Amino Acids (1994) 6: 141–153

Amino
Acids

Structure property relationships of amino acids and some dipeptides

L. Pogliani

Dipartimento di Chimica, Università della Calabria, Rende, Italy

Accepted June 22, 1993

Summary. A molecular connectivity model of the crystal densities and specific rotations of some natural amino acids and of the longitudinal relaxation rates of some natural amino acids and cyclic dipeptides is presented. While crystal densities and relaxation rates are better described by a set of three valence molecular connectivity indices $\{D^v, {}^0X^v, {}^1X^v\}$, specific rotations are better described by a set of two simple molecular connectivity indices $\{{}^1X, {}^0X\}$. Relaxation rates are, also, well described by the simple molecular connectivity $\{D, {}^1X\}$ index set. Use of orthogonal indices, derived from the corresponding ordinary indices shows, in the case of specific rotations, the possibility to condense the information by the aid of a single high quality descriptor underlining, thus, the versatility of these indices and also their dependence on the orthogonalisation process.

Keywords: Amino acids – QSPR – Molecular connectivity – Properties

Background

The structure-property relationships quantify the connection between the structure and properties of molecules, as it is well-known in chemistry that structural characteristics of a molecule are responsible for its properties (Trinajstić, 1983; Kier and Hall, 1986; Turro, 1986; Hansen and Jurs, 1988; Rouvray, 1989; Mihalić and Trinajstić, 1992). These relationships and the corresponding structure-activity relationships are mathematical models that allow the prediction of properties and activities from structural parameters. Structure-property-activity studies start by representing compounds by suitable molecular descriptors, which can be chosen among physicochemical properties, quantum mechanically computed parameters or graph theoretically derived topological indices, which indicate the characterization of a molecule by a single number, like most molecular properties also recorded as single numbers. We refer to these indices as structure explicit parameters because they have a direct, often simple, structural

interpretation and they are calculated directly from the chemical graphs of a given set of compounds. These structure parameters are then correlated to the observed physicochemical properties by using single, multiple linear or power type regression analysis to obtain predictive equations which are termed quantitative structure-property or -activity relationships (QSPR or QSAR). The most widely used example of nonempirical structure-explicit topological indices in QSPR/QSAR is the set of simple and valence molecular connectivity indices developed by Randić (1975), Kier et al., (1975) and Kier and Hall (1976, 1981). These structure descriptors have been quite successful in analyzing molecular properties in quantitative structure-property-activity relationships. The simple and valence molecular connectivity X and X^v indices are rooted on the two cardinal δ and valence δ^v numbers respectively. These atom-level structure delta indices are generated from hydrogen-suppressed graphs as, in chemical graph theory, more commonly, hydrogen atoms are ignored. For a skeletal atom, δ^v is the count of valence electrons ($Z^v - h$) and δ is the count of sigma electrons ($\sigma - h$), where Z^v is the number of valence electrons, σ is the count of electrons in σ orbitals and h is the number of bonded hydrogen atoms. Cardinal number δ can also be seen as the count of nearest neighbors in the skeleton (for sp^3 skeletal carbons $\delta = \delta^v$) and thus as a measure of the degree of mantle-atom status. Skeletal atoms with $\delta = 3$ or 4 are relatively buried within the molecule whereas terminal atoms, for which $\delta = 1$, lie on the surface of the molecule. Randić in 1991 introduced, in multivariate regression analysis, the orthogonal ${}^i\Omega$ connectivity indices to bypass collinearity among different indices and to identify the role of individual descriptors, disentangling, thus, contributions of indices that duplicate information.

Introduction

The graph-theoretical approach to QSPR of α-amino acids and cyclic dipeptides used in this study is based on the use of a set of molecular connectivity X and X^v indices for encoding the structural information. We will, here, illustrate a multivariate analysis of the following physicochemical P properties of natural amino acids: crystal CD densities, specific SR rotations, and longitudinal *nmr* relaxation $R_1 = 1/T_1$ rates of the Cα of amino acids and cyclic dipeptides which are an important dynamical and structural parameter of these molecules in solution (Pogliani, 1993b). A subset of up to three simple or valence molecular connectivity indices, chosen in the following $X = \{D, D^v, {}^0X, {}^0X^v, {}^1X, {}^1X^v\}$ set, will be used as basis descriptors for the present structure-property study. Corresponding orthogonalized connectivity indices $\{({}^i\Omega)$ with $i = 1\text{–}3\}$ will also be considered to exemplify their use and meaning. The two molecular D and D^v indices of the X set have, recently, been added by Pogliani (1992a, 1993a) to the set of the known molecular connectivity indices to encode the pH at the isoelectric point of amino acids. Pogliani (1992b, 1993a,c) also showed that the two cardinal numbers δ and δ^v correlate fairly well with the atomic charges in amino acids.

