Available online at www.sciencedirect.com
SCIENCE DIRECT®
Bioorganic & Medicinal Chemistry 12 (2004) 4815–4822
Bioorganic & Medicinal Chemistry

Stochastic-based descriptors studying peptides biological properties: modeling the bitter tasting threshold of dipeptides

Ronal Ramos de Armas,[a,b,*] Humberto González Díaz,[a,c] Reinaldo Molina,[a,d] Maykel Pérez González[e] and Eugenio Uriarte[c]

[a]*Chemical Bioactives Center, Central University of 'Las Villas', 54830, Cuba*
[b]*Department of Chemistry, Central University of 'Las Villas' 54830, Cuba*
[c]*Department of Organic Chemistry, Faculty of Pharmacy, University of Santiago de Compostela 15706, Spain*
[d]*Universität Rostock, FB Chemie, Albert-Einstein-Str. 3a, D 18059 Rostock, Germany*
[e]*Experimental Sugar Cane Center Villa Clara-Cienfuegos, Cuba*

Received 23 April 2004; revised 5 July 2004; accepted 7 July 2004
Available online 5 August 2004

Abstract—MARCH-INSIDE methodology was applied to the prediction of the bitter tasting threshold of 48 dipeptides by means of pattern recognition techniques, in this case linear discriminant analysis (LDA), and regression methods. The LDA models yielded a percentage of good classification higher than 80% with the two main families of descriptor generated by this methodology (95.8% for self return probability and 83.3% using electronic delocalization entropy). The regression models can explain more than 80% of the experimental variance of the independent variable. Two regression models were obtained with R^2 values of 0.82 and 0.88 for the whole data and the data without two outliers, respectively; having a standard deviation of 0.27 and 0.23. The predictive power of the obtained equations was assessed by the Leave-One-Out cross validation procedures, giving the same percentages of good classification as in the training set, in the LDA models, and yielding values of q^2 of 0.78 and 0.86 in the regression model, respectively. The validation of this methodology was also carried out by comparison with previous reports modeling this data with other well-known methodologies, even 3-D molecular descriptors.

1. Introduction

The sense of taste plays an important role in life and nutritional status of human beings and other organisms. This perception can be categorized into four well-known and accepted descriptors: sweet, bitter, salty and sour and two more controversial categories: fat taste and amino acid taste. The ability to identify the sweet-tasting foodstuffs is particularly important, giving the vertebrates an important tool for seeking our necessary, highly nutritive and energetic carbohydrates. The perception of bitter taste, on the other hand, protects us avoiding potentially poisonous plants and other environmental toxins.[1]

Nowadays, the development of low-calories sweeteners substituting sugar is very important and can solve several problems related to modern medicine and nutrition. For instance, these sweeteners can be employed in diabetes and obesity management allowing the control and prevention of many other diseases.[2]

The mechanism of transduction of bitter and sweet taste comprises complex interactions involving G-protein coupled receptors in the taste receptor cells (TCR)[3,4] and a number of G-proteins deriving in several intracellular signals.[5,6]

This complex system cannot be reproduced in vitro in a successful way yet, so in order to taste a new sweetener several panels of taster are used and these procedures are not always as reproducible and confidents as we wish, yielding a time–money consuming and not always accurate procedure. In these sense, quantitative structure properties relationships (QSPR) aim to find quantitative models relating a given property with molecular structure,[7] providing a useful tool to solve the ancestral problem of the *TRIAL AND ERROR* system for selecting a compound with a desired property.[8]

Keywords: Stochastic process; Markov chain; QSPR; Bitter taste; Dipeptides.
* Corresponding author. Tel.: +53-42281473; fax: +53-42281230; e-mail addresses: ronalr@qf.uclv.edu.cu; ronal@medinews.com

0968-0896/$ - see front matter © 2004 Elsevier Ltd. All rights reserved.
doi:10.1016/j.bmc.2004.07.017

Several QSAR studies have been done related to the prediction of the taste of a compound. A very extend and complete review of these subject was given by Katritzky et al.[2] According to this author, the main classes of sweeteners studied by QSPR are dihydrochalcones,[9] perillartines,[2,10–13] sulfamates,[2,14–17] five-membered rings and diverse compounds.[2,14] In these studies, several recurrent molecular descriptors appeared such as constitutional descriptors (number of oxygen atoms),[2] descriptors derived from quantum mechanic (Coulombic interactions, reactivity indices related to atoms and electrons,[2] antibinding contributions,[2] electrophilic parameters,[2] describing hydrophobicity ($\log P$, STERIMOL parameters),[13,16] accessible area by different kinds of atoms and solvents,[2] connectivity indices.[10–12,15]

Peptides were also object of great attention studying their taste properties. This property depends greatly on the stereochemistry of the peptidic backbone.[2] The QSPR studies reported previously by this author showed the influence of energetic factors (total enthalpy), Zefirov maximum partial charge number of chlorine atoms and net charge over the nitrogen atom. The diversity of conformation of these compounds is the main cause for the lack of good correlation parameters (at least not as good as the former reports with other families of chemicals) in these studies.

