MARSplines approach for quantitative relationships between structure and pharmacological activity of potential drug candidates

Marcin Gackowski [1]*, Karolina Szewczyk-Golec [2] and Marcin Koba [1]

1 Department of Toxicology and Bromatology, Faculty of Pharmacy, L. Rydygier Collegium Medicum in Bydgoszcz, Nicolaus Copernicus University in Torun, A. Jurasza 2 Street, PL–85089 Bydgoszcz, Poland;

2 Department of Medical Biology and Biochemistry, Faculty of Medicine, L. Rydygier Collegium Medicum in Bydgoszcz, Nicolaus Copernicus University in Torun, Karłowicza 24 Street, PL–85092 Bydgoszcz, Poland;

*Correspondence: marcin.gackowski@cm.umk.pl (M.G)

Introduction:

Stevioside, one of the natural sweeteners extracted from stevia leaves, and its derivatives are considered to have numerous beneficial pharmacological properties, including the inhibition of activated coagulation factor X (FXa). FXa-PAR signaling is a possible therapeutic target to enhance impaired metabolism and insulin resistance in obesity. Acidic hydrolysis of stevioside affords a structural isomer of steviol, a tetracyclic diterpenoid isosteviol (ISV). Isosteviol-related compounds, possessing an ent-beyerane skeleton, have aroused interest because of their numerous pharmacological effects, including antibacterial, anticancer, anti-inflammatory, glucocorticoid agonist and cardioprotective properties [1]. Anthrapyrazoles are synthetic anticancer drugs, synthesized in order to retain high levels of wide spectrum of the antitumor activity of anthracyclines (e.g. doxorubicin) but at the same time to diminish cardiotoxicity by reducing the potential to generate semiquinone free radicals in cardiac cells. Apart from a broad range of antitumor activity in model tumors, they revealed diversified activity in doxorubicin-resistant cells. Attempts to reduce the toxicity of anthracyclines have led to the development of various anthrapyrazole derivatives with reduced side effects and increased efficacy in patients with breast cancer. Some of them even underwent clinical trials, where in phase II trials exhibited significant response rates in women with metastatic breast cancer [2].

The multivariate adaptive regression splines (MARSplines) was presented by Friedman as a method for flexible regression modeling of high dimensional data. This modern machine learning algorithm was successfully applied in QSAR and QSRR modeling approach in studies for drug activity prediction. MARSplines procedure was used for development of predictive QSAR models of various compounds with diverse pharmacological activities [1,2].

The goal of current investigation is to create models predicting the Fxa inhibitory activity of 20 isosteviol derivatives bearing thiourea fragments and antitumor activity of 73 anthrapyrazole derivatives as well as to evaluate the usefulness of MARSplines procedure for quantitative structure–activity relationship (QSAR) studies.

Molecular modeling and statistical analysis:

Geometry optimization was accomplished using semiempirical calculation with molecular mechanics (MM+) and Austin Model 1 (AM1) force fields as implemented in Hyper-Chem 8.0 (Hypercube Inc., Gainesville, FL, USA). The geometry of each compound was smoothly optimized with the MM+ molecular mechanics method and the resulting structure became an initial structure for the AM1 semiempirical method with the application of the Polak–Ribiere algorithm to a maximum energy gradient of 0.01 kcal (Å·mol)$^{-1}$. The optimization was performed for up to 30,000 steps. The examples of geometrically optimized structures are depicted in Figure 1. Calculation of molecular descriptors was performed using Dragon 7 (Talete, Milano, Italy) software.

