Abstract

Shen, Yang. Prediction of peptide maps in CZE and MEKC systems(Under the direction of Dr. Morteza G. Khaledi).

A new Quantitative Structure-Migration Relationships(QSMR) model was developed to predict the electrophoretic mobilities of peptides in capillary zone electrophoresis(CZE). A three-step strategy was used: first, select the best charge-size term from the existing models; second, develop a muilti-linear regression(MLR) model to study the linear characteristics of peptide mobility using the best charge-size term and other descriptors; third, generate an artificial neural network(ANN) to investigate the nonlinear behavior of peptide mobility and use this ANN model to predict peptide migration behavior in CZE. To test the robustness of the QSMR model, it was applied to the data published by another research group. Very accurate predictions were achieved.

To study the influence of peptide sequence on the migration of a peptide in CZE, a series of "sequence-related" descriptors were developed. These descriptors were used to develop MLR models for peptide mobility prediction. With the "sequence-related" descriptor, more accurate mobility could be predicted for peptides with same amino acid composition but different sequences.

Group contribution approach(GCA) was used to determine the individual contribution of each amino acid residue and both N-, C- terminal to the peptide mobility in Tween20 system. Data of a relatively small number of peptides were used for this purpose. The sum of individual contributions was calculated for each peptide and used as a new descriptor in developing MLR models for the prediction of peptide mobilities in Tween20 system. Good preliminary results were achieved.

Prediction of peptide maps in CZE and MEKC systems

by

Yang Shen

A thesis submitted to the Graduate Faculty of
North Carolina State University
in partial fulfillment of the
requirements for the Degree of
Master of Science

Department of Chemistry

Raleigh, NC 27695

2005

APPROVED BY:

Charles B. Boss	**Edmond F. Bowden**

Morteza G. Khaledi
(Chair of Advisory Committee)

Biography

Yang Shen attended Fudan University from 1992-1997 where he obtained a B.S degree in chemistry. In the fall of 2000, he began graduate study at North Carolina State University where he received a M.S degree in Analytical Chemistry under the direction of Dr. Morteza G.Khaledi.

Acknowledgements

First of all, I'd like to thank my wife, Ying, for her love and support throughout all these years. Likewise, my parents, Tongsheng Shen and Qinglai Li, have always been there when I need them. I'd like to thank Dr. Morteza G.Khaledi for his guidance and invaluable advice. Also, I'd like to thank Dr. Mehdi Jalali-Heravi for his help and enlightening discussion. And I'd like to show my appreciation to Dr. Damian Shea for letting me use his Beckman CE instrument. Finally, I'd like to thank the whole Khaledi research group for their friendship and useful discussion.

Table of Contents

Multi-dimensional separation setup

Appendix 1: Matlab program for the calculation of "sequence-related" descriptors

List of Tables

List of Figures

Page

Chapter 2: Introduction

Chapter 3: Artificial neural network modeling of peptide mobility and peptide mapping in capillary zone electrophoresis

Chapter 1 Introduction

Keywords: Capillary Electrophoresis, Peptide Mapping, Capillary Zone Electrophoresis, QSMR, Micellar Electrokinetic Chromatography, QSPR, Multi-dimensional

1. Peptide separation using Capillary Electrophoresis

Peptides are a large group of complex biomolecules that play very important roles in living organisms. Peptides can function as hormones, coenzymes, toxins, drugs and antibiotics. Separation and characterization of complex peptide samples pose a great challenge to peptide scientists, and also leads to the rapid and intensive development of capillary electrophoresis technique.

Capillary electrophoresis(CE) has received a great deal of attention for peptide separations due to several advantages such as high efficiency, speed, small sample size, automation, and high-throughput capability. Different capillary electrophoresis techniques has been used for this purpose such as capillary zone electrophoresis(CZE), capillary isotachophoresis(CITP), capillary isoelectric focusing(CIEF), micellar electrokinetic chromatography(MEKC) and affinity electrophoresis. These techniques can combine with SDS-PAGE or HPLC to form two-dimensional separation systems. Using mass spectrometry as on-line detection technique, they can even form three-dimensional separation systems.

Capillary electrophoresis has thus attracted more attention and become a recognized counterpart and complement to most frequently used techniques for peptide separations.

2. Peptide mapping

Peptide mapping is the separation of peptide fragments originating from specific chemical or enzymatic hydrolysis of proteins. It is a very important tool for protein identification, sequence determination, monitoring of post-translational modifications and

protein structure elucidation. Due to the high complexity of the peptide maps that are generated, for complete resolution, multidimensional separation methods are usually used. It will be very time-consuming to identify each peak in the peptide map. To facilitate peptide research, different theoretical or semi-empirical models have been developed to predict either mobility or partition coefficient of peptides in certain separation systems[1-4]. Relatively accurate prediction has been achieved in some of the published results.

3. Peptide mobility prediction in capillary zone electrophoresis

Capillary zone electrophoresis(CZE) is the simplest mode of capillary electrophoresis. In CZE, several theoretic or semi-empirical models have been developed to explain the dependence of peptide electrophoretic mobility on charge and size[1-4]. All these models are derived from Stoke's law that describes the motion of an ion in an electric field(Eq.1):

$$\mu_{ef} = \frac{q}{6\pi\eta r} \qquad \text{(Eq.1)}$$

Where μ_{ef} is the effective electrophoretic mobility at a given ionic strength, r is effective ion radius, q is net charge of the ion and η is viscosity of the solution. Stoke's law is mainly valid for rigid spherical molecules in low ionic strength buffers which doesn't fit peptide very well. Offord, Jokl, Grossman and Cifuentes developed their own models by modifying either charge term or size term[1-4]. With these models, more accurate peptide mobility can be predicted. When applied to hydrophobic and highly charged peptides, neither of these models can give accurate predictions. In this work, new Quantitative Structure-Migration Relationships(QSMR) models were developed to take into consideration of steric effect and peptide bulkiness. Artificial neural networks were also

developed to study the nonlinear characteristic of peptide mobility in CZE. With this QSMR model, more accurate predictions have been achieved[5]. To test the robustness of our QSMR model, it was applied to the peptide data set from Janini's work and very good results were achieved[6]. With all previous developed models, peptides with same amino acid composition but different sequence will always have the same electrophoretic mobility while their experimental mobilities are different. A new QSMR model was developed to improve the prediction accuracy for the mobilities of peptides with different sequence[7].

4. Peptide mobility and partition coefficient prediction in micellar electrokinetic chromatography

Since its introduction by Terabe et al in 1984, Micellar Electrokinetic Chromatography (MEKC) has been widely studied and applied to different areas of separation science[8]. It has many advantages such as high efficiency, short analysis time, automation, small sample size, little solvent consumption and no requirement for sample purity. It separates solutes on the basis of their differential partitioning into the micellar pseudo-stationary phases. A limited elution window exists in MEKC. All uncharged solutes will be separated between the migration time of unretained solute (T_{eo}) and the migration time of micelles (T_{mc}). MEKC can be viewed as the hybrid of RPLC and CZE, as the separation process incorporates hydrophobic and polar interactions, partitioning mechanism and electromigration. It's the only capillary electrophoresis method that can separate mixtures of charged and uncharged solutes. And MEKC can offer much higher separation efficiency than RPLC.

Another big advantage of MEKC over conventional chromatographic techniques is its flexibility and ease of changing the chemical composition of pseudo-stationary phases. By simply rinsing with a new micellar solution, the new pseudo-stationary phase is introduced into the system, and the equilibration time is usually very short. There are many kinds of surfactants that can offer very different selectivity such as bile salts, anionic alkyl chain surfactants, cationic alkyl chain surfactants, nonionic surfactants, zwitterionic surfactants, chiral surfactants and fluorocarbon surfactants. In addition to that, many modifiers such as organic solvents, urea, glucose and cyclodextrins can be added to MEKC systems to further improve the selectivity.

Since CZE can't resolve all peptides in complex peptide mixtures and protein digests, several MEKC systems were tried to improve the separation. SDS/Hexanol system and Tween20 system system were selected in this work. SDS is anionic surfactant while Tween20 is nonionic surfactant. They offer distinctively different selectivity in peptide separations.

As to the best of our knowledge, no effort has been made toward developing peptide mobility or partition coefficient prediction models in MEKC systems. In this work, a good Quantitative Structure-Migration Relationships(QSMR) model have been develop for the prediction of peptides' electrophoretic mobilities in 10mM Tween20 system[9]. Good preliminary results have also been achieved in developing a Quantitative Structure-Partition Relationships(QSPR) model for the micelle-water partition coefficient prediction for peptides in SDS/Hexanol system.

5. Multi-dimensional separation techniques for peptide mapping

In spite of the high separation power of individual electrophoresis techniques, for complete separation of complex peptide mixtures and protein digests, multi-dimensional separation systems are often necessary. 2-D PAGE is the most commonly used 2-D technique that uses IEF as the first dimension and SDS-PAGE as the second dimension[10]. An alternative 2-D technique is LC-CZE[11]. It uses RPLC to separate peptides according to their different hydrophobicity in the first dimension and uses CZE to further separate peptides according to their different charge-to-size ratio in the second dimension. In this work, MEKC and CZE will be combined together to form a new 2-D technique for peptide mapping. MEKC will be used as the first dimension, which will separate peptides according to their different interaction with the micelles. Peptides will be further separated using CZE as the second dimension. This technique can be easily coupled with Mass Spectrometry to form a 3-D technique.

A home-made 2-D(MEKC-CZE) system is under development in our group. With this system, we can record 2-D peptide maps and use these maps to test and validate our QSMR and QSPR prediction models.

References

[1] Offord, R. E., Nature (London) 1966, 211, 591-593.

[2] Jokl, V., J. Chromatogr. 1964, 13, 451-458.

[3] Grossman, P. D., Colburn, J. C., Lauer, H. H., Anal. Biochem. 1989, 179(1), 28-33.

[4] Cifuentes, A., Poppe, H., J. Chromatogr. A 1994, 680, 321-340.

[5] Jalali-Heravi, M., Shen, Y., Hassanisadi, M. and Khaledi, M.G., Manuscript submitted to Electrophoresis.

[6] Jalali-Heravi, M., Shen, Y. and Khaledi, M.G., Manuscript in preparation

[7] Shen, Y., Jalali-Heravi, M. and Khaledi, M.G., Manuscript in preparation

[8] Terabe, S., Otsuka, K., Ichikawa, K., Tsuchiya, A., Ando, T., Anal. Chem. 1984, 56, 111-113

[9] Shen, Y., Jalali-Heravi, M. and Khaledi, M.G., Manuscript in preparation

[10] Smithies, O., Poulik, M. D., Nature 1956, 177, 1033-1035

[11] Moore, A. V., Jorgenson, J. W., Anal. Chem. 1995, 67, 3448-3455

Chapter 2 Prediction of electrophoretic mobilities of peptides in Capillary Zone Electrophoresis by Quantitative Structure Mobility Relationships using Offord Model and Artificial Neural Networks

Keywords: Capillary Electrophoresis, Peptide Mapping, QSMR

1. **Abstract**

The aim of this work was to explore the usefulness of empirical models and multivariate analysis techniques in predicting electrophoretic mobilities of small peptides in capillary zone electrophoresis (CZE). The data set consists electrophoretic mobilities, measured at pH 2.5, for 125 peptides ranging in size between 2 and 14 amino acids. Among the existing empirical models, the Offord model (i.e. $\mu \equiv Q/M^{2/3}$) gave the best correlation for the data set. A Quantitative Structure Mobility Relationship (QSMR) was developed using the Offord's charge over mass term ($Q/M^{2/3}$) as one descriptor combined with the corrected steric substituent constant ($E_{s,c}$) and molar refractivity (MR) descriptors to account for the steric effects and bulkiness of the amino acids side chains. The multi-linear regression (MLR) of the data set showed an improvement in the predictive ability of the model over the simple Offord's relationship. A 3-4-1 Back Propagation Artificial Neural Networks (BP-ANN) model resulted in a significant improvement in the predictive ability of the QSMR over the multi-linear regression (MLR) treatment, especially for peptides of higher charges that contain basic amino acids arginine, histidine and lysine. The improved correlations by the BP-ANN analysis suggest the existence of nonlinear characteristic in the mobility – charge relationships.

2. **Introduction**

Capillary Electrophoresis (CE) has received a great deal of attention for peptides separations due to several advantages such as high efficiency, speed, small sample size, automation, and high-throughput capability [1]. A less noticed characteristic of capillary zone electrophoresis (CZE), however, is its simplicity and predictable migration patterns

as charged molecules are resolved due to their differences in electrophoretic mobilities that are proportional to their charge to size ratio. The simple relationship between mobility and molecular properties of charged solutes provides an opportunity for prediction of migration patterns from quantitative structure – mobility relationships (QSMR) that could facilitate method development and optimization of separations.

In CZE, electrophoretic mobilities of charged compounds is expressed as:

$$\mu_{ef} = \frac{q}{6\pi\eta r} \qquad \text{(Eq. 1)}$$

where μ_{ef} is effective electrophoretic mobility at a given ionic strength, r is effective ion radius, q is ion charge, and η is solution viscosity.

Several empirical models have been developed for the calculation/prediction of electrphoretic mobilities [2-8]. Most of these models are based on Stoke's law for ion mobility in an electric field that is mainly valid for rigid spherical molecules in low ionic strength buffers [9]. Generally, it has been shown that the electrophoretic mobility is proportional to the charge *Q* and inversely proportional to the molecular mass *M* as:

$$\boxed{\mu = a\frac{Q}{M^b}} \qquad \text{Eq. 2}$$

Where *a and b are constants.* The main difference between various reported models is the *b* value that depends upon the assumptions involved in the derivation of the models and the conditions under which these assumptions are valid. Cross and Cao have presented an interesting discussion in this respect [10]. Jokl [11] reported *b*=0.5 using paper electrophoresis, whereas Offord by considering peptides as large non-spherical ions

determined b=2/3 [3]. On the other hand, Grossman et al. considered peptides as classical linear polymer with n amino acid residues and arrived at an equation that correlated mobility with a function in the form of ln $[(Q+1)/n^{0.43}]$ [2]. Cifuentes and Poppe [12] modified the classical linear model of Grossman et al. by retaining the logarithmic dependence of mobility on charge but substituted molecular mass, M, for n for the size dependence. Janini and coworkers 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 [13]. They concluded that although the Offord model gives the best overall mobility, but it fails when applied to hydrophobic and highly charged peptides. These researchers also showed that peptides electrophoretic mobilities cannot be successfully predicted with reasonable degree of accuracy for all different categories of peptides by relying on two parameter-models, namely charge (Q) and size dependence (n or M) [13].

Inspection of all previous works performed for calculation of peptides electrophoretic mobilities reveals three factors that could be the reasons behind the variations in the reported models and their predictive abilities. One is an assumption that the relationships between electrophoretic mobility with charge and size in the Stoke or Offord – based models are inherently linear, which ignores possible non-linear behavior. Secondly, in order to accurately determine the peptides charges, an accurate knowledge of the ionization constants of the amino acid residues in peptides is required [14]. The available pKa values for the ionizable functional groups in the amino acids are used for all peptides, regardless of their composition and sequence. This could lead to erroneous determination of peptides charges as the ionization of certain amino acid residues in a sequence could be affected by electrostatic and steric interactions with the nearest

neighboring residues [15]. The steric and electrostatic interactions occur according to the composition and sequence of amino acids residues in peptides; which in turn could be a source of error in determination of pKa and subsequently peptides charges depending on the nature of peptides used as the training and/or test sets. Finally, there exists a disagreement in the literature about the dependence of mobility on molar mass. Compton has shown that the mobilities of small molecules in low ionic strength buffer are more closely correlated with $1/M^{1/3}$ while large molecules in high ionic strength buffer correlated with $1/M^{2/3}$ [4]. Molecules of intermediate-size and in moderate ionic strength buffers show dependence on ($1/M^{1/2}$) [4]. Janini et al. based on a data set of 58 peptides concluded that except for the highly-charged and the hydrophobic peptides the Offord model is superior to the other models [13].

In this work, we first identified the best model reported in the literature correlating electrophoretic mobility with charge over size ratios (i.e. $\mu \equiv Q/M^{1/3}$, $Q/M^{1/2}$, $Q/M^{2/3}$ and $log\ (1+0.297Q)/M^{0.411}$ for a set of 125 peptides ranging in size between 2 and 14. The next step involved development of a Quantitative Structure Mobility Relationships (QSMR) model that incorporated the charge over mobility term from the best model in the first part as a descriptor along with other structural descriptors to further improve the predictive ability. The QSMR model is based on the multiple linear regression (MLR) of mobility as a function of peptides structural descriptors for predicting mobilities of peptides and investigating their linear characteristics. Finally, a hybrid method consisting of MLR and artificial neural network (MLR-ANN) was developed to study the non-linear characteristics of electrophoretic mobilities of 125 small peptides at pH 2.5 in bare silica capillaries in CZE. In the latter strategy, the MLR

served as a feature selection technique for choosing the most suitable inputs (i.e. peptides descriptors) for the neural network.

Artificial neural networks (ANNs) are a compact group of connected, ordered in layer elements, which are able to process information. ANN-based approaches have several advantages that include a capacity to self-learn and to model complex data without the need for a detailed understanding of the underlying phenomena [16-22].

To the best of our knowledge, the application of ANN for modeling and prediction of peptides' electrophoretic mobilities has not been reported previously. The MLR and ANN models were derived based on the assumption that the peptide electrophoretic mobility should substantially depend on amino acid compositions. The development of robust models for calculation of electrophoretic mobilities of peptides will be of great use for prediction and simulation of peptide maps of proteins. It is demonstrated that ANNs are capable of efficiently reproducing the electrophoretic mobilities of peptides in CZE.

3. Experimental section

3.1 Peptides, Chemicals and Materials

A total of 125 standard peptides were purchased from either Sigma Chemical Co. (St. Louis, MO) or Bachem Co. (Torrance, CA). These peptides are listed in Table 1. Also, the values of the charge over mass term in Offord model ($Q/M^{2/3}$), corrected steric constituent constant, $E_{s,c}$, and molar refractivity, MR, are given in Table 1 for all peptides. The net charge of a peptide was calculated as the sum of the net charges of all charged amino acid residues and the N- and C-terminals. The net charge of each charged amino acid residue, N-terminal and C-terminal at pH 2.5 was determined using Henderson-

Hasselbach equation [14]. Each arginine(R), histidine (H), lysine (K) and N-terminal contribute a charge of +1; while each aspartic acid (D) (pKa 3.5), glutamic acid (E) (pKa 4.5) and the C-terminal (pKa 3.2) contributes a charge of –0.091, –0.01 and –0.166, respectively.

