CROATICA CHEMICA ACTA CCACAA **69** (1) 95–109 (1996)

ISSN 0011-1643
CCA–2291 *Original Scientific Paper*

A Strategy for Molecular Modeling of a Physicochemical Property using a Linear Combination of Connectivity Indexes

Lionello Pogliani

Dipartimento di Chimica, Università della Calabria, 87030 Rende (CS) Italy

Received November 21, 1994; revised June 15, 1995; accepted June 20, 1995

A strategy to employ a linear combination of the connectivity indexes (LCCI) method to model a physicochemical property of a series of molecules is outlined throughout. The chosen physicochemical property is the solubility of 19 natural amino acids. This property is modeled with the aid of normal LCCI and of linear combinations of special constructions of connectivty indexes relating to different numbers of amino acids, including and excluding extreme outliers. The employed indexes are analyzed following their descriptive power of solubility of natural amino acids. A linear combination of reciprocal connectivity indexes (LCRCI) showed the best mapping of the water solubility of 16 amino acids while a model based on LCRCI, along with the use of supraconnectivity indexes for the amino acids proline, serine and arginine, achieved very good modeling of the water solubility of the entire set of n = 19 natural amino acids. Linear combinations of fragment connectivty indexes were also analyzed while linear combinations of orthogonal connectivty indexes (LCOCI) were used to improve the modeling and to detect dominant descriptors for this physicochemical property. Non connectivity '*ad hoc*' indexes, already used in a previous study with good success, did not achieve the same good quality as the linear combination of reciprocal connectivity indexes.

INTRODUCTION

Recently, a graph theoretical molecular modeling (MM) method formed from linear combinations of molecular connectivity indexes, LCCI-MM, was successfully applied to the modeling of seven different physicochemical properties of different numbers of natural amino acids.[1–3] This method has been

recently refined and used to encode other physicochemical properties of amino acids, plus some different properties of inorganic salts as well as different size- and shape-dependent physicochemical properties of different sets of organic saturated and unsaturated compounds, including properties dependent on the conformational isomerism in unsaturated molecules.[4–7] Linear combinations of orthogonal indexes (LCOCI) have also been applied to improve the modeling and to derive the dominant descriptors of a specific property, whenever possible. The parallelism between the LCCI-MM method and the LCAO-MO (linear combination of atomic orbitals – molecular orbital) method from quantum theory is somewhat stunning but while the LCAO-MO constructs, with non-structure-explicit basis functions, the molecualr orbitals (MO) that are subsequently used to derive the values of the physical properties, the LCCI-MM method uses structure-explicit topological indexes[8–17] to estimate physical properties or biological activities.

The aim of the present paper will be to delineate a strategy of modeling the solubility of 19 natural amino acids using a linear combination of connectivity indexes (LCCI) or linear combinations of special $X = f(\chi)$ constructions of connectivity indexes (LCXCI). It should be mentioned that modeling of the solubility of 13 amino acids has already been tried[1,2] with different kinds of connectivity and non-connectivity indexes and that recently[6] it was suggested that connectivity indexes are defective in modeling sets of compounds containing species that undergo association or strong solvation phenomena in solution.

METHOD

The molecular connectivity indexes are computed using the following relations

$$D = \Sigma_i \delta_i \tag{1}$$

$$^m\chi_t = \Sigma_p(\delta_1 \ldots \delta_{m+1})^{-1/2} \; . \tag{2}$$

Delta and valence delta values are atom-level numbers[9] that describe, respectively, the numbers of nearest-neighbours and the number of valence electrons of a non-hydrogen (heteroatom) atom of a molecule. In these equations, m = 0, 1, 2,..., is the order and t the type (path, path-cluster, cluster, *etc.*; for more details see Ref. 9.) of the molecular connectivity index. Summation in Eq. (1) runs over the different delta values of a molecule while in Eq. (2) it runs over the m-order paths; subscripts 1, 2, m+1 describe the delta values of adjacent atoms in a molecule. The corresponding valence molecular connectivity indexes are obtained by introducing the valence delta

values δ^v of the heteroatoms in amino acids into Eqs. (1) and (2). The following minimal set of χ connectivity indexes will be used in this study (in ${}^0\chi$ the summation runs over the number of heteroatoms and in ${}^1\kappa$ it runs over the number of σ bonds. The type t index is meaningless here since zeroth- and first-order indexes only are used)

$$\{\chi\} = \{D, D^v, {}^0\chi, {}^0\chi^v, {}^1\chi, {}^1\chi^v\} \ .$$

Modeling of every value of solubility S is accomplished with the aid of the following dot product

$$S = \boldsymbol{C} \cdot \chi \tag{3}$$

where $\boldsymbol{C}$ is the row correlation vector resulting from the multivariate analysis and χ is the best connectivity vector, made up of the parameters of the given minimal χ set plus the unitary connectivity index $\chi^0 \equiv 1$.[2,5,6] The LCCI-MM method searches the entire space of possible combinations of the given six χ indexes (63 combinations), sorting the best $Q = r/s^2$ combinations with their F ($F = f \cdot r^2/[v \cdot (1 - r^2)]$, f = freedom degrees and v = number of variables) values. While, the quality Q factor minimizes the standard deviation of estimates s for a given r (correlation coefficient), F controls that, for a given r, the number of χ variables does not grow excessively. As some of the estimated values are sometimes negative, it is advantageous to use the following modeling equation

$$S = | \boldsymbol{C} \cdot \chi | \tag{4}$$

where bars stand for absolute values. Normally, Eq. (4) improves the description of the estimated property.

