Prediction of Sweetness by Multilinear Regression Analysis and Support Vector Machine

Min Zhong, Yang Chong, Xianglei Nie, Aixia Yan, and Qipeng Yuan

Abstract: The sweetness of a compound is of large interest for the food additive industry. In this work, 2 quantitative models were built to predict the logSw (the logarithm of sweetness) of 320 unique compounds with a molecular weight from 132 to 1287 and a sweetness from 22 to 22500000. The whole dataset was randomly split into a training set including 214 compounds and a test set including 106 compounds, represented by 12 selected molecular descriptors. Then, logSw was predicted using a multilinear regression (MLR) analysis and a support vector machine (SVM). For the test set, the correlation coefficients of 0.87 and 0.88 were obtained by MLR and SVM, respectively. The descriptors found in our quantitative structure–activity relationship models are prone to a structural interpretation and support the AH/B System model proposed by Shallenberger and Acree.

Keywords: food properties, multilinear regression (MLR), quantitative structure–activity relationships (QSAR), support vector machine (SVM), sweeteners

Practical Application: In this study, 2 quantitative models were built based on multilinear regression and support vector machine to predict the logSw of 320 compounds. The sweet taste system of a sweetener has extensively been investigated but much still needs clarification. The quantitative models for predicting sweetness built in this work can be helpful for research in food additives.

Introduction

The sweetness is a key benchmark of a sweetener since the sweet taste is one of the most important properties of a sweetener in food industry. But the high cost always limits the determination of the sweetness; therefore, it is necessary to build a quantitative model to predict the sweetness of compounds for the food additive industry.

Structure–activity relationship (SAR) and quantitative structure–activity relationship (QSAR) studies are widely undertaken for modeling the bioactivities such as bioactivity of Aurora-A kinase inhibitors (Yan and others 2011) and bioactivity of HCV NS5B polymerase inhibitors (Wang and others 2012), and for predicting toxicities such as classifying compounds inducing and noninducing toxic myopathy (Hu and others 2012) or rhabdomyolysis (Hu and others 2011) on the basis of a set of compounds. Several QSAR models have been proven useful in rationalizing and predicting structure–taste relationship, in which sweetness potency is statistically correlated with structural properties. In 1966, a QSAR research on a series of 9 sweet compounds replaced by nitroaniline has been built by Deutsch and Hansch, and they found that octanol/water distribution coefficient and hydrophobicity of the substituents play an important role in the sweetness of these compounds (Deutsch and Hansch 1966). From 1980 to 1981, Iwamura investigated the sweetness of 49 perillartine and 20 aniline derivatives and 217 L-aspartyl dipeptide analogues with QSAR models, who proposed that the sweetness of these kinds of compounds had a high correlation with the size of the substituents (Iwamura 1980; Iwamura 1981). Using the Property-Evaluation by Class Variables (PRECLAV) software, Tarko studied the sweetness of the amino-succinic acid derivatives in 2006. He found that the sweetness power of a compound is favorably influenced by the size of a molecule. The calibration set includes 97 molecules ($r^2 = 0.83$) and the validation set includes 24 molecules ($r^2 = 0.86$) (Tarko and Lupescu 2006). In our former work, a QSAR model with 103 sugars and sweeteners was built and for the test set, $r = 0.943$ was obtained (Yang and others 2011).

Since any quantitative structure–property relationship (QSPR) model is constructed on a particular dataset, it is important to build a model using a large, diverse dataset, and test the model using other different datasets. This study aims at building suitable models for predicting sweetness of a dataset of 320 compounds, which is larger than the dataset of our former studies (Yang and others 2011). The procedure of this study for building QSAR models includes 4 steps: (1) preparing a dataset containing 320 compounds having experimental sweetness. Experimental sweetness is defined that sucrose in solution has a sweetness perception rating of 100, and the sweetness of another substance is rated relative to this (Yang and others 2011); (2) calculating the molecular 2D and 3D descriptors and selecting the appropriate descriptors for sweetness prediction; (3) dividing the whole dataset into a training set and a test set by a random process; (4) building models using a multilinear regression (MLR) method and a support vector machine (SVM) analysis.

