Peptide Property Modeling by Cluj Indices

Authors: D. Opris a; M. V. Diudea a

Affiliation: a Department of Organic Chemistry, Faculty of Chemistry and Chemical Engineering, "Babes-Bolyai" University. Cluj-Napoca, 3400. Romania

DOI: 10.1080/10629360108035377

Publication Frequency: 8 issues per year

Published in: SAR and QSAR in Environmental Research, Volume 12, Issue 1 & 2 April 2001 , pages 159 - 179

Subjects: Applied & Industrial Chemistry; Chemistry; Environmental & Ecological Toxicology; Environmental Sciences; History & Philosophy of Mathematics;

Formats available: PDF (English)

Abstract

The novel Cluj property indices are used for modeling the biological properties of dipeptides: the ACE inhibition activity of a set of 58 dipeptides and the bitter tasting activity of a set of 48 dipeptides, taken from the literature. The results are compared to those reported in some previous works.

Keywords: Molecular graph; Topological index; Peptide; Cluj indices
PEPTIDE PROPERTY MODELING BY CLUJ INDICES

D. OPRIS and M. V. DIUDEA*

Department of Organic Chemistry, Faculty of Chemistry and Chemical Engineering, “Babeș-Bolyai” University, 3400 Cluj-Napoca, Romania


Dedicated to Professor Alexandru T. Balaban, for his bright contribution to Chemical Graph Theory and development of this science branch in Romania

* Author to whom all correspondence should be addressed.
   e-mail address: diudea@chem.ubbcluj.ro

Abstract

The novel Cluj property indices are used for modeling the biological properties of dipeptides: the ACE inhibition activity of a set of 58 dipeptides and the bitter tasting activity of a set of 48 dipeptides, taken from the literature. The results are compared to those reported in some previous works.

Keywords: molecular graph; topological index; peptide
Quantitative Structure-Activity Relationships (QSARs) establish a mathematical relation between the biological activity of chemical compounds and their molecular structure. They provide quantitative models aimed to accurately predict a certain activity from the structural attributes. This topic has become a well-delimited branch in chemistry and was favored by the progress in computer science.

Many biological and physico-chemical properties can be correlated with the topostructural and topochemical features. Topological indices, TIs, (encoding information regarding the size, shape, branching or centricity) are among the simplest and efficient descriptors of molecular structure. Biological properties, more than physical or chemical properties, depend on the three-dimensional (3D) arrangements of the atoms in a molecule.

Coding of 3D structural information can be achieved in many different ways but all procedures make use of the Euclidean distances, irrespective they are considered explicitly or are involved in some more cryptic (3D sensitive) molecular descriptors. Since the actual conformation of a biological receptor is known only in fortunate cases, the researcher has to choose between the use of a single conformation (e.g. the minimum energy isomer) of the actives and of rotamer-library conformations. Excepting cases of very specific receptors (where more elaborated 3D grid descriptors are strongly needed for drawing virtual receptor maps - see the CoMFA-like procedures) [1-4], biological activity can be satisfactorily described by using a single conformation of actives and, moreover, by using 2D topological descriptors. Indeed, topological indices encode some information regarding the spatiality of molecules (e.g. as topological distance). Such a description (that is invariant to rotation and torsion) can sometimes be assimilated to the so-called "extended" conformation [5].

As the computer technology developed, novel descriptors, with enhanced ability in coding the 3D structural information and modeling a certain molecular property/activity could be designed. Large pools of descriptors were thus created.

In this paper a new approach, leading to a fragmental property index family, \( FPIF \), is presented. These indices are calculated as local descriptors of some fragments of the molecule and, a global index is then obtained by summing the fragmental contributions. The modeling ability of \( FPIF \) is demonstrated on two sets of dipeptides, taken from literature [5-8].
The graph-theoretical descriptors \( CJ, CF \) and \( SZ \) \(^{[9-14]} \) represent the theoretical ground for counting the fragmental property indices. These descriptors are derived from the cardinality of the vertex sets defined by:

\[
CJ_{i,j,p} = \{ v \mid v \in V(G); \ d(G)_{v,i} < d(G)_{v,j}, \text{ and } \exists w \in W_{v,i}: V(w) \cap V(p) = \{ i \} \} \quad (1)
\]

\[
CF_{i,j,p} = \{ v \mid v \in V(G); \ d(G_p)_{v,i} < d(G_p)_{v,j}; \ G_p = G - p \} \quad (2)
\]

\[
SZ \_{i,j} = \{ v \mid v \in V(G); \ d(G)_{v,i} < d(G)_{v,j} \} \quad (3)
\]

In the above relations, \( G_p = G - p \) is the spanning subgraph, resulted by deleting the path \( p \) joining the vertices \( i \) and \( j \) (except its endpoints), \( d(G) \) and \( d(G_p) \) denote the topological distances measured in \( G \) and \( G_p \), respectively.

The sets \( CJ_{i,j,p} \) and \( CF_{i,j,p} \) represent subgraphs (connected or not) in \( G \), referred to the endpoint \( i \) and related to \( j \) and path \( p \).

In defining Cluj indices, the path \( p \) plays the central role in selecting the subgraphs (eqs 1 and 2), particularly in cycle-containing graphs, where more than one path could join the pair \((i,j)\). In such graphs, more than one subgraph (i.e. fragment), referred to \( i \), can be counted. By this reason, the nondiagonal entries \([UM]_{ij}\) in Cluj matrices are defined as the maximum cardinality of the sets supplied by eq 1 or 2

\[
[UM]_{ij} = \max_p |V_{i,j,p}| \quad (4)
\]

where \( V_{i,j,p} \) is either \( CJ_{i,j,p} \) or \( CF_{i,j,p} \) and consists of vertices \( v \) lying closer to the vertex \( i \) than to the vertex \( j \). When \( p \in D(G) \), (i.e. the set of all topological distances, or geodesics in \( G \)) then \( M = CJD \) (Cluj-Distance) or \( CFD \) (Cluj-Fragmental-Distance). When \( p \in A(G) \), (i.e. the set of all topological detours, or the longest distances in \( G \)) \( M = CJ \Delta \) (Cluj-Detour) or \( CFA \) (Cluj-Fragmental-Detour). The diagonal entries are zero. The Cluj matrices are square arrays, of
dimension \( N \times N \), usually *unsymmetric* (excepting some symmetric regular graphs). Figures 1 and 2 illustrate the construction of matrices \( \text{CJD} \) and \( \text{CJ} \Delta \) respectively.

