Journal of the Serbian Chemical Society
JSCS-info@shd.org.rs • www.shd.org.rs/JSCS

ACCEPTED MANUSCRIPT

Please cite this article as S. Torabi, F. Honarasa, S. Yousefinejad, *J. Serb. Chem. Soc.* (2020) https://doi.org/10.2298/JSC200611065T

J. Serb. Chem. Soc. 85 (0) 1-11 (2020)
JSCS–9591

UDC
Original scientific paper

Prediction of retardation factor of protein amino acids in reversed phase TLC and ethanol–sodium azide solution as mobile phase using QSRR

SUSAN TORABI[1], FATEMEH HONARASA[2] and SAEED YOUSEFINEJAD[3]*

[1]*Deputy of Food and Drug Control, Shiraz University of Medical Sciences, Shiraz, Iran;* [2]*Department of Chemistry, Shiraz Branch, Islamic Azad University, Shiraz, Iran;* [3]*Research Center for Health Sciences, Institute of Health, Department of Occupational Health Engineering, School of Health, Shiraz University of Medical Sciences, Shiraz, Iran*

(Received 11 June; revised 30 August; accepted 6 October 2020)

Abstract: Because of the importance of amino acids as the basic tiles of protein and their application in drug and food industries, there is a lot of interest in their separation and identification using simple and inexpensive approaches. Application of predictive models for determination of the behavior of AAs can reduce trial-and-error experiments. Here, the retardation factor (RF) of 21 protein AAs were studied using the quantitative structure-retardation factor (QSRR) model. The RF of the AAs in ethanol–sodium azide solution as the mobile phase of reversed phase thin layer chromatography (RP-TLC) was correlated with the AAs structural properties. The suggested QSRR indicated excellent fitting and prediction ability (R2train=0.95 and R2test=0.94). Furthermore, other statistical tests such as y-scrambling, cross validation and Williams plot confirmed the stability, absence of chance and the suitable applicability domain, respectively. It was shown that the sum of geometrical distances between oxygen and nitrogen atoms in AA molecule is an important factor in RF values of AAs in the ethanol–sodium azide.

Keywords: Natural amino acids; descriptors; structural property; thin layer chromatography; QSPR.

INTRODUCTION

Finding correlation and relationship between chromatographic retention time or retardation indices and structural parameters of the desired analysts has been an interesting subject because it is a way to obtain basic information on the impact of the structural features on the retention/retardation indices and to give an insight to possible mechanisms for separation forces and elution process[1,2].

The ideal objective of quantitative structure-retention relationships (QSRR) models, as a subfield in quantitative structure-retention relationships (QSPR), is

*Corresponding author E-mail: yousefisa@sums.ac.ir; Tel. +98 71 37251001 (Ext. 374); Fax: +98 71 37256006
https://doi.org/10.2298/JSC200611065T

its application to predict the retention behavior of newly identified molecules and similar synthesized derivatives or to estimate the involved mechanisms during various interactions such as solute–solute, solute–stationary phase and solute–mobile phase[3,4]. The prediction of retention/retardation of compounds in chromatographic systems using QSRR can reduce trial-and-error experiments which is important in some series of species such as polychlorinated biphenyls (PCBs), polybrominated, diphenyl ethers (PBDEs), polychlorinated dibenzofurans (PCDFs), *etc.* This can be significant in such compounds due to their toxicity or few information about some of their derivatives[5,6].

Different QSRR studies have been conducted on various chromatography methods such as reverse phase- and normal phase- high performance chromatography (RP-HPLC and NP-HPLC), micellar HPLC, gas chromatography, high performance affinity chromatography, immobilized artificial membrane, and planar chromatography which is reviewed in different articles[5,7,8].

RP-TLC is one of the most popular techniques and its retention mechanism is based on partitioning of the compounds between hydrophilic mobile phase and hydrophobic stationary phase. Because of the nature and mechanism of RP-TLC, the lipophilicity of the analytes has effect on their retention with a strict correlation[9].

Analysis and identification of amino acids (AAs) is of high importance because the AAs are the basic units of biomolecules, enzymes, peptides, and protein and have impacts in food and drug industries. Separation and identification of AAs (protein or non-protein) have been intensively performed with column chromatography but their separation using simple methods such as thin layer chromatography (TLC) have been considered in many studies[10,11]. In a previous study, a QSRR model was reported for separation of AAs in Normal Phase TLC (NP-TLC) by considering both properties of mobile phase and AAs structures[12]. To show the ability of QSRR in modeling the behavior of AAs in RP-TLC, a QSRR modeling was recently designed in two different mobile phase and the potential of structure-chromatography was confirmed in this case[13]. But because the recent work was done by considering only AAs structure, more studies is required to show the capability of QSRR approach in separation of AAs in RP-TLC in various mobile phases. Thus, in the current study, a QSRR model was developed for the retardation factor (R_F) of 21 protein AAs in RP-TLC using the ethanol–sodium azide solution as a well-known mobile phase to identify the significant features in this elution process.

