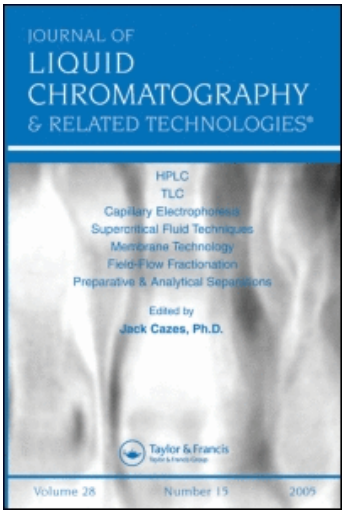


**Journal of Liquid Chromatography & Related Technologies**
Publication details, including instructions for authors and subscription information:
http://www.informaworld.com/smpp/title~content=t713597273

# Review of QSPR Modeling of Mobilities of Peptides in Capillary Zone Electrophoresis

K. P. Liu[a]; B. B. Xia[a]; X. Y. Zhang[a]
[a] Department of Chemistry, Lanzhou University, Lanzhou, Gansu, P. R. China

Online publication date: 22 June 2010

**To cite this Article** Liu, K. P. , Xia, B. B. and Zhang, X. Y.(2008) 'Review of QSPR Modeling of Mobilities of Peptides in Capillary Zone Electrophoresis', Journal of Liquid Chromatography & Related Technologies, 31: 11, 1808 — 1822
**To link to this Article: DOI:** 10.1080/10826070802129001
**URL:** http://dx.doi.org/10.1080/10826070802129001

*Journal of Liquid Chromatography & Related Technologies*®, 31: 1808–1822, 2008

ISSN: 1082-6076 print/1520-572X online
DOI: 10.1080/10826070802129001

# Review of QSPR Modeling of Mobilities of Peptides in Capillary Zone Electrophoresis

**K. P. Liu, B. B. Xia, and X. Y. Zhang**

Department of Chemistry, Lanzhou University, Lanzhou, Gansu, P. R. China



**Abstract:** Quantitative-structure property relationships, as related to peptide electrophoretic mobility, are presented in this review. The methods of discussion ranged from linear to non-linear method. It is the intent that the review will provide the present state of knowledge and current trends in this area for a new investigator in this field.

**Keywords:** CZE, Mobility, Peptide, QSPR, Review

## INTRODUCTION

Peptides, which are composed of amino acids linked through the peptide bonds between each other, belong to the most important biologically active substances in the living organisms. We can find many kinds of peptides in the natural world. They play a significant role in control and regulation of many vitally important living processes, acting as hormones, neurotransmitters, immunomodulators, coenzymes, enzyme substrates, and inhibitors, receptor, ligands, drugs, toxins, and antibiotics. In the era of proteomics, the comprehensive analysis of proteome currently represents the main road for a new drug discovery, since both the structures and functions of many proteins are identified via their peptide fragments.[1] According to the pool of these fragments, a peptide map which

Correspondence: X. Y. Zhang, Department of Chemistry, Lanzhou University, Lanzhou 730000, Gansu, P. R. China. E-mail: xyzhang@lzu.edu.cn

serves as fingerprint for protein identification can be accomplished. From the peptide map, one can obtain the whole peptide set of a cell or a peptidome and then understand the living cell function.[2] So, this peptidic approach is becoming one of the main directions in proteome research. Now, as we know, because of their nutritional and biological properties, studies of peptides have become a hot spot in pharmaceutical and cosmetic industries all over the world.[3] Thus, it can be seen that separation and analysis of peptides become more and more important and requires a powerful analytical technique to separate the peptides and to identify them.

Among the numerous separation techniques, capillary zone electrophoresis (CZE) is the most widely used method for peptide separation because of its simplicity, versatility, high-resolution power, high sensitivity, and a low analysis time.[4] As long as a molecule is charged, it can be separated by CZE. This makes the applications for CZE very diverse, being used for peptide, ion, enantiomer, pharmaceuticals, proteins, polymers, amines and food constituents' analysis.[5] Due to these peculiar advantages of CZE, CZE is also efficient to obtain some information about the identity, the purity, and some structural changes of peptides. The basic mechanism in electrophoresis is the differences in the analytes' mobilities; so, the electrophoretic mobility, which can be converted to migration time, is the most important parameter governing the separation of solutes in capillary electrophoresis. However, during the method development in CZE to develop an optimized separation, the analysts generally have to employ a large number of experiments, which is often costly and time-consuming, to analyze and identify the peptides from real samples. So, there is a necessity to develop a computational method for calculate the electrophoresis mobility in a certain practical conditions to shorten the long time normally needed for CZE peptide identification and, at the same time, to facilitate the improvement of the quality for CZE peptides separations.