Materials and Methods

Dataset

A total of 320 unique compounds with molecular weight from 132 to 1287 and the sweetness from 22 to 22500000 were collected. The experimental logSw value of these compounds was

MS 20120543 Submitted 4/14/2012, Accepted 5/29/2013. Authors are with State Key Laboratory of Chemical Resource Engineering, Dept. of Pharmaceutical Engineering, Beijing Univ. of Chemica Technology, Beijing, China. Direct inquiries to author Yan (E-mail: aixia_yan@yahoo.com).

doi: 10.1111/1750-3841.12199

compiled from 7 references (Iwamura 1981; Douglas and Soejarto 2002; Zheng and others 2003; Gao and Fang 2005; Spillane 2006; Wilson 2007; Yang and others 2011) (as shown in Table S1). In our former work, 103 sugars and common sweeteners were collected. In this work, we extended the dataset to 320 compounds by adding 157 L-aspartyl dipeptide analogues, 23 isovanillic, and some other compounds. The duplicate compounds were removed and the average sweetness value was used if the experimental values for 1 compound were different in different references. Due to the wide range of sweetness values, the logarithm of sweetness (logSw) was used ranging from 1.3424 to 7.3522. Where if 2 bibliographic sources indicate different values of sweetness for the same molecule, the average value was used.

Structure building

The input of 2D structures was carried out using the software molecular operating environment (MOE) (MOE Chemical Computing Group, Montreal, Canada) and the 3D structures were built using the program CORINA (CORINA Molecular Networks GmbH, Erlangen, Germany). Three-dimensional structure of a molecule is closely related to a large variety of chemical, physical, and biological properties. The need for computer-generated 3D molecular structures has clearly been recognized in drug design and many other areas.

Molecular descriptors calculation

All the 1235 descriptors were calculated with the program ADRIANA.Code (ADRIANA.Code Molecular Networks GmbH, Erlangen, Germany), which include 19 global molecular descriptors (Miller 1990; Lipinski and others 1997; Ertl and others 2000), 8 shape descriptors (Tanford 1963; Volkenshteĭn 1963; Petitjean 1992), 88 two-dimensional autocorrelation vectors (Moreau and Broto 1980; Wagener and others 1995), 88 three-dimensional autocorrelation vectors (Broto and others 1984), and 1024 three-dimensional property-weighted radial distribution functions (RDFs) descriptors (Hemmer and others 1999; Yan and Gasteiger 2003). All the descriptors were calculated by ADRIANA.Code (ADRIANA.Code Molecular Networks GmbH, Erlangen, Germany) with the hydrogen atoms excluded.

A global molecular descriptor represents a chemical structure by a structural, chemical, or physicochemical feature or property of the molecule expressed by a single value. Global descriptors include molecular weight (Weight), the numbers of H-bond donors (HDon) and acceptors (HAcc), octanol/water distribution coefficient (XlogP) (Lipinski and others 1997), topological polar surface area (TPSA) (Ertl and others 2000), mean molecular polarizability (Polariz) (Miller 1990), aqueous solubility (LogS), the number of atoms (NAtoms), and so on.

A size or shape descriptor also represents a molecule by a single value and it is derived from the 3D structure of the molecule. Shape descriptors include maximum distance between 2 atoms in the molecule (Diameter) (Petitjean 1992), principal component of the inertia tensor (Inertia), molecular span (Span), molecular radius of gyration (Rgyr) (Tanford 1963; Volkenshteĭn 1963), molecular eccentricity (Eccentric), and molecular asphericity (Aspheric).

The 2D property autocorrelation uses the molecular 2D structure and atom pair properties as a basis to obtain vectorial molecular descriptors (Moreau and Broto 1980; Wagener and others 1995). The atom pair properties are summed up for certain topological distances that count the number of bonds on the shortest path between 2 atoms. The 2D autocorrelation vectors were calculated based on the following 8 atomic properties: atom identity, σ charge (2DACorr_SigChg), π charge (2DACorr_PiChg), total charge (2DACorr_TotChg), σ electronegativity (2DACorr_SigEN), π electronegativity (2DACorr_PiEN), lone-pair electronegativity (2DACorr_LpEN), and atomic polarizability (2DACorr_Polariz). The 2D autocorrelation vectors of a molecule for each 1 of the above 8 physicochemical atomic properties were calculated using Eq. (1):

$$A(d) = \frac{1}{2}\sum_{i,j} p_i p_j \delta(d - d_{ij}) \tag{1}$$

Where *A(d)* is the autocorrelation coefficient for a certain topological distance *d* (number of bonds between 2 atoms), whereas p_i and p_j are the atom properties of the atoms *i* and *j*. If the distance $d = d_{ij}$ (d_{ij} is the shortest path between the atoms *i* and *j*), the distance function $\delta_{ij} = 1$, otherwise $\delta_{ij} = 0$. The atom property *p* used for the calculation of autocorrelation functions can be either simply the identity (identity: $p_i = p_j = 1$) or any physicochemical atom property. Eleven distance values from distance of $d = 0$ to $d = 10$ were considered in this study. Thus, for each molecule, 88 two-dimensional autocorrelation vectors were obtained.