MATERIALS AND METHODS

The retardation factor (R_F) the steudied AAs using RP-TLC in ethanol–sodium azide were adopted from literature [10] which are shown in Table I.

TABLE I. Experimental and predicted R_F of 21 AAs in ethanol–sodium azide and related residual values.

code	Name	R_F (Exp)	R_F (Pred)	Residual
AA 1	Glycine	0.82	0.84	0.02
AA 2	Alanine	0.82	0.78	-0.04
AA 3	Aspartic acid	0.25	0.35	0.10
AA 4	Arginine	0.13	0.14	0.01
AA 5	Proline	0.78	0.65	-0.13
AA 6	Hydroxyproline	0.84	0.88	0.04
AA 7	Lysine	0.02	0.02	0.00
AA 8	Glutamic acid	0.86	0.74	-0.12
AA 9[a]	Serine	0.8	0.74	-0.06
AA 10	Tryptophan	0.83	0.87	0.04
AA 11	Valine	0.83	0.89	0.06
AA 12	Phenyl alanine	0.83	0.90	0.07
AA 13	Isoleucine	0.88	0.84	-0.04
AA 14[a]	Leucine	0.88	0.97	0.09
AA 15	Asparagine	0.62	0.60	-0.02
AA 16	Methionine	0.83	0.84	0.01
AA 17	Cysteine	0.84	0.88	0.04
AA 18[a]	Histidine	0.33	0.35	0.02
AA 19	Threonine	0.83	0.80	-0.03
AA 20[a]	Tyrosine	0.83	0.83	0.00
AA 21[a]	Glutamine	0.77	0.72	-0.05

[a]AA samples used as the test set

The structural properties were produced form 22 categories of descriptors such as topological, constitutional, topological charge indices, geometrical, connectivity, RDF, 3D MoRSE, WHIM, GETAWAY, functional group counts, and some other groups were extracted[14]. The extracted descriptors were form variety kind of features to cover structural details of natural AAs. The structural descriptors were arranged in a matrix (D) with size of 21×c where c denotes number of total utilized descriptors. Then, constant and near constant columns from this matrix was deleted from to remove redundant information. Collinear column in D were also removed after calculating the correlation of descriptors with R_F vector and with other descriptors. Finally, among a detected pair of collinear columns, one with the lowest correlation with the R_F vector was eliminated from D matrix. After these refining process, 404 descriptors were retained in D (size=21×404) for further process and variable selection. In the current study, the structures of natural AAs were drawn using Hyperchem software (Version 7, Hypercube Inc., http://www.hyper.com, USA) and AM1 semi-empirical method were applied during optimization. Different categories of the structural features were extracted using DRAGON (http://michem.disat.unimib.it/chm/; Milano Chemometrics and QSAR research group) for all AAs.

In the next step, the constructed data matrix (D) was divided in to test and training subsets and the training set was targeted by variable selection and further model development. Stepwise multiple linear regression (SMLR) was applied as the variable selection method and squared correlation coefficient of training set (R^2_{train}) and cross validation (Q^2_{CV}) was criteria for choosing final model[15–17]. In addition to cross validation, different statistical approaches such as y-

scrambling and prediction of a small portion of AAs, as the external test set, were utilized to evaluate the prediction ability and stability of suggested QSRR model in RP-TLC[18,19].

All statistical and calculations tasks were carried out via MATLAB software (version 7.7, R2008b, Math work, Inc., http://mathworks.com, USA). A personal computer under the Windows 7 operating system was used to run all software.

RESULTS AND DISCUSSION

After spiliting data in to the training and test sets of samples using random selection from PCA space (Fig. 1a) and variable selection using stepwise MLR, different models were constructed based on 1 to 8 descriptors of AAs and the results of fitting and cross validatio is represented in Fig. 1-b (supproting information). As it can be observed in Fig. 1-b, a QSRR contained 5 descriptors was utillized as the optimum one, after autoscaling the X-matrix (descriptors) and y-vector (RTLC retardation factor), which is shown in equation 1.