---

**Figure 1. Construction of Cluj Distance Matrix, UCJD**

![Cluj Distance Matrix](image)

Cluj Distance Sets \( CJD_{i,j,p} \); pair (6,8):

(a) (6, 8) \[ 6, 2, 5, 8 \] \{ 3, 6, 9, 12 \}

(b) (6, 8) \[ 6, 9, 11, 8 \] \{ 2, 3, 6 \}

(a) (8, 6) \[ 8, 5, 2, 6 \] \{ 1, 4, 7, 8, 10, 11 \}

(b) (8, 6) \[ 8, 11, 9, 6 \] \{ 1, 4, 5, 7, 8, 10 \}

**Cluj Distance Matrix UCJD**

\[
\begin{array}{cccccccccccc}
0 & 3 & 5 & 9 & 3 & 5 & 5 & 2 & 3 & 5 & 3 & 5 \\
5 & 0 & 7 & 7 & 5 & 7 & 5 & 3 & 4 & 5 & 4 & 7 \\
1 & 1 & 0 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\
3 & 3 & 4 & 0 & 2 & 3 & 6 & 2 & 3 & 2 & 3 & 2 \\
9 & 7 & 7 & 8 & 0 & 6 & 8 & 6 & 6 & 5 & 4 & 6 \\
5 & 5 & 11 & 4 & 0 & 6 & 4 & 4 & 3 & 6 \\
2 & 2 & 3 & 6 & 2 & 3 & 0 & 2 & 3 & 3 & 3 & 4 \\
5 & 4 & 6 & 8 & 6 & 6 & 8 & 6 & 9 & 7 & 7 \\
4 & 3 & 6 & 5 & 4 & 6 & 5 & 4 & 0 & 5 & 5 & 11 \\
5 & 3 & 5 & 5 & 2 & 3 & 9 & 3 & 5 & 0 & 3 & 5 \\
5 & 4 & 7 & 5 & 3 & 4 & 7 & 5 & 7 & 5 & 0 & 7 \\
1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 0 \\
\end{array}
\]

\( IP2(CJD) = 1024 \)

\( IE2(CJD) = 378 \)
Figure 2. Construction of Cluj Detour matrix, UCJΔ

Cluj Detour Sets $CJΔ_{i,j,p}$; pair (5,8):

(a) $(5, 8) \ [ 5, 1, 4, 7, 10, 8 ] \ { 2, 3, 5, 6 }$
(b) $(5, 8) \ [ 5, 2, 6, 9, 11, 8 ] \ { 1, 4, 5 }$

(a) $(8, 5) \ [ 8, 10, 7, 4, 1, 5 ] \ { 8, 9, 11, 12 }$
(b) $(8, 5) \ [ 8, 11, 9, 6, 2, 5 ] \ { 7, 8, 10 }$

Cluj Detour Matrix $UCJΔ$

\[
\begin{array}{cccccccccccc}
0 & 1 & 2 & 1 & 1 & 2 & 1 & 2 & 1 & 2 & 1 & 1 \\
1 & 0 & 1 & 2 & 1 & 1 & 1 & 3 & 1 & 1 & 3 & 3 \\
1 & 1 & 0 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\
1 & 3 & 3 & 0 & 1 & 3 & 1 & 2 & 2 & 1 & 2 & 2 \\
1 & 1 & 2 & 1 & 0 & 1 & 2 & 4 & 2 & 2 & 2 & 2 \\
5 & 2 & 2 & 1 & 1 & 1 & 1 & 1 & 2 & 1 & 0 & 3 \\
1 & 2 & 2 & 1 & 2 & 2 & 0 & 1 & 3 & 1 & 3 & 3 \\
2 & 2 & 2 & 2 & 4 & 2 & 1 & 0 & 1 & 1 & 1 & 2 \\
3 & 2 & 2 & 3 & 4 & 2 & 5 & 2 & 0 & 5 & 2 & 11 \\
2 & 1 & 1 & 1 & 1 & 1 & 1 & 2 & 0 & 1 & 2 & 1 \\
1 & 3 & 3 & 1 & 3 & 1 & 2 & 1 & 1 & 1 & 0 & 1 \\
1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 0 \\
\end{array}
\]

$IP2(CJΔ) = 247$

$IE2(CJΔ) = 53$
The entries in the unsymmetric Szeged distance matrix, \( \text{USZD} \), are supplied by the cardinality of the sets in eq 3. Note that in defining the Szeged fragments, the path joining the vertices \( i \) and \( j \) is irrelevant. Thus, for each pair \((i, j)\) it results one and only one fragment.

When the distance criterion \( d(G)_{v,i} < d(G)_{v,j} \) (eq 3) is changed by the detour criterion \( \delta(G)_{v,i} < \delta(G)_{v,j} \), the cardinality of the sets thus supplied represent entries in the unsymmetric Szeged detour matrix, \( \text{USZ} \Delta \).

The above definitions hold for any connected graph

The unsymmetric matrices can be symmetrized, e.g., by the Hadamard product with their transposes

\[
\begin{align*}
\text{SM}_p &= \text{UM} \bullet (\text{UM})^T \tag{5} \\
\text{SM}_e &= \text{SM}_p \bullet \text{A} \tag{6}
\end{align*}
\]

The symbol \( \bullet \) indicates the Hadamard (pairwise) matrix product (i.e. \([M_a \bullet M_b]_{ij} = [M_a]_{ij} [M_b]_{ij}\)).