Molecules are spatial objects in 3D space. Autocorrelation can also be applied to the 3D structure of a molecule (Broto and others 1984). Thus, the resulting autocorrelation vectors code not only for the spatial arrangement of the atoms but also for the spatial distribution of physicochemical properties in a molecule. Since the distances *d* and d_{ij} (in Eq. (1) are continuous distances in 3D space between the atoms *i* and *j* (in Å), an additional binning of *d* into certain distance intervals (for example, in steps of 1 Å) is necessary to transform the function *A(d)* into a vector $A(d_n)$ of size *n*.

$$A(d_n) = \frac{1}{2L_n}\sum_{\substack{i,j \\ i\neq j}} p_i p_j \tag{2}$$

In Eq. (2), L_n is the number of distances occurring in a certain distance interval and p_i and p_j are the atom properties of the atoms *i* and *j*. The sampling of all distances in *n* equidistant intervals (for example, in distance bins of 1 to 2 Å, 2 to 3 Å, 3 to 4 Å . . .) results in an *n*-dimensional vector of autocorrelation coefficients. The atom property *p* used for the calculation of the 3D autocorrelation coefficients can either be simply the identity (identity: $p_i = p_j = 1$) or any physicochemical atom property, such as charge distributions or polarizability effects. For each molecule, all the hydrogen atoms were excluded. For each of the 8 three-dimensional autocorrelation coefficients, a series of 11 vectors were computed, where L_n corresponds to the 11 three-dimensional distance intervals from 1 to 2 Å, 2 to 3 Å . . . to 11 to 12 Å. Thus, for each molecule, 88 autocorrelations of 3D properties can be obtained.

Property-weighted RDFs use the 3D structure of a molecule and atom pair properties as the basis to derive vectorial molecular descriptors (Hemmer and others 1999; Yan and Gasteiger 2003). The products of atom pair properties are summed up within a certain distance range and weighed by a Gaussian term as distance function using Eq. (3). The resulting radial distribution function that was digitized using the formula in predefined steps leads to a vector of RDF coefficients.

$$g(r) = f \sum_{i}^{N-1} \sum_{j>1}^{N} A_i A_j e^{-B(r-r_{ij})^2}$$

$$f = \frac{1}{\sqrt{\sum_{r} [g(r)]^2}} \qquad (3)$$

where *f* is a scaling factor and *N* is the number of atoms. By including characteristic atomic properties *A* of atoms *i* and *j*, the RDF code can be used in different tasks to meet the requirements of the information to be represented. The exponential term contains the distance r_{ij} between 2 atoms *i* and *j* and the smoothing parameter *B*, which defines the probability distribution of the individual distances. *g(r)* was calculated at a number of discrete points with intervals of 0.1 Å. The smoothing parameter *B* is 100; the minimum 3D distance and maximum 3D distance are 0 and 12.8 Å, respectively. For a certain atomic property, 128 RDF codes can be obtained. Here, for each molecule, 8 atomic properties (atom identity, σ charge (SigChg), π charge (PiChg), total charges (TotChg), σ electronegativity (SigEN), π electronegativity (PiEN), lone-pair electronegativity (LpEN), and atomic polarizability (Apolariz)) were considered. Thus, for each molecule, 1024 RDF codes can be obtained.

SVM analysis

The LIBSVM (Chang and Lin 2011; LIBSVM 2011) program was used to build a SVM model. This software is based on the function of classification. After certain improvement, it can also be applied to the regression problem well (Chang and Lin 2011). More introductions and implementations about LIBSVM can be found on its website.

The LIBSVM regression was realized by the ε-support vector regression (ε-SVR) with a radial basis function (RBF) kernel function. The kernel function is used to convert the data into a higher dimensional space in order to account for nonlinearities in the estimate function. A commonly used kernel is the RBF kernel:

$$k(x, y) = \exp(-\gamma \|x - y\|^2) \qquad (4)$$

The parameter γ is selected by the user.

