

Computational studies of sweet-tasting molecules*

Jodie S. Barker, Channa K. Hattotuwigama, and
Michael G. B. Drew[‡]

Department of Chemistry, University of Reading, Whiteknights, Reading RG6 6AD, UK

Abstract: Quantitative structure–activity relationships (QSARs) are developed for two separate families of sweet-tasting molecules for which sweetness values relative to sucrose (RS) have been measured. For these two families of sucrose and guanidine derivatives, the molecules were divided into training and test sets. Linear multiple regression equations have been generated to relate separately $\log(\text{RS})$ to two types of parameters, namely molecular descriptors and energies derived via molecular field analysis (MFA). The parameters used in the development of linear multiple regression equations were selected by the genetic algorithm. The equations obtained show high predictive quality, which is confirmed by statistical parameters obtained with the test sets.

The data for these two families were then combined with data from two other families previously studied, namely the sulfamates and isovanillates, to make a set of 149 compounds. These molecules were also studied by QSAR methods. The generated equations show remarkable predictive power, and the quality of the results suggest that the mechanism of sweet taste receptor is similar and, therefore, that there could well be only one receptor site for sweet taste, particularly for the four sweet taste families considered in this work.

INTRODUCTION

Sweet taste transduction is thought to arise from the interaction of a molecule with a G protein-coupled taste receptor at the taste receptor cells, which generate a sensation of pleasant sweetness. Recent studies have identified the T1R3 receptor as a probable candidate for the sweet taste receptor [1–6]. While the residue sequence is published, the detailed structure has not yet been determined. When that structure is available, then it will be possible to investigate how sweet-tasting molecules interact with the active site and hence establish a direct correlation with relative sweetness (RS). In the meantime, it is possible to investigate sweet taste by studying molecules with particular taste properties and establishing the molecular features that are responsible for taste.

Previous studies on structure–sweet taste relationships, recently reviewed [7], have been characterized by the glucophore models, starting at the AH-B theory of Shallenberger and Acree [8] and the AH, B, X theory of Kier [9] and moving on to the multipoint attachment theory of Tinti and Nofre [10,11]. These models describe the sweetness of a molecule as arising from functional groups with specific physicochemical properties and, in particular, geometric arrangements. These models, though qualitative or semiquantitative, have been widely used over the last decade to explain the taste characteristics of large numbers of sweet-tasting compounds.

**Pure Appl. Chem.* Vol. 74, No. 7, 2002. A special topic issue on the science of sweeteners.

[‡]Corresponding author: E-mail: m.g.b.drew@reading.ac.uk

By contrast, we have concentrated on quantitative studies of sweet taste. Here, we detail our development of quantitative structure–activity relationships (QSARs) for sweet-tasting molecules in which we analyze molecules with known RS values. QSARs can be established most readily with families of molecules, and we have studied several series with known sweetness value. In previous work, we have studied 21 sulfamates that have RS values of 0.6 to 70.5 [12] and 41 isovanillates with RS values between 1 to 10 000 [13]. For the sulfamates, we used principal component analysis to distinguish between molecules that tasted sweet and bitter, and also developed linear multiple regression equations with good predictive power. For the isovanillates, we also developed similar high-quality equations [13].

In this paper, we detail QSARs for 40 sucrose derivatives with RS values of 0.2 to 7500 and 47 guanidine molecules with RS values of 350 to 205 000. We also show combined QSARs for sweet taste using data from 149 molecules in all four families of molecules.

EXPERIMENTAL

An equivalent methodology was used for the sucrose and guanidine derivatives and the combined set of molecules, and this is equivalent to that detailed in our published work with the isovanillates [13]. For each study, the molecules were divided into a training set to develop the QSAR, and a test set to validate the QSAR.

Two types of parameters were used separately in the calculations. Molecular descriptors were derived from the 2-dimensional and 3-dimensional structures of the molecules. For the guanidine derivatives, as previously for the sulfamates and isovanillates, the lowest energy conformations of each molecule were used for the generation of descriptors, and these were obtained by using the grid scan method in Cerius² [14]. However, for the sucrose derivatives, the conformations were generated following a different protocol. As is known from the crystal structures of sucrose and sucralose, their conformations are very different with respect to the glycosidic bridge torsion angles which have been determined by X-ray crystallography as 108.2, –45.1° in sucrose [15] and 91.4, –162.2° in sucralose [16]. The differences occur because in sucrose there are O(2)...O(1') and O(5)...O(6') intramolecular hydrogen bonds between the two rings, while in sucralose these hydrogen bonds cannot occur because of hydroxide substitution, and, therefore, the O(2)...O(5') hydrogen bond is found which requires a different conformation. However, it is debatable whether these hydrogen bonds, and therefore the conformation, would persist in solution and indeed what conformation would exist when the molecules interact with the receptor site. Molecular mechanics calculations suggest that the conformation could readily change to maximize the interaction in a receptor site [17]. We considered three different choices of conformation, using the sucrose bridge conformation or the sucralose bridge conformation or obtaining and then using the lowest energy conformation for each molecule. Taking into account that it has been established by X-ray crystallography that drugs do not often take up their lowest energy conformation in their receptor sites, we rejected the third option of using the lowest energy conformation of each molecule. Given that the interaction of the molecules with the receptor site is likely to be mediated with solvent water, a fact that has implications for changing the conformational preferences of the derivatives, there seems no obvious reason to choose one conformation or the other. We, therefore, selected the first option and for each derivative constrained the torsion angles in the glycosidic bridge to 108.2 and –45.1° the values found in sucrose and carried out a grid scan analysis on the remaining rotatable bonds in the molecule to find the lowest energy conformation.