The buffer solution was 50 mM sodium phosphate (pH 2.5) throughout the study. It was prepared by adjusting the pH of phosphoric acid to 2.5 using sodium hydroxide. High purity phosphoric acid was purchased from Aldrich Chemical Company (Milwaukee, WI), and sodium hydroxide was purchased from Fisher Scientific (Pittsburgh, PA). The sodium phosphate buffer was filtered through a 0.2 μm Acrodisc filter (STRL, Eatontown, NJ) before use.

Bare-silica capillary used in this study was purchased from Polymicro Technologies (Phoenix, AZ), with the inner diameter of 50 μm and the outer diameter of 375 μm. The total length and the length from the injection port to the detection window were 37 and 30 cm, respectively.

3.2 Equipment

A Beckman P/ACE 2200 instrument with a UV detector was used in this work. The detection wavelength was set at 214 nm and 37 °C was used as the running temperature. The voltage was set at 15 kV. Low-pressure injection (0.5 psi, 2 seconds) was used to introduce the sample into the capillary. Before any injection, the bare-silica capillary was conditioned by the following rinsing procedure: MilliQ water for 10 minutes**,** 1M NaOH solution for 10 minutes, pure Methanol for 10 minutes, MilliQ water for 10 minutes and buffer for 20 minutes. Between injections, the capillary was pressure rinsed (20 psi) with buffer solution for 5 minutes.

3.3 Determination of the effective mobility of peptides.

In the present work, mesityl oxide was selected as the EOF marker. Each standard was prepared into a solution of about 1mg peptide/ml. Each solution was introduced into the CE instrument by low pressure (0.5 psi) for 2 seconds and its individual retention time was recorded.

Migration times and mobilities of the peptides in this study were measured by preparing several mixtures, each containing 5-6 peptides (200 μl each) and 20 μl of mesityl oxide. Identity of each peptide peak was confirmed by matching its retention time with the individual retention time of the peptides in this mixture. Three injections were made for each sample mixture.

The effective mobility of peptides was determined by Eq. 3.

$$\mu_{ef} = \frac{L_t L_d}{V}\left(\frac{1}{t_r} - \frac{1}{t_{eo}}\right) \qquad (3)$$

where L_t is the total length of the capillary, L_d is the length from the capillary inlet to the detection point, V is the applied voltage, t_r is the peptide retention time and t_{eo} is the retention time of the EOF marker.

3.4. Reproducibility of experiments.

Each sample mixture was injected for three times. Excellent reproducibility was achieved. The range of RSD of the peptides effective mobilities was from 0.02 to 4.46, with an average of 1.95 for the whole data set.

3.5. Data Analysis Methodology

The main goal of the present work was development of multivariate analysis models for accurate prediction of electrophoretic mobilities of peptides. For this purpose, our strategy consisted of the following steps:

1. Measuring the electrophoretic mobilities of the training set of peptides with different compositions and sequence ranging in size between 2 and 14 amino acids at pH 2.5 using bare silica capillaries. 2. Using a subjective method to select the proper descriptors as inputs for the generation of the QSMR models. This method defines the relation between the dependent variable (i.e. electrophoretic mobility of the peptides) and the peptides structural descriptors as independent variables. 3. Development of a linear regression method as a feature selection technique and a calibration method to study the linear characteristics of the mobility of the model peptides. 4. Generation of an artificial neural network as a nonlinear method to investigate the nonlinear behavior of mobility and development of a model to accurately predict peptides migration behavior.

3.6. Generation of multiple linear regression model

The stepwise multiple linear regression procedure was used for model generation. The charge over mass term in the Offord model (i.e. $Q/M^{2/3}$) was chosen as a hybrid descriptor for the QSMR model. Then, the stepwise addition method implemented in the software of Minitab [23] was used for choosing other descriptors contributing to the electrophoretic mobility. In this step, all previously reported descriptors by different research groups were used as input for the regression analysis using the stepwise procedure. The list of different descriptors that were examined is given in the Results and Discussion section. This procedure was repeated until all descriptors were entered in the model. Various models were compared and the best MLR model was chosen on the basis

of correlation coefficient, F statistics, and a standard error. The best MLR model was consisted of three descriptors of Offord's charge over size term [3], steric substituent constant [24] and molar refractivity [25-26] as Eq. 4:

$$\mu = p\frac{Q}{M^{2/3}} + e\sum E_s + m\sum MR \qquad (4)$$

3.7. Neural network generation

A feed forward back propagation of error artificial neural network (BP-ANN) was written in C++. A set of three descriptors appearing in the MLR model was used as input parameters for generation of the network. The selection of the same descriptors for the MLR and ANN models was for the sake of comparison between the linear and nonlinear characteristics of these methods in predicting the electrophoretic mobilities of the model peptides. The signals from the output layer represent the electrophoretic mobilities of the peptides. Such an ANN may be designed as a 3-n_h-1 net, indicating that the number of nodes in the hidden layer should be optimized. The input signal to a node is modulated by a weight (w) along each link. The net input to a node is thus a function of all signals to a node and all of its associated weights. The net input for a node j is given by

$$\text{Net}_j = \sum w_{ji}\, Oi \qquad (3)$$

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 the node i. The net input to a neuron undergoes an additional transformation using a transfer function. In this investigation, the log-sigmoid function, i.e., $f(x) = 1/(1+\exp(-x))$, was used as a transfer function.

The initial weights were chosen randomly, and were optimized based on the delta rule through back propagation of errors. Basically the sum of the initial weights for each layer should be 1. The program is written in such a way that the randomized weights depend on the number of input, hidden and output nodes. Before training, the output and inputs (except for the values of the Offord model) were normalized between 0 and 1. Also, the network parameters of the number of neurons at the hidden layer, learning rate and momentum were optimized. Procedures for optimization of these parameters were reported in previous papers [27-30].

4. Results and Discussion.

The main goals of the present work were as follows: (1) to accurately predict the electrophoretic mobilities of relatively small peptides in CZE; (2) to achieve a better understanding of the physicochemical basis of the motion of a peptide in an electric field and (3) to investigate the linear and non-linear relationship between electrophoretic mobilities of peptides and their size and charge. To fulfill these goals a diverse data set consisting of 125 peptides, ranging in size from 2 to 14 amino acids and charge from 0.743 to 5.843 was chosen and their electrophoretic mobility were measured at pH 2.5 using CZE. Table 1 shows the experimental values of electrophoretic mobility for all peptides ranging from 13.762 x 10^{-5} to 56.533 x 10^{-5} $cm^2 v^{-1} s^{-1}$. In the present work, these peptides were randomly divided into three groups of training, test and validation sets (Table 1). The training set consists of 90 peptides and was used for model generation. The test and validation sets consist of 20 and 15 peptides, respectively. The test set is useful for considering the over-fitting of the neural network and the validation set

evaluates the generated models. The peptides in the validation set have not been used in the optimization procedure of the neural network and were chosen in such a way that represented the training set adequately.

4.1 Identification of the best empirical model.

Several empirical models based on Stoke's law for motion of ions in an electric field have been developed to explain the dependence of electrophoretic mobility on charge and size [2-8]. Among them, the most common models are: $Q/M^{1/3}$, $Q/M^{1/2}$, $Q/M^{2/3}$ and log $(1+0.297\ Q)/M^{0.411}$. Much work with conflicting conclusions has been reported on applying these models for peptide electrophoretic mobility modeling. Janini and coworkers [13] and Rickard et al. [15] have shown that the Offord model (μ_{ef} vs $Q/M^{2/3}$) offers the best fit to their experimental data. On the other hand, Grossman et al. based on 40 peptides obtained an excellent correlation for μ_{ef} versus ln $[(1+Q)/n^{0.43}]$ [2]. While Hilser et al. [31] and Cifuentes and Poppe [12] offered the Grossman et al. model, Gaus and coworkers [32] showed that neither the Offord model nor the Grossman et al. model is suitable. Chen and coworkers showed an excellent correlation between the peptide mobility data and $Q/M^{1/2}$ [33]. On the other hand, Survay and coworkers [34] concluded that none is uniquely suitable for peptide mobility modeling.

One of the objectives of the present work was to test which empirical model most accurately reproduces the experimental values of the electrophoretic mobility for the model peptides. Therefore, the electrophoretic mobilities of 125 peptides (Table 1) were measured at pH 2.5 by CZE. The acidic pH 2.5 is a suitable condition for this application, because the silanol groups on the capillary walls are protonated that reduces or eliminates peptides wall adsorption problems. The peptides are positively charged at

this pH, and in the absence of EOF migrate towards the cathode. Fig. 1a shows the correlation between the peptides mobilities and the values of charge-to-size ratios obtained using different models. It should be noted that for all models, including Cifuentes et al model, the charge of peptides was calculated employing the Henderson-Hasselbach equation [14]. However, Cifuentes and coworkers in their original work, have used different charges obtained using their computer program [12]. To calculate charges, they have employed the pK values obtained considering the electrostatic effects among neighboring charges [12]. Fig. 1b shows the plot of the values of charge-to-size ratios versus the peptide mobilities obtained using the Offord model. Table 2 shows the linear regression equations for the data shown in Fig. 1 and the corresponding correlation factors (R^2). Inspection of the plots and the R^2 values demonstrate the superiority of the Offord model. However, Cifuentes et al. model (log (1+0.297 Q)/$M^{0.411}$) also shows a reasonable correlation, which is in contradiction to Janini et al. results [13]. This could be due to the logarithmic dependence of Q in this model, which compensates for electrostatic charge suppression due to mutual electrostatic interactions of the charged group that is significant for highly charged peptides [2,12]. On the other hand, in agreement with Janini et al. [13], our results show a poor correlation with the charge-to-size values obtained using $Q/M^{1/3}$ model ($R^2 = 0.63$). The R^2 of 0.78 for the $Q/M^{1/2}$ model (Table 2) is in good agreement with the value of 0.80 of Janini et al. based on 58 peptides [13]. However, inspection of the R^2 values in Table 2 reveals that in agreement with Survay et al. [34], none of the models is able to reproduce adequately the electrophoretic mobility of small peptides. Our results confirm the conclusion that electrophoretic mobility cannot be successfully predicted for all different categories of peptides by

relying on only two parameters of charge and size. It seems that some other factors affecting the mobility of peptides in CZE.

4.2 Multiple linear regression analysis

In order to improve the predictive ability of the Offord model, we investigated various QSMR models that incorporated other structural descriptors of peptides in addition to charge and size. Thus, in the QSMR models, the ratio of $Q/M^{2/3}$ was used as a hybrid descriptor. As a first step in developing a MLR model, one has to choose the most suitable descriptors contributing to the motion of a peptide in an electric field. Several physico-chemical parameters such as effective net charge, molar mass, number of amino acid residue, average residue mass, molecular volume, surface area, hydrophobicity, isoelectric point value, strain parameter, Z-scale and alpha-helix content have been used for peptide mobility modeling [35-36]. Also, an attempt was made by Sak et al. to derive a set of parameters that would quantify the influence of a single amino acid on the conformation of a peptide/protein molecule [36]. Variation of amino acids influences the conformation of a peptide, which in turn characterizes the interaction between the peptide molecule and the surrounding environment. Among different parameters suggested by Sak et al. as molecular descriptors for QSAR studies of peptides, two descriptors of steric constant (Es) and molar refractivity (MR) have appeared in the MLR model generated in the present work. The molar refractivity is a constitutive-additive property that is calculated by the Lorenz-Lorentz formula [25]. MR is strongly related to the volume of the molecules (molecular bulkiness). The steric substituent constant (Es) has been defined by Taft as $\log (k/k_0)$, where k and k_0 are the rate constants for the acidic hydrolysis of a

substituted ester and of a reference ester, respectively [37]. This parameter represents the steric interactions.

The Offord's charge/mass parameter was chosen as the first input for the software package of Minitab [23] to generate the MLR model. Then, the stepwise addition method was used for choosing the other descriptors contributing to the electrophoretic mobilities of model peptides. All of above-mentioned descriptors suggested by previous researchers were used for model generation. The specifications for the best MLR model are shown in Table 3. Also the mean effect for each descriptor is included in this table. The calculated values of electrophoretic mobilities using this model are presented in Table 1 for all peptides studied in this work (training, prediction and validation sets). It can be seen from Table 3 that in addition to the charge-to-size ratio parameter of Offord, corrected steric substituent constant ($E_{s,c}$) and molecular refractivity (MR) have also appeared in the MLR model.

The Offord model shows a mean effect of 28.102, which is the largest among the descriptors appearing in the model. This indicates that the net charge of the peptide and its size play the major roles in the migration mechanism of the peptides in an electric field. The contribution of $E_{s,c}$ and MR to the electrophoretic mobility is almost the same, but in opposite direction. The electrophoretic mobility decreases as the absolute value of $E_{s,c}$ increases. The calculated values of the three descriptors appearing in the best MLR model are given in Table 1 for all peptides included in our data set. Inspection of Tables 1 and 3 reveals that larger peptides show a higher steric constant and therefore, have a smaller mobility in a CZE system. On the other hand, the peptide electrophoretic mobility increases as MR increases. One might expect an opposite trend given the fact that MR

represents molecular bulkiness, however, molecules with larger MR also engage in stronger dispersion interactions with the surrounding water molecules which could have an influence on the effective charge as well as hydration radius. Thus, the exact reason for this behavior is not immediately clear. Fig. 2 shows the plot of the MLR calculated electrophoretic mobility against the experimental values for the validation and test sets. The resulting R^2 of 0.895 shows some improvements compared with the R^2 of 0.878 for the Offord model (Table 2). However, there are some MLR calculated electrophoretic mobilities, which show a large deviation from the experimental values (Table 1).

Inspection of Table 1 reveals that the MLR model overestimates the electrophoretic mobility of peptides containing arginine (R), histidine (H) and lysine (K) amino acids. These amino acids contribute a charge of +1 to the peptide. Therefore, one may conclude that the linear models are not able to predict the mobility of the peptides with high charges. This can be due to the nonlinear behavior of the mobility of highly charged peptides in an electric field. To examine the nonlinear characteristics of the electrophoretic mobility, an artificial neural network was developed in the present work.

4.3 Artificial neural network analysis

The purpose for developing the MLR and ANN models was to compare the abilities of linear and nonlinear models in predicting small peptides electrophoretic mobilities. For the sake of comparison, the descriptors used in the MLR model should be the same as the input parameters for generating the network. Therefore, a BP-ANN was generated by using the three descriptors appearing in the MLR model as its inputs. The experimental values of peptide electrophoretic mobility were considered as target values

and the back-propagation algorithm was used to reduce the output error. According to the neural network methodology, the number of hidden nodes and iterations and the learning and momentum terms were optimized in the present work. The values of these parameters were optimized with the procedure that was reported in our previous works [27-30]. The hidden layer could be increased to a large number of nodes without the back-propagation algorithm being able to reduce the errors. We used a hidden layer with four nodes for the present work, because it reduces the error to a near optimal minimum without potentially sacrificing generality. To maintain the predictive power of the network at a desirable level, typically a test set is used to stop the training when the error for the data set ceases to decrease. Going beyond this point suggests that the ANNs learn from the noise in the training set that is not present in the test set. Therefore, the training was stopped at 5000 epochs. The specifications for the final 3-4-1 neural network are given in Table 4. The predicted electrophoretic mobilities for all peptides by ANN in this work together with their percent deviations are shown in Table 1. To compare the chemometric methods of MLR and ANN in predicting the electrophoretic mobilities of peptides, some statistics for these models are included in Table 5. It can be seen that all statistics have improved considerably in the case of the ANN analysis. The MLR standard error values of 2.78, 3.74 and 3.52 were obtained for the training, test and validation sets, respectively, which should be compared with the corresponding values of 1.86, 2.93 and 2.75 for the ANN model. The standard error of the ANN model is comparable with the level of uncertainty of some experimental deviation of electrophoretic mobility [38]. Fig. 3 shows the plot of the ANN predicted versus the experimental values of the electrophoretic mobility of the validation and test sets. In order to demonstrate the absence of bias, zero intercept unit

slope lines are shown in Figures 2 and 3. Comparison of the zero intercept plots for the MLR and ANN models shows the superiority of the ANN over that of the MLR model. Fig. 4 shows the plot of residuals against the experimental values of peptide mobility for the ANN model. The propagation of the residuals in both sides of zero indicates that no systematic error exists in the development of the ANNs. Inspection of the ANN calculated mobilities (Table 1) reveals a remarkable improvement for those peptides containing the highly charged amino acids of arginine, histidine and lysine. For example, the deviation of –19.40% for the MLR calculated value of KYK peptide should be compared with the 3.95% for the ANN model. Also, for the relatively larger peptides such as CGYGPKKKRKVGG and DRVYIHPFHLVIHN, the MLR deviations of –25.58% and –39.21%, respectively, which should be compared with the deviations of 1.42% and –1.12% for the ANN model. The weakness of the Offord model in predicting the electrophoretic mobility of highly charged peptides has been attributed to inaccurate determination of charges [12,13]. It has been suggested that the calculated charges might deviate from the actual ones due to mutual electrostatic interaction of charge groups in proximity of each other. Although, this could be true, our results reveal that the nonlinear characteristics of the highly charged peptides play the major role in the mechanism of their mobility.

4.4 Conclusions

A three-step strategy was chosen to develop a simple computer-assisted model for accurate prediction of the electrophoretic mobilities of peptides. As a first step, the abilities of the existing empirical models in reproducing of μ_{ef} were examined. Among different models, the Offord and Cifuentes et al. models showed a reasonable correlation.

Although the Offord model is superior to others, this model does not account for several factors that affect peptide mobility. Our results confirm the conclusion that μ_{ef} cannot be successfully predicted for all categories of peptides relying only on the two parameters of charge and size. As a second step, a MLR model was generated by considering different parameters that may affect μ_{ef}. Appearance of the three parameters of steric constant, molar refractivity and Offord charge/mass parameter in this model indicates the importance of the steric and dispersion interactions in the electromigration mechanism. Examination of the MLR results reveals no systematic deviation with respect to the experimental values. A notable exception to this observation is peptides containing highly charged amino acids of arginine, histidine and lysine. As a third step, a nonlinear BP-ANN model was generated. The 3-4-1 ANN model with the three descriptors appearing in the MLR model as its inputs showed a remarkable improvement for highly charged peptides. The weakness of the Offord model in reproducing of the mobility for highly charged peptides has been attributed to the inaccurate determination of the peptide charge or the absence of hydration. However, all previous researchers have assumed a linear relation between their models and the peptide mobility. Our results show that the structure - migration relationships for highly charged peptides follow nonlinear patterns. Despite the simplicity of our ANN model the average accuracy achieved was ~2 %. The capability of this method in prediction of electrophoretic mobility of peptides adds another dimension of information for peptide mapping. In the derived models, we only considered the effect of peptides amino acids composition on electrophoretic mobility. Work is underway to incorporate newly developed set of descriptors to account for the amino acids sequence in peptides. We are also investigating the present methodologies

for much larger peptides. The preliminary results, using literature data, have been quite promising, which indicate the robustness of the models described in this work [39].