Alternatively, quantitative structure-property relationships (QSPR) provided a promising method for the estimation of compounds' electrophoretic behavior based on the descriptors derived solely from the molecular structure to fit experimental data. The advantage of this approach over other methods lies in the fact that once a reliable model was built, it required only the knowledge of chemical structure and was not dependent on the experiment data. The QSPR approach had become a very useful tool in the prediction of many physicochemical properties. This approach was based on the assumption that the variation of the behavior of compounds, as expressed by any measured physicochemical properties, could be correlated with changes in molecular features of the compounds, termed descriptors.[6] Noteworthily, the success of a QSPR will depend on the quality of the data set and on the suitability of the

descriptor(s) selected. This method could predict the properties of new compounds that had not been synthesized or found. It can also identify important structural features of the molecules that are relevant to variations in molecular properties, and thus gain some insight into the structural factors affecting the molecular properties. Furthermore, the application of QSPR, maybe, can reduce the number of chemicals released into the environment and greatly lessen the impact of these hazardous chemicals on the ecosystem. However, the main problems encountered in this kind of research are the description of the molecular structure using appropriate molecular descriptors and selection of suitable modeling methods. QSPR studies, as germane to peptides' electrophoretic mobility, will be discussed in this review. This review is written with the purpose of providing the present state of knowledge and current trends for peptides' electrophoretic mobilities in QSPR methodologies.



## MAJOR STAGES IN QSPR/QSAR MODEL OF MOBILITIES OF PEPTIDE IN CZE

The main steps in QSPR/QSAR modeling process include: data collection, structure input, structure optimization, molecular descriptors calculation, descriptors selection, modeling, and model validation.[7,8] This process can be illustrated as in Fig. 1.[9–11]

In QSPR/QSAR studies, the data's reliability is the key to obtaining a precise model. In most cases, the number of the compounds should be more than ten. To have the computer identify the compound, the structure of a compound must be converted to a special code which is a type of data that the computer can deal with. At present, there is much software to draw the compound structures, such as Chemdraw,[12] and ISIS DRAW,[13] and others. Then, the structure optimization is enforced using a molecular modeling software package such as HyperChem,[14] Gaussian,[15] Sybyl,[16] and MOPAC.[17] They employ the optimization methods consistent with molecular mechanics force field method,

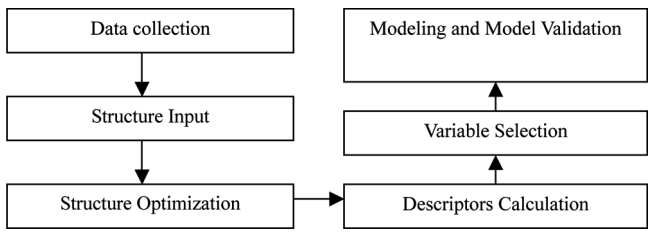


***Figure 1.*** Main steps in QSPR studies.

semi-empirical quantum-chemical method, and ab initio method, and others. The aim of the optimization is to search the optimal conformation of the compound whose energy is lowest in all conformations. To obtain a QSPR/QSAR model, compounds are often represented by molecular descriptors. The optimal structures of compounds are then exported to CODESSA[18] or DRAGON[19] to calculate the molecular descriptors.

Once molecular descriptors are generated, the variable selection method must be implemented to reduce the pool of the descriptors because, after the calculation of descriptors, there will be several hundreds of descriptors and so many descriptors can not directly be used to construct the model. Furthermore, according to the QSPR/QSAR theory, the number of the compounds must be five or more greater than the number of the descriptors. At present, the most important used variable selection methods include: stepwise regression,[20] genetic algorithm (GA),[21] and the heuristic method (HM).[22] Among these methods, the HM is widely used, owing to its high speed and no software restrictions on the size of the data set. Then, using the selected descriptors and the properties or activities of the compounds, one can build the mathematic linear or non-linear models, finally validating the model by various methods. Generally, the following methods of validation are applied to the models:

1. Leave-One-Out (LOO) – Standard Leave-One-Out cross-validation is performed on the data.
2. Leave-Many-Out (LMO) – Leave-Many-Out validation is performed on the data set by randomly splitting into a number of disjoint subsets. For each subset, the standard Leave-Many-Out cross-validation procedure is performed.
3. Test Set Validation – An external test set is used for validation of the models.



## METHODS

Various linear and nonlinear chemometrics or chemoinformatics methods can be used to model the relationships between the structural factors and properties/activities.

