Trade Science Inc.

January 2008 Volume 7 Issue 3

Analytical CHEMISTRY
An Indian Journal

Full Paper

ACAIJ, 7(3) 2008 [122-130]

A simple approach to the prediction of electrophoretic mobilities of some peptides

Rahmatollah Rajabzadeh[1]*, Maziar Ahmadi Golsefidi[2], Gozal Inchebroni[3]

[1]Chemistry Department, Faculty of Sciences, K.N.Toosi University of Technology, Tehran, (IRAN)

[2]Chemistry Department, Faculty of Sciences, Islamic Azad University, Gorgan Branch, Gorgan, (IRAN)

[3]Chemistry Department, Faculty of Sciences, University of Mazandaran, Babolsar, (IRAN)

E-mail: rahmat_rajabzadeh@yahoo.com

Received: 18th April, 2007 ; Accepted: 23rd April, 2007

ABSTRACT

In this study, based on molecular structure descriptors, by the use of partial least squares analysis, a good prediction quantitative structure-property relationship for the mobility of some Peptides was obtained. Five classes of molecular descriptors including topological, geometrical, functional group, empirical and properties descriptors were generated for each peptide. Constant and near constant variables exclude from descriptors and then descriptors with pair correlations 0.95 exclude, the total remaining descriptors were 72 and we use these descriptors for Genetic Algorithm (GA) variable selection. A series of 12 peptides were selected randomly to check the prediction ability of the obtained model. The RMSEP was 0.492 for selected model.

KEYWORDS

QSMR;
Peptide;
Electrophoretic mobility;
Partial least squares.

INTRODUCTION

Capillary electrophoresis(CE) has become an important separation technique and alternate to other analytical methods like high performance liquid chromatography (HPLC) due to its simple preprocessing, high separation efficiency, low operating costs and solvent consumption. It has been widely applied in the analysis of both small and large molecules, such as inorganic ions, organic acids, carbohydrates, pharmaceuticals, and even living cells[1].

Electrophoretic separation is based on the difference in migration of ions under a certain applied electric field. Generally, the migration behavior is denoted by electrophoretic mobility(μ), which is dependent on both the molecular structure and the separation condition. Therefore, the prediction of the mobilities of ions by theoretical calculation will relieve analysts of a large number of costly and time consuming experiments to develop a faster optimization process in CE. More and more investigators have paid attention to this problem and some papers have contributed to the study of quantitative relationship between molecular structures and electrophoretic mobilities. Based on the published reports, two principal methods can be summarized, i.e., the mechanistic and the statistical methods.

The mechanistic models are closely related to the mechanism of electrophoretic separation. The basic expression of such method is Max Born's model[2].

$$\mu = \frac{q}{fh + fdl} \qquad (1)$$

where q is the effective charge on the ion, f_h is the hydrodynamic friction related to molecular size and shape, and f_{dl} is the dielectric friction acused by the orientation of the solvent dipoles in response to the ionic charge.

In most cases, the dielectric friction is overlooked and only the hydrodynamic friction is considered, such as Huckel equation[3]:

$$\mu = \frac{q}{6\pi\eta\lambda} \qquad (2)$$

where η is the viscosity of medium and γ is the Strokes's radius of the ion($\gamma=[V/(4/3\pi)]^{1/3}$, where V is the van der Waals volume of the molecule). Due to the dependence of hydrodynamic frictional drag on the molecular size, the electrophoretic mobilities were also modeled by charge/mass or charge/volume ratio[4-10].

The statistical models are based on the quantitative structure mobility relationship(QSMR). This approach aims to get high predictive performance with relatively less consideration to the mechanism of separation. One of the most important factors governing the quality of QSMR model is the quantification of structural features, i.e., the extraction of molecular descriptors. Both new descriptors developed by oneself and existing descriptors embodied in commercial special softwares can be used to build linear or nonlinear models by techniques such as multiple linear regression(MLR), artificial neural networks(ANNs) and support vector machines (SVM). The electrophoretic mobilities of a variety of compounds have been investigated in this way[11-16].

According to the present chemometric theory, relevant data should be considered in QSMR studies as many as possible because this increases the probability of a good characterization of compounds[17]. As a consequence of the increase of the number of descriptors, the inter-correlation of independent variables (multicol linearity) will become more important. Under these circumstances, regression analysis(a method frequently used in QSMR studies) will not be useful, specially when the number of observations in the training set is less than four or five times the number of independent variables in a model. To overcome this problem, the partial least squares(PLS) method, a widely used chemometric method first developed by Wold et al.[18], will be used in this study. PLS finds the relationship between a matrix Y(containing dependent variables-usually only one for QSMR studies) and a matrix X (predictor variables) by reducing the dimension of the independent and dependent variables, and at the same time, maximizing the relationship between the independent and dependent matrices.

This work intents to construct a QSMR predictive model with a set of molecular structural descriptors to provide the analysts with a clear and convenient tool to estimate the absolute electrophoretic mobilities(μ_0) of peptides. The absolute electrophoretic mobility is a constant characteristic of an ion, which is measured experimentally either by extrapolating the mobilities observed over a range of ionic strength to infinite dilution or by measuring their limiting equivalent conductance.