Method

The molecular connectivity indices used in this study are:

— the sum delta index (the summations run over the non-hydrogen atoms of the molecular skeleton)

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

— the simple connectivity index of the zeroth-order

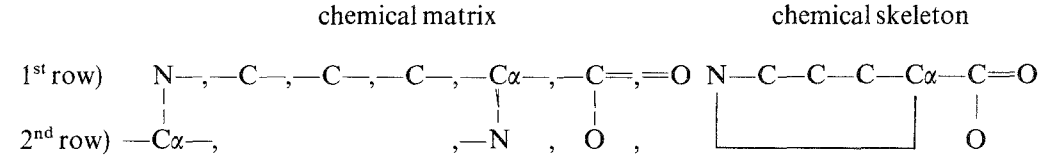
$${}^0X = \sum(\delta_i)^{-1/2} \tag{2}$$

— the simple connectivity index of the first-order (i and j denote adjacent atoms forming a bond)

$${}^1X = \sum(\delta_i\delta_j)^{-1/2} \tag{3}$$

and the corresponding valence molecular connectivity indices D^v, ${}^0X^v$ and ${}^1X^v$, where δ^v substitutes δ in expressions 1–3. Given indices are atom or bond additive properties and while for the calculation of ${}^0X$ and ${}^1X$, terminal atoms or bonds play a larger role than a atom or a bond buried inside a molecule for which ${}^0X$ and ${}^1X$ are smaller the contrary is valid for D. For the valence indices, atoms with smaller valence play a larger role in the description of ${}^0X^v$ and ${}^1X^v$ and a minor role in the description of D^v. Multivariate regressions of P in terms of orthogonalised molecular connectivity indices, ${}^i\Omega$ have been elaborated as indicated by Randić (1991a and b): by deriving stepwise regression equations for ordinary $\{X_i\}$ and then using the diagonal coefficients as those of the sought-after regression equation, the constant term of the equations being given by the first single-X linear regression equation. Numerical values of the orthogonal indices can be derived by the aid of an orthogonalization procedure of the corresponding ordinary indices as outlined by Randić (1991a and b), the first step of which consist by selecting an ordinary index as the first orthogonal index.

Neutral forms of amino acids have been represented by $(2 \cdot n)$ δ and δ^v matrices. These delta matrices, which have been collected in Table 1, constitute the input data to derive the different molecular connectivity indices, which have been collected in Table 2. While the number of columns n of these matrices represent the number of skeletal atoms in the main backbone, zero symbolizes an empty space and c, in cyclic amino acids, symbolizes connected atoms due to ring closure. Connections take place horizontally, only between first-row elements, and vertically, between first-row and second-row elements. Following example explains the passage from the delta matrices of Pro given in Table 1 to the hydrogen suppressed chemical formula through the chemical matrix, where delta values have been substituted by the corresponding connected atoms,



Connectivity values for the cyclic dipeptides $c(AA_1\text{–}AA_2)$ (or diketopiperazines), which have been collected in Table 3, are the sum of the corresponding indices of AA_1 and AA_2 delta corrected delta matrices. Corrected delta matrices of amino acids in a peptidic moiety are obtained in the following way: the head of the amino acid delta matrices (the $2 \cdot 3$ portion at the right side of the matrix) is substituted by the following head

$$\begin{matrix}\delta(\text{head}) = 3,3,1 \\ \quad\quad\quad 2,p,0\end{matrix} \qquad \begin{matrix}\delta^v(\text{head}) = 3,4,6 \\ \quad\quad\quad 4,p,0\end{matrix}$$

where p means connection, due to the peptidic bond, between the delta value at its top and the delta value at its left.