### Linear Methods

Currently, there are many linear methods used in QSPR studies, such as multiple linear regression (MLR),[23] linear discriminant analysis (LDA),

principle component regression (PCR), and partial least squares (PLS).[24] Herein, only a few methods are discussed.

## Multiple Linear Regression (MLR)

Multiple Linear Regression (MLR) is a commonly used statistical method in traditional 2D-QSPR. In MLR analysis, the descriptors in the regression equation must be independent variables. To reduce the number of the descriptors and minimize the information overlap in the descriptors, the concept of non-redundant descriptors (NRD)[25] is used. The linear correlation coefficients value between two descriptors should be less than a pre-determined threshold (e.g., 0.8 or 0.9). Once descriptors are generated, a forward stepwise regression method is used to develop the linear model of the property of interest, which is shown as follows:

$$Y = b_0 + b_1X_1 + b_2X_2 + \cdots + b_nX_n \quad (1)$$

where, Y is the property, that is, the dependent variable, $X_1,\ldots\ldots,X_n$ represent the specific descriptors, while $b_1,\ldots\ldots,b_n$ represent the coefficients of those descriptors, and $b_0$ the intercept of this equation.

## Linear Discriminant Analysis (LDA)

Linear Discriminant Analysis (LDA) is used in statistics to find the linear combination of features which best separate two or more classes of objects or events. LDA classifies the dependent by dividing an n-dimensional descriptor space into two regions that are separated by a hyperplane which is defined by a linear discriminant function;[26] for more than two groups, a set of discriminant functions are generated. The regions formed by the hyperplane correspond to the classes to which individual compounds are predicted to belong.

## Principle Component Regression (PCR)

Principle Component Regression (PCR) is a useful dimensional reduction method for original data sets, and containing principle component analysis and regression. Firstly, a statistical technique of changing the many variables in a data matrix so that the new components are correlated with the original components but not with each other, that is, so that they are now independent of each other. It is a technique used to change a set of original variables into a number of basic dimensions. The principle component analysis (PCA) is used to extract the abstract factors, and then construct the mathematic model through a normal regression method. For the concept of principle component, it can be considered

that it is a new variable which is the linear combination of the original variable $x_{ij}$. The main step of the PCR consists of: 1) the normalization of the data set; 2) the computation of the eigenvalue vector from covariance matrix; 3) selection of principle component and multiple linear regression analysis.

However, PCR only considers the independent variable but not the dependent variable, which may include more useful information. Fortunately, PLS can overcome this drawback of the PCR approach. PLS considers not only the independent variable, but also the dependent variable and simultaneously describes the independent variable and the dependent variable better through compromising the factors in each feature space. The main advantages of PLS are: 1) no rigorous limit for the correlation between variables; 2) meaningful result can be obtained when the number of variables is larger than the number of samples; 3) the information of the independent and dependent variables are simultaneously considered and can obtain a more meaningful result; 4) chance correlation can be reduced owing to use the cross-validation to select the optimal number of the principle components in the model.



## Non-Linear Methods

At present, there are many non-linear methods; here, we will only introduce the two primary methods which are used most frequently. One is artificial neural network (ANN)[27] in which a back propagation neural network (BPNN)[28] and a radial basis function neural network (RBFNN)[29] are the most useful methods. Another is the support vector machine (SVM).[30]

### Back Propagation Neural Network (BPNN)

The Back Propagation Neural Network (BPNN) is represented schematically in Fig. 2. The BPNN model is composed of a large number of simple processing elements (PE) or neuron nodes, organized into a sequence of layers. The first layer is the input layer with one node for each variable or feature of the data. The last layer is the output layer consisting of one node for each variable to be investigated. In between these two layers are a series of one or more hidden layer(s) consisting of a number of nodes, which are responsible for learning. Nodes in any layer are fully or randomly connected to nodes of a succeeding layer. Each connection is represented by a number called a weight (w). BPNN are most often used to analyze non-linear multivariable data. In these networks, signals are propagated from the input layer through the hidden layer(s) to the output layer. A node thus receives signals via connections from other nodes