After establishing the conformations of the molecules, over 100 such parameters were obtained, which could be categorized as conformational, electronic constants, receptor, topological, information-content, molecular shape analysis, spatial, structural, and thermodynamic descriptors. A second set of parameters was then obtained via molecular field analysis (MFA). First, the molecules in each group were overlapped using common structural features. This is straightforward for the sucrose and guanidine descriptors because they share common structural features, less so for the combined set of mole-

cules. The overlapped molecules were then placed within a box of appropriate size (*vida supra*). The fragments H^+ , CH_3 , CH_3^- , and CH_3^+ were used as probes and positioned at 2\AA intervals within the box. The energies of interaction between the probes and the molecules were then calculated at each point to give, for example, for the 149 compounds in the 4 families, 990 points for each probe and therefore 3960 interaction energies for each molecule. The 10 % interaction energies with the highest variance were then selected to give 396 points for each probe.

With both types of parameters, we next derived multiple linear regression equations of the type $\log(RS) = a_0 + a_1p_1 + a_2p_2 + a_3p_3 \dots a_n p_n$, where a_i are constants and p_i are the parameters. The genetic algorithm was then used to select the most appropriate parameters to use in the equation using the methodology previously described [13]. Both the genetic factor algorithm (GFA) [18] and genetic partial least squares (GPLS) methods were used within the QSAR modules of Cerius² [14]. In the GFA, the best equations are selected via Friedman's lack of fit (LOF) factor, which takes into account the number of terms used in the equation and is not biased, as are other indicators toward large numbers of parameters. Other statistical measures such as r^2 , $r^2(CV)$ the cross-validation r^2 value obtained via the leave-one-out method, and the predicted residual sum of squares (PRESS) value were used to validate the equation and assess its predictive ability.

RESULTS AND DISCUSSION

Sucrose derivatives

RS values were available in the literature for 40 sucrose derivatives [19], which are shown in Fig. 1. Over 20 years ago, it was discovered that the replacement of hydroxide groups by halides led to increased sweetness. Of particular interest is sucralose, 4,1',6'-trichloro-4,1',6'-trideoxygalactosucrose, which has an RS value of 650. This compound is currently used as a high-intensity sweetener in many countries such as Australia and Canada and is being considered as a food ingredient in the EC [20,21]. However, there are many other such compounds with RS values of up to 7500 as shown in Fig. 1. For the QSAR study, the molecules were divided into a training set of 30 and a test set of 10 molecules. The molecules in the test set were chosen to provide a varied range of RS values and substituents. Molecular descriptors were then calculated. For the MFA, the molecules were overlapped (Fig. 2) using the O(5)-C(1)-O(1)-C(2')-O(5') linkage and placed within a box of size $6*6*6\text{\AA}^3$.

Multiple linear regression equations were then developed using both molecular descriptors and energy values obtained from MFA. Equations were generated using different numbers of descriptors. The best equations, assessed via the statistical parameters listed above, using both GFA and GPLS methods are reported below. The descriptors abbreviated in the equations are given in full in the Appendix. Plots of calculated $\log(RS)$ against experimental $\log(RS)$ values for both the training and test sets from the GFA methods are given in Figs. 3 and 4.

$$\begin{aligned} \text{Log(RS)} = & -2.708 - 0.016*\text{MW} + 0.500*\text{CHI-V-1} - 2.352*\text{ROG} + 0.181*\text{IAC-Total} - \\ & 0.053*\text{Dipole-Y} - 0.922*\text{Kappa-2} \text{ [GFA, } r^2 = 0.902, \text{ LOF} = 0.270, r^2(\text{CV}) = 0.849, \\ & \text{PRESS}_{(\text{training})} = 4.725, \text{PRESS}_{(\text{test})} = 5.453] \end{aligned} \quad (1)$$

$$\begin{aligned} \text{Log(RS)} = & -3.528 - 0.355*\text{RotlBond} + 0.157*\text{IAC-Total} - 0.072*\text{Dipole-Y} - 0.032*\text{JURS-} \\ & \text{WPSA-1} + 0.032*\text{JURS-PPSA-3} \text{ [GPLS, } r^2 = 0.892, r^2(\text{CV}) = 0.809, \text{PRESS}_{(\text{training})} = 5.953, \\ & \text{PRESS}_{(\text{test})} = 5.611] \end{aligned} \quad (2)$$

$$\begin{aligned} \text{Log(RS)} = & -0.084 + 0.045*\text{H}^+/406 + 0.056*\text{H}^+/181 + 0.065*\text{CH}_3/243 + 0.037*\text{H}^+/269 \text{ [GFA,} \\ & r^2 = 0.904, \text{ LOF} = 0.186, r^2(\text{CV}) = 0.860, \text{PRESS}_{(\text{training})} = 4.378, \text{PRESS}_{(\text{test})} = 11.353] \end{aligned} \quad (3)$$

$$\begin{aligned} \text{Log(RS)} = & -0.530 + 0.040*\text{CH}_3^+/150 + 0.030*\text{H}^+/181 + 0.025*\text{CH}_3^-/182 + 0.023*\text{H}^+/269 + \\ & 0.033*\text{CH}_3/243 + 0.015*\text{H}^+/180 + 0.029*\text{CH}_3^+/356 - 0.023*\text{CH}_3^-/181 \text{ [GPLS, } r^2 = 0.942, \\ & r^2(\text{CV}) = 0.893, \text{PRESS}_{(\text{training})} = 3.345, \text{PRESS}_{(\text{test})} = 8.134] \end{aligned} \quad (4)$$