MATERIALS AND METHODS

The absolute electrophoretic mobilities of thirty five

TABLE 1: Mobility(μ_0) values of peptides

^{T}Peptid	$\mu_{0(obs)}$	$\mu_{0(pred)}$	Diff	E-pred%
^{T}Gly-Gly	31.500	30.450	1.050	3.330
^{P}Gly-Leu	25.100	25.499	-0.399	-1.930
^{T}Gly-Thr	26.300	26.871	-0.571	-1.870
^{P}Gly-Ser	28.100	27.720	0.380	2.190
^{T}Gly-Asn	27.500	27.072	0.428	2.100
^{T}Gly-Phe	24.800	24.438	0.362	2.350
^{T}Gly-Trp	23.600	23.322	0.278	3.150
^{P}Ala-Gly-Gly	25.000	25.532	-0.532	-2.180
^{T}Gly-Gly-Ileu	21.900	21.323	0.577	2.000
^{T}Gly-Gly-Phe	21.900	22.154	-0.254	-2.380
^{T}Gly-His-Gly	22.500	23.108	-0.608	-2.660
^{P}Gly-Gly-Gly	26.100	26.081	0.019	2.460
^{T}Gly-Gly-Val	22.600	22.634	-0.034	-2.240
^{P}Ala-Ala	27.000	27.018	-0.018	-2.590
^{T}Ala-Ala-Ala	22.200	22.984	-0.784	-2.830
^{T}Ala-Leu	23.900	24.238	-0.338	-1.450
^{P}Ala-Val	25.200	24.438	0.762	2.380
^{T}Ala-Ser	26.200	26.743	-0.543	-1.950
^{T}Ala-Asn	25.500	27.046	-1.546	-2.500
^{P}Ala-Met	24.200	23.624	0.576	2.560
^{T}Ala-Phe	23.900	23.815	0.085	2.490
^{P}Gly-Gly-Gly-Gly	23.600	22.950	0.650	3.200
^{P}Ala-Leu-Gly	21.300	23.002	-1.702	-2.570
^{T}Leu-Leu	21.600	21.977	-0.377	-1.770
^{T}Leu-Leu-Leu	17.600	17.883	-0.283	-4.450
^{T}Leu-Val	22.300	22.494	-0.194	-3.000
^{T}Leu-Phe	21.800	21.938	-0.138	-2.370
^{P}Gly-Leu-Tyr	21.000	20.206	0.794	3.010
^{T}Gly-Phe-Phe	19.700	19.390	0.310	3.710
^{T}Leu-Gly-Phe	19.300	19.088	0.212	4.280
^{P}Ser-Ser-Ser	22.000	21.717	0.283	4.100
^{T}Gly-Pro-Ala	22.500	22.117	0.383	2.790
^{T}Ala-Gly	28.800	28.505	0.295	3.210
^{T}Gly-Val	26.000	25.675	0.325	2.710
^{P}Gly-Ileu	25.200	24.648	0.552	1.450

^{T}Training; ^{P}Prediction. μ_0(Obs): observed values determined by Wronski et al.[7]; μ_0(Pred): predicted values by model (10) of this study; Diff: $_0$(Obs) -μ_0(Pred); E-Pred: relative errors of the predicted values.

peptides were considered in the study, μ_0 values of the peptides were obtained directly by Wronski et al.[7]. The 23 peptides for which μ_0 values were determined directly served as the training set of the study, and the other 12 peptides for which μ_0 values were used as the validation set to test and verify the QSMR models. These μ_0 values are listed in TABLE 1 to aid discussion.

A Pentium IV personal computer with Windows XP operating system was used. The programs needed Genetic Algorithm(GA) variable selection was used from genpls(MATLAB code written by R.Leardi). Partial least squares regression was performed by the XLSTAT 2006 version 2006.2 Add-in software (XLSTAT company). For the calculation of molecular descriptors, the Hyper chem version 7.5[19] and Dragon(Milano chemometrics group, version 3.0)[20] softwares were used.

Molecular descriptors define the molecular structure and physicochemical properties of molecules by a single number. A wide variety of descriptors have been reported for using in QSAR analysis[21-26]. Here, 463 descriptors, 5 classes of Dragon descriptors including topological, geometrical, functional group, empirical and properties descriptors were generated for each compound(TABLE 2). Constant and near constant variables exclude from descriptors and then descriptors with pair correlations 0.95 exclude, the total remaining descriptors were 72 and we use these descriptors for Genetic Algorithm(GA) variable selection.

GAPLS

The GA algorithm is described in[27,28], and is implemented in MATLAN 4.0. Reference[28] covers the details of the settings required in the software, which are summarized in TABLE 3. GAPLS is a sophisticated hybrid approach that combines GA[29] as a powerful optimization method with PLS[30-33] as a robust statistical method for variable selection. The combination of variables and the internal predictivity of the derived PLS model in GAPLS correspond a chromosome and its fitness in GA, respectively. GAPLS consists of three basic steps. (1) An initial population of chromosomes is created. Each chromosome is a binary bit string, by which the existence of a variable is represented. (2) A fitness of each chromosome in the population is evaluated by the internal predictivity of PLS. (3) The population of chromosomes in the next generation is reproduced. Three operations, i.e., selection, cross-over and mutation of chromosomes, are made in this step. In the overall scheme, steps 2 and 3 are continued until the number of the repetitions is reached at the designated number of generations.

