CROATICA CHEMICA ACTA CCACAA **68** (3) 435–450 (1995)

ISSN 0011-1643
CCA–2244

Original Scientific Paper

A Novel QSPR Approach to Physicochemical Properties of the α-Amino Acids

Bono Lučić, Sonja Nikolić and Nenad Trinajstić

The Rugjer Bošković Institute, P.O.B. 1016,
HR-10 001 Zagreb, The Republic of Croatia

Davor Juretić

Department of Natural Sciences and Arts, The University of Split,
HR-58 000 Split, The Republic of Croatia

Albin Jurić

Department of Chemistry, Agricultural Institute,
HR-43 260 Križevci, The Republic of Croatia

Received November 29, 1994; revised February 6, 1995; accepted February 6, 1995

A novel multivariate linear regression, based on the ordered orthogonalized connectivity basis and dominant descriptor analysis, is applied to several physicochemical properties of α-amino acids. Our results compare favourably with those based on the non-orthogonalized connectivity basis by Pogliani.[1,2]

INTRODUCTION

An interesting achievement of QSPR (quantitative structure-property relationships) studies[3] was the effort to encode numerically molecules according to their structural features and to use these numbers in structure-property studies.[4,5] The conversion of the structural formula into a numerical value, often called topological or graph-theoretical index,[6,7] can be achieved in many ways.[8–11] The rules were even set how to search for topological indices of (bio)chemical interest.[3,12,13]

It appears that among many topological indices[14] that have been proposed since the Wiener index in 1947,[15] the connectivity index, introduced by Randić twenty years ago,[16] is the most often used index in QSPR and QSAR (quantitative structure-activity relationships).[17–19] Recently, Randić has also initiated two important advances in this field, that is, he introduced the connectivity basis[20,21] and the orthogonalized descriptors.[22–24]

Randić and other authors[25] observed that the multiple linear regression with orthogonalized descriptors was a more stable model, but not statistically better than the model with non-orthogonalized descriptors, since the values of the correlation coefficient (R), the standard error (S) and the F-test remained the same for both models. However, we recently succeeded in showing that the structure-property-activity models can be improved using the orthogonalized descriptors.[26] We discovered that the orthogonalization orderings are important because certain orderings of orthogonalized descriptors in multiple linear regressions lead to models with higher values of R than the corresponding models with non-orthogonalized descriptors. This novel procedure is used in the present report and is applied to physicochemical properties of α-amino acids. We selected this particular set of molecules and some of their properties because in the literature there are already reports by Pogliani[1,2] on QSPR models using the connectivity basis for predicting physicochemical properties of α-amino acids. This offered an opportunity to test our proposal with already existing good models, which represent the best possible models that can be obtained with a non-orthogonalized basis.

MULTIPLE LINEAR REGRESSION IN AN ORTHOGONAL CONNECTIVITY BASIS

The degree of relationship between three or more variables is called the multiple correlation. A regression equation is an equation for estimating a dependent variable, say P' (physicochemical property) by means of independent variables D_1, D_2,... (topological indices, molecular descriptors) and is called the regression equation of P' on D_1, D_2,...,*etc*. For the case of three variables, the simplest regression equation of P' on D_1 and D_2 is of the form: $P' = a\ D_1 + b\ D_2 + c$, where a, b and c are constants. P' represents an approximation to property P by means of descriptors D_1 and D_2 according to the above equation.

The correlation coefficient and the standard error between the experimental property P and the computed property P' by two non-orthogonal descriptors D_1 and D_2 are given by:[27]

$$R(\mathrm{P,P'}) = \left[\left(R^2_{P,D_1} + R^2_{P,D_2} - 2\,R_{P,D_1} R_{P,D_2} R_{D_1,D_2}\right) / \left(1 - R^2_{D_1,D_2}\right)\right]^{1/2} \tag{1}$$

$$\begin{aligned} S(\mathrm{P,P'}) = \left[m\ /\ (m - \mathrm{I} - 1)\right]^{1/2} S_p \Big[(1 - R^2_{P,D_1} - R^2_{P,D_1} - R^2_{D_1,D_2} + \\ + 2\,R_{P,D_1} R_{P,D_2} R_{D_1 D_2})\ /\ (1 - R^2_{D_1 D_2})\Big]^{1/2} \end{aligned} \tag{2}$$

where m is the number of molecules.

The above expresssions in the orthogonal basis, because of $\mathrm{R}_{\mathrm{D}_1,\mathrm{D}_2} = 0$, reduce to:

$$R(\mathrm{P,P'}) = \left[(R^2_{P,D_1} + R^2_{P,D_2}\right]^{1/2} \tag{3}$$

$$S(\mathrm{P,P'}) = \left[m\ /\ (m - \mathrm{I} - 1)\right]^{1/2} S_p \left[(1 - R^2_{P,D_1} - R^2_{P,D_2}\right]^{1/2} \tag{4}$$

where I stands for all I-tuples ($I = 2,3,\ldots,n$) that can be obtained from a given set of n descriptors. In the above case $n = 2$.

The above is the reason why the use of the orthogonalized descriptors is advantageous in the multiple linear regression. Here we have used the concept of orthogonalized descriptors as defined by Randić.[22–24]

A BRIEF DESCRIPTION OF THE APPROACH WHICH IMPROVES THE QSPR MODELS USING ORTHOGONALIZED DESCRIPTORS

The approach starts with the connectivity basis $^{\ell}\chi$ ($\ell = 0,1,\ldots,\ell{+}1$) which can be generated using the following formula:[28]

$$^{\ell}\chi = \sum_{\text{paths}} \left[\mathrm{d(i)\ d(j)\ ..\ d}(\ell + 1)\right]^{-1/2} \tag{5}$$

where d(i), d(j),...,d(ℓ+1) are valencies of vertices i, j, ..., ℓ+1 in the considered path of length ℓ. This is a non-orthogonal connectivity basis.

