The Structure−Property Models Can Be Improved Using the Orthogonalized Descriptors

Bono Lučić, Sonja Nikolić, and Nenad Trinajstić*

The Rugjer Bošković Institute, P.O.B. 1016, HR-41001 Zagreb, The Republic of Croatia

Davor Juretić

Department of Natural Sciences and Arts, The University of Split, Nikole Tesle 12, HR-58000 Split, The Republic of Croatia

Received November 23, 1994[⊗]

In this report we describe an approach of how one can with the use of orthogonalized descriptors achieve a better structure−property−activity model. This is illustrated using the truncated connectivity basis ${}^{l}\chi$ (l = 0, 1, ..., 6). The molecular property used to test the approach was the boiling points of octanes. We first developed the algorithm which produces absolutely the best models with I descriptors (I = 1−7) in nonorthogonalized basis. These models were always better than the models that most authors achieve by the use of the stepwise/inclusion−exclusion procedure. The next step was the development of the computer program by which we could realize all possible orthogonalization orderings of a given set of I descriptors. In doing that we discovered that the certain orderings of the orthogonalized descriptors lead to models with higher values of the correlation coefficient (R) than the corresponding models with nonorthogonalized descriptors. Because of that we selected among all the possible orthogonalization orderings (there are $I!$ possibilities for I descriptors) that ordering which leads to the descriptor which gives the highest value of R. We call this descriptor the dominant descriptor. After we located the first dominant descriptor, we have chosen the second dominant descriptor among the remaining (I − 1) descriptors following the same procedure. In the identical way are obtained the third, the fourth, etc. dominant descriptor. In this manner the selection of the dominant descriptors necessarily minimize the contributions of those descriptors which contribute small amounts to the total correlation coefficient, because the total R is for any fixed set of I descriptors constant and independent of the orthogonalization order. These descriptors appear to be insignificant and are removed from the consideration. With this act we only negligibly diminished the total R, but the value of S as well as F-test were significantly improved, since we obtained the model with less descriptors.

INTRODUCTION

Several authors, including Lukovits (who first used orthogonalized quantum chemical indices in QSAR studies),[1] Randić (who first introduced orthogonalized molecular descriptors),[2] and ourselves,[3] who used multiple linear regression with orthogonalized descriptors, pointed out that these regressions models are more stable but not better than the models which use nonorthogonalized descriptors, that is, both models possess the same values for the correlation coefficient (R), the standard error (S), and the F-test. The present report deals with the investigation on whether the structure−property−activity models could be improved in some way using the orthogonalized descriptors.

If we wish to represent a certain physical, chemical, or biological property of a group of molecules as the linear combination of molecular descriptors, there is a practical problem of how to select the set I (I = 1, ..., N) from a collection of the N descriptors,[4] which approximate best (with the smallest standard error)[5,6] a given property. This problem has, of course, been considered by a number of people;[7] the most systematic efforts being by Randić.[8]

In this work we have utilized the concept of orthogonalized descriptors, introduced by Randić,[2,9−11] because the expressions for the standard error and the correlation coefficient are much simpler when the orthogonalized basis is used instead of nonorthogonalized basis.

The correlation coefficient (R) and the standard error (S) between the experimental property P and computed property P' by means of two nonorthogonalized descriptors D_1 and D_2 (I = 2) are given by[12]

$$R(P,P') = [(R_{P,D_1}{}^2 + R_{P,D_2}{}^2 - 2R_{P,D_1}R_{P,D_2}R_{D_1D_2})/(1 - R_{D_1D_2}{}^2)] \quad (1)$$

$$S(P,P') = [m/(m - I - 1)]^{1/2}S_P[(1 - R_{P,D_1}{}^2 - R_{P,D_2}{}^2 - R_{D_1D_2}{}^2 + 2R_{P,D_1}R_{P,D_2}R_{D_1D_2})/(1 - R_{D_1D_2}{}^2)]^{1/2} \quad (2)$$

where m is the number of molecules.

The above expressions in the orthogonal basis, because of R_{D_1,D_2} = 0, reduce to

$$R(P,P') = [(R_{P,D_1}{}^2 + R_{P,D_2}{}^2)]^{1/2} \quad (3)$$

$$S(P,P') = [m/(m - I - 1)]^{1/2}S_P[(1 - R_{P,D_1}{}^2 - R_{P,D_2}{}^2)]^{1/2} \quad (4)$$

Two descriptors can be selected out of the set of, for example, 10 descriptors in $\binom{10}{2}$ = 45 ways. The optimum pair of descriptors is that one which gives the smallest

[⊗] Abstract published in *Advance ACS Abstracts*, April 1, 1995.

0095-2338/95/1635-0532$09.00/0

standard error. Consequently three descriptors can be selected in 120 ways, four descriptors in 210 ways, etc. Most of the available commercial programs in which the program has to start from the beginning for each selection of the I-tuple of descriptors, the optimization is unwieldly and is a slow process for any of larger initial sets of descriptors. Because of that we developed an algorithm which computes for any given set of orthogonalized descriptors the correlation coefficient for each selection of I-tuple of descriptors and chooses that I-tuple which produces the highest value of the correlation coefficient between a given property P and its approximation P' obtained by means of this I-tuple of descriptors. The increase of the number of descriptors (from I to $I + 1$) necessarily increases the value of R, but it needs not decrease the value of S, because the increased number of descriptors enters into the S-computation.

THE STRUCTURE OF THE ALGORITHM

The structure of the algorithm, on which our computer program is based, consists of the following steps:

(1) Initial data: The set of N descriptors and a property to be considered.

