Journal of Computer-Aided Molecular Design (2006) 20: 1–11
DOI 10.1007/s10822-005-9031-1

Prediction of standard Gibbs energies of the transfer of peptide anions from aqueous solution to nitrobenzene based on support vector machine and the heuristic method

Luan Feng[a], Zhang Xiaoyun[a], Zhang Haixia[a], Zhang Ruisheng[b,]*, Liu Mancang[a], Hu Zhide[a] & Fan Botao[c]
[a]*Department of Chemistry, Lanzhou University, 730000, Lanzhou , China;* [b]*Department of Computer Science, Lanzhou University, 730000, Lanzhou , China;* [c]*Université Paris 7-Denis Diderot, ITODYS 1, Rue Guy de la Brosse, 75005, Paris, France*

Received 29 May 2005; accepted 13 November 2005

Key words: HM, QSPR, peptide, SVM, ΔG^{θ}

Summary

Quantitative structure-property relationship (QSPR) method was performed for the prediction of the standard Gibbs energies (ΔG^{θ}) of the transfer of peptide anions from aqueous solution to nitrobenzene. Descriptors calculated from the molecular structures alone were used to represent the characteristics of the peptides. The four molecular descriptors selected by the heuristic method (HM) in COmprehensive DEscriptors for Structural and Statistical Analysis (CODESSA) were used as inputs for support vector machine (SVM) and radial basis function neural networks (RNFNN). The results obtained by the novel machine learning technique, SVM, were compared with those obtained by HM and RBFNN. The root mean squared errors (RMS) of the training, predicted and overall data sets are 2.192, 2.541 and 2.267 unit (kJ/mol) for HM, 1.604, 2.478 and 1.817 unit (kJ/mol) for RBFNN and 1.5621, 2.364 and 1.756 unit (kJ/mol) for SVM, respectively. The prediction results were in agreement with the experimental values. This paper provided a potential method for predicting the physiochemical property (ΔG^{θ}) of various small peptides.

Introduction

Standard Gibbs energies of transfer of compounds across the interface of immiscible solvents are highly important thermodynamic data. They are a measure of the compounds' lipophilicity and allow assessing the compounds' biological activity, membrane permeability, intra- and intermolecular interactions, and so on [1–3]. Before a drug can elicit any effect it usually has to pass through at least one biological membrane by passive diffusion or by carrier-mediated uptake [4]. Transport and distribution processes of drugs within biological systems are to a large extent controlled by their lipophilicity. The lipophilicity is one of the crucial parameters included in empirical methods of quantitative structure property relationships (QSPR) and quantitative structure activity relationships (QSAR) [5, 6]. Consequently, it is a useful parameter in understanding the behavior of drug molecules. The usual measure of the lipophilicity of a compound is its partition coefficient

*To whom correspondence should be addressed. Phone: +86-931-891-2578; Fax: +86-931-891-2582; E-mail: liumc@lzu.edu.cn

Table 1. Compounds, descriptors, experiment and calculated ΔG^{0} (kJ/mol).