In eq 6, the Hadamard product between the path-defined matrix \( \text{SM}_p \) and the adjacency matrix \( \text{A} \) (i.e. the matrix having the non-diagonal entries unity for two adjacent vertices and zero otherwise) provides the corresponding edge-defined matrix, \( \text{SM}_e \), which is a weighted adjacency matrix. For the symmetric matrices, the letter \( \text{S} \) is usually missing.

In trees, \( \text{CJD}_e = \text{CFD}_e = \text{SZD}_e \). In cyclic graphs, \( \text{CJD}_p \neq \text{CFD}_p \neq \text{SZD}_p \), \( \text{CJ}_\Delta_p \neq \text{CF}_\Delta_p \neq \text{SZ}_\Delta_p \).

The above-discussed matrices allow the calculation of integer value indices by relations given for the fragmental property indices (see below) [14].
FragmenTal Property Indices

Model Parameters

In physical phenomena, the macroscopic interactions are often interactions of field-type. The field is produced by a scalar function of potential. Let \( f(x, y, z) \) be such a scalar function. The field induced by this function can be written as:

\[
\vec{\nabla} \cdot \mathbf{f} = \left( \frac{\partial f}{\partial x} \hat{i} + \frac{\partial f}{\partial y} \hat{j} + \frac{\partial f}{\partial z} \hat{k} \right) \cdot \mathbf{f}(x, y, z) = \frac{\partial f}{\partial x} \hat{i} + \frac{\partial f}{\partial y} \hat{j} + \frac{\partial f}{\partial z} \hat{k}
\]  

(7)

For the potential of type

\[ f(x, y, z) = pz \]  

(8)

the associated field can be derived as

\[
\vec{\nabla} \cdot \mathbf{f} = \frac{\partial f}{\partial x} \hat{i} + \frac{\partial f}{\partial y} \hat{j} + \frac{\partial f}{\partial z} \hat{k} = \frac{\partial (pz)}{\partial x} \hat{i} + \frac{\partial (pz)}{\partial y} \hat{j} + \frac{\partial (pz)}{\partial z} \hat{k} = 0 \hat{i} + 0 \hat{j} + pk = pk = \vec{p}
\]

(9)

This is the case of the well-known uniform gravitational field:

\[
\vec{G} = m\vec{g}
\]

(10)

with the corresponding potential given by

\[ E_p = E_p(z) = mgz \]  

(11)

where \( m \) is the mass of the probe and \( z \) is the reference coordinate.

Note that eq 9 is applicable both to the Newtonian (gravitational) interactions and the Coulombian (electrostatic) interactions. In both cases the relation is valid if the mass \( m \) (or the charge \( q \)) that generates the potential \( f \) and associated field \( \vec{\nabla} \cdot \mathbf{f} \) is far enough \((r >> z)\) so that the approximation \((r + z)^2/r^2 = (r^2 + 2rz + z^2)/r^2 = 1 + 2z/r + (z/r)^2 \approx 1\) is valid.

For the potential of type:

\[ f(x, y, z) = p/z \]  

(12)

eq 7 leads to the associated field:

\[
\vec{\nabla} \cdot \mathbf{f} = \frac{\partial f}{\partial x} \hat{i} + \frac{\partial f}{\partial y} \hat{j} + \frac{\partial f}{\partial z} \hat{k} = \frac{\partial (p/z)}{\partial x} \hat{i} + \frac{\partial (p/z)}{\partial y} \hat{j} + \frac{\partial (p/z)}{\partial z} \hat{k} = 0 \hat{i} + \frac{-p}{z^2} \hat{k} = \frac{-p}{z^2} \hat{k} = \frac{-p}{z^2} \]  

(13)
This is the case of well-known (non-uniform) gravitational field:

\[ \vec{G} = \vec{G}(m, r) = -k \frac{m}{r^3} \vec{r} \]  

(14)

and the associated potential of the form:

\[ U = U(m, r) = k \frac{m}{r} \]  

(15)

where \( m \) is the mass of the probe and \( r \) is the position relative to the location of the point producing the field.

For the Coulombian field eq 13 becomes:

\[ \vec{F_C} = \vec{F_C}(r) = -k \frac{q}{r^3} \vec{r} \]  

(16)

and the potential associated to the Coulombian field:

\[ U = U(q, r) = k \frac{q}{r} \]  

(17)

Four models were implemented in the view of building the fragmental property indices: two of them topological (dense topological and rare topological) and two others geometric (dense geometric and rare geometric). In these models a weak dependence on distance for the potential of the type (8) generating a uniform field (9), and a strong dependence on distance for the potential of the type (12) that generates a non-uniform field (13) were considered.

The variables in the models are: property \( \Phi \) (mass \( M \), electronegativity \( E \), cardinality \( C \), partial charge or any other atomic property \( P \)), property descriptor \( \Omega \) (\( p \), \( d \), \( pd \), \( 1/p \), \( 1/d \), \( p/d \), \( p/d^2 \), \( p^2/d^2 \)) and superposition \( \Psi \) (\( S \), \( P \), \( A \), \( G \), \( H \)).

The expressions for the property descriptors are:

\[ \Omega : p = p ; d = d ; pd = p \cdot d ; 1/p = \frac{1}{p} ; 1/d = \frac{1}{d} ; p/d = \frac{p}{d} ; p/d^2 = \frac{p}{d^2} ; p^2/d^2 = \frac{p^2}{d^2} \]  

(18)

where \( p \) is any property (\( p \in \Phi \)) and \( d \) is any metric of distance.