(2) Two descriptors are unbiasly selected from the initial set of descriptors, then they are orthogonalized, and after that the correlation coefficient is computed between the property P and each individual orthogonalized descriptor. From this individual computations of the total correlation coefficient is obtained which is equal to the correlation coefficient between the property P and its computed value P'. After that the same procedure is repeated for three, four, ..., n descriptors, where $n \leq N$ and $n < m$ (m = the number of molecules). The value of n is fixed at the beginning of the computation.

(3) For I-tuple of descriptors ($I = 2, 3, ..., n$), the score with the highest R is singled out, and this score possesses necessarily the smallest value of S among all possibilities generated by use of the I-tuple of the same class (there are exactly $\binom{N}{I}$ of them). In this way one obtains n best I-tuples.

(4) Among the n best I-tuples ($I = 2, 3, ..., n$) we choose one that gives the smallest S, which is at the same time the best total solution. This solves the problem of selecting the optimum number of descriptors and detecting the optimum I-tuple of descriptors for producing the best approximation of the property P.

(5) If the optimum solution is of class I, then there are $I!$ possible realizations. Because of that the orthogonalization must be carried out separately for each ordering of considered descriptors. In this case the correlation coefficients between property P and each single orthogonal descriptor change, because with the change of the order of orthogonalization the individual orthogonalized descriptors also change, but the correlation coefficient between P and P' remains the same for each order with the same I-tuple. Usually the optimum ordering is selected as the one which gives the greatest contributions to the dominant descriptors (dominant component analysis).[13]

COMPARISON WITH VECTORS

Let us consider an initial set of N nonorthogonalized descriptors (N nonorthogonal vectors with m components) by which we wish to approximate the property P (or vector). Since the descriptors (vectors) are nonorthogonal, we cannot know in advance the dimension of the space which spans N descriptors. Similarly, we cannot know the dimension of the subspace which is spanned by I descriptors ($I = 2, 3, ..., n$; $n < N$) selected from the set of N nonorthogonal descriptors. Because of linear dependence of descriptors, it is possible to express one (or more) descriptor(s) from the set of I descriptors in terms of the remaining $I - 1$ descriptors. Therefore, we use the term "I-dimensional nonorthogonal space", but we keep in mind that this term is only partially correct. The correct term to use would be "space which is spanned by I nonorthogonal vectors". Only when all I vectors are mutually orthogonal, then we can talk about I-dimensional vector (descriptor) space. In general P has m components. Note that $m > n$. Let the approximation of the property P be a certain property (vector) P' which is analogous to the projection of the vector with m components in the space which spans I vectors with m components (descriptor space). If we increase I, the angle between vectors P and P' diminishes. In reality we wish to minimize the dimension of the space of the projection P' by minimizing a certain functional. In the case of the multivariate linear relationship such functional is often the standard error, which permits the increase of the space dimensions in which P is projected. Thus, the optimum approximation according to the criterion of the minimum standard error will be in general in the space of dimension $I < n$ ($n < N$), that is, in the I-dimensional subspace of the N-dimensional space.

In the space of dimension N, there are $\binom{N}{I}$ subspaces of dimension I. For instance, for $N = 3$ there are three subspaces of dimension 2 (three planes). In the orthogonalized basis we can describe the property P with I descriptors in $I!$ ways. In the expression for the standard error all these representations will possess the same sanction factor because of the number of descriptors used for approximating the property P, if we are within the same class of descriptors (i.e., within the same I-tuple with the same I).

If the orthogonal descriptor basis is used, the correlation coefficient between the property P and its approximation P' obtained with I descriptors is given by

$$R_{P,P'} = [\sum_{i=1}^{I} R_{PP'_i}{}^2]^{1/2} \tag{5}$$

where $R_{PP'_i}$ is the correlation coefficient between the property P and its approximation P' with only one descriptor, that is, its projection on the unit axis of the ith descriptor.

In the orthogonal vector space the cosine of the angle between N-dimensional vector P and its projection P' in the vector subspace with dimension $I < N$ is

$$\cos(P,P') = \{\sum_{i=1}^{I} [\cos(P,P'_i)]^2\}^{1/2} \tag{6}$$

where P'_i are the projections of the vector P onto unit vectors which span the subspace of dimension I.

Among $\binom{N}{I}$ possible selections of descriptors, the best is that selection which gives the projection P' with the greatest value of the correlation coefficient with property P (that is, with the smallest angle between vector P with m components and its projection P' in the I-dimensional subspace). At the

Table 1. Connectivity Indices $^{l}\chi$ (l = 0, 1, ..., 6) for Octanes

octane	connectivity indices						
	$^{0}\chi$	$^{1}\chi$	$^{2}\chi$	$^{3}\chi$	$^{4}\chi$	$^{5}\chi$	$^{6}\chi$
n-octane	6.242 64	3.914 21	2.414 21	1.457 10	0.853 55	0.478 55	0.250 00
3-ethylhexane	6.405 77	3.846 06	2.471 19	1.851 62	1.105 17	0.408 24	0.000 00
4-methylheptane	6.405 77	3.808 06	2.682 52	1.562 94	1.129 93	0.288 67	0.144 33
3-methylheptane	6.405 77	3.808 06	2.655 64	1.747 40	0.756 71	0.492 79	0.144 33
2-methylheptane	6.405 77	3.770 05	2.889 62	1.385 02	0.802 58	0.433 01	0.288 67
3,4-dimethylhexane	6.568 91	3.718 74	2.117 06	2.259 30	0.804 73	1.666 66	0.000 00
3-ethyl-2-methylpentane	6.568 91	3.718 74	2.820 59	1.991 56	1.231 48	0.000 00	0.000 00
3-ethyl-3-methylpentane	6.621 32	3.681 98	2.871 32	2.560 66	0.750 00	0.000 00	0.000 00
2,3-dimethylhexane	6.568 91	3.680 73	3.009 97	1.882 08	0.788 67	0.333 33	0.000 00
2,4-dimethylhexane	6.568 91	3.663 90	3.142 96	1.570 69	0.971 40	0.333 33	0.000 00
2,5-dimethylhexane	6.568 91	3.625 89	3.365 04	1.321 36	0.666 66	0.666 66	0.000 00
3,3-dimethylhexane	6.621 32	3.621 32	3.267 76	1.883 88	0.853 55	0.176 77	0.000 00
2,2-dimethylhexane	6.621 32	3.560 66	3.664 21	1.280 33	0.707 10	0.530 33	0.000 00
2,3,4-trimethylpentane	6.732 05	3.553 41	3.347 15	2.103 13	0.769 80	0.000 00	0.000 00
2,3,3-trimethylpentane	6.784 45	3.504 03	3.496 83	2.474 17	0.408 24	0.000 00	0.000 00
2,2,3-trimethylpentane	6.784 45	3.481 38	3.675 32	2.090 77	0.612 37	0.000 00	0.000 00
2,2,4-trimethylpentane	6.784 45	3.416 50	4.158 63	1.020 62	1.224 74	0.000 00	0.000 00
2,2,3,3-tetramethylbutane	7.000 00	3.250 00	4.5000 00	2.250 00	0.000 00	0.000 00	0.000 00