One of the first reports related to these compounds and their sweet taste prediction was found in the Benson's work[18] studying sulfamate sweeteners. Rodriguez et al.[19,20] found significant electronic and topologic contribution as well the size of the C-terminal residue in a series of chemically modified tripeptides. Miyashita et al.[21] applied a pattern recognition technique yielding an 87% and 81% of good classification for the sweet and bitter peptides, respectively, employing connectivity indices and STERIMOL parameters.

Asao was one of the first modeling the bitter tasting threshold of dipeptides employing hydrophobic and steric parameters reported for the side chain residues. The length of the zig-zag dipeptide's conformation showed to be an important factor modeling this property. The importance of hydrophobicity was also reported by Tamura et al.[22] in a later report. Spillane et al.[14] also reported the influence of apparent specific volume, apparent molar volume related to the sweetener power of peptides.

For a data of 48 dipeptides, their bitter tasting threshold was determined.[23,24] These data have been widely employed assessing new descriptors for describing peptides and protein properties. Collantes and Dunn[24] introduced their *ISA* and *ECI* descriptors (isotropic surface area and electronic charge index, respectively) modeling this data yielding a correlation coefficient 0.848 and an error of 0.24. They used PLS as statistical technique. The *z*-scale was also employed getting in this case regression coefficients of 0.71 and 0.82 and errors of 0.34 and 0.26. These later studies showed the influence of hydrophobic residues in both positions increasing the bitter taste, as well as polar residues in position 1 (Asp, Arg).[24,25]

Other methodologies have been applied to this data such as the MAPS methodology[23] obtaining a two-component model, extended *z*-scale[26] with a one component PLS model explaining the 78% of the total experimental variance, Cluj indices,[27] descriptors based on holographic distance vectors[28] providing a regression coefficient of 0.91 as well as a q^2 value of 0.86 and 3-D descriptors such as MS-WHIM[25] with an R^2 value of 0.824. All these reports reinforce the thesis of hydrophobicity and size as the main factor affecting the bitter taste properties of these compounds.

The **MARCH-INSIDE** (**Mar**kovian **Ch**emicals **In Si**lico **De**sign) methodology has been developed by our research group to generate molecular descriptors based on the Markov Chain Theory. This approach has been successfully employed in QSPR and QSAR studies, including studies related to proteomics and nucleic acid–drug interactions. The approach describes changes in the electron distribution and vibrational decay with time throughout the molecular backbone. The method allowed us to introduce physically meaningful stochastic graph invariants for the study molecular properties. The method has also demonstrated flexibility in relation to many different problems. One of the applications involved the prediction of the fluckicidal activity of novel drugs (flukes are tiny intestinal parasites)[29] More recently, the MARCH-INSIDE approach has been applied to the fast-track experimental discovery of novel anticancer compounds.[30] Additionally, promising results have been found in the modeling of the interaction between drugs and HIV-packaging-region RNA in the field of bioinformatics.[31] An alternative formulation of our approach in terms of negentropies gives more physical sense to our models for drug–RNA interactions.[32] The prediction of the biological activities of peptides and NMR shifts in proteins are problems that can also be addressed using this approach.[33] Codification of chirality and other 3-D structural features constitutes another advantage of this method[34] applied in the estimation of the level of agranulocytocis chemically induced by drugs.[35]

A precise definition of the descriptors generated by this methodology can be found in several reports of its application in the study of several biological properties.[29–35] Briefly we can say that MARCH-INSIDE methodology considers as states of the MCH the electrons layers of any atom in the molecule. The method uses as source of molecular descriptor the ${}^1\Pi$ matrix (the one-step electron-transition stochastic matrix) built up as a squared matrix $n \times n$ (n number of atoms in the molecule) whose elements (p_{ij}) are calculated as the ratio between the withdrawing of the jth atom and the sum over all the atoms covalently linked to the ith atom. Also a new matrix, ${}^A\Pi_k$ matrix, can be defined as the product of a $1 \times n$ vector (${}^A\Pi_0$) whose elements (${}^A\pi_k(j)$) are calculated in the similar way as p_{ij} but the sum is carried out now over the all the atoms in the molecule and the ${}^k\Pi$, which is the kth power of ${}^1\Pi$matrix.

These matrices can then be used to generate three families of molecular descriptors:

Self return probabilities ($^{SR}\pi_k$): it can be defined as the trace (sum over the p_{ii} values) of the kth power of the $^{1}\Pi$matrix.

$$^{SR}\pi_k(S_m) = \sum_{i=1}^{g} {}^{k}p_{ii} = \mathrm{Tr}[(^{1}\Pi)^{k}] \quad (1)$$

Codify the attraction of an atom or group of atoms for its electrons (the electrons that were at the atom or the given group of atoms in the time t_0) at any time t_k located at kth steps away or less.

Absolute probabilities ($^{Abs}\pi_k(j)$): Codify the attraction of the jth atom over any electron in the molecule at any time t_k after traveling by different paths of less than k steps.