References

[1] (a) Kasicka, V., Electrophoresis 2003, 24, 4013-4046; (b) Simo, C., Soto-Yarritu, P. L., Cifuentes, A., Electrophoresis 2002, 23, 2288-2295; (c) Simo, C., Cifuentes, A., Electrophoresis 2003, 24, 834-842.

[2] Grossman, P. D., Colburn, J. C., Lauer, H. H., Anal. Biochem. 1989, 179(1), 28-33.

[3] Offord, R. E., Nature (London) 1966, 211, 591-593.

[4] Compton, B. J., J. Chromatogr. 1991, 559, 357-366.

[5] Adamson, N. J., Reynolds, E. C., J. Chromatogr. B 1997, 699, 133-147.

[6] Messana, I., Rossetti, D. V., Cassiano, L., Misiti, F., Giardina, B., Castagnola, M., J. Chromatogr. B 1997, 699, 149-171.

[7] Kasicka, V., Electrophoresis 1999, 20, 3084-3105.

[8] Cifuentes, A., Poppe, H., Electrophoresis 1997, 18, 2362-2376.

[9] Grossman, P. D., Soane, D. S., Anal. Chem. 1990, 62, 1592-1596.

[10] Cross, R. F., Cao, J., J. Chromatogr. A 1997, 786, 171-180.

[11] Jokl, V., J. Chromatogr. 1964, 13, 451-458.

[12] Cifuentes, A., Poppe, H., J. Chromatogr. A 1994, 680, 321-340.

[13] Janini, G. M., Metral, C. J., Issaq, H. J., Muschik, G. M., J. Chromatogr. A 1999, 848, 417-433.

[14] Skoog, B., Wichman, A., Trends. Anal. Chem. 1986, 5(4), 82-83.

[15] Rickard, E. C., Strohl, M. M., Nielsen, R. G., Anal. Biochem. 1991, 197(1), 197-207.

[16] Haykin, S., Neural Network, Prentice Hall, Englewood Cliffs, NJ, 1994.

[17] Zupan, J., Gasteiger, J., Neural Networks in Chemistry and Drug Design, Wiley-VCH, Weinheim, 1999.

[18] Bose, N. K., Liang, P., Neural Network, Fundamentals, McGraw-Hill, New York, 1996.

[19] Anker, S. L., Jurs, P. C., Anal. Chem. 1992, 64, 1157-1164.

[20] Beal, M. T., Hagan, H. B., Demuth, M., Neural Network Design, PWN, Boston,

1996.

[21] Zupan, J., Gasteiger, J., Neural Networks for Chemists: An introduction, VCH, Weinheim, 1993.

[22] Hopke, P. K., Song, X., Anal. Chim. Acta. 1997, 348, 375-388.

[23] Minitab Release 12, http://www.minitab.com

[24] Hancock, C. K., Meyers, E. A., Yager, B. J., J. Am. Soc. 1961, 83, 4211-4213.

[25] Padron, J. A., Carrasco, R., Pellon, R. F., J. Pharm. Pharmaceut. Sci. 2002, 5(3), 258-265.

[26] Hansch, C., Leo, A., Hoekman, D., (1995) Exploring QSAR. Hydrophobic, Electronic and Steric constants, ACS Professional Reference Book, ACS, Washington D.C.

[27] Jalali-Heravi, M., Kyani A., J. Chem. Inf. Comput. Sci. 2004, 44, 1328-1335.

[28] Jalali-Heravi, M., Garkani-Nejad, Z., J. Chromatogr. A 2001, 927, 211-218.

[29] Jalali-Heravi, M., Fatemi, M. H., Anal. Chim. Acta. 2000, 415, 95-103.

[30] Jalali-Heravi, M., Parastar, F., J. Chromatogr. A 2000, 903, 145-154.

[31] Hilser Jr, V. J., Worosila, G. D., Rudnick, S. E., J. Chromatogr. 1993, 630, 329-336.

[32] Gaus, H. J., Beck-Sickinger, A. G., Bayer, E., Anal. Chem. 1993, 65, 1399-1405.

[33] Chen, N., Wang, L., Zhang, Y. K., Chromatographia 1993, 37, 429-432.

[34] Survay, M. A., Goodall, D. M., Wren, S. A. C., Rowe, R. C., J. Chromatogr. 1993, 636, 81-86.

[35] Baczek, T., Bucinski, A., Ivanov, A. R., Kaliszan, R., Anal. Chem. 2004, 76, 1726-1732.

[36] Sak, K., Karelson, M., Jarv, J., Bioorganic Chem. 1999, 27, 434-442.

[37] Taft Jr, R. W., Steric Effects in Organic Chemistry, Newman, M. S. ed., John Wiley and Sons, Inc., New York, 1956.

[38] Janini, G. M., Metral, C. J., Issaq, H. J., J. Chromatogr. A 2001, 924, 291-306.

[39] Jalali-Heravi, M., Shen, Y., and Khaledi, M.G., Manuscript in preparation.

Table 1. The experimental, MLR and ANN calculated values of electrophoretic mobility ($cm^2\ v^{-1}s^{-1}$) for model peptides together with the values of descriptors appearing in the models

		Descriptors			Mobility				
No.	Peptide sequence	QM	Es'c	MR	Experimental	MLR		ANN	
Training set					μ_{ef} x 10^5	μ_{ef} x 10^5	% Deviation	μ_{ef} x 10^5	% Deviation
1	AF	0.0218	-0.700	35.700	30.706	30.147	1.82	29.049	5.40
2	A_LY_L	0.0209	0.700	37.500	29.368	31.009	-5.59	28.581	2.68
3	AY	0.0209	0.700	37.500	29.581	31.009	-4.83	28.743	2.83
4	D_LF_L	0.0173	-1.480	41.600	24.016	23.289	3.03	25.693	-6.98
5	DF	0.0173	-1.480	41.600	23.221	23.289	-0.29	25.737	-10.83
6	DW	0.0159	-1.440	51.400	22.379	21.889	2.19	23.819	-6.44
7	FA	0.0218	-0.700	35.700	33.135	30.147	9.02	29.087	12.22
8	F_LF_L	0.0181	-1.400	60.000	26.975	25.352	6.02	25.241	6.43
9	FF	0.0181	-1.400	60.000	27.908	25.352	9.16	25.538	8.49
10	FG	0.0227	-0.500	31.000	32.634	31.427	3.70	30.784	5.67
11	FI	0.0196	-2.310	49.600	30.097	25.465	15.39	28.831	4.21
12	FL	0.0196	-1.940	49.600	29.428	26.001	11.64	29.422	0.02
13	FM	0.0188	-1.530	53.100	27.798	25.690	7.59	29.086	-4.63
14	FV	0.0202	-1.790	45.000	29.665	26.914	9.27	31.331	-5.62
15	FW	0.0167	-1.360	69.800	24.649	24.056	2.41	27.405	-11.18
16	GW	0.0204	-0.460	40.800	30.013	28.842	3.90	31.404	-4.64
17	GY	0.0217	-0.500	32.800	27.063	30.130	-11.34	32.247	-19.16
18	HW	0.0375	-1.320	63.600	45.709	51.824	-13.38	48.136	-5.31
19	IF	0.0196	-2.310	49.600	28.114	25.465	9.42	28.749	-2.26
20	IW	0.0179	-2.270	59.400	27.850	23.804	14.53	25.537	8.31
21	K_LF_L	0.0415	-1.320	55.100	49.893	56.776	-13.80	48.045	3.70
22	LF	0.0196	-1.940	49.600	28.115	26.001	7.52	26.739	4.90
23	L_LW_L	0.0179	-1.900	59.400	27.678	24.340	12.06	24.455	11.64
24	PW	0.0185	-0.660	53.800	30.439	26.706	12.26	25.562	16.02
25	RW	0.0362	-1.280	69.900	47.713	50.398	-5.63	46.057	3.47
26	VF	0.0202	-1.790	45.000	29.135	26.914	7.62	28.783	1.21
27	VY	0.0195	-1.790	46.800	28.535	25.956	9.04	28.729	-0.68
28	WA	0.0197	-0.660	45.500	30.077	27.845	7.42	29.420	2.18
29	WE	0.0171	-1.280	56.000	25.778	24.015	6.84	27.228	-5.63
30	WF	0.0167	-1.360	69.800	26.829	24.056	10.34	27.033	-0.76
31	WG	0.0204	-0.460	40.800	30.322	28.842	4.88	31.468	-3.78
32	WL	0.0179	-1.900	59.400	28.017	24.340	13.12	28.819	-2.86
33	WM	0.0173	-1.490	62.900	25.261	24.235	4.06	28.256	-11.85
34	WP	0.0185	-0.660	53.800	28.436	26.706	6.08	29.584	-4.04
35	WR	0.0362	-1.280	69.900	46.964	50.398	-7.31	48.417	-3.09
36	WS	0.0190	-0.940	51.600	28.416	26.762	5.82	28.998	-2.05
37	WV	0.0185	-1.750	54.800	28.681	25.067	12.60	27.826	2.98
38	WW	0.0156	-1.320	79.600	27.448	23.073	15.94	24.354	11.27

39	WY	0.0163	-1.360	71.600	25.686	23.486	8.56	25.015	2.61
40	YA	0.0209	-0.700	37.500	30.774	28.981	5.83	30.134	2.08
41	YG	0.0217	-0.500	32.800	30.404	30.130	0.90	30.948	-1.79
42	YI	0.0188	-2.310	51.400	28.884	24.591	14.86	28.415	1.62
43	YL	0.0188	-1.940	51.400	27.192	25.127	7.59	28.394	-4.42
44	YV	0.0195	-1.790	46.800	29.069	25.956	10.71	29.218	-0.51
45	YW	0.0163	-1.360	71.600	25.656	23.486	8.46	25.418	0.93
46	YY	0.0170	-1.400	63.600	19.869	23.996	-20.77	26.167	-31.69
47	FFF	0.0140	-2.100	90.000	24.028	20.296	15.53	22.039	8.28
48	FGG	0.0195	-0.300	32.000	28.754	27.441	4.57	28.175	2.01
49	$G_LF_LL_L$	0.0173	-1.740	50.600	27.495	23.262	15.40	25.433	7.50
50	GFL	0.0173	-1.740	50.600	27.014	23.262	13.89	25.335	6.22
51	GGF	0.0195	-0.300	32.000	28.204	27.441	2.71	27.828	1.33
52	KYK	0.0492	-1.940	82.000	56.533	67.503	-19.40	54.299	3.95
53	MLF	0.0151	-2.770	72.700	23.171	19.970	13.81	22.536	2.74
54	WGG	0.0179	-0.260	41.800	27.808	25.792	7.25	26.736	3.85
55	WWW	0.0120	-1.980	119.400	22.711	19.285	15.08	21.504	5.31
56	YAG	0.0182	-0.500	38.500	27.362	25.745	5.91	27.594	-0.85
57	YGG	0.0188	-0.300	33.800	27.687	26.572	4.03	28.470	-2.83
58	YPF	0.0147	-1.400	75.800	22.280	21.597	3.07	23.811	-6.87
59	YYL	0.0140	-2.640	83.200	20.727	19.231	7.22	22.392	-8.03
60	YYY	0.0131	-2.100	95.400	21.390	19.357	9.51	21.852	-2.16
61	FFFF	0.0116	-2.800	120.000	19.813	17.587	11.23	20.800	-4.99
62	FGFG	0.0147	-1.000	62.000	23.711	21.459	9.50	23.732	-0.09
63	GGFL	0.0156	-1.540	51.600	25.440	21.288	16.32	24.565	3.44
64	RFDS	0.0268	-2.380	83.500	38.294	36.849	3.77	37.611	1.78
65	WGGY	0.0136	-0.960	73.600	21.691	20.553	5.24	22.372	-3.14
66	YGGF	0.0144	-1.000	63.800	20.935	21.068	-0.64	23.089	-10.29
67	TRSAW	0.0252	-2.090	99.200	36.698	35.894	2.19	36.427	0.74
68	YGGFM	0.0121	-1.830	86.900	20.091	17.941	10.70	20.148	-0.28
69	YGGWL	0.0118	-2.200	93.200	19.891	17.333	12.86	19.622	1.35
70	RGPFPI	0.0236	-2.730	108.700	34.937	33.218	4.92	33.352	4.54
71	RRPYIL	0.0324	-4.790	145.200	44.047	43.967	0.18	45.280	-2.80
72	RYLGYL	0.0216	-4.300	133.900	34.445	29.490	14.39	29.661	13.89
73	VEPIPY	0.0103	-4.020	110.600	16.976	13.537	20.26	18.378	-8.26
74	YSGFLT	0.0107	-3.250	106.000	16.728	14.996	10.36	19.229	-14.95
75	RPKPQQF	0.0304	-3.180	151.400	41.665	43.878	-5.31	45.868	-10.09
76	RVYIHPI	0.0305	-6.290	153.900	42.713	39.588	7.32	43.051	-0.79
77	RVYVHPF	0.0300	-4.860	159.700	42.533	41.349	2.78	43.480	-2.23
78	YGGFMRF	0.0200	-3.150	147.000	30.735	29.714	3.32	30.948	-0.69
79	DRVYIHPF	0.0266	-6.160	175.900	37.065	35.685	3.72	37.945	-2.37
80	NRVYVHPF	0.0278	-5.640	174.200	39.989	37.900	5.22	39.148	2.10
81	PPGFSPFR	0.0196	-2.100	144.900	30.685	30.590	0.31	30.831	-0.47
82	RPPGFSPFR	0.0273	-2.720	175.000	39.746	41.484	-4.37	38.300	3.64
83	YLEPGPVTA	0.0085	-3.980	129.100	13.762	12.176	11.52	17.364	-26.18

84	DRVYIHPFHL	0.0315	-8.060	219.300	40.920	41.645	-1.77	40.584	0.82
85	SYSMEHFRWG	0.0237	-5.150	219.400	35.601	35.405	0.55	32.963	7.41
86	RPKPQQFFGLM	0.0232	-5.750	225.100	32.614	34.148	-4.70	33.191	-1.77
87	CGYGPKKKRKVGG	0.0471	-4.090	195.300	53.565	67.268	-25.58	52.802	1.42
88	ELYENKPRRPFIL	0.0270	-9.370	280.800	33.525	36.832	-9.87	32.781	2.22
89	AGCKNFFWKTFTSC	0.0204	-5.920	236.600	29.038	30.652	-5.56	31.447	-8.29
90	DRVYIHPFHLVIHN	0.0325	-12.200	292.200	29.238	40.701	-39.21	29.564	-1.12
Prediction set									
91	EW	0.0171	-1.280	56.000	26.854	24.015	10.57	24.334	9.38
92	HY	0.0393	-1.360	55.600	43.268	53.779	-24.29	46.540	-7.56
93	KF	0.0415	-1.320	55.100	49.966	56.776	-13.63	48.008	3.92
94	LW	0.0179	-1.900	59.400	27.931	24.340	12.86	24.877	10.94
95	VW	0.0185	-1.750	54.800	29.979	25.067	16.39	25.636	14.49
96	WD	0.0159	-1.440	51.400	25.063	21.889	12.66	23.538	6.08
97	FGGF	0.0147	-1.000	62.000	22.161	21.459	3.17	22.246	-0.38
98	GGFM	0.0151	-1.130	55.100	24.086	21.438	10.99	22.766	5.48
99	FFFFF	0.0101	-3.500	150.000	17.927	15.954	11.00	19.511	-8.84
100	WGGGY	0.0126	-0.760	74.600	19.263	19.578	-1.64	20.621	-7.05
101	WHWLQL	0.0199	-5.080	161.700	30.945	27.548	10.98	27.803	10.15
102	DRVYIHP	0.0294	-5.460	145.900	38.118	39.018	-2.36	39.909	-4.70
103	MEHFRWG	0.0290	-3.890	164.000	33.569	41.560	-23.81	40.793	-21.52
104	YMEHFRW	0.0270	-4.790	194.800	39.519	39.160	0.91	37.696	4.61
105	YPFVEPI	0.0091	-4.720	140.600	15.574	12.395	20.41	17.571	-12.82
106	ASTTTNYT	0.0092	-3.880	111.000	14.376	12.344	14.14	17.329	-20.54
107	RPGFSPFR	0.0291	-2.720	161.000	41.648	43.215	-3.76	41.842	-0.47
108	IARRHPYFL	0.0345	-6.150	204.700	46.710	47.728	-2.18	45.155	3.33
109	YRPPGFSPFR	0.0248	-3.420	206.800	36.381	38.711	-6.41	33.719	7.31
110	ELYENKPRRPYIL	0.0269	-9.370	282.600	29.361	36.692	-24.97	32.712	-11.41
Validation set									
111	AW	0.0197	-0.660	45.500	30.097	27.845	7.48	27.026	10.21
112	GF	0.0227	-0.500	31.000	31.713	31.427	0.90	29.800	6.03
113	IY	0.0188	-2.310	51.400	26.571	24.591	7.45	26.290	1.06
114	KW	0.0382	-1.280	64.900	48.799	52.852	-8.31	46.882	3.93
115	LY	0.0188	-1.940	51.400	22.889	25.127	-9.78	26.135	-14.18
116	MW	0.0173	-1.490	62.900	28.476	24.235	14.89	24.246	14.86
117	WGY	0.0148	-1.160	72.600	23.958	21.817	8.94	22.042	8.00
118	FLEEI	0.0108	-4.790	101.600	18.159	12.734	29.87	17.855	1.67
119	RKDVY	0.0355	-3.810	113.600	45.995	47.926	-4.20	48.682	-5.84
120	YAGFL	0.0121	-2.440	88.100	20.664	17.193	16.80	19.287	6.67
121	PPGFSP	0.0117	-0.780	84.800	20.266	18.866	6.91	20.010	1.26
122	RPPGFSP	0.0221	-1.400	114.900	34.114	33.434	2.00	32.738	4.03
123	WQPPRARI	0.0279	-4.130	172.400	40.150	40.192	-0.10	39.109	2.59
124	DRVYVHPFHL	0.0317	-7.540	214.700	41.955	42.478	-1.25	42.003	-0.12
125	ELYENKPRRPY	0.0296	-6.520	243.400	31.971	42.507	-32.96	38.266	-19.69

Table 2. Correlation equations for different empirical models of peptide mobility

Model	Equation	R^2
$Q/M^{2/3}$	$\mu_{ef} = 1378.19\ (Q/M^{2/3})$	0.878
$\text{Log}\ (1+0.297\ Q)/M^{0.411}$	$\mu_{ef} = 2743.66 \log (1+0.297\ Q)/M^{0.411}$	0.861
$Q/M^{1/2}$	$\mu_{ef} = 476.18\ (Q/M^{1/2})$	0.781
$Q/M^{1/3}$	$\mu_{ef} = 158.68\ (Q/M^{1/3})$	0.627

Table 3. Specifications of the best selected MLR model

Descriptor	Notation	Coefficient	Mean effect[a]
Charge to size ratio	QM	1347.04(±31.51)	28.102
Corrected steric substituent constant	Es'c	1.4476(±0.4161)	-3.252
Molecualr refractivity	MR	0.04979(±0.01466)	4.266
Constant		0.0	

[a] The mean effect of a descriptor is the product of its mean and the regression coefficient in the MLR model.