Table 1. The $(2 \cdot n)$ δ and δ^v matrices of amino acids

AA	δ matrix	δ^v matrix
Gly	$\begin{bmatrix} 2, 3, 1 \\ 1, 1, 0 \end{bmatrix}$	$\begin{bmatrix} 2, 4, 6 \\ 3, 5, 0 \end{bmatrix}$
Ala	$\begin{bmatrix} 1, 3, 3, 1 \\ 0, 1, 1, 0 \end{bmatrix}$	$\begin{bmatrix} 1, 3, 4, 6 \\ 0, 3, 5, 0 \end{bmatrix}$
Ser	$\begin{bmatrix} 1, 2, 3, 3, 1 \\ 0, 0, 1, 1, 0 \end{bmatrix}$	$\begin{bmatrix} 5, 2, 3, 4, 6 \\ 0, 0, 3, 5, 0 \end{bmatrix}$
Pro	$\begin{bmatrix} 2, 2, 2, 2, 3, 3, 1 \\ c, 0, 0, 0, c, 1, 0 \end{bmatrix}$	$\begin{bmatrix} 4, 2, 2, 2, 3, 4, 6 \\ c, 0, 0, 0, c, 5, 0 \end{bmatrix}$
Val	$\begin{bmatrix} 1, 3, 3, 3, 1 \\ 0, 1, 1, 1, 0 \end{bmatrix}$	$\begin{bmatrix} 1, 3, 3, 4, 6 \\ 0, 1, 3, 5, 0 \end{bmatrix}$
Thr	$\begin{bmatrix} 1, 3, 3, 3, 1 \\ 0, 1, 1, 1, 0 \end{bmatrix}$	$\begin{bmatrix} 5, 3, 3, 4, 6 \\ 0, 1, 3, 5, 0 \end{bmatrix}$
Leu	$\begin{bmatrix} 1, 3, 2, 3, 3, 1 \\ 0, 1, 0, 1, 1, 0 \end{bmatrix}$	$\begin{bmatrix} 1, 3, 2, 3, 4, 6 \\ 0, 1, 0, 3, 5, 0 \end{bmatrix}$
Ile	$\begin{bmatrix} 1, 2, 3, 3, 3, 1 \\ 0, 0, 1, 1, 1, 0 \end{bmatrix}$	$\begin{bmatrix} 1, 2, 3, 3, 4, 6 \\ 0, 0, 1, 3, 5, 0 \end{bmatrix}$
Asp	$\begin{bmatrix} 1, 3, 2, 3, 3, 1 \\ 0, 1, 0, 1, 1, 0 \end{bmatrix}$	$\begin{bmatrix} 6, 4, 2, 3, 4, 6 \\ 0, 5, 0, 3, 5, 0 \end{bmatrix}$
Lys	$\begin{bmatrix} 1, 2, 2, 2, 2, 3, 3, 1 \\ 0, 0, 0, 0, 0, 1, 1, 0 \end{bmatrix}$	$\begin{bmatrix} 3, 2, 2, 2, 2, 3, 4, 6 \\ 0, 0, 0, 0, 0, 3, 5, 0 \end{bmatrix}$
Glu	$\begin{bmatrix} 1, 3, 2, 2, 3, 3, 1 \\ 0, 1, 0, 0, 1, 1, 0 \end{bmatrix}$	$\begin{bmatrix} 6, 4, 2, 2, 3, 4, 6 \\ 0, 5, 0, 0, 3, 5, 0 \end{bmatrix}$
Met	$\begin{bmatrix} 1, 2, 2, 2, 3, 3, 1 \\ 0, 0, 0, 0, 1, 1, 0 \end{bmatrix}$	$\begin{bmatrix} 1, .67, 2, 2, 3, 4, 6 \\ 0, \quad 0, 0, 0, 3, 5, 0 \end{bmatrix}$
His	$\begin{bmatrix} 2, 2, 2, 2, 3, 2, 3, 3, 1 \\ c, 0, 0, 0, c, 0, 1, 1, 0 \end{bmatrix}$	$\begin{bmatrix} 3, 5, 3, 4, 4, 2, 3, 4, 6 \\ c, 0, 0, 0, c, 0, 3, 5, 0 \end{bmatrix}$
Phe	$\begin{bmatrix} 2, 2, 2, 2, 2, 3, 2, 3, 3, 1 \\ c, 0, 0, 0, 0, c, 0, 1, 1, 0 \end{bmatrix}$	$\begin{bmatrix} 3, 3, 3, 3, 3, 4, 2, 3, 4, 6 \\ c, 0, 0, 0, 0, c, 0, 3, 5, 0 \end{bmatrix}$
Arg	$\begin{bmatrix} 1, 3, 2, 2, 2, 2, 3, 3, 1 \\ 0, 1, 0, 0, 0, 0, 1, 1, 0 \end{bmatrix}$	$\begin{bmatrix} 3, 4, 4, 2, 2, 2, 3, 4, 6 \\ 0, 4, 0, 0, 0, 0, 3, 5, 0 \end{bmatrix}$
Tyr	$\begin{bmatrix} 2, 2, 3, 2, 2, 3, 2, 3, 3, 1 \\ c, 0, 1, 0, 0, c, 0, 1, 1, 0 \end{bmatrix}$	$\begin{bmatrix} 3, 3, 4, 3, 3, 4, 2, 3, 4, 6 \\ c, 0, 5, 0, 0, c, 0, 3, 5, 0 \end{bmatrix}$
Trp	$\begin{bmatrix} 3, 2, 2, 2, 2, 3, 2, 2, 3, 2, 3, 3, 1 \\ c, 0, 0, 0, 0, c, 0, 0, c, 0, 1, 1, 0 \end{bmatrix}$	$\begin{bmatrix} 4, 3, 3, 3, 3, 4, 4, 3, 4, 2, 3, 4, 6 \\ c, 0, 0, 0, 0, c, 0, 0, c, 0, 3, 5, 0 \end{bmatrix}$

Table 2. D, D^v, ${}^0X$, ${}^0X^v$, ${}^1X$, ${}^1X^v$, values of amino acids

AA	D	D^v	${}^0X$	${}^0X^v$	${}^1X$	${}^1X^v$
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
Leu	16	28	7.439	5.795	4.036	3.021
Ile	16	28	7.439	5.795	4.091	3.076
Asp	16	38	7.439	4.572	4.037	2.239
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

Table 3. T_1(Cα) (in s), ${}^0X$, ${}^1X$, D, ${}^0X^v$, ${}^1X^v$ and D^v in cyclic dipeptides