same time such a selection will also minimize the standard error between the approximation P' and the I-tuple of the descriptor.

THE ORDER OF ORTHOGONALIZATION AND THE DOMINANT DESCRIPTOR

Let us select one realization of nonorthogonal subspace which is spanned by I nonorthogonal vectors in the space which spans N nonorthogonal vectors (that is, we select a subset of I descriptors out of the set of N descriptors). The orthogonalization of the subspace of dimension I can be carried out in $I!$ ways (this is really the number of possible orderings of orthogonalization). Among these possibilities we select that ordering which gives the highest projections on the dominant unit vectors of the basis. What we wish to do is to maximize the projection of a given vector P on the dominant unit vectors, that is, to reduce the angle between P and the dominant unit vectors. The cosine of this angle, as we have seen, is analogous to the correlation coefficient.

A SELECTION OF THE DOMINANT DESCRIPTOR IN THE ORTHOGONAL BASIS

If we have successfully detected the I-tuple of descriptors in orthogonal basis which approximate best the property P (that is, the structure−property correlation which has the highest value of R and smallest value of S), this is also the best I-tuple for approximating P in nonorthogonal basis. Orthogonalization of selected I-tuple gives in all possible orderings (there are, of course, $I!$ possible orderings of selected descriptors) always equally accurate approximate value of P; that is, the values of R and S do not change. However, the contribution of the dominant descriptor (and all other ($I = 1$) descriptors) changes in the total value of R according to expression (3). If we compute contributions of the dominant descriptor for every ordering of descriptors in the orthogonal basis, we can reveal the ordering at which the contribution of the dominant descriptor is optimum. This contribution appears to be greater than the contribution when the orthogonalization was initiated at this particular descriptor (and this contribution of the dominant descriptor is equal to its contribution in the nonorthogonal descriptor basis).

With optimum ordering of the orthogonalization we maximize the contribution of the dominant descriptor in such a way that mutual nonorthogonality between descriptors is transfered to the dominant descriptor before the orthogonalization in order to increase the correlation coefficient with the property P.

Table 2. Boiling Points of Octanes taken from Ref 17

octane	boiling point in °C
n-octane	125.665
3-ethylhexane	118.534
4-methylheptane	117.709
3-methylheptane	118.925
2-methylheptane	117.647
3,4-dimethylhexane	117.725
3-ethyl-2-methylpentane	115.650
3-ethyl-3-methylpentane	118.259
2,3-dimethylhexane	115.607
2,4-dimethylhexane	109.429
2,5-dimethylhexane	109.103
3,3-dimethylhexane	111.969
2,2-dimethylhexane	106.840
2,3,4-trimethylpentane	113.467
2,3,3-trimethylpentane	114.760
2,2,3-trimethylpentane	109.841
2,2,4-trimethylpentane	99.238
2,2,3,3-tetramethylbutane	106.470

EXAMPLE

For detailed discussion about the application of our procedure we selected octanes as suggested by Randić and Trinajstić,[14] since this class of alkanes shows considerable structural diversity and is modest in the size. We used the connectivity basis of Randić,[15] which can be generated using the following formula[16]

$$^{l}\chi = \sum_{\text{paths}} [d(i)\, d(j) \ldots d(l+1)]^{-1/2} \qquad (7)$$

where $d(i)$, $d(j)$, ..., $d(l + 1)$ are valencies of vertices, $i, j, \ldots, l + 1$ in the considered path of the length l. We considered the following set of connectivity indices: $^{0}\chi$, $^{1}\chi$, $^{2}\chi$, $^{3}\chi$, $^{4}\chi$, $^{5}\chi$, and $^{6}\chi$. They represent a nonorthogonal basis set. Their numerical values are reported in Table 1.