Table 4. Architecture and specifications of the generated ANN

Number of nodes in the input layer	3
Number of nodes in the hidden layer	4
Number of nodes in the output layer	1
Number of iterations	5000
Learning rate	0.5
Momentum	0.5

Table 5. Comparison of the statistics for the MLR and ANN models

Model	$R^2_{Training}$	R^2_{Test}	$R^2_{Validation}$	$SE_{Training}$	SE_{Test}	$SE_{Validation}$	F
MLR	0.8892	0.8983	0.8953	2.777	3.739	3.522	1111
ANN	0.9385	0.9206	0.9302	1.865	2.933	2.753	1441

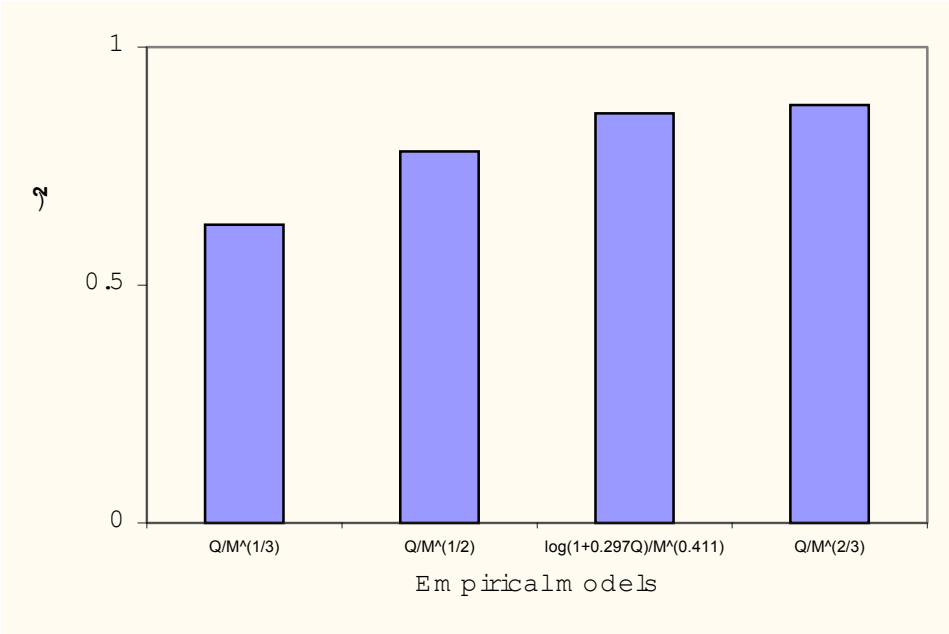

Fig.1a Correlation between the peptides mobilities and the values of charge-to-size ratios obtained using different empirical models.

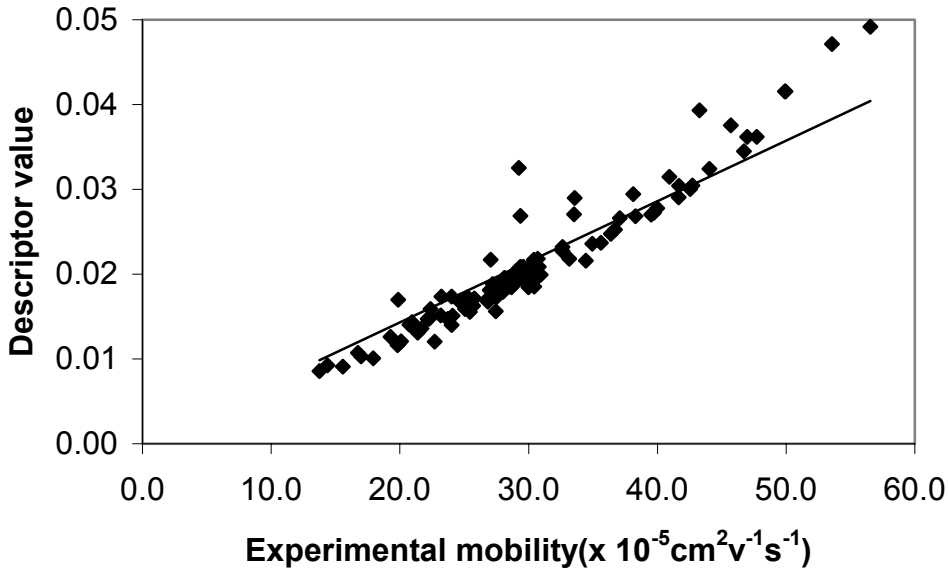

Fig.1b Plot of the values of charge-to-size ratios obtained using Offord model versus peptide mobilities

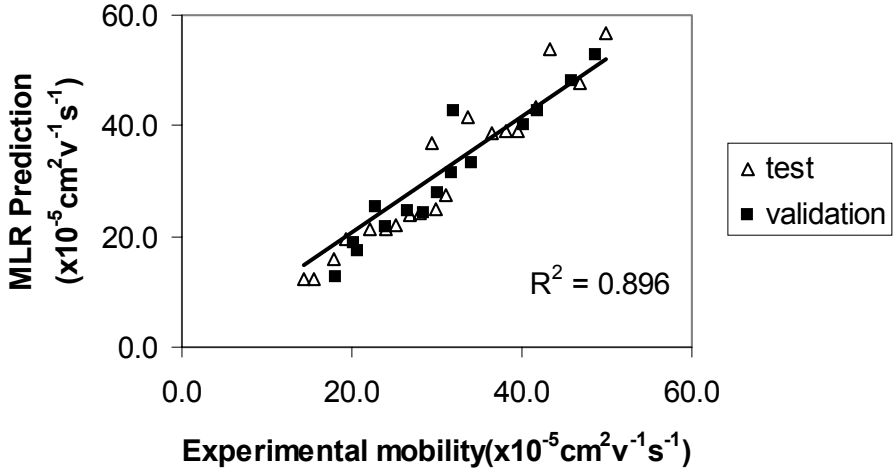

Fig.2 Plot of the MLR calculated electrophoretic mobilities against the experimental values for the test an validation sets.

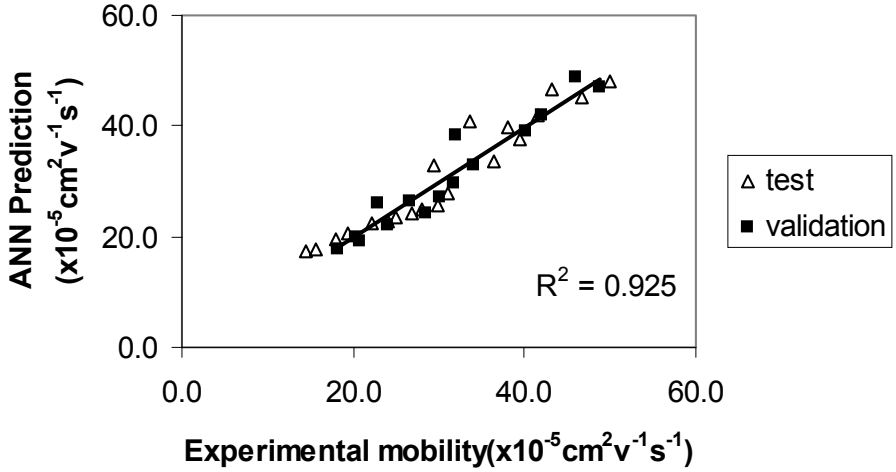

Fig.3 Plot of the ANN calculated electrophoretic mobilities against the experimental values for the test and validation sets.

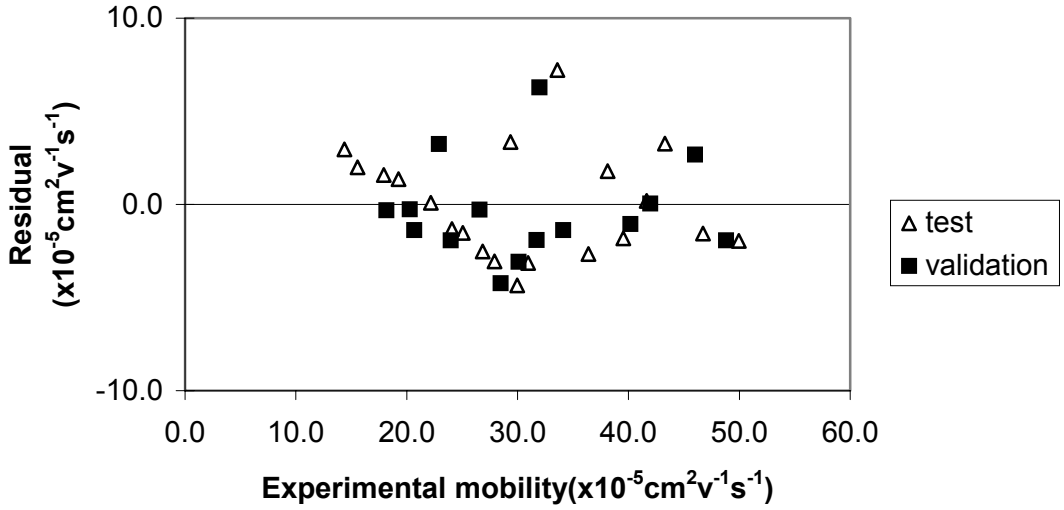

Fig.4 Plot of residuals against the experimental values of peptide mobility for the test and validation sets.

Chapter 3. Artificial neural network modeling of peptide mobility and peptide mapping in capillary zone electrophoresis

Keywords: Capillary Electrophoresis, Peptide Mapping, QSMR

1. Abstract

Recently, we had developed an artificial neural network model, which was able to predict accurately the electrophoretic mobilities of relatively small peptides. To examine the robustness of this model, a 3-3-1 BP-ANN model is developed using the same inputs as the previous one, such as Offord's charge over mass term ($Q/M^{2/3}$), corrected steric substituent constant ($E_{s,c}$) and molar refractivity (MR). This examination relied on a data set consisting of 102 peptides ranging in size from 2 to 42 amino acid residues, including highly charged and hydrophobic ones. The results of this model are compared with those obtained using multiple linear regressions (MLR) model developed in this work and the multi-variable model released by Janini et al. Superiority of the BP-ANN model over the MLR indicates the non-linear characteristics of the electrophoretic mobility. The present model exhibits robustness when predicting electrophoretic mobilities of a diverse data set obtained by CZE in different experimental conditions, unlike the multi-variable model. To explore the utility of the ANN model in simulation of the CZE peptide maps, the profiles for the endoproteinase digests of melittin, glucagon and horse cytochrome C is studied in the present work.

2. Introduction

Peptides are protein metabolites that reveal overall system function. Peptide mapping is a technique for analyzing proteins by separating and detecting the mixture of fragments (peptides) generated when a protein is broken up with chemicals or enzymes. The resulting peptide maps then serve as "fingerprints" that can be applied to rapid

protein identification and the detection of mutation and other modifications that may degrade protein performance.

Capillary liquid chromatography (combined with MS/MS) is currently one of the most commonly used techniques for peptide mapping [1]. While this method provides excellent resolution, it is often slow and generally consumes relatively large quantities of peptides. Capillary zone electrophoresis (CZE) has received considerable attention as peptide mapping method because of its high speed and high resolution for peptide analysis and also its small sample size requirement. Among other advantages, CZE can be multiplexed and automated so that many experiments can be performed in parallel. In this technique, peptide can be identified based on the time it takes to migrate through CZE system. However, electrophoretic peptide profiles obtained are sometimes very complex owing to the complicated nature of the samples. Generally, CE separation of peptides is more successful than proteins because the smaller molecules tend to interact less compared with the capillary wall.

The key parameter for separation of peptides, especially in low ionic strength buffers, is their electrophoretic mobilities. This parameter can be converted to migration time and a CZE electropherogram can be simulated using a Gaussian function. Therefore, calculation/prediction of this parameter is very useful in peptide mapping studies.

Several empirical models, based on Stoke's law, have been developed for the prediction of electrophoretic mobilities [2-8]. Investigation of these models reveals two reasons behind their relatively poor predictive ability: (1) ignoring non-linear characteristics of electrophoretic mobilty and assuming it as an inherent linear

phenomenon; (2) inaccuracy of pK_a values for the ionizable functional group in the amino acids and therefore, peptide charges. These models also differ by their dependence on molar mass.

Janini and coworkers 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 [9]. Based on this data set, they concluded that the Offord model gives the best overall mobility, but it fails when applied to hydrophobic and highly charged peptides. These researchers also showed that peptide electrophoretic mobility cannot be successfully predicted with reasonable degree of accuracy for all different categories of peptides by relying on two-parameter models, namely charge (Q) and size dependence (N or M) [9].

In two other efforts, Janini et al. have developed a multi-variable computer model, based on a closest-neighbor algorithm for the prediction of the electrophoretic mobilities of peptides [10,11]. Their model is based on assumption that the electrophoretic mobilities can be calculated by a product of three functions representing peptide charge, length and width [11]. Although, these researchers have obtained excellent results, but the main drawback of their model is that all functions appearing in the model are data dependent. In other words, their model is based on a closest-neighbor algorithm and is not robust.

We have recently developed an artificial neural network model to explore: (1) the linear/non-linear characteristics of the motion of a peptide in a capillary column under the influence of an electric field; (2) to identify the best model (among different existing models; and (3) to investigate different factors affecting the motion of a peptide in an

electric field [12]. This work was based on a Quantitative Structure Mobility Relationships (QSMR) model using the Offord's charge over mass term ($Q/M^{2/3}$) as a hybrid descriptor combined with the corrected steric substituent constant (Es,c) and molar refractivity (MR) descriptors. The last two parameters account for the steric effects and the bulkiness of the amino acids side chains, respectively. The ANN model showed a significant improvement in the predictive ability over the empirical models and the multiple linear regression (MLR) treatment [12]. This improvement was especially pronounced for the peptides of higher charges that contain basic amino acids arginine, histidine and lysine. This model was based on a data set consisted of the electrophoretic mobilities for 125 peptides ranging in size between 2 and 14 amino acids. These electrophoretic mobilities were obtained using capillary zone electrophoresis (CZE), at 37 °C, with a 50 mM sodium phosphate buffer (pH 2.5).

The main aim for developing a theoretical model is producing a robust model to accurately predict the property of interest. In order to prove the robustness of our ANN model, we decided to apply it to a more diverse data set, preferably with different experimental conditions. Therefore, in the present work, a data set released by Janini and coworkers was chosen [11]. This data set consists of 102 peptides that varies in charge from 0.65 to 16 and in size between 2 and 42 amino acid residues. The main reason behind choosing this data set was its diversity compared with our previous basis set for generating the ANN model [12]. Also, Janini et al. have obtained their electrophoretic mobilities in a different experimental condition namely at 22 °C and a column coated with a dense layer of 10% polyacrylamide [11]. Whereas, our ANN model was relied on the electrophoretic mobilities obtained at 37 °C using a bare silica column [12]. However,

the ability of the ANN model to predict accurately the electrophoretic mobilities of a diverse data set with a different experimental condition exhibits its robustness.

The present results are quite important from point of accurate calculation/prediction of the electrophoretic mobility of peptides. By having a robust model for predicting the electrophoretic mobility of peptides, one can theoretically simulate the electropherograms, which are the main key for peptide mapping in protein studies.

3. Experimental

3.1 Data set

Development of the multiple linear regression (MLR) and artificial neural networks (ANN) in the present work relies on a data set taken from reference [11]. This data set (Table 1) consists of 102 peptides ranging in size from 2 to 42 amino acid residues, which varies in charge from 0.65 to 16. The data set includes 18 dipeptides, 32 peptides with five or less amino acid residues and 72 peptides with six or more amino acid residues. The electrophoretic mobilities of these peptides were obtained using capillary zone electrophoresis (CZE), at 22 °C with a 50 mM phosphate buffer, at pH 2.5 [11]. The mobilities were accurately measured at these fixed experimental conditions using 10% polyacrylamide-coated columns [11]. Coating with polyacrlamide provided stability and migration time reproducibility throughout the experiments to within 1% RSD [11]. However, in order to determine the uncertainty of the experimental electrophoretic mobilities, Janini et al. have measured the mobilities of six peptides and the reference standard using two independent columns with different buffer preparations in separate days. They obtained an average RSD of 2.34%, which represents the uncertainty of the

experimental values of electrophoretic mobilities. The electrophoretic mobilities of peptides in data set fall in the range of 4.73-33.03 (10^{-5} cm^2 s^{-1} V^{-1}) for FIGITEAAANLVPMVATV and KKKKK peptides, respectively. The data set was randomly divided into three groups of training, test and validation sets consisting of 70, 20 and 12 peptides, respectively. The training set was used for the model generation. However, the test set plays a different role in the cases of the MLR and the ANN models. For the ANN model, this set was used for early stopping to optimize learning iteration size and avoid overtraining. On the other hand in the case of the MLR model, the test set together with the validation set was used to evaluate the generated model. As can be seen from Table 1, the test and validation sets were chosen in a way that adequately represents the training set.