In this study, orthogonal Ω connectivity indexes and the corresponding LCOCI will be also used to improve the modeling of the solubility values. Use of these orthogonal indexes allows bypasing the problem of partial collinearity of the given connectivity χ indexes, a collinearity that, nevertheless, does not prevent optimal modeling of the studied properties.[19–22,2,3,5–7] Normally, the orthogonalization procedure chooses the best single χ index as the first orthogonal ${}^1\Omega$ index and goes on deriving from the next best ${}^i\chi$ index the corresponding ${}^i\Omega$ index orthogonal to every previous ${}^j\Omega$ index with $j = 1, 2..., i-1$ (for more details see Refs. 19–22.). The degree of collinearity between two χ and χ' indexes is detected following the collinearity criterion[14] that states that the correlation coefficient r of the linear regression, $\chi = a\chi' + b$, is taken as a measure of collinearity and a strong collinearity is characterized by $r(\chi,\chi') > 0.98$. To measure the interrelation of a full

set of indexes used to describe a property, the mean correlation coefficient $< r_{IM}(P:\{\chi\})>$ of the interrelation matrix[2] can be used.

RESULTS AND DISCUSSION

In Tables I and II, the connectivity and fragment connectivity indexes, respectively, for 19 natural amino acids, together with their water solubility values (Table I) taken from Ref. 18. (Table I), have been collected.

In the course of a previous study[1,2] on amino acids it was realized that the water solubility of a set of $n = 13$ natural amino acids was well modeled by a linear combination of fragment connectivity indexes (LCFCI), that is, of χ_f indexes based mainly on the functional groups of the amino acids. In fact, the following χ_f and correlation $\boldsymbol{C}$ vectors achieved satisfactory modeling

$$\chi_f = (D, D^v, {}^0\chi, {}^0\chi^v, \chi^0)_f, \quad \boldsymbol{C} = (49.69, 57.83, 127.5, -1070, 282.8)$$

$$Q = 0.122, \; F = 87, \; r = 0.989, \; s = 8.1, \; n = 13 \; .$$

TABLE I

Experimental water solubility S at 25 °C in units of grams per kilogram of water of 19 amino acids (AA) and their molecular connectivity index values

AA	S	D	D^v	${}^0\chi$	${}^0\chi^v$	${}^1\chi$	${}^1\chi^v$
Pro	1622	16	28	5.983	4.554	3.805	2.767
Ser	422	12	28	5.862	3.664	3.181	1.774
Gly	251	8	20	4.284	2.640	2.270	1.190
Arg	181	22	42	9.560	6.709	5.537	3.600
Ala	167	10	22	5.155	3.510	2.643	1.627
Thr	97	14	30	6.732	4.535	3.553	2.219
Val	58	16	28	6.732	5.088	3.553	2.538
Met	56	16	26.7	7.276	6.146	4.181	4.044
His	43	24	42	8.268	5.819	5.198	3.155
Gln	42	18	38	8.146	5.410	4.537	2.804
Ile	34	16	28	7.439	5.795	4.091	3.076
Phe	29	24	42	8.975	6.604	5.698	3.722
Asn	25	16	36	7.439	4.703	4.037	2.304
Leu	23	16	28	7.439	5.795	4.036	3.021
Trp	12	32	54	10.836	8.104	7.182	4.716
Glu	8.6	18	40	8.146	5.280	4.537	2.739
Lys	6	18	32	7.983	5.916	4.681	3.366
Asp	5	16	38	7.439	4.572	4.037	2.239
Tyr	0.5	26	48	9.845	6.974	6.092	3.857

TABLE II

Fragment χ_f connectivity values of 19 natural amino acids (AA)

AA	D	D^v	${}^0\chi$	${}^0\chi^v$	${}^1\chi$	${}^1\chi^v$
Pro	7	19	3.284	1.855	1.896	1.005
Ser	7	23	4.577	2.380	2.772	1.366
Gly	6	18	3.577	1.933	2.065	1.050
Arg	4	14	2.870	1.433	1.249	0.546
Ala	6	18	3.577	1.933	2.065	1.050
Thr	7	23	4.577	3.380	2.642	1.308
Val	6	18	3.577	1.933	2.065	1.050
Met	6	18	3.577	1.933	2.065	1.050
His	4	13	2.870	1.486	1.565	0.792
Gln	9	25	4.154	2.264	2.473	1.319
Ile	6	18	3.577	1.933	2.065	1.050
Phe	6	18	3.577	1.933	2.065	1.050
Asn	9	25	4.154	3.264	2.473	1.319
Leu	6	18	3.577	1.933	2.065	1.050
Trp	6	18	3.577	1.933	2.065	1.050
Glu	11	33	6.154	3.288	3.628	1.831
Lys	5	15	2.577	1.356	1.358	0.642
Asp	11	33	6.154	3.288	3.628	1.831
Tyr	7	23	4.577	2.380	2.642	1.274

If fragment χ_f indexes are used to model the recently published[18] solubility of $n = 19$ amino acids, the result achieved is, however, very poor. The best linear combination of $\{D, D^v, {}^0\chi^v, {}^1\chi, {}^1\chi^v, \chi^0\}_f$ indexes rates: $Q = 0.00093$, $F = 0.42$, $r = 0.374$, $s = 403$, $n = 19$.