$c(AA_1-AA_2)$	T_1	${}^0X$	${}^1X$	D	${}^0X^v$	${}^1X^v$	D^v
c(Leu-Gly)	0.9	9.138	5.592	24	7.385	4.613	40
c(Phe-Gly)	0.64	10.673	7.254	32	8.195	5.314	54
c(Phe-Val)	0.51	13.121	8.575	38	10.642	6.224	60
c(Trp-Gly)	0.38	12.535	8.737	40	9.695	6.308	66
c(Leu-Trp)	0.31	15.690	10.541	48	12.850	7.701	74
c(Trp-Trp)	0.18	19.087	13.686	64	15.159	9.396	100

Results and discussion

Specific $SR = [\alpha]$ rotations and crystal CD densities of natural amino acids, taken from the CRC Handbook of Chemistry and Physics, and the relaxation T_1 times (from which the relaxation rates are derived) of natural amino acids taken from Pogliani (1993b) have been collected in Table 4. The relaxation times of cyclic dipeptides (diketopiperazines) have been collected in Table 3 together with their connectivity values. In Tables 5–7 the regression equations and the statistical parameters for the given physicochemical properties of natural amino acids based on the molecular connectivity indices and on the corresponding orthogonalised indices have been collected. Randić (1991a,b), recently, demonstrated the critical importance of the statistical parameter S (standard error of estimate), near the regression R coefficient, in determining the quality of a set of basis descriptors; to take account of these two factors we introduce the internal quality $Q = R/S$ factor, which encodes information on both R and S statistical parameters (next to last row in Tables 5–7) and which gives an idea of the quality

Table 4. Specific rotations [α] (in angular degrees) using sodium light (5893 Å), crystal *CD* densities and longitudinal relaxation T_1(Cα) (in s.) of amino acids

AA	[α]	*CD*	T_1(Cα)
Gly		1.601	6.0
Ala	2.7	1.401	3.1
Ser	−6.83	1.537	3.5
Pro	−85.0		4.3
Val	6.42	1.230	
Thr	28.4		2.2
Leu	−10.8	1.165	1.9
Ile	11.29		2.2
Met	−8.11	1.340	
Asp	4.7	1.660	
Lys	14.6		1.4
Glu	11.5	1.538	
His	−39.01		
Phe	−35.14		
Arg	12.5	1.100	
Tyr		1.456	
Trp	−31.5		

Table 5. The regression equations and the statistical parameters for crystal densities *CD* of amino acids based on the molecular conectivity indices and on the corresponding orthogonalised molecular connectivity indices

Regression equations	Constant
$CD = -0.0756\ {}^0X^v$	1.7836
$CD = -0.1540\ {}^0X^v + 0.0167\ D^v$	1.6428
$CD = -0.4918\ {}^0X^v + 0.4481\ {}^1X^v + 0.0259\ D^v$	1.8566
$CD = -0.0756\ {}^1\Omega_a$	1.7836
$CD = -0.0756\ {}^1\Omega_a + 0.0167\ {}^2\Omega_a$	1.7836
$CD = -0.0756\ {}^1\Omega_a + 0.0167\ {}^2\Omega_a + 0.4481\ {}^3\Omega_a$	1.7836

normal descriptors:	${}^0X^v$	${}^0X^v, D^v$	${}^0X^v, D^v, {}^1X^v$
orthogonal descrs.:	${}^1\Omega_a$	${}^1\Omega_a, {}^2\Omega_a$	${}^1\Omega_a, {}^2\Omega_a, {}^3\Omega_a$
Coeff. of regr. *R*:	0.570	0.799	0.934
Standard error *S*:	0.166	0.130	0.083
$Q = R/S$:	3.433	6.146	11.25
number of observations $n = 10$			

of the set of connectivity indices used: the higher the *Q* value the better the combination of connectivity indices used to describe the investigated property. The molecular connectivity indices belonging to one of the two $\{D, {}^0X, {}^1X\}$ and $\{D^v, {}^0X^v, {}^1X^v\}$ subsets were chosen starting with the index, which produces the best description of the *P* property, finding, then, the best two-index description

Table 6. The regression equations and the statistical parameters for relaxation rates R_1 ($=1/T_1$ s^{-1}) of amino acids and diketopiperazines based and on the molecular connectivity indices and on the corresponding orthogonalised molecular connectivity indices

Regression equations	Constant
$R_1 = 0.0638\ D^v$	−1.4078
$R_1 = 0.0833\ D^v - 0.1909\ {}^1X^v$	−1.4610
$R_1 = 0.0823\ D^v + 0.2249\ {}^0X^v - 0.5174\ {}^1X^v$	−1.6614
$R_1 = 0.0638\ {}^1\Omega^b$	−1.4078
$R_1 = 0.0638\ {}^1\Omega_b - 0.1909\ {}^2\Omega_b$	−1.4078
$R_1 = 0.0638\ {}^1\Omega_b - 0.1909\ {}^2\Omega_b + 0.2249\ {}^3\Omega_b$	−1.4078

normal descriptors:	D^v	D^v, ${}^1X^v$	D^v, ${}^1X^v$, ${}^0X^v$
orthogonal descrs.:	${}^1\Omega_b$	${}^1\Omega_b$, ${}^2\Omega_b$	${}^1\Omega_b$, ${}^2\Omega_b$, ${}^3\Omega_b$
Coeff. of regr. R:	0.982	0.984	0.986
Standard error S:	0.300	0.295	0.292
$Q = R/S$:	3.278	3.338	3.374
number of observations $n = 14$			