3.2 Regression analysis

The main aim of the present work was to investigate the robustness of our previous Quantitative Structure-Mobility Relationships(QSMR) model [12]. Therefore, as first step for developing the regression model, the three descriptors appearing in our previous QSMR models were chosen as the most suitable parameters contributing to the motion of a peptide in an electric field [12]. These descriptors consisted of $Q/M^{2/3}$ (Offord empirical model) as a hybrid parameter, corrected steric substituent constant (E s,c) and molar refractivity (MR). These descriptors were chosen from a large set of parameters such as effective net charge, molar mass, number of amino acid residue, average residue mass, molecular volume, surface area, hydrophobicity, isoelectric point value, strain parameter, Z-scale and alpha-helix content [13,14]. A detailed description of the stepwise multiple linear regression procedure used for choosing these descriptors is given in our previous

paper [12]. The calculated values of these parameters for all peptides in the training, test and validation sets are given in Table 1. The best MLR model is one that has high correlation coefficient (R) and F-values, low standard deviation and high ability for prediction. The specifications for the best MLR model (Eq. 1) are presented in Table 2.

$$\mu = p\frac{Q}{M^{2/3}} + e\sum E_s + m\sum MR \qquad (1)$$

3.3 Artificial Neural Networks Model

A detailed description of theory behind artificial neural networks has been adequately described elsewhere [15-18]. A three-layer back-propagation network with a sigmoidal transfer function was designed in the present work. This network is written in C++ in our laboratory. The three descriptors appearing in our previous QSMR model were used as input parameters for generation of the network. The signals from the output layer represent the electrophoretic mobilities of the peptides. Such an ANN may be designed as 3-y-1 net to indicate the number of nodes in input, hidden and output layers, respectively. Generally, the neural network methodology has several empirically determined parameters. These include: (1) when to stop training (i.e. the number of iterations or the convergence criterion), (2) the number of hidden nodes, and (3) learning rate and momentum terms. The values of constructed ANNs parameters were optimized with the procedure that was reported in our previous works [19-21]. The initial weights were chosen randomly. The program is written in such a way that the randomized weights depend on the number of input, hidden and output nodes. Before training, the output and inputs (except for the values of the Offord model) were normalized between 0 and 1. To evaluate the performance of the ANN, standard error of calibration (SEC) and

standard error of prediction (SEP) were used [22]. The number of neurons of the hidden layer with the minimum value of SEC was selected as the optimum number. Learning rate and momentum were optimized in a similar way. We have used the validation set to examine the validity of the ANN model. Also, to further test the utility of the model, we have simulated the theoretical peptide maps of the digests of melittin, glucagon and horse cytochrome C polypeptide and proteins.

4. Results and Discussion

The main goal of the present study was developing a theoretical model that can be used as a tool for the simulation of the peptide maps. In other words, our objective was developing a robust model for predicting accurately the electrophoretic mobilities of all categories of peptides. Recently, we released a 3-4-1 BP-ANN model, which was able to predict the electrophoretic mobilities of relatively small peptides ranging in size between 2 and 14 amino acid residue [12]. This model showed a better predictive ability compared with all previous existing empirical models and the multiple linear regression treatment. To inspect the robustness of this model, one should choose a very diverse data set consisting of all categories of peptides, including highly charged and hydrophobic. Therefore, in the present work we have applied a similar strategy to predict the electrophoretic mobilities of a diverse peptide set obtained by Janini et al. using capillary zone electrophoresis [11].

4.1 Multiple Linear Regression analysis

The MLR calculated values of electrophoretic mobilities of all peptides are shown in Table 1. Also, the specifications for the selected MLR model are presented in Table 2.

It can be seen that the most important descriptor is the hybrid parameter of charge-to-size ratio (Offord model). This descriptor shows a mean effect of 9.845, which is the largest among the parameters appearing in the model. This is in agreement with our previous work and confirms the conclusion reached by other researchers [3, 10-12]. However, a major problem with this parameter is the accuracy of the calculated charge, which is controversial in the literature [23-25]. The most common method for calculating peptide charges has been released by Henderson-Hasselbach [26]. According to this method, the net charge of a peptide at pH 2.5 can be calculated as the algebraic sum of all charged amino acid residues and carboxy- and amino-terminals. Each arginine (R), histidine (H), lysine (K) and N-terminal contribute a charge of +1; while each aspartic acid (D) (pKa 3.5), glutamic acid (E) (pKa 4.5) and the C-terminal (pKa 3.2) contributes a charge of –0.091, -0.01 and –0.166, respectively. Although Rickard and coworker have claimed that at pH 2.5 a good agreement is expected between the calculated and actual charge, this is not the case for hydrophobic or highly charged peptide [24]. At this pH, for hydrophobic peptides variations in the ionization constants of D and E due to their local environment might be significant, while for highly charged peptides the calculated charges might deviate from the actual one due to mutual electrostatic interactions of charged groups in proximity of each other. However, Cifuentes et al. have presented a model for predicting pKa values of peptides considering the mutual electrostatic interaction of charged groups, but their model is empirical in nature and time consuming [25]. Since the charge-to-mass ratio parameter plays the major role in the mechanism of electrophoretic mobilities, the MLR calculated mobilities for hydrophobic and highly charged peptides might deviate considerably from the experimental values. Inspection of Table 1 shows that this is true

for most of the highly charged or hydrophobic peptides. The peptides 48, 68 and 70 with charges of 11.83, 15.83 and 12.72, respectively are the most highly charged peptides in Table 1. The MLR calculated values for these peptides show a deviation of –17.51, -18.21 and –6.24%, respectively. On the other hand, among the peptides listed in Table 1 peptides 2, 30 and 94 show the lowest charges of 0.65, 0.73 and 0.74, respectively. The MLR calculated mobilities of these peptides also show high deviations of –26.33, -43.70 and –48.99%, respectively. Jean Luc et al. have released a hydrophobic parameter, which can be considered as a measure for the hydrophobicity of peptides [27]. They have obtained this parameter from the partitioning of N-acetyl-amino acid amides in octanol/water system [27]. Based on this parameter, peptides 66, 90 and 102 of Table 1 are the most hydrophobic ones with large deviations of –66.33, 4.92 and 12.47%, respectively in their MLR calculated values. Fig. 1 shows the correlation between the MLR calculated and the experimental values of the electrophoretic mobilties of peptides included in the test and validation sets. The correlation of R^2 = 0.906 indicates a reasonable agreement between these values and also demonstrates some improvements over those obtained using the Offord model. These improvements can be attributed to the inclusion of the parameters of corrected steric constant and molecular refractivity into the MLR model. However, this model shows weakness in predicting the electrophoretic mobilities of highly charged and hydrophobic peptides.

4.2 Artificial Neural Network Analysis

The most important advantage of the artificial neural networks over regression analyses is their ability to allow for the flexible mapping of the selected features by manipulating their functional dependence implicitly. Developing networks and

comparing them with the MLR models provides us the opportunity to investigate the nonlinear characteristics of the eletophoretic mobilities of peptides. In order to have a meaningful comparison, the variables for the linear and nonlinear treatments should be the same. Therefore, the three descriptors appearing in the MLR model have been considered as inputs for generating the networks. After optimizing the parameters needed for constructing ANNs, a neural network with architecture of 3-3-1 was obtained, which its specifications are given in Table 3. We used the test set consisting of twenty peptides to optimize the learning iteration size and avoid overtraining. To evaluate the network, the electrophoretic mobilities of peptides included in the validation set were predicted and are shown in Table 1. Fig. 2 shows the plot of the ANN predicted versus the experimental values of the electrophoretic mobilities for the test and validation sets. A high value of 0.970 for the correlation between the ANN calculated and the experimental values of the electrophoretic mobilities, indicates the ability of this model in predicting the mobility. Comparison of this value with a correlation of 0.930 for the same peptides using linear regression with the same descriptors reveals the superiority of ANNs over the MLR model. Inspection of the results given in Table 1 shows that this improvement is mostly due to a better ability of the nonlinear model in predicting the electrophoretic mobilities of highly charged or hydrophobic peptides. A detailed discussion of this improvement is given in the next section. Fig. 3 shows a plot of the residuals of ANN predicted values of the electrophoretic mobilities against the experimental values. The propagation of the residuals on both sides of zero indicates that no systematic error exists in the development of the neural network.

To compare the chemometric methods of MLR and ANN in predicting the electrophoretic mobilities of peptides, some statistics for these models are summarized in Table 4. It can be seen from this table that R^2 values for the ANN model are considerably higher than those for the MLR model for all three sets: training, test and validation. Also, the ANN predicted values show much lower standard errors compared with those of the MLR model. The ANN model also reveals a higher value for the F-statistic. The superiority of the ANN model demonstrates the nonlinear characteristics for the electrophoretic mobility.

4.3 Prediction of highly charge and hydrophobic peptides

Despite the voluminous amount of attempt reported for calculating the electrophoretic mobilities of peptides [5-11, 25], there is still no robust model to be able to predict accurately this parameter for all categories of peptides, especially highly charged and hydrophobic peptides. A notable exception to the previous works are a multi-variable computer model presented by Janini et al. [10,11] and our recent ANN model [12]. Janini and coworkers, based on a data set consisted of the electrophoretic mobilities of 58 peptides that varied in size from 2 to 39 amino acids, examined the existing empirical models that correlates electrophoretic mobility with physical parameters. They reached to the conclusion that the charge-to-size parameter of Offord offers the best fit to their experimental data [9]. However, inspection of their results revealed a systematic deviation for small peptides with large positive charge, such as the lysine homologous with n=2-5 [9]. To address this deficiency, based on a basis set of 64 peptides, Janini et al. presented a multi-variable model that takes into account other physical properties that were neglected by the Offord model [10]. These researchers

using a purely phenomenological and empirical approach, assumed that the electrophoretic mobility can be presented by a product of four functions representing the length, n(N), average residue mass, w(W), charge, q(Q) and the position of the center of charge relative to the center of mass, c(CC) of peptides [10]. However, the success of their model was dependent upon the accurate measurement of the electrophoretic mobilities of a large number of peptides with a wide range of charge and molar mass [10]. Therefore, to improve the predictive ability, they extended their basis set from 64 to 102 peptides ranging in size between 2 and 42 amino acid residues [11]. Despite a significant improvement for all categories of peptides, including highly charged and hydrophobic, the multi-variable model suffers from the lack of robustness. Calculations of different functions of w(W), n(N) and q(Q) depends entirely on the data set and changing the basis set requires new equations for these functions. Also, this model is based on closest-neighbor algorithm that matches an unknown peptide to its closest neighbor in the basis set [11]. However, the ANN model has overcome the problem of robustness by incorporating three general descriptors as its inputs. To demonstrate the predictive ability of the ANN model, we have highlighted the results for some of the highly charged and hydrophobic peptides in Table 5. For the sake of comparison the Offord, multi-variable [11] and MLR calculated values for the electrophoretic mobilities together with some statistics are also included in this table. The data included in Table 5 represent peptides with different sizes and were chosen based on a charge larger than 4 and a pai parameter larger than 3. The results of this table clearly demonstrate a significant improvement for the multi-variable and ANN models compared with the MLR and the Offord models. Although there is no significant differences between the R^2 values of multi-variable and

ANN models, the later shows a lower standard error (SE) and relative standard deviation (RSD). Despite the improvements, for both models several inconsistencies exist between the sequences of the real and predicted electrophoretic mobilities of highly charged and hydrophobic peptides. A summary of these inconsistencies is shown in Fig. 4. Peptides 10, 19 and 12 show an incorrect order for the ANN model, while the order of electrophoretic mobilities for peptides 9, 19, 13 and 22 is not correct for the multi-variable model. Comparison of the results in Table 5 reveals superiority for ANN model over the MLR and the Offord model. The R^2 value of 0.990 for the ANN should be compared with a value of 0.910 for the MLR and Offord models. Also, SE and RSD values of 0.80 and 31.30%, respectively for the ANN model should be compared with 2.60, 3.67 and 36.73% and 46.43% for the MLR and Offord models, respectively. Large deviations of –17.51, -18.21 and –6.24% for the MLR calculated values of the most highly charged peptides 48, 68 and 70, respectively should be compared with the values of –1.34, 1.52 and –4.64% for their ANN calculated counterparts. Also, deviations of 4.92 and 12.4% for the MLR calculated electrophoretic mobilities of highly hydrophobic peptides of 90 and 102, respectively should be compared with the values of –3.41 and 0.16% for their ANN counterparts. A notable exception is the hydrophobic peptide FIGITEAAANLVPMVATV (Table 1) with a large deviation of –66.33 and –88.22% for the MLR and ANN calculated values, respectively. Although, we are uncertain about the origin of these deviations, they could be due to the experimental uncertainty.

4.4 Simulation of peptide maps of protein digests

The long-range goal of developing theoretical models, which can accurately predict the CZE parameters, such as retention time or electrophoretic mobility, is the

construction of a database of peptide maps. Reaching this goal means that one can easily identify unknown proteins by submitting the experimental maps to the database and searching for the closest match in terms of the migration times for the major peaks. To explore the utility of the ANN model in simulation of the CZE peptide maps, the profiles for the endoproteinase Lys-C digests of a peptide sequencing standard, melittin GIGAVLKVLTTGLPALISWIKRKRQQ, and two more complicated proteins namely glucagon and horse cytochrome C were studied in this work. Choosing these peptides were based on possibility of comparison of the ANN simulated maps with the experimental and multi-variable simulated ones. In order to simulate an electropherogram, first the ANN calculated electrophoretic mobilities were converted to migration times using the same values for the experimental parameters as reported by Janini et al. [9]. Values of 37 and 30 cm were used for the total length of the column and injector-to-detector length, respectively [9]. Also, a running voltage of 8 kV was used for the purpose of this conversion [9]. To simulate the peak of each theoretical fragment, it was assumed that the area for each peak is proportional to the number of peptide bonds. It is shown that at 200 nm, the absorbance of a peptide is largely attributed to the peptide bonds while the contribution of amino acids residues can be ignored [28,29]. For the sake of convenience, same peak width was used for each simulated peak in this work, which makes the peak height proportional to the number of peptide binds.

Fig. 5 demonstrates the experimental and ANN simulated maps for the endoproteinase Lys-C digest of the peptide sequencing standard of melittin. The correct migration order of peptides and corresponding retention times agrees fairly with the

experimental electropherogram. However, multi-variable results also show a correct order and more accurate retention times [11].

Next, we have considered glucagon, a polypeptide with 29 amino acid residues. This polypeptide can be used as a control for proteolytic digestion, sequencing and amino acid analysis [11]. Janini and coworkers digested this protein with endoproteinase Glu-C with characteristic cleavage at the C-terminal of aspartic acid (D) and glutamic acid (E) residues [11]. Therefore, after a complete digestion, four fragments are expected for this protein. These fragments are listed in Table 6. Also, the values for the three descriptors together with the calculated MLR and ANN values for the electrophretic mobilities of glucagon fragments are summarized in this table. The ANN calculated mobilities were converted to the migration times and simulated electrophoregram is shown in Fig. 6 For the purpose of comparison, the experimental and simulated electropherograms reported by Janini et al., are also shown in this figure. Inspection of Fig. 6 reveals an excellent matching between the line positions and a reasonable agreement between the relative heights of the experimental and simulated peaks. It seems that both models of multi-variable and ANN overestimates the mobility for the FVQWLMNT fragment and therefore, a smaller value for the corresponding migration time. Validity of this conclusion depends upon the accuracy of assignments of the peaks. However, one should consider the possibility of several shortcomings in experiment such as imperfect enzymatic digestion, impurity and autolysis of the endoproteinase.

Finally, we studied the relatively complex protein of horse cytochrome C with 104 amino acid residues. This protein also was digested with endoproteinase Lys-C with specificity of cleavage at the C-terminal of lysine residues [11]. The theoretical fragments

of this protein together with the values of the descriptors and corresponding MLR and ANN calculated mobilities are demonstrated in Table 6. Fig. 7 shows the experimental and ANN simulated electropherogram of cytochrome C. For the sake of comparison, the simulated electropherogram of this protein obtained by Janini et al. is also included in this table. Inspection of Fig. 7 Shows that the experimental electropherogram of Cytochrome C have a striking similarity to the ANN simulated electropheroram. Almost each simulated peak has a counterpart in the experimental electropherogram with a good agreement between their migration times. However, Janini et al. have obtained only 13 peaks for the 15 theoretical fragments. They believed that the pairs of fragments (17, 19), (5,7) and (10,11) co-migrated because of the closeness of their migration time. Our result confirms this conclusion except for the pairs of (5,7), which is replaced with the pairs of (7,8) in this work.

5. Conclusions

Recently, our research is focused on developing a multivariate model, which can be used as a tool in the construction of a database of peptide maps. As a long-range aim we have considered the specifications of simplicity, accuracy and robustness for the model in predicting the CZE parameters of peptides. Also to get this dream closer to reality, the model must be able to predict accurately the CZE parameters of peptides with identical amino acid residues but different sequences. Three steps were considered for reaching this goal: (1) Developing a model by using the descriptors as simple and general as possible and examining its ability in accurate prediction of the electrophoretic mobilities of peptides. We had already generated an artificial neural network model using

three simple descriptors of the charge-over-mass ratio of $Q/M^{2/3}$, corrected steric substituent constant and molar refractivity as its inputs [12]. Application of this model in calculating the electrophoretic mobilities of peptides ranging in size between 2 and 14 amino acid residues showed its predictive ability. (2) To investigate the robustness of the model, a more diverse data set obtained in different experimental conditions was needed. To fulfill this, a data set consisted of 102 peptides ranging in size from 2 to 42 amino acid residues including hydrophobic and highly charged ones was chosen in the present work [11]. The robustness of the neural network model was exhibited by accurate ANN calculated electrophoretic mobilities of all categories of peptides, obtained in different experimental conditions. Also, simulating the endoproteinase digests of melittin, glucagon and horse cytochrome C maps showed the utility of the model in simulation of the CZE peptide maps. (3) A research is under way in our group to explore the use of a new series of sequence-descriptors. The preliminary results are promising. Success in this step improves the capability of the model in simulating proteins maps consisted of isomeric peptides.

References

[1] (a) A.J. Link, J. Eng, D.M. Schieltz, E. Carmack, G.J. Mize, D.R. Moris, B.M. arvik, J.R. Yates, III. Nat. Biotechnol. 17 (1999) 767. (b) M.P. Wasburn, D. Wolters, J.R. Yates, III. Nat. Biotechnol. 19 (2001) 242.

[2] P.D. Grossman, J.C. Colburn, H.H. Lauer, Anal. Biochem. 179 (1989) 28.

[3] R.E. Offord, Nature (London) 211 (1966) 591.

[4] B.J. Compton, J. Chromatogr. 599 (1991) 357.

[5] N.J. Adamson, E.C. Reynolds, J. Chromatogr. B 699 (1997) 133.

[6] I. Messana, D.V. Rossetti, L. Cassiano, F, Misiti, B. Giardina, M. Castagnola, J. Chromatogr. B 699 (1997) 149.