As the property, the boiling points of octanes are taken into consideration. The boiling points used are taken from Needham et al.[17] and are given in Table 2. We selected boiling points because they are "difficult" molecular proper-

Table 3. Best Possible Multiple Linear Regression Models for Boiling Points of Octanes (with I Descriptors) with the Nonorthogonal Connectivity Basis

$I = 1$
$R = 0.88165$, $S = 2.9772$, $F = 55.8$
bp = $(143.95152 \pm 4.107) + (-9.51488 \pm 1.273)\ ^{2}\chi$

$I = 2$
$R = 0.96468$, $S = 1.7164$, $F = 100.6$
bp = $(316.7375 \pm 15.698) + (-33.4307 \pm 2.476)\ ^{0}\chi + (9.5569 \pm 1.042)\ ^{3}\chi$

$I = 3$
$R = 0.99061$, $S = 0.9223$, $F = 244.9$
bp = $(-67.4405 \pm 6.857) + (53.4156 \pm 2.050)\ ^{1}\chi + (-13.1286 \pm 0.969)\ ^{4}\chi + (-12.7727 \pm 1.231)\ ^{5}\chi$

$I = 4$
$R = 0.99181$, $S = 0.8937$, $F = 196.1$
bp = $(327.0834 \pm 30.978) + (-26.7655 \pm 5.072)\ ^{0}\chi + (-7.4430 \pm 1.161)\ ^{2}\chi + (-12.4854 \pm 1.020)\ ^{4}\chi + (-13.6157 \pm 1.916)\ ^{5}\chi$

$I = 5$
$R = 0.99256$, $S = 0.8867$, $F = 159.6$
bp = $(2583.4164 \pm 450.924) + (-182.7563 \pm 25.593)\ ^{0}\chi + (-310.4976 \pm 67.625)\ ^{1}\chi + (-46.1495 \pm 12.544)\ ^{2}\chi + (8.3044 \pm 1.872)\ ^{3}\chi + (-5.6664 \pm 1.703)\ ^{5}\chi$

$I = 6$
$R = 0.99288$, $S = 0.9062$, $F = 127.4$
bp = $(1418.7410 \pm 895.483) + (-100.3094 \pm 58.706)\ ^{0}\chi + (-147.8996 \pm 124.870)\ ^{1}\chi + (-28.8510 \pm 18.986)\ ^{2}\chi + (-11.1293 \pm 2.531)\ ^{4}\chi + (-15.0798 \pm 3.410)\ ^{5}\chi + (-7.3298 \pm 7.589)\ ^{6}\chi$

$I = 7$
$R = 0.99288$, $S = 0.9062$, $F = 127.4$
bp = $(1368.8749 \pm 2345.269) + (-96.7300 \pm 160.092)\ ^{0}\chi + (-141.0455 \pm 323.114)\ ^{1}\chi + (-28.1233 \pm 37.149)\ ^{2}\chi + (-0.3283 \pm 14.149)\ ^{3}\chi + (-11.5584 \pm 18.686)\ ^{4}\chi + (-15.4209 \pm 15.127)\ ^{5}\chi + (-7.5350 \pm 11.896)\ ^{6}\chi$

ties according to the classification of Randić[5] and a lot of work is done on structure-boiling point relationships for alkanes.[11,17-20]

All results will be given in terms of orthogonalized connectivity indices $^{I}\Omega$. It should be noted that in the case of orthogonal basis the total correlation coefficient R is equal to the square root of the sum of squares of correlation coefficients $R_{PP'_i}$ ($i = 1, ..., I$) between each individual orthogonal descriptor and computed physical (chemical, biological) property P' (the boiling point in our case). Because of that in all tables based on the orthogonalized descriptors we give this number above each of the coefficients of the regression equations. This number is not a square of $R_{PP'_i}$, but just $R_{PP'_i}$, that is, the correlation coefficient between the orthogonalized descriptor i and property P'. This piece of information was not previously given in our structure–property papers, because we did not pay particular attention to R in terms of orthogonal basis. However, we have now established that $R_{PP'_i}$ is a very good criterion for appraising the significance of an individual descriptor.

RESULTS AND DISCUSSION

First we have found the best possible models in nonorthogonal connectivity basis, which produced the best possible approximation for boiling points (bp) of octanes. We used the first seven connectivity indices which produced six best models ($I = 1, 2, ..., 6$). Additionally, we investigated the model with all seven connectivity indices, though this was the only 7-tuple possible. All these seven model are given in Table 3.

As the criteria of the goodness of the model are listed, respectively, the correlation coefficient, the standard error, and the F-test. We also considered it useful to list the standard error of all coefficients in the model, though many of other authors usually do not give these data. From these data it is possible to assess the significance of the descriptor as reflected in the stability of each coefficient at a given descriptor. Stability of coefficients determine the power of the predictability of the model. This will be analyzed later in the text when the comparison is made with models based on orthogonal connectivity basis which are better than the best models that can be obtained in nonorthogonal connectivity basis. If the criteria for selecting the best model are used (either the standard error or the F-test), one sees that a different model can be obtained as the best. Criterion of the smallest standard error indicate the model with five descriptors to be the best, while if the F-test is considered, then the model with three descriptors appears to be better.

Dependence of the model on the order of orthogonalization (see Table 4) is investigated on the model based on the three descriptors from Table 3 ($^{1}\chi$, $^{4}\chi$, $^{5}\chi$). This is the model which gives the best approximation with three descriptors in nonorthogonal connectivity basis. Each of six possible orderings of orthogonalization gives a model which possesses the same R, S, and F-test. However, the coefficients at each orthogonal descriptor differ in the model, this being the consequence of different contributions of separate descriptors in the model (that is, the difference between various correlation coefficients between the property which is approximated—in our case boiling point—and individual descriptors) depending on the order of orthogonalization by which the descriptors have been obtained.