The next step is to orthogonalize the connectivity basis to obtain the orthogonalized set of connectivity indices $^{\ell}\Omega$ ($\ell = 0,1,\ldots,\ell{+}1$). The computation of the orthogonal connectivity indices has been outlined in several papers.[22–26] Perhaps the most clear presentation was given recently in an article by Randić and Trinajstić.[29]

The key step is the building up of QSPR models for different orderings of orthogonalized descriptors. In our recent report,[26] we discovered that some orthogonalization orderings lead to QSPR models with smaller values of standard errors than the corresponding models with non-orthogonalized descriptors.

The QSPR models are generated by investigating all possible orthogonalization orderings in the set of n (n = 6) descriptors taken from Pogliani.[1] Among all possible orderings, we selected those which maximize the contributions of one descriptor (the model with I = 1 in the tables), two descriptors (I = 2), three descriptors (I = 3), four descriptors (I = 4) and five descriptors (I = 5). Besides, the result with all six descriptors is always shown. This model is trivial because it can be shown that each model with n descriptors obtained by any orthogonalization ordering of these n descriptors gives identical values of R, S and F.

The algorithm by which we select the best models is detailed in Ref. 26. We first carry out orthogonalization in all possible orderings of the set of n descriptors (there are n! of these). Then, for example, we want to find the best possible model in an orthogonalized basis with three descriptors. Consequently, among the n! orthogonalization orderings we choose that ordering in which these *three* descriptors are with the highest contributions in the total correlation coefficient (since their total contribution is the square root of the sum of squares of the correlation coefficient of each individual descriptor and the property which is approximated). In other words, we search for the orthogonalization ordering in which there are three most dominant descriptors. The remaining (n – 3) descriptors, extracted from one such orthogonalization ordering, contain »the minimum information content«, that is, they are insignificant descriptors. This can also be stated as follows: Let be n = 6 and let, for example, the orthogonalization ordering $D_2D_1D_5D_4D_3D_6$ be the ordering which gives the best model with three descriptors. Let also $D_1D_4D_6$ be descriptors with highest contributions to the total correlation coefficient. Their total correlation coefficient is given by:

$$R' = \left[R^2_{P,D_1} + R^2_{P,D_4} + R^2_{P,D_6} \right]^{1/2} \tag{6}$$

where R_{P,D_1}, R_{P,D_4} and R_{P,D_6} are the correlation coefficients between property P and each individual dominant descriptor. The contribution of the remaining three descriptors is:

$$R'' = \left[R^2_{P,D_2} + R^2_{P,D_3} + R^2_{P,D_5}\right]^{1/2} \tag{7}$$

The value of R'' is considerably smaller than R'. The total R for all 6!=720 orthogonalization orderings is the same and given by:

$$\left[R'^{\,2} + R''^{\,2}\right]^{1/2} \tag{8}$$

With our algorithm we choose that orthogonalization ordering by which we maximize R' (and with this we necessarily minimize R''). We repeat the same procedure with the selection of the best models with 1, 2, 4 and 5 descriptors.

Insignificant descriptors can be removed from the model, since their removal leads to the model with the smaller number of descriptors. Removal of the insignificant descriptors results in improvement of the S values and F-tests, although the correlation coefficients become slightly smaller in the reduced model. Changes in the S and F-test during the passing over from the approximation with I descriptors to the approximation with $(I - 1)$ descriptors are leap-like. Because of that, the transition from the approximation with I descriptors to the approximation with $(I - 1)$, which is carried out by removing one insignificant descriptor, necessarily leads to a significant decrease in S and a significant increase in the F-test. The reason why R is only slightly affected by the removal of insignificant descriptors is that their numerical contributions often influence only the third or fourth decimal place in the correlation coefficient. Therefore, the orthogonalization ordering based on the dominant descriptor analysis produces QSPR models with smaller values of standard errors than the corresponding models with non-orthogonalized descriptors.

This finding represents a step forward in the use of orthogonalized descriptors. This is so because other authors, who did not investigate the effect of orthogonalization ordering on the quality of the QSPR models, found that the models with orthogonalized descriptors are more stable but statistically equal to the models with non-orthogonalized descriptors, since their statistical parameters (R, S and F-test) remained the same for both models. However, it is important to emphasize that the above can be achieved *only* if computation of the dominant descriptor for every ordering of descriptors in the orthogonal basis is carried out.

RESULTS AND DISCUSSION

We have used the same set of non-orthogonal molecular connectivity indices for numerical encoding of α-amino acids as Pogliani.[1] They are given in Table I and are denoted as D, D^v, $^\circ\chi$, $^\circ\chi^v$, $^1\chi$ and $^1\chi^v$.