[7] V. Kasicka, Electrophoresis 20 (1999) 3084.

[8] A. Cifuentes, H. Poppe, Electrophoresis 18 (1997) 2362.

[9] G.M. Janini, C.J. Metral, H.J. Issaq, G.M. Muschik, J. Chromatogr. A 848 (1999) 417.

[10] C.J. Metral, G.M. Janini, G.M. Muschik, H.J. Issaq, J. high RESOL. Chromatogr. 22 (1999) 373.

[11] G.M. Janini, C.J. Metral, H.J. Issaq, J. Chromatogr. A 924 (2001) 291.

[12] M. Jalali-Heravi, Y. Shen, M. Hassanisadi, M.G. Khaledi, Electrophoresis, Submitted.

[13] T. Baczek, A. Bucinski, A.R. Ivanov, R. Kaliszan, Anal. Chem. 76 (2004) 1726.

[14] K. Sak, M. Karelson, J. Jarv, Bioorganic Chem. 27 (1999) 434.

[15] J. Zupan, J. Gasteiger, Neural Networks in Chemistry and Drug Design, VCH: Weinheim, 1999.

[16] N.K. Bose, P. Liang, Neural Network, Fundamentals, McGraw Hill: New York, 1996.

[17] D.W. Patterson, Artificial Neural Networks: Theory and Applications. Simon and Schuster: New York, 1996.

[18] P.K. Hopke, X. Song, Anal. Chim. Acta 348 (1997) 375.

[19] M. Jalali-Heravi, F. Parastar, J. Chromatogr. A 903 (2000) 145.

[20] M. Jalali-Heravi, Z. Garkani-Nejad, J. Chromatogr. A 927 (2001) 211.

[21] M. Jalali-Heravi, M. Kyani, J. Chem. Inf. Comput. Sci. 44 (2004) 1328.

[22] T.B. Blank, S.T. Brown, Anal. Chem. 65 (1993) 3084.

[23] H.J. Issaq, G.M. Janini, I.Z. Atamna, G.M. Muschik, J. Lukso, J. Liq. Chromatogr. 15 (1992) 1129.

[24] E.C. Rickard, M.M. Strohl, R.G. Nielson, Anal. Biochem. 197 (1991) 197.

[25] A. Cifuentes, H. Poppe, J. Chromatogr. A 680 (1994) 321.

[26] B. Skoog, A. Wichman, Trends. Anal. Chem. 5 (1986) 82.

[27] F. Jean Luc, P. Vladimir, European J. Med. Chem. 18 (1983) 369.

[28] R.K. Scopes, Anal. Biochem. 59 (1974) 277.

[29] M. Herold, G.A. Ross, R. Grimm, D.N. Heiger, in: K.D. Altria (Ed.), Capillary Electrophoresis Guidebook, Principles, Operation, and Application, Methods in Molecular Biology, Vol. 52, Humana Press, Totowa, NJ, 1996, pp.285-308.

Table 1. Experimental and calculated values of electrophoretic mobilities using MLR and ANN models together with the values of the descriptors

No.	Peptide sequence	Descriptors[a]			Exp	MLR		ANN	
Training set		QM	$E_{s,c}$	MR	μ_{ef} x 10^5	μ_{ef} x 10^5	% Dev	μ_{ef} x 10^5	% Dev
1	AA	0.0283	0.00	11.3	18.77	17.85	4.91	19.84	-5.68
2	DD	0.0165	-1.56	23.2	10.31	13.02	-26.33	11.92	-15.58
3	EE	0.0192	-1.24	32.5	12.52	14.22	-13.55	13.58	-8.46
4	FA	0.0218	-0.70	35.7	14.86	15.40	-3.63	15.20	-2.29
5	FF	0.0181	-1.40	60.0	12.81	13.99	-9.25	13.06	-1.95
6	FG	0.0227	-0.50	31.0	15.16	15.76	-3.98	15.57	-2.69
7	FL	0.0196	-1.94	49.6	13.33	14.30	-7.26	13.51	-1.35
8	FV	0.0202	-1.79	45.0	13.90	14.56	-4.77	13.89	0.05
9	GG	0.0321	0.40	2.1	21.70	19.36	10.79	21.70	0.00
10	HG	0.0515	-0.46	24.8	27.04	26.67	1.36	29.35	-8.53
11	LL	0.0213	-2.48	39.2	14.55	14.73	-1.21	14.61	-0.43
12	PG	0.0269	0.20	15.0	18.43	17.43	5.41	18.83	-2.15
13	RK	0.0629	-1.24	55.1	32.00	31.02	3.05	30.98	3.17
14	VV	0.0231	-2.18	29.9	15.39	15.42	-0.20	16.17	-5.08
15	WW	0.0156	-1.32	79.6	11.05	13.23	-19.77	12.80	-15.84
16	YY	0.0170	-1.40	63.7	12.10	13.59	-12.34	13.29	-9.81
17	FFF	0.0140	-2.10	90.0	10.38	12.48	-20.28	11.92	-14.80
18	SSS	0.0195	-0.84	35.5	13.22	14.48	-9.55	14.34	-8.44
19	AAAA	0.0185	0.00	22.6	13.87	14.24	-2.64	13.72	1.10
20	ANSK	0.0328	-1.68	57.0	20.91	19.46	6.93	22.53	-7.76
21	HMTE	0.0283	-2.64	75.0	18.91	17.64	6.70	19.73	-4.34
22	PARR	0.0451	-1.24	79.7	27.65	24.46	11.55	27.43	0.78
23	KKKKK	0.0770	-3.10	125.3	33.03	36.45	-10.37	31.76	3.86
24	RPPGF	0.0266	-1.12	89.0	18.36	17.55	4.42	18.72	-1.95
25	YGGFL	0.0123	-2.24	83.5	9.75	11.76	-20.57	10.58	-8.49
26	GIGAVLK	0.0243	-4.16	86.9	15.50	15.76	-1.67	16.69	-7.66
27	AAGIGILTV	0.0096	-5.68	98.9	6.50	9.83	-51.20	8.78	-35.00
28	ACHGRDRRT	0.0453	-3.63	144.0	26.54	24.39	8.08	26.90	-1.37
29	AFLPWHRLF	0.0253	-5.82	212.5	16.55	16.73	-1.09	17.53	-5.93
30	MLDLQPETT	0.0071	-6.39	146.8	6.33	9.10	-43.70	8.36	-32.11
31	RPPGFSPFR	0.0273	-2.72	174.8	19.71	18.07	8.35	18.47	6.30
32	VLQELNVTV	0.0082	-8.30	145.7	6.97	8.93	-28.10	8.14	-16.80
33	VVRRYPHHE	0.0429	-6.06	199.6	27.38	23.24	15.10	25.55	6.70
34	YLSGADLNL	0.0076	-6.06	135.1	6.23	9.28	-49.00	8.39	-34.67
35	KLVVVGAAGV	0.0195	-5.82	117.8	14.10	13.72	2.67	13.32	5.55
36	KLVVVGADGV	0.0180	-6.60	123.7	13.13	12.96	1.29	12.52	4.67
37	ACLGRDRRTEE	0.0312	-5.45	172.3	20.97	18.73	10.66	21.16	-0.88

38	CRHRRRHRRGC	0.0676	-4.84	228.9	29.68	33.25	-12.04	30.72	-3.52
39	LLGRNSFEMRV	0.0234	-7.82	210.9	17.02	15.44	9.31	17.22	-1.18
40	RPKPQQFFGLM	0.0232	-5.75	225.0	16.98	16.07	5.34	17.77	-4.64
41	YAEGDVHATSK	0.0245	-5.08	159.4	17.40	16.20	6.89	18.51	-6.37
42	ACPGKDRRTGGGN	0.0316	-3.15	146.7	19.11	19.36	-1.28	22.64	-18.49
43	ACPGTDRRTGGGN	0.0235	-3.06	133.5	15.08	16.18	-7.32	16.56	-9.84
44	ACPGRNRRTEEENL	0.0273	-6.85	219.8	19.40	17.26	11.03	20.01	-3.14
45	MGGMNWRPILTIIT	0.0134	-10.45	248.6	10.20	11.17	-9.46	12.08	-18.47
46	SPALNKMFCELAKT	0.0211	-7.46	222.0	15.71	14.73	6.25	16.32	-3.90
47	VLTTGLPALISWIK	0.0139	-10.45	233.9	10.50	11.24	-7.07	11.83	-12.63
48	HRSCRRRKRRSCRHR	0.0733	-7.46	336.7	30.27	35.57	-17.51	30.68	-1.34
49	YSPALNKMCCQLAKT	0.0201	-7.46	226.7	14.90	14.41	3.27	15.38	-3.23
50	IITLEDSSNLLGRNSF	0.0118	-12.39	263.1	11.33	10.12	10.68	10.50	7.36
51	LAPPQHHLIQVGNLRV	0.0260	-11.27	273.2	15.01	15.92	-6.07	18.07	-20.40
52	LDDRNTFRRSVVVPYE	0.0232	-11.54	307.9	18.30	15.08	17.57	16.53	9.66
53	PPPGTRVRVMAIKQSQ	0.0262	-8.33	268.2	18.20	16.84	7.47	18.26	-0.34
54	TYSPALNRMFCQLAKT	0.0188	-8.69	273.5	14.77	13.97	5.44	14.66	0.74
55	DGLAPPQHRIRVEGNLR	0.0305	-10.10	284.7	18.98	18.10	4.65	20.40	-7.47
56	KSSQYIKANSKFIGITE	0.0248	-11.28	299.5	17.05	15.70	7.94	17.48	-2.53
57	LGRNSFEVCVCACPGRD	0.0183	-7.42	215.4	13.66	13.63	0.21	14.19	-3.88
58	NHQLLSPAKTGWRIFHP	0.0304	-10.02	323.1	19.42	18.40	5.24	20.28	-4.41
59	NTFRHSVVEPYEPPEVG	0.0179	-9.20	290.2	13.55	13.61	-0.43	14.52	-7.13
60	SSCMGGMNQRPILTIIT	0.0123	-10.97	251.5	10.66	10.62	0.36	10.99	-3.09
61	YKLVVVGACGVKGSALT	0.0202	-8.99	218.9	14.33	13.92	2.88	14.95	-4.36
62	YKLVVVGANGVGKSALT	0.0201	-9.77	233.4	14.36	13.78	4.04	14.92	-3.89
63	YKLVVVGARGVGKSALT	0.0267	-9.61	249.0	17.80	16.48	7.40	18.72	-5.16
64	YKLVVVGAVGVGKSALT	0.0202	-10.08	233.9	15.06	13.74	8.77	14.97	0.59
65	YNYMCNSSGMGGMNRRP	0.0182	-7.43	277.2	14.29	14.13	1.10	15.08	-5.53
66	FIGITEAAANLVPMVATV	0.0055	-11.52	248.7	4.73	7.87	-66.33	8.62	-82.22
67	VPYEPPEVGSVYHHPLQLHV	0.0219	-12.16	354.2	15.13	14.80	2.17	16.54	-9.32
68	RTHGQSHYRRRHCSRRRLHRIHRRQ	0.0708	-15.18	565.2	29.01	34.29	-18.21	28.57	1.52
69	KSSQYIKANSKFIGITEAAANLVPMVATV	0.0182	-17.93	449.9	14.21	12.52	11.92	14.54	-2.32
70	DRVIEVVQGAYRAIRHIPRRIRQGLERRIHIGPGRAFYTTKN	0.0437	-28.09	788.1	20.83	22.13	-6.24	21.80	-4.64
Test set									
71	FD	0.0173	-1.48	41.6	13.00	13.52	-4.02	12.82	1.41
72	MM	0.0195	-1.66	46.2	13.86	14.31	-3.28	14.14	-1.99
73	AAA	0.0221	0.00	17.0	14.96	15.56	-4.04	15.83	-5.78
74	RQQ	0.0322	-1.86	68.3	24.00	19.27	19.70	22.13	7.79
75	KKK	0.0703	-1.86	75.2	33.03	33.83	-2.43	31.55	4.49
76	SSQYIK	0.0227	-4.11	119.2	16.71	15.46	7.46	16.39	1.92
77	YMDGTMSQV	0.0073	-5.46	148.4	6.62	9.45	-42.69	9.04	-36.58
78	ACSGRDRRTEE	0.0316	-4.49	164.5	21.91	19.11	12.79	21.50	1.85
79	NSFCMGGMNRR	0.0241	-5.04	179.2	18.30	16.24	11.27	17.47	4.51
80	AAANLVPMVATV	0.0076	-6.65	150.4	6.15	9.23	-50.01	8.86	-44.14
81	DAEKSDICTDEY	0.0124	-7.32	173.0	9.91	11.05	-11.53	10.81	-9.07

82	GSDCTTIHCNYM	0.0143	-6.50	160.9	12.41	11.92	3.96	11.72	5.59
83	PHRERCSDSDGL	0.0295	-5.68	181.5	19.33	18.11	6.30	20.34	-5.22
84	TTIHYNYICNSS	0.0145	-8.46	202.8	10.59	11.80	-11.38	11.90	-12.41
85	HMTEVRRYPHHER	0.0469	-8.23	289.7	26.42	24.92	5.67	26.54	-0.44
86	YAEGDVHATSKPARR	0.0338	-6.32	239.1	21.38	20.04	6.25	22.15	-3.60
87	LAKTCPVRLWVDSTPP	0.0187	-8.68	258.5	15.13	13.77	9.00	14.65	3.20
88	VVRRCPHQRCSDSDGL	0.0311	-8.48	244.3	20.75	18.44	11.12	20.92	-0.82
89	YKLVVVGAAGVGKSALT	0.0204	-8.99	224.6	14.22	14.07	1.09	15.08	-6.03
90	KQIINMWQEVGKAMYAPPISGQIRRIHIGPGRAFYTTKN	0.0288	-24.13	702.2	17.78	16.91	4.92	18.39	-3.41
Validation set									
91	KKKK	0.0737	-2.48	100.2	33.03	35.17	-6.47	31.88	3.47
92	AAAAA	0.0161	0.00	28.3	12.34	13.36	-8.27	11.40	7.64
93	YGGFM	0.0121	-1.83	87.0	9.53	11.81	-23.89	9.56	-0.35
94	VISNDVCAQV	0.0072	-7.34	127.1	5.83	8.69	-48.99	6.75	-15.81
95	CRHHRRRHRRGC	0.0710	-5.50	252.7	29.68	34.54	-16.38	30.89	-4.06
96	HMTEVRHCPHHER	0.0484	-7.57	251.6	26.41	25.35	4.02	26.36	0.21
97	LAKTCPVRLWVDS	0.0210	-8.15	218.8	10.510	9.91	5.71	16.03	-52.51
98	RTHCQSHYRRRHCSR	0.0560	-7.49	308.0	28.96	28.72	0.84	27.64	4.54
99	EPPEVGSDYHHPLQLHV	0.0238	-10.06	290.1	16.91	15.60	7.77	16.62	1.73
100	KLVVVGAGDVGKSALTI	0.0198	-10.68	218.3	13.69	13.29	2.94	13.76	-0.52
101	TPPPGTRVQQSQHMTEV	0.0184	-8.44	270.6	14.17	13.86	2.15	14.02	1.08
102	FLTPKKLQCVDLHVISNDVCAQVHPQKVTK	0.0295	-20.82	494.9	18.68	16.35	12.47	18.65	0.16

[a] Definitions of descriptors are given in the text

Table 2. Specifications of the selected MLR model

Descriptor	Notation	Coefficient	Mean effecta
Charge to size ratio	QM	380.080(±17.360)	9.845
Corrected steric substituent constant	$E_{s,c}$	0.292(±0.170)	-1.775
Molecualr refractivity	MR	0.009(±0.006)	1.514
Constant		7.009(±0.512)	7.009

[a] Mean effect of a descriptor is the product of its mean and regression coefficient in the MLR model.

Table 3. Architecture and specifications of the ANN model

Number of nodes in the input layer	3
Number of nodes in the hidden layer	3
Number of nodes in the output layer	1
Number of iterations	35000
Learning rate	0.7
Momentum	0.9

Table 4. Comparison of the statistics for the MLR and ANN models

Model	$R^2_{Training}$	R^2_{Test}	$R^2_{Validation}$	$SE_{Training}$	SE_{Test}	$SE_{Validation}$	F
MLR	0.903	0.907	0.949	1.89	1.77	0.91	872
ANN	0.959	0.960	0.993	1.04	0.96	0.65	3153

Table 5. Comparison of different models for highly charged and hydrophobic peptides

No.	Peptide	Exp	Offord	Multi-var*	MLR	ANN
		μ_{ef} x 10^5	μ_{ef} x 10^5	μ_{ef} x 10^5	μ_{ef} x 10^5	μ_{ef} x 10^5
1	FF [a]	12.81	12.98	13.18	13.99	13.06
2	FL [a]	13.33	13.82	13.91	14.30	13.51
3	LL [a]	14.55	14.85	14.58	14.73	14.61
4	WW [a]	11.05	11.52	10.91	13.23	12.80
5	FFF [a]	10.38	10.59	10.76	12.48	11.92
6	KKKK [b]	33.03	45.34	33.53	35.17	31.88
7	KKKKK [b]	33.03	47.26	33.18	36.45	31.76
8	YGGFL [a]	9.75	9.62	9.70	11.76	10.58
9	ACHGRDRRT [b]	26.54	28.80	28.54	24.39	26.90
10	VVRRYPHHE [b]	27.38	27.40	25.46	23.24	25.55
11	CRHRRRHRRGC [b]	29.68	41.78	29.75	33.25	30.72
12	CRHHRRRHRRGC [b]	29.68	43.74	29.61	34.54	30.89
13	HMTEVRRYPHHER [b]	26.42	30.32	27.11	24.92	26.54
14	HMTEVRHCPHHER [b]	26.41	29.55	25.07	25.35	26.36
15	HRSCRRRKRRSCRHR [b]	30.27	45.09	31.41	35.57	30.68
16	RTHCQSHYRRRHCSR [b]	28.96	34.99	26.39	28.72	27.64
17	YAEGDVHATSKPARR [b]	21.38	22.10	21.94	20.04	22.15
18	VVRRCPHQRCSDSDGL [b]	20.75	20.54	19.22	18.44	20.92
19	DGLAPPQHRIRVEGNLR [b]	18.98	20.21	20.49	18.10	20.40
20	NHQLLSPAKTGWRIFHP [c]	19.42	20.14	18.85	18.40	20.28
21	RTHGQSHYRRRHCSRRRLHRIHRRQ [b]	29.01	43.60	28.32	34.29	28.57
22	FLTPKKLQCVDLHVISNDVCAQVHPQKVTK [c]	18.68	19.59	19.83	16.35	18.65
23	KQIINMWQEVGKAMYAPPISGQIRRIHIGPGRAFYTTKN [c]	17.78	19.21	17.71	16.91	18.39
24	DRVIEVVQGAYRAIRHIPRRIRQGLERRIHIGPGRAFYTTKN [c]	20.83	27.85	21.79	22.13	21.80
	R^2		0.91	0.98	0.91	0.99
	Standard Error		3.67	1.09	2.60	0.80
	RSD		46.43	33.38	36.73	31.30