While in the following paragraph an attempt will be made to find the optimal modeling of the solubility values of amino acids without leaving the frame of the molecular connectivity theory, an outlined by Randić, Kier and Hall (RKH), it should be added that during the mentioned study it was found that the solubility of amino acids was well modeled by a set of 6 low collinear characteristic properties, designed as Γ indexes[2] and collected in Table III. These Γ indexes, derived by Kidera, Konisci, Oka, Ooi and Scheraga[2,24] and designated with the acronym KOKOS, are *'ad hoc'* indexes since they have been derived by applying several multivariate statistical analyses to 188 physical properties (many directly related to solubility) of amino acids in proteins. When these indexes are used to model the solubility of $n = 19$ amino amino acids, the following satisfactory best linear combinations (LCΓI) are obtained

$\{{}^1\Gamma, {}^2\Gamma, {}^3\Gamma, {}^4\Gamma, {}^5\Gamma, {}^6\Gamma\}$: $Q = 0.0061$, $F = 18.0$, $r = 0.935$, $s = 154.3$.

TABLE III

KOKOS ${}^{i}\Gamma$ indexes of $n = 19$ natural amino acids (AA)

AA	${}^{1}\Gamma$	${}^{2}\Gamma$	${}^{3}\Gamma$	${}^{4}\Gamma$	${}^{5}\Gamma$	${}^{6}\Gamma$
Pro	−0.71	0.9	0.21	−0.72	−1.26	0.86
Ser	−1.21	−1.19	−0.33	−0.46	−0.54	0.22
Gly	−2.16	−1.02	−0.19	−0.03	−0.84	−0.99
Arg	1.16	−0.57	−1.52	−1.07	−0.28	−0.13
Ala	−1.44	−0.47	0.11	0.32	−0.51	−0.86
Thr	−0.67	−0.97	0.1	−0.36	0.57	0.86
Val	−0.34	0.42	0.77	1.38	1.84	1.66
Met	0.44	0.2	0.72	1	0.45	0.24
His	0.52	−0.46	−0.18	−0.13	−0.56	−0.1
Gln	0.22	−1.24	−0.46	−1.05	0.19	−0.42
Ile	0.21	1.37	0.97	1.52	1.91	1.27
Phe	1.09	1.6	1.24	1.16	0.88	0.48
Asn	−0.34	−1.25	−0.06	−0.96	−1	−1.19
Leu	0.25	1.06	1.01	1.14	0.69	0.02
Trp	2.08	2.06	1.55	0.67	0.61	0.42
Glu	0.17	−0.62	−1.65	−1.03	−1.74	−1.78
Lys	0.68	−0.16	−1.62	−1.76	−0.86	−0.19
Asp	−0.54	−0.75	−1.74	−1.07	−1.17	−1.72
Tyr	1.34	1.16	1.04	−0.07	1.02	1.21

If LCFCI are poor descriptors of the full $n = 19$ set of solubility values, the best way is to revert, as a starting point, to the normal χ indexes and to the corresponding LCCI. The search for the Q-best LCCI ends up with the following insufficient LCCI modeling

$$\{{}^{0}\chi, {}^{1}\chi\} : \quad Q = 0.0022,\ F = 5.93,\ r = 0.652,\ s = 297,\ n = 19$$

$$\{D, D^{v}, {}^{1}\chi, {}^{1}\chi^{v}\} : \quad Q = 0.0023,\ F = 3.25,\ r = 0.694,\ s = 302,\ n = 19\ .$$

To gain a deeper insight into the modeling power of the given χ indexes, it is worth analyzing the variability of these indexes with the solubility. Figure 1 should help us in this task. In this figure, the experimental water solubility values of the given 19 amino acids have been plotted *versus* D, and D^{v} (other χ indexes show the same kind of variability). Striking features of this figure are: i) the similarity of the variability for normal and valence connectivity indexes and ii) the hyperbolic character of this variability ($S \cdot \chi \approx$ cost, dashed lines) a fact that becomes more evident if outliers Pro, Ser and Arg are left out of the plot. While the partial collinearity of the given set of indexes with $< r_{IM}(S;\{\chi\}> = 0.905$ explains their similar behaviour along the

