIF values. Indeed, statistical insignificant values have all descriptors that describe aromatic properties: logS, opr_nring, vsurf_CW5 and vsurf_HL1. (Table 2, bolded values).

Table 2. Tolerance and VIF value calculated for the variables used in the model.

Descriptor	r^2 for each variable	Tolerance ($1-r^2$) (0.20 min. value)	VIF 1/Tolerance (4-20 max value)
E_nb	0.1756	0.824	1.213
E_stb	0.5473	0.452	2.212
Gcut_Peoe_2	0.7116	0.288	3.372
Gcut_SlogP_0	0.3399	0.660	1.515
logS	0.8998	**0.100**	**10.000**
opr_nring	0.8967	**0.103**	**9.708**
opr_nrot	0.7543	0.247	4.048
opr_violation	0.3205	0.679	1.472
Radius	0.7516	0.243	4.115
SlogP_VSA9	0.7427	0.257	3.891
SMR_VSA6	0.1871	0.812	1.231
vsurf_CW5	0.9698	**0.032**	**31.250**
vsurf_DD13	0.4878	0.512	1.950
vsurf_HL1	0.9628	**0.037**	**27.027**
vsurf_IW6	0.6796	0.320	3.125

Note that, in Figure 1, a “region” between 1 and 3 units where the Ames values are dispersed. It was assumed that relatively low value of r^2 is due to the insufficient description of the phenomena involved in toxicity.

As anticipated in the Methods section, a docking study, performed on a presumable target, transferase DNA fragment 3KHH, retrieved the results shown in Figure 2 . It is observed that the compound #53 has the favorable energy; its complex with 3KHH is represented in Figure 3.

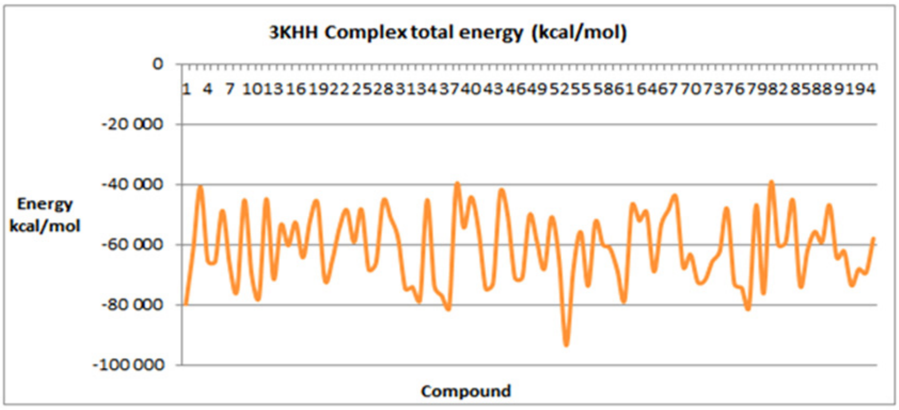

Figure 2. Docking energy data: the smooth lines represent total energy of the complex of amine compounds with the transferase DNA fragment (3KHH).

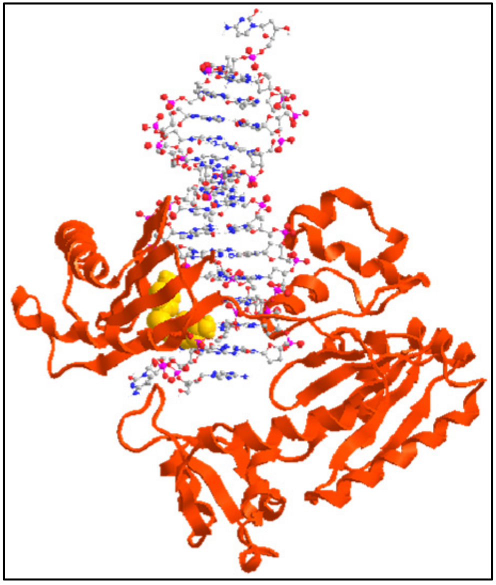

Figure 3. Ligand #53(space filling –yellow) in the complex with its presumable target 3KHH.

A linear model using Cluj topological descriptors and including docking data was computed. Docking data were explained 2% by the Cluj descriptors (in single variable), totally unsatisfactory. These descriptors better describe log P (in ligand aligned/oriented approach):

$y = 1.01 + SD_{logP}$(fragmental mass); $n = 92$, $r^2 = 0.77$ (three molecules were found as outliers).