[a] hydrophobic peptide, [b] highly-charged peptide, [c] hydrophobic and highly-charged peptide

* multi-variable model from Janini's work

Table 6. Descriptor values, together with MLR and ANN calculated mobilities of theoretical fragments of Lys-C digest of Cytochrome C and Glu-C digest of Glucagon

No.	Peptide sequence	Descriptors			MLR	ANN
		QM	$E_{s,c}$	MR	μ_{ef} x 10^5	μ_{ef} x 10^5
Glucagon Digest						
1	SRRAQD	0.0338	-2.92	103.6	19.88	22.05
2	YSKYLD	0.0204	-4.32	127.1	14.60	14.41
3	HSQGTFTSD	0.0177	-4.18	123.6	13.56	12.75
4	FVQWLMNT	0.0081	-6.45	172.9	9.69	7.90
Cytochrome C Digest						
5	GK	0.0530	-0.42	26.08	27.27	29.52
6	HK	0.0657	-1.28	48.84	32.02	31.39
7	NK	0.0450	-1.40	39.5	24.03	27.27
8	GGK	0.0450	-0.22	27.11	24.27	27.29
9	ATNE	0.0144	-1.93	48.2	12.32	10.18
10	GDVEK	0.0259	-2.91	68.8	16.60	17.67
11	GITWK	0.0257	-3.22	97.3	16.66	17.61
12	IFVQK	0.0249	-4.64	108.7	16.03	17.00
13	YIPGTK	0.0238	-3.26	103.3	15.97	16.48
14	MIFAGIK	0.0217	-5.17	124.0	14.79	15.02
15	CAQCHTVEK	0.0279	-4.14	144.4	17.64	18.86
16	TEREDLIAYLK	0.0223	-8.58	207.2	14.74	15.51
17	EETLMEYLENPK	0.0137	-8.42	224.3	11.68	10.92
18	TGPNLHGLFGRK	0.0322	-5.79	191.4	19.20	20.81
19	TGQAPGFTYTDANK	0.0135	-5.39	194.8	12.22	11.11

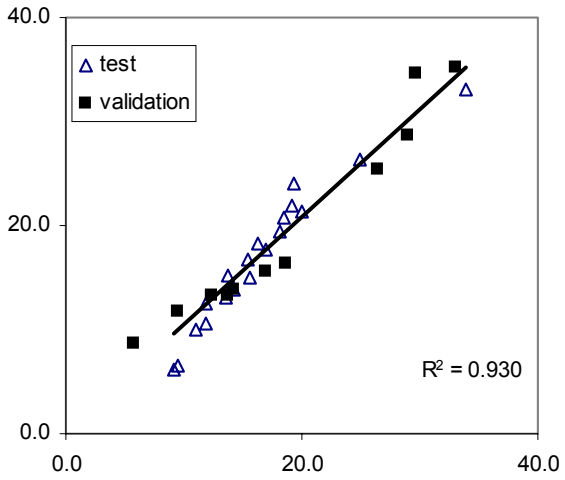

Fig1 Plot of the MLR calculated electrophoretic mobilities against the experimental values for the test and validation sets.

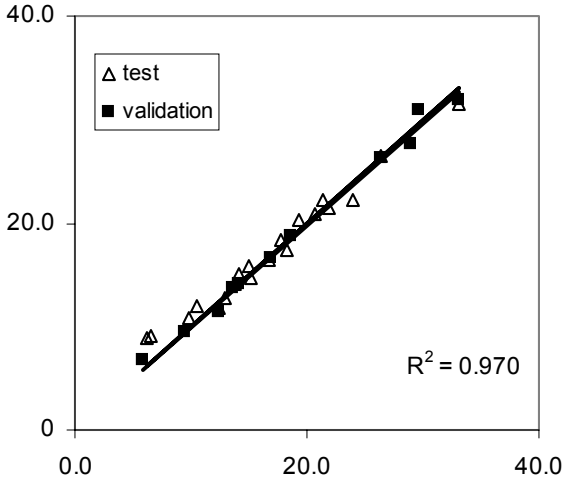

Fig 2 Plot of the ANN calculated electrophoretic mobilities against the experimental values for the test and validation sets.

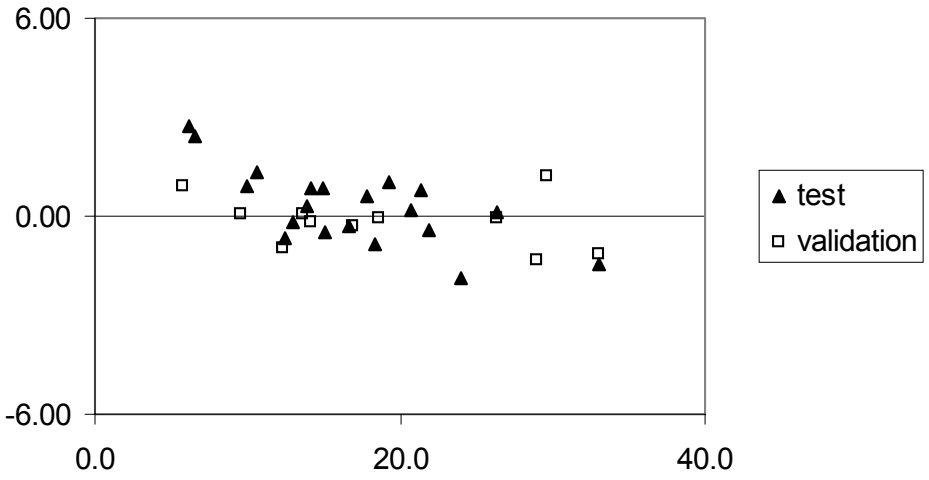

Fig 3 Residual plot of the ANN calculated electrophoretic mobilities for the test and validation sets.

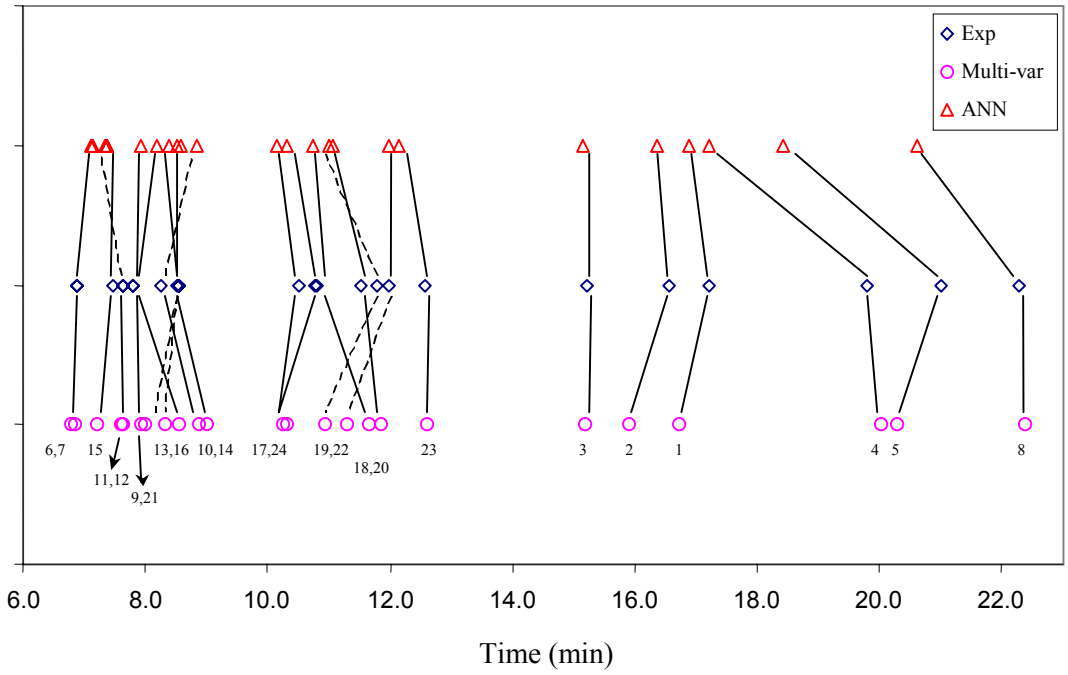

Fig 4 Comparison of different models for highly charged and hydrophobic peptides**.**

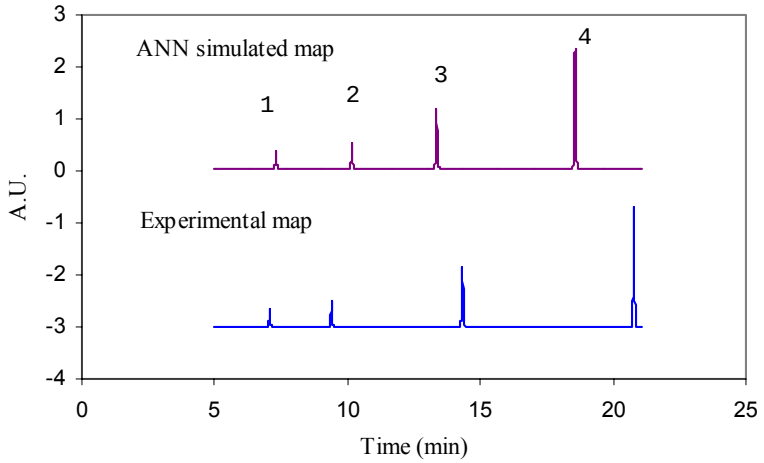

Fig 5 Comparison of ANN simulated map with experimental map of the Lys-C digest of melittin.

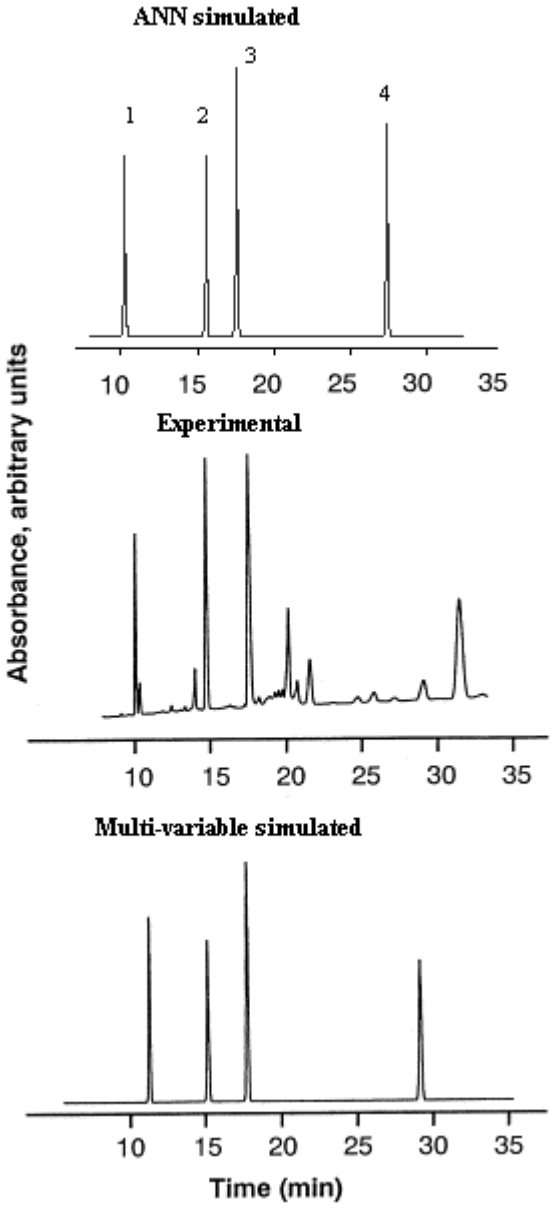

Fig 6 Comparison of ANN simulated map with Multi-variable simulated and experimental map of Glu-C digest of glucagons.

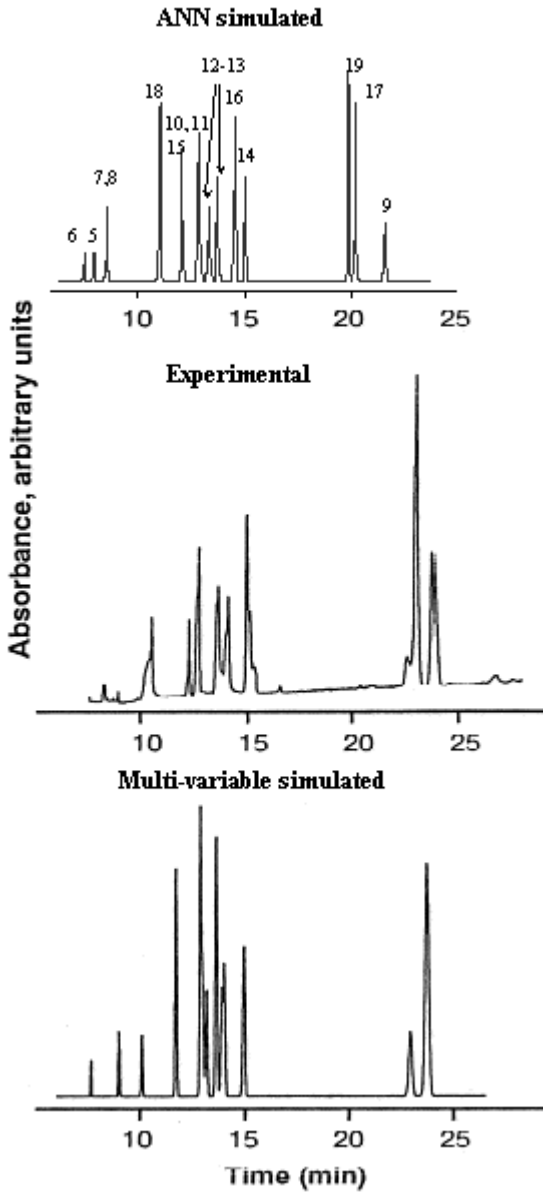

Fig 7 Comparison of ANN simulated map with Multi-variable simulated and experimental map of Lys-C digest of horse cytochrome c.

Chapter 4 Development of a "sequence-related" descriptor $\sum \frac{m}{\sqrt{n}}$ for peptide mobility prediction in CZE

Keywords: Capillary Zone Electrophoresis, Peptide, QSMR, Sequence, Peptide mapping

1. Abstract

The aim of this work was to further explore multivariate analysis techniques in predicting peptide mobilities in capillary zone electrophoresis(CZE). A new "sequence-related" descriptor $\sum \frac{m}{\sqrt{n}}$ is introduced to predict and differentiate the mobilities of peptides with the same amino acid composition but different sequences. A new Quantitative Structure-Mobility Relationships(QSMR) model was developed based on electrophoretic mobility of 125 standard peptides using Offord's charge over mass term($Q/M^{2/3}$), corrected steric substituent constant($E_{s,c}$) and sequence-related descriptor ($\sum \frac{m}{\sqrt{n}}$). The first two descriptors account for the charge-to-size ratio and steric effect, and the sequence descriptor represents the influence of both sequence and bulkiness of amino acid side chains on peptide mobilities in CZE. The new QSMR model based on the same data set showed improvement over our earlier QSMR model based on Offord's charge over mass term($Q/M^{2/3}$), corrected steric substituent constant($E_{s,c}$) and molar refractivity(MR). It also showed a significant improvement in predicting mobilities of the peptides with different sequences.

2. Introduction

Capillary Zone Electrophoresis(CZE) has been widely used in peptide separations due to its high efficiency, high speed, automation, simple experimental setup and high-throughput capability. In CZE, electrophoretic mobilities of charged solutes are proportional to their charge-to-size ratio. This simple relationship between mobility and molecular properties of charged solutes makes it possible to predict migration patterns from quantitative strcture-mobility relationships(QSMR) that could greatly facilitate method development and separation optimization.

Several empirical models have been developed by different research groups for the prediction of peptide mobilities in CZE[1-7]. Most of these models are based on Stoke's law that is mainly valid for rigid spherical molecules in low ionic strength buffers. In these models, electrophoretic mobility of peptide is proportional to the peptide net charge and inversely proportional to the peptide mass as:

$$\mu = a \frac{Q}{M^b} \qquad \text{Eq. 1}$$

Where a and b are constants. The main difference between these models is the b value. Janini and coworkers studied the electrophoretic mobility of 58 peptides ranging in size from 2 to 39 amino acid residues with net charge ranging from 0.65 to 7.82 units. They concluded that Offord model gave the best results among all the previous models, but it still can't accurately predict the mobilities of hydrophobic and highly charged small peptides. They also concluded that electrophoretic mobilities of peptides can't be accurately predicted with only charge(Q), molar mass(M) or number of amino acid residues(N).

Inspection of all previous reported results reveals that three factors may lead to the controversy of different models in predicting peptide mobility. First, the method most research groups used to calculate the net charge of peptides is not very accurate. Instead of using the ionization constants of the peptide to calculate the net charge, net charge of each amino acid residue in the peptide is calculated and added to the net charge of N- and C-terminal to get the net charge of this peptide. With this method, electrostatic and steric interactions of neighboring residues are not taken into consideration. Second, there's still argument about which descriptor of mass dependence to use in these models. Based on the peptide size and buffer ionic strength, either $1/M^{1/3}$, $1/M^{1/2}$ or $1/M^{2/3}$ may be used. Last, all previous models did not consider the influence of amino acid sequence on the electrophoretic mobility of peptides. Based on our experimental results, peptides with different sequence do have different electrophoretic mobilities in CZE.

In our previous work, Offord's mass to charge term($Q/M^{2/3}$) turns out to be the best among all the empirical models in predicting peptide mobility in CZE base on our data of 125 standard peptides. Two more descriptors namely corrected steric substutuent constant($E_{s,c}$) and molar refractivity(MR) was used with Offord model($Q/M^{2/3}$) to form a better MLR model[8]. But peptide with different sequence will have same electrophoretic mobilities if predicted using our MLR model.

In this work, a series of new "sequence-related" descriptors was generated and used to develop QSMR prediction models along with Offord's charge over mass term and other structural descriptors. To the best of our knowledge, no effort has been made toward introducing "sequence-related" descriptor into peptide mobility prediction models. Our results demonstrate that "sequence-related" descriptor can greatly improve the

accuracy of mobility prediction for peptides with same amino acid composition but different sequences.