No.	Name	*KHI*2	ΔH_f	V_{avg}	E_{e-n}	ΔG^{0}(kJ/mol)			
						Experiment	HM	RBFNN	SVM
Training set									
1	Trp-Phe-	6.24	15.64	2.02	220.85	5.30	6.038	4.193	5.043
2	Trp-Tyr-	6.42	−27.95	2.03	221.29	7.40	9.707	10.087	10.182
3	Trp-	3.49	20.11	2.07	228.88	10.80	13.960	11.508	10.705
4	Trp-	3.49	20.46	2.10	229.15	10.80	14.958	10.742	11.056
5	Trp-Val-	5.72	−19.97	2.02	221.38	11.60	11.178	10.563	11.339
6	Trp-Gly-Tyr-	7.20	−56.43	2.08	228.02	15.00	11.919	14.385	14.744
7	Trp-Gly-Gly-Tyr-	7.97	−103.48	2.01	219.61	15.50	12.965	14.713	15.240
8	Trp-Gly-	4.26	−16.72	2.10	228.86	15.60	15.448	14.994	15.345
9	Trp-Gly-Gly-	5.04	−43.95	1.98	219.82	15.80	16.034	17.443	16.059
10	Gly-Trp-Gly-	4.98	−31.49	2.07	226.54	15.80	14.964	14.673	15.705
11	Tyr-Ala-Gly-Phe-Leu-	10.24	−199.08	2.08	235.98	16.60	21.841	20.663	22.051
12	Tyr-Ala-Gly-Leu-Arg-	10.08	−153.89	1.96	228.58	17.10	18.576	17.196	16.843
13	Tyr-Ala-Gly-Phe-Met-	10.25	−180.16	2.08	236.48	18.40	20.195	16.436	18.654
14	Gly-Gly-Trp-	4.98	−49.38	2.06	226.92	19.00	17.277	17.930	18.436
15	Gly-Phe-Phe-	6.08	−49.23	2.09	237.77	20.15	22.565	24.430	23.844
16	Gly-Phe-Tyr-	6.26	−92.02	2.09	238.29	20.20	18.570	20.929	20.405
17	Phe-	2.62	−1.89	2.12	233.94	21.00	20.071	22.145	21.008
18	Tyr-	2.80	−45.66	2.12	234.18	21.20	24.025	23.995	25.276
19	Leu-Leu-Leu-	7.94	−156.47	2.08	236.96	23.20	24.194	23.234	23.046
20	Gly-Leu-Tyr-	6.20	−128.42	2.09	241.01	23.40	27.640	25.227	26.176
21	Leu-	2.56	−47.10	2.12	234.47	23.90	24.942	24.586	26.024
22	Met-	2.69	−33.70	2.12	234.01	24.50	23.054	23.880	24.352
23	Tyr-Ala-Gly-	4.80	−123.45	2.10	234.49	24.90	20.168	23.025	24.640
24	Tyr-Ala-Gly-Met-Phe-Glycinol-	6.15	−93.88	2.08	234.49	24.90	27.291	27.442	27.377
25	Gly-Phe-	3.33	−38.03	2.10	235.59	25.00	23.295	25.393	24.765
26	Gly-Phe-Gly-	4.10	−68.47	2.11	234.59	25.60	23.406	24.328	25.340
27	Gly-Gly-Phe-	4.11	−77.08	2.09	236.15	26.00	25.722	27.116	26.259
28	Gly-Leu-Phe-	6.02	−83.37	2.09	240.84	26.20	23.594	23.806	25.332
29	Leu-Gly-Phe	6.08	−86.49	2.08	239.84	26.50	23.619	24.645	23.992
30	Gly-	0.58	−26.94	2.12	233.42	26.60	27.669	26.846	26.860
31	Gly-Gly−Leu-	4.05	−111.05	2.09	234.89	26.80	28.611	27.552	27.053
32	Val-	2.09	−42.21	2.12	233.92	26.80	25.432	24.792	25.765
33	Gly-Leu-Gly-	1.35	−61.37	2.11	234.69	27.00	28.125	26.667	26.747
34	Gly-Tyr-Gly-	4.28	−112.27	2.11	234.57	27.10	27.236	25.729	27.359
35	Lys-	2.27	−81.39	1.81	208.80	27.30	27.856	27.310	27.046
36	Gly-Ala-Phe-	4.56	−74.29	2.08	240.63	27.40	26.816	27.288	27.831
37	Ala-	4.56	−77.72	2.09	240.00	27.50	27.702	27.428	27.245
38	His-	2.25	−35.13	1.89	220.66	27.70	26.230	27.189	27.446
39	Leu-Leu-Gly-	6.02	−137.46	2.09	234.41	28.00	25.777	28.418	26.959
40	Leu-Leu-Ala-	6.48	−142.41	2.09	240.39	28.20	27.976	25.215	25.611
41	Gly-Gly-Tyr-	4.16	−71.38	2.10	234.96	29.00	29.471	27.511	28.062
42	Pro-	1.97	−29.69	2.11	236.80	29.50	26.290	28.254	28.797
43	Lys-Tyr-Thr-	6.87	−183.53	2.07	232.45	30.00	27.844	29.831	30.255
Prediction set									
44	Trp-Trp-	6.97	10.69	2.09	229.07	4.80	6.907	6.301	4.014
45	Trp-Leu-	6.18	−27.98	2.02	220.88	9.50	9.775	10.802	10.793

Table 1. (Continued).

No.	Name	*KHI2*	ΔH_f	V_{avg}	E_{e-n}	ΔG^{θ}(kJ/mol)			
						Experiment	HM	RBFNN	SVM
46	Trp-Ala-	4.72	−9.25	2.08	227.27	15.75	14.162	12.011	12.754
47	Leu-Leu-Phe-	8.00	−110.17	2.08	238.97	17.50	17.870	17.794	18.800
48	Leu-Leu-Tyr-	8.18	−151.61	2.08	241.42	19.70	21.523	17.511	19.780
49	Leu-Leu-	5.25	−98.71	2.11	237.60	23.70	23.348	25.778	27.235
50	Tyr-Lys-Thr-	6.87	−183.45	2.07	233.63	24.60	23.730	29.749	29.229
51	Gly-Gly-Val-	3.59	−111.77	2.09	235.06	26.40	28.346	27.009	27.207
52	Gly-Gly-e	4.04	−114.94	2.09	233.06	27.00	30.487	25.918	27.940
53	Gly-Phe-Ala-	1.11	−33.49	2.11	234.28	27.50	25.781	27.167	28.223
54	Phe-Gly-Gly-	4.29	−120.84	2.10	236.72	29.00	23.522	25.494	25.458

[7, 8], which is connected to the standard Gibbs energy of transfer of a compound by the following equation:

$$\log P = \frac{-\Delta G^{\theta}}{2.3RT} \qquad (1)$$

There are some experimental techniques developed to measure the standard Gibbs energies [9, 10]. However, there are some inherent limitations for these methods. For example, to the voltammetry at the interface of two immiscible electrolyte solutions (ITIES) with the help of a four-electrode electrochemical measurements, the narrow potential window (300–500 mV) as a result of the presence of electrolytes in both phases and the nonpolarizability of many organic solvents [11, 12] render its wider usage. Beside the above mentioned, the experimental determination of standard Gibbs energies is time-consuming and expensive. Alternatively QSPR provides a promising method for the estimation of ΔG^{θ} values based on descriptors derived solely from the molecular structure to fit experimental data. The advantages of this approach lie in the fact that it requires only the knowledge of chemical structure and is not dependent on any experiment properties. Once the structure of a compound is known, any descriptor can be calculated no matter whether it is obtained or not. So once a reliable model is established, we can predict the property of compounds and know which structural factors play an important role to the property.