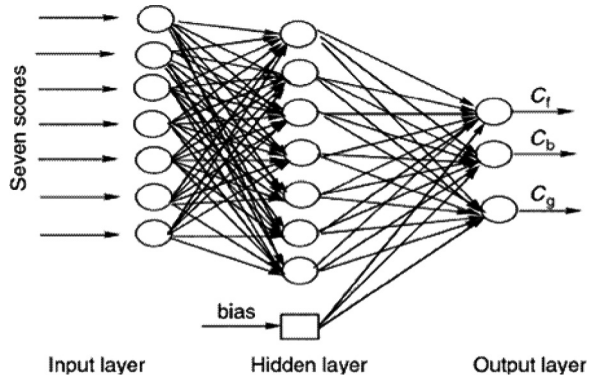


***Figure 2.*** The typical architecture of the ANN.



(or the outside world in the case of the input layer). The net input for a node j is given by:

$$net_j = \sum_i w_{ji} o_i \tag{2}$$

where, i represents the nodes in the previous layer, $w_{ji}$ is the weight associated with the connection from node i to node j, and $o_i$ is the output of node i. The output of a node is determined by the transfer function and net input of the node. A popular transfer function is the sigmoid:

$$o_j = f(net_j) = \frac{1}{1 + \exp[-(net_j + \theta_j]} \tag{3}$$

where, $\theta_j$ is a bias term or threshold value of node j responsible for accommodating non-zero offsets in the data. The adequate functioning of neural networks depends strongly on the way the signals are propagated through the networks. The weights play an important role in this propagation and a proper setting of these weight factors is essential. Generally, such a setting is not known beforehand and the weights are initially given small, random values. The process of adapting the weights to an optimum set of values is called training and is usually done by means of supervised learning. A representative training set with examples is presented iteratively to the neural network and the difference between the desired solution and the one obtained is used to adapt the weights in small steps, according to a learning algorithm. There are a number of learning algorithms used to train a neural network. A frequently used one is the back propagation (BP) learning rule.[31]

Radial Basis Function Neural Network (RBFNN)

The typical RBFNN architecture is similar to Fig. 2. It also consists of three layers. The first layer is made up of input nodes that transmit unweighted inputs to each node in the hidden layer. Each hidden node contains a radial basis function as the transfer function. The outputs of these nodes are weighted and summed to produce the final output. In contrast to the sigmoid function, the radial basis function is classified as a local activation function. The most often used is the Guassian function:

$$Z_{i,j}(x_j, \alpha_i, \beta_i) = \exp(-\|\alpha_i - x_j\|^2/\beta_i^2) \quad (4)$$

where, $x_j = \{x_1, x_2, \ldots\ldots, x_M\}_j$ is the jth input vector of dimension M presented to the net, $Z_{i,\,j}(x_j, \alpha_i, \beta_i)$ is the activation of the ith node in the hidden layer in response to the jth input vector $x_j$. $M + 1$ parameters are associated with each node, viz. $\alpha_j = \{\alpha_1, \alpha_2, \ldots\ldots, \alpha_M\}_i$, as well as $\beta_j$, a distance scaling parameter which determines the distance in the input space over which the node will have a significant influence. The parameters $\alpha_i$ and $\beta_j$ function in much the same way as the mean and standard deviation in a normal distribution. The closer is the input vector to the pattern of a hidden unit (i.e., the smaller the distance between these vectors), the stronger is the activity of the unit. The hidden layer can thus be considered to be a density function for the input space and can be used to derive a measure of the probability that a new input vector is part of the same distribution as the training vectors. After selection of the centers and radius, the connections between the radial basis units and the output node are weighted. The output of the net is, consequently, given by:

$$Fr_k = \sum W_{ik} Z_i + b_i \quad (5)$$

where, $b_i$ is the bias; i represents the ith node in the hidden layer; $w_{ik}$ is the weight associated with the connection from node i to node k; $Z_i$ is the output of the ith hidden layer node.

Support Vector Machine (SVM)

The Support Vector Machine (SVM), developed by Vapnik[32] as a novel type of machine learning method, is gaining popularity due to many attractive features and promising empirical performance. Compared with traditional neural networks, the SVM possesses prominent advantages: 1) strong theoretical background provides SVM with high generalization capability and can avoid local minima; 2) SVM always has a solution, which can be quickly obtained by a standard algorithm (quadratic programming); 3) SVM need not determine network topology in advance,

which can be automatically obtained when the training process ends; 4) SVM builds a result based on a sparse subset of training samples, which reduces the workload. Originally, SVMs are developed for pattern recognition problems, and now, with the introduction of an insensitive loss function, SVMs have been extended to solve nonlinear regression estimation and time-series prediction with excellent performance.[33] The basic principle of support vector regression is described below.