The (mathematical) superposition \( \Psi \), given by

\[ \Psi^\Phi S = \sum_{i=1}^{n} x_i ; P = \prod_{i=1}^{n} x_i ; A = S / n ; \ G = (\text{sgn}(P))^n \cdot \sqrt[n]{\text{abs}(P)} ; \ H = \left( \sum_{i=1}^{n} \frac{1}{x_i} \right)^{-1} \]  

(19)
is applied upon a string of vertex descriptors to give a fragment descriptor. The used symbols are: 

\( S = \text{sum}; \ P = \text{product}; \ A = \text{arithmetic mean}; \ G = \text{geometric mean} \) and \( H = \text{harmonic sum} \). The summation is suitable in case of any additive property (mass, volume, partial charges, electric capacities, etc.) [18]. The multiplication occurs in concurrent phenomena (probabilistically governed) [19-21]. Other operators find appropriate justification [22-26].

**Model Description**

Let \((i,j)\) be a pair of vertices and \(Fr_{i,j}\) any fragment referred to \(i\) and related to \(j\).

**Dense Topological Model**

Let \(v\) be a vertex in the fragment \(Fr_{i,j}\). The property descriptor applies to the vertex property \(p_v\) and topological distance \(d_{Tv,j}\). The fragmental property descriptor \(PD\), resulting by the vertex descriptor superposition, gives the interaction of all the points belonging to the fragment \(Fr_{i,j}\) with the point \(j\):

\[
PD(Fr_{i,j}) = \sum_{v \in Fr_{i,j}} \Omega(d_{Tv,j}, p_v)
\]  

(20)

The \(j\) point can be conceived as an internal probe atom (see the CoMFA approach) [1-4]. However, the chemical identity of \(j\) is not considered.

**Rare Topological Model**

Within this model, the property descriptor applies to the fragmental property and topological distance \(d_{T_{ij}}\). The fragmental property descriptor models the interaction of the whole fragment \(Fr_{i,j}\) with the point \(j\) and looks the global property being concentrated in the vertex \(i\):

\[
PD(Fr_{i,j}) = \Omega(d_{T_{ij}}, \sum_{v \in Fr_{i,j}} p_v)
\]  

(21)

**Dense Geometric Model**

The fragmental property descriptor is the vector sum of the vertex descriptor vectors. It applies the property descriptor to the vertex property \(p_v\) and the Euclidean distance \(d_{Ev,j}\) in providing a point of equivalent (fragmental) property located at the Euclidean distance \(d_{E CP,j}\) (with \(d_{E CP,j}\) being the distance of property). The vector of the fragmental property has the orientation of this
distance vector. The model simulates the interactions in non-uniform fields (gravitational, electrostatic, etc):

\[
PD(F_{ri,j}) = \frac{\sum_{v \in F_{ri,j}} \tilde{\Omega}(d_{E_{v,j}}, p_v)}{\tilde{\Omega}} = \frac{\Omega}{d_{E_{v,j}}}; \quad P(F_{ri,j}) = \sum_{v \in F_{ri,j}} \Psi(p_v);
\]

\[d_{E_{CP,j}} = \Omega^{-1}_p(DG(F_{ri,j}), P(F_{ri,j})),\]  \hspace{1cm} (22)

where \(d_{E_{CP,j}}\) is the distance that satisfies: \(\Omega(d_{E_{CP,j}}, P(F_{ri,j})) = PD(F_{ri,j})\)

**Rare Geometric Model**

The scalar fragmental descriptor applies the property descriptor to the center of fragment property and Euclidean distance between this center and the vertex \(j\).

The model simulates the interactions in uniform fields (uniform gravitational, electrostatic, etc.):

\[
PD(F_{ri,j}) = \Omega(d_{E_{CP,j}}, \sum_{v \in F_{ri,j}} \Psi(p_v)); \quad CP_i(x_{CP,i,j}, y_{CP,i,j}, z_{CP,i,j});\]  \hspace{1cm} (23)

\[
x_{CP,i,j} = \sum_{v \in F_{ri,j}} x_v \cdot p_v / \sum_{v \in F_{ri,j}} p_v ; \quad y_{CP,i,j} = \sum_{v \in F_{ri,j}} y_v \cdot p_v / \sum_{v \in F_{ri,j}} p_v ; \quad z_{CP,i,j} = \sum_{v \in F_{ri,j}} z_v \cdot p_v / \sum_{v \in F_{ri,j}} p_v
\]

Some Particular Fragmental Property Models were discussed elsewhere [14].

**Fragmental Property Matrices**

The fragmental property matrices are square matrices of the order \(N\) (i.e. the number of non-hydrogen atoms in the molecule). The non-diagonal entries in such matrices are fragmental properties corresponding to any pair of vertices \((i,j)\) by a chosen model.

In case of Cluj criteria, the fragmentation can supply more than one maximal fragment for the pair \((i,j)\). In such cases, the matrix entry is the arithmetic mean of the individual values.

Thus, if \(i, j \in V(G), i \neq j\) and \(P_{ij} = \{ p_{i,j}^1, p_{i,j}^2, ..., p_{i,j}^m \}\) paths joining \(i\) and \(j\), then cf. \(CJ\) or \(CF\) definition (eqs 1-3), the fragments \(F_{ri,j}^1, F_{ri,j}^2, ..., F_{ri,j}^m\) are generated. Let \(m\) be the number of
maximal fragments among all the \( k \) fragments, \( 1 \leq m \leq k \), and let \( \sigma_1, ..., \sigma_m \) be the index for the maximal fragments.

By applying any of the above models, for all \( m \) maximal fragments we obtain \( m \) values and, consequently, the matrix entry associated to the pair \((i,j)\) is the mean value, e.g.

\[
PD_{i,j} = \frac{\sum_{t=1}^{m} PD(F_{i,j}^{\sigma_t})}{m}
\]

(24)

The resulting matrices are in general \textit{unsymmetric} but they can be symmetrized (see eqs 5, 6). The symbols for the fragmental property matrices will be detailed below.