According to the program guide, 2 necessary steps had to be taken in advance: the scaling of input data and searching for best parameters.

The input data were compressed into [0.1, 0.9] through the formula:

$$x^* = \frac{x - x_{\min}}{x_{\max} - x_{\min}} \times 0.8 + 0.1 \qquad (5)$$

where x was the original value and x^* is the scaled value. $x_{\min}$ and $x_{\max}$ are the corresponding minimum and maximum values of the descriptor variable, respectively. There are 3 parameters used to adjust the efficiency of LIBSVM program: C, γ, and ε. An automatic searching program named "grid regression" was adopted. It could be used to search for best parameters C, γ, and ε through a leave-*k*-out cross-validation method. Meanwhile, overfitting of training set could be prevented. Here, a leave-10%-out cross validation was carried out (Yan and others 2008). When searching the best C, γ, and ε for training set, the parameters can be adjusted in the scope of 0.2.

Molecular descriptors selection

Based on training set, molecular descriptors for building the QSAR models were selected. The correlation coefficients between any 2 descriptors and between the sweetness and each descriptor were calculated. If the correlation coefficient between a descriptor and the sweetness is more than 0.70, the descriptor remained; if the pairwise correlation coefficient between any 2 descriptors is larger than 0.80, the one having the higher correlation with the sweetness was kept.

Results and Discussion

Descriptors selection

In this study, molecular descriptors for building the QSAR models were selected based on the compounds in dataset. For the 2D autocorrelation descriptors, 6776 new combination descriptors were defined by Yang's method (Yang and others 2011). The 8011 descriptors (1235 descriptors are from the computed result by ADRIANA.Code and 6776 descriptors are from the combination for the 2D autocorrelation vectors) were chosen using stepwise linear regression variable selection method. Stepwise variable entry and removal examines the variables in the block at each step for entry or removal. According to the criteria, 12 descriptors were selected, which include 2DACorr_Sigchg_2/2DAcorr_Polariz_0 (comb_desc), logS, 3DACorr_PiChg_5, RDF_SigChg_59, RDF_PiChg_9, RDF_PiChg_33, RDF_PiChg_124, RDF_TotChg_20, RDF_PiEN_41, RDF_PiEN_46, RDF_PiEN_81, and RDF_Polariz _22. The intercorrelations between the 12 descriptors and the logSw are shown in Table 1. The selected descriptors were used in the following study.

Build a model by MLR analysis

The dataset was randomly split into a training set including 214 compounds and a test set including 106 compounds. A MLR analysis was performed using 12 descriptors as input variables. The 214 compounds in the training set were used to build a model, and 106 compounds in the test set were used for the prediction of logSw. The logSw was represented by the following equation:

$$\log Sw = \sum (C_i D_i) + D_c \qquad (6)$$

In the equation, D_c is a constant, D_i is a descriptor, and C_i is its corresponding regression coefficient in the MLR model. The coefficients of 12 selected descriptors and the constant are shown in Table 2.

For the training set, $r = 0.902$, $sd = 0.958$, $F = 73.275$, and $n = 214$ were achieved (*r* is the correlation coefficient and sd is the standard deviation).

By using the equation built for training set, the logSw of the 106 compounds in the test set was calculated. For the test set, $r = 0.879$, $sd = 1.029$, and $n = 106$. The results are shown in Figure 1.

Build a model by SVM analysis

A model was built by the SVM with the LIBSVM program (LIBSVM 2011). The 214 compounds in the training set were used to build a SVM model. The optimum parameters were set as: $C = 97.006$, $\gamma = 0.082$, and $\varepsilon = 0.0625$. For the training set, $r = 0.911$ and $sd = 0.979$ were obtained.

By using the built SVM model, the logSw of the 106 compounds in the test set was predicted. For the test set, $r = 0.882$ and $sd = 0.994$. The results are shown in Figure 2.