TABLE I

Molecular connectivity indices of amino acids (AA)

AA	D	D^{v}	$^{0}\chi$	$^{0}\chi^{v}$	$^{1}\chi$	$^{1}\chi^{v}$
Gly	8	20	4.284	2.640	2.270	1.190
Ala	10	22	5.155	3.510	2.643	1.627
Ser	12	28	5.862	3.664	3.181	1.774
Pro	16	28	5.983	4.554	3.805	2.767
Val	16	28	6.732	5.088	3.553	2.538
Thr	14	30	6.732	4.535	3.553	2.219
Cys	12	23.6	5.862	4.554	3.181	2.403
Leu	16	28	7.439	5.795	4.036	3.021
Ile	16	28	7.439	5.795	4.091	3.076
Hyp	18	34	6.853	4.872	4.198	2.842
Asn	16	36	7.439	4.703	4.037	2.304
Asp	16	38	7.439	4.572	4.037	2.239
Gln	18	38	8.146	5.410	4.537	2.804
Lys	18	32	7.983	5.916	4.681	3.366
Glu	18	40	8.146	5.280	4.537	2.739
Met	16	26.7	7.276	6.146	4.181	4.044
His	22	42	8.268	5.819	5.198	3.155
Phe	24	42	8.975	6.604	5.698	3.722
Arg	22	42	9.560	6.709	5.537	3.600
Tyr	26	48	9.845	6.974	6.092	3.857
Trp	32	54	10.836	8.104	7.182	4.716

These indices can be computed as follows. D is the sum-delta index which is given by:[30]

$$D = \sum \delta_i \tag{9}$$

where δ_{i} is the delta index which represents the count of non-hydrogen σ-bond electrons contributed by atom i:[18,31]

$$\delta_i = \sigma_i - \mathrm{H}_i \tag{10}$$

where H_i is the number of hydrogen atoms attached to i.

D^{v} is the sum-delta valence index which is defined by:[30]

$$D^{v} = \sum \delta_i^{v} \tag{11}$$

where δ_i^{v} represents the count of all non-hydrogen valence electrons contributed by atom i:[1,32]

$$\delta_i^{\mathrm{v}} = \delta_i + \mathrm{p}_i + \mathrm{n}_i \tag{12}$$

where σ, p and n are sigma, p and lone-pair electrons, respectively. $^{\circ}\chi$ is the zero-order connectivity index:[10,18]

$$^{\circ}\chi = \sum (\delta_i)^{-1/2} \tag{13}$$

Similarly, $^{\circ}\chi^{\mathrm{v}}$ is the zero-order valence connectivity index:[10,18]

$$^{\circ}\chi^{\mathrm{v}} = \sum (\delta_i^{\mathrm{v}})^{-1/2} \tag{14}$$

Finally, the first-order connectivity index $^{1}\chi$ and the first-order valence connectivity index $^{1}\chi^{\mathrm{v}}$ are defined as:[10,18]

TABLE II

Molecular weights of amino acids (M_{r}), side-chain molecular volumes (V) and densities of crystalline amino acids (CD)

	M_{r}	V	CD
gly	75	36.3	1.601
ala	89	52.6	1.401
ser	105	54.9	1.537
pro	115	73.6	
val	117	85.1	1.230
thr	119	71.2	
cys	121		1.165
leu	131	102.0	
ile	131	102.0	
hyp	132		
asn	132	72.4	
asp	133	68.4	1.660
gln	146	92.7	
lys	146	105.1	
glu	147	84.7	1.538
met	149		1.340
his	155	91.1	
phe	165	113.9	
arg	174	109.1	1.100
tyr	181	116.2	1.456
trp	204	135.4	

$$ {}^1\chi = \sum (\delta_i \delta_j)^{-1/2} \tag{15} $$

$$ {}^1\chi^{\mathrm{v}} = \sum (\delta_i^{\mathrm{v}} \delta_j^{\mathrm{v}})^{-1/2} \tag{16} $$

The properties considered are molecular weights (M_{r}), side-chain molecular volumes V and crystal densities CD of α-amino acids. They are taken from Pogliani[1] and reported in Table II.

We compared our results with the best models obtained using multiple linear regression with the non-orthogonal connectivity basis and Pogliani's

TABLE III

The best combinations of non-orthogonal connectivity indices used in a regression *vs.* CD of amino acids and the corresponding values of *R*, *S* and *Q*

***I* = 1**

$\mathrm{CD} = (1.7836 \pm 0.201) +(-0.0756 \pm 0.038)\ {}^{\circ}\chi^{\mathrm{v}}$

$n = 10, R = 0.5703, S = 0.1660, Q = 3.4355$

***I* = 2**

$\mathrm{CD} = (1.7901 \pm 0.150) + (0.04212 \pm 0.010)\ D^{\mathrm{v}} +(-0.2383 \pm 0.051)\ {}^{\circ}\chi$

$n = 10, R = 0.8698, S = 0.1066, Q = 8.1595$

***I* = 3**

$\mathrm{CD} = (2.0310 \pm 0.120) + (0.0697 \pm 0.011)\ D^{\mathrm{v}} +(-0.4909 \pm 0.082)\ {}^{\circ}\chi +$
$+ (0.2476 \pm 0.074)\ {}^1\chi^{\mathrm{v}}$

$n = 10, R = 0.9565, S = 0.0681, Q = 14.0455$

***I* = 4**

$\mathrm{CD} = (1.7439 \pm 0.088) +(-0.1099 \pm 0.021)\ D + (0.1402 \pm 0.016)\ D^{\mathrm{v}} +$
$+ (-0.9270 \pm 0.116)\ {}^{\circ}\chi + (0.7090 \pm 0.109)\ {}^{\circ}\chi^{\mathrm{v}}$

$n = 10, R = 0.9872, S = 0.0407, Q = 24.2555$

***I* = 5**

$\mathrm{CD} = (1.7876 \pm 0.086) +(-0.1171 \pm 0.019)\ D + (0.1338 \pm 0.015)\ D^{\mathrm{v}} +$
$+ (-0.9407 \pm 0.105)\ {}^{\circ}\chi + (0.6664 \pm 0.1103)\ {}^{\circ}\chi^{\mathrm{v}} + (0.1468 \pm 0.101)\ {}^1\chi$