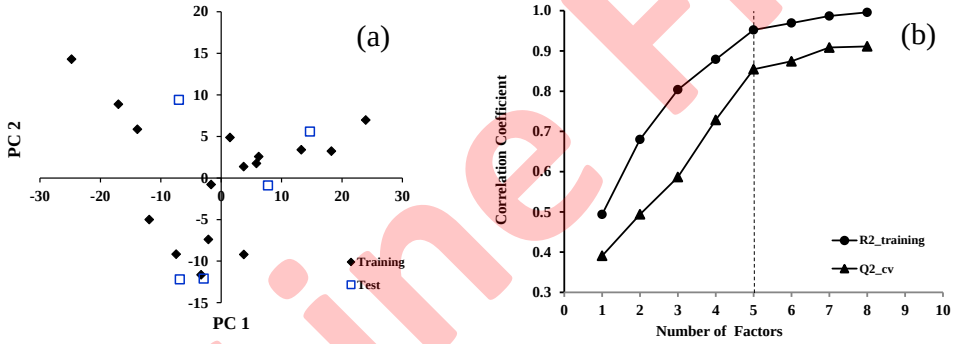

Fig. 1. Two dimensional-PCA plot of the space of total descriptors in AA chromatographic samples (a), Correlation coefficient of the training set (R^2_{train}) and cross-validation (Q^2) versus number of descriptors to select the optimum number of factors (b) for suggested RP-TLC model in ethanol–sodium azide solution.

$$R_{F(\text{ethanol-sodium azide})} = -2.34\ (\pm 0.584) - 0.19\ (\pm 0.003)\ G_{(N..O)} - -1.594\ (\pm 0.211)\ \text{Mor24u+}\ 6.661\ (\pm 1.080)\ \text{PW2} + 1.018\ (\pm 0.200)\ \text{Mor28u} - 0.619\ (\pm 0.158)\ \text{SEige} \quad (1)$$

In the above suggested QSRR in ethanol–sodium azide solution $G_{(N..O)}$ denotes the sum of geometrical distances between Nitrogen and Oxygen atoms in AA structure. Mor24u is the unweighted three dimentional Molecular Representation of Structures based on Electronic diffraction (3D-MoRSE) descriptors of signal 24, PW2 denotes the path/walk 2-Ranic shape index, Mor28u corresponds to the unweighted 3D-MoRSE information of signal 28 and SEige is the Eigenvalue sum from electronegativity weighted distance matrix[14]. The definitions and categories of the utilized descriptors are also summarized in Table II.

TABLE II. Class and definition of descriptors used in model 1 and 2

No.	Name	Class	Definition
1	$G_{(N..O)}$	3D Atom Pairs	Sum of geometrical distances between N..O
2	Mor24u	3D-MoRSE descriptors	3D-MoRSE descriptors signal 24 / unweighted
3	PW2	Topological indices	path/walk 2 - Randic shape index
4	Mor28u	3D-MoRSE descriptors	3D-MoRSE descriptors signal 28 / unweighted
5	SEige	Eigenvalue-based indices	Eigenvalue sum from electronegativity weighted distance matrix

To have better prediction of R_F values, all five included descriptors and experimental vector of R_F were targeted by mean-centering and scaling (autoscaled)[12,20]. After these pre-treatment, the linear model was re-computed based on these prepared data and an equation 2 contained standardized MLR-coefficients were obtained.

$$R_{F(\text{ethanol-sodium azide})} = -0.72\ (\pm 0.098)\ G_{(N..O)} - 0.540\ (\pm 0.072)\ \text{Mor24u} + \\ + 0.461\ (\pm 0.075)\ \text{PW2} + 0.457\ (\pm 0.090)\ \text{Mor28u} - \\ 0.410\ (\pm 0.105)\ \text{SEige} \qquad (2)$$

$$R^2_{train}=0.95,\ F=39.85,\ RMSE_{train}=0.22,\ F_{crit(95\ \%)}=3.33$$

As it is shown in the above results, the significant higher value of F-statistic in comparison F_{crit}, confirms the significance of the above QSRR developed in ethanol–sodium azide solution. In addition to R^2_{train} of model 2, which was equal to 0.95, Q^2_{LOO} of cross validation was equal to 0.85 which shows suitable fitness and enough stability of model 2[21]. In Eqs. (1) and (2), the standard 3of the five included descriptors because they have very low value than the calculated coefficients. Magnitude of the coefficient of each descriptor of AAs can be used to conclude about the effect of that descriptor on R_F of natural AAs in the studied mobile phase (ethanol–sodium azide solution).

Details and explains of all structural descriptors in the suggested models will be presented in the next sections. The numerical value of $G_{(N..O)}$, Mor24u, PW2, Mor28u and SEige is shown in Supplementary material (Table S-1).