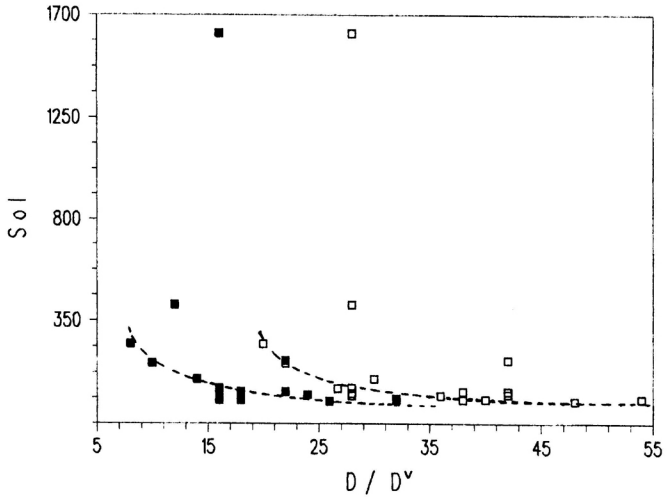


Figure 1. Experimental water solubility values of 19 amino acids *versus* D and D^v connectivty indexes (■: normal χ values; □: valence connectivty values).

solubility dimension, point ii) suggests that introduction of reciprocal connectivity $1/\chi = R$ indexes should improve the modeling of S. The following set of reciprocal connectivty indexes is, then, selected to build the linear combinations (LCRCI) to model the solubility of amino acids

$$\{R\} = \{{}^{D}R, {}^{D}R^{v}, {}^{0}R, {}^{0}R^{v}, {}^{1}R, {}^{1}R^{v}\} .$$

Now, the best successive LCRCI appropriate for modeling S (n = 19) are

$$\{{}^{0}R\} : \quad Q = 0.0011.\ F = 2.75,\ r = 0.373,\ s = 353$$

$$\{{}^{D}R, {}^{0}R\} : \quad Q = 0.0037,\ F = 16.8,\ r = 0.823,\ s = 222$$

$$\{{}^{D}R, {}^{0}R, {}^{0}R^{v}, {}^{1}R^{v}\} : \quad Q = 0.0046,\ F = 12.8,\ r = 0.886,\ s = 194 .$$

The chosen R and C simulating vectors are

$$\boldsymbol{R} = ({}^{D}R, {}^{0}R, {}^{0}R^{v}, {}^{1}R^{v}, R^{0}),\ \boldsymbol{C} = (-36121,\ 18636,\ 22730,\ -7432.8,\ -1830.7) .$$

From the given series of combinations we notice that i) there is a nice statistical improvement from the single- to the two-index combination, ii) the Q-best combination is the last one. Thus, while the poor statistical score

TABLE IV

Orthogonal reciprocal connectivity indexes of n = 19 amino acids (AA)

AA	$^1\Omega \equiv {}^0R$	$^2\Omega$	$^3\Omega$	$^4\Omega$
Pro	0.16714	–0.01759	–0.00121	–0.02074
Ser	0.17059	0.00109	0.017792	–0.01401
Arg	0.104603	0.004459	0.000677	–0.00075
Gly	0.233427	0.003478	0.013459	0.001466
Ala	0.193986	0.003132	–0.01302	0.01859
Thr	0.148544	0.002965	–0.0001	0.004005
Val	0.148544	–0.00596	–0.01045	0.037834
Met	0.137438	0.000979	–0.03604	–0.03385
His	0.120948	–0.00576	0.011339	0.008074
Gln	0.12276	0.00321	0.007582	–0.00858
Ile	0.134427	0.002862	–0.02395	0.007863
Phe	0.111421	–0.00359	0.003762	0.006883
Asn	0.134427	0.002862	0.016117	–0.00208
Leu	0.134427	0.002862	–0.02395	0.013782
Trp	0.092285	–0.00205	0.005827	0.012363
Glu	0.12276	0.00321	0.012133	–0.01361
Lys	0.125266	0.001643	–0.01009	–0.0155
Asp	0.134427	0.002862	0.022209	–0.00755
Tyr	0.101574	–0.00064	0.007926	0.005811

of the single-index combination points to the possible existence of a dominant orthogonal descriptor, the introduction of LCRCI seems to be a good move in the right direction. The only negative point concerns the very high values of the standard deviation of the estimate s. It is now worthy deriving the corresponding orthogonal reciprocal connectivity indexes to detect better dominant orthogonal descriptors and a better LCOCI. It is noticeable that R indexes of the best combination are to some extent interrelated with $< r_{\mathrm{IM}}(S;\{R\}_{best}>$. Table IV presents the values of the orthogonal indexes derived by an orthogonalization procedure performed on the $\{{}^0R, D^R, {}^0R^{\mathrm{v}}, {}^1R^{v}\}$ ordered combination (first, second, third and fourth beest descriptors). The best LCOCI are

$$\{^2\Omega\} : \quad Q = 0.0028.\ F = 19.8,\ r = 0.734,\ s = 258$$

$$\{^1\Omega, {}^2\Omega, {}^4\Omega\} : \quad Q = 0.0047,\ F = 18.1,\ r = 0.885,\ s = 188$$

$$\{^i\Omega, i = 1\text{–}4\} : \quad Q = 0.0046,\ F = 12.8,\ r = 0.886,\ s = 194\ .$$

While the improvement of ${}^2\Omega$ index relative to ${}^0R$ index clearly indicates that this orthogonal index is the dominant component in the modeling of the solubility of amino acids, the second orthogonal combination is the best overall descriptor of the solubility of the amino acdis. The last Ω combination can be used as a validity test for the orthogonalization procedure since it should have the same statistical score of the parent LCRCI, as it really does. In Figure 2, the S values calculated with the aid of the following Ω and $\boldsymbol{C}$ vectors and of Eq. (4) are plotted *versus* the corresponding experimental ones.