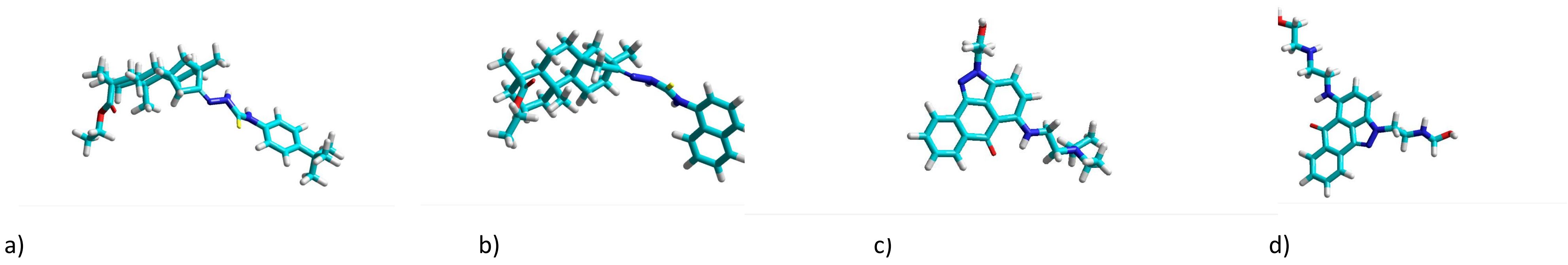

Figure 1. Geometrically optimized structures of selected isosteviol thiourea (a,b) and antrapyrazole (c,d) analogues.

The statistical analysis is based on the following data: descriptors encoding molecular properties of a particle and the values of the negative decimal logarithm of the half-maximal inhibitory concentration (IC_{50}) denoting FXa inhibitory activity and antitumor activity, obtained from the literature data. Statistica 13.3 software (StatSoft, Cracow, Poland) was used for this purpose. Raw data comprising about 5000 descriptors (acting as independent variables) and negative decimal logarithm values of the IC_{50} (pIC_{50}, dependent variable) underwent a process of standardization and pre-selection. In this step, almost half of descriptors with constant and near constant values, with standard deviation less than 0.0001 and with at least one missing value was excluded. The analyses were conducted at the 5% significance level (α = 0.05). The sets of isosteviol and anthrapyrazole compounds were divided into a training and a test sets on the basis of random sample selection in STATISTICA 13.3 Data Miner (StatSoft, Cracow, Poland). Building quantitative structure–activity relationship (QSAR) models involved applying a multivariate adaptive regression splines (MARSplines) algorithm, as implemented in STATISTICA 13.3 Data Miner. Initial evaluation of elaborated submodels led to the selection of a theoretical model suitable for predictive purposes. This assessment was performed on the basis of basic validation parameters calculated for each model (R^2, Q^2, MAE), which provided minimal but satisfactory information about model performance. The best submodel for each set of compounds, chosen for predictive purposes on the basis of abovementioned parameters, underwent full validation procedure with the parameters as follows: R^2, Q^2, Q_{F1}^2, Q_{F2}^2, Q_{F3}^2, CCC, Δr_m^2, $\overline{r_m^2}$, PRES, SDEP and *MAE*, which were calculated according to Roy et al. [3].

Results:

The MARS nonparametric procedure allowed for the establishment of a portfolio of QSAR submodels and a subsequent analysis of calculated validation parameters led to the selection of a submodel for each set of compounds that best describes the quantitative structure–activity relationship and may be employed to predict the FXa inhibitory activity of thiourea ISV derivatives as well as the antitumor activity of anthrapyrazole derivatives. Elaborated models reveal, which molecular properties affect the most the pharmacological activity of anthrapyrazole and isosteviol compounds. Among the independent variables appearing in the statistically significant MARS models, descriptors belonging 2D Atom Pairs, 2D autocorrelations, 3D-MoRSE, GETAWAY, burden eigenvalues, RDF and WHIM descriptors, may be distinguished. Reasonably high predictive performances of obtained models are confirmed by validation metrics. It should be emphasized that all calculated parameters exceed the threshold, what means that MARS models meet validation requirements for QSAR models (see Table 1).