Electronic delocalization entropy (Θ_k):

$$\Theta_k = \sum_{j=1}^{n} \Theta_k(j) = -\sum_{j=1}^{n} [^{A}\pi_k(j)] \ln[^{A}\pi_k(j)] \quad (2)$$

It describes the entropy involved in the electron attraction at least k steps beginning with the j atom.

This article is aimed to apply the MARCH-INSIDE methodology to the prediction and study of the bitter tasting threshold of 48 dipeptides by means of pattern recognition techniques (in these case linear discriminant analysis) and regression methods. The results obtained in this study will be compared with the previous reports of QSPR studies related to this data of dipeptides. In addition, several well-known families of molecular descriptors were calculated for this data and applied to the prediction of this property in order to compare the accuracy of our methodology with other well-known methodologies usually employed in QSAR studies.

2. Results

2.1. Linear discriminant analysis

In order to determine the limits between both groups in the LDA analysis a piece wise linear regression was developed explaining in this case the 87.4% of the experimental variance with a break point value of pT = 1.98. This model set a clear differentiation between two groups around the pT value of 2.0 as can be seen in Figure 1.

Afterwards, the application of the linear discriminant analysis yielded the following models:

$$\text{Bitter taste} = 22.99 \cdot \Theta_3 - 51.16 \quad (3)$$

$$N = 48 \quad \lambda = 0.49 \quad F = 47.62 \quad D^2 = 4.15 \quad p < 0.0001$$

$$\text{Bitter taste} = 4.64 \cdot {}^{SR}\pi_0 - 20.39 \cdot {}^{SR}\pi_3 - 17.19 \quad (4)$$

$$N = 48 \quad \lambda = 0.33 \quad F = 46.5 \quad D^2 = 8.28 \quad p < 0.0001$$

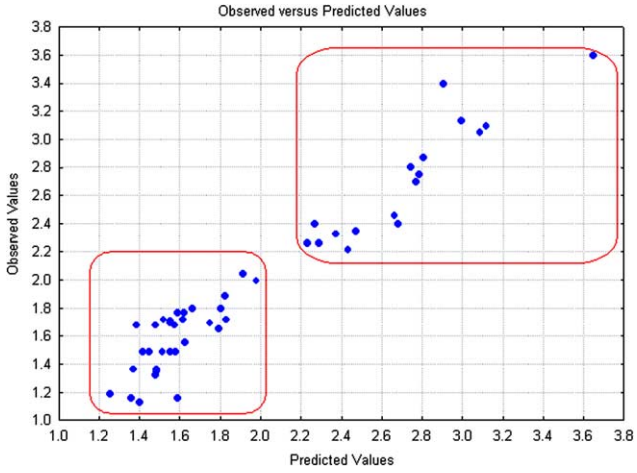

Figure 1. Piece wise linear regression method sets a clear differentiation between less and more bitter compounds around the pT value of 2.0.

Eq. 3 classified correctly the 89.3% (25/28) of the compounds having a log(1/T) lower than 2.0 (less bitter) and the 78.9% (15/19) of the compounds with a log(1/T) higher than 2.0 (more bitter); giving an overall percentage of good classification of 83.3 (40/48) with only one compound as nonclassified (differential percentage of posterior probability < 5). In the case of Eq. 4 the figures were: 93.1% (27/29) for the 'less bitter' group, 100% (19/19) for the 'bitterer' group with an overall percentage of 95.8 (46/48) and the absence of nonclassified compounds. All these figures are depicted in Table 1.

Analyzing the descriptors in the above equations highlights that, according to Eq. 3, all factors that increase the electronic delocalization entropy will increase the bitter taste of these compounds; it means, residues whose chemical structure gives a higher absolute probability of electronic delocalization such as aromatic residues (Trp, Phe, Tyr, His). These findings agree with previous reports[24] indicating that aromatic residues increase the bitter taste of dipeptides.

In the case of Eq. 4, the combination of a positive contribution of $^{SR}\pi_0$ and a negative contribution of $^{SR}\pi_3$, points to a hydrophobic interaction ruling the process. The higher the size of the residue (evaluated by $^{SR}\pi_0$) and the lower the value of higher range self-return probabilities, indicate that the electronegativity of the atoms in the residue will decrease and hence the possibility of electrostatic interaction will decrease in the same proportion allowing only hydrophobic interaction to take place. This hydrophobic contribution was also reported in several previous works about bitter taste QSPR studies.[13,19,21,24,25]

Afterward a Leave-One-Out cross validation procedure was developed. The same percentages of good classification were obtained as for the training set showing the stability and predictive power of both equations. The differential posterior probabilities for both equations and for the cross validation LOO are depicted in Table 2.