Above each model in Table 4 are given correlation coefficients between each individual descriptor and the property that is approximated. From this piece of information it is possible to observe that the correlation coefficient between bp and, for example, descriptor $^{5}\Omega$ does not depend on the order of orthogonalization of descriptors that were orthogonalized before it. The orthogonal descriptor $^{5}\Omega$ obtained by the orthogonalization in the order $^{1}\chi$, $^{4}\chi$, $^{5}\chi$ is identical to the orthogonal descriptor $^{5}\Omega$ obtained by orthogonalization in the different order $^{4}\chi$, $^{1}\chi$, $^{5}\chi$. For certain orthogonalization orderings of some descriptors it is possible to obtain higher correlation coefficients between the property that is approximated and a given orthogonal descriptor than

Table 4. Best Possible Multiple Linear Regression Models for Boiling Points of Octanes with Three Descriptors ($^1\chi$, $^4\chi$, $^5\chi$) with the Orthogonalized Basis[a]

orthogonalization ordering: $^1\chi$, $^4\chi$, $^5\chi$
$R = 0.99061$, $S = 0.9222$, $F = 244.9$
$R_{PP'i} = (^1\Omega$: 0.821; $^4\Omega$: 0.404; $^5\Omega$: 0.379)
bp = 113.7132 + 30.3282 $^1\Omega$ + 12.2696 $^4\Omega$ − 12.7727 $^5\Omega$

orthgonalization ordering: $^1\chi$, $^5\chi$, $^4\chi$
$R = 0.99061$, $S = 0.9222$, $F = 244.9$
$R_{PP'i} = (^1\Omega$: 0.821; $^4\Omega$: 0.495; $^5\Omega$: 0.249)
bp = 113.7132 + 30.3282 $^1\Omega$ − 13.1287 14Ω + 8.0285 $^5\Omega$

orthogonalization ordering: $^4\chi$, $^1\chi$, $^5\chi$
$R = 0.99061$, $S = 0.922$, $F = 244.9$
$R_{PP'i} = (^1\Omega$: 0.905; $^4\Omega$: 0.138; $^5\Omega$: 0.379)
bp = 113.7132 − 40.7258 $^1\Omega$ + 2.8794 $^4\Omega$ − 12.7727 $^5\Omega$

orthogonalization ordering: $^4\chi$, $^5\chi$, $^1\chi$
$R = 0.99061$, $S = 0.9222$, $F = 244.9$
$R_{PP'i} = (^1\Omega$: 0.952; $^4\Omega$: 0.138; $^5\Omega$: 0.236)
bp = 113.7132 + 53.4160 $^1\Omega$ + 2.8792 $^4\Omega$ − 6.3686 $^5\Omega$

orthogonalization ordering: $^5\chi$, $^1\chi$, $^4\chi$
$R = 0.99061$, $S = 0.9222$, $F = 244.9$
$R_{PP'i} = (^1\Omega$: 0.820; $^4\Omega$: 0.495; $^5\Omega$: 0.251)
bp = 113.7132 − 36.4850 $^1\Omega$ − 13.1287 $^4\Omega$ + 6.7317 $^5\Omega$

orthogonalization ordering: $^5\chi$, $^4\chi$, $^1\chi$
$R = 0.99061$, $S = 0.9222$, $F = 244.9$
$R_{PP'i} = (^1\Omega$: 0.952; $^4\Omega$: 0.108; $^5\Omega$: 0.251)
bp = 113.7132 + 53.4161 $^1\Omega$ − 2.2619 $^4\Omega$ + 6.7313 $^5\Omega$

[a] It shows the dependence of significance of the descriptor and coefficients in the model on the orthogonalization order.

when the nonorthogonalized connectivity basis is used. The meaning of this is that by orthogonalization of a certain descriptor in a specific order we can obtain the orthogonalized descriptor which approximates better a given property than the related nonorthogonalized descriptor.

The models given in Table 3 were then computed using the orthogonalized connectivity basis. They are shown in Table 5.

The ordering of orthogonalization was selected in such a way to maximize the contributions of dominant descriptors. In this manner the number of models that remain to be considered diminishes. Among all models that maximize the contributions by the first dominant descriptor we select only those which maximize the contributions by the second dominant descriptor. This procedure repeats for each descriptor used in building up the structure−property model. For example, for the model with four descriptors we can choose at the most three dominant descriptors, since the fourth is in this way automatically determined. The above procedure offers one way for the orthogonalization ordering. In this way the orthogonalization orderings of models with up to seven descriptors in Table 5 were chosen. The comparison between the corresponding models in Tables 3−5 reveals that coefficients in the models with orthogonalized connectivity basis are in most cases more stable, that is, these coefficients have a smaller ratio between the individual standard error and its absolute value. This is especially so for the constants in all models.

A closer inspection of the correlation coefficients between individual descriptors, in particular in models with four, five, and six descriptors from Table 5, reveals that each of these models contains insignificant descriptors. They are the consequence of the selection of the orthogonalization ordering which maximizes the contributions of dominant descriptors. The insignificant descriptors can be removed from the model, because in this way we obtain the model with a smaller number of descriptors the outcome of which is the betterment of the S values and F-tests, although the correlation coefficients become slightly smaller. This is so because the numerical contributions of these insignificant descriptors are often found only in the third or fourth decimal place in the correlation coefficient. Models obtained by omitting the insignificant descriptors are given in Table 6.

The removal of descriptor $^0\Omega$ from the model with four descriptors results in the model with much higher value of the F-test, but the standard error also increases and this does not represent a particular improvement. However, in the case of the model with five descriptors we can remove $^5\Omega$ achieving the model with four descriptors with a much higher value of the F-test and a considerably smaller value of S. The same can never be attained if we work with nonorthogonal connectivity basis. An even better model can be obtained if we remove $^6\Omega$ and $^2\Omega$ insignificant descriptors from the model with six descriptors. The obtained model with four descriptors gives the S the value of 0.8553, while the best model with nonorthogonal basis following the criterion of minimizing the standard error produces for S the value of 0.8868, and this result is achieved with five descriptors.

A closer inspection of the standard errors of the coefficients in the models before and after the removal of insignificant descriptors indicates that the coefficients become more stable, that is, their standard error reduces compared to the absolute value of the coefficient. This points to increasing power of the predictability of such a model.