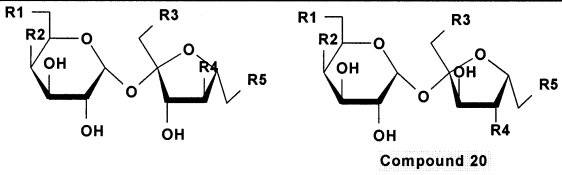
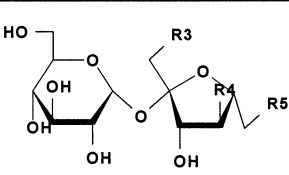
											
Cpd	R1	R2	R3	R4	R5	RS	Cpd	R3	R4	R5	RS
1	OH	OH	OH	OH	OH	0.2	34	OH	OH	OH	1
2	OH	H	OH	OH	OH	1	35	Cl	OH	OH	20
3	OH	OH	OH	CL	OH	2	36*	OH	OH	Cl	20
4*	Cl	Cl	OH	OH	Cl	4	37	Cl	Cl	OH	30
5	OH	Cl	OH	OH	OH	5	38	Cl	OH	Cl	80
6	OH	OH	OH	Cl	Cl	5	39*	Br	OH	Br	80
7	OH	OH	Cl	OH	OH	20	40*	Cl	Cl	Cl	100
8	OH	OH	OH	OH	Cl	20					
9*	Cl	OH	Cl	OH	Cl	25					
10	OH	OH	Cl	Cl	OH	30					
11*	OH	F	F	OH	F	40					
12	OH	Cl	OH	OH	Cl	50					
13	OH	OH	Cl	OH	Cl	76					
14	OH	OH	Cl	Cl	Cl	100					
15	OH	Cl	Cl	OH	OH	120					
16	OH	I	I	OH	I	120					
17	OH	Cl	Cl	H	Cl	150					
18	OH	Cl	OH	Cl	Cl	160					
19	Cl	Cl	Cl	OH	Cl	200					
20*	OH	Cl	Cl	Cl	Cl	200					
21	OH	Cl	Cl	Cl	OH	220					
22	OH	Cl	Cl	OMe	Cl	300					
23	OH	Br	Cl	OH	Cl	375					
24	H	Cl	Cl	OH	Cl	400					
25*	OMe	Cl	Cl	OH	Cl	500					
26	OH	Cl	Cl	OH	Cl	650					
27	OH	Cl	Br	OH	Br	800					
28	OH	Br	Br	OH	Br	800					
29	OH	Cl	Cl	F	Cl	1,000					
30	OH	Cl	Cl	Cl	Cl	2,200					
31	OH	Cl	Cl	Br	Cl	3,000					
32*	OH	Cl	Cl	I	Cl	7,500					
33*	OH	Br	Br	Br	Br	7,500					

Fig. 1 List of the sucrose derivatives used in this work together with their RS values. Molecules in the test set are marked *.

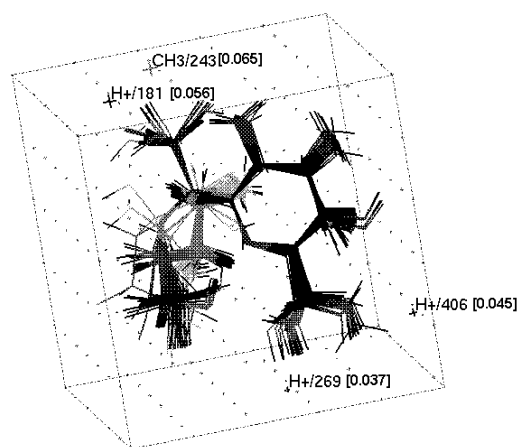
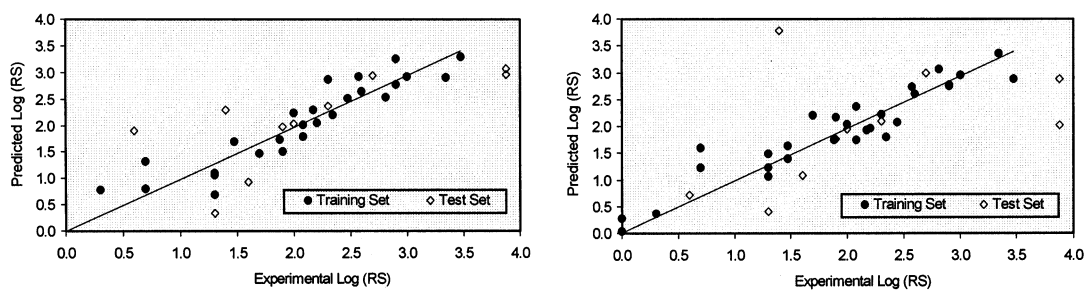


Fig. 2 Overlapping of the sucrose derivatives as used in the MFA. The positions of the probes used in eq. 3 are identified.