Table 7. The regression equations and the statistical parameters for specific rotations SR of amino acids based on the molecular connectivity indices and on the corresponding orthogonalised molecular connectivity indices

Regression equations	Constant
$SR = -1.7479\ D$	22.943
$SR = -90.969\ {}^1X + 67.389\ {}^0X$	−119.921
$SR = -2.5392\ D - 75.931\ {}^1X + 64.797\ {}^0X$	−120.964
$SR = -1.7479\ {}^1\Omega_c$	22.943
$SR = -6.5875\ {}^1\Omega' + 67.389\ {}^2\Omega'$ (${}^1\Omega' \equiv {}^1X$)	20.661
$SR = -0.8918\ {}^1\Omega'' - 90.969\ {}^2\Omega''$ (${}^1\Omega'' \equiv {}^0X$)	−1.5182
$SR = -1.7479\ {}^1\Omega_c + 43.475\ {}^2\Omega_c - 75.931\ {}^3\Omega_c$	22.943

normal descriptors:	D	${}^1X$, ${}^0X$	D, ${}^0X$, ${}^1X$
orthogonal descriptors:	${}^1\Omega_c$	${}^1\Omega'$, ${}^2\Omega'$; ${}^1\Omega''$, ${}^2\Omega''$	${}^1\Omega_c$, ${}^2\Omega_c$, ${}^3\Omega_c$
Coeff. of regression R:	0.325	0.882	0.885
Standard error S:	28.35	14.684	15.166
$Q = R/S$:	0.014	0.060	0.058
number of observations $n = 15$			

of P, to end up with the three-index description of P (ordering of descriptors reflects their quality as single descriptors). The ordering of the orthogonal indices reflects, instead, the gradual inclusion of new ordinary indices which can be traced back by the aid of their mutual coefficients. The gradual inclusion of the next index, normally, increases Q by increasing R while simultaneously decreasing S.

Let us analyze found regressions of Tables 5–7, starting with the properties mapped by the ordinary molecular connectivity indices. The best descriptors for

CD (see Table 5) are the valence molecular connectivity indices. Best single descriptor being ${}^0X^v$, which gives a rather poor description of *CD*. The best two-index regression shows an interesting improvement in quality, while further inclusion of ${}^1X^v$, which is the second best single descriptor, produces a further conspicuous improvement in statistical parameters *R*, *S* and *Q*. The growing of the *Q* value, which nearly doubles at every inclusion of a new descriptor, indicates that each descriptor is statistically significant. The fact that *CD* is essentially described by valence molecular connectivity indices tell us that electronic distribution in sp^2 carbons, —OH, =O and in amino or imino groups play an important role in determining the solid state structure of amino acids.

Valence molecular connectivity indices are also the best descriptors for the R_1 relaxation rates of amino acids and diketopiperazines (see Table 6). The best single descriptor for this property is D^v, which alone produces an impressive regression, while inclusion of ${}^1X^v$ and specially of ${}^0X^v$ (the second best single descriptor, but ${}^1X^v$ and ${}^0X^v$ are very similar) improves the regression in a negligible way. Non-valence connectivity indices are also good descriptors of R_1 as indicated by the following *R*, *S* and *Q* values of the following sets

$$\{D\}\text{: } R = 0.977,\ S = 0.339 \text{ and } Q = 2.766$$

$$\{D, {}^1X\}\text{: } R = 0.979,\ S = 0.341 \text{ and } Q = 2.871$$

$$\{D, {}^1X, {}^0X\}\text{: } R = 0.979,\ S = 0.356 \text{ and } Q = 2.750$$

Here, inclusion of the last ${}^0X$ index is, clearly, statistically insignificant. The fact that simple connectivity indices produce also good regressions for R_1 indicates that the valence information encoded by these indices (for sp^3 fragments $\delta = \delta^v$) is enough to give good $R_1(C\alpha)$ projections.

Specific *SR* rotations of amino acids are, by far, better described by the simple molecular connectivity indices (best valence indices describe *SR* with $R = 0.686$ and $Q = 0.026$) and this fact indicates that this property is mainly determined by structural parameters, valence electrons playing a minor role. The first good descriptor of *SR* is index *D*, which gives, anyway, a very poor description of *SR*. Together with the index *D*, in a two-index regression, the best index is ${}^0X$ (the coefficient of which is 43.474, i.e., the coefficient of ${}^2\Omega_c$) with $R = 0.826$, $S = 17.60$ and $Q = 0.047$. But it is the combination of ${}^1X$ and ${}^0X$ indices which produces the best two-index regression with an impressive improvement (more than four-fold) of the *Q* factor. This two-index regression is even better than the three-index regression, the *Q* factor of which diminishes due to a growing *S*. To notice is the fact that indices ${}^1X$ and especially ${}^0X$ alone produce astonishingly poor correlations:

$${}^1X \text{ describes } SR \text{ with } R = 0.260,\ S = 28.94 \text{ and } Q = 0.009$$

$${}^0X \text{ describes } SR \text{ with } R = 0.045,\ S = 29.94 \text{ and } Q = 0.0015$$

This underlines the fact, already noticed by Randić (1991a and b), that successive inclusion of new descriptors to the best ones from a previous step, may miss the optimal combination of descriptors.