In this model, $RMSE_{test}$ and R^2_{test} were equal to 0.26 and 0.94 respectively. Thus, the results showed a good agreement between predicted and experimental R_F in 21 AAs in training and test samples. Y-scrambling was applied and Q^2_{MP} of this test was 0.25, which confirms that a chancy QSRR model in ethanol–sodium azide was not constructed. The predicted values of retardation factor of 21 free amino acids using model are presented in Table I and the statistical parameters of this QSRR model are shown in Table III. As it is shown in the Fig. 2-a, a suitable correlation between experimental and predicted R_F values of amino acids was obtained. In addition to the denoted statistics, some other criteria suggested by Golbraikh and Tropsha was calculated for the developed QSRR model[21,22]. Some of the important and well-known metrics was calculated for the suggested model is denoted in the following and their threshold value for a valid model is also mentioned in eqs 3-5.

$$|R_0^2 - R'_0{}^2| = 0.00597 \text{ (Threshold value} < 0.3\text{, Passed!)} \quad (3)$$

$$k = 1.00073 \text{ and } (|R_0^2 - R'_0{}^2|/R^2) = 0.00094$$
$$\text{(Threshold value: } [0.85<k<1.15 \text{ and } |R_0^2 - R'_0{}^2| / R^2 < 0.1]\text{, Passed!)} \quad (4)$$

$$k' = 0.99132 \text{ and } (|R'_0{}^2 - R_0^2|/R^2) = 0.00766$$
$$\text{(Threshold value: } [0.85< k' <1.15 \text{ and } |R'_0{}^2 - R_0^2|/R^2 < 0.1]\text{, Passed!)} \quad (5)$$

The details of calculation of these parameters can be found in original references [21] but as can be seen the constructed QSRR passed these validation criteria. Other criteria were also suggested by Roy *et al.* to show the prediction ability known as average r^2_m or $\overline{r^2_m}$ and delta r^2_m or Δr^2_m [22]. $\overline{r^2_m}$ should be higher than 0.5 and Δr^2_m must be lower than 0.2 in a valid method[23]. For the suggested QSRR shown in Eq. (2), $\overline{r^2_m}$ and Δr^2_m of test set were 0.845 and 0.0402 respectively which indicate the validity of agreement between predicted and actual solubility in the test set. However, it is noteworthy that the small number of AAs in test set is a limitation but was done because of limited number of experimental data.

TABLE III. Statistical performance of the proposed QSRR of the RP-TLC samples of amino acids (in ethanol–sodium azide) using three different random-divided test and training sets.

Training-test set	N_{train}[a]	N_{test}[b]	R^2_{train}[c]	RMSE$_{trai}$[d]	Q^2_{LOO}[e]	RMSE$_{cv}$[f]	R^2_{test}[g]	RMSEP[h]	Q^2_{MP}[i]
1[A]	16	5	0.95	0.22	0.85	0.35	0.94	0.26	0.25
2[B]	16	5	0.94	0.23	0.84	0.40	0.91	0.35	0.31
3[C]	16	5	0.95	0.23	0.84	0.41	0.91	0.33	0.16

[a]Number of RP-TLC runs in training set; [b]Number of RP-TLC runs in test set; [c]Correlation coefficient of the training RP-TLC runs; [d]Root mean square error of training RP-TLC runs (Calibration); [e]Correlation coefficient of leave-one-out cross-validation; [f]Root-mean-square errors of leave-one-out cross-validation; [g]Correlation coefficient of the test RP-TLC runs; [h]Root-mean-square errors of the test RP-TLC runs; [i]Maximum cross-validation correlation coefficient for 30 Y-randomization test
[A]Number of solvents in the test set as indicated in Table I: AA9, AA14, AA18, AA20, AA21
[B]Number of solvents in the test set: AA2, AA10, AA12, AA18, AA20
[C]Number of solvents in the test set: AA1, AA2, AA12, AA16, AA18

Another important point should be considered in evaluation of QSRR is its independency from the AAs which has been used as the training set or reserved as the test set. To ensure about this independency, two other subsets of TLC runs were randomly chosen from the RP-TLC samples as the training sets and two rest subsets (contained 5 AAs) as the test sets which was shown in results as train-test 2 and train-test 3. As it is represented in Table III, change in the AAs of training or test set has not effect on the goodness of fit or prediction of QSRR model for the RP-TLC using ethanol–sodium azide.

Applicability domain and Pair/Multi correlation

As a well-known recommendation, the suggested QSRR model must exclude any linear dependency between the AAs descriptors. This necessity is because of this serious limitation of accuracy in models with collinearity and cannot be

useful to justify the chromatographic behavior of analytes in RP-TLC during elution with ethanol–sodium azide. Moreover the presentation of collinear descriptors led to obtaining wrong signs for the coefficients of the QSRR[16,24]. The correlation between each pair of descriptors in five utilized AAs descriptors was done and the matrix of pair correlation for our QSRR model is shown in Table IV which indicates no high correlation in this model.