Figs. 3 and 4 Plot of observed $\log(\text{RS})$ value against calculated $\log(\text{RS})$ value for the 40 sucrose derivatives using eqs. 1 and 3, respectively. Data for molecules in the training set are shown as filled circles, test set as open diamonds.

Both sets of parameters have led to equations with good predictive quality with r^2 values in excess of 0.90 and $r^2(\text{CV})$ values in excess of 0.85. Figures 2 and 3 show that there is a good agreement for the test set, also indicating that the equations have good predictive power. The molecular descriptors in the equations include a wide range of parameters including electronic (Dipole-Y), information content (IAC-Total), molecular weight, and JURS descriptors, which represent partial charges mapped on surface area. The terms used in MFA equations provide important information concerning how the molecules interact with the receptor site. Indeed, for the sucrose derivatives, it is particularly noticeable from Fig. 2 that there are two H^+ probes located close to the pyranose ring at positions O(6), O(4), and O(1'), suggesting that these are the crucial hydroxides to be substituted by halogens for enhanced sweetness.

The other probes are to be found around the furanose ring at positions close to O(1'). It is interesting that no probes are located close to the 3' and 4' positions, indicating that these positions are relatively unimportant for sweet taste.

Guanidine derivatives

The guanidine derivatives show a very wide range of RS values ranging up to 205 000 [16] (Fig. 5). The molecules were divided into a training set of 39 and a test set of 8. Molecules in the test set were chosen to provide a varied range of RS values and substituents. Molecular descriptors were then calculated. The molecules were overlapped using the Ph-N=C(N)N moiety (Fig. 6), and MFA parameters were calculated using a $7 \times 7 \times 7 \text{ \AA}^3$ grid.

The best equations obtained from the linear multiple regression are shown below.

$$\begin{aligned} \text{Log}(\text{RS}) = & 3.248 + 0.007 \cdot \text{Wiener} + 0.898 \cdot \text{CHI-V-3-C} - 0.259 \cdot \text{Rotlbonds} - 0.384 \cdot \text{SC3-C} + \\ & 4.081 \cdot \text{BIC} - 0.002 \cdot \text{V-Dist-MAG} \quad [\text{GFA}, r^2 = 0.705, \text{LOF} = 0.244, r^2(\text{CV}) = 0.534, \\ & \text{PRESS}_{(\text{training})} = 7.209, \text{PRESS}_{(\text{test})} = 1.690] \end{aligned} \quad (5)$$

$$\begin{aligned} \text{Log}(\text{RS}) = & 8.820 - 0.017 \cdot \text{E-ADJ-MAG} + 0.004 \cdot \text{Wiener} - 0.256 \cdot \text{Rotlbonds} + 1.133 \cdot \text{Rad. Of} \\ & \text{Gyration} - 2.018 \cdot \text{JX} - 1.081 \cdot \text{CIC} \quad [\text{GPLS}, r^2 = 0.706, r^2(\text{CV}) = 0.519, \text{PRESS}_{(\text{training})} = 7.435, \\ & \text{PRESS}_{(\text{test})} = 1.819] \end{aligned} \quad (6)$$

$$\begin{aligned} \text{Log}(\text{RS}) = & 3.498 - 0.029 \cdot \text{CH}_3^-/269 + 0.030 \cdot \text{CH}_3^-/352 + 0.050 \cdot \text{CH}_3/205 + 0.067 \cdot \text{H}^+/167 - \\ & 0.030 \cdot \text{H}^+/147 - 0.016 \cdot \text{H}^+/156 \quad [\text{GFA}, r^2 = 0.758, \text{LOF} = 0.200, r^2(\text{CV}) = 0.631, \text{PRESS}_{(\text{training})} = \\ & 5.713, \text{PRESS}_{(\text{test})} = 8.883] \end{aligned} \quad (7)$$

$$\begin{aligned} \text{Log}(\text{RS}) = & 3.672 - 0.026 \cdot \text{CH}_3/223 + 0.025 \cdot \text{CH}_3^+/177 + 0.041 \cdot \text{H}^+/202 + 0.030 \cdot \text{H}^+/324 + \\ & 0.023 \cdot \text{CH}_3/371 + 0.018 \cdot \text{H}^+/167 \quad [\text{GPLS}, r^2 = 0.778, r^2(\text{CV}) = 0.506, \text{PRESS}_{(\text{training})} = 7.637, \\ & \text{PRESS}_{(\text{test})} = 5.439] \end{aligned} \quad (8)$$

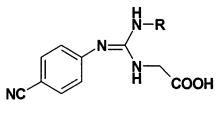
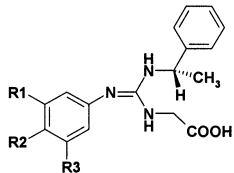
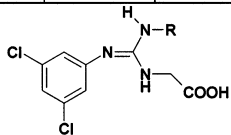
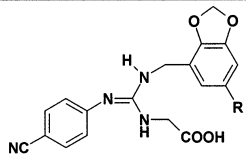
							