A support vector machine is first trained on a sample with objects having known target values. After training, the machine is used to predict or estimate target values for objects where these values are unknown. A kernel-induced feature space with function K(x, $x_i$) is used for the mapping of objects onto target values. Thus, a non-linear feature mapping will allow the treatment of non-linear problems in a linear space. The prediction or approximation function used by a basic SVM is

$$f(x) = \sum_{i=1}^{l} \alpha_i K(x, x_i) + b \tag{6}$$

where, $\alpha_i$ is some real value, $x_i$ is a feature vector corresponding to a training object. The components of vector $\alpha$ and the constant b represent the hypothesis and are optimized during training. K (x, $x_i$) is a kernel function, which value is equal to the inner product of two vectors x and $x_i$ in the feature space $\Phi$ (x) and $\Phi$ ($x_i$). That is, K (x, $x_i$) = $\Phi$ (x)·$\Phi$ ($x_i$). The elegance of using kernel function lies in the fact that one can deal with feature spaces of arbitrary dimensionality without having to compute the map $\Phi$ (x) explicitly and it may be useful to think of the kernel, K (x, $x_i$) as comparing patterns, or as evaluating the proximity of objects in their feature space. Thus, a test point is evaluated by comparing it to all training points. Training points with nonzero weight $\alpha_i$ are called the support vectors.

For a given dataset, only the kernel function and the regularity parameter C must be selected to specify one SVM. Any function that satisfies Mercer's condition can be used as the kernel function. In support of vector regression, the Gaussian kernel $K(u, v) = \exp(-|u - v|^2/\delta^2)$ is most commonly used.

## RECENT APPLICATIONS

The linear methods are commonly used methods in the QSPR study for peptide electrophoretic mobility. Most of these models are based on the Offord model, which has shown that the electrophoretic mobility is proportional to the charge Q and inversely proportional to the molecular mass M.[34–36] Cifuentes et al.[37] considered ten peptides as classical linear

polymers with n amino acid residues and arrived at an equation that correlated mobility with a function in the form of ln $[(0.297q + 1)/M^{0.411}]$, and obtained an R of 0.9993. Rickard et al.[38] studied 33 diverse peptides from enzymatic digests which have a function with $q/M^{2/3}$ and obtained an R of 0.948. On the other hand, Janini et al.[39] have obtained the electrophoretic mobility of 58 peptides ranging in size from 2 to 39 amino acids and charge from 0.65 to 7.82. They also investigated the correlation between mobility and $q/M^{2/3}$ which gave an R of 0.96; the obtained regression equation can be shown as:

$$\mu_{ef} = 2.44 + 581.85 \times q/M^{2/3} \quad (7)$$

Then, they concluded that, although the Offord model gave the best overall mobility, it fails when applied to hydrophobic and highly charged peptides. In 2004, Veronika et al.[40]studied the correlation between mobility and $q/M^{2/3}$ of 20 synthetic insect oostatic peptides (IOPs) and their derivatives and fragments. They established several models which gave a range of R from 0.888 to 0.936. The result indicated that the peptides which have three or more amino acid residues gave an unsatisfactory result. And then, in 2007, they studied 12 synthetic gonadotropin-releasing hormones (GnRHs) and their analogs and fragments again.[41] They also used the $q/M^{2/3}$ as the descriptor to construct the model which gave and a range of R from 0.995 to 0.999 and obtained a satisfactory experimental result. From the above literature, we can see that the main deficiency of the Offord model is that it takes into account only two physicochemical properties of peptides, the charge (Q) and the relative molecular mass (M) and the data set cannot be larger, which may result in a bad regression result.

To improve the predicted accuracy and to deal with a large data set, many other regression methods are used, in a stepwise manner, such as MLR, ANN, and SVM. The application of these methods impact the QSPR studies of peptides' electrophoretic mobility to a great extent. Jalali-Heravi et al.[42] considered 125 peptides ranging in size between 2 and 14 amino acids. Their aim was to explore the usefulness of empirical models and multivariate analysis techniques in predicting electrophoretic mobilities of small peptides in capillary zone electrophoresis (CZE). They used the charge-to-size ratio (QM), using the corrected steric substituent constant (Es, c) and molecular refractivity (MR) as the descriptors to construct the MLR and BPANN models. Two models gave squared correlation coefficients ($R^2$) of 0.895 and 0.930, respectively. Such a BPANN model can be designed as 3-4-1 net to indicate the number of the units in the input, the hidden, and the output layer, respectively. By comparing two models, it can be found that the BPANN model is better than the MLR model. In this work, the 125 peptides are all small peptides, so they

then studied 102 large peptides in which the largest peptide had 42 amino acids to validate the stability of the BPANN.[43] They also used the same three descriptors as the input and constructed a MLR and a BPANN which had a 3-3-1 net structure. The obtained squared correlation coefficient ($R^2$) was 0.930 and 0.970, respectively. The result could reflect the relationship between structure of peptides and electrophoretic mobilities more accuracy. Then, they used the obtained BPANN model to predict the other 24 high-charged and hydrophobic peptides and obtained higher accuracy than the former literature. The squared correlation coefficient ($R^2$) predicted was up to 0.990. This result indicated that the BPANN model was more stable and accurate than the MLR model. By comparing two BPANN in these two works, it can be found that the latter BPANN model can include more types of peptides, has larger applicability, and has more powerful predictive ability.