It is important to mention that, for example, for assembling the structure−property model consisting of four descriptors, which is obtained after the removal of insignificant orthogonal descriptor $^5\Omega$ from the model with five descriptors in Table 6, there are at first necessary five descriptors: $^3\chi$, $^0\chi$, $^5\chi$, $^1\chi$, and $^2\chi$. These five descriptors in the presented order of orthogonalization lead to the result that the orthogonal descriptor $^5\Omega$ is insignificant, and by its removal we obtain the model with four descriptors with a smaller value of S and a greater value of the F-test. Similarly, one can see that for obtaining $^1\Omega$ and $^2\Omega$ orthogonal descriptors in the same model, the following nonorthogonal descriptors are needed: $^3\chi$, $^0\chi$, and $^5\chi$. The meaning of this is that only when the orthogonalization is carried out in the designated order, then the obtained descriptors produce the model with properties given in Table 6. This leads to the conclusion that for approximating a given molecular property there are also important those descriptors which do not take direct part in building up the model, but they are needed in the process of orthogonalization since we want to obtain the best possible set of orthogonal descriptors which are directly involved in the model.

CONCLUDING REMARKS

The application of the multiple linear regression allowed the construction of the structure−property models, based on the connectivity basis, which approximate well the boiling points of octanes. We have examined the best possible models based on the nonorthogonal connectivity basis and the corresponding models based on the orthogonal basis. Our models, based on the nonorthogonal connectivity basis, are the best possible models of all models that can be obtained

Table 5. Best Possible Multiple Linear Regression Models for Boiling Points of Octanes (with I Descriptors) with the Orthogonal Connectivity Basis[a]

$I = 1$; orthogonalization ordering: $^{2}\chi$
$R = 0.88165$, $S = 2.9772$, $F = 55.8$
bp = $(143.95152 \pm 4.107) + (-9.51488 \pm 1.273)\ ^{2}\Omega$

$I = 2$; orthogonalization ordering: $^{3}\chi$, $^{0}\chi$
$R = 0.96468$, $S = 1.7164$, $F = 100.6$
$R_{PP'i}$ = ($^{0}\Omega$: 0.294; $^{3}\Omega$: 0.919)
bp = $(113.7132 \pm 0.405) + (33.4306 \pm 2.476)\ ^{0}\Omega + (4.1724 \pm 0.963)\ ^{3}\Omega$

$I = 3$; orthogonalization ordering: $^{5}\chi$, $^{4}\chi$, $^{1}\chi$
$R = 0.99061$, $S = 0.9222$, $F = 244.9$
$R_{PP'i}$ = ($^{1}\Omega$: 0.952; $^{4}\Omega$: 0.108; $^{5}\Omega$: 0.251)
bp = $(113.7132 \pm 0.217) + (53.4161 \pm 2.050)\ ^{1}\Omega + (-2.2619 \pm 0.768)\ ^{4}\Omega + (6.7313 \pm 0.980)\ ^{5}\Omega$

$I = 4$; orthogonalization ordering: $^{4}\chi$, $^{2}\chi$, $^{0}\chi$, $^{5}\chi$
$R = 0.99181$, $S = 0.8937$, $F = 196.1$
$R_{PP'i}$ = ($^{0}\Omega$: 0.043; $^{2}\Omega$: 0.948; $^{4}\Omega$: 0.138; $^{5}\Omega$: 252)
bp = $(113.7132 \pm 0.211) + (3.3541 \pm 2.786)\ ^{0}\Omega + (11.9875 \pm 0.448)\ ^{2}\Omega + (2.8793 \pm 0.739)\ ^{4}\Omega + (13.6157 \pm 1.916)\ ^{5}\Omega$

$I = 5$; orthogonalization ordering: $^{3}\chi$, $^{0}\chi$, $^{5}\chi$, $^{1}\chi$, $^{2}\chi$
$R = 0.99256$, $S = 0.8868$, $F = 159.6$
$R_{PP'i}$ = ($^{0}\Omega$: 0.919; $^{1}\Omega$: 0.194; $^{2}\Omega$: 0.129; $^{3}\Omega$: 0.295; $^{5}\Omega$: 0.020)
bp = $(113.7133 \pm 0.209) + (33.4304 \pm 1.279)\ ^{0}\Omega + (65.5945 \pm 11.909)\ ^{1}\Omega + (-46.1457 \pm 12.546)\ ^{2}\Omega + (4.1723 \pm 0.497)\ ^{3}\Omega + (-0.8723 \pm 1.529)\ ^{5}\Omega$

$I = 6$; orthogonalization ordering: $^{5}\chi$, $^{4}\chi$, $^{1}\chi$, $^{2}\chi$, $^{6}\chi$, $^{0}\chi$
$R = 0.99288$, $S = 0.9062$, $F = 127.4$
$R_{PP'i}$ = ($^{0}\Omega$: 0.061; $^{1}\Omega$: 0.952; $^{2}\Omega$: 0.027; $^{4}\Omega$: 0.108; $^{5}\Omega$: 0.251; $^{6}\Omega$: 0.003)
bp = $(113.7133 \pm 0.214) + (100.2785 \pm 58.701)\ ^{0}\Omega + (53.4162 \pm 2.014)\ ^{1}\Omega + (-2.3143 \pm 3.054)\ ^{2}\Omega + (-2.2620 \pm 0.755)\ ^{4}\Omega + (6.7314 \pm 0.964)\ ^{5}\Omega + (0.5225 \pm 6.039)\ ^{6}\Omega$