Cpd	R	RS	Cpd	R1	R2	R3	RS
3	c-C ₉ H ₁₇	200,000	28*	H	NO ₂	H	7,000
4*	c-C ₈ H ₁₅	170,000	29	H	H	H	5,000
5	c-C ₁₀ H ₁₉	150,000	30	Br	H	H	25,000
6	CH-(C ₆ H ₅) ₂	150,000	31	CH ₃	H	H	12,000
7*	c-C ₇ H ₁₃	60,000	32	CF ₃	H	H	7,500
8	1-naphthyl	60,000	33	NO ₂	H	H	6,000
9	(S)CH(CH ₃)-c-C ₆ H ₁₁	50,000	34	CN	H	H	5,500
10	CH ₂ -c-C ₆ H ₁₁	35,000	35*	Cl	H	Cl	120,000
11*	CH ₂ -C ₆ H ₅	30,000	36	CH ₃	H	CH ₃	30,000
12	(S)CH(CH ₃)C ₆ H ₅	28,000	37	F	H	F	15,000
13	CH ₂ -1-adamantyl	23,000	38*	CH ₃	CN	H	50,000
14	N(CH ₃)C ₆ H ₅	18,000	39	CH ₃	CN	CH ₃	50,000
15	c-C ₆ H ₁₁	12,000	40	Cl	Cl	Cl	35,000
16*	C ₆ H ₄ (3-Cl)	10,000					
17	(R)CH(CH ₃)C ₆ H ₅	9,000	Cpd	R		RS	
18	C ₆ H ₄ (3-CH ₃)	9,000	41	CH ₂ C ₆ H ₅		80,000	
19	CH ₂ CH ₂ C ₆ H ₅	8,500	42	(S)CH(CH ₃)-c-C ₆ H ₁₁		70,000	
20	C ₆ H ₄ (4-CH ₃)	7,000	43	c-C ₈ H ₁₅		60,000	
21	(CH ₂) ₅ CH ₃	6,000	44*	CH ₂ -c-C ₆ H ₁₁		35,000	
22	C ₆ H ₄ (2-CH ₃)	5,000	45	1-naphthyl		30,000	
23	1-indanyl	5,000	46	c-C ₇ H ₁₃		20,000	
24	C ₆ H ₅	4,000	47	C ₆ H ₅ (3,5-dicl)		1,000	
25	1-adamantyl	3,500					
26	H	2,700	48	H		205,000	
27	CH ₂ CH ₃	350	49	CH ₃		200,000	

Fig. 5 List of the guanidine derivatives used in this work together with their RS values. Molecules in the test set are marked *.

The observed and calculated values of $\log(\text{RS})$ for the training and test sets are shown in Figs. 7 and 8 for eqs. 5 and 7 which were obtained using the GFA method.

The statistical measures are not as high as for the sucrose derivatives with r^2 values around 0.70, while $r^2(\text{CV})$ values are between 0.50 and 0.60. However, it is interesting that the PRESS(test) values using molecular descriptors are particularly good, showing that the equations have good predictive quality. It is noteworthy that the majority of descriptors selected for the equations are 2-D connectivity terms. By default, this may suggest that the 3-D conformations are particularly flexible and those found in the receptor site may not be those used for the MFA. Supporting evidence for this view is provided by the high PRESS(test) values observed in eqs. 7 and 8 and illustrated in Fig. 8, which show that the MFA equations do not have good predictive quality. In the MFA equations, it can be observed from Fig. 6 that there are no probe positions located around the common features of the guanidines. This is

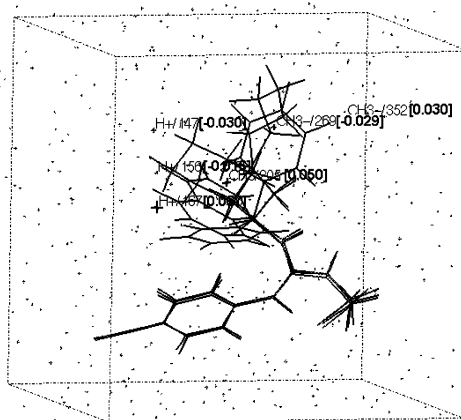
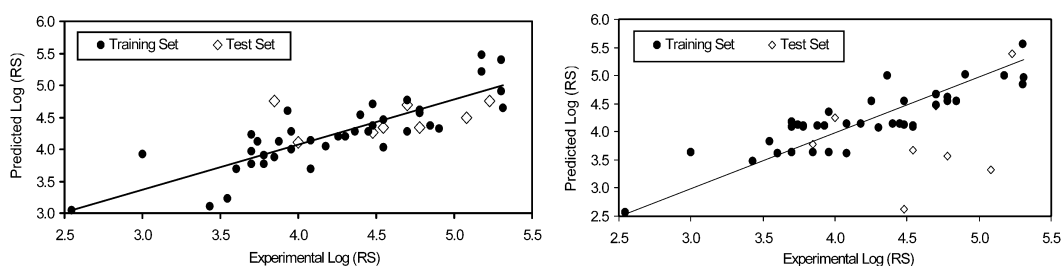


Fig. 6 Overlapping of selected guanidine derivatives as used in the MFA. Not all the guanidine derivatives are shown for reasons of clarity, but the overlapping groups are clearly indicated. The positions of the probes used in eq. 7 are identified.