TABLE 2 : The calculated descriptors used in this study

Descriptor type	Molecular descriptors	No. of descriptors
Topological indices	Molecular size index, molecular connectivity indices, information contents, Kier shape indices, path/walk-Randic shape indices, Zagreb indices, Schultz indices, Balaban J index, Wiener indices, information contents …	266
Geometrical	3D-Wiener index, average geometrical distances, molecular eccentricity, spherocity, average shape profile index ..	70
Functional group	Numbers of different types of carbons, number of allenes groups, number of esters (aliphatic or aromatic), number of amides, number of different functional groups, number of CH3R, number of CR4, number of different halogens attached to different type of carbons, number of PX3, number of PR3 and …	121
Empirical descriptors	Unsaturation index, hydrophilic factor, aromatic ratio	3
Properties	Molar refractivity, polar surface area, LogP	3

TABLE 3 : Parameters of the GA

Population size	30 chromosomes(on average, five variables per chromosome in the original population)
Regression method	PLS
Response	Cross-validated % explained variance (five deletion groups; the number of components is determined by cross-validation)
Maximum number of variables selected in the same chromosome	30
Probability of mutation	1%
Maximum number of components	The optimal number of components determined by cross-validation on the model containing all the variables (no higher than 15)
Number of runs	100 Backward elimination after every 100th evaluation and at the end (if the number of evaluations is not a multiple of 100)

Cross validation is used during the GA procedure. In this case, the data are split into five deletion groups. The deletion groups are created by taking every fifth