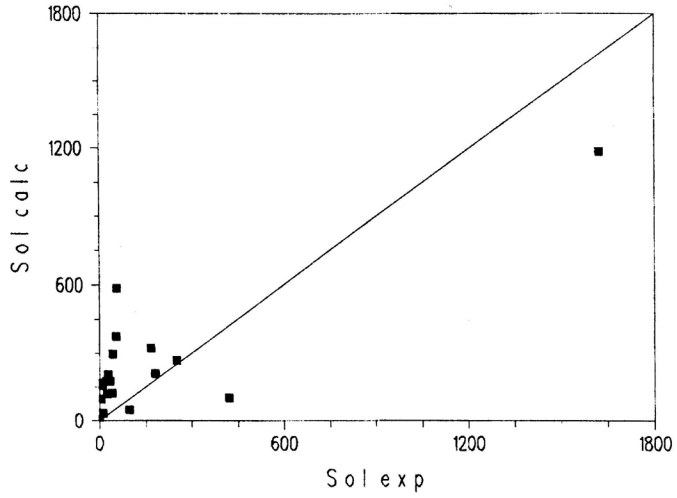


Figure 2. Calculated (using orthogonal LCRCI) *versus* experimental water solubility of n = 19 amino acids.

$$\Omega = ({}^1\Omega, {}^2\Omega, {}^4\Omega, \Omega^0),\ \boldsymbol{C} = (4079.1, -51211, -7459.1, -404.28)$$

Equation (4), where Ω replaces χ, is used here because 6 estimated S values are negative; furthermore, Eq. (4) gives a better estimation of S (Eq. (3): Q = 0.0030, F = 22.2; Eq. (4): Q = 0.0040, F = 39.1).

Figure 2 shows, nevertheless, that the modeling is far from being optimal. An analysis of Figure 1 shows that the solubility values of Pro, Ser and Arg (highly soluble compounds, especially Pro) are strong outliers, and thus, interesting hints could be obtained about the modeling power of a LCCI or a LCRCI if they are excluded from the modeling. The following are the best LCCI and LCRCI for a description of the restricted solubility n = 16 set

$$\{{}^0\chi\}: \quad Q = 0.0202,\ F = 27.8,\ r = 0.816,\ s = 40.3$$

$$\{{}^0\chi,\ {}^1\chi\}: \quad Q = 0.0371,\ F = 46.8,\ r = 0.937,\ s = 25.24$$

$$\{{}^0\chi,\ {}^1\chi,\ {}^1\chi^v\}: \quad Q = 0.0374,\ F = 31.7,\ r = 0.942,\ s = 25.18$$

$$\{{}^0R\}: \quad Q = 0.038,\ F = 97.4,\ r = 0.935,\ s = 24.7$$

$$\{{}^0R,\ {}^1R\}: \quad Q = 0.044,\ F = 66.2,\ r = 0.954,\ s = 21.7$$

$$\{{}^DR,\ {}^DR^v,\ {}^0R,\ {}^0R^v\}: \quad Q = 0.053,\ F = 47.9,\ r = 0.972,\ s = 18.3$$

$$\{R\}: \quad Q = 0.058,\ F = 37.6,\ r = 0.981,\ s = 17.0\ .$$

The differences between the two sets of linear combinations are evident: i) LCRCI scores better than LCCI, ii) the single-index LCRCI shows a good quality, iii) Q-score of LCRCI improves with the introduction of the next R index while LCCI deteriorates after the three-index combination and ivi) the two-R-index LCRCI is better than the Q-best LCCI. To derive the calculated S (n = 16) values plotted *versus* the corresponding experimental values in Figure 3, the following $\boldsymbol{R}$ and $\boldsymbol{C}$ vectors together with Eq. (4) where $\boldsymbol{R}$ replaces χ) have been used

$$\boldsymbol{R} = ({}^DR,\ {}^DR^v,\ {}^0R,\ {}^0R^v,\ R^0),\ \boldsymbol{C} = (8113.1,\ -26638,\ 11696,\ -5227.3,\ -199.08)\ .$$

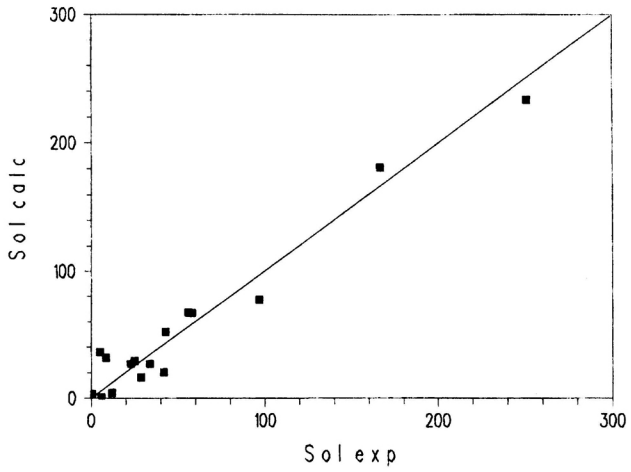


Figure 3. Calculated (using LCRCI) *versus* experimental water solubility of n = 16 amino acids.