Quantitative structure–property relationships are mathematical relationships linking chemical structure and thermodynamic property in a quantitative manner for a series of compounds. The main steps involved in this method include the following: data collection, molecular descriptor obtaining and selection, correlation model development, and finally model evaluation. Owing to there are no considerable or precise methods to determine standard Gibbs energies values and these fundamental data are not available for most of the simple organic anions, such as the anions of the aliphatic and aromatic acids, phenols, and so on [13]. We think QSPR method may be a potential way to predict the property and to our knowledge, there are no QSPR studies focused on the standard Gibbs energies of peptides anions.

Recently, the standard Gibbs energies of 54 peptides anions were measured by the three-phase electrode technique [14]. As we know, peptides acting as hormones, neurotransmitters, immunomodulators, coenzymes, enzyme substrates and inhibitors, receptor ligands, drugs, toxins, and antibiotics, play a significant role in controlling and regulating many vitally important processes in living organisms [15]. And they usually should be in the ionization state such as peptide anions to play their role and have the required physical properties such as standard Gibbs energies to further study.

In the present study, the support vector machine (SVM), as a novel type of learning machine, was used to develop a QSPR model of 54 peptides based on molecular descriptors calculated from the structure alone. Meanwhile, the artificial neural networks (ANN), radial basis function neural networks (RBFNN), was developed for the purposes. The heuristic method (HM) was also utilized to select descriptors, to establish linear model, and to compare the results with that obtained by SVM and RBFNN. The aim of this study is to establish a relatively accurate model to predict the standard Gibbs energies (ΔG^{θ}) of the transfer of peptide anions from aqueous solution to nitrobenzene, and to seek for the important structural features related to ΔG^{θ} property of peptides anions.

Experiment section

Data set

A total of 54 of the standard Gibbs energies (ΔG^{θ}) of the transfer of peptide anions from aqueous solution to nitrobenzene were obtained from the literature [14]. All of the data of the ΔG^{θ} of the investigated peptides are presented in Table 1. The entire set of compounds was randomly divided into two subsets through diversity analysis: a training set, whose information was used to build the models, and a prediction set, consisting of molecules not found in the training set, which was used to evaluate the models once they were built.

Diversity analysis

Two fundamental research themes in chemical database analysis are similarity and diversity sampling [16]. The diversity problem involves defining a diverse subset of "representative" compounds so that researchers can scan only a subset of the huge database each time. In this study, diversity analysis was performed for the data set to make sure the structures of the training or test cases can represent those of the whole ones.

We consider a database of n compounds generated from m highly correlated chemical descriptors $\{x_j\}^{m}{}_{j=1}$. Each compound X_i is represented as a vector

$$X_i = (x_{i1}, x_{i2}, x_{i3} \ldots x_{im})^T \quad for\ i = 1, 2, \ldots, n$$

where x_{ij} denotes the value of descriptor j of compound X_i. The collective database $X = \{X_I\}_{I=1}{}^{N}$ is represented by the $n \times m$ matrix X:

$$X = (X_1, X_2, \ldots, X_N)^T = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1m} \\ x_{21} & x_{22} & \cdots & x_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nm} \end{bmatrix}$$

Here the superscript T denotes the vector/matrix transpose.

A distance score for two different compounds X_i and X_j can be measured by the Euclidean distance norm based on the compound descriptors:

$$d_{ij} = \|X_i - X_j\| = \sqrt{\sum_{k=1}^{m} (x_{ik} - x_{jk})^2}$$

The mean distances of one sample to the remaining ones were computed as follow:

$$\overline{d_i} = \frac{\sum_{j=1}^{n} d_{i,j}}{n-1} \quad i = 1, 2, \ldots, n$$

And then the mean distances were normalized within the interval [0, 1]. The closer to one the distance is, the more diverse to each other the compound is. For the data sets, the mean distances of samples vs. the experimental ΔG^{θ} was shown in Figure 1, which illuminates the diversity of the molecules in the training and prediction sets. As can be seen from the figure, the structures of the compounds are diverse in both sets. The training set with a broad representation of the chemistry space was adequate to ensure models' stability and the diversity of prediction set can prove the predictive capability of the model.