TABLE 4 : Molecular structural descriptors of the peptides

Peptid	SEigZ	SEigv	VEA2	SIC0	IC1	CIC1	Ms	IC3	IC4	T (N..O)	SPAN	X2A	PHI	X1A	S3K	nCrHR	MAXDN	PW2	PW3	CIC2
Gly-Gly	1.036	-3.738	0.312	0.442	3.102	0.986	3.65	3.735	3.735	21	3.921	0.387	3.793	0.52	5.099	0	2.739	0.539	0.249	0.588
Gly-Leu	1.036	-3.738	0.248	0.327	3.073	1.785	3.04	4.047	4.047	21	4.619	0.359	5.465	0.497	5.469	0	2.685	0.561	0.281	0.811
Gly-Thr	1.286	-4.691	0.258	0.377	3.235	1.35	3.5	4.22	4.22	30	3.726	0.35	4.618	0.499	3.979	0	2.952	0.569	0.311	0.448
Gly-Ser	1.286	-4.691	0.273	0.406	3.327	1.065	3.65	4.107	4.107	30	3.707	0.347	4.566	0.511	3.798	0	2.95	0.549	0.314	0.381
Gly-Asn	1.429	-5.13	0.248	0.397	3.253	1.332	3.6	4.252	4.252	48	3.528	0.359	5.026	0.497	5.163	0	2.986	0.561	0.281	0.583
Gly-Phe	1.036	-3.738	0.227	0.334	3.121	1.786	2.86	4.482	4.707	21	5.233	0.325	4.797	0.477	4.247	0	2.731	0.557	0.304	0.812
Gly-Trp	1.179	-4.177	0.189	0.345	3.429	1.525	2.79	4.696	4.761	37	5.521	0.296	3.627	0.453	2.705	0	2.76	0.574	0.331	0.516
Ala-Gly-Gly	1.429	-5.13	0.243	0.373	3.171	1.584	3.38	4.282	4.357	53	4.957	0.367	5.831	0.495	7.041	0	2.808	0.568	0.271	0.769
Gly-Gly-Ileu	1.429	-5.13	0.204	0.319	3.113	2.057	3.07	4.628	4.684	53	5.842	0.336	7.5	0.495	6.238	0	2.752	0.56	0.316	0.931
Gly-Gly-Phe	1.429	-5.13	0.193	0.324	3.28	1.929	2.92	4.811	4.993	53	5.489	0.327	6.7	0.476	6.206	0	2.799	0.559	0.301	0.875
Gly-His-Gly	1.714	-6.008	0.2	0.356	3.37	1.718	3.05	4.852	4.852	111	6.347	0.327	6.147	0.475	5.609	0	2.836	0.559	0.294	0.739
Gly-Gly-Gly	1.429	-5.13	0.256	0.397	3.042	1.542	3.5	4.085	4.252	53	4.601	0.369	5.837	0.505	7.238	0	2.807	0.547	0.261	1.063
Gly-Gly-Val	1.429	-5.13	0.213	0.335	3.112	1.933	3.17	4.271	4.332	53	5.519	0.347	6.686	0.492	6.157	0	2.77	0.568	0.305	1.015
Ala-Ala	1.036	-3.738	0.278	0.373	3.012	1.511	3.32	3.762	4.023	21	4.445	0.36	3.891	0.495	3.829	0	2.737	0.588	0.308	0.935
Ala-Ala-Ala	1.429	-5.13	0.226	0.335	2.945	2.1	3.19	3.998	4.552	53	5.545	0.343	5.996	0.483	5.522	0	2.806	0.592	0.32	1.456
Ala-Leu	1.036	-3.738	0.24	0.31	2.912	2.088	2.95	4.179	4.179	21	4.924	0.358	5.538	0.488	5.565	0	2.687	0.58	0.289	1.142
Ala-Val	1.036	-3.738	0.25	0.327	2.909	1.949	3.06	4.021	4.021	21	4.736	0.351	4.771	0.488	4.127	0	2.705	0.589	0.316	1.191
Ala-Ser	1.286	-4.691	0.263	0.377	3.235	1.35	3.5	4.22	4.22	30	4.411	0.348	4.618	0.499	3.979	0	2.952	0.573	0.319	0.531
Ala-Asn	1.429	-5.13	0.259	0.421	3.209	1.25	3.47	4.278	4.278	42	4.242	0.321	2.849	0.462	2.402	2	2.919	0.593	0.327	0.364
Ala-Met	1.661	-3.657	0.236	0.356	3.114	1.792	2.95	4.39	4.39	21	6.097	0.35	6.747	0.499	5.848	0	2.687	0.557	0.298	0.584
Ala-Phe	1.036	-3.738	0.221	0.317	3.124	1.921	2.8	4.574	4.779	21	5.595	0.327	4.974	0.471	4.447	0	2.733	0.573	0.309	0.882
Gly-Gly-Gly-Gly	1.821	-6.522	0.221	0.369	2.983	1.972	3.42	4.002	4.309	108	4.354	0.361	7.876	0.497	9.436	0	2.847	0.552	0.267	1.517
Ala-Leu-Gly	1.179	-4.177	0.222	0.3	2.973	2.237	2.83	4.446	4.446	39	5.162	0.364	7.09	0.488	8.095	0	2.621	0.569	0.262	1.041
Leu-Leu	1.036	-3.738	0.221	0.283	2.792	2.456	2.77	4.451	4.451	21	4.6	0.348	7.139	0.492	6.444	0	2.675	0.562	0.294	1.317
Leu-Leu-Leu	1.429	-5.13	0.171	0.246	2.678	3.229	2.62	3.88	4.185	53	6.254	0.346	10.96	0.476	10.453	0	2.742	0.578	0.285	2.252
Leu-Val	0.893	-3.298	0.222	0.274	2.539	2.631	2.72	3.753	4.086	6	5.53	0.359	6.376	0.489	6.369	0	2.636	0.574	0.284	1.695
Leu-Phe	0.893	-3.298	0.2	0.269	2.795	2.527	2.56	4.415	4.634	6	6.122	0.336	6.405	0.473	6.277	0	2.665	0.563	0.284	1.335
Gly-Leu-Tyr	1.679	-6.083	0.177	0.284	3.239	2.404	2.83	5.014	5.134	83	5.998	0.325	8.553	0.469	7.647	0	2.852	0.573	0.303	1.03
Gly-Phe-Phe	1.429	-5.13	0.17	0.28	2.968	2.676	2.65	4.774	5.244	53	6.097	0.311	7.987	0.464	6.851	0	2.809	0.562	0.312	1.615
Leu-Gly-Phe	0.643	-2.345	0.219	0.25	2.468	2.89	2.19	4.503	4.668	0	5.355	0.333	6.624	0.479	6.374	0	1.994	0.549	0.285	1.504
Ser-Ser-Ser	2.179	-7.989	0.203	0.342	3.203	1.966	3.55	4.308	4.816	98	4.81	0.321	8.209	0.493	5.951	0	3.133	0.564	0.341	1.327
Gly-Pro-Ala	1.429	-5.13	0.206	0.332	3.158	1.93	3.01	4.653	4.653	53	4.558	0.308	4.902	0.469	3.411	1	2.77	0.579	0.345	0.787
Ala-Gly	0.893	-3.298	0.292	0.368	2.843	1.549	3.23	3.88	3.88	6	4.348	0.387	4.719	0.518	6.273	0	2.676	0.528	0.241	0.703
Gly-Val	1.036	-3.738	0.258	0.348	3.086	1.614	3.17	3.873	3.873	21	4.844	0.35	4.664	0.499	4.009	0	2.702	0.569	0.311	0.827
Gly-Ileu	1.036	-3.738	0.243	0.327	3.073	1.785	3.04	4.323	4.323	21	5.045	0.346	5.465	0.499	4.728	0	2.732	0.566	0.306	0.742

sample(e.g. 1,6,11,16,… as one group; 2,7,12,17,… as the next group). The data are sorted by the mobility of the component of interest before deletion groups are formed in order that the deletion groups uniformly sample from the data space. The cross validation is then five steps, each time leaving out one group, making a model on the other four, and then predicting the left out group. The number of components is selected according to the minimum value of root mean square error of cross validation(RMSECV). The validation data are not used at all until the GA procedure is complete and a final model is selected. The validation data are then predicted using this final model.