\textbf{Fragmental Property Indices}

Fragmental property indices are calculated at any fragmental property matrices above discussed, by applying four types of index operators: \( P_-, P^2, E_-, E^2 \) according to the relations:

\[
P_-(M) = \frac{1}{2} \sum \sum [M]_{i,j} \quad ; \quad P^2(M) = \frac{1}{2} \sum \sum [M]_{i,j} [M]_{j,i};
\]

\[
E_-(M) = \frac{1}{2} \sum \sum [M]_{i,j} [A]_{i,j} \quad ; \quad E^2(M) = \frac{1}{2} \sum \sum [M]_{i,j} [M]_{j,i} [A]_{i,j}
\]

(25)

where \( M \) is any property matrix, symmetric or unsymmetric.

\textbf{Symbolism of the Fragmental Property Matrices and Indices}

The name of fragmental property matrices is of the general form:

\[
ABcDdEffffffG
\]

(26)

where:

- \( A \in \{ D, R \}; \ D = \text{Dense}; \ R = \text{Rare} \);
- \( B \in \{ T, G \}; \ T = \text{Topological}; \ G = \text{Geometric} \);
- \( c \in \{ f, j, s \}; \ f = \text{CF-type}; \ j = \text{CJ-type}; \ s = \text{Sz-type} \);
- \( Dd \in \{ Di, De \}; \ Di = \text{Distance}; \ De = \text{Detour} \);
- \( E \in \Phi \) (i.e. \( E \in \{ M, E, C, P \} \) where \( M = \text{mass}; \ E = \text{electronegativity}; \ C = \text{cardinality} \);
- \( P = \) other atomic property - implicitly, \textit{partial charge}; explicitly, a property given by manual input);
- \( fffffff \in \Omega \) (i.e. \( fffffff \in \{ p, \_1/p, \_d, \_1/d, \_p.d, \_p/d, \_p/d2, \_p2/d2 \} \)
\( G \in \Psi \) (i.e. \( G \in \{S, P, A, G, H\} \) with the known meaning (see above).

The name of fragmental property indices is of the general form:

\[
ABcDdEffffffGii
\]  

(27)

where:

\( ii \in \{P\_, P2, E\_, E2\} \) with the known meaning (eq 25).

If an operator, such as \( f(x) = 1/x \) (inverse operator) or \( f(x) = \ln(x) \), is applied the indices are labeled as follows:

\[
\ln\!ABcDdEffffffGii := \ln(ABcDdEffffffGii);
\]

\[
1/ABcDdEffffffGii := \frac{1}{ABcDdEffffffGii} \quad (28)
\]

For example, index \( \ln\!DGfDeM\_p\_SP \) is the logarithm of index \( DGfDeM\_p\_SP \) computed on the property matrix \( DGfDeM\_p\_S \). The model used is dense, geometric, on fragment of type \( CF \), with the cutting path being detour. The chosen property is the mass, the descriptor for property is even the property (mass) and the sum operator counts the vertex descriptors.

The fragmental indices were calculated by the aid of \textit{Cluj3Cmd} 16-bit windows computer programs.

**CORRELATING STUDIES**

The mathematical models of a certain property can be achieved by MLR (Multiple Linear Regression), CNN (Computational Neural Networks) [27-31], or other mathematical procedures. In the following, the MLR procedure is presented.

MLR, for \( n \) observations and \( m \) independent variables is represented by

\[
Y_i = b_0 + \sum_{j=1}^{m} b_{ij} X_{ij}
\]

(29)

or in matrix form as

\[
Y = bX
\]

(30)
where \( \mathbf{Y} \) is the \( n \times 1 \) vector of responses, \( \mathbf{X} \) is an \( n \times (m + 1) \) matrix of independent variables and \( \mathbf{b} \) is the \( (m + 1) \times 1 \) vector of regression coefficients. The regression coefficients can be determined by the least-squares solution of (30)

\[
\mathbf{b} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{Y}
\]

(31)

With \( \mathbf{b} \) calculated, eq 30 can be used for estimating the chosen property for other chemical structures.

We tested the correlating ability of \( \text{FPIF} \) on two sets: Dipeptide ACE Inhibitors. The set consists of 58 dipeptides and was taken from Cocchi's report [6]. The molecular structure of these peptides was input and optimized by using the MM+ and then by semiempirical AM1 procedure of the HyperChem Program (HyperCube Inc.). Table I includes the dipeptide names by using the one-letter code for aminoacids, the observed ACE inhibitory activity (biological activity, \( BA \), as \( \log(1/\text{IC}_{50}) \)), the calculated \( BA \) according to the best model (eq 32) and the corresponding residuals. As above mentioned, \( \text{FPIF} \) descriptors take explicitly into account 3D-structural features of the whole molecule of dipeptides.

Table I.