To further increase the correlation value, a new set of Cluj topological descriptors (considering the heteroatoms) was computed; then, different types of models were generated. The models using only Cluj topological descriptors provided unsatisfactory results, irrespective what technique was used (e.g., MLR, NNR, SVM); among these, the best values, $r^2 = 0.583$, $p = 0.506$, $q^2 = 0.334$ were given by the NNR model.

In ligand orientated approach[5], a cluster correlation mapping[6] of the entire Cluj topological descriptors was performed. Correlations and disturbance in data dimensionality were observed (boded continuous red regions – Figure 4). These regions suggest that there is yet information that needs to be explored (eventually by using other descriptors). Receptor aligned/oriented approach is not appropriate manly because the real target and consecutively mutagenicity mechanism is not known.

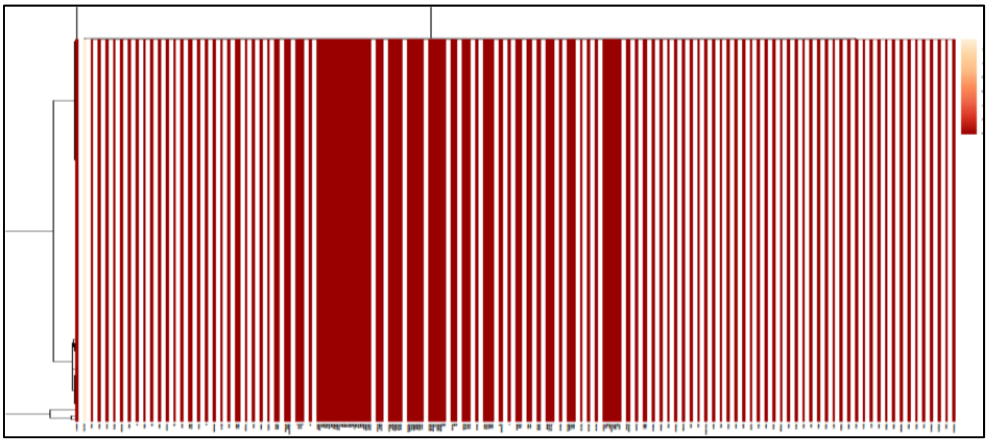

Figure 4. Cluj topological descriptors cluster correlation space. Confluent lines suggest correlation

[5] Deng Z, Chuaqui C, Singh J, *Journal of Medicinal Chemistry*. **2004,** 47 (2), 337**.**

[6] Campbell M.K., Grimshaw J.M., Elbourne D.R., *BMC Medical Research Methodology*, **2004,** *4*, 9.

In order to prove that data dimensionality can be improved by new descriptors, a predictive model, based on interactions between descriptors was developed. PCA was calculated for all 95 compounds TopoCluj descriptors set. A further selection algorithm was used to choose the independent variables; the number of descriptors used was 19. The descriptors are: C[Sh[CjMin]]; IP[CjMin]; PC10; PC11; PC12; PC13; PC16; PC17; PC2; PC22; PC3; PC4; PC5; PC6; PC7; PC8; PC9; X[LM[Electronegativity]]; X[LM[Mass]]. Model was computed using MLR: RMSD=0.000337, q^2=0.99, r^2=0.99 and plotted in Figure 5.

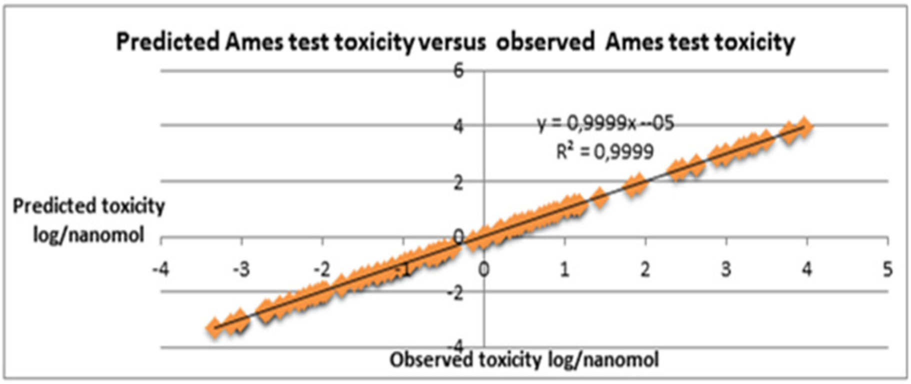

Figure 5. Plot of Observed toxicity vs predicted toxicity (mutagenicity), with 19 descriptors.