Table 1. Values of validation parameters of the best MARS submodel

Parameter	Value		Threshold	Meaning [3]
	ISV-compounds	anthrapyrazoles		
$R^2 = 1 - \frac{\sum(Y_{obs}-Y_{cal})^2}{\sum(Y_{obs}-\bar{Y}_{training})^2}$	0.9985	0.9617	~1	a measure of the variation of observed with the predicted data
$Q^2 (or Q^2_{LOO}) = 1 - \frac{\sum(Y_{obs(training)}-Y_{pred(training)})^2}{\sum(Y_{obs(training)}-\bar{Y}_{(training)})^2}$	0.7922	0.9016	≥0.5	cross-validated R^2 (Q^2) tested for internal validation
$Q_{F1}^2 = 1 - \frac{\sum(Y_{obs(test)}-Y_{pred(test)})^2}{\sum(Y_{obs(test)}-\bar{Y}_{(training)})^2}$	0.9874	0.9119	≥0.5	it measures the correlation between the observed and predicted data of the test set
$Q_{F2}^2 = 1 - \frac{\sum(Y_{obs(test)}-Y_{pred(test)})^2}{\sum(Y_{obs(test)}-\bar{Y}_{(test)})^2}$	0.7927	0.90163	≥0.5	almost equal or closer values of $Q^2_{(F2)}$ and $Q^2_{(F1)}$ infer that the training set mean lies in the close propinquity to that of the test set
$Q_{F3}^2 = 1 - \frac{[\sum(Y_{obs(test)}-Y_{pred(test)})^2]/n_{test}}{[\sum(Y_{obs(train)}-\bar{Y}_{(train)})^2]/n_{train}}$	0.9706	0.7959	≥0.5	it is a measure of the model predictability
$CCC = \frac{2\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sum_{i=1}^{n}(x_i-\bar{x})^2+\sum_{i=1}^{n}(y_i-\bar{y})+n(\bar{x}-\bar{y})}$	0.9635	0.9496	~1	concordance correlation coefficient (CCC) measures both precision and accuracy, detecting the distance of the observations from the fitting line and the degree of deviation of the regression line from that passing through the origin, respectively
$\overline{r_m^2} = \frac{(r_m^2+r'^2_m)}{2}$ and $\Delta r_m^2 = \lvert r_m^2 - r'^2_m \rvert$, where $r_m^2 = r^2 \times \left(1-\sqrt{r^2-r_0^2}\right)$ $r'^2_m = r^2 \times \left(1-\sqrt{r^2-r'^2_0}\right)$ And parameters r^2 and r_0^2 are denoted as follows: $r_0^2 = 1 - \frac{\sum(Y_{obs}-k\times Y_{pred})^2}{\sum(Y_{obs}-\bar{Y}_{obs})^2}$ and $r'^2_0 = 1 - \frac{\sum(Y_{pred}-k'\times Y_{obs})^2}{\sum(Y_{pred}-\bar{Y}_{pred})^2}$ The terms k and k' are explained as follows: $k = \frac{\sum(Y_{obs}\times Y_{pred})}{\sum(Y_{pred})^2}$ and $k' = \frac{\sum(Y_{obs}\times Y_{pred})}{\sum(Y_{obs})^2}$	0.0196 and 0.9216	0.0173 and 0.9181	Δr_m^2 < 0.2 provided that the value of $\overline{r_m^2}$ 2 > 0.5	they reflect the overall predictability of the model for the whole data set
$PRESS = \sum(Y_{obs}-Y_{pred})^2$	0.8154	0.3446		assesses the model using the predicted residual sum of squares
$SDEP = \sqrt{\frac{PRESS}{n}}$	0.2020	0.1252		standard deviation of error of prediction (SDEP) is calculated from PRESS
$MAE = \frac{\sum\lvert Y_{obs}-Y_{pred}\rvert}{n}$	0.1017	0.0772		index of errors in the context of predictive modeling studies

Values of pIC_{50} (pIC_{50calc}) computed by the elaborated model were compared with the experimental data pIC_{50exp}) in the scatter plot, where a positive relationship is shown (see Figure1).