Generation of molecular descriptors

All structures of the peptides were drawn with the HyperChem program [17] and exported in a file format suitable for MOPAC [18]. All calculations were carried out at restricted Hartree-Fock level

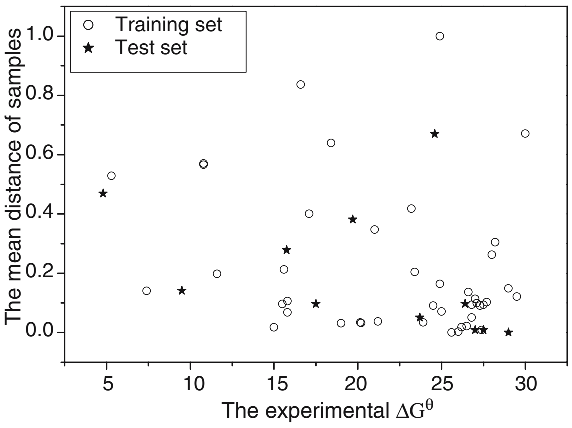

Figure 1. Scatter plot of samples for training and test set.

with no configuration interaction. The molecular structures of anion were optimized using the Polak-Ribiere algorithm until the root-mean-square gradient was 0.1. A more precise optimization is done with the semiempirical PM3 method in MOPAC6.0. The resulting geometry was transferred into software CODESSA, developed by the Katritzky group [19, 20], which can calculate constitutional, topological, geometrical, electrostatic, and quantum chemical descriptors and has been successfully used in various QSPR and QSAR researches. Constitutional descriptors are related to the number of atoms and bonds in each molecule. Topological descriptors include valence and nonvalence molecular connectivity indices calculated from the hydrogen-suppressed formula of the molecule, encoding information about the size, composition, and the degree of branching of a molecule. The topological descriptors describe the atomic connectivity in the molecule. The geometrical descriptors describe the size of the molecule and require 3D-coordinates of the atoms in the given molecule. The electrostatic descriptors reflect characteristics of the charge distribution of the molecule. The quantum chemical descriptors offer information about binding and formation energies, partial atom charge, dipole moment, and molecular orbital energy levels. In the present work, five classes of structural descriptors were obtained and about 480 descriptors were provided.

Methodology

Theory of HM [20, 21]

After the calculation of a large number of descriptors, HM in CODESSA was used to select descriptors and built the linear model. Successful QSPR depends on good descriptors selection. If molecular structures are represented by improper descriptors, they will not lead to reasonable predictions. In recent years, methodology for a general QSPR approach has been developed and coded as the CODESSA software package, which combines different ways of quantifying the structural information about the chemicals with advanced statistical analyses for the establishment of molecular structure–property relationships. To find the best QSPR model, the correlation analysis was carried out using HM, which is based on the scale forward selection technique.

The advantages of this method are: the high speed usually produces correlations 2–5 times faster than other methods, with comparable quality; and no software restrictions on the size of the data set. HM can either quickly give a good estimation about what quality of correlation to expect from the data, or derive several best regression models. Besides, it will demonstrate which descriptors have bad or missing values, which descriptors are insignificant (from the standpoint of a single-parameter correlation), and which descriptors are highly inter-correlated.

The HM provides collinearity control (i.e., any two descriptors intercorrelated above 0.8 are never involved in the same model) and implement heuristic algorithms for rapid selection of the best correlation, without testing all the possible combinations of the available descriptors. HM of the descriptors selection proceeds with a pre-selection of the descriptors to eliminate: (1) those descriptors that are not available for each structure; (2) descriptors having a small variation in magnitude for all structures; (3) descriptors that provide a F-test's value below 1.0 on the one-parameter correlation, and (4) and the descriptors whose t-values are less than the user-specified value, etc.

Following the pre-selection of descriptors, multiple linear regression models are developed. The selection of best correlations proceeds as follows: (1) Beginning with the top descriptor from the pre-selected list of descriptors, the two-parameter correlations are calculated using the following pairs: the first descriptor with each of the remaining descriptors, second descriptor with each of the remaining descriptors, etc. This procedure is continued until for some n-th descriptor no correlations with an F-test value above one-third of the maximum F-test value for a given set are found. (2) The best pairs of branching criteria (number of descriptors sets to select for next recursion level) with highest F-test values in the two-parameter correlations are selected and processed further as the working sets. (3) If not correlated over rsig (descriptors are considered to be noncollinear below the value of their pair correlation coefficient) with the descriptors already included, each of the remaining descriptors is added to the selected working set of descriptors. If the resulting correlation gives F-test value above F working $n/(n+1)$ (where n is a number of descriptors in the working set plus one), i.e., if this correlation is more significant than the working correlation, then this extended set of descriptors is considered for further treatment. (4) After all descriptors have been applied one-by-one and if the maximum number of descriptors, allowed by the user, is not yet achieved, then best extended working sets, i.e., the sets with the highest *F*-values, are submitted to the procedure from step (3). Otherwise the procedure is completed and the maximum number of descriptors best correlations found. The goodness of the correlation is tested by the coefficient regression (R^2) and the *F*-test values (F).

Theory of SVM [22, 23]

SVM, as a novel type of learning machine, is gaining popularity due to many attractive features and promising empirical performance. It can solve high-dimension problems and therefore avoid the "curse of dimensionality". Originally, SVM is developed for pattern recognition problems. And now, with the introduction of ε-insensitive loss function, SVM has been extended to solve regression estimation and time series prediction [24]. Comparing with traditional neural networks, SVM possesses the prominent advantages including high generalization capability, avoiding local minima, always having solution by a standard algorithm (quadratic programming), automatically obtaining network topology structure, and lower workload. A detailed description of the theory of SVM can be referred in several excellent books and tutorials. In our previous study, the theory of SVM for regression has been introduced in detail [25]. So here, we introduce the basic idea and its performance simply in the following.