Table 1–The correlations between the selected descriptors and the logSw.

	logSw	comb_desc	LogS	3DACorr_PiChg_5	RDF_SigChg_59	RDF_PiChg_9	RDF_PiChg_33	RDF_PiChg_124	RDF_TotChg_20	RDF_PiEN_41	RDF_PiEN_46	RDF_PiEN_81
comb_desc[a]	− 0.705	1.000										
LogS[b]	− 0.687	0.791	1.000									
3DACorr_PiChg_5[c]	0.084	0.050	− 0.063	1.000								
RDF_SigChg_59[d]	− 0.061	− 0.117	− 0.196	0.018	1.000							
RDF_PiChg_9[e]	− 0.367	0.107	0.155	− 0.048	− 0.011	1.000						
RDF_PiChg_33[f]	− 0.262	0.170	0.028	0.162	0.018	0.216	1.000					
RDF_PiChg_124[g]	0.005	− 0.015	− 0.041	0.017	0.012	0.336	0.314	1.000				
RDF_TotChg_20[h]	− 0.265	0.086	0.058	− 0.115	− 0.066	− 0.037	− 0.121	0.007	1.000			
RDF_PiEN_41[i]	0.493	− 0.338	− 0.267	− 0.156	− 0.150	− 0.106	− 0.461	− 0.063	0.110	1.000		
RDF_PiEN_46[j]	0.445	− 0.385	− 0.192	− 0.246	− 0.091	− 0.310	− 0.271	− 0.135	0.078	0.539	1.000	
RDF_PiEN_81[k]	0.298	− 0.279	− 0.356	0.091	0.155	− 0.532	− 0.116	− 0.240	0.127	0.326	0.431	1.000
RDF_Polariz_22[l]	0.274	− 0.227	− 0.181	0.081	− 0.115	− 0.147	− 0.166	− 0.088	0.087	0.131	0.236	0.146

[a]comb_desc: the combined descriptor denoting 2DACorr_Sigchg_2/2DAcorr_Polariz_0; [b]LogS: aqueous solubility; [c]3D_ACorr_PiChg_5: spatial autocorrelation π charge in intervals of 4 to 5 Å; [d]RDF_SigChg_59: radial distribution function weighted by σ charge, where r is in the range of 5.8 to 5.9 Å; [e-g]RDF_PiChg_9, 33, 124: radial distribution functions weighted by π charges, where r are in the range of 0.8 to 0.9 Å, 3.2 to 3.3 Å, 12.3 to 12.4 Å; [h]RDF_TotChg_20: radial distribution function weighted by the total atom charge (sum of σ and π charges), where r is in the range of 1.9 to 2.0 Å; [i to k]RDF_PiEN_41, 46, 81: radial distribution functions weighted by π atom electronegativities, where r are in the range of 4.0 to 4.1 Å, 4.5 to 4.6 Å, and 8.0 to 8.1 Å; [l]RDF_Polariz_22: radial distribution function weighted by effective atom polarizability, where r is in the range of 2.1 to 2.2 Å.

Table 2–The coefficients of 12 selected descriptors and the constant of the MLR analysis.

Descriptor D_i	Coefficient C_i
comb_desc	− 1594.599
LogS	− 0.223
3DACorr_PiChg_5	9.311
RDF_SigChg_59	− 1.943
RDF_PiChg_9	− 270749.711
RDF_PiChg_33	− 15.980
RDF_PiChg_124	71.895
RDF_TotChg_20	− 52.568
RDF_PiEN_41	0.004
RDF_PiEN_46	0.004
RDF_PiEN_81	− 0.006
RDF_Polariz_22	0.002
Constant D_c	3.529

A comparison of the MLR and SVM models

According to the MLR and SVM prediction figures (Figure 1 and 2), both the MLR and SVM models perform well in this study, except for small differences between the MLR and SVM results in these datasets. SVM exhibits the better performance than MLR due to embodying the structural risk minimization principle and some advantages over linear method of MLR, such as more easily overcome the "high-dimensionality problem."

The relationship between logSw and descriptors

In this study, a broad range of molecular descriptors (8011 descriptors) based on 2D and 3D autocorrelation and RDF molecular structures were calculated, and 12 descriptors were selected by statistical methods for predicting the logSw of compounds.