TABLE IV. Pair correlation matrix for descriptors in developed QSRR model for ethanol–sodium azide and related VIF values as the index of multi-collinearity

	$G_{(N..O)}$	Mor24u	PW2	Mor28u	SEige	**VIF**
$G_{(N..O)}$	1.00					2.03
Mor24u	0.00	1.00				1.07
PW2	0.00	0.05	1.00			1.17
Mor28u	0.35	0.00	0.00	1.00		1.68
SEige	0.45	0.00	0.03	0.42	1.00	2.30

Not only the existence of pair-correlation, but also the multi-collinearity (*i.e.* collinearity of one with all others) in the QSRRs is also a risk for model accuracy and variance inflation factor (VIF) is a good metrics to evaluate such collinearity[25]. The high multi-collinearity can hide some of structural information because of overlapping in independent variables[25]. As can be observed in Table IV, the calculated VIF of all utilized descriptors ($G_{(N..O)}$, Mor24u, PW2, Mor28u and SEige*)* are lower than critical value 5.0[25] which shows that our QSRR does not suffer from risk of multi-collinearity. Thus, we can trust the sign of coefficient and their magnitude in suggested QSRR to justify the effect of selected structural properties of AAs on their R_F.

After different statistical evaluation of QSRR, the leverage and standardized residual of AAs samples were calculated to represent the applicability domain (AD)[26]. Considering AD can clarify the limitations and potential of the developed QSRR for these AAs which might be useful for RP-TLC of similar structures derived from AAs[19]. AA samples with leverage below the cut-off value and with standardized residual within the logical range can be considered to be in normal AD. Standardized residual bigger than -3σ or lower than $+3\sigma$ is a suitable value[27,28]. Also, cut-off value of leverage is $h^*=3(d+1)/n_{train}$[27] in which d denotes number of descriptors in QSRR model (here equal to 5) and n_{train} shows the number of AAs in the training subset (here equal to 16). According to what was explained, h^* of our QSRR was calculated equal to 1.125.

According to Fig. 2b which is the Williams plot, for representation of both standardized residual and leverage, all the studied chromatographic samples were within the AD of model, suggested for RP-TLC using ethanol-sodium azide.

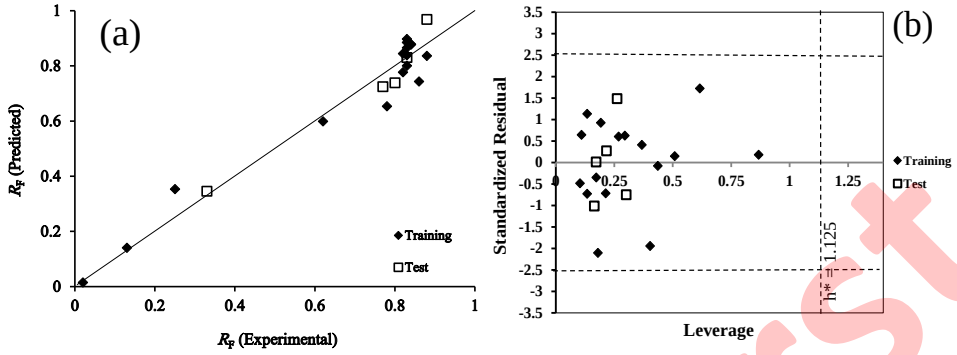

Fig. 2. Plot of predicted R_F versus experimental values in 21 investigated AA samples using the five parametric RP-TLC model in ethanol –sodium azide (**a**) Applicability domain of the set of AAs shown by Williams plot (b) Cut off values of the standardized residual (±2.5 times the standards deviation) and the leverage (h^*) are illustrated by horizontal and vertical dashed lines, respectively. All samples are located within the applicability domain.

Interpretation of model

In this model, $G_{(N..O)}$ was the first descriptor with negative effect on the R_F of AAs in the ethanol-sodium azide which confirmed the importance of "Sum of geometrical distances between N and O" ($G_{(N..O)}$). It should be emphasized that the effect of $G_{(N..O)}$ was also illustrated in our previous research which was significant in separation of amino acids in normal phase TLC (NP-TLC) with negative effect on R_F of amino acids[12].