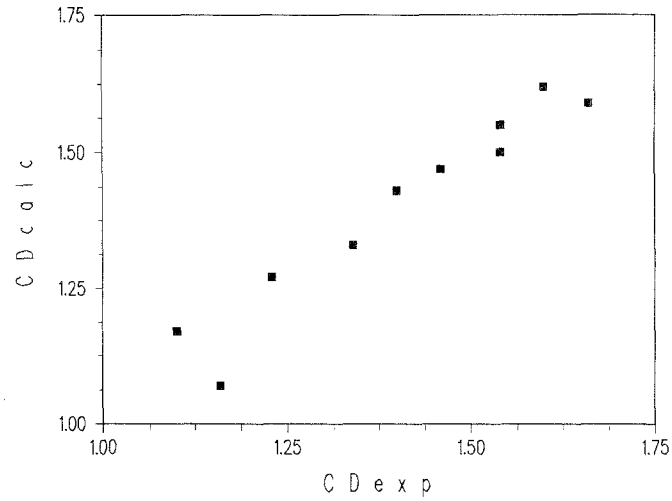


Fig. 1. Plot of the calculated versus the experimental crystal density, *CD*, for 10 amino acids

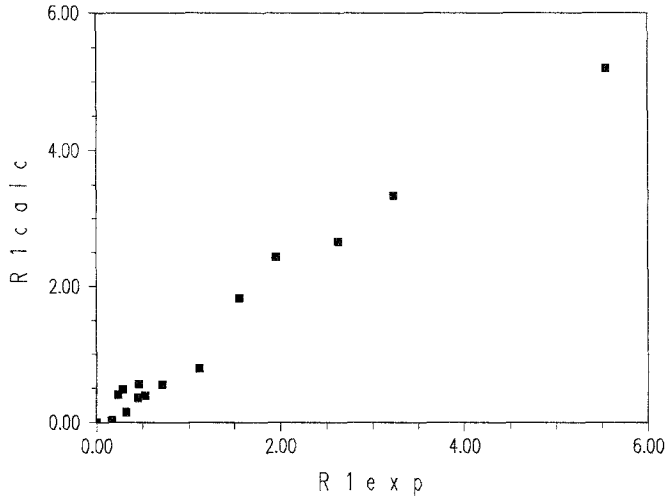


Fig. 2. Plot of the experimental versus the calculated relaxation rates, $R_1 = 1/T_1$, of 8 amino acids and 6 diketo-piperazines

In Fig. 1–3 given experimental properties have been plotted, by the aid of the three-term regressions, vs. the corresponding calculated ones. These figures clearly indicate, near the rather poor description of the *SR* property, the good description of the R_1 and *CD* properties.

Analyzing, in Tables 5–7, the coefficients of the regressions and their constant term we observe that, after inclusion of a new index, they fluctuate randomly. This fact and the worsening of the Q value with inclusion of a new index is due to the partial collinearity among the different molecular connectivity indices which may, in some extent, duplicate one another. To analyze the extent to which the connectivity indices of each set are linearly intercorrelated we investigate the linear relationship between two pair of indices X_i and X_j ($i = 1, 2$ and $j = i + 1$, 3) of our sets

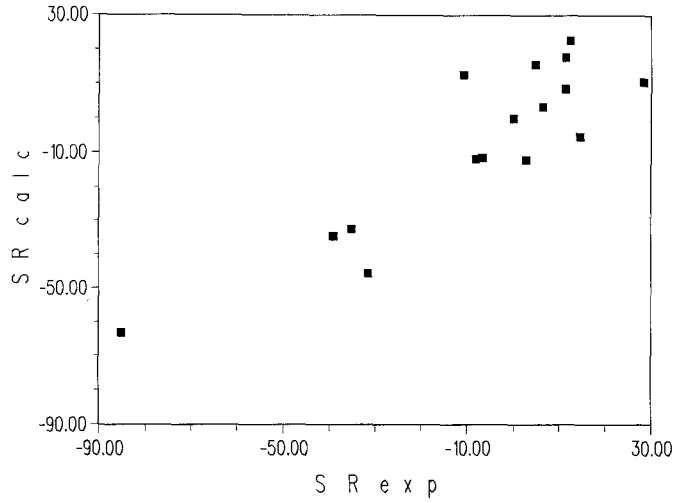


Fig. 3. Plot of the calculated versus the experimental specific rotation, $[\alpha] = SR$, for 16 amino acids

$$X_i = aX_j + b \tag{4}$$

and we will consider (Mihalic et al., 1992) the correlation coefficient R as the criterion for the intercorrelation and as strongly intercorrelated indices those with the $R \geq 0.98$. For CD, R_1 and SR the resulting R values of the two sets of indices are CD:

$$R(D) = 1,\ R({}^0X) = 0.98,\ R({}^1X) = 0.99;\ R({}^0X) = 1,\ R({}^1X) = 0.99$$

$$R(D^v) = 1,\ R({}^0X^v) = 0.73,\ R({}^1X^v) = 0.61;\ R({}^0X^v) = 1,\ R({}^1X^v) = 0.97$$