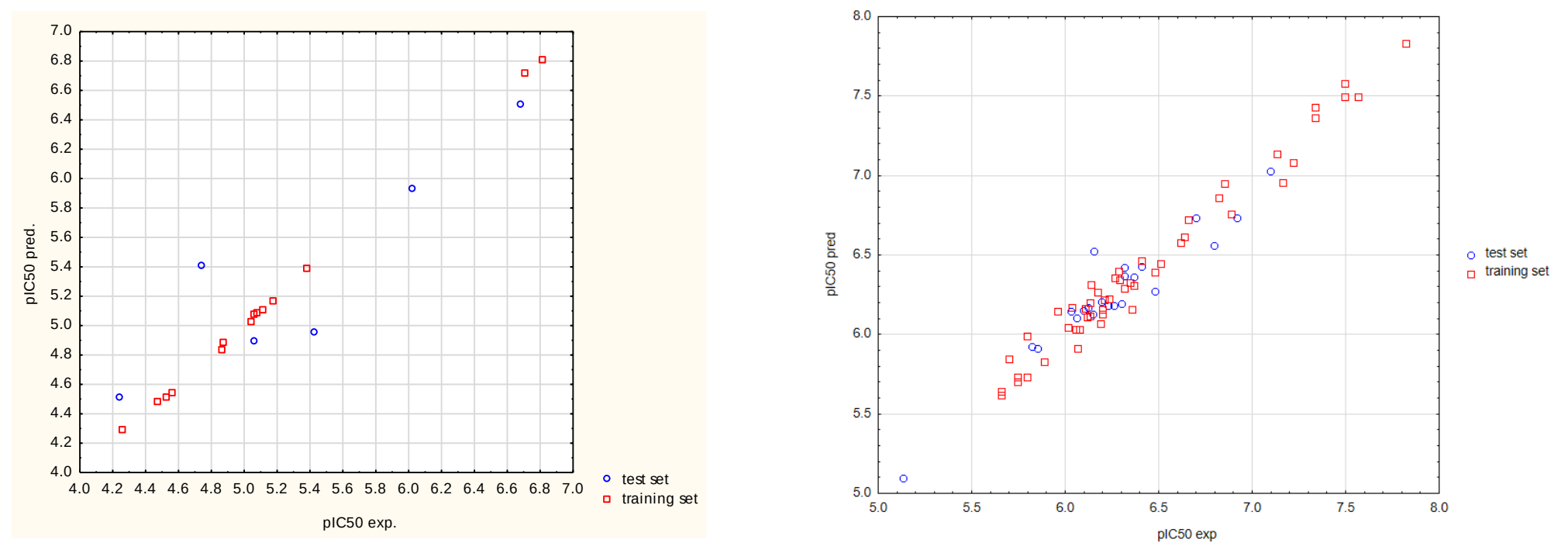

Figure 1. Correlation between the calculated and experimental FXa inhibitory activity of thiourea isosteviol analogues for training and test data sets. (pIC50 - negative decimal logarithm of the half-maximal inhibitory concentration)

Most of relevant descriptors describe the molecule's three-dimensional geometrical properties. For details see Table 2.

Table 2. Selected descriptors and the number of their appearances in the basis functions of the MARS model.

ISV analogues				
Symbol	Definition	Block	Dimensionality	Number in the Basis Function
B01[C-Cl]	Presence/absence of C-Cl at topological distance 1	2D Atom Pairs	2D	1
E2m	2nd component accessibility directional WHIM index/weighted by mass	WHIM * descriptors	3D	1
L3v	3rd component size directional WHIM index/weighted by van der Waals volume	WHIM descriptors	3D	1
Mor06i	signal 06/weighted by ionization potential	3D-MoRSE ** descriptors	3D	1
RDF070i	Radial Distribution Function—070/weighted by ionization potential	RDF *** descriptors	3D	1
HATS7s	leverage-weighted autocorrelation of lag 7/weighted by I-state	GETAWAY **** descriptors	3D	1

Anthrapyrazole analogues				
Symbol	Definition	Block	Dimensionality	Number in the Basis Function
Mor05s	signal 05/weighted by I-state	3D-MoRSE descriptors**	3D	9
Mor19m	signal 19/weighted by mass	3D-MoRSE descriptors**	3D	6
MATS8e	Moran autocorrelation of lag 8 weighted by Sanderson electronegativity	2D autocorrelations	2D	4
H1e	H autocorrelation of lag 1/weighted by Sanderson electronegativity	GETAWAY**** descriptors	3D	3
ATSC7v	Centred Broto–Moreau autocorrelation of lag 7 weighted by van der Waals volume	2D autocorrelations	2D	2
ATSC1e	Centred Broto–Moreau autocorrelation of lag 1 weighted by Sanderson electronegativity	2D autocorrelations	2D	2
SpMax8_Bh(s)	largest eigenvalue n. 8 of Burden matrix weighted by I-state	Burden eigenvalues	2D	2
Mor21e	signal 21/weighted by Sanderson electronegativity	3D-MoRSE descriptors**	3D	2
Mor13s	signal 13/weighted by I-state	3D-MoRSE descriptors**	3D	2
R5p	R autocorrelation of lag 5/weighted by polarizability	GETAWAY**** descriptors	3D	2
ATSC1s	Centred Broto–Moreau autocorrelation of lag 1 weighted by I-state	2D autocorrelations	2D	1
ATSC8s	Centred Broto–Moreau autocorrelation of lag 8 weighted by I-state	2D autocorrelations	2D	1
RDF135e	Radial Distribution Function—135/weighted by Sanderson electronegativity	RDF*** descriptors	3D	1
HATS5s	leverage-weighted autocorrelation of lag 5/weighted by I-state	GETAWAY**** descriptors	3D	1