Two unweighted 3D-MoRSE descriptors named Mor24u and Mor28u (signal 26 and signal 24)[14] are imported in the model with negative and positive sign of coefficients. The presence of 3D-MoRSE indices in this model and previous work on this subject show the potential of this category of descriptors in prediction of the retardation factors of amino acids in RP-TLC[13]. On the other hand, different signs of these two 3D-MoRSE descriptors indicate their complex contribution in retardation of samples using ethanol-sodium azide. The other descriptor in model is *PW*2 which is categorized in topological indices and shows path/walk 2 - Randic shape index of the amino acids[14]. The positive contribution of *PW*2 in model illustrates the direct relationship of *PW*2 and R_F of amino acids during separation with RP-TLC using ethanol-sodium azide. SEige is the last parameter in model which is an Eigenvalue-based index with negative sign. SEige denotes the Eigenvalue sum from electronegativity weighted distance matrix[14] and thus the increasing this index in amino acids can enhance their retardation factor in ethanol-sodium azide system.

CONCLUSION

QSRR as a basic field in chromatography is a tool for showing the effect of the molecular structure of analytes on their chromatographic behavior. On the

other hand because of the effect of other parameters such as stationary and mobile phase in separation, this work focused on the modeling the R_F of protein AAs in RP-TLC during elution with ethanol-sodium azide.

One of findings in our study was the impact of sum of geometrical distances between N and O on R_F value of AAs in RP-TLC using ethanol-sodium azide. It was found that decreasing the sum of this distance can increase the remaining of AAs on TLC plate and also their R_F in ethanol-sodium azide. This fact was in accordance with previous report on the Normal phase TLC of AAs. Eigenvalue sum from electronegativity weighted distance matrix and two 3D-MoRSE properties form AAs had also important effect on R_F in the investigated system.

Moreover, different statistical evaluation on training, cross validation, prediction, y-randomization and applicability domain confirmed the stability and accuracy of the suggested QSRR. However small number of compounds in training and test sets can be denoted as a limitation of work but it is noteworthy that the goal of current modeling was not only external prediction but also was on the chemical/structural description of the chromatographic behavior of AAs. This work can give more information for explaining the separation of AAs, in continue to previous studies on other mobile phases and can be completed with more studies in future.

SUPPLEMENTARY MATERIAL

Table S1 (Numerical vales of original descriptor used in model Eq. 1) is presented in Supplementary Material available from Journal Web site http://www.shd.org.rs/JSCS/, or from the corresponding author on request.

Acknowledgments: Supported of Shiraz University of Medical Sciences (Grant No. 99-01-04-23865) is gratefully acknowledged.

ИЗВОД

ПРЕТСКАЗИВАЊЕ ФАКТОРА ЗАДРЖАВАЊА ПРОТЕИНСКИХ АМИНО КИСЕЛИНА У РЕВЕРСНО-ФАЗНОЈ ТАНКОСЛОЈНОЈ ХРОМАТОГРАФИЈИ СА ЕТАНОЛ-НАТРИЈУМ АЗИДОМ КАО МОБИЛНОМ ФАЗОМ КОРИСТЕЋИ КВАНТИТАТИВНУ РЕЛАЦИЈУ СТРУКТУРЕ И ЗАОСТАЈАЊА (QSRR)

SUSAN TORABI[1], FATEMEH HONARASA[2] и SAEED YOUSEFINEJAD3

[1]Deputy of Food and Drug Control, Shiraz University of Medical Sciences, Shiraz, Iran; [2]Department of Chemistry, Shiraz Branch, Islamic Azad University, Shiraz, Iran; [3]Research Center for Health Sciences, Institute of Health, Department of Occupational Health Engineering, School of Health, Shiraz University of Medical Sciences, Shiraz, Iran

Због значаја аминокиселина као основних циглица протеина и њихове примене у индустрији лекова и хране, постоји велико занимање за њихово раздвајање и идентификацију коришћењем простих и јефтиних приступа. Примена предиктивних модела за одређивање понашања АК може скратити експерименте покушаја-и-грешке. Овде су фактори заостајања (R_F) 21 протеинске амино киселине проучавани користећи квантитативни структура-фактор заостајања (QSRR) модел. R_F-ови аминокиселина у раствору етанола–натријум азида као мобилне фазе танкослојне

реверсно-фазне хроматографије (RP-TLC) су корелисани са структурним особинама аминокиселина. Сугерисани QSRR указује на изврсно фитовање и способност предвиђања (R_{2train} = 0,95 и R_{2test} = 0,94). Надаље, остали статистички тестови као што су 'y-scrambling', унакрсна валидација, Уиљемсов график, потврђују стабилност, одсуство случајности, односно погодан домен оримењивости. Показано је да је збир геометријских удаљености атома кисеоника и азота у аминокиселинама значајан фактор за RF вредности аминокиселина у етанол–натријум азиду.