$n = 10, R = 0.9916, S = 0.369, Q = 26.8726$

***I* = 6**

$\mathrm{CD} = (1.7419 \pm 0.093) +(-0.1624 \pm 0.045)\ D + (0.1594 \pm 0.027)\ D^{\mathrm{v}} +$
$+ (-1.2107 \pm 0.264)\ {}^{\circ}\chi + (1.0189 \pm 0.333)\ {}^{\circ}\chi^{\mathrm{v}} + (0.3016 \pm 0.171)\ {}^1\chi +$
$+ (-0.19122 \pm 0.173)\ {}^1\chi^{\mathrm{v}}$

$n = 10, R = 0.9941, S = 0.0359, Q = 27.6908$

TABLE IV

The best combinations of orthogonalized connectivity indices used in a regression *vs.* CD of amino acids and the corresponding values of R, S and Q. Bold letters denote dominant descriptors

$\boldsymbol{I} = \mathbf{1}$; orthogonalization ordering: D^v, ${}^{\circ}\chi$, $\boldsymbol{D}$, ${}^{\circ}\chi^v$, ${}^1\chi$, ${}^1\chi^v$
CD = $(1.4028 \pm 0.032) + (0.2383 \pm 0.048)\ {}^2\Omega$
$n = 10$, $R = 0.8693$, $S = 0.0999$, $Q = 8.7017$

$\boldsymbol{I} = \mathbf{2}$; orthogonalization ordering: D^v, ${}^{\circ}\chi$, $\boldsymbol{D}$, ${}^{\circ}\boldsymbol{\chi}^{\mathbf{v}}$, ${}^1\chi$, ${}^1\chi^v$
CD = $(1.4028 \pm 0.011) + (0.2383 \pm 0.017)\ {}^2\Omega + (-0.709{}^{\circ}\chi 0.095)\ {}^3\Omega$
$n = 10$, $R = 0.9862$, $S = 0.0358$, $Q = 27.5475$

$\boldsymbol{I} = \mathbf{3}$; orthogonalization ordering: D^v, ${}^{\circ}\chi$, $\boldsymbol{D}$, ${}^{\circ}\boldsymbol{\chi}^{\mathbf{v}}$, ${}^1\chi^v$, ${}^1\chi$
CD = $(1.4028 \pm 0.009) + (0.2382 \pm 0.014)\ {}^2\Omega + (-0.709 \pm 0.077)\ {}^3\Omega +$
$+(-\ 0.3016 \pm 0.137)\ {}^4\Omega$
$n = 10$, $R = 0.9924$, $S = 0.0288$, $Q = 34.4583$

$\boldsymbol{I} = \mathbf{4}$; orthogonalization ordering: D^v, ${}^{\circ}\boldsymbol{\chi}^{\mathbf{v}}$, ${}^{\circ}\boldsymbol{\chi}$, ${}^1\chi$, $\boldsymbol{D}$, ${}^1\chi^v$
CD = $(1.4028 \pm 0.009) + (-0.1171 \pm 0.015)\ {}^{\circ}\chi + (-0.611 \pm 0.070)\ {}^2\Omega +$
$+\ (0.154 \pm 0.010)\ {}^3\Omega + (0.1912 \pm 0.139)\ {}^5\Omega$
$n = 10$, $R = 0.9936$, $S = 0.0289$, $Q = 34.3599$

$\boldsymbol{I} = \mathbf{5}$; orthogonalization ordering: $\boldsymbol{D}^{\mathbf{v}}$, ${}^{\circ}\boldsymbol{\chi}^{\mathbf{v}}$, ${}^{\circ}\chi$, ${}^1\chi$, $\boldsymbol{D}$, ${}^1\chi^v$
CD = $(1.4028 \pm 0.010) +(-0.1171 \pm 0.016)\ {}^{\circ}\Omega\ (-0.0006 \pm 0.001)\ {}^1\Omega +$
$+\ (-\ 0.6111 \pm 0.076)\ {}^2\Omega + (0.1540 \pm 0.011)\ {}^3\Omega + (0.1912 \pm 0.150)\ {}^5\Omega$
$n = 10$, $R = 0.9941$, $S = 0.0311$, $Q = 31.9646$

$\boldsymbol{I} = \mathbf{6}$; orthogonalization ordering: ${}^1\boldsymbol{\chi}$, $\boldsymbol{D}^{\mathbf{v}}$, $\boldsymbol{D}$, ${}^{\circ}\chi^v$, ${}^1\boldsymbol{\chi}^{\mathbf{v}}$, ${}^{\circ}\chi$
CD = $(1.413 \pm 0.038) +(-0.033 \pm 0.047)\ {}^{\circ}\chi + (-0.0372 \pm 0.005)\ {}^1\Omega +$
$+\ (1.4811 \pm 0.982)\ {}^2\Omega + (0.2034 \pm 0.265)\ {}^3\Omega + (-0.035 \pm 0.059)\ {}^4\Omega +$
$+\ (0.5738 \pm 0.148)\ {}^5\Omega$
$n = 10$, $R = 0.9942$, $S = 0.0430$, $Q = 23.1209$

results for each of the three properties taken into account.[1,2] In Tables III and IV models are given for predicting crystal densities of α-amino acids. Table III contains QSPR models based on the best combinations of non-orthogonal connectivity indices, while Table IV our results.

The best model in Table III is the last model with all non-orthogonal connectivity indices considered (S = 0.0359). The best model with orthogonalized connectivity basis is the third model in Table IV (S = 0.0288). The orthogonalization ordering for this model is D^v, ${}^{\circ}\chi$, D, ${}^{\circ}\chi^v$, ${}^1\chi^v$ and ${}^1\chi$. Only dominant descriptors are given in the model.