<table>
<thead>
<tr>
<th>No.</th>
<th>Peptide</th>
<th>( BA_{\text{obs}} )</th>
<th>( BA_{\text{calc(eq 32)}} )</th>
<th>Residuals</th>
<th>Peptide</th>
<th>( BA_{\text{obs}} )</th>
<th>( BA_{\text{calc(eq 33)}} )</th>
<th>Residuals</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>YG</td>
<td>2.7</td>
<td>2.859794</td>
<td>-0.159799</td>
<td>YL</td>
<td>2.4</td>
<td>2.282458</td>
<td>0.117542</td>
</tr>
<tr>
<td>2</td>
<td>YA</td>
<td>3.34</td>
<td>3.411811</td>
<td>-0.07181</td>
<td>WW</td>
<td>3.6</td>
<td>3.627912</td>
<td>-0.02791</td>
</tr>
<tr>
<td>3</td>
<td>WG</td>
<td>2.23</td>
<td>3.297515</td>
<td>-1.06751</td>
<td>WE</td>
<td>1.56</td>
<td>1.768665</td>
<td>-0.20867</td>
</tr>
<tr>
<td>4</td>
<td>VY</td>
<td>4.66</td>
<td>4.519753</td>
<td>0.140247</td>
<td>VV</td>
<td>1.71</td>
<td>1.93277</td>
<td>-0.22277</td>
</tr>
<tr>
<td>5</td>
<td>VW</td>
<td>5.8</td>
<td>5.149804</td>
<td>0.650196</td>
<td>VL</td>
<td>2.117623</td>
<td>-0.11762</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>VP</td>
<td>3.38</td>
<td>3.828656</td>
<td>-0.448666</td>
<td>VA</td>
<td>1.19</td>
<td>1.384127</td>
<td>-0.19413</td>
</tr>
<tr>
<td>7</td>
<td>VG</td>
<td>2.96</td>
<td>2.652658</td>
<td>0.307342</td>
<td>VA</td>
<td>1.16</td>
<td>1.534544</td>
<td>-0.37454</td>
</tr>
<tr>
<td>8</td>
<td>VF</td>
<td>4.28</td>
<td>4.366133</td>
<td>-0.08613</td>
<td>SL</td>
<td>1.49</td>
<td>1.412394</td>
<td>0.077606</td>
</tr>
<tr>
<td>9</td>
<td>TG</td>
<td>2.426741</td>
<td>-0.42674</td>
<td></td>
<td>PY</td>
<td>1.8</td>
<td>2.199467</td>
<td>-0.39947</td>
</tr>
<tr>
<td>10</td>
<td>SG</td>
<td>2.07</td>
<td>2.06538</td>
<td>0.00462</td>
<td>PL</td>
<td>2.22</td>
<td>2.011651</td>
<td>0.208349</td>
</tr>
<tr>
<td>11</td>
<td>RW</td>
<td>4.8</td>
<td>4.77005</td>
<td>0.02995</td>
<td>PI</td>
<td>2.33</td>
<td>1.956997</td>
<td>0.373003</td>
</tr>
<tr>
<td>12</td>
<td>RP</td>
<td>3.74</td>
<td>3.427529</td>
<td>0.312471</td>
<td>PF</td>
<td>2.8</td>
<td>2.595868</td>
<td>0.204132</td>
</tr>
<tr>
<td>13</td>
<td>RF</td>
<td>3.64</td>
<td>3.966233</td>
<td>-0.32623</td>
<td>PA</td>
<td>1.32</td>
<td>1.413292</td>
<td>-0.09329</td>
</tr>
<tr>
<td>14</td>
<td>RA</td>
<td>3.34</td>
<td>2.663339</td>
<td>0.676661</td>
<td>LY</td>
<td>2.46</td>
<td>2.358832</td>
<td>0.101168</td>
</tr>
<tr>
<td>15</td>
<td>QG</td>
<td>2.13</td>
<td>2.023636</td>
<td>0.106364</td>
<td>LW</td>
<td>3.4</td>
<td>3.094146</td>
<td>0.305854</td>
</tr>
<tr>
<td>16</td>
<td>PG</td>
<td>1.77</td>
<td>2.208404</td>
<td>-0.4384</td>
<td>LL</td>
<td>2.35</td>
<td>2.151635</td>
<td>0.198365</td>
</tr>
<tr>
<td>17</td>
<td>MG</td>
<td>2.32</td>
<td>2.708227</td>
<td>-0.38823</td>
<td>LG</td>
<td>1.72</td>
<td>1.507413</td>
<td>0.212587</td>
</tr>
<tr>
<td>18</td>
<td>LG</td>
<td>2.06</td>
<td>2.572545</td>
<td>-0.51254</td>
<td>LF</td>
<td>2.75</td>
<td>2.874128</td>
<td>-0.12413</td>
</tr>
<tr>
<td>19</td>
<td>LA</td>
<td>3.51</td>
<td>3.052476</td>
<td>0.457524</td>
<td>LA</td>
<td>1.72</td>
<td>1.677078</td>
<td>0.042922</td>
</tr>
<tr>
<td>20</td>
<td>KG</td>
<td>2.49</td>
<td>2.208121</td>
<td>0.281879</td>
<td>IW</td>
<td>3.05</td>
<td>2.885932</td>
<td>0.164068</td>
</tr>
<tr>
<td>21</td>
<td>KA</td>
<td>3.42</td>
<td>2.663597</td>
<td>0.756403</td>
<td>IV</td>
<td>2.05</td>
<td>1.992664</td>
<td>0.057336</td>
</tr>
<tr>
<td>22</td>
<td>IY</td>
<td>5.43</td>
<td>4.720289</td>
<td>0.709711</td>
<td>IT</td>
<td>1.49</td>
<td>1.49369</td>
<td>-0.00369</td>
</tr>
<tr>
<td>23</td>
<td>IW</td>
<td>5.7</td>
<td>5.366469</td>
<td>0.333531</td>
<td>IS</td>
<td>1.49</td>
<td>1.425665</td>
<td>0.064335</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
<td></td>
</tr>
<tr>
<td>----</td>
<td>-----</td>
<td>-----</td>
<td>---------</td>
<td>---------</td>
<td>---------</td>
<td>---------</td>
<td>---------</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Index</td>
<td>DTlDeM_p/d2GP_</td>
<td>lnDGsDeE_1/p_GE_</td>
<td>DTJlDeM_p/d2GP_</td>
<td>lnDTlDeEp2/d2AE_</td>
<td>DTsDeP_1/d_GP2</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>r</td>
<td>0.78192</td>
<td>0.88696</td>
<td>0.78843</td>
<td>0.87536</td>
<td>0.79228</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>r²</td>
<td>0.61140</td>
<td>0.78670</td>
<td>0.62162</td>
<td>0.76626</td>
<td>0.62770</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>s</td>
<td>0.630</td>
<td>0.471</td>
<td>0.471</td>
<td>0.493</td>
<td>0.616</td>
</tr>
</tbody>
</table>

**TABLE II.** Statistics for ACE inhibitors set.
Table II collects the statistics of monovariate and bivariate regression in modeling the ACE inhibiting potency of dipeptides by FPIF. Cross-validation tests (Leave-20%-out L20%o or Leave-one-out Loo procedures) are given here only for bivariate regressions.