As two S values estimated using Eq. (3) are negative, Eq. (4) is to be preferred; furthermore, its modeling power is more convincing

$$\text{Eq. (3): } Q = 0.060,\ F = 243.7,\ \text{Eq. (4): } Q = 0.062,\ F = 262.7\ .$$

The 4-R-index combination has been chosen for the modeling as, even if it does not show the betst Q value, it nevertheless shows a better F value than the following two combinations. It is noticeable that this description of S (n = 16) is nearly as good as the S (n = 13) description wiht a LCFCI.

Let us retrieve our fragment connectivity indexes and examine their descriptive power. The best Q-LCFCI for the solubility of n = 16 amino acids is rather poor

$$\{D, D^v, {}^0\chi^v, {}^1\chi\}_f: \quad Q = 0.0073,\ F = 0.895,\ r = 0.496,\ s = 68.3\ .$$

To check if reciprocal fragment connectivity indexes are good descriptors of the solubility of amino acids let us analyze the variability of χ_f indexes with the solubility, as previously done for χ indexes. Figure 4 shows the variability for D_f and D_f^v indexes (other indexes behave similarly). Leaving out the last two points, the variability seems totally random, a fact that would exclude the use of reciprocal fragment connectivity indexes for achieving better

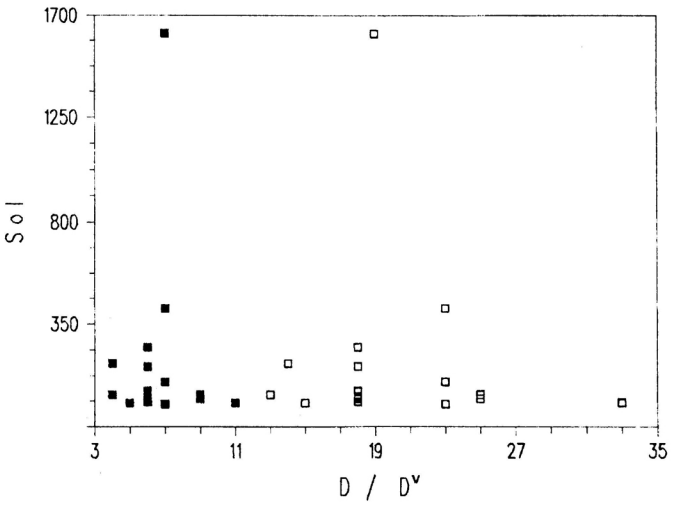


Figure 4. Experimental water solubility values od n = 19 amino acids *versus* fragment D and D^v indexes (■: normal fragment χ_f values; □: fragment connectivity valence values).

modeling. Anyway, the statistical score of the best Q-LCRFCI (linear combination of reciprocal fragment connectivity indexes) for $n = 19$ is

$$\{{}^{D}R^{v}, {}^{0}R^{v}, {}^{1}R, {}^{1}R^{v}\}_{f}: \quad Q = 0.0013,\ F = 1.0,\ r = 0.474,\ s = 369,\ n = 19\ .$$

The only fairly acceptable modeling of a LCRFCI is with $n = 14$ low points, that is, excluding the most soluble amino acids, Pro, Ser, Gly, Arg and Ala,

$$\{{}^{D}R, {}^{0}R, {}^{0}R^{v}, {}^{1}R^{v}\}_{f}: \quad Q = 0.060,\ F = 8.3,\ r = 0.887,\ s = 14.8,\ n = 14\ .$$

The KOKOS indexes, instead, release a more satisfactory description of this $n = 16$ solubility values, even if not as good as the LCRCI description,

$$\{{}^{1}\Gamma, {}^{3}\Gamma, {}^{4}\Gamma, {}^{5}\Gamma\}: \quad Q = 0.029,\ F = 14.6,\ r = 0.917,\ s = 31.3\ .$$

To model also the high solubility values of Pro, Ser and Arg, suprareciprocal connectivity indexes, $a \cdot R$ (where $a > 1$ is an association parameter) are introduced, as recently done for caffeine homologues, where supraconnectivity $a \cdot \chi$ indexes have been successfully used.[6] The reason for doing this is here one inferential and due to the abnormally high solubility values of Pro, Ser and Arg. Even if there is no experimental evidence that Pro, Ser and Arg undergo intermolecular association in aqueous solutions, their high solubility (especially Pro) can be better grasped conceiving either self-association or strong solvation phenomena, which can be simulated by introduction of supramolecular connectivity indexes[6,23] that take the solvated or self-associated molecules as supramolecular species. Using LCRCI with supra-$(a \cdot R)$-indexes for Pro ($a = 4$), Ser ($a = 1.5$) and Arg ($a = 2$) it is possible to derive the following best single-index, Q-best and F-best combinations for $n = 19$