Ma et al.[44] studied 183 peptides. Their aim was to predict electrophoretic mobilities of peptides in capillary zone electrophoresis using the HM and a new nonlinear method of RBFNN. The whole data set was divided into two subsets: data set 1, which consisted of 125 peptides ranging in size between 2 and 14 amino acids, and data set 2, which consisted of 58 peptides ranging in size between 2 and 39 amino acids. The HM method was a selection method of variables and was usually used to obtain preliminary screening of the library of descriptors in order to select a subset of descriptors that may be of interest and importance for the study under consideration. Through the HM, they selected four descriptors which consisted of $q/M^{2/3}$, the Wiener index (W), the relative number of O atoms (RNO), and the relative number of N atoms (RNN) for data set 1 and two descriptors which consisted of $q/M^{2/3}$ and the Wiener index (W) for data set 2. For two data sets, they used both MLR an RBFNN to construct the linear and non-linear model by dividing the data set into training set and test set. The MLR equation for two data sets shown as follows:

$$\mu_{ef}(\text{data set } 1) = 10.1 + 984 \times q/M^{2/3} - 50600 \times W - 45.8 \times RNO + 52.7 \times RNN \quad (8)$$

$$\mu_{ef}(\text{data set } 2) = 2.32 + 592 \times q/M^{2/3} - 70900000 \times W \quad (9)$$

The result indicated that two RBFNN models which gave a squared correlation coefficients ($R^2$) 0.9740 and 0.9773, respectively, were all better than two MLR models which gave squared correlation coefficients ($R^2$) 0.9414 and 0.9671, respectively. So, the RBFNN method is a useful and successful method to predict the electrophoretic mobilities of peptides for large data set.

SVM is a new algorithm developed from the machine learning community. Due to its remarkable generalization performance, the SVM has attracted attention and gained extensive application. In 2005, for the first time, Liu et al.[45] used SVM to predict the electrophoretic mobilities of 139 polypeptides using the nine descriptors calculated from the molecular structure alone. The whole data set was divided into training set and test set to construct a validate model. The optimized parameters of SVM were $C = 100$, $\varepsilon = 0.04$, and $\gamma = 0.002$, respectively. The obtained SVM model gave squared correlation coefficients ($R^2$) of 0.925, which was better than the MLR model which gave squared correlation coefficients ($R^2$) of 0.904. From the *t*-test value, one can find the three descriptors: average information content (order 1), number of benzene rings, and relative number of H atoms have the largest influence on the electrophoretic mobilities, and then obtained some insight to the electrophoretic behavior of the peptides.

Recently, Yu et al.[46] made a further study for 102 large peptides based on the studies of Jalali-Heravi et al.[43] They used four methods, which consisted of MLR, BPANN, RBFNN, and SVM, to establish their model. The four models gave squared correlation coefficients ($R^2$) of 0.913, 0.970, 0.980, and 0.980, respectively, which were better than the result of Jalali-Heravi et al.[43] The results showed that these machine learning techniques, especially RBF-ANN and SVM, were effective and efficient for the development of the accurate and reliable QSPR models, which was helpful for peptide separations.



## CONCLUSION

The diverse studies discussed herein clearly show the importance of QSPR studies as related to peptides. With the improvement of computational chemistry and molecular modeling methods, the theoretical descriptor can comprehensively describe the feature structures of molecules. The use of a new regression algorithm, such as RBFNN and SVM, can effectively establish the relationship between molecular structure feature and property. These models can provide some theoretical guide for the fast experimental condition optimization.

## ACKNOWLEDGMENT

The authors thank the financial support of the key program of National Natural Science Foundation of China (NSFC, No. 90612016)