The best model of Pogliani is the following:[1]

$$\mathrm{CD} = -0.16\,D + 0.16\,D^{\mathrm{v}} - 1.21\,{}^{\circ}\chi + 1.02\,{}^{\circ}\chi^{\mathrm{v}} + 0.30\,{}^{1}\chi - 0.19\,{}^{1}\chi^{\mathrm{v}} + 1.74 \tag{17}$$

$$n = 10,\ R = 0.994,\ S = 0.035,\ F = 42,\ Q = 28.4$$

where Q is equal to R/S and is called the quality factor.[33]

It appears that, among the three models listed, our model given in Table IV is the simplest and possesses the lowest value of the standard error.

TABLE V

The best combinations of non-orthogonal connectivity indices used in a regression *vs.* M_{r} of amino acids and the corresponding values of R, S and Q

$\boldsymbol{I = 1}$

$M_r = (-2.2735 \pm 6.752) + (18.65391 \pm 0.889)\ {}^{\circ}\chi$

$n = 21,\ R = 0.9791,\ S = 6.2715,\ Q = 0.1561$

$\boldsymbol{I = 2}$

$M_r = (3.8913 \pm 4.9748) + (13.8828 \pm 1.2401)\ {}^{\circ}\chi + (10.2668 \pm 2.300)\ {}^{1}\chi^{\mathrm{v}}$

$n = 21,\ R = 0.9901,\ S = 4.4389,\ Q = 0.2231$

$\boldsymbol{I = 3}$

$M_r = (9.8235 \pm 4.8459) + (0.8102 \pm 0.305)\ D^{\mathrm{v}} + (8.1103 \pm 2.421)\ {}^{\circ}\chi +$
$+ (13.6586 \pm 2.363)\ {}^{1}\chi^{\mathrm{v}}$

$n = 21,\ R = 0.9930,\ S = 3.8385,\ Q = 0.2587$

$\boldsymbol{I = 4}$

$M_r = (0.8791 \pm 4.470) + (-3.2006 \pm 0.685)\ D + (2.6794 \pm 0.286)\ D^{\mathrm{v}} +$
$+ (10.0662 \pm 2.672)\ {}^{\circ}\chi^{\mathrm{v}} + (16.7048 \pm 3.765)\ {}^{1}\chi^{\mathrm{v}}$

$n = 21,\ R = 0.9962,\ S = 2.9453,\ Q = 0.3382$

$\boldsymbol{I = 5}$

$M_r = (-2.4176 \pm 4.592) + (-7.2297 \pm 1.339)\ D + (4.213 \pm 0.813)\ D^{\mathrm{v}} +$
$+ (-19.5079 \pm 5.258)\ {}^{\circ}\chi + (34.9044 \pm 5.830)\ {}^{\circ}\chi^{\mathrm{v}} + (19.3406 \pm 6.297)\ {}^{1}\chi$

$n = 21,\ R = 0.9963,\ S = 2.9777,\ Q = 0.3346$

$\boldsymbol{I = 6}$

$M_r = (-0.6986 \pm 4.540) + (-5.7369 \pm 1.611)\ D + (3.6116 \pm 0.872)\ D^{\mathrm{v}} +$
$+ (-11.2852 \pm 7.366)\ {}^{\circ}\chi + (23.15836 \pm 9.494)\ {}^{\circ}\chi^{\mathrm{v}} + (11.9399 \pm 7.733)\ {}^{1}\chi +$
$+ (8.8853 \pm 5.807)\ {}^{1}\chi^{\mathrm{v}}$

$n = 21,\ R = 0.9968,\ S = 2.8524,\ Q = 0.3496$

TABLE VI

The best combinations of orthogonalized connectivity indices used in a regression *vs.* M_r of amino acids and the corresponding values of R, S and Q. Bold letters denote dominant descriptors.

$\boldsymbol{I = 1}$; orthogonalization ordering: $\boldsymbol{{}^{\circ}\chi}$, $\boldsymbol{D}$, D^v, ${}^{\circ}\chi^v$, ${}^1\chi$, ${}^1\chi^v$
$M_r = (136.5239 \pm 1.369) + (18.6539 \pm 0.889)\ {}^2\Omega$
$n = 21$, $R = 0.9791$, $S = 6.2715$, $Q = 0.1561$

$\boldsymbol{I = 2}$; orthogonalization ordering: $\boldsymbol{{}^{\circ}\chi}$, D^v, $\boldsymbol{{}^1\chi^v}$, $\boldsymbol{D}$, ${}^{\circ}\chi^v$, ${}^1\chi$
$M_r = (1365239 \pm 0.8212) + (18.6539 \pm 0.533)\ {}^2\Omega + (13.6586 \pm 2.316)\ {}^5\Omega$
$n = 21$, $R = 0.9929$, $S = 3.7631$, $Q = 0.2639$

$\boldsymbol{I = 3}$; orthogonalization ordering: $\boldsymbol{{}^{\circ}\chi}$, D^v, $\boldsymbol{{}^{\circ}\chi^v}$, ${}^1\chi$, $\boldsymbol{D}$, ${}^1\chi^v$
$M_r = (136.5239 \pm 0.620) + (-7.2297 \pm 1.278)\ {}^{\circ}\chi + (18.6539 \pm 0.403)\ {}^2\Omega +$
$+ (13.9691 \pm 2.119)\ {}^3\Omega$
$n = 21$, $R = 0.9962$, $S = 2.8431$, $Q = 0.3504$

$\boldsymbol{I = 4}$; orthogonalization ordering: $\boldsymbol{{}^{\circ}\chi^v}$, $\boldsymbol{{}^1\chi}$, ${}^1\chi^v$, ${}^{\circ}\chi$, D, $\boldsymbol{D^v}$
$M_r = (136.5237 \pm 0.589) + (-3.6116 \pm 0.826)\ {}^1\chi + (-16.8785 \pm 2.616)\ {}^2\Omega +$
$+ (22.7478 \pm 0.477)\ {}^3\Omega + (-17.021 \pm 1.479)\ {}^4\Omega$
$n = 21$, $R = 0.9968$, $S = 2.7003$, $Q = 0.3691$