The best-found model was:

$$BA_{\text{calc}} = 35.992 + 0.822 * \frac{DTfDiM_p}{d2GP} - 11.802 * \ln \frac{DGsDiE_1}{p GE}$$

$$n = 58; r = 0.88696; s = 0.471; F = 101.426$$ (32)

Both topology ($T$ - in the index symbol) and geometry ($G$) contribute to the best model. As local property, the atomic mass ($M$) and electronegativity ($E$) modulate the structure-activity relationship.

For the best model (see also column 3, Table II) the L20%o cross-validation was averaged on 25 randomly chosen 20% objects. The drop in $r$ is around 1.6 % that proves a good predicting ability of the models. The plot of observed $BA$ vs calculated $BA$ (eq 33) is presented in Figure 3.

The model given by eq 32 is superior, both in estimation and prediction, to those reported in literature (see Table III). Note that the Zaliani's results refer both to a single conformation (i.e. extended) of amino acids and to a library conformation family (i.e. rotameric). The correlation recorded in case of extended conformation surpasses the correlation obtained in rotameric case.
Figure 3. The plot of observed vs calculated BA (eq 32).

### TABLE III. Comparative statistics of QSAR models of 58 ACE inhibitors and 48 sweeteners dipeptides

<table>
<thead>
<tr>
<th>No</th>
<th>Peptide Set (Reference)</th>
<th>Descriptors per Residue</th>
<th>No. Components</th>
<th>$r^2$ (fitting)</th>
<th>$r^2$ (cross-validated)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>ACE (Cocchi et al.) [6]</td>
<td>7</td>
<td>1</td>
<td>0.744</td>
<td>nd$^{a}$</td>
</tr>
<tr>
<td>2</td>
<td>ACE (Collantes et al.) [8]</td>
<td>2</td>
<td>nd</td>
<td>0.700</td>
<td>nd</td>
</tr>
<tr>
<td>3</td>
<td>ACE (Zaliani et al. -extended) [5]</td>
<td>3</td>
<td>2</td>
<td>0.708</td>
<td>0.637</td>
</tr>
<tr>
<td>4</td>
<td>ACE (Zaliani et al. -rotameric) [5]</td>
<td>3</td>
<td>6</td>
<td>0.657</td>
<td>0.541</td>
</tr>
<tr>
<td>5</td>
<td>ACE (<em>FPIF</em>) [this work]</td>
<td>2</td>
<td>2</td>
<td><strong>0.787</strong></td>
<td><strong>0.759$^{b}$</strong></td>
</tr>
<tr>
<td>6</td>
<td>Sweeteners (Jonsson et al.) [7]</td>
<td>3</td>
<td>1</td>
<td>nd</td>
<td>0.780</td>
</tr>
<tr>
<td>7</td>
<td>Sweeteners (Collantes et al.) [8]</td>
<td>2</td>
<td>2</td>
<td>0.847</td>
<td>nd</td>
</tr>
<tr>
<td>8</td>
<td>Sweeteners (Zalini et al. - extended) [5]</td>
<td>3</td>
<td>3</td>
<td>0.754</td>
<td>0.710</td>
</tr>
</tbody>
</table>
In general, a model is built up by using a training set of structures (that provides a calibration equation) and further it is validated by a cross-validation procedure and also by using an external prediction set. Due to the fixed mode of selection, the Loo procedure strongly requires an external set for prediction. It is not the case of averaged L20%o procedure, when the predicting sets (and implicitly the 80% training sets) can be randomly selected, thus getting enough statistical meaning for the model. A similar procedure was used in Zaliani’s report [5].

Table IV shows the occurrence of descriptors in the best 10 regression equations. All indices of the first variable in bivariate regression are topological (T in index symbol) while only six of ten of the second variable are geometric (G in index symbol). This result correlates with the Zaliani’s best result when used extended conformations (see Table III).

As local property, the atomic mass (M) occurs five times in the first variable while the electronegativity (E) seven times in the second variable. Other occurring properties are the partial charge (P) and cardinality (C). Clearly, the chemical features play an important role in their occurrence, which is also consistent with the Zaliani’s best result.

<table>
<thead>
<tr>
<th>No.</th>
<th>Score</th>
<th>Score 2</th>
<th>Index 1</th>
<th>Index 2</th>
<th>( r )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>89</td>
<td>5831</td>
<td>DTfDiM_p/d2GP_</td>
<td>InDGsDiE_1/p_GE_</td>
<td>0.88696</td>
</tr>
<tr>
<td>2</td>
<td>54</td>
<td>1771</td>
<td>DTjDeM_p/d2GP_</td>
<td>InDTjDeEp2/d2AE_</td>
<td>0.87536</td>
</tr>
<tr>
<td>3</td>
<td>29</td>
<td>7894</td>
<td>DTsDeP_1/d_GP2</td>
<td>InRGsDeMp2/d2AE_</td>
<td>0.87171</td>
</tr>
<tr>
<td>4</td>
<td>29</td>
<td>2644</td>
<td>DTsDeP_1/d_GP2</td>
<td>InRTjDeEp2/d2AE_</td>
<td>0.86856</td>
</tr>
<tr>
<td>5</td>
<td>18</td>
<td>8213</td>
<td>DTsDeM_p/d_GP2</td>
<td>InRGsDeEp2/d2AE_</td>
<td>0.86812</td>
</tr>
<tr>
<td>6</td>
<td>18</td>
<td>7725</td>
<td>DTsDeM_p/d_GP2</td>
<td>InRGsDeE_p/d2AE_</td>
<td>0.86237</td>
</tr>
<tr>
<td>7</td>
<td>15</td>
<td>6476</td>
<td>DTfDeM_p/d_PP_</td>
<td>InRTsDiEp2/d2AE2</td>
<td>0.86181</td>
</tr>
<tr>
<td>8</td>
<td>1</td>
<td>15876</td>
<td>RTfDeE_1/p_PP2</td>
<td>InRGsDeCp2/d2HP2</td>
<td>0.86136</td>
</tr>
<tr>
<td>9</td>
<td>1</td>
<td>8719</td>
<td>RTfDeE_1/p_PP2</td>
<td>DGfDiP_p/d_GP_</td>
<td>0.85178</td>
</tr>
<tr>
<td>10</td>
<td>1</td>
<td>6485</td>
<td>RTfDeE_1/p_PP2</td>
<td>InRTsDiEp2/d2GE2</td>
<td>0.84654</td>
</tr>
</tbody>
</table>
discriminating vertices (i.e. atoms or atom groups), fragments and whole molecules of dipeptides. They are strongly involved in modeling the biological activity of dipeptide ACE inhibitors.