Figs. 7 and 8 Plot of observed log(RS) value against calculated log(RS) value for the 47 guanidine derivatives using eqs. 5 and 7, respectively. Data for molecules in the training set are shown as filled circles, test set as open diamonds.

to be expected, as these common features are clearly not responsible for the variations in RS values, although of course they are necessary for the generic sweet taste in the guanidine family. The probe positions are located around the R groups (Fig. 5) and illustrate how variations in R affect the sweet taste.

Four families of sweetener molecules

We next considered whether it was possible to generate a QSAR for all sweet-tasting molecules. Clearly, the answer could only be yes, if there was one mode of action in one sweet taste receptor. We derived QSARs for molecules in 4 families, the sucrose and guanidine derivatives discussed above together with sulfamate and isovanillate families. The molecules used in these latter 2 families, and their RS values are previously published [12,13]. There were 149 molecules in all, and these were divided into a training set of 120 molecules and a test set of 29. The test set was distributed in proportion throughout the 4 families, and molecules were selected to provide a varied range of RS values and of derivatives. There is a significant problem in applying MFA to these 149 molecules, namely, how to overlap the molecules. As we have shown, it was relatively straightforward for the 4 families separately,

but much more difficult for the 4 families together. It was decided to overlap the families using the Tinti–Nofre sweetness model [22] Each family was overlapped with its pharmacophore coincident with the AH, B, and G sites of the model to obtain the overlap shown in Fig. 9. A box of size $10 \times 9 \times 8 \text{ \AA}^3$ was used. The equations obtained are shown below, and the agreement between observed and calculated $\log(\text{RS})$ values are illustrated in Figs. 10 and 11.

$$\begin{aligned} \text{Log}(\text{RS}) = & -3.766 - 0.002 \cdot \text{Wiener} + 0.390 \cdot \text{AlogP} + 0.537 \cdot \text{CHI-O} - 0.936 \cdot \text{CHI-3-C} + \\ & 0.031 \cdot \text{Dipole-Y} \text{ [GFA, } r^2 = 0.833, \text{ LOF} = 0.376, r^2(\text{CV}) = 0.817, \text{ PRESS}_{(\text{training})} = 41.497, \\ & \text{PRESS}_{(\text{test})} = 11.236] \end{aligned} \quad (9)$$

$$\begin{aligned} \text{Log}(\text{RS}) = & -4.526 - 0.002 \cdot \text{Wiener} - 0.176 \cdot \text{SC-3-C} + 0.314 \cdot \text{SC-0} + 0.240 \cdot \text{CHI-V-0} + \\ & 0.290 \cdot \text{AlogP} + 0.034 \cdot \text{Dipole-Y} \text{ [GPLS, } r^2 = 0.840, r^2(\text{CV}) = 0.823, \text{ PRESS}_{(\text{training})} = 40.148, \\ & \text{PRESS}_{(\text{test})} = 9.186] \end{aligned} \quad (10)$$

$$\begin{aligned} \text{Log}(\text{RS}) = & 3.721 - 0.021 \cdot \text{CH}_3^+/730 + 0.085 \cdot \text{H}^+/282 + 0.271 \cdot \text{CH}_3^+/531 + 0.022 \cdot \text{CH}_3^-/573 - \\ & 0.339 \cdot \text{CH}_3^+/426 - 0.038 \cdot \text{CH}_3^-/778 \text{ [GFA, } r^2 = 0.830, \text{ LOF} = 0.396, r^2(\text{CV}) = 0.808, \\ & \text{PRESS}_{(\text{training})} = 43.605, \text{PRESS}_{(\text{test})} = 27.279] \end{aligned} \quad (11)$$

$$\begin{aligned} \text{Log}(\text{RS}) = & -1.071 + 0.078 \cdot \text{CH}_3^+/265 + 0.100 \cdot \text{CH}_3^+/534 - 0.028 \cdot \text{CH}_3^+/730 + 0.020 \cdot \text{CH}_3^-/573 \\ & + 0.118 \cdot \text{CH}_3^-/815 \text{ [GPLS, } r^2 = 0.819, r^2(\text{CV}) = 0.800, \text{ PRESS}_{(\text{training})} = 45.449, \text{PRESS}_{(\text{test})} = \\ & 28.577] \end{aligned} \quad (12)$$