Table 1. Percentages of good classification according to Eqs. 3 and 4

	Eq. 3				Eq. 4		
	%	Less bitter	More bitter	NC	%	Less bitter	More bitter
Less bitter	89.3	25	3	1	93.1	27	2
More bitter	78.9	4	15	0	100.0	0	19
Total	83.3	29	18	1	95.8	27	21

Table 2. Posterior probabilities in classification[a] and LOO cross validation[b] according to Eqs. 3 and 4; as well as the observed values[c] and predicted in multivariate linear regression according to Eqs. 7,[d] 8[e] and PLS model[f]

Peptide	Linear discriminant analysis				Multiple linear regression			
	Eq. 3		Eq. 4		Obs.[c]	Pred. Ec. 7[d]	Pred. Ec. 8[e]	Pred. PLS[f]
	Prob. Pred.[a]	CV[b]	Prob. Pred.	CV				
GV	−99.0	−99.0	−100.0	−100.0	1.13	1.12	1.09	1.12
GL	−92.1	−92.1	−99.6	−99.7	1.68	1.51	1.48	1.58
GI	−92.1	−92.1	−99.6	−99.7	1.70	1.45	1.42	1.54
GP	−99.8	−99.8	−99.6	−99.7	1.35	1.39	1.42	1.30
GF	−90.5	−90.5	−99.6	−99.7	1.80	1.78	1.77	1.87
GW	−47.5	−47.5	−52.4	−57.7	1.89	2.22	2.24	2.33
GY	−89.5	−89.5	−99.9	−100.0	1.77	1.64	1.64	1.73
AV	−83.6	−83.6	−99.8	−99.8	1.16	1.49	1.44	1.50
AL	−23.2	−23.2	−96.6	−97.3	1.70	1.86	1.81	1.91
AF	−15.7	−15.7	−97.2	−97.8	1.72	2.13	2.10	2.19
VG	−99.3	−99.3	−100.0	−100.0	1.19	0.92	0.95	0.82
VA	−95.1	−95.1	−99.8	−99.8	1.16	1.16	1.18	1.12
VV	−6.1	−6.1	−76.1	−79.6	1.71	1.65	1.67	1.68
VL	61.8	61.8	36.7	31.7	2.00	1.96	1.99	2.00
LG	−93.5	−93.5	−99.5	−99.7	1.72	1.45	1.44	1.43
LA	−66.2	−66.2	−96.6	−97.3	1.72	1.68	1.66	1.69
LL	91.8	91.8	94.3	94.0	2.35	2.44	2.45	2.50
LF	92.2	92.2	92.9	92.5	2.75	2.69	2.73	2.76
LW	92.5	92.0	99.4	99.4	3.40	2.76	2.82	2.79
LY	92.2	92.2	65.1	61.9	2.46	2.54	2.58	2.58
IG	−95.5	−95.5	−99.6	−99.7	1.68	1.24	1.26	1.20
IA	−75.9	−75.9	−96.8	−97.4	1.68	1.46	1.48	1.45
IV	55.8	55.0	33.6	28.5	2.05	1.90	1.93	1.94
IL	87.4	86.7	93.9	93.6	2.26	2.20	2.24	2.24
II	87.4	86.7	93.5	93.2	2.26	2.14	2.18	2.19
IP	−8.3*	−8.0*	93.1	92.9	2.40	2.25	2.31	2.24
IW	96.7	96.4	100.0	100.0	3.05	2.80	2.91	2.78
IN	−30.1	−29.4	−98.6	−99.0	1.49	1.48	1.51	1.53
ID	−60.8	−59.8	−99.8	−99.9	1.37	1.31	1.34	1.37
IQ	43.1*	42.4*	−81.1	−84.2	1.49	1.77	1.81	1.83
IE	9.0*	8.9*	−97.3	−97.9	1.37	1.60	1.65	1.68
IK	93.3*	92.9*	98.5*	98.5*	1.65	2.38	Outlier	1.38
IS	−74.8	−73.9	−99.0	−99.2	1.49	1.36	1.38	1.65
IT	−13.7	−13.3	−92.8	−94.2	1.49	1.61	1.64	1.45
PA	−99.5	−99.4	−93.1	−94.3	1.32	1.55	1.63	2.40
PL	−27.4*	−26.8*	97.2	97.2	2.22	2.38	2.47	2.36
PI	−27.4*	−26.8*	97.0	97.0	2.33	2.33	2.41	2.67
PY	−21.9	−21.4	81.4*	80.2*	1.80	2.48	Outlier	1.75
PF	−23.0*	−22.4*	96.5	96.4	2.80	2.63	2.75	2.80
FG	−93.1	−92.6	−99.6	−99.7	1.77	1.74	1.76	2.80
FL	91.0	90.5	92.9	92.5	2.87	2.72	2.75	3.06
FP	10.6	10.5	92.0	91.7	2.70	2.76	2.82	2.88
FF	91.5	91.0	91.3	90.7	3.10	2.97	3.03	2.13
FY	91.5	91.0	58.4	54.4	3.13	2.82	2.88	3.20
WE	−0.4**	−0.3**	−57.0	−62.7	1.56	2.06	2.14	2.63
WW	94.8	94.4	100.0	100.0	3.60	3.23	3.39	1.67
YL	90.9	90.4	65.1	61.9	2.40	2.56	2.60	1.12
SL	−63.3	−62.3	−98.9	−99.2	1.49	1.62	1.60	1.58

*Miss-classified compound.
**Non-classified compound.