(Примљено 11. јуна; ревидирано 30. августа; прихваћено 6. октобра 2020)

REFERENCES

1. S. H. Park, M. De Pra, P. R. Haddad, S. Grosse, C. A. Pohl, F. Steiner, *J. Chromatogr. A* **1609** (2020) 460508 (https://dx.doi.org/10.1016/j.chroma.2019.460508)
2. A. M. Ramezani, S. Yousefinejad, A. Shahsavar, A. Mohajeri, G. Absalan, *J. Chromatogr. A* (2019) (https://dx.doi.org/10.1016/j.chroma.2019.03.063)
3. J. M. Sutter, T. A. Peterson, P. C. Jurs, *Anal. Chim. Acta* **342** (1997) 113 (https://dx.doi.org/10.1016/S0003-2670(96)00578-8)
4. Y. Marrero-Ponce, S. J. Barigye, M. E. Jorge-Rodríguez, T. Tran-Thi-Thu, *Chem. Pap.* **72** (2018) 57 (https://dx.doi.org/10.1007/s11696-017-0257-x)
5. C. Giaginis, A. Tsantili-Kakoulidou, *Chromatographia* **76** (2013) 211 (https://dx.doi.org/10.1007/s10337-012-2374-6)
6. J. Dai, L. Jin, S. Yao, L. Wang, *Chemosphere* **42** (2001) 899 (https://dx.doi.org/10.1016/S0045-6535(00)00181-8)
7. K. Héberger, *J. Chromatogr. A* **1158** (2007) 273 (https://dx.doi.org/10.1016/J.CHROMA.2007.03.108)
8. R. Kaliszan, *Chem. Rev.* **107** (2007) 3212 (https://dx.doi.org/10.1021/cr068412z)
9. R. Kaliszan, *J. Chromatogr. A* **220** (1981) 71 (https://dx.doi.org/10.1016/S0021-9673(00)98504-2)
10. D. Kaźmierczak, W. Ciesielski, R. Zakrzewski, *JPC - J. Planar Chromatogr. - Mod. TLC* **18** (2005) 427 (https://dx.doi.org/10.1556/JPC.18.2005.6.5)
11. T. Hudaib, S. Brown, D. Wilson, P. E. Eady, *JPC - J. Planar Chromatogr. - Mod. TLC* **29** (2016) 145 (https://dx.doi.org/10.1556/1006.2016.29.2.9)
12. S. Yousefinejad, F. Honarasa, N. Saeed, *J. Sep. Sci.* **38** (2015) 1771 (https://dx.doi.org/10.1002/jssc.201401427)
13. S. Yousefinejad, F. Honarasa, S. Akbari, M. Nekoeinia, *J. Liq. Chromatogr. Relat. Technol.* (2020) 1 (https://dx.doi.org/10.1080/10826076.2020.1774388)
14. R. Todeschini, V. Consonni, *Molecular Descriptors for Chemoinformatics*, Second, WILEY-VCH, Weinheim, 2009 (ISBN: 9783527318520)
15. P. Gramatica, *Mol. Inform.* **33** (2014) 311 (https://dx.doi.org/10.1002/minf.201400030)
16. S. Yousefinejad, B. Hemmateenejad, *Chemom. Intell. Lab. Syst.* **149** (2015) 177 (https://dx.doi.org/10.1016/j.chemolab.2015.06.016)
17. S. Yousefinejad, F. Honarasa, A. Solhjoo, *J. Chem. Eng. Data* **61** (2016) 614 (https://dx.doi.org/10.1021/acs.jced.5b00768)
18. J. U. N. Shao, *J. Am. Stat. Assoc.* **88** (1993) 486 (https://dx.doi.org/10.2307/2290328)
19. P. Gemperline, *Practical Guide to Chemometrics*, 2nd ed., Taylor & Francis Group, Boca Raton, 2006 (ISBN: 1574447831)
20. F. Honarasa, S. Yousefinejad, S. Nasr, M. Nekoeina, *J. Mol. Liq.* **212** (2015) 52 (https://dx.doi.org/10.1016/j.molliq.2015.08.055)