SVM can be applied to regression by the introduction of an alternative loss function and results appear to be very encouraging. In support vector regression (SVR), the input x is first mapped into a higher dimensional feature space by the use of a kernel function, and then a linear model is constructed in this feature space. Any function that satisfies Mercer's condition (kernel function must be symmetric, and it must be positive definite) can be used as the kernel function. The kernel functions often used in SVM include linear, polynomial, radial basis function, and sigmoid function, etc. For regression tasks, the Gaussian radial basis function kernel is often used because of its effectiveness and speed in training process. The form of the Gaussian function in R is

$$\exp\left\{-\gamma(u-v)^2\right\} \tag{2}$$

where γ is the parameter of the kernel, u and v are two independent variables.

The basic idea of SVR is it approximates the function by minimizing the regularized risk function

$$R(C) = C\frac{1}{n}\sum_{i=1}^{n} L_{\in}(d_i, y_i) + \frac{1}{2}\|w\|^2 \tag{3}$$

where

$$L_{\in}(d,y) = \left\{\begin{array}{cc} |d-y| - \varepsilon & |d-y \geq \in| \\ 0 & \text{otherwise} \end{array}\right\} \tag{4}$$

and $\in$ is a prescribed parameter. In Equation 3, $C(1/n)\sum_{i=1}^{n} L_{\in}(d_i, y_i)$ is the so-called empirical error (risk), which is measured by $\in$-insensitive loss function L_ε *(d, y)*, which indicates that it does not penalize errors below $\in$. The second term, $1/||w||^2$, is used as a measurement of function flatness. C is a regularized constant determining the tradeoff between the training error and the model flatness.

The generalization performance of SVR depends on a good setting of parameters: C, ε and the kernel type and corresponding kernel parameters. The selection of the kernel function and corresponding parameters is very important because they define the distribution of the training set samples in the high dimensional feature space. The Gaussian radial basis function kernel was used for the SVR model in our study due to advantages mentioned above. γ, the parameter of the kernel, controls the amplitude of the Gaussian function and, further, controls the generalization ability of SVM. The optimal value for ε depends on the type of noise present in the data, which is usually unknown. Even if enough knowledge of the noise is available to select an optimal value for ε, there is the practical consideration of the number of resulting support vectors. ε-insensitivity prevents the entire training set meeting boundary conditions, and so allows for the possibility of sparsity in the dual formulation's solution. The value of ε can affect the number of support vectors used to construct the regression function. The bigger ε, the fewer support vectors are selected.

C is a regularization parameter, which controls the tradeoff between maximizing the margin and minimizing the training error. If C is too small, then insufficient stress will be placed on fitting the training data. If C is too large, then the algorithm will overfit the training data. To make the learning process stable, a large value should be set up for C.

To select the proper values for the regulation parameter C, γ and ε, different values for these parameters should be tried; the set of values with the best leave-one-out (LOO) cross-validation performance will be selected for further analysis. The overall performances of SVM were evaluated in terms of root mean square error (RMS), which was defined as below

$$\mathrm{RMS} = \sqrt{\frac{\sum_{i=1}^{n}(d_i - o_i)^2}{n}} \tag{5}$$

where d_i is the desired output, o_i is the actual output of the model, and n is the number of samples in the analyzed set.

Theory of RBFNN [26, 27]

RBFNN consists of an input layer, a hidden layer, and an output layer. The input layer does not process the information; it only distributes the input vectors to the hidden layer. The hidden layer of RBFNN consists of a number of RBF units (n_h) and bias (b_k). Each hidden layer unit represents a single radial basis function, with associated center position and width. Each neuron on the hidden layer employs a radial basis function as a nonlinear transfer function to operate on the input data. The most often used RBF is a Gaussian function that is characterized by a center (c_j) and a width (r_j). In this study, the Gaussian was selected as a radial basis function. The operation of the output layer is linear, which is given as below

$$y_k(x) = \sum_{j=1}^{n_k} w_{kj}h_j(x) + b_k \tag{6}$$

where y_k is the kth output unit for the input vector x, w_{kj} is the weight connection between the kth output unit and the jth hidden layer unit, and b_k is the bias. The training procedure when using RBF involves selecting centers, width and weights. In this paper, the forward subset selection routine was used to select the centers from training set samples. The adjustment of the connection weight between the hidden layer and the output layer was performed using a least-squares solution after the selection of centers and width of radial basis functions.

Algorithm implementation and computation environment

All calculation programs implementing SVM were written in an R-file based on the R script for SVM. All calculation programs implementing RBFNN were written in M-file based on a basis MATLAB script for RBFNN. The scripts were run on a Pentium IV PC with 256M RAM.

Results and discussion

Results of HM

Through HM implemented in CODESSA, the best linear model with four parameters was obtained, which is shown in Table 2. The correlation matrix of the four selected descriptors was shown in Table 3. From Table 3, it can be seen that the linear correlation coefficient value of each of the two descriptors is <0.80, which means the descriptors are independent in the analysis and this approximate correction device avoid serious overestimation of chance correlation effects [20, 28]. This model gave an RMS error of 2.191 for the training set, 2.541 for the prediction set, and 2.267 for the whole set, and the corresponding correlations coefficients (R) were 0.938, 0.911 and 0.920, respectively. The calculated and experimental values of ΔG^{0} were given in Table 1, the scatter plot was shown in Figure 2.