Most of the mis-classified and nonclassified compounds had proline and isoleucine residues. Taking these peptides out of the data the percentages and statistical parameter improved considerably to 95.0 and 100 for electronic delocalization entropies and self-return probabilities, respectively.

No reports, until our concern, were found related to the application of pattern recognition techniques to these data of peptides. Only the works of Kier[12] with an 85% of good classification for the training set and 89% for the prediction set of perillartines and Miyashita's work[21] yielding 87% and 81% for different groups of L-aspartyl-dipeptides.

2.2. Linear regression models

For the quantitative prediction of the bitter tasting threshold, the best equation found by means of multivariate linear regression were

$$\log\left(\frac{1}{T}\right) = 3.98 \cdot \Theta_1 - 6.67 \qquad (5)$$

$$N = 48 \quad R^2 = 0.61 \quad s = 0.40 \quad F = 71.4$$

$$\log\left(\frac{1}{T}\right) = 0.08 \cdot {}^{SR}\pi_0 - 0.90 \qquad (6)$$

$$N = 48 \quad R^2 = 0.58 \quad s = 0.40 \quad F = 71.4$$

Due to the poor quality of both models we try to get a new equation combining both sets of descriptors, yielding the following equation:

$$\log\left(\frac{1}{T}\right) = 3.13 \cdot \Theta_1 + 0.19 \cdot {}^{SR}\pi_0 - 1.79 \cdot {}^{SR}\pi_{10} - 4.39 \qquad (7)$$

$$N = 48 \quad R^2 = 0.82 \quad s = 0.27 \quad F = 68.0 \quad q^2 = 0.78$$

As can be seen from the statistical parameters of the above equation a considerable improvement was achieved by combining both descriptors. Eq. 7 can explain the 82% of the experimental variance of the dependent variable (20% more than Eqs. 5 and 6) with almost the half of the standard error of determination and the same F value. Despite having three times more variables the ratio number of cases/number of variables gives an excellent value of 16 (should be higher than 5 for linear regression methods).

A careful look at the variables in the equation resembles the same influence of the above explained equations (Eqs. 3 and 4) with a positive contribution for the entropic variables and ${}^{SR}\pi_0$ while the ${}^{SR}\pi_k$ has a negative contribution indicating the importance of hydrophobic characteristic of the peptide and the presence of aromatic residues in bitter-tasting characteristic of a given dipeptide.

From this data, after the application several statistical techniques for the detection of outliers, two compounds were extracted as outliers (in this case IK and PY). These compounds were also misclassified by our previous classification models (EQ) at least once (see Table 2) and were reported also as outliers by Liu et al.[28] using molecular holographic distance vector (MHDV) descriptors. No structural explanation can be found so far for this regular behavior. It seems that dipeptides with isoleucine and proline in the position 1 have higher probability of being misclassified or an outlier (see Table 2). After that, a new linear regression analysis was performed and the equation obtained was

$$\log\left(\frac{1}{T}\right) = 2.49 \cdot \Theta_1 + 0.21 \cdot {}^{SR}\pi_0 - 1.91 \cdot {}^{SR}\pi_{10} - 3.50 \qquad (8)$$

$$R^2 = 0.88 \quad s = 0.23 \quad F = 104.5 \quad q^2 = 0.86$$

The introduction of a fourth variable does not improve considerably the correlation as can be seen in Table 3.

Again the same variables appeared in the model with the same contribution as in the former model described in this paper. The observed and predicted values of log(1/T) according to Eqs. 7 and 8 are shown in Table 2 and Figure 1.

A partial least squares model was obtained with the whole set of descriptors (22 columns X-matrix). The model explained 85.3% of the experimental variance with three components that account for the 56.3%, 26.3% and 5.9% of the variance, respectively, with an error of estimation of 0.226. These values were also reported in Table 2 and Figures 2–4.

When we compared the above result with the previously reported studies for this data (Table 4) we concluded that despite the use of strongest statistical techniques such as partial least squares (PLS) and principal component regression (PCR) and well-known molecular descriptors (even 3-D descriptors), MARCH-INSIDE methodology reach to similar results employing multiple linear regression (MLR) methods, allowing a better interpretation of the influence of each variable in the taste of the peptides. Besides, the PLS model obtained by these methodology gives results as significant as the previously reported employing this technique.[24]

Table 3. Forward stepwise analysis in the linear regression analysis obtaining Eq. 8

Step	Variables	R^2	S	F
1	Θ_1	0.649	0.385	80.7
2	Θ_1 ${}^{SR}\pi_{10}$	0.687	0.367	47.1
3	Θ_1 ${}^{SR}\pi_{10}$ ${}^{SR}\pi_0$	0.881	0.228	104.5
4	Θ_1 ${}^{SR}\pi_{10}$ ${}^{SR}\pi_0$ ${}^{SR}\pi_1$	0.893	0.220	85.16