$I = 7$; orthogonalization ordering: $^{5}\chi$, $^{4}\chi$, $^{1}\chi$, $^{2}\chi$, $^{6}\chi$, $^{0}\chi$, $^{3}\chi$
$R = 0.99288$, $S = 0.9504$, $F = 99.3$
$R_{PP'i}$ = ($^{0}\Omega$: 0.061; $^{1}\pm$: 0.952; $^{2}\Omega$: 0.027; $^{3}\Omega$: 0.001; $^{4}\Omega$: 0.108; $^{5}\Omega$: 0.251; $^{6}\Omega$: 0.003)
bp = $(113.7133 \pm 0.224) + (100.2784 \pm 61.565)\ ^{0}\Omega + (53.4162 \pm 2.113)\ ^{1}\Omega + (-2.3143 \pm 3.203)\ ^{2}\Omega + (-0.3246 \pm 14.150)\ ^{3}\Omega + (-2.2620 \pm 0.792)\ ^{4}\Omega + (6.7314 \pm 1.010)\ ^{5}\Omega + (0.5225 \pm 6.333)\ ^{6}\Omega$

[a] The orthogonalization order was selected in such a way to maximize the contributions of the first, then the second, then the third, etc. dominant descriptor.

Table 6. Structure–Property Models Obtained by Removing Insignificant Descriptors (That Is, Descriptors with the Smallest $R_{PP'i}$) from the Models Presented in Table 5

$I = 4$; orthogonalization ordering: $^{4}\chi$, $^{2}\chi$, $^{0}\chi$, $^{5}\chi$
$R = 0.99181$, $S = 0.8937$, $F = 196.1$
$R_{PP'i}$ = ($^{0}\Omega$: 0.043; $^{2}\Omega$: 0.948; $^{4}\Omega$: 0.138; $^{5}\Omega$: 0.252)
bp = $(113.7132 \pm 0.211) + (3.3541 \pm 2.786)\ ^{0}\Omega + (11.9875 \pm 0.448)\ ^{2}\Omega + (2.8793 \pm 0.739)\ ^{4}\Omega + (13.6157 \pm 1.916)\ ^{5}\Omega$

model with removed $^{0}\Omega$
$R = 0.99090$, $S = 0.9080$, $F = 252.8$
bp = $(113.7132 \pm 0.214) + (11.9875 \pm 0.455)\ ^{2}\Omega + (2.8793 \pm 0.750)\ ^{4}\Omega + (13.6157 \pm 1.946)^{5}\Omega$

$I = 5$; orthogonalization ordering: $^{3}\chi$, $^{0}\chi$, $^{5}\chi$, $^{1}\chi$, $^{2}\chi$
$R = 0.99256$, $S = 0.8868$, $F = 159.6$
$R_{PP'i}$ = ($^{0}\Omega$: 0.919; $^{1}\Omega$: 0.194; $^{2}\Omega$: 0.129; $^{3}\Omega$: 0.295; $^{5}\Omega$: 0.020)
bp = $(113.7133 \pm 0.209) + (33.4304 \pm 1.279)\ ^{0}\Omega + (65.5945 \pm 11.909)\ ^{1}\Omega + (-46.1457 \pm 12.546)\ ^{2}\Omega + (4.1723 \pm 0.497)\ ^{3}\Omega + (-0.8723 \pm 1.529)\ ^{5}\Omega$

model with removed $^{5}\Omega$
$R = 0.99236$, $S = 0.86349$, $F = 210.3$
bp = $(113.7133 \pm 0.204) + (33.4304 \pm 1.245)\ ^{0}\Omega + (65.5946 \pm 11.596)\ ^{1}\Omega + (-46.1456 \pm 12.216)\ ^{2}\Omega + (4.1723 \pm 0.484)\ ^{3}\Omega$

$I = 6$; orthogonalized ordering: $^{5}\chi$, $^{4}\chi$, $^{1}\chi$, $^{2}\chi$, $^{6}\chi$, $^{0}\chi$
$R = 0.99288$, $S = 0.9062$, $F = 127.4$
$R_{PP'i}$ = ($^{0}\Omega$: 0.061; $^{1}\Omega$: 0.952; $^{2}\Omega$: 0.027; $^{4}\Omega$: 0.108; $^{5}\omega$: 0.251; $^{6}\Omega$: 0.003)
bp = $(113.7133 \pm 0.214) + (100.2785 \pm 58.701)\ ^{0}\Omega + (53.4162 \pm 2.014)\ ^{1}\Omega + (-2.3143 \pm 3.054)\ ^{2}\Omega + (-2.2620 \pm 0.755)\ ^{4}\Omega + (6.7314 \pm 0.964)\ ^{5}\Omega + (0.5225 \pm 6.039)\ ^{6}\Omega$

model with removed $^{6}\Omega$
$R = 0.99288$, $S = 0.8679$, $F = 166.7$
bp = $(113.7133 \pm 0.205) + (100.2795 \pm 56.221)\ ^{0}\Omega + (53.4162 \pm 1.929)\ ^{1}\Omega + (-2.3143 \pm 2.925)\ ^{2}\Omega + (-2.2620 \pm 0.723)\ ^{4}\Omega + (6.7314 \pm 0.923)\ ^{5}\Omega$

model with removed $^{6}\Omega$, $^{2}\Omega$
$R = 0.99250$, $S = 0.8553$, $F = 214.4$
bp = $(113.7133 \pm 0.202) + (100.2795 \pm 55.407)\ ^{0}\Omega + (53.4162 \pm 1.901)\ ^{1}\Omega + (-2.2620 \pm 0.712)\ ^{4}\Omega + (6.7314 \pm 0.909)\ ^{5}\Omega$

with nonorthogonal descriptors because they are obtained by investigating all possible combinations between descriptors for all I-tuples generated from the set of seven connectivity indices ($^{l}\chi$, $l = 0, 1, ..., 6$). This procedure allows the selection of the best I-tuples according to the chosen criterion, that is, the criterion of the minimum standard error. The structure–property models based on the nonorthogonal connectivity basis are essentially better than the models

which can be generated by the stepwise/inclusion−exclusion or some other "greedy" algorithm.[11] Among all possible orthogonalization orderings we selected the one which maximizes the contribution of one descriptor and which necessarily minimizes the contributions of other descriptors some of which may during this procedure become insignificant. This is so because the square of the total correlation coefficient between the experimental value of the molecular property, and its calculated value by means of I orthogonalized descriptors is equal to the sum of the squares of the correlation coefficients between the value of the molecular property and each individual orthogonal descriptor in the model and is constant for each orthogonalization ordering. This is the reason why we obtained, by removing the insignificant descriptors from the models based on the orthogonal basis, considerably better models (judging by their statistical characteristics). This, of course, is not possible if we work with models based on nonorthogonal connectivity basis.