CONCLUSION

Data dimensionality can be explored using PCA. Models based on descriptors interactions include information of all descriptors of the chemical space. Models built using descriptors based on culture media simulations are superior in predicting occurrence of toxicity compared with the models developed on the basis of Cluj topological descriptors.

EXPERIMENTAL

In order to explore data dimensionality, a set of 95 amine compounds with observed Ames mutagenicity test (logC; nM) were used. QSAR methodology with related regression models was implemented for exploring data dimensionality. Two type of models were consider: (i) discriminant (perceptive) models, where collinearity and multicollinearity are avoided by using statistical tests applied to descriptors (like variability, tolerance and value of inflation (VIF)); (ii) predictive models, where collinearity and multicollinearity were not taken into account, the target being the r^2 value, witch in this case is not influenced by descriptors dimensionality. Correlation between observed and predicted data was studied.

Descriptors used for characterizing the data set were topological descriptors based on adjacency, connectivity and distance matrix and Cluj matrices, respectively. Using this methodology, 185 topological descriptors were computed for each compound using TopoCluj software. A future selection algorithm was used to select topological descriptors with relevant information regarding mutagenicity explored by Ames test.

Regression models were built using distinct methodologies: multiple linear regression (MLR), partial least square regression (PLS), support vector regression (SVR) and neural network regression (NNR). Models were validated internally, using the leave-one-out technique, and externally, by evaluating the test set. Compounds were randomly divided into a training and a test set. For the predictive model, interactions between descriptors were computed providing multiplicative cross-terms and principal component analysis (PCA).

Docking studies were performed on a hypothetical complex (DNA-protein-ligand) binding site located on DNA strings. Strings were retrieved form literature and from PDB data: 3KHH. Complex total energy (kcal/mol) was chosen to generate a new QSAR model in order to obtain a better r^2 using combined docking energy and Cluj topological descriptors. To explore deeper in data dimensionality a set of commercially available descriptors was computed and a regression model was compiled. Models from both descriptor type were compared.

ACKNOWLEDGEMENT

This work was supported by a grant of the Romanian National Authority for Scientific Research and Innovation, CCCDI – UEFISCDI, project number 8/2015, acronym GEMNS (under the frame of the ERA-NET EuroNanoMed II European Innovative Research and Technological Development Projects in Nanomedicine).

REFERENCES

1. Jolliffe I.T. Principal Component Analysis, Series: Springer Series in Statistics, 2nd ed., Springer, NY, , XXIX, 487 p. 28 . ISBN 978-0-387-95442-4, **2002.**
2. Hair, J. F. Jr., Anderson, R. E., Tatham, R. L. & Black, W. C. Multivariate Data Analysis (3rd ed). New York: Macmillan, **1995.**
3. Wallace A.D. *Progress in Molecular Biology and Translational Sciences*, **2012,***112*, 89.
4. Basak, S.C.; Vraćko, M.; Witzmann, F.A. *Current Computer* Aided Drug Design, **2016,** *12*(4), 259**.**
5. Deng Z, Chuaqui C, Singh J, *Journal of Medicinal Chemistry*. **2004,** 47 (2), 337**.**
6. Campbell M.K., Grimshaw J.M., Elbourne D.R., *BMC Medical Research Methodology*, **2004,** *4*, 9.
7. Norman R. Draper, Smith H., Applied Regression Analysis. Wiley, New York, **1998.**
8. Wold, S; Sjöström, M.; Eriksson,L., *Chemometrics and Intelligent Laboratory Systems* **2001,** *58*, 109.
9. Fisher, R.A., *Annals of Eugenics* **1936,** *7*, 179.
10. Fernández, S., Graves, A., Schmidhuber, J.*, In Proc. 20th Int. Joint Conf. on Artificial Intelligence, Ijcai:* **2007,** 774**.**
11. Kohavi, R., *Mateo*, C.A., Morgan K., *Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence. San*. **1995,** *2* (12), 1137**.**
12. San-Martin A1, Donoso V, Leiva S, Bacho M, Nunez S, Gutierrez M, Rovirosa J, Bailon-Moscoso N, Camacho SC, Aviles OM, Cazar ME, *Current Topics Medicinal* Chemistry, **2015,** *15*(17), 1743**.**
13. Gramatica P. *QSAR &Combinatorial Sc*ience **2007**