Result of SVM and RBFNN

To obtain more accurate model, after the linear model was established, we built the nonlinear prediction model by SVM and RBFNN to further study the correlation between the molecular structure and the ΔG^{0} based on the same subset of descriptors.

The selection of the value for SVM was performed by systemically changing its value on the training step. The value which gives the best

Table 2. Descriptors, coefficients, standard error, and *t*-test values for the linear model.

No.	Descriptor	Coefficient	Standard error	*t*-test
0	Intercept	−1.1343e+1	1.2672e+1	−0.8951
1	Kier and Hall index (order 2) (*KHI*2)	−2.7152e+0	1.9996e−1	−13.5786
2	Final heat of formation (ΔH_f)	−8.7861e−2	9.6064e−3	−9.1461
3	Min e−n attraction for a N atom (E_{e-n})	5.1117e−1	9.8032e−2	5.2143
4	Avg valency of a O atom (V_{avg})	−3.7953e+1	1.0448e+1	−3.6327

$R^2=0.881$, $N=43$, $F=70.11$, RMS = 2.192.

LOO cross-validation result was used in the model. The parameters here include capacity parameter C, ε of ε-insensitive loss function, and γ controlling the amplitude of the Gaussian function. Since three parameters exhibits strong interactions, grid search (GS), which has been used either formally or informally for SVM parameter selection, was performed in this study. In the grid search, we considered the parameter γ from 0.0001 to 0.1 with 0.0001 as the increment. The parameter C was chosen from values between 100 and 1000 with 100 as the increment and 1000 as the increment within 1000 to 10000. The parameter ε was searched with 0.001 increments within 0.001 to 0.1. LOO cross-validation was performed for parameters selection, the overall performances of SVM were evaluated in terms of RMS. The γ, ε, and C for this data set were finally fixed to 0.04, 0.04, and 800, respectively. The predicted results of the nonlinear models are shown in Table 1 and Figure 3. The models gave RMS of 1.562 for the training set, 2.364 for the prediction set and overall data sets are 1.756, and the corresponding correlation coefficients (R) were 0.970, 0.960 and 0.966, respectively.

Table 3. Correlation matrix of the four descriptors used in this work.

	*KHI*2	ΔH_f	V_{avg}	E_{e-n}
*KHI*2	1.000			
ΔH_f	−0.5762	1.000		
V_{avg}	−0.1368	−0.0005	1.000	
E_{e-n}	0.0760	−0.4038	0.7601	1.000

The parameters that influence the performance of RBFNN were also optimized using the same procedure on training set. For this data set, the optimal spread value was determined as 1.1 and the corresponding number of centers (hidden layer nodes) of RBFNN is 15. The predicted results of the nonlinear models are shown in Table 1 and Figure 4. The RMS errors of the training set, the prediction set and the whole data are 1.604, 2.478

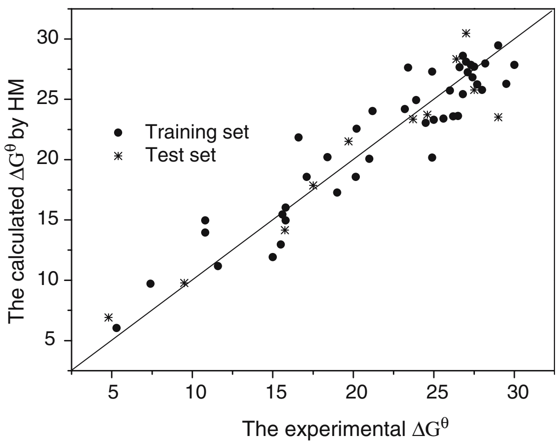

Figure 2. Predicted vs. experimental ΔG^{θ} (kJ/mol) by HM.

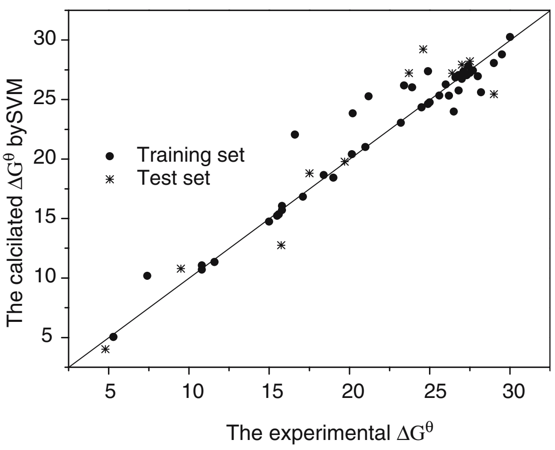

Figure 3. Predicted vs. experimental ΔG^{θ} (kJ/mol) by SVM.