$\boldsymbol{I = 5}$; orthogonalization ordering: $\boldsymbol{D^v}$, $\boldsymbol{{}^{\circ}\chi^v}$, ${}^{\circ}\chi$, ${}^1\chi$, $\boldsymbol{D}$, ${}^1\chi^v$
$M_r = (136.5239 \pm 0.601) + (-7.2296 \pm 1.239)\ {}^{\circ}\Omega + (0.1406 \pm 0.184)\ {}^1\Omega +$
$+ (18.6539 \pm 0.391)\ {}^2\Omega + (13.9691 \pm 2.054)\ {}^3\Omega + (-8.8854 \pm 5.611)\ {}^5\Omega$
$n = 21$, $R = 0.9968$, $S = 2.7561$, $Q = 0.3616$

$\boldsymbol{I = 6}$; orthogonalization ordering: $\boldsymbol{{}^1\chi}$, $\boldsymbol{D^v}$, $\boldsymbol{D}$, ${}^{\circ}\chi^v$, $\boldsymbol{{}^1\chi^v}$, ${}^{\circ}\chi$
$M_r = (136.5239 \pm 0.622) + (55.7369 \pm 1.611)\ {}^{\circ}\chi + (0.1406 \pm 0.191)\ {}^1\Omega +$
$+ (18.6539 \pm 0.404)\ {}^2\Omega + (7.103 \pm 5.263)\ {}^3\Omega + (-8.0475 \pm 5.318)\ {}^4\Omega +$
$+ (13.6585 \pm 1.756)\ {}^5\Omega$
$n = 21$, $R = 0.9968$, $S = 2.8524$, $Q = 0.3495$

Models for computing the molecular weights of α-amino acids are given in Tables V and VI.

The best model based on the non-orthogonalized connectivity basis is again the last one and the most complex model ($S = 2.8524$) in Table V. Somewhat better is the model based on the orthogonalized connectivity basis (S = 2.7003), viz. the fourth model in Table VI. The orthogonalization ordering for this model is ${}^{\circ}\chi$, ${}^1\chi$, ${}^1\chi^v$, ${}^0\chi$, D and D^v Only dominant descriptors are given in the model.

TABLE VII

The best combinations of non-orthogonal connectivity indices used in a regression *vs.* V of amino acids and the corresponding values of R, S and Q

***I* = 1**

$V = (-12.7793 \pm 3.907) + (18.7801 \pm 0.714)\ {}^{\circ}\chi^{v}$

$n = 18,\ R = 0.9886,\ S = 3.9450,\ Q = 0.2506$

***I* = 2**

$V = (-9.8463 \pm 2.476) + (-0.6119 \pm 0.118)\ D^{v} + (22.2194 \pm 0.795)\ {}^{\circ}\chi^{v}$

$n = 18,\ R = 0.9960,\ S = 2.4342,\ Q = 0.4092$

***I* = 3**

$V = (-13.2558 \pm 3.409) + (-1.7775 \pm 0.330)\ D + (18.9233 \pm 2.715)\ {}^{\circ}\chi^{v} +$
$+ (11.1145 \pm 4.361)\ {}^{1}\chi^{v}$

$n = 18,\ R = 0.9964,\ S = 2.3833,\ Q = 0.4181$

***I* = 4**

$V = (-10.3973.542) + (-1.0406 \pm 0.308)\ D^{v} + (5.402 \pm 3.565)\ {}^{\circ}\chi +$
$+ (13.5556 \pm 5.806)\ {}^{\circ}\chi^{v} + (7.2991 \pm 5.656)\ {}^{1}\chi^{v}$

$n = 18,\ R = 0.9966,\ S = 2.4002,\ Q = 0.4152$

***I* = 5**

$V = (-10.7335 \pm 3.675) + (-0.7619 \pm 0.562)\ D^{v} + (5.5624 \pm 3.666)\ {}^{\circ}\chi +$
$+ (12.8804 \pm 6.060)\ {}^{\circ}\chi^{v} + (-4.2809 \pm 7.146)\ {}^{1}\chi + (11.4738 \pm 9.067)\ {}^{1}\chi^{v}$

$n = 18,\ R = 0.9967,\ S = 2.4617,\ Q = 0.4049$

***I* = 6**

$V = (-10.7266 \pm 4.4991) + (0.0062 \pm 2.112)\ D + (-0.765 \pm 1.194)\ D^{v} +$
$+ (5.5879 \pm 9.442)\ {}^{\circ}\chi + (12.8496 \pm 12.177)\ {}^{\circ}\chi^{v} + (-4.2949 \pm 8.849)\ {}^{1}\chi +$
$+ (11.481 \pm 9.781)\ {}^{1}\chi^{v}$

$n = 18,\ R = 0.9967,\ S = 2.5711,\ Q = 0.3877$

Pogliani gives two QSPR models for predicting molecular weights. The first model is given by:[2]

$$M_{\mathrm{r}} = -0.6986 - 5.47\,D + 3.61\,D^{\upsilon} - 11.29\ {}^{\circ}\chi + 23.16\ {}^{0}\chi^{v} + 11.94\ {}^{1}\chi + 8.89\ {}^{1}\chi^{v} \tag{18}$$

$$n = 21,\ R = 0.9968,\ S = 2.85,\ Q = 0.350$$

Pogliani's second model uses orthogonalized descriptors:[2]

TABLE VIII

The best combination of orthogonalized connectivity indices used in a regression *vs.* V of amino acids and the corresponding values of R, S and Q. Bold letters denote dominant descriptors.