R_1:

$$R(D) = 1,\ R({}^0X) = 0.99,\ R({}^1X) = 0.999;\ R({}^0X) = 1,\ R({}^1X) = 0.996$$

$$R(D^v) = 1,\ R({}^0X^v) = 0.97,\ R({}^1X^v) = 0.98;\ R({}^0X^v) = 1,\ R({}^1X^v) = 0.99$$

SR:

$$R(D) = 1,\ R({}^0X) = 0.93,\ R({}^1X) = 0.98;\ R(0X) = 1,\ R({}^1X) = 0.97$$

$$R(D^v) = 1,\ R({}^0X^v) = 0.72,\ R({}^1X^v) = 0.62;\ R({}^0X^v) = 1,\ R({}^1X^v) = 0.96$$

From these R values we notice that i) valence indexes are less intercorrelated than simple indices which are normally strongly intercorrelated, ii) valence indexes are, with the exception in R_1, poorly intercorrelated and iii) both sets of indices in R_1 show nearly the same strong intercorrelation. The fact that SR is better described by the stronger intercorrelated set shows that intercorrelation cannot be taken as the criterion for the inclusion or exclusion of an ordinary descriptor and even with an intercorrelation as high as 0.98 the fraction of structural information not in common between two descriptors is always important. The strong intercorrelation among the indices which describe SR tell us that orthogonal indices should play an essential role in the description of specific rotation of amino acids.

Orthogonal descriptors were introduced by Randić (1991a and b) in applications using multivariate regressions to bypass the problem of mutual relatedness (collinearity) among different descriptors, which results in highly unstable estimated regression coefficients subject to consistent changes when a single descriptor is deleted or included in the regression analysis.

Analyzing *CD*, R_1 and *SR* orthogonal regressions (regressions with orthogonal connectivity indices) in Table 5–7, three features are evident:

(i) Sets of orthogonal connectivity indices $\{({}^1\Omega_i, {}^2\Omega_i, {}^3\Omega_i)$ with $i = a, b, c\}$ for *CD*, *R*1 and *SR* (and ${}^1\Omega'$, ${}^2\Omega'/{}^1\Omega''$, ${}^2\Omega''$ for *SR*) are different from each other, i.e., the ordering in which the subsequent ordinary molecular connectivity indices are added in the normal regressions fixes the orthogonalization process.

(ii) The numerical stability of the regression equations with the inclusion or deletion of a new orthogonal index. This stability lies in the constancy of the equation coefficients and of the constant of the regression.

(iii) Statistical coefficients *R*, *S* and *Q* are the same whether we use a set of ordinary or corresponding orthogonal descriptors. In fact, the latter set, stemming from the former, cannot expand the information content of the former. This means that the predictive capability of the normal and of the corresponding orthogonal regressions are the same, i.e., Fig. 1–3 could also have been obtained from the corresponding orthogonal regressions (last regressions in Tables 5–7) if ${}^i\Omega$ values were known.

Due to their constancy, under inclusion or deletion of a new orthogonal index, the numerical values for the coefficients acquire now a clear meaning: indices with the largest coefficient, in the optimal regression, are the dominant orthogonal variable (Randić, 1991a and b). Optimal orthogonal regressions for *SR* are the second and third one of Table 7 derived from the second ordinary regression, once selecting ${}^1X$ as the first ${}^1\Omega'$ orthogonal descriptor and once choosing ${}^0X$ as ${}^1\Omega''$ and making the remaining index orthogonal to ${}^1X$ or to ${}^0X$ respectively. By the aid of these two correlations it is possible to illustrate how versatile the orthogonal indices are. The dominant variables in these two regression are ${}^2\Omega'$ and ${}^2\Omega''$ the values of which have been collected in Table 8. By the aid of these values we are able to derive the following correlations

$$SR = 67.389\ {}^2\Omega' - 8.285 \tag{5}$$

Table 8. Orthogonal connectivity indices ${}^2\Omega'$ and ${}^2\Omega''$ used in the *SR* regressions of amino acids and based on ${}^0X$ (with ${}^1X \equiv {}^1\Omega'$) and on ${}^1X$ (with ${}^0X \equiv {}^1\Omega''$) ordinary indices respectively

AA:	Thr	Lys	Arg	Glu	Ile	Val	Asp	Ala
${}^2\Omega'$:	0.197	0.036	0.541	0.379	0.230	0.197	0.298	−0.241
${}^2\Omega''$:	−0.98	−0.009	−0.337	−0.276	−0.191	−0.198	−0.245	0.075
AA:	Ser	Met	Leu	Trp	Phe	His	Pro	
${}^2\Omega'$:	−0.207	−0.045	0.299	−0.243	−0.246	−0.327	−0.868	
${}^2\Omega''$:	0.083	0.021	−0.246	0.350	0.263	0.294	0.616	

$$R = 0.843, \quad S = 16.12, \quad Q = 0.052$$
$$SR = -90.969\ {}^{2}\Omega'' - 8.285 \tag{6}$$
$$R = 0.881, \quad S = 14.17, \quad Q = 0.062$$

As expected the coefficients of the variables in eqs. (5) and (6) and in the corresponding two-Ω-index regressions of Table 7 are the same. Comparison between the statistical parameters of regressions with ${}^{1}X$ or ${}^{0}X$ (see text above) with regressions (5) and (6) show the astounding quality of the orthogonal indices. Linear regression (6) shows an even higher Q value than the two-index regressions. We have, thus, found the dominant variable for encoding SR and condensed the sought information in a simple linear equation.