- Weighted Holistic Invariant Molecular descriptors
- ** Molecular Representation of Structures based on Electron diffraction
- *** Radial Distribution Function
- **** Geometry, Topology and Atom-Weights Assembly

Conclusions:

A set of isosteviol thiourea derivatives was subjected to a molecular modeling study and an approach of MARSplines was employed for predicting FXa inhibitory activity. The developed QSAR model reveals information about the importance of the presence of chlorine atoms (B01[C-Cl]), the uniform distribution of the atomic mass (E2m), the molecular volume (L3v), the 3D molecular distribution of ionization potential (Mor06i and RDF070i) and the intrinsic properties of a molecule (HATS7s). Five out of six descriptors are geometrical descriptors quantifying three-dimensional aspects of molecular structure. Despite a relatively small set of studied compounds, the high application value of the obtained model was confirmed through an extensive validation protocol typical of QSAR models. Consequently, all calculated validation coefficients reflect the predictive power of regression. As FXa-PAR signaling is a possible therapeutic target to enhance impaired metabolism and insulin resistance in obesity, the predictive model may represent a valuable tool in searching for new active isosteviol analogues. Finally, the results of the present study confirmed an enhancement in pharmacological activity of isosteviol analogues by the presence of chlorine in the phenyl ring. Nevertheless, future studies are necessary to investigate the influence of a wider variety of substituents.

A quantitative structure–activity relationship study was also applied to a large set of anthrapyrazole compounds presenting antitumor activity against murine leukemia L1210. The approach of MARSplines was employed for prediction purposes, and was able to describe more than 96% of the variance in the experimental activity. This study has shown that fourteen parameters appearing in the statistically significant and extensively validated MARS model (i.e., descriptors belonging to 3D-MoRSE, 2D autocorrelations, GETAWAY, burden eigenvalues and RDF descriptors) significantly affect the antitumor activity of anthrapyrazole compounds. Moreover, this study confirmed the benefit of using the modern machine learning algorithm, namely, the MARSplines procedure, because the elaborated flexible model was also used in the prediction of antitumor activity against murine leukemia L1210 using an external set of seven anthrapyrazole compounds. Finally, in the light of the potential laying in such a large set of anthrapyrazole compounds, which still may be tested on various cell lines, and the high predictive power of the MARS model, the MARSplines procedure may be useful in the selection of the anticancer compounds of anthrapyrazoles for future clinical studies.

Both studies confirmed the benefit from using MARSplines algorithm, since high predictive power of obtained models make them useful for the prediction of antitumor and FXa inhibitory activity and possibly this approach can be considered as a tool for searching new drug candidates.

Acknowledgements:

With many thanks to Robert Pluskota, Katarzyna Mądra-Gackowska and Alina Woźniak for their support.

References:

1. Gackowski, M.; Golec, K.S.; Katarzyna, M.; Pluskota, R.; Koba, M. Quantitative Structure – Activity Relationship Analysis of Isosteviol - Related Compounds as Activated Coagulation Factor X (FXa) Inhibitors. 2022.

2. Gackowski, M.; Szewczyk-Golec, K.; Pluskota, R.; Koba, M.; Madra-Gackowska, K.; Woźniak, A. Application of Multivariate Adaptive Regression Splines (MARSplines) for Predicting Antitumor Activity of Anthrapyrazole Derivatives. Int. J. Mol. Sci. 2022, 23, doi:10.3390/ijms23095132.

3. Roy, K.; Ambure, P.; Kar, S.; Ojha, P.K. Is It Possible to Improve the Quality of Predictions from an "Intelligent" Use of Multiple QSAR/QSPR/QSTR Models? J. Chemom. 2018, 32, 1–18, doi:10.1002/cem.2992.

ECMC 2022

The 8th International Electronic Conference on Medicinal Chemistry

01–30 NOVEMBER 2022 | ONLINE