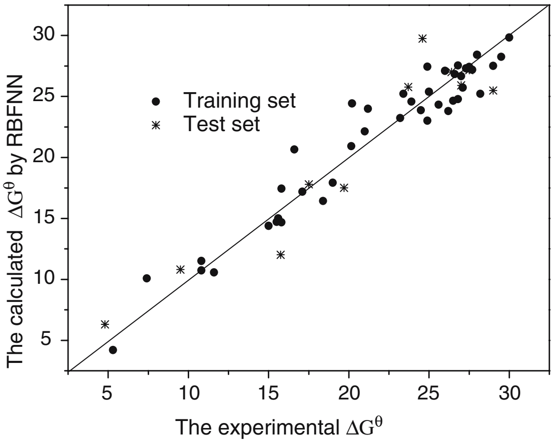

Figure 4. Predicted vs. experimental ΔG^{θ} (kJ/mol) by RBFNN.

and 1.817, and the correlation coefficient are 0.968, 0.947 and 0.962, respectively.

The statistical parameters of different QSPR models were listed in Table 4. Comparison the results obtained by HM, RBFNN and SVM, it can be seen that the results of nonlinear model are better than linear model. And of the nonlinear model, SVM is better than RBFNN.

Discussion of the input descriptors

By interpreting the descriptors in the QSPR model, it is possible to gain some insight into factors that are likely to govern the ΔG^{θ} of the transfer of peptide anions from aqueous solution to nitrobenzene. This model contains one topological (Kier and Hall index (order 2)) and three quantum chemical (Final heat of formation, Avg valency of a O atom and Min e–n attraction for a N atom), descriptors. These descriptors encoded different aspects of the molecular structure.

The Kier and Hall index (order 2), *KHI*2, belongs to the well-known valence connectivity indices. The Kier and Hall indices ${}^{m}X^{v}$ defined by Equations 7 and 8 belong to the same "family" of descriptors.

$${}^{m}X^{v} = \sum_{i=1}^{N_S} \prod_{k=1}^{m+1} \left(\frac{1}{\delta_k^v} \right)^{1/2} \qquad (7)$$

$$\delta_k^v = \frac{z_k^v - H_k}{Z_k - Z_k^v - 1} \qquad (8)$$

In Equations 7 and 8, Z_k is the total number of electrons in the *k*th atom, Z_k^v is the total number of valence electrons in the *k*th atom, H_k is the number of hydrocarbon atoms directly attached to the *k*th nonhydrocarbon atom, $m = 0$ is the atomic valence connectivity indices, $m = 1$ – one bond path valence connectivity indices, $m = 2$ – two bond valence connectivity indices, and $m = 3$ – three contiguous bond fragment valence connectivity indices, etc. Kier and Hall have recently interpreted the molecular connectivity in terms of intermolecular accessibility starting from the interpretation of the bond contributions. Thus, they have concluded that [29] "the molecular connectivity index is the contribution of one molecule to the bimolecular interactions arising from encounters of bonds among two molecules". The significant negative coefficient of size related descriptors in the models indicate the higher probability of interaction leads to lower ΔG^{θ} values.

The *final heat of formation* (ΔH_f) is a quantum mechanical energy-related descriptor that gives the energy of the molecule in the thermodynamic standard scale (elements in ideal gas state at 298.15 K and 101,325 Pa). The negative values combined with the negative coefficient of this descriptor in the model indicated that increasing the value of this descriptor can lead to the larger values of standard Gibbs energies.

The *Min e–n attraction for a N atom* ($E_{e\text{–}n}$) is another quantum mechanical energy-related descriptor that is defined as extreme (maximum or

Table 4. Comparison between the different statistical parameters of HM, RBFNN and SVM models.

Method	R_t	R_p	$R^2{}_t$	$R^2{}_p$	RMS_t	RMS_p	RMS_{tot}
HM	0.938	0.911	0.881	0.830	2.192	2.541	2.267
RBFNN	0.968	0.947	0.937	0.897	1.604	2.478	1.817
SVM	0.970	0.960	0.941	0.922	1.562	2.364	1.756

minimum) values of the nuclear-electron attraction energy for a given atomic species (atom A) in the molecule, calculated as follows:

$$E_{ne}(A) = \sum_{B} \sum_{\mu,\nu\varepsilon a} p_{\mu\nu} \langle \mu | Z_B / R_{iB} | \nu \rangle \qquad (9)$$

The first summation in Equation 9 is performed over all atomic nuclei in the molecule and $\langle \mu | Z_B / R_{iB} | \nu \rangle$ denote the nuclear–electron attraction integrals on the given atomic basis. This energy describes the nuclear–electron attraction driven processes in the molecule and may be related to the conformational (rotational, inversional) changes or atomic reactivity in the molecule. The descriptor bears positive sign in the linear model indicating a large (in magnitude) value of it should and does lead to higher ΔG^{θ} values.

The final descriptor, *average valency of a O atom* (V_{avg}) is a quantum mechanical valency-related descriptor. These descriptors relate to the strength of intramolecular bonding interactions and characterize the stability of the molecules, their conformational flexibility and other valency-related properties. The negative coefficient in the linear model indicates that ΔG^{θ} values are proportional to this descriptor.

Analysis of the results indicated that the models we proposed correctly represent the structural-property relationship of these peptides and that molecular descriptors calculated solely from structures can represent the structural features of the compounds responsible for their ΔG^{θ} values.