Conclusions

In this work we carried out a QSPR study of three different physicochemical properties of amino acids: crystal densities, relaxation rates (for this last property cyclic peptides were included) and specific rotations. These properties are encoded by two qualitatively different sets of molecular connectivity indices: simple and valence connectivity indices. CD is exclusively encoded by valence connectivity indices, R_1 nearly by both types of indices and SR only by simple connectivity indices which have a direct structural meaning. While best correlations for CD and R_1 are the ones with all the three basic $\{D^v, {}^{0}X^v, {}^{1}X^v\}$ descriptors included, for SR the two-index basis set $\{{}^{1}X, {}^{0}X\}$ shows the best quality even if the three-index set shows the highest R value. Making use of simple indices to describe R_1 the simple basis set $\{D, {}^{1}X\}$ shows a better quality than the $\{D, {}^{1}X, {}^{0}X\}$ set. QSPR model of SR gives us possibility to detect i) how strongly intercorrelated ordinary indices can always give a good description of a given property and ii) how important orthogonal indices are. Orthogonal connectivity indices, bypassing collinearity which exists among ordinary indices give rise to stable regressions which are formally more appealing and which allow, once the optimal regression is known, a meaningful interpretation of the coefficients and the detection of the dominant variable as is the case with the description of SR. Clearly orthogonal connectivity indices do not have any direct structural or electronic meaning and, as they are strongly dependent on the orthogonalization process, calculations to derive them, in the case of more than three ordinary indices and three properties to fit, could become a rather delicate and arduous task.

References

CRC Handbook of Chemistry and Physics, 64th edn. (1984–1985) In: Weast RC (ed) CRC Press Inc., Boca Raton, Florida, pp C-722–C-727

Hansen PJ, Jurs PC (1988) Chemical applications of graph theory. J Chem Ed 65: 574–580

Kier LB, Hall LH (1976) Molecular connectivity VII: specific treatment of heteroatoms. J Pharm Sci 65: 1806–1809

Kier LB, Hall LH (1981) Derivation and significance of valence molecular connectivity. J Pharm Sci 70: 583–589

Kier LB, Hall LH (1986) Molecular connectivity in structure-activity analysis. Wiley, New York
Kier LB, Hall LH, Murray WJ, Randić M (1975) Molecular connectivity I: relationships to nonspecific local anesthesia. J Pharm Sci 64: 1971–1974
Mihalić Z, Trinajstić N (1992) A graph-theoretical approach to structure-property relationships. J Chem Ed 69: 701–712
Mihalić Z, Nikolić S, Trinajstić N (1992) Comparative study of molecular descriptors derived from the distance matrix. J Chem Inf Comput Sci 32: 28–37
Pogliani L (1992a) Molecular connectivity model for determination of isoelectric point of amino acids. J Pharm Sci 81: 334–336
Pogliani L (1992b) Molecular connectivity: treatment of the electronic structure of amino acids. J Pharm Sci 81: 967–969
Pogliani L (1993a) A QSAR model of the isoelectric points and of the atomic charges of amino acids in chemistry and properties of biomolecular systems, vol. 2. In: Russo N et al. (eds) Kluwer, Dordrecht (in press)
Pogliani L (1993b) Molecular connectivity model for determination of T_1 relaxation times of α-carbons of amino acids and cyclic dipeptides. Computers Chem (in press)
Pogliani L (1993c) Molecular connectivity model for determination of physicochemical properties of α-amino acids. J Phys Chem 97: 6731–6736
Randić M (1975) On characterization of molecular branching. J Am Chem Soc 97: 6609–6615
Randić M (1991a) Orthogonal molecular descriptors. N J Chem 15: 517–525
Randić M (1991b) Resolution of ambiguities in structure-property studies by use of orthogonal descriptors. J Chem Inf Comput Sci 31: 311–320
Rouvray DH (1989) The limits of applicability of topological indices. J Mol Struct (Theochem) 185: 187–201
Trinajstić N (1983) Chemical graph theory, vol 2, ch 4. CRC Press, Boca Raton, FL
Turro NJ (1986) Geometric and topological thinking in organic chemistry. Angew Chem (Int Edn Engl) 25: 882–901

Author's address: Prof. L. Pogliani, Dipartimento di Chimica, Universitá della Calabria, I-87030 Rende (CS), Italy. (On sabbatical leave for the Centro de Quimica-Fisica Molecular, Instituto Superior Técnico, P-1096 Lisboa Codex, Portugal).

Received May 5, 1993