The last step of the GA is a stepwise approach in which the variables are entered according to the smoothed value of the frequencies of selection, each time computing the % cross-validated explained variance and the RMSECV. A crucial point in the previous algorithm is the detection of the number of variables to be taken into account(the selection of the solution corresponding to the global maximum often leads to overfitting); this decision is usually made by visually inspecting the plot of the % cross-validated explained variance(or of the RMSECV) versus the number of variables in the model, looking for the number of variables beyond which no "significant" increase of the response(decrease in RMSECV) takes place. Of course, this analysis required some time and the decision about the selected model sometimes could involve a high degree of subjectivity. To have a sounder statistical approach, the following algorithm has been implemented:

- Detect the global minimum of RMSECV;
- By using an F-test($P<0.1$, d.f.=number of samples in the training set 1, both in numerator and in the denominator) select a "threshold value" corresponding to the highest RMSECV being not significantly different from the global minimum;
- Look for the solution with the lowest number of variables having a RMSECV lower than the "threshold value".

In such a way, the most parsimonious model among all the models being not significantly different from the global optimum is selected. This modification generally leads to models having a slightly better root mean square error of prediction(RMSEP). Because each GA gives a slightly different model, some GA runs are performed on each data set, with the goal of verifying the robustness of the predictive ability and of the selected variables.

As a rule of thumb, it has been found that the performance of the algorithm decreases when >200 variables are used[34]. This is due to the fact that a higher variables/objects ratio increases the risk of overfitting and also due to the fact that the size of the search domain becomes too great.

It has been suggested that an adequate model should include descriptors as many as possible to increase the probability of a good characterization of compounds[17]. Therefore, a total 20 Dragon derived descriptors was selected by GA in the study. They are, eigenvalue sum from Z weighted distance matrix(Barysz matrix) (SEigZ), eigenvalue sum from Van der Waals weighted distance matrix(SEigv), average eigenvector coefficient sum from adjacency matrix(VEA2), structural information content(neighborhood symmetry of 0-order)(SIC0), information content index(neighborhood symmetry of 1-order)(IC1), complementary information content (CIC1), mean electrotopological state(Ms), information content index(neighborhood symmetry of 3-order) (IC3), information content index(neighborhood symmetry of 4-order)(IC4), sum of topological distances between N..O(T(N..O)), span R(SPAN), average connectivity index chi-2(X2A), kier flexibility index (PHI), average connectivity index chi-1(X1A), 3-path Kier alpha modified-shape index(S3K), number of ring tertiary C(sp^3)(nCrHR), maximal electrotopological negative variation(MAXDN), path/walk 2-Randic shape index(PW2), path/walk 3-Randic shape index(PW3), complementary information content(neighborhood symmetry of 2-order) (CIC2). The values of all the 20 descriptors are listed in TABLE 4.

The criterion used to determine the model dimentionality-the number of significant PLS component-is cross validation (CV). With CV, when the fraction of the total variation of the dependent variables that can be predicted by a component, Q^2, for the whole data set is larger than a significance limit (0.097), the tested PLS component is considered significant. When the cumulative Q^2 for the extracted components, Q^2_{cum}, is larger than 0.5, the model is considered to have a good prediction ability. Model adequacy was mainly measured as the number of PLS principal components (A), Q^2_{cum}, the correlation coefficient between observed values and fitted values (R). Besides the relative error of predicted values (E-Pred.) given by PLS analysis, root mean squares error of prediction (RMSEP) was adopted to compare the prediction precision of different models. RMSEP was defined as in multiple regression

analysis, i.e. $$\mathbf{RMSEP = [(Yi - \hat{Y}i)^2]^{0.5} \frac{1}{n}\sum_{i=1}^{n}} \qquad (3)$$

Where n stands for the number of compounds in the training set.

RESULTS AND DISCUSSION

PLS analysis for the 23 peptides in the training set, with μ_0 as a dependent variable and the 20 chemical descriptors as independent variables, resulted in PLS model(1), for which the results are listed in TABLE 5. In TABLE 5, $R^2_{X(adj)(cum)}$ and $R^2_{Y(adj)(cum)}$ stand for cumulative variance of all X’s and Y’s, respectively, explained by all extracted components. The optimum number of components(latent variables) to be included in the calibration model was also determined by computing the prediction error sum of squares(PRESS) for cross-validated models using a high number of factors (half the number of total standard + 1), which is defined as follows:

$$\mathbf{PRESS = \sum(Yi - \hat{Y}i)^2} \qquad (4)$$

Where y_i is the observed mobility for the *i*th peptide and $\hat{\mathbf{Y}}\mathbf{i}$ represents the estimated mobility A cross-validation method was employed to eliminate only one peptide at a time and then PLS algorithm modelS the remaining Y matrix and corresponding X matrix. By using the established calibration model the mobility of the peptide, left out was predicted. This process was repeated until each peptide had been out once.

One reasonable choice for the optimum number of factors would be that number which yielded the minimum PRESS. Since there are a finite number of peptides in the training set, in many cases the minimum PRESS value causes overfitting for unknown peptides that were not included in the model. A solution to this problem has been suggested by Haaland and Thomas[36] in which the PRESS values for all previous factors are compared to the PRESS value at the minimum. The F-statistical test can be used to determine the significance of PRESS values greater than the minimum.