## REFERENCES


1. Wolters, D.A.; Washburn, M.P.; Yates, J.R. An automated multidimensional protein identification technology for shotgun proteomics. Anal. Chem. **2001**, *73*, 5683–5690.
2. Kasicka, V. Recent advances in capillary electrophoresis and capillary electrochromatography of peptides. Electrophoresis **2003**, *24*, 4013–4046.
3. Tessier, B.; Blanchard, F.; Vanderesse, R.; Harscoat, C.; Marc, I. Applicability of predictive models to the peptide mobility analysis by capillary electrophoresis-electrospray mass spectrometry. J. Chromatogr A **2004**, *1024*, 255–266.
4. Kuhn, R.; Hoffstetter-Kuhn, S. *Capillary Electrophoresis—Principle and Practice*, Springer-Verlag: Berlin, 1993.
5. Fatemi, M.H.; Goudarzi, N. Quantitative structure property relationship study of the electrophoretic mobilities of some benzoic acids derivatives in different carrier electrolyte compositions. Electrophoresis **2005**, *26*, 2968–2973.
6. Yao, X.J.; Liu, M.C.; Zhang, X.Y.; Hu, Z.D.; Fan, B.T. Radial basis function network-based quantitative structure-property relationship for the prediction of Henry's law constant. Anal. Chim. Acta **2002**, *462*, 101–117.
7. Xue, C.X.; Zhang, R.S.; Liu, H.X.; Yao, X.J.; Liu, M.C.; Hu, Z.D.; Fan, B.T. QSAR models for the prediction of binding affinities to human serum albumin using the heuristic method and a support vector machine. J. Chem. Inf. Comp. Sci. **2004**, *44*, 1693–1700.
8. Luan, F.; Ma, W.P.; Zhang, H.X.; Zhang, X.Y.; Liu, M.C.; Hu, Z.D.; Fan, B.T., Prediction of pK(a) for neutral and basic drugs based on radial basis function neural networks and the heuristic method. Pharmaceut. Res. **2005**, *22*, 1454–1460.
9. Katritzky, A.R.; Lobanov, V.S.; Karelson, M. QSPR: The Correlation and quantitative prediction of chemical and physical properties from structure. Chem. Soc. Rev. **1995**, 279–287.
10. Karelson, M.; Maran, U.; Wang, Y.; Katritzky, A.R. QSPR and QSAR models derived using large descriptor spaces. A review of CODESSA applications. Collect. Czech. Chem. Commun. **1999**, *64*, 1551–1571.
11. Katritzky, A.R.; Maran, U.; Lobanov, V.S.; Karelson, M. Structurally diverse quantitative structure-property relationship correlations of technologically relevant physical properties. J. Chem. Inf. Comput. Sci. **2000**, *40*, 1–18.
12. *Chemdraw 2006*; Cambridge Soft Corporation, 2006.
13. *MDL ISIS Draw 2.5*; MDL Information Systems, Inc., 2002.
14. *HyperChem 4.0*; Hypercube, Inc., 1994.
15. *Gaussian 98*, Revision A.7; Gaussian Inc., 1998.
16. *Sybyl 7.2*; Tripos Associates, Inc., 2003.
17. *MOPAC 2000*; Fujitsu Limited, 1999.
18. Xia, B.B.; Ma, W.P.; Zhang, X.Y.; Fan, B.T. Quantitative structure-retention relationships for organic pollutants in biopartitioning micellar chromatography. Anal. Chim. Acta **2007**, *598*, (1), 12–18.