$\boldsymbol{I} = \mathbf{1}$; orthogonalization ordering: $^\circ\chi^v$, $\boldsymbol{D}$, D^v, $^\circ\chi$, $^1\chi$, $^1\chi^v$
$V = (87.0389 \pm 0.930) + (18.7801 \pm 0.714)\ {}^3\Omega$
$n = 18$, $R = 0.9886$, $S = 3.9450$, $Q = 0.2506$

$\boldsymbol{I} = \mathbf{2}$; orthogonalization ordering: $^\circ\chi^v$, $^1\chi^v$, $\boldsymbol{D}$, D^v, $^\circ\chi$, $^1\chi$
$V = (87.0389 \pm 0.557) + (-1.7775 \pm 0.327)\ {}^\circ\chi + (18.7801 \pm 0.428)\ {}^3\Omega$
$n = 18$, $R = 0.9962$, $S = 2.3629$, $Q = 0.4216$

$\boldsymbol{I} = \mathbf{3}$; orthogonalization ordering: $^\circ\chi^v$, $\boldsymbol{D^v}$, $^1\chi$, $^1\chi^v$, $^\circ\chi$, D
$V = (87.0389 \pm 0.545) + (0.6119 \pm 0.112)\ {}^1\Omega + (5.5624 \pm 3.446)\ {}^2\Omega +$
$+ (18.7801 \pm 0.419)\ {}^3\Omega$
$n = 18$, $R = 0.9966$, $S = 2.3135$, $Q = 0.4308$

$\boldsymbol{I} = \mathbf{4}$; orthogonalization ordering: $\boldsymbol{D^v}$, $^\circ\chi^v$, $^1\chi$, $^\circ\chi$, $^1\chi^v$, D
$V = (87.0389 \pm 0.557) + (2.1268 \pm 0.063)\ {}^1\Omega + (-3.3061 \pm 3.078)\ {}^2\Omega +$
$+ (-22.2194 \pm 0.772)\ {}^3\Omega + (11.4736 \pm 8.712)\ {}^5\Omega$
$n = 18$, $R = 0.9967$, $S = 2.3652$, $Q = 0.4214$

$\boldsymbol{I} = \mathbf{5}$; orthogonalization ordering: $\boldsymbol{D^v}$, $^\circ\chi^v$, $^1\chi$, $l\chi^v$, $^\circ\chi$, D
$V = (87.0389 \pm 0.580) + (2.1268 \pm 0.066)\ {}^1\Omega + (5.5624 \pm 3.666)\ {}^2\Omega +$
$+ (-22.2194 \pm 0.804)\ {}^3\Omega + (0.1306 \pm 3.854)\ {}^4\Omega + (-4.7833 \pm 7.923)\ {}^5\Omega$
$n = 18$, $R = 0.9967$, $S = 2.4617$, $Q = 0.4049$

$\boldsymbol{I} = \mathbf{6}$; orthogonalization ordering: $\boldsymbol{D^v}$, $^\circ\chi^v$, $^1\chi$, $^\circ\chi$, $\boldsymbol{D}$, $^\circ\chi^v$
$V = (87.0389 \pm 0.606) + (-0.6139 \pm 2.044)\ {}^\circ\chi + (2.1268 \pm 0.069)\ {}^1\Omega +$
$+ (-3.3062 \pm 3.346)\ {}^2\Omega + (-22.2194 \pm 0.840)\ {}^3\Omega + (0.1307 \pm 4.025)\ {}^4\Omega +$
$+ (-11.4813 \pm 9.782)\ {}^5\Omega$
$n = 18$, $R = 0.9967$, $S = 2.5711$, $Q = 3877$

$$M_r = 5.268\ {}^1\Omega + 14.22\ {}^3\Omega + 36.25\ {}^4\Omega + 21.17\ {}^5\Omega + 9.038\ {}^6\Omega + 45.36$$

$$n = 21,\ R = 0.9968,\ S = 2.7603,\ Q = 0.3611 \qquad (19)$$

As expected, models based on the orthogonalized connectivity basis are superior to other models. Our model is only slighly better and less complex than Pogliani's model, thus confirming that the orthogonalization ordering is important in the QSPR modelling with orthogonalized descriptors. This is also supported by the quality factor, which is also slightly higher for our model (Q = 0.3691).

Finally, the modelling of the side-chain molecular volumes V is reported in Tables VII and VIII.

The best model with the non-orthogonalized connectivity basis is the third model (S = 2.3833) in Table VII. A better model is based on the orthogonalized connectivity basis, which is the third model (S = 2.3135) in Table VIII with the following orthogonalization ordering: ${}^{\circ}\chi$, D^{v}, ${}^{1}\chi$, ${}^{1}\chi^{\circ}$, ${}^{\circ}\chi$, and D. Only dominant descriptors are given in the model.

Pogliani has also produced two models for approximating the side-chain molecular volumes. The first model is given by:[1]

$$V = 1.28 + 0.075\, D^{v} - 0.27\, {}^{\circ}\chi + 19.8\, {}^{\circ}\chi^{v} - 3.61\, {}^{1}\chi + 11.3\, {}^{1}\chi^{v} - 12.2 \tag{20}$$

$$n = 18,\ R = 0.997,\ S = 2.62,\ F = 266$$

Pogliani's second model is less complex:[2]

$$V = -\ 1.777\ D + 18.92\ {}^{\circ}\chi^{v} + 11.11\ {}^{1}\chi^{v} - 13.26 \tag{21}$$

$$n = 18,\ R = 0.9964,\ S = 2.3833,\ Q = 0.4181$$

If we accept that the standard eror of estimate S is the critical quantity for determining the quality of a set of basis descriptors, then our model (the third model in Table VIII) appears to be the best among the presented QSPR models. This is also supported by the quality factor Q, which is also the highest for our model (0.4308).