TABLE 5 : Model fitting results

Models	A^a	$R_{X(adj)(cum)}$	$R_{Y(adj)(cum)}$	Q^2_{cum}	R	RMSEP
1	1	0.602	0.934	0.884	0.965	0.708
2	1	0.585	0.935	0.872	0.968	0.701
3	3	0.823	0.943	0.867	0.971	0.612
4	5	0.940	0.952	0.893	0.973	0.601
5	5	0.941	0.958	0.886	0.976	0.592
6	5	0.946	0.959	0.902	0.977	0.568
7	5	0.946	0.959	0.896	0.980	0.494
8	2	0.688	0.957	0.869	0.978	0.581
9	2	0.724	0.963	0.869	0.981	0.495
10	5	0.980	0.968	0.871	0.984	0.492
11	2	0.718	0.935	0.843	0.967	0.718
12	5	0.988	0.954	0.892	0.977	0.603
13	3	0.916	0.914	0.836	0.956	0.827
14	3	0.912	0.916	0.829	0.957	0.814
15	3	0.938	0.906	0.846	0.952	0.866
16	2	0.866	0.824	0.755	0.908	1.181
17	4	1.000	0.856	0.735	0.902	1.070
18	2	0.893	0.659	0.534	0.812	1.646
19	1	0.678	0.618	0.513	0.786	1.743

[a]Number of component

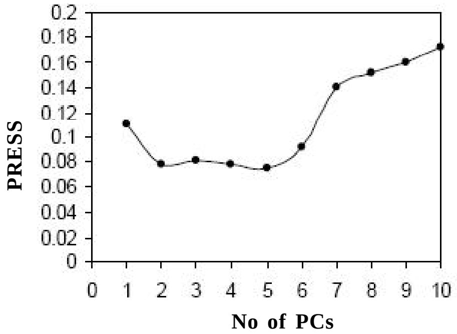

Figure 1 : PRESS vs. Number of principal components

The maximum number of factors used to calculate the optimum PRESS was selected as 10 and the optimum number of factors obtained by the application of PLS models are summarized in TABLE 5. In all instances, the number of factors for the first PRESS values whose F-ratio probability drops below 0.75 was selected as the optimum. The figure 1 shows the PRESS obtained by optimizing the calibration matrix of the descriptors with PLS. It can be concluded from TABLE 5 that five PLS components were selected in model (10), and the five PLS components explained 98.0% of the variance of the independent variables, and 96.8% of the variance of the dependent variable.

Variable importance in the projection(VIP) is a parameter that shows the importance of a variable in a model. Terms with a large value of VIP, larger than 0.8, are the most relevant for explaining the dependent variable. In a model, the smaller VIP value of the descriptor, the less significant the descriptor is in explaining the μ_0. Although the PLS method offers the advantage of handling data sets where the number of independent variables is great, it can be seen that considerably worse predictions are obtained if many irrelevant descriptors are included in the PLS model[35]. So it is necessary to perform a PLS analysis that excludes the least significant descriptor. Such a PLS analysis resulted in model (2) by excluding the least significant descriptor from model (1). Following the same method, by removing the least significant, descriptor from the former model step by step, until only 2 descriptors left in the model, models (2)-(19) were obtained successively, as shown in TABLE 5.

Since Q^2_{cum} value was determined by the CV method, the greater the Q^2_{cum} value, the more robust or stable the model. The RMSEP in this study was a measure of prediction precision, the lower the RMSEP, the better the prediction precision. Comparing with these

TABLE 6 : The PLS weights, VIPs and pseudo-regression coefficients[a]

Variables	W[a][1]	W[a][2]	W[a][3]	W[a][4]	W[a][5]	VIP	Coefficients(α)	Coefficients(b)
SEigZ	-0.109	0.405	-0.137	-0.193	-0.162	1.671	-0.052	-0.474
SEigv	0.110	-0.410	0.121	0.153	0.038	1.416	0.018	0.046
VEA2	0.504	0.241	0.363	0.313	0.549	1.315	0.465	39.836
IC1	0.208	0.314	-0.105	0.234	-0.082	1.222	0.145	1.904
IC3	-0.276	-0.070	-0.051	0.466	-0.142	0.955	-0.144	-1.234
IC4	-0.368	-0.012	0.098	0.749	0.731	0.916	0.062	0.490
PHI	-0.427	0.108	0.109	-0.129	-0.022	0.691	-0.198	-0.346
S3K	-0.288	0.299	0.426	-0.029	-0.252	0.547	-0.125	-0.204
MAXDN	0.165	0.461	-0.049	0.073	0.409	0.437	0.255	4.185
PW3	-0.132	-0.455	-0.910	-0.591	-0.159	0.366	-0.287	-33.718
CIC2	-0.396	-0.138	0.070	-0.135	0.498	0.362	-0.103	-0.697
Constants							31.104	15.250

[a]Coefficients (α)-coefficients scaled and centered; coefficients (b)-coefficients unscaled.