The AH/B system theory was founded by Shallenberger and Acree (1967). This model suggests that all sweet-tasting compounds contain a hydrogen bond donor group (AH) and a hydrogen bond acceptor (B), separated by a distance of 2.5 to 4.0 Å. They inferred that the receptor must contain a complementary B-AH pair that forms 2 hydrogen bonds when the sweetener interacts with its receptor. It is found in our work (Yang and others 2011) that the combination descriptor (2DACorr_Sigchg_2/2DAcorr_Polariz_0) is the most important

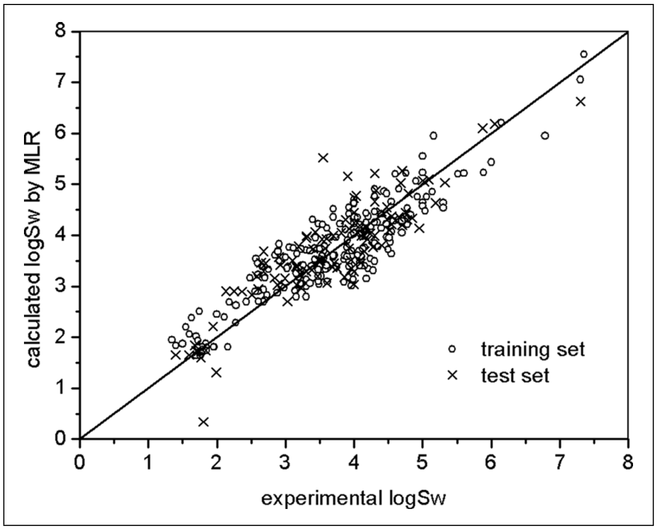

Figure 1–Calculated compared with experimental values of logSw by using MLR analysis. (The calculated compared with experimental values of logSw based on a training set and a test set by using MLR analysis.)

descriptor for predicting the logSw of compounds since it is highly correlated to logSw (as shown in Table 1). We concluded that the sweetness was highly associated with the properties of atoms themselves and the pairwise atoms separated by 2 bonds in a molecule. For example (as shown in Figure 3), in Saccharin, the –NH group (as an H-bond donor) and an oxygen atom (as an H-bond acceptor) form a AH/B group, the nitrogen atom and the oxygen atom are separated by 2 bonds. Lugduname has the highest logSw in this dataset; in this compound, at least 4 representative AH/B groups can be found (OH/O, NH/O, and 2 NH/N) and these H-bond donor/acceptor are separated by 2 or 2 bonds. The high correlation coefficient between LogS and logSw means that aqueous solubility of a compound highly impacts the sweetness. Since sweet taste chemoreception occurs in water (saliva), any solvation-related properties of sweeteners in water, such as solubility, may play a role in sweet taste mechanisms and sweeteners differentiation (Mathlouthi 1994). All the above result coincided with the conclusion of our former work (Yang and others 2011).

Furthermore, in this research, we found that vectorial molecular descriptors by radial distribution functions of atom pair properties worked well in determining the sweetness of a compound because 9 of 12 selected descriptors were based on atom RDF properties. Generally speaking, RDF descriptors describe a relationship between properties of 2 atoms and the distance among these atoms. The generation of them is shown in Equation 3, including atom charge, polarizability, and so on. But it has shown its superiority to other descriptors in our previous work (Yan and others 2011; Yan and Wang 2012). In this work, 5 out of 9 selected RDF descriptors (RDF_SigChg_59, RDF_PiChg_9, RDF_PiChg_33, RDF_PiChg_124, and RDF_TotChg_20) were based on atom charges. This indicated that the representation of the compounds structures for predicting logSw was much attribute to the atom charge. In addition, the selection of 3DACorr_PiChg_5 means that the 3D autocorrelation π charge could also impact the sweet taste of a compound. The selected RDF_PiEN_41, RDF_PiEN_46, and RDF_PiEN_81 relate to atom electronegativity which means atom electronegativity also played an important role in the sweet taste of a compound. Moreover, the logSw was correlated with the atom polarizability of a molecule due to the selection of the descriptors RDF_Polariz_22.

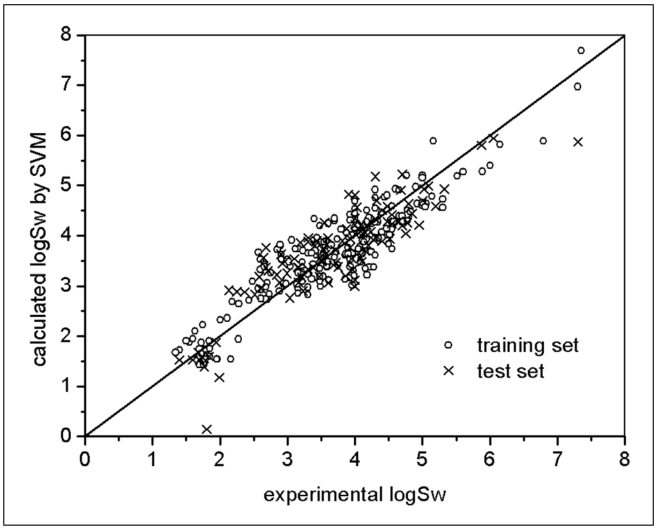

Figure 2–Calculated compared with experimental values of logSw by using SVM analysis. (The calculated compared with experimental values of logSw based on a training set and a test set by using SVM analysis.)