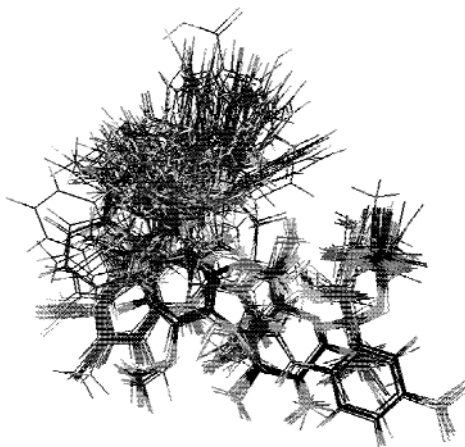
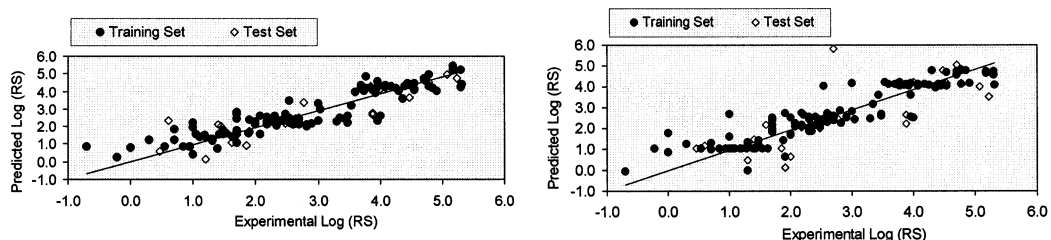


Fig. 9 Overlap of the 149 molecules used in the study of the 4 sweet-tasting families.



Figs. 10 and 11 Plot of observed $\log(\text{RS})$ value against calculated $\log(\text{RS})$ value for the 149 sweet-tasting molecules using eqs. 9 and 11, respectively. Data for molecules in the training set are shown as filled circles, test set as open diamonds.

While the statistical tests show that both sets of parameters have led to equations with good predictive quality, the values using the molecular descriptors give higher statistical measures. It is interesting that the majority of terms used in the equations derive from 2-D connectivity and molecular graph terms, although AlogP and Dipole-Y are also found indicating the importance of the charge distribution in the molecules. From Fig. 11, it can be concluded that the equations developed via the MFA do not show much predictive quality. The RS values of a few of the test set molecules are particularly badly predicted. This may well indicate that the method that was used for the overlap of the molecules was unsatisfactory. However, in the absence of knowledge of the structure of the receptor, it is difficult to decide upon a better method. It is interesting, however, to note that studies using the pseudo-receptor, where the methodology allows for the minor realignment of the overlapped molecules, does provide activity relationships with better predictive quality [23].

It can, of course, be argued that we should have incorporated a wide range of sweet-tasting molecules in our study rather than limiting ourselves to 4 families, and this indeed is work that we are planning to carry out in the near future. However, even so, the results of calculations using molecular descriptors on 149 molecules in four different families are very encouraging. In particular, the high predictive quality of this equation is indicated by the statistical values and indeed by the values predicted for the test set. This shows that this equation or a variation thereof may well be able to predict the relative sweetness of a wide range of sweet compounds, not just those in the 4 families. The success of this equation in explaining the sweet taste response of 4 families may well confirm the view that there is indeed just one taste receptor that acts in a comparable manner for the majority of sweet-tasting molecules.

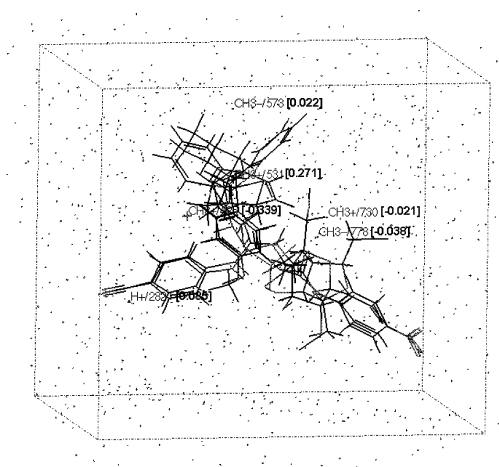


Fig. 12 Overlap of the 149 molecules used in the study of the 4 sweet-tasting families. Not all the molecules are shown for reasons of clarity but the overlapping groups are clearly indicated. The positions of the probes used in eq. 11 are identified.

ACKNOWLEDGMENTS

We thank EPSRC for a studentship (CKH) and the EC for support via grant (FAIR-CT98-4040). We thank G. Morini, University of Milano, for providing the overlapped structures of the guanidine derivatives.

APPENDIX

A listing of descriptors used in the equations reported in this work [24] is given below.

MW: molecular weight.

Rotlbonds: the number of rotatable bonds in the molecule.

DIPOLE: the Dipole moment. DIPOLE-Y is the dipole in the Y direction.

Kier and Hall CHI molecular connectivity indices are numerical indices that represent structure via chemical graph theory with emphasis on the molecular skeleton. Two types of CHI indices are used here; CHI that is dependent just on the connectivity and CHI-V that takes into account the atom type [25,26].

Kier and Hall subgraph count index (SC): This is the number of subgraphs of a given type and order. SC-3_C counts the number of clusters.

Kier's shape indices [κ_n ($n = 1, 2, 3$)]: These indices compare the molecular graph with "minimal" and "maximal" graphs where the meaning of "minimal" and "maximal" depends on the order n [25,26]. Kappa-2 indicates the degree of linearity, or star-shape, of the bonding pattern.

Wiener index (W) is the sum of the chemical bonds existing between all pairs of heavy atoms in the molecule [27].

IAC-Mean, IAC-Total: Information of atomic composition index. The atoms in the molecule are partitioned into equivalence classes corresponding to their atomic numbers.

Information indices based on the distance and edge adjacency matrices.

V_DIST_mag: Vertex distance/magnitude

E_ADJ_mag: Edge adjacency/magnitude

Multigraph information content indices.