TABLE 7 : Correlation coefficient between some descriptors (p<0.05)

Variables	SEigZ	SEigv	VEA2	IC1	IC3	IC4	PHI	S3K	MAXDN	PW3	CIC2
SEigZ	1.000										
SEigv	-0.936	1.000									
VEA2	-0.358	0.365	1.000								
IC1	0.516	-0.497	0.062	1.000							
IC3	0.288	-0.259	-0.670	0.384	1.000						
IC4	0.387	-0.389	-0.812	0.216	0.895	1.000					
PHI	0.415	-0.392	-0.704	-0.360	0.227	0.415	1.000				
S3K	0.256	-0.250	-0.418	-0.461	-0.018	0.142	0.866	1.000			
MAXDN	0.679	-0.724	0.048	0.692	0.003	0.038	-0.067	-0.166	1.000		
PW3	0.286	-0.295	-0.331	0.387	0.358	0.398	-0.125	-0.567	0.324	1.000	
CIC2	0.007	-0.070	-0.543	-0.746	-0.132	0.182	0.757	0.707	-0.319	-0.140	1.000

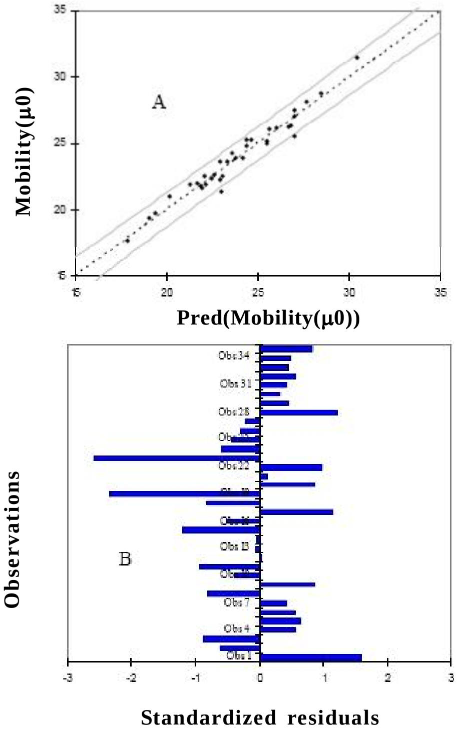

Figure 2 : Plot of (A) observed and predicted mobility (μ_0) and (B) standardized residuals and predicted μ_0

statistical indices of models (1)-(19), it can be found that the Q^2_{cum} of model (10) was the largest, and the RMSEP of model (10) was the smallest, indicating that model (10) was the most robust and best prediction precision QSPR model among all the 19 PLS models. In model (10), 9 descriptors, SIC0, CIC1, Ms, T(N..O), SPAN, X2A, X1A, nCrHR and PW2 have been kept out, so it can be concluded that these 9 descriptors are of less importance to the μ_0 of peptides. If these descriptors are included in PLS models, they can increase the “background noise” of the models, resulting in less robust and poor significance of PLS models as indicated by models (1)-(9). Nevertheless, the other 11 descriptors are necessary to modeling μ_0 of peptides. If they are not considered in PLS models, the molecular structure character relevant to μ_0 cannot be well described, leading to PLS models with bad prediction precision-such as models (11)-(19).

For the 35 peptides contained in the training and validation set, the correlation between observed and predicted μ_0 values is very significant figure 2, as indicated by R values. As the cross-validated Q^2_{cum} value of model (10) is remarkable larger than 0.50, model (10) is surely stable and has good prediction ability. Based on model (10), μ_0 for the other 12 peptides were calculated(TABLE 1). As shown by TABLE 1 and fig-

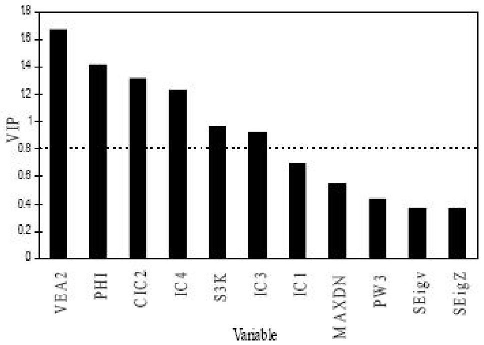

Figure 3 : The plot of VIP for different variables

ure 2, the predicted values were consistent with the corresponding μ_0 values determined by Wronski et al.[7]. So it has been validated that model (10) can be used to predict μ_0 values of the other peptides.

From the PLS weights($W^a[1]$, $W^a[2]$, $W^a[3]$, $W^a[4]$ and $W^a[5]$) listed in TABLE 6, it can be seen how much a single variable contributes in each PLS component to the modeling of the μ_0. The first PLS component is mainly related to the descriptors VEA2 and PHI. The absolute values of $W^a[1]$ for these descriptors are larger than 0.400 and larger than the absolute values of $W^*[1]$ for the other descriptors and the second PLS component is mainly related to the descriptors SEigZ, SEigv, MAXDN and PW3. The absolute values of $W^a[2]$ for these descriptors are larger than 0.400 and larger than the absolute values of $W^a[2]$ for the other descriptors also we can see the same observations for other PLS components and related descriptors. As shown in TABLE 7, all these descriptors are inter-correlated.

The VIP values for the independent variables in model (10) are listed in TABLE 6. Moreover the figure 3 displays the VIP values for each explanatory variable, on first PLS component. Similarly, it can be seen the pseudo-regression coefficients of the independent variables and constants transformed from PLS results from TABLE 6. From the positive and negative symbols of the coefficients of the independent variables, one can evaluate the effects of each independent variable on μ_0. Based on the unscaled coefficients and constants, QSPR equations like those obtained from multiple regression analysis can be obtained, as follows:

Mobility(m0)=15.2496-0.4737*SEigZ+4.5702 E02*SEigv +39.8356*VEA2+1.9041*IC1-1.2335* IC3+0.4904*IC4-0.3459*PHI-0.2044* S3K+ 4.1848*MAXDN-33.7180*PW3-0.6970 *CIC2.