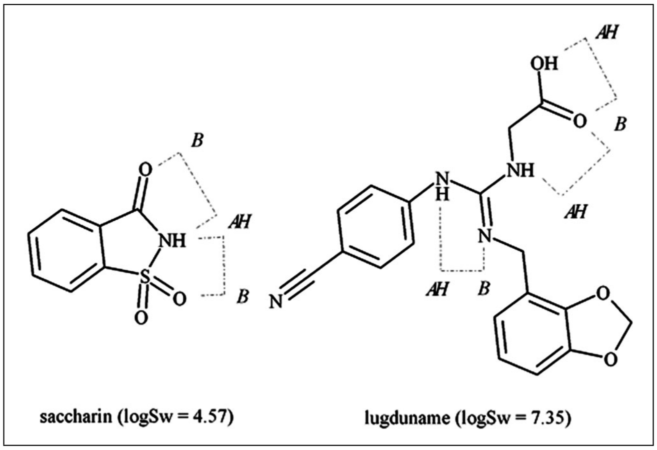

Figure 3–Two compounds structures, which have high sweetness. (Two representative compounds in dataset with high experimental and predicted sweetness by MLR and SVM models.)

In addition, 5 of the 9 RDF descriptors involve the combination of electronic properties of atoms in a distance range of 2.0 to 4.6 Å. This gives further support to the hypothesis of the structural model of Shallenberger and Acree of an AH/B system, which suggests that a hydrogen bond donor group (AH) and a hydrogen bond acceptor (B) are in the range of 2.5 to 4.0 Å. It indicated that the selected molecular descriptors can well interpret the AH/B system theory.

Conclusion

In this study, 2 quantitative models were built based on MLR and SVM regression to predict the logSw of 320 compounds. Twelve descriptors (including 2DACorr_Sigchg_2/2DAcorr_Polariz_0, LogS, 3DACorr_PiChg_5, and 9 RDF descriptors) were used to develop these models. The selected molecular descriptors can well predict the sweetness, and they can also well interpret the sweet taste system theory of a sweetener, such as the AH/B System Theory founded by Shallenberger and Acree, has extensively been investigated but much still needs clarification. The quantitative models for predicting sweetness built in this work can be helpful for research in food additives.

Acknowledgments

This work was supported by the Natl. Natural Science Foundation of China (20605003 and 20975011). The authors thank Prof. J. Gasteiger and the Molecular Networks GmbH, Erlangen, Germany for providing the programs ADRIANA.Code and CORINA for our scientific work.

References

ADRIANA.Code. Molecular Networks GmbH, Erlangen, Germany. v2.2.2, http://www.molecular-networks.com/products/adrianacode.

Broto P, Moreau G, Vandycke C. 1984. Molecular structures: perception, autocorrelation descriptor and sar studies. Autocorrelation descriptor. Eur J Med Chem Chim Ther 19(1):66–70.

Chang C-C, Lin C-J. 2011. LIBSVM: a library for support vector machines. ACM Trans Intell Syst Technol 2(3):1–27.

CORINA. Molecular Networks GmbH, Erlangen, Germany. v3.4, http://www.molecular-networks.com/products/corina.

Deutsch EW, Hansch C. 1966. Dependence of relative sweetness on hydrophobic bonding. Nature 211(5044):75.

Douglas KA, Soejarto DD. 2002. Discovery of terpenoid and phenolic sweeteners from plants. Research Triangle Park, NC: Pure and Applied Chemistry.

Ertl P, Rohde B, Selzer P. 2000. Fast calculation of molecular polar surface area as a sum of fragment-based contributions and its application to the prediction of drug transport properties. J Med Chem 43(20):3714–7.