21. A. Golbraikh, A. Tropsha, *Mol. Divers.* **5** (2000) 231 (https://dx.doi.org/10.1023/A:1021372108686)
22. K. Roy, R. N. Das, P. Ambure, R. B. Aher, *Chemom. Intell. Lab. Syst.* **152** (2016) 18 (https://dx.doi.org/10.1016/j.chemolab.2016.01.008)
23. K. Roy, P. Chakraborty, I. Mitra, P. K. Ojha, S. Kar, R. N. Das, *J. Comput. Chem.* **34** (2013) 1071 (https://dx.doi.org/10.1002/jcc.23231)
24. L. Eriksson, J. Jaworska, A. P. Worth, M. T. D. Cronin, R. M. McDowell, P. Gramatica, *Environ. Health Perspect.* **111** (2003) 1361 (https://dx.doi.org/10.1289/ehp.5758)
25. T. A. Craney, J. G. Surles, *Qual. Eng.* **14** (2002) 391 (https://dx.doi.org/10.1081/QEN-120001878)
26. R. Todeschini, V. Consonni, P. Gramatica, M. Descriptors, H. Approach, G. C. Methods, C. S. Analysis, R. Approach, M. Descriptors, M. D. Selection, V. Reduction, V. S. Selection, C. Modeling, U. M. Algorithm, A. Domain, M. D. Interpretability, *Chemometrics in QSAR*, in S. D. Tauler, R., Walczak, B., & Brown (Ed.), *Compr. Chemom. Chem. Biochem. Data Anal.*, Elsevier B.V., Amsterdam, 2009, pp. 129–172 (ISBN: 9780444641663)
27. T. I. Netzeva, A. P. Worth, T. Aldenberg, R. Benigni, M. T. D. Cronin, P. Gramatica, J. S. Jaworska, S. Kahn, G. Klopman, C. A. Marchant, G. Myatt, N. Nikolova-Jeliazkova, G. Y. Patlewicz, R. Perkins, D. W. Roberts, T. W. Schultz, D. T. Stanton, J. J. M. van de Sandt, W. Tong, G. Veith, C. Yang, *Altern. to Lab. Anim.* **33** (2005) 155 (https://dx.doi.org/10.1177/026119290503300209)
28. S. Yousefinejad, R. Eftekhari, F. Honarasa, Z. Zamanian, F. Sedaghati, *J. Mol. Liq.* **241** (2017) 861 (https://dx.doi.org/10.1016/j.molliq.2017.06.081).

SUPPLEMENTARY MATERIAL TO

Prediction of retardation factor of protein amino acids in reversed phase TLC and ethanol–sodium azide solution as mobile phase using QSRR

SUSAN TORABI[1], FATEMEH HONARASA[2] and SAEED YOUSEFINEJAD[3]

[1]Deputy of Food and Drug Control, Shiraz University of Medical Sciences, Shiraz, Iran; [2]Department of Chemistry, Shiraz Branch, Islamic Azad University, Shiraz, Iran; [3]Research Center for Health Sciences, Institute of Health, Department of Occupational Health Engineering, School of Health, Shiraz University of Medical Sciences, Shiraz, Iran

Table S-1. Numerical vales of original descriptor used in Eq. 1 (before auto scaling)

code	Name	$G_{(N..O)}$	Mor24u	PW2	Mor28u	SEige
AA 1	Glycine-III	6.5	-0.105	0.517	0.092	0.637
AA 2	Alanine-III	6.52	0.06	0.578	-0.115	0.637
AA 3	Aspartic acid-III	15.5	0.095	0.572	0.03	1.134
AA 4	Arginine-III	48.96	-0.125	0.557	0.142	1.058
AA 5	Proline-III	6.43	0.034	0.563	-0.18	0.637
AA 6	Hydroxyproline-III	10.15	-0.172	0.581	-0.181	0.886
AA 7	Lysine-III	20.33	0.09	0.533	-0.183	0.778
AA 8	Glutamic acid-III	17.94	-0.135	0.568	0.124	1.134
AA 9	Serine-III	10.26	-0.135	0.557	-0.017	0.886
AA 10	Tryptophan-III	16.85	-0.024	0.582	0.09	0.778
AA 11	Valine-III	6.47	-0.032	0.588	-0.219	0.637
AA 12	Phenyl lalanine-III	6.53	-0.056	0.564	-0.086	0.637
AA 13	Isoleucine-III	6.47	-0.013	0.571	-0.126	0.637
AA 14	Leucine-III	6.56	-0.142	0.572	-0.228	0.637
AA 15	Asparagine-III	20.62	-0.111	0.572	-0.023	1.026
AA 16	Methionine-III	6.48	-0.253	0.537	-0.233	0.709
AA 17	Cysteine-III	6.57	-0.092	0.557	-0.071	0.709
AA 18	Histidine-III	25.23	0.038	0.57	0.028	0.918
AA 19	Threonine-III	14.31	-0.052	0.579	0.021	0.886
AA 20	Tyrosine-III	10.26	0.018	0.588	-0.002	0.886
AA 21	Glutamine-III	4.22	0.018	0.568	-0.282	0.637