CONCLUSION

In this study, based on molecular structural descriptors, by the use of PLS analysis, a significant QSPR was obtained for the mobility of peptides. The QSPR can be used for prediction. 11 descriptors including SEigZ, SEigv, VEA2, IC1, IC3, IC4, PHI, S3K, MAXDN, PW3 and CIC2 were used in selected PLS model. The results showed the ability of the obtained model in the determination of mobility of peptides.

REFERENCES

[1] D.A.Kevin; J.Chromatogr., **A856**, 443 **(1999)**.

[2] R.L.Kay; Pure.Appl.Chem., **63**, 1393 **(1991)**.

[3] P.D.Grossman, J.C.Colburn (Eds.); 'Capillary Electrophoresis: Theory and Practice', Academic Press, San Diego, 112 **(1992)**.

[4] R.Weinberger; 'Practical Capillary Electrophoresis', Academic Press, London, 48 **(1993)**.

[5] R.E.Offord; Nature, **211**, 591 **(1966)**.

[6] M.Wronski; J.Chromatogr., **288**, 206 **(1984)**.

[7] M.Wronski; J.Chromatogr., **A657**, 165 **(1993)**.

[8] B.J.Compton; J.Chromatogr., **559**, 357 **(1991)**.

[9] D.M.Li, L.A.Lucy; Anal.Chem., **73**, 1324 **(2001)**.

[10] S.L.Fu, D.M.Li, C.A.Lucy; Analyst, **123**, 1487 **(1998)**.

[11] C.X.Xue, H.X.Liu, X.J.Yao, M.C.Liu, Z.D.Hu, B.T. Fan; J.Chromatogr., **A1048**, 233 **(2004)**.

[12] M.Jalali-Heravi, Z.Garkani-Nejad; J.Chromatogr., **A927**, 211 **(2001)**.

[13] M.Jalali-Heravi, Z.Garkani-Nejad; J.Chromatogr., **A971**, **(2002)** 207.

[14] H.R.Liang, H.Vuorela, P.Vuorela, M.L.Riekkola, R. Hiltunen; J.Chromatogr., **A798**, 233 **(1998)**.

[15] A.G.McKillop, R.M.Smith, R.C.Rowe, S.A.C.Wren; Anal.Chem., **71**, 497 **(1999)**.

[16] A.Jouyban, B.H.Yousefi; Comput.Biol.Chem., **27**, 297 **(2003)**.

[17] R.Kaliszan; J.Chromatogr., **A656**, 417 **(1993)**.

[18] S.Wold, H.Wold, W.J.Dunn; Report UMINF-83, Department of Chemistry, University of Umea, Sweden **(1984)**.

[19] Hypercube, http://www.hyper.com

[20] R.Todeschini; Milano Chemometrics, QSAR Group, http://www.disat.unimib.it/vhm/

[21] R.Todeschini, V.Consonni; 'Handbook of Molecular Descriptors', Wiley-VCH, Weinheim, Germany, **(2000)**.

[22] R.Todeschini, V.Consonni; 'Handbook of Molecular Descriptors', Wiley-VCH, Weinheim, Germany, **(2000)**.

[23] E.V.Kostantinora; J.Chem.Inf.Comp.Sci., **36**, 54**(1997)**.

[24] G.Rucker, C.Rucker; J.Chem.Inf.Comp.Sci., **33**, 683 **(1993)**.

[25] J.Galvez, R.Garcia, M.T.Salabert, R.Soler; J.Chem. Inf.Comp.Sci., **34**, 520 **(1994)**.

[26] P.Broto, G.Moreau, M.Fortin, C.Turpin; Eur.J.Med. Chem., **23**, 275 **(1988)**.

[27] R.Leardi, R.Boggia, M.Terrile; J.Chemomet., **6**, 267 **(1992)**.

[28] R.Leardi; J.Chemomet., **14**, 643 **(2000)**.

[29] D.E.Goldberg; Addison Wesley, New York, **(1989)**.

[30] B.Kowalski, R.Gerlach; In K.G.Joreskog, H.Wold (Eds.); 'Systems Under Indirect Observation', North Holland, Amsterdam, 191-209 **(1982)**.

[31] R.Q.Yu; 'Introduction to Chemometrics', Human Education Publishing House, Changsha, **(1992)**.

[32] B.S.Dayal, J.F.MacGregor; J.Chemomet., **11**, 73 **(1997)**.

[33] P.Geladi, B.R.Kowalski; Anal.Chim.Acta, **185**, 1 **(1986)**.

[34] R.Leardi, A.L.Gonzalez; Chemom.Intell.Lab.Syst., **41**, 195 **(1998)**.

[35] J.M.Lucco; J.Chem.Inf.Comput.Sci., **39**, 396 **(1999)**.

[36] D.M.Haaland, E.V.Thomas; Anal.Chem., **62**, 1091 **(1990)**.