In all the three cases studied our models, that is, models based on the orthogonalized connectivity basis which are obtained by taking care of the orthogonalization ordering and the dominant descriptor analysis, proved to be better models than those based either on the non-orthogonalized connectivity basis or Pogliani's models representing the best combinations of several kinds of topological indices.

CONCLUDING REMARKS

A recently introduced approach,[26] which improves the structure-property models using the orthogonalized descriptors by carefully accounting of the orthogonalization orderings and dominant descriptor analysis, has proved to be a reliable method for building up the QSPR models for predicting the physicochemical properties of α-amino acids. It compares favourably with the best results obtained using multiple linear regression in a non-orthogonal connectivity basis and models proposed by Pogliani.[1,2]

Acknowledgment. – This work was supported by the Ministry of Science and Technology of the Republic of Croatia through Grants 1-07-159 and 1-03-171.

REFERENCES

1. L. Pogliani, *J. Phys. Chem.* 97 (1993) 6731.
2. L. Pogliani, *J. Phys. Chem.* 98 (1994) 1494.
3. M. Randić, Z. Mihalić, S. Nikolić and N. Trinajstić, *Croat. Chem. Acta* **66** (1993) 411.
4. D. H. Rouvray, *Sci. Am.* **254** (1986) 40.
5. M. I. Stankevich, I. V. Stankevich and N. S. Zefirov, *Russ. Chem. Rev.* **57** (1988) 191
6. H. Hosoya, *Bull. Chem. Soc. Japan* **44** (1971) 2332.
7. P. J. Hansen and P.C. Jurs, *J. Chem. Educ.* **65** (1988) 574.
8. A. Sabljić and N. Trinajstić, *Acta Pharm.* **31** (1981) 189.
9. A. T. Balaban, I. Motoc, D. Bonchev and O. Mekenyan, *Topics Curr. Chem.* **114** (1983) 21.
10. N. Trinajstić, *Chemical Graph Theory*, 2nd revised edition, CRC Press, Boca Raton, FL, 1992, chapter 10.
11. O. Mekenyan and S.C. Basak, in: *GraphTheoretical Approaches to Chemical Reactivity*, D. Bonchev and O. Mekenyan (Eds.), Kluwer Academic Press, Dordrecht, 1994, p. 221.
12. M. Randić, *J. Math. Chem.* **7** (1991) 155.
13. M. Randić and N. Trinajstić, *J. Mol. Struct. (Theochem)* **300** (1993) 551.
14. D. H. Rouvray, *J. Mol. Struct. (Theochem)* **185** (1989) 187.
15. H. Wiener, *J. Amer. Chem. Soc.* **69** (1947) 17.
16. M. Randić, *J. Amer. Chem. Soc.* **97** (1975) 6609.
17. L. B. Kier and L. H. Hall, *Molecular Connectivity in Chemistry and Drug Research*, Academic Press, New York, 1976.
18. L. B. Kier and L. H. Hall, *Molecular Connectivity in Structure – Activity Analysis,* Wiley, New York, 1986.
19. L. H. Hall and L. B. Kier, in: *Reviews in Computational Chemistry II*, K. B. Lipkowitz and D. B. Boyd, Eds., VCH, New York, 1991, p. 367.
20. M. Randić, *J. Chem. Inf. Comput. Sci.* **32** (1992) 57.
21. M. Randić and N. Trinajstić, *J. Mol. Struct. (Theochem)* **284** (1993) 209.
22. M. Randić, *J. Chem. Inf. Comput. Sci.* **31** (1991) 311.
23. M. Randić, *New J. Chem.* **15** (1991) 517.
24. M. Randić, *J. Chem. Educ.* **69** (1992) 713.
25. D. Amić, D. Davidović-Amić and N. Trinajstić, *J. Chem. Inf. Comput. Sci.* **35** (1995) 136.
26. (a) B. Lučić, S. Nikolić, N. Trinajstić and D. Juretić, *J. Chem. Inf. Comput. Sci.* **35** (1995) 532.
    (b) B. Lučić, work in preparation.
27. M. R. Spiegel, *Statistics*, Schaum Publ. Co., New York, 1961.

28. L. B. Kier, L. H. Hall, W. J. Murray and M. Randić, *J. Pharm. Sci.* **65** (1976) 1226.
29. M. Randić and N. Trinajstić, *New J. Chem.* **18** (1994) 179.
30. L. Pogliani, *J. Pharm. Sci.* **81** (1992) 334.
31. L. B. Kier and L. H. Hall, *J. Pharm. Sci.* **70** (1981) 583.
32. L. B. Kier and L. H. Hall, *J. Pharm. Sci.* **65** (1976) 1806.
33. L. Pogliani, *Amino Acids* **6** (1994) 141.

SAŽETAK

Novi QSPR pristup predviđanju fizikalno-kemijskih svojstva α-amino kiselina

B. Lučić, S. Nikolić, N. Trinajstić, D. Juretić i A. Jurić

Primijenjena je nova višestruka linearna regresijska analiza, koja se temelji na uređenim ortogonalnim indeksima povezanosti, na predviđanje fizikalno-kemijskih svojstava α-aminokiselina. Postignuti rezultati bolji su od onih koji se dobiju kada se upotrijebe neortogonalizirani indeksi povezanosti i od objavljenih rezultata Poglianija, koji su dobiveni kombiniranjem različitih molekularnih deskriptora.