19. Dragon 5.4, Talete srl, Milan, Italy, 2006.
20. Turner, J.V.; Glass, B.D.; Agatonovic-Kustrin, S. Prediction of drug bioavailability based on molecular structure. Anal. Chim. Acta **2003**, *485*, 89–102.
21. Zhang, D.R. QSPR studies of PCBs by the combination of genetic algorithms and PLS analysis. Comput. Chem. **2001**, *25*, 197–204.
22. Oblak, M.; Randic, M.; Solmajer, T. Quantitative structure-activity relationship of flavonoid analogues. 3. Inhibition of p56(lck) protein tyrosine kinase. J. Chem. Inf. Comp. Sci. **2000**, *40*, (4), 994–1001.
23. Deconinck, E.; Coomans, D.; Heyden, Y.V. Exploration of linear modelling techniques and their combination with multivariate adaptive regression splines to predict gastro-intestinal absorption of drugs. J. Pharmaceut. Biomed. **2007**, *43*, 119–130.
24. Tantishaiyakul, V.; Worakul, N.; Wongpoowarak, W. Prediction of solubility parameters using partial least square regression. Int. J. Pharm. **2006**, *325*, 8–14.
25. Katritzky, A.R.; Gordeeva, V. Traditional topological induces vs electronic, geometrical, and combined molecular descriptors in QSAR QSPR research. J. Chem. Inf. Comput. Sci. **1993**, *33*, 835–857.
26. Kachigan, S.K. *Statistical Analysis*, Radius Press: New York, 1986.
27. Garg, P.; Verma, J. In silico prediction of blood brain barrier permeability: An artificial neural network model. J. Chem. Inf. Comp. Sci. **2006**, *46*, 289–297.
28. Li, Q.F.; Yao X.J.; Chen X.G.; Liu, M.C.; Zhang, R.S.; Zhang, X.Y.; Hu, Z.D. Application of artificial neural networks for the simultaneous determination of a mixture of fluorescent dyes by synchronous fluorescence. Analyst **2000**, *125*, 2049–2053.
29. Xue, C.X.; Liu, H.X.; Yao, X.J.; Liu, M.C.; Hu, Z.D.; Fan, B.T. Study of quantitative structure-mobility relationship of carboxylic and sulphonic acids in capillary electrophoresis. J. Chromatogr. A **2004**, *1048*, 233–243.
30. Ma, W.P.; Zhang, X.Y.; Luan, F.; Zhang, H.X.; Zhang, R.S.; Liu, M.C.; Hu, Z.D.; Fang, B.T. Support vector machine and the heuristic method to predict the solubility of hydrocarbons in electrolyte. J. Phys. Chem. A **2005**, *109*, 3485–3492.
31. Maggiora, G.M.; Elrod, D.W.; Trenary, R.G. Computational neural networks as nodel-free mapping devices. J. Chem. Inf. Comput. Sci. **1992**, *32*, 732–741.
32. Cortes, C.; Vapnik, V. Support-vector networks. Mach. Learn. **1995**, *20*, 273–297.
33. Wang, W.J.; Xu, Z.B.; Lu, W.Z.; Zhang, X.Y. Neurocomputing **2003**, *55*, 643–663.
34. Cifuentes, A.; Poppe, H. Behavior of peptides in capillary electrophoresis: Effect of peptide charge, mass and structure. Electrophoresis **1997**, *18*, 2362–2376.
35. Sanz-Nebot, V.; Benavente, F.; Barbosa, J. Liquid chromatography-mass spectrometry and capillary electrophoresis combined approach for separation and characterization of multicomponent peptide mixtures – Application to crude products of leuprolide synthesis. J. Chromatogr A **2002**, *950*, 99–111.

36. Offord, R.E. Nature (London) **1966**, *211*, 591–593.
37. Cifuentes, A.; Poppe, H. Simulation and optimization of peptide separation by capillary electrophoresis. J. Chromatogr A **1994**, *680*, 321–340.
38. Rickard, E.C.; Strohl, M.M.; Nielsen, R.G. Correlation of electrophoretic mobilities from capillary electrophoresis with physicochemical properties of proteins and peptides. Anal. Biochem. **1991**, *197*, 197–207.
39. Janini, G.M.; Metral, C.J.; Issaq, H.J.; Muschik, G.M. Peptide mobility and peptide mapping in capillary zone electrophoresis – Experimental determination and theoretical simulation. J. Chromatogr A **1999**, *848*, 417–433.
40. Solinova, V.; Kasicka, V.; Koval, D.; Hlavacek, J. Separation and investigation of structure-mobility relationships of insect oostatic peptides by capillary zone electrophoresis. Electrophoresis **2004**, *25*, 2299–2308.
41. Solinova, V.; Kasicka, V.; Sazelova, P.; Barth, T.; Miksik I. Separation and investigation of structure-mobility relationship of gonadotropin-releasing hormones by capillary zone electrophoresis in conventional and isoelectric acidic background electrolytes. J. Chromatogr A **2007**, *1155*, 146–153.
42. Jalali-Heravi, M.; Shen, Y.; Hassanisadi, M.; Khaledi, M.G. Prediction of electrophoretic mobilities of peptides in capillary zone electrophoresis by quantitative structure-mobility relationships using the offord model and artificial neural networks. Electrophoresis **2005**, *26*, 1874–1885.
43. Jalali-Heravi, M.; Shen, Y.; Hassanisadi, M.; Khaledi, M.G. Artificial neural network modeling of peptide mobility and peptide mapping in capillary zone electrophoresis. J Chromatogr A **2005**, *1096*, 58–68.
44. Ma, W.P.; Luan, F.; Zhang, H.X.; Zhang, X.Y.; Liu, M.C.; Hu, Z.D.; Fan, B.T. Accurate quantitative structure-property relationship model of mobilities of peptides in capillary zone electrophoresis. Analyst **2006**, *131*, 1254–1260.
45. Liu, H.X.; Yao, X.J.; Xue, C.X.; Zhang, R.S.; Liu, M.C.; Hu, Z.D.; Fan, B.T. Study of quantitative structure-mobility relationship of the peptides based on the structural descriptors and support vector machines. Anal. Chim. Acta **2005**, *542*, 249–259.
46. Yu, K.; Cheng, Y.Y. Machine learning techniques for the prediction of the peptide mobility in capillary zone electrophoresis. Talanta **2007**, *71*, 676–682.

Manuscript 62731