2. Dipeptide Sweeteners. The set including 48 dipeptides was taken from Jonsson's paper [7]. The molecular structures were input and optimized by using MM+ and then by semiempirical AM1 procedure of the HyperChem Program (HyperCube Inc.). Table I includes the dipeptide names by using the one-letter code for aminoacids, the observed bitter tasting activity (biological activity, BA, as log(1/T)), the calculated BA (according to eq 33) and the corresponding residuals.

Table V collects the statistics of monovariate and bivariate regression in modeling BA of dipeptide sweeteners by FPIF. The same remark holds for the cross-validation tests.

The best-found model was:

\[
BA_{\text{calc}} = 1.142 + 0.474 \cdot \frac{RT_{DiM}}{p_{SP}} - 0.043 \cdot \frac{DG_{DiE}}{p_{AP}}
\]

\(n = 48; r = 0.92272; s = 0.248; F = 128.922\) (33)

As in the previous test, both topology and geometry contribute to the best model and again the local property, was the atomic mass (\(M\)) and electronegativity (\(E\)).

In predicting tests, (see Table V, columns 3, 5 and 7) the drop in \(r\) was around 1 %, proving a good stability of the models. The plot of observed BA vs calculated BA (eq 33) is presented in Figure 4.

The model given by eq 33 surpasses those reported in literature (see Table III).
Figure 4. The plot of observed vs calculated BA (eq 33).

TABLE V. Statistics for sweeteners dipeptides.

<table>
<thead>
<tr>
<th>Index</th>
<th>$r$</th>
<th>$r^2$</th>
<th>$s$</th>
<th>$F$</th>
<th>$b_0$</th>
<th>$b_1$</th>
<th>$b_2$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.81448</td>
<td>0.66337</td>
<td>0.369</td>
<td>90.650</td>
<td>0.106</td>
<td>0.898</td>
<td>4.479</td>
</tr>
<tr>
<td></td>
<td>0.91688</td>
<td>0.84067</td>
<td>0.257</td>
<td>118.714</td>
<td>0.482</td>
<td>-0.058</td>
<td>1.364</td>
</tr>
<tr>
<td></td>
<td>0.80079</td>
<td>0.64126</td>
<td>0.381</td>
<td>82.231</td>
<td>0.115</td>
<td>0.315</td>
<td>0.315</td>
</tr>
<tr>
<td></td>
<td>0.91525</td>
<td>0.83768</td>
<td>0.259</td>
<td>116.116</td>
<td>0.471</td>
<td>-0.006</td>
<td>1.364</td>
</tr>
<tr>
<td></td>
<td>0.77192</td>
<td>0.59586</td>
<td>0.404</td>
<td>67.821</td>
<td>-0.058</td>
<td>0.085</td>
<td>0.474</td>
</tr>
</tbody>
</table>

Cross-validated

<table>
<thead>
<tr>
<th>$r$</th>
<th>$r^2$</th>
<th>$s$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.90670</td>
<td>0.82210</td>
<td>0.268</td>
</tr>
</tbody>
</table>

$L20\%_o$ (aver.)

<table>
<thead>
<tr>
<th>$r$</th>
<th>$r^2$</th>
<th>$s$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.90473</td>
<td>0.81854</td>
<td>0.271</td>
</tr>
<tr>
<td>0.91286</td>
<td>0.83331</td>
<td>0.259</td>
</tr>
</tbody>
</table>

$a$ average of twenty five 20% sets of randomly chosen objects.
Table VI shows the occurrence of descriptors in the best 10 regression equations. Seventeen indices in bivariate regression are topological while only three geometric. This result proves that the topology is the main feature in describing this dipeptide activity. In fact, topological indices are descriptors invariant to rototranslation, so that it is not surprising that Zaliani obtained the best correlation when used extended conformations of amino acids (see Table III).

As local property, the electronegativity ($E$) occurs nineteen times while the atomic mass ($M$) only once, in bivariate regression.

It appears that the bitter tasting activity is controlled by electronic factors.

Concluding Remarks

The fragmental property indices take into account the chemical nature of atoms (mass, electronegativity and partial charge), various kinds of interactions between the fragments of molecules as generated by Cluj and Szeged criteria and the 3D geometry of molecular structures as well.

There exist an analogy between CoMFA and FPIF: both of them calculate the interaction of a chemical structure (or substructure) with a probe atom in the 3D space. The fragmental property $Fr_{ij}$ is viewed as an interaction of atoms forming the fragment $Fr_{ij}$ with the atom $j$. The major difference is that in CoMFA the probe atoms (with well defined chemical identity) is
external whereas FPIF considers internal probe atoms with no chemical identity. Only the fragments (i.e. substructures) are chemically well defined.

FPIF offer good description and modeling of dipeptides activity, such as the ACE inhibition or bitter tasting. As it is known, a correlational model does not involve a causal relationship between descriptors and a molecular property. However, a look upon the occurrence of indices with some best scores (and implicitly best structure description) can highlight some aspects of intra- and/or intermolecular interactions.

The above results demonstrate the usefulness of our descriptors in modeling peptide structures and properties. For other FPIF modeling examples the reader can consult [14].

Acknowledgement

This work in under financial support of GRANT CNSIS, T 34, 2000. The authors are grateful to the referees for their valuable comments.

References