Conclusion

Linear and nonlinear QSPR models to predict ΔG^{θ} of 54 peptides anions were built based on HM, RBFNN and SVM using the calculated topological and quantum chemical descriptors alone. Comparison the results obtained by the models, it is proved that nonlinear SVM model gave best results. It gave RMS of 1.756 for the whole data set, which means QSPR is a potential method for predicting the physiochemical property (ΔG^{θ}) of various small peptides. Furthermore, the four descriptors selected in the model may help to understand transfer actions and to predict the ΔG^{θ} values of many new small peptides anions.

Acknowledgements

The authors thank the National Natural Science Foundation of China (NSFC) Fund (NO.20305008) for financial support. The authors also thank the Association Franco-Chinoise pour la Recherche Scientifique & Technique (AFCRST) for supporting this study (Programme PRA SI 02–03).

References

1. Volkov, A.G. Liquid Interfaces in Chemical, Biological, and Pharmaceutical Applications. Marcel Dekker, New-Yorks Basel, pp. 2001.
2. Lyman, W.J., Reehl, W.F. and Rosenblatt, D.H. (Eds.), Handbook of Chemical Property Estimation, American Chemical Society, Washington, DC, 1990.
3. Hansch, C., Quinlan, J.E. and Lawrence, G.L., J. Org. Chem., 33 (1968) 347.
4. Li, X., Glen, R.C. and Clark, R.D., J. Chem. Inf. Comput. Sci., 43 (2003) 870.
5. Plass, M. Habilitation Thesis, Martin Luther University, Halle-Wittenberg, Germany, 2000 Chapters 1 and 2.
6. Testa, B., van de Waterbeemd, H., Folkers, G. and Gay, R. Pharmacokinetic Optimization in Drug Research Chapter 6. Wiley-WCH, Weinheim, Germany, 2001, pp. 591–613.
7. Reymond, F., Steyaert, G., Carrupt, P.A., Testa, B. and Girault, H.H., J. Am. Chem. Soc., 118 (1996) 11951.
8. Testa, B., van de Waterbeemd, H., Folkers, G. and Gay, R. Pharmacokinetic Optimization in Drug Research Chapter 6. Wiley-WCH, Weinheim, Germany, 2001, pp. 275–304.
9. Gulaboski, R., Mirceski, V. and Scholz, F., Amino Acids, 24 (2003) 149.
10. Komorsky-Lovrić, S., Riedl, K., Gulaboski, R., Mircjeski, V. and Scholz, F., Langmuir, 18 (2002) 8000.
11. Marcus, Y. Ion Properties. Marcel Dekker, New York, 1997, pp. 212–219.
12. Volkov, A.G. Liquid Interfaces in Chemical, Biological and Pharmaceutical Applications Vol. 95. Marcel Dekker, New York, 2001, pp. 729–773 Chapter 3.
13. Marcus, Y. Ion Properties. Marcel Dekker, New York, 1997.
14. Gulaboski, R. and Scholz, F., J. Phys. Chem. B, 107 (2003) 5650.
15. Liu, H.X., Xue, C.X., Zhang, R.S., Yao, X.J., Liu, M.C., Hu, Z.D. and Fan, B.T., J. Chem. Inf. Comput. Sci., 44 (2004) 1979.
16. Maldonado, A.G., Doucet, J.P., Petitjean, M. and Fan, B.T., Mol Divers, 10 (2006) 39.
17. HyperChem 6.01, Hypercube, Inc., 2000.
18. MOPAC, v.6.0 Quantum Chemistry Program Exchange, Program 455, Indiana University, Bloomington, IN.
19. Katritzky, A.R., Lobanov, V.S. and Karelson, M. CODESSA: Training Manual. University of Florida, Gainesville, FL, 1995.
20. Katritzky, A.R., Lobanov, V.S. and Karelson, M. CODESSA: Reference Manual. University of Florida, Gainesville, FL, 1994.

21. Luan, F., Xue, C.X., Zhang, R.S., Zhao, C.Y., Liu, M.C., Hu, Z.D. and Fan, B.T., Analytica Chimica Acta, 537 (2005) 101.
22. Vapnik, V.N. Statistical Learning Theory. John Wiley & Sons, New York, 1998.
23. Schölkopf, B. and Smola, A. Learning with Kernels. MIT Press, Cambridge, MA, 2002.
24. Tay, F.E.H. and Cao, L.J., Neurocomputing, 48 (2002) 847.
25. Luan, F., Zhang, R.S., Liu, M.C., Hu, Z.D. and Fan, B.T., QSAR Comb. Sci., 24 (2005) 227.
26. Derks, E.P.P.A., Sanchez Pastor, M.S. and Buydens, L.M.C., Chemom. Int. Lab. Sys., 28 (1995) 49.
27. Xiang, Y.H., Liu, M.C., Zhang, X.Y., Zhang, R.S. and Hu, Z.D., J. Chem. Inf. Comput. Sci., 42 (2002) 592.
28. Topliss, J.G. and Edwards, R.P., J. Med. Chem., 22 (1979) 1238.
29. Kier, L.B. and Hall, L.H., J. Chem. Inf. Comput. Sci., 40 (2000) 792.