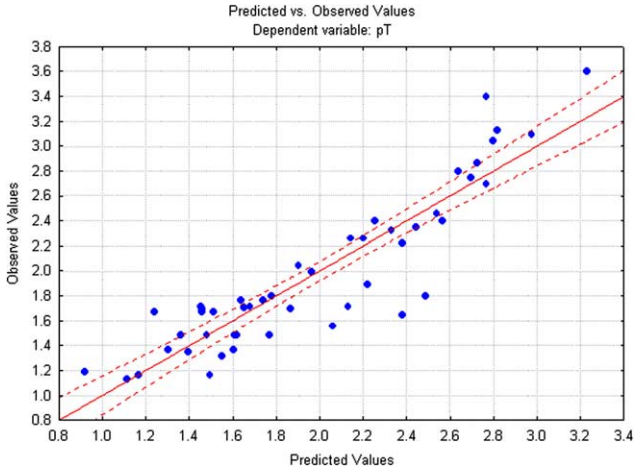

Figure 2. Plots of observed versus predicted values according to Eq. 7.

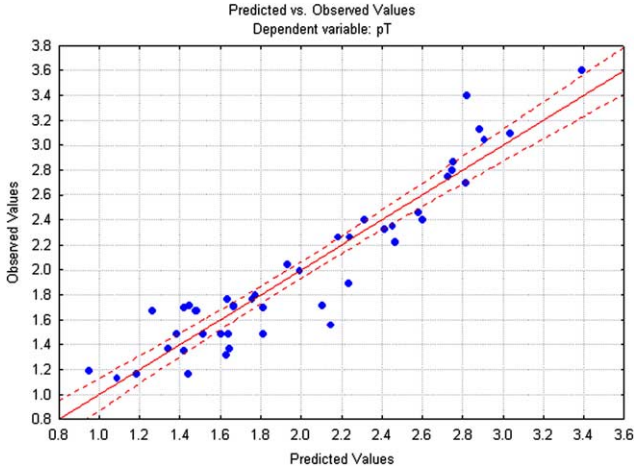

Figure 3. Plots of observed versus predicted values according to Eq. 8.

Finally 12 families of molecular descriptors[7] were generated with the DRAGON software and a multiple linear regression analysis was developed yielding the results summarized in Table 5. These results are another proof of the applicability of MARCH-INSIDE methodology in the study of this kind of compounds. Even when some descriptors have better statistical parameters modeling this property they employed more variables than our models (topologic descriptors and 3-D molecular descriptors such as RDF, 3-D MORSE y WHIM).

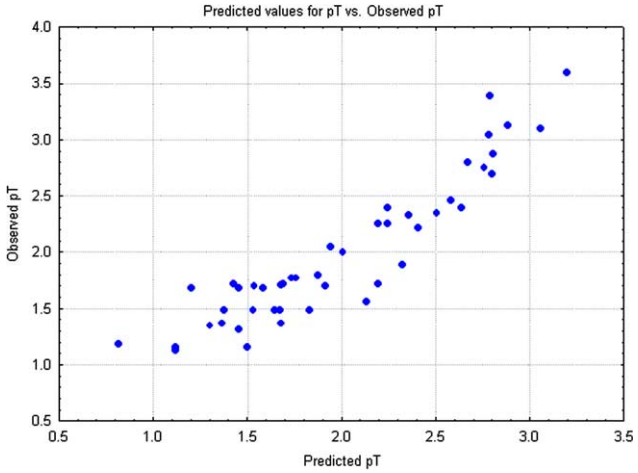

Figure 4. Plots of observed versus predicted values according to the PLS model.

3. Conclusion

The MARCH-INSIDE methodology applied to peptides gives a powerful tool to study and predict the biological activities of this family of chemicals. These approach allow us to predict even in a qualitative and quantitative way many properties, in these case modeling the bitter tasting threshold of dipeptides, having several advantages referring to the former reports such as: it permits the determination of local descriptors over important residues (which could be focused in a forthcoming paper); this approach also makes use of linear regression method, which permits a simple and direct way of analyze the results and the influence of the variables in the model in a better way than partial least squares (PLS) and principal component regression (PCR). The results obtained can be compared with those reported previously in the literature with the later mentioned techniques and well-known molecular descriptors. These methodology could also be applied to the study of more large and complex peptides even proteins and a new branch of this approach is been completing taken into account also the 3-D structure of the peptides and proteins, essential factor in the biological properties of these complex biopolymers.