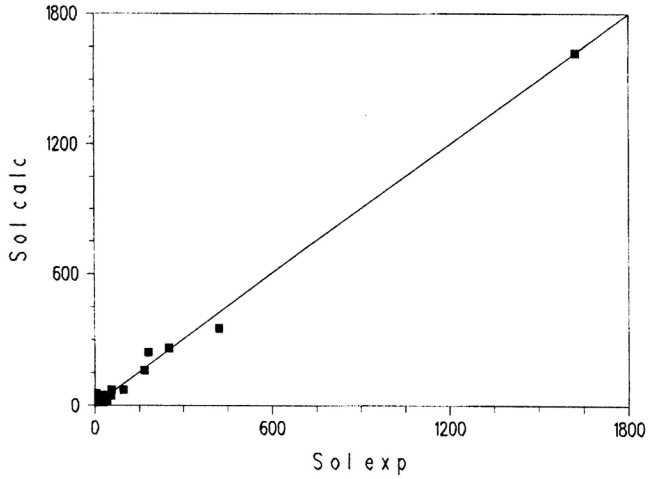
$$\{{}^{0}R\}: \quad Q = 0.022,\ F = 1141,\ r = 0.9926,\ s = 46.1$$

$$\{{}^{0}R, {}^{1}R\}: \quad Q = 0.029,\ F = 1028,\ r = 0.9961,\ s = 34.4$$

$$\{{}^{D}R^{v}, {}^{0}R, {}^{1}R^{v}\}: \quad Q = 0.030,\ F = 732,\ r = 0.9966,\ s = 33.3\ .$$

The description of the solubility of amino acids is now optimal and, while the Q and s values are not as good as in the former description ($n = 16$), the correesponding F and r values are much better. As two solubility values are negative, Eq. (4) is used here ($Q = 0.034$, $F = 2795$, Eq. (3): $Q = 0.032$, $F = 2488$) to model, together with the following $\boldsymbol{R}$ and $\boldsymbol{C}$ vectors, the solubility of amino acids (see Figure 5)

$$\boldsymbol{R} = ({}^{D}R^{v}, {}^{0}R, {}^{1}R^{v}, R^{0}),\ \boldsymbol{C} = (-8018.9,\ 5271.0,\ -316.16,\ -299.06)\ .$$



Figurte 5. Calculated (with LCRCI plus supraconnectivity R values for Pro, Ser and Arg *versus* experimental water solubility values of n = 19 natural amino acids.

The use of supra-($a \cdot \Gamma$)-indexes for Pro, Ser and Arg gives the following Q-best LCΓI, which has no better modeling ability than the previous LCRCI with supraindexes

$$\{{}^1\Gamma, {}^5\Gamma, {}^6\Gamma\} : \quad Q = 0.0139,\ F = 158,\ r = 0.985,\ s = 70.9\ .$$

Aa expected, reciprocal 1 / Γ indexes, with and without suprareciprocal indexes, provide less satisfactory modeling since Γs are '*ad hoc*' indexes directly derived from the physical properties they describe.

Figure 1 could raise the suspicion that the variability of connectivity indexes is exponential ($S \cdot e^{\chi} \approx$ cost). Let us, then, find out the best linear combination of connectivity indexes (LCCI) apt to model ln(1/S) instead of S

$$\{D^v, {}^0\chi, {}^1\chi\} : \quad Q = 0.48,\ F = 4.5,\ r = 0.69,\ s = 1.45,\ n = 19\ .$$

Clearly, a comparison with previous results should be based on F and r parameters, since the good scores of s and, consequently, of Q are due to the use of the logarithmic form of solubility.

CONCLUSION

The value of the proposed LCCI-MM strategy for modeling the solubility of 19 amino acids goes beyond the specific property analyzed since it delineates a general method for modeling every kind of molecular property with a self-consistent set of graph theoretical indexes, defined within the frame

of the RKH molecular connectivity theory. The central moment of this strategy is the analysis of the variability of the used indexes along the solubility dimension. This variability provides precious information on the derivation of new kinds of indexes from the original ones. The second important aspect can be phrased in the following way: *outliers are useful*, since they can help to design indexes that are able to describe an entire set of values of a given property. The third not unimportant aspect has to do with the modeling equation: form Eq. (4) of the modeling equation normally achieves a better description than Eq. (3) since it allows handling positive calculated values and excludes negative values that have no physical meaning.