Gao P, Fang Z. 2005. The relationship between the sweetness and structure of sucrose derivatives. Jiangsu Chem Ind 33:73–5.

Hemmer MC, Steinhauer V, Gasteiger J. 1999. Deriving the 3D structure of organic molecules from their infrared spectra. Vibrat Spectrosc 19(1):151–64.

Hu X, Yan A. 2011. In silico prediction of rhabdomyolysis of compounds by self-organizing map and support vector machine. Toxicol In Vitro 25(8):2017–24.

Hu X, Yan A. 2012. In Silico models to discriminate compounds inducing and non-inducing toxic myopathy. Mol Inf 31(1):27–39.

Iwamura H. 1980. Structure-taste relationship of perillartine and nitro- and cyanoaniline derivatives. J Med Chem 23(3):308–12.

Iwamura H. 1981. Structure–sweetness relationship of L-aspartyl dipeptide analogues. A receptor site topology. J Med Chem 24(5):572–83.

LIBSVM. 2011. v3.1, Chang, CC, Lin, CJ. http://www.csie.ntu.edu.tw/∼cjlin/libsvm/. Accessed June 2013.

Lipinski CA, Lombardo F, Dominy BW, Feeney PJ. 1997. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. Adv Drug Deliv Rev 23(1–3):3–25.

Mathlouthi M. 1994. Physico-chemical aspects of sweeteners: the ideal sugar substitute. Ind Alimentaires Agricoles 111(7–8):402–10.

Miller KJ. 1990. Additivity methods in molecular polarizability. J Am Chem Soc 112(23):8533–42.

MOE. Chemical Computing Group, Montreal, Canada. v2010.10, http://www.chemcomp.com/.

Moreau G, Broto P. 1980. The autocorrelation of a topological structure: a new molecular descriptor. Nouv J Chim 4:359–60.

Petitjean M. 1992. Applications of the radius-diameter diagram to the classification of topological and geometrical shapes of chemical compounds. J Chem Inf Comput Sci 32(4): 331–7.

Shallenberger RS, Acree TE. 1967. Molecular theory of sweet taste. Nature 216(5114): 480–2.

Spillane WJ. 2006. Optimising sweet taste in foods. Cambridge, England: Woodhead.

Tanford C. 1963. Physical chemistry of macromolecules. Hoboken, N.J.: John Wiley & Sons, Inc.

Tarko L, Lupescu I. 2006. QSAR Studies on amino-succinamic acid derivatives sweeteners. Arkivoc 13: 22–40.

Volkenshteĭn MV. 1963. Configurational statistics of polymeric chains. New York, N.Y.: Interscience Publishers.

Wagener M, Sadowski J, Gasteiger J. 1995. Autocorrelation of molecular surface properties for modeling corticosteroid binding globulin and cytosolic Ah receptor activity by neural networks. J Am Chem Soc 117:7769–75.

Wang M, Wang K, Yan A, Yu C. 2012. Classification of hcv ns5b polymerase inhibitors using support vector machine Intl J Mol Sci 13:4033–47.

Wilson R. 2007. Sweeteners. London, England: Blackwell Publishing.

Yan A, Chong Y, Wang L, Hu X, Wang K. 2011. Prediction of biological activity of Aurora-A kinase inhibitors by multilinear regression analysis and support vector machine. Bioorg Med Chem Lett 21(8):2238–43.

Yan A, Gasteiger J. 2003. Prediction of aqueous solubility of organic compounds based on a 3D structure representation. J Chem Inf Comput Sci 43(2):429–34.

Yan A, Wang K. 2012. Quantitative structure and bioactivity relationship study on human acetylcholinesterase inhibitors. Bioorg Med Chem Lett 22(9):3336–42.

Yan A, Wang Z, Cai Z. 2008. Prediction of human intestinal absorption by GA feature selection and support vector machine regression. Intl J Mol Sci 9(10):1961–76.

Yang X, Chong Y, Yan A, Chen J. 2011. In-silico prediction of sweetness of sugars and sweeteners. Food Chem 128(3):653–8.

Zheng J, Rao Z, Jia C. 2003. Study on the relationship between structure and sweetness of sucrose derivatives. Food Sci 24:29–33.

Supporting Information

Additional Supporting Information may be found in the online version of this article at the publisher's website:

Table S1. The sweetness values of 320 compounds as taken from literature.

Table S2. The data and result of training set.

Table S3. The data and result of test set.