Table 4. Comparison between MARCH-INSIDE methodology and previously reported QSPR studies with this data set of dipeptides

#	Descriptor	Statistical technique	N	R^2	RMS	Ref.
1	ISA-ECI	PLS	48	0.85	0.24	24
2	Z Scale	—	48	0.71	0.34	25
3	Z Scale	PLS	48	0.82	0.26	24
4	MHDV	PCR	48	0.91	0.18	28
5	Extended Z Scale	PLS	48	0.78	—	26
6	MS-WHIM	—	48	0.824	0.26	25
7	MARCH-INSIDE	RLM	48	0.823	0.27	—
8	MARCH-INSIDE	RLM	46	0.881	0.22	—
9	MARCH-INSIDE	PLS	48	0.858	0.23	—

Table 5. Comparison between QSPR models obtained from different kinas of descriptors not previously employed for modeling this dataset

Descriptor employed	R^2	S	F	# Var[a]	q^2
Constitutional	0.846	0.256	123.0	2	0.82
Topological	0.910	0.195	110.8	4	0.89
Molecular walk counts	0.618	0.393	74.47	1	0.58
BCUT	0.783	0.303	53.0	3	0.72
Gálvez topological charges indices	0.617	0.398	36.0	2	0.56
2D autocorrelations	0.753	0.320	68.5	2	0.71
Randic molecular profiles	0.559	0.423	58.2	1	0.512
Geometrical	0.909	0.197	146.0	3	0.895
RDF	0.851	0.254	61.5	4	0.814
3D-Morse	0.914	0.195	89.5	5	0.880
WHIM	0.861	0.248	52.2	5	0.799
GETAWAY	0.889	0.217	117.5	3	0.857

[a] Number of variables.

3.1. Descriptors generation and statistical process

The bitter tasting threshold (log(1/T) for the data of 48 dipeptides was taken from the Collantes report.[24] The molecular descriptors ($^{SR}\pi_k$ Θ_k $^{Abs}\pi_k$) were calculated with the experimental software MARCH-INSIDE version 2.0.[36] The chemical structure is directly introduced by using the Draw Mode of the software. The structure can then be saved and is possible to select the Calculus Mode of the software and to obtain the first 10th-order local (as desired) and total molecular descriptors. In this case, the only total descriptors were calculated.

Besides, seven families of molecular descriptors (comprises 0-D, 2-D and 3-D descriptors) were calculated by using the DRAGON software.[37] These descriptor families were: constitutional, topological, molecular walk counts, BCUT, Gálvez topological charges indices, 2D autocorrelations, Randic molecular profiles, geometrical, RDF, 3D-Morse, WHIM and GETAWAY. The molecular structure of the dipeptide was drawn by using the CHEMDRAW software[38] and saved as a .mol file. The structure is modified with the MOPAC software,[39] geometry was optimized with the AM1 method and the structure was saved as a .hin file, which can be processed by the software DRAGON.

Linear discriminant analysis (LDA) was done in order to classify into two groups: one having log(1/T) > 2.0 and the other group with log(1/T) < 2.0. This value was selected as a threshold between both groups taking into account the break point value obtained after applying a piece wise linear regression[40] to this data. For the LDA analysis, we employed the linear discriminant analysis module of the STATISTICA software.[40] The assessment of the statistical quality of the models was done by some well-known parameters such as Wilk's lambda (λ), Fischer ratio (F), squared Mahlanobic distance (D^2) and the percentage of good classification for the training set as well as for cross validation procedure Leave-One-Out.

Afterward, a multiple linear regression was applied including the whole set of compounds. The statistical and predictive power of these models was assessed by means of the values of the regression coefficient (R^2), standard error of estimation (sd), Fischer ratio and level of significance (F, p) and cross validation regression coefficient (q^2).

References and notes

1. Margolskee, R. F. *J. Biol. Chem.* **2002**, *277*, 1.
2. Katritzky, A. R.; Petrukhin, R.; Perumal, S.; Karelson, M.; Prakash, I.; Desaic, N. *Chem. Act.* **2002**, *75*, 475.
3. Adler, E.; Hoon, M. A.; Mueller, K. L.; Chandrashekar, J.; Ryba, N. J. P.; Zuker, C. S. *Cell* **2000**, *10*, 693.
4. Matsunami, H.; Montmayeur, J.-P.; Buck, L. B. *Nature* **2000**, *404*, 601.
5. Huang, L.; Shanker, Y. G.; Dubauskaite, J.; Zheng, J. Z.; Yan, W.; Rosenzweig, S.; Spielman, A. I.; Max, M.; Margolskee, R. F. *Nat. Neurosci.* **1999**, *2*, 1055.
6. Rossler, P.; Boekhoff, I.; Tareilus, E.; Beck, S.; Breer, H.; Freitag, J. G. *Chem. Senses* **2000**, *25*, 413.
7. Todeschini, R.; Consonni, V. In *Handbook of Molecular Descriptors*; Mannhold, R., Kubinyi, H., Timmerman, H., Eds.; Wiley-ECH, 2000; p 667.
8. Font-Arellano, M.; Monge-Vega, A. Métodos Variacionales. In *Diseño de Fármacos (Monografías)*; Mosqueira-Toribio, A., Eds.; Farmaindustria: Madrid, 1994.
9. Shin, W.; Kim, S. J.; Shin, J. M.; Kim, S. H. *J. Med. Chem.* **1995**, *38*, 4325.
10. Takahashi, Y.; Miyashita, Y.; Tanaka, Y.; Abe, H.; Sasaki, S. *J. Med. Chem.* **1982**, *25*, 1245.
11. Takahashi, Y.; Miyashita, Y.; Tanaka, Y.; Hayasaka, H.; Abe, H.; Sasaki, S. *J. Pharm. Sci.* **1984**, *73*, 737.
12. Kier, L. B. *J. Pharm. Sci.* **1980**, *69*, 416.
13. Iwamura, H. *J. Med. Chem.* **1980**, *23*, 308.
14. Spillane, W. J.; Birch, G. G.; Drew, M. G. B.; Bartolo, I. *J. Chem. Soc., Perkin Trans. 2* **1992**, 497.
15. Spillane, W. J.; Morini, G.; Birch, G. G. *Food Chem.* **1992**, *44*, 337.
16. Spillane, W. J.; Sheahan, M. B. *J. Chem. Soc., Perkin Trans. 2* **1989**, 741.
17. Birch, G. G.; Parke, S.; Siertsema, R.; Westwell, J. M. *Pure Appl. Chem.* **1997**, *69*, 685.
18. Benson, G. A.; Spillane, W. J. *J. Med. Chem.* **1976**, *19*, 869.
19. Rodriguez, M.; Bland, J. M.; Tsang, J. W.; Goodman, M. *J. Med. Chem.* **1985**, *28*, 1527.
20. Rodriguez, M.; Goodman, M. *J. Med. Chem.* **1984**, *27*, 1668.
21. Miyashita, Y.; Takahashi, Y.; Takayama, C.; Sumi, K.; Nakatsuka, K.; Ohkubo, T.; Abe, H.; Sasaki, S. *J. Med. Chem.* **1986**, *29*, 906.