Modeling of the water solubility of the $n = 19$ amino acids confirms a recent general observation that claims that normal connectivity χ indexes are generally good descriptors of low solubility and gas phase properties,[6] where association phenomena are negligible. This modeling confirms also, through the analysis of the variability of χ indexes and through the analysis of outliers, that the supra-χ, $a \cdot \chi$ or $a \cdot f(\chi)$ (with $a > 1$) types of indexes contribute to improvement of the quality of molecular modeling. Linear combinations of reciprocal connectivity indexes LCRCI (where, $f(\chi) = 1 / \chi$) are satisfactory descriptors of the full solubility set of values and optiomal descriptors of the restricted $n = 16$ solubility space. Introduction of suprareciprocal connectivity indexes for Pro, Ser and Arg as well as employment of LCRCI that include these kinds of indexes produce an exceptional description of the $n = 19$ solubility values. Thus, the aim of descrbing the solubility of amino acids with graph theoretical indexes has not only been achieved but has ended with the finding of new $f(\chi)$ connectivity descriptors that are even better than the γ indexes. The suprareciprocal indexes for Pro, Ser and ASrg are, clearly, not experimentally grounded and the value of the association α constant is just inferential and based on rational arguments, but, even if it is very important that these supraindexes should be experimentally grounded, the inferential power of the method should not be underscored or obliterated.

Acknowledgements. – This paper is dedicated to the 20th anniversary of Milan Randić's original paper, *On Characterization of Molecular Branching*, published in *J. Amer. Chem. Soc.* **97** (1975) 6609.

REFERENCES

1. L. Pogliani, *J. Phys. Chem.* **97** (1993) 6731 (and references therein).
2. L. Pogliani, *J. Phys. Chem.* **98** (1984) 1494.
3. L. Pogliani, *Amino Acids* **6** (1994) 141.
4. L. Pogliani, *J. Chem. Inf. Comput. Sci.* **34** (1994) 801.
5. L. Pogliani, *Curr. Top. Pept. & Prot. Res.* **1** (1994) 119.
6. L. Pogliani, *J. Phys. Chem.* **99** (1995) 925.

7. L. Pogliani, unpublished and to to be published results.
8. M. Randić, *J. Amer. Chem. Soc.* **97** (1975) 6609.
9. L. B. Kier and L. H. Hall, *Molecular Connectivity in Structure-Activity Analysis*, Wiley, New York, 1986 and references therein.
10. N. Trinajstić, *Chemical Graph Theory*, vol. 2, CRC Press, Boca Raton, FL (1983).
11. M. Randić, *J. Math. Chem.* **7** (1991) 155.
12. D. H. Rouvray, *J. Mol. Struct. (Theochem)* **185** (1989) 187.
13. P. J. Hansen and P. C. Jurs, *J. Chem. Ed.* **65** (1988) 574.
14. Z. Mihalić, S. Nikolić, and N. Trinajstić, *J. Chem. Inf. Comput. Sci.* **32** (1992) 28.
15. S. C. Basak, G. H. Niemi, and G. D. Veith, *J. Math. Chem.* **7** (1991) 243.
16. D. E. Needham, I. C. Wei, and P. G. Seybold, *J. Amer. Chem. Soc.* **110** (1988) 4186.
17. a) L. B. Kier, L. H. Hall, and J. W. Frazer, *J. Chem. Inf. Comput. Sci.* **33** (1993) 143 and b) L. H. Hall, L. B. Kier, and W. J. Frazer, *J. Chem. Inf. Comput. Sci.* **33** (1993) 148.
18. CRC *Hanbook of Chemistry and Physics* 72nd Ed. 1991–1992: Boca Raton, FL.
19. M. Randić, *N. J. Chem.* **15** (1991) 517.
20. M. Randić, *J. Chem. Inf. Comput. Sci.* **31** (1991) 311.
21. M. Randić, *J. Mol. Struct. (Theochem)* **233** (1991) 45.
22. M. Randić, *Croat. Chem. Acta* **64** (1991) 43.
23. L. Pogliani, *Comput. Chem.* **17** (1993) 283.
24. A. Kidera, Y. Konisci, T. Ooi, and H. A. Scheraga, *J. Prot. Chem.* **4** (1985) 23.

SAŽETAK

Pristup molekularnom modeliranju s pomoću linearne kombinacije indeksa povezanosti

Lionello Pogliani

Prikazan je pristup modeliranju fizikalno-kemijskih svojstava s pomoću linearne kombinacije indeksa povezanosti. Upotrijebljeno fizikalno svojstvo bila je topljivost 19 prirodnih aminokiselina. To je svojstvo modelirano s pomoću linearne kombinacije indeksa povezanosti (LKIP) i s pomoću linearne kombinacije specijalno konstruiranih indeksa povezanosti. Model utemeljen na linearnoj kombinaciji recipročnih indeksa povezanosti (LKRIP) pokazao se najboljim u predviđanju topljivosti 16 aminokiselina. Taj je model znatno poboljšan kada su uz LKRIP upotrijebljeni i indeksi superpovezanosti za aminokiseline prolin, serin i arginin. Također je proučavana upotrebljivost linearne kombinacije indeksa povezanosti fragmenata. Upotrijebljena je i linearna kombinacija ortogonalnih indeksa povezanosti, koja je poboljšala model i pomogla da se odrede dominantni deskriptori za topljivost. *Ad hoc* indeksi nisu proizveli modele koji su usporedljivi s navedenima.