To our knowledge this is the first example which shows that the structure−property modeling within the orthogonalized descriptor basis produces much better models than the modeling with the nonorthogonalized basis. Albeit the orthogonal descriptors are used already for several years in QSPR and QSAR studies, their use never produced a better and more accurate model than the modeling within the nonorthogonal basis. The usual conclusion of structure−property−activity studies with the orthogonalized basis was that the model with the orthogonalized descriptors is more stable, but the statistical parameters (R, S, and F-test) remained the same as they were produced by models with nonorthogonalized descriptors. The reason for the above are clear after the present report. All that was needed was to prepare a computer program for detailed study about the dependence of the coefficients in the multiple linear regression with the orthogonal basis on the orthogonalization order. After we succeeded to do that we discovered a number of possibilities to improve this already classical method for constructing the structure−property−activity models. The results presented, obtained by the multiple linear regression within the orthogonal descriptor basis, clearly indicate that we are developing a very powerful novel method which could possibly completely substitute the multiple linear regression within the nonorthogonal basis in QSPR and QSAR studies, simply because it always produces better (more accurate, more stable) models with higher power of the predictability.

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. We are thankful to Milan Randić, Sanja Sekušak, Tomislav Došlić, and the reviewers for their constructive comments.

REFERENCES AND NOTES

(1) Lukovits, I. Quantitative Structure−Activity Relationships Employing Independent Quantum Chemical Indexes. *J. Med. Chem.* **1983**, *26*, 1104−1109.

(2) Randić, M. Orthogonal Molecular Descriptors. *New J. Chem.* **1991**, *15*, 517−525.

(3) E.g.: Amić, D.; Davidović-Amić, D.; Trinajstić, N. Calculation of Retention Times of Anthocyanins with Orthogonalized Topological Indices. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 136−139.

(4) There are at present more than 120 descriptors in the literature: See, for example: Rouvray, D. H. The Limits of Applicability of Topological Indices. *J. Mol. Struct. (Theochem)* **1989**, *185*, 187−201.

(5) Randić, M. Comparative Regression Analysis. Regressions Based on a Single Descriptor. *Croat. Chem. Acta* **1993**, *66*, 289−312.

(6) Garbalena, M.; Herndon, W. C. Optimum Graph-Theoretical Models for Enthalpic Properties of Alkanes. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 37−42.

(7) E.g.: Kier, L. B.; Hall, L. H. Molecular Connectivity in Structure−Activity Analysis. Wiley: New York, 1986.

(8) Randić, M. On Computation of Optimal Parameters for Multivariate Analysis of Structure−Property Relationship. *J. Comput. Chem.* **1991**, *12*, 970−980.

(9) Randić, M. Resolution of Ambiguities in Structure−Property Studies by Use of Orthogonal Descriptors. *J. Chem. Inf. Comput. Sci.* **1991**, *31*, 317−320.

(10) Randić, M. Chemical Structure—What is "She"? *J. Chem. Educ.* **1992**, *63*, 713−718.

(11) Randić, M.; Trinajstić, N. Isomeric Variations in Alkanes. Boiling Points of Nonanes. *New J. Chem.* **1994**, *18*, 179−189.

(12) Spiegel, M. R. *Statistics*; Schaum Publ. Co.: New York, 1961.

(13) Randić, M.; Trinajstić, N. Viewpoint 4—Comparative Structure−Property Studies: The Connectivity Basis. *J. Mol. Struct. (Theochem)* **1993**, *284*, 209−221.

(14) Randić, M.; Trinajstić, N. In Search of Graph Invariants of Chemical Interest. *J. Mol. Struct. (Theochem)* **1993**, *300*, 551. Other authors also use octanes as a convenient set of alkanes for initial testing of various topological indices in structure−property−activity procedures. E.g.: Kirby, E. C. Sensitivity of Topological Indices to Methyl Group Branching in Octanes and Azulenes, or What Does a Topological Index? *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 1030−1035.

(15) Randić, M. Representation of Molecular Graphs by Basic Graphs. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 57−69.

(16) Kier, L. B.; Murray, W. J.; Randić, M.; Hall, L. H. Molecular Connectivity V: Connectivity Series Applied to Density. *J. Pharm. Sci.* **1976**, *65*, 1226−1230.

(17) Needham, D. E.; Wei, I.-C.; Seybold, P. G. Molecular Modeling of the Physical Properties of the Alkanes. *J. Am. Chem. Soc.* **1988**, *110*, 4186−4194.

(18) Rouvray, D. H. In *Graph Theory and Topology in Chemistry*; King, R. B., Rouvray, D. H., Eds.; Elsevier: New York, 1987; pp 177−193.

(19) Mihalić, Z.; Nikolić, S.; Trinajstić, N. Comparative Study of Molecular Descriptors Derived from the Distance Matrix. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 28−37.

(20) Cherqaoui, D.; Villemin, D. Use of a Neural Network to Determine the Boiling Points of Alkanes. *J. Chem. Soc., Faraday Soc.* **1994**, *90*, 97−102.

CI940131H