22. Tamura, M.; Miyoshi, T.; Mori, N.; Kinomura, K.; Kawaguchi, M.; Ishibashi, N.; Okai, H. *Agric. Biol. Chem.* **1990**, *54*, 1401.
23. Hellberg, S.; Ericsson, L.; Jonson, J. *Int. J. Pept. Protein Res.* **1991**, *37*, 414.
24. Collantes, E.; Dunn, W. *J. Med. Chem.* **1995**, *38*, 2705.
25. Zaliani, A.; Gancia, E. *J. Chem. Inf. Comp. Sci.* **1999**, *39*, 525.
26. Jonsson, J.; Eriksson, L.; Hellberg, S.; Sjostrom, M.; Wold, S. *QSAR* **1989**, *8*, 204.
27. Opris, D.; Diudea, M. V. *SAR QSAR Environ. Res.* **2001**, *12*, 159.
28. Liu, S. S.; Yin, Ch.; Cai, S.; Li, Z. A. *J. Chin. Chem. Soc.* **2001**, *48*, 253–260.
29. González, D. H.; Olazábal, E.; Castañedo, N.; Hernádez, S. I.; Morales, A.; Serrano, H. S.; González, J.; Ramos de Armas, R. *J. Mol. Mod.* **2002**, *8*, 237.
30. González, D. H.; Gia, O.; Uriarte, E.; Hernádez, I.; Ramos, R.; Chaviano, M.; Seijo, S.; Castillo, J. A.; Morales, L.; Santana, L.; Akpaloo, D.; Molina, E.; Cruz, M.; Torres, L. A.; Cabrera, M. A. *J. Mol. Mod.* **2003**, *9*, 395.
31. González, D. H.; Ramos de, A. R.; Molina, R. *Bull. Math. Biol.* **2003**, *65*, 991.
32. González, D. H.; Ramos de, A. R.; Molina, R. *Bioinformatics* **2003**, *19*, 2079.
33. González, D. H.; Ramos de, A. R.; Uriarte, E. *Online J. Bioinf.* **2002**, *1*, 83.
34. González, D. H.; Hernández, S. I.; Uriarte, E.; Santana, L. *Comput. Biol. Chem.* **2003**, *27*, 217.
35. González, D. H.; Marrero, Y.; Hernández, I.; Bastida, I.; Tenorio, I.; Nasco, O.; Uriarte, E.; Castañedo, N.; Cabrera, M.; Aguila, E.; Marrero, O.; Morales, A.; Pérez, M. *Chem. Res. Toxicol.* **2003**, *16*, 1318.
36. González D.-H.; Molina-Ruiz, R.; Hernández, I. MARCH-INSIDE version 2.0 (Markovian Chemicals 'In Silico' Design), Chemicals Bio-actives Center, Central University of 'Las Villas', Cuba 2003. This is a preliminary experimental version, future professional version shall be available to the public. For any information about it send an e-mail to the corresponding author: humbertogd@vodafone.es.
37. Todeschini, R.; Consonni, V.; Pavan, M. DRAGON software version 2.1, 2002.
38. CambrigdeSoft Corporation. ChemDraw® Ultra. Chemical Structure Drawing Standard, 2003.
39. Dewar, M. J. S. *J. Mol. Struct.* **1983**, *41*, 100.
40. StatSoft, Statistica 6.0. Copyright ©1984–2002.