BIC: Bonding information content, number of bonds counting bond orders.

CIC: Complementary information content measures the deviation of IC from its maximum possible value corresponding to the partition into classes containing one element each.

JURS descriptors are based on partial charges mapped on surface area. This set of descriptors combines shape and electronic information to characterize the molecules. The descriptors are calculated by mapping atomic partial charges on solvent accessible surface areas of individual atoms [28]. A total of 30 different descriptors are included in the set of which 2 were used in the equations, namely, JURSWPSA-1 and JURSPPSA-3, which represent, respectively, surface-weighted charged partial surface areas and atomic charge-weighted positive surface area.

ROG: Radius of gyration (\AA)

AlogP: LogP, the octanol/water partition coefficient is a molecular descriptor that can be used to relate chemical structure to observed chemical behavior. LogP is related to the hydrophobic character of the molecule describing effects such as the solvent behavior, polarizability, and partitioning through a cell wall. It is calculated via an atom-based approach where each atom of the molecule is assigned to a particular class, with additive contributions to the total [29].

REFERENCES

1. B. Lindemann. *Physiol. Rev.* **76**, 718 (2001).
2. A. A. Bachmanov, M. G. Tordoff, G. K. Beauchamp. *Chem. Senses* **26**, 905 (2001).
3. M. Kitagawa, Y. Kusakabe, H. Mius. *Biochem. Biophys. Res. Commun.* **283**, 236 (2001).
4. M. Max, Y. G. Shanker, L. Q. Huang. *Nat. Genet.* **28**, 58 (2001).
5. J. P. Montmayeur, S. D. Liberles, H. Matsunami, L. B. Buck. *Nat. Neurosci.* **4**, 492 (2001).
6. E. Sainz, J. N. Korley, J. F. Battey. *J. Neurochem.* **77**, 896 (2001).
7. D. E. Walters. *J. Chem. Educ.* **72**, 680 (1995).
8. R. S. Shallenberger and T. E. Acree. *Nature* **216**, 480 (1967).
9. L. B. Kier. *J. Pharm. Sci.* **61**, 1394 (1972).
10. J. M. Tinti and C. Nofre. In *Sweeteners: Discovery, Molecular Design and Chemoreception*, D. E. Walters, F. T. Orthoefer, G. DuBois (Eds.), pp. 206–213, American Chemical Society, Washington, DC (1991).
11. C. Nofre, J. M. Tinti, D. Glaser. *Chem. Senses* **21**, 747 (1996).
12. M. G. B. Drew, G. R. H. Wilden, W. J. Spillane, R. M. Walsh, C. A. Ryder, J. M. Simmie. *J. Agric. Food Chem.* **46**, 3016 (1998).
13. A. Bassoli, M. G. B. Drew, C. K. Hattotuwigama, L. Merlini, G. Morini, G. R. H. Wilden. *Quant. Struct.–Act. Rel.* **20**, 3 (2001).
14. Cerius² software. Molecular Simulations, Inc., San Diego, CA.
15. G. M. Brown and H. A. Levy. *Acta Cryst.* **B29**, 790 (1973).
16. J. A. Kanters, R. L. Scherrenberg, B. R. Leafiang, J. Kroon. *Carbohydr. Res.* **180**, 175 (1988).
17. V. H. Tran and J. W. Brady. In *Computer Modelling of Carbohydrate Molecules*, A. D. French and J. W. Brady (Eds.), pp. 213–226, ACS Symposium Series 430, American Chemical Society, Washington, DC (1990).
18. D. Rogers and A. J. Hopfinger. *J. Chem. Inf. Comput. Sci.* **34**, 854 (1994).
19. L. Hough and R. Khan. In *Sweet Taste Chemoreception*, M. Mathlouthi, J. A. Kanters, G. G. Birch (Eds.), pp. 91–102, Elsevier, New York (1991).
20. I. Knight. *Can. J. Physiol. Pharmacol.* **72**, 435 (1994).
21. C. Meyer, S. Perez, C. H. Dupenhoat, V. Michon. *J. Am. Chem. Soc.* **115**, 10300 (1995).
22. C. Nofre and J.-M. Tinti. In *Sweet Taste Chemoreception*, M. Mathlouthi, J. A. Kanters, G. G. Birch (Eds.), pp. 205–236, Elsevier, New York (1991).
23. L. Merlini, A. Bassoli, G. Morini. *Pure Appl. Chem.* **74**, 1181 (2002).
24. All descriptors are detailed in the Cerius² QSAR+ manual, Molecular Simulations, Inc., San Diego, CA (1999).
25. L. B. Kier. *Quant. Struct.–Act. Relat.* **5**, 1 (1986).
26. L. H. Hall and L. B. Kier. In *Reviews in Computational Chemistry*, Vol. 2, K. B. Lipkowitz and D. B. Boyd (Eds.), pp. 367–422, VCH, New York (1991).
27. H. Weiner. *J. Am. Chem. Soc.* **69**, 17 (1947).
28. D. T. Stanton and P. C. Jurs. *Anal. Chem.* **62**, 2323 (1990).
29. J. E. Leffler and E. Grunwald. *Rates and Equilibrium Constants of Organic Reactions*, Wiley, New York (1963).