3. Experimental Section

3.1 Peptides.

A total of 125 standard peptides were purchased from either Sigma Chemical Co. (St. Louis, MO) or Bachem Co. (Torrance, CA). These peptides are listed in Table.1. Also, the values of the charge over mass term in Offord model ($Q/M^{2/3}$), corrected steric constituent constant, $E_{s,c}$, and "sequence-realated" descriptor, $\sum \frac{m}{\sqrt{n}}$, are given in Table.1 for all peptides. The net charge of a peptide was calculated as the sum of the net charges of all charged amino acid residues and the N- and C-terminals. The net charge of each charged amino acid residue, N-terminal and C-terminal at pH 2.5 was determined using Henderson-Hasselbach equation [14]. Each arginine(R), histidine(H), lysine(K) and N-terminal contribute a charge of +1; while each aspartic acid(D) (pKa 3.5), glutamic acid(E) (pKa 4.5) and the C-terminal (pKa 3.2) contributes a charge of –0.091, –0.01 and –0.166, respectively.

3.2 Chemicals and materials.

The buffer solution was 50 mM sodium phosphate (pH 2.5) throughout the study. It was prepared by adjusting the pH of phosphoric acid to 2.5 using sodium hydroxide. High purity phosphoric acid was purchased from Aldrich Chemical Company (Milwaukee, WI), and sodium hydroxide was purchased from Fisher Scientific (Pittsburgh, PA). The sodium phosphate buffer was filtered through a 0.2 μm Acrodisc filter (STRL, Eatontown, NJ) before use.

Bare-silica capillary used in this study was purchased from Polymicro Technologies (Phoenix, AZ), with the inner diameter of 50 μm and the outer diameter of 375 μm. The total length and the length from the injection port to the detection window were 37 and 30 cm, respectively.

3.3 Equipment.

A Beckman P/ACE 5000 instrument with a UV detector was used in this work. The detection wavelength was set at 214 nm and 37 °C was used as the running temperature. The voltage was set at 15 kV. Low-pressure injection (0.5 psi, 2 seconds) was used to introduce the sample into the capillary. Before any injection, the bare-silica capillary was conditioned by the following rinsing procedure: MilliQ water for 10 minutes**,** 1M NaOH solution for 10 minutes, pure Methanol for 10 minutes, MilliQ water for 10 minutes and buffer for 20 minutes. Between injections, the capillary was pressure rinsed (20 psi) with buffer solution for 5 minutes.

3.4 Determination of the effective mobility of peptides.

In the present work, mesityl oxide was selected as the EOF marker. Each standard was prepared into a solution of about 1mg peptide/ml. Each solution was introduced into the CE instrument by low pressure (0.5 psi) for 2 seconds and its individual retention time was recorded.

Migration times and mobilities of the peptides in this study were measured by preparing several mixtures, each containing 5-6 peptides (200 μl each) and 20 μl of mesityl oxide. Identity of each peptide peak was confirmed by matching its retention time with the individual retention time of the peptides in this mixture. Three injections were made for each sample mixture.

Effective mobilities of peptides were determined by Eq.2.

$$\mu_{ef} = \frac{L_t L_d}{V}\left(\frac{1}{t_r} - \frac{1}{t_{eo}}\right) \qquad \text{Eq.2}$$

where L_t is the total length of the capillary, L_d is the length from the capillary inlet to the detection point, V is the applied voltage, t_r is the peptide retention time and t_{eo} is the retention time of the EOF marker.

3.5 Reproducibility of experiment.

Each sample mixture was injected three times. Excellent reproducibility was achieved. The range of RSD(%) of the peptides effective mobilities was from 0.02% to 4.46%, with an average of 1.95% for the whole data set.

3.6 Development of "sequence-related" descriptors.

A total of 14 "sequence-related" descriptors were developed in this work. They are listed in Table.2. Descriptor values of 125 peptides were calculated by a short Matlab program(Appendix A) that greatly reduced the time for the calculation. In these sequence descriptors, m is the mass of the amino acid, n is the position of the amino acid residue in the sequence(starting from N-terminal), w is the mass of the amino acid side chain and z is the number of bonds in the peptide backbone from the residue to the N-terminal(z = 3n-1). To calculate the $\sum \frac{m}{\sqrt{n}}$ value for peptide AF, for example, the mass of alanine is 89.1, and it is the first residue in sequence(n=1); the mass of phenylalanine is 165.2, and it is the second residue in sequence(n=2). Therefore, the $\sum \frac{m}{\sqrt{n}}$ value for AF would be $[89.1/(1^{1/2}) + 165.2/(2^{1/2})] = 205.9$.

3.7 Generation of QSMR model

Stepwise multi-linear regression procedure was used to generate the QSMR model. Since the charge over mass term in Offord model($Q/M^{2/3}$) is the most important descriptor, it was the first one to be included into the QSMR model. After that, stepwise MLR procedure was implemented to the rest of the descriptors. Es,c was the second descriptor that appeared in the stepwise regression. Various models were compared and the best MLR model was chosen based on correlation coefficient, F statistics and standard error. In this study, the best MLR model was consisted of three descriptors: Offord's charge over mass term($Q/M^{2/3}$), corrected steric substituent constant($E_{s,c}$) and "sequence-related" descriptor($\sum \frac{m}{\sqrt{n}}$) as in Eq.3.

$$\mu = a \frac{Q}{M^{2/3}} + b\, E_{s,c} + c \sum \frac{m}{\sqrt{n}} \qquad \text{Eq.3}$$

4. Results and Discussion

The main goal of this work was to develop a QSMR model in CZE that can accurately predict electrophoretic mobilities of peptides with the same amino acid composition but different sequences. For this purpose, a dataset consisting of 125 peptides ranging in size from 2 to 14 amino acid residues was chosen. The net charge of these peptides ranges from 0.743 to 5.843, while their mobility ranges from 13.762 to 56.533 x $10^{-5}cm^2v^{-1}s^{-1}$. There were 23 pairs of peptides with the same amino acid composition but different sequences. One peptide from each pair was randomly picked to form the test set(total 23 peptides). The remaining 102 peptides formed the training set which was used to generate the models.

The mobilities of all the 125 peptides was measured at pH 2.5. At this pH, silanol groups on the capillary wall are protonated that greatly reduce or even eliminate EOF. All peptides carry positive charge at this pH and move toward cathode. The adsorption of peptides on the capillary wall is minimized at this pH that makes it ideal for the peptide mobility study.

4.1 Development of QSMR model

In our previous work on the same peptide dataset, it was demonstrated that Offord's charge over mass term($Q/M^{2/3}$) is superior to other empirical models(eg. $Q/M^{1/3}$, $Q/M^{1/2}$ and $\log(1+0.297Q)/M^{0.411}$) in fitting our experimental data[8]. In this work, Offord's charge over mass term was also chosen as the first descriptor to be included in our QSMR model. Descriptors used by other groups such as number of amino acid residues, molar mass, average residue mass, peptide net charge, surface area, hydrophobicity scale, isoelectric point value, corrected steric substituent constant, strain parameter and Z-scales had been tried together with our newly developed "sequence-related" descriptors for peptide mobility modeling(Table.2). Using Minitab software, correlation coefficient matrix of all descriptors was generated. To avoid bringing in redundant information to the QSMR model, one descriptor will be eliminated if it is highly correlated with other descriptors but contribute less to the mobility(smaller R^2). Only three "sequence-related" descriptors that are not correlated namely $\Sigma(z/w)$, $\Sigma(w/z)$ and $\sum \frac{m}{\sqrt{n}}$ were left after this step. $\sum \frac{m}{\sqrt{n}}$ was the best "sequence-related" descriptor among the three as shown in Figure 1. Stepwise addition method was processed on the remaining descriptors and the descriptors that can contribute more to peptide mobility were picked to generate the QSMR model. The specifications of the best MLR model

generated in this work were listed in Table.3. In this work, three descriptors appeard in the MLR model: Offord's charge over mass term($Q/M^{2/3}$), corrected steric substituent constant($E_{s,c}$) and "sequence-related" descriptor $\sum \frac{m}{\sqrt{n}}$. And their correlation matrix was listed in Table.4.

Mean effect of each descriptor was also calculated and listed in Table.3 to describe its contribution to peptide mobility. The mean effect of a descriptor is the product of its regression coefficient in the MLR equation with its average value in the dataset. In this MLR model, Offord's charge over mass term($Q/M^{2/3}$) shows the largest mean effect of 25.48, which means charge-to-size ratio plays the most important part in determining peptide mobility in CZE. Corrected steric substituent constant($E_{s,c}$) has the smallest mean effect value of –4.95 among these three descriptors. When the $E_{s,c}$ value of a peptide is more negative which means it have more and bigger residues, it tends to experience larger resistance and have a relatively smaller mobility. The mean effect of $\sum \frac{m}{\sqrt{n}}$ is 9.56, which means sequence is not negligible in peptide mobility prediction. It agrees well with our experimental data in which difference between the mobilities of peptide pairs different only in sequences can be as high as 19%.

The predicted peptide mobilities using both MLR model developed in this work and our previous work[8] were compared with experimental values and correlation coefficients, standard error and F-test value were listed in Table 5.

4.2 Correlation between Molar Refractivity and "sequence-related" descriptor $\sum \frac{m}{\sqrt{n}}$

Multi-collinearility between various descriptors was checked. Using the Correlation function of Minitab software, the correlation coefficient matrix of the descriptors was calculated based on the descriptor values of the 102 peptides in the training set.

It was noticed that the correlation coefficient between molar refractivity and "sequence-related" descriptor $\sum \frac{m}{\sqrt{n}}$ was very high($R^2 = 0.99$). Molar refractivity is a constitutive-additive property that is calculated by Lorenz-Lorentz formula. It is strongly related to the volume of the molecule(bulkiness parameter). Since peptides tend to have similar density(~ 0.73 g/cm^3), molar refractivity is also highly correlated with the molecule weight of peptides.

"Sequence-related" descriptor $\sum \frac{m}{\sqrt{n}}$ is a descriptor that can describe not only the "bulkiness" but also the "compactness" of a peptide. The biggest advantage of this descriptor over molar refractivity is it can assign different descriptor value to peptides with the same amino acid composition but different sequences while molar refractivity values for these peptides are identical. In this case, it makes $\sum \frac{m}{\sqrt{n}}$ an ideal descriptor for the prediction of the mobilities of peptides with different sequence. For example, with our previous MLR model[8], the predicted mobilities of peptide AF and FA are the same while their experimental values are 30.71 and 33.13 x $10^{-5}cm^2v^{-1}s^{-1}$respectively. Their molar refractivity values are the same(35.7), but the $\sum \frac{m}{\sqrt{n}}$ values of these two peptides are 205.90 and 228.19 respectively. As expected, with descriptor $\sum \frac{m}{\sqrt{n}}$ we are able to

predict the electrophoretic mobilities of peptides with different sequences in CZE more accurately.

4.3 Comparison of the experimental and MLR predicted mobilities of peptides with different sequences

A total of 46 peptides with different sequences(23 pairs) was listed in Table.6 with corresponing experimental mobility and MLR calculated mobilities. With our previous MLR model, the calculated mobilities of both peptides in one pair are the same, while their experimental mobilities are different. With the current MLR model, the correlation coefficient and standard error between the calculated and experimental mobility showed improvement. It was shown that for most of these peptide pairs, the relative magnitude of the MLR predicted mobilities for the two peptides in a pair are in good agreement with the experimental value. The MLR calculated mobility of a peptide will be largger than the other peptide in a pair if its experimental mobility is larger. The reason for this is because $\sum \frac{m}{\sqrt{n}}$ describes not only the bulkiness of the peptide but also the "compactness" of the peptide. When a peptide has a larger $\sum \frac{m}{\sqrt{n}}$ value, it will have bigger residues close to the N-terminal positive charge which effectively makes the positive charge more localized and it also reduces the interaction of the N-terminal positive charge with the negative charged groups either in the solution(eg. phosphate group) or on the capillary wall(eg. deprotonated silanol group). For these reason, its mobility will be larger than the other peptide in the pair.

4.4 Comparison of the experimental and MLR predicted maps of the tryptic digest of horse cytochrome C

The ultimate goal for this project is to find models that can accurately predict peptide maps for biological samples. For this purpose, the generated MLR model was used to predict the peptide map of the tryptic digest of horse cytochrome C, and the new map was compared with both the experimental and previously simulated maps[8]. Since the absorbance coefficient of a peptide is proportional to the number of its peptide bonds, and the concentration of the protein digest was low, three dipeptides and one tripeptide with no aromatic side chain were removed because they can't be detected at this wavelength.

It was observed that the elution orders of all the theoretical fragments are the same for both MLR models. With the MLR model developed in this work, the retention times of the theoretical fragments agree better with those of the peptide peaks in the experimental electrophoregram(Fig.2a,2b). It works especially well for fragment 13 to 16. With our previous MLR model, we will get four evenly spaced peaks for the fragments in the predicted map. However, with current MLR model, peak 13-15 appear closer to each other which is in agreement with the experimental electroporegram.

The experimental and MLR predicted retention times for each peptide fragment were listed in Table.7. The correlation coefficient between the new MLR model predicted and experimental retention time is 0.99(Fig.3).

5. Conclusions

A new Quantitative Structure-Migration Relationships(QSMR) model was developed using a data set of 125 standard peptides in capillary zone electrophoresis(CZE). In this work, a "sequence-related" descriptor $\sum \frac{m}{\sqrt{n}}$, Offord's

charge over mass term($Q/M^{2/3}$) and corrected steric substituent constant($E_{s,c}$) appeared in the MLR model. $\sum \frac{m}{\sqrt{n}}$ is highly correlated with the molar refractivity(MR) we used in our previous MLR model. It can indicate not only the "bulkiness" of the peptide but also can improve the mobility prediction accuracy of peptides with difference sequences. It is a better descriptor than molar refractivity to be used in developing QSMR models for peptides. By inspecting the data of 23 pair of peptides that only differ in sequences, it was observed that in most cases the relative magnitude of predicted mobility of two peptides in a pair is in agreement with the experimental data. Although there are no peptides with different sequence in tryptic digest of horse cytochrome C, the new MLR model was used to predict the peptide map of this digest. More accurate prediction was observed compared with our previous MLR model using $Q/M^{2/3}$, $E_{s,c}$ and MR. The correlation(R^2) between MLR predicted and experimental retention time was 0.99.

References

[1] Grossman, P. D., Colburn, J. C., Lauer, H. H., Anal. Biochem. 1989, 179(1), 28-33.

[2] Offord, R. E., Nature (London) 1966, 211, 591-593.

[3] Compton, B. J., J. Chromatogr. 1991, 559, 357-366.

[4] Adamson, N. J., Reynolds, E. C., J. Chromatogr. B 1997, 699, 133-147.

[5] Messana, I., Rossetti, D. V., Cassiano, L., Misiti, F., Giardina, B., Castagnola, M., J. Chromatogr. B 1997, 699, 149-171.

[6] Kasicka, V., Electrophoresis 1999, 20, 3084-3105.

[7] Cifuentes, A., Poppe, H., Electrophoresis 1997, 18, 2362-2376.

[8] Jalali-Heravi, M., Shen, Y., Hassanisadi, M. and Khaledi, M.G., Manuscript submitted to Electrophoresis.

Table 1. Experimental and two MLR calculated mobilities for standard peptides with their descriptor values

No.	Peptide	Descriptors				Exp	MLR 1		MLR 2	
Training set		$Q/M^{2/3}$	Es,c	$\Sigma(m/n^{1/2})$	MR	$\mu_{ef} \times 10^5$	$\mu_{ef} \times 10^5$	% Dev	$\mu_{ef} \times 10^5$	% Dev
1	AW	0.0197	-0.66	233.50	45.50	30.10	27.16	-9.77	27.29	**-9.32**
2	A_LY_L	0.0209	-0.70	217.21	37.50	29.37	28.22	-3.90	28.26	**-3.79**
3	AY	0.0209	-0.70	217.21	37.50	29.58	28.22	-4.59	28.26	**-4.48**
4	D_LF_L	0.0173	-1.48	249.91	41.60	24.02	22.69	-5.52	23.40	**-2.56**
5	DF	0.0173	-1.48	249.91	41.60	23.22	22.69	-2.29	23.40	**0.77**
6	EW	0.0171	-1.28	291.54	56.00	26.85	23.46	-12.63	24.44	**-8.99**
7	FA	0.0218	-0.70	228.19	35.70	33.13	29.34	-11.44	29.60	**-10.66**
8	F_LF_L	0.0181	-1.40	282.00	60.00	26.97	24.77	-8.17	25.16	**-6.71**
9	FF	0.0181	-1.40	282.00	60.00	27.91	24.77	-11.24	25.16	**-9.84**
10	FI	0.0196	-2.31	257.94	49.60	30.10	24.80	**-17.60**	24.70	-17.92
11	FM	0.0188	-1.53	270.70	53.10	27.80	25.06	-9.84	25.44	**-8.48**
12	FV	0.0202	-1.79	248.03	45.00	29.66	26.20	-11.67	26.23	**-11.59**
13	GF	0.0227	-0.50	191.88	31.00	31.71	30.57	**-3.60**	30.23	-4.69
14	GW	0.0204	-0.46	219.48	40.80	30.01	28.11	-6.34	28.17	**-6.15**
15	GY	0.0217	-0.50	203.19	32.80	27.06	29.32	8.36	29.26	**8.13**
16	HW	0.0375	-1.32	299.57	63.60	45.71	50.44	10.35	48.66	**6.45**
17	HY	0.0393	-1.36	283.28	55.60	43.27	52.30	20.87	50.34	**16.34**
18	IW	0.0179	-2.27	275.58	59.40	27.85	23.23	-16.58	23.23	**-16.58**
19	K_LF_L	0.0415	-1.32	263.00	55.10	49.89	55.21	10.67	52.57	**5.37**
20	KF	0.0415	-1.32	263.00	55.10	49.97	55.21	10.50	52.57	**5.21**
21	KW	0.0382	-1.28	290.60	64.90	48.80	51.45	5.44	49.33	**1.09**
22	LF	0.0196	-1.94	247.98	49.60	28.12	25.33	**-9.89**	25.15	-10.56
23	L_LW_L	0.0179	-1.90	275.58	59.40	27.68	23.77	-14.13	23.90	**-13.66**
24	LW	0.0179	-1.90	275.58	59.40	27.93	23.77	-14.91	23.90	**-14.44**
25	MW	0.0173	-1.49	293.62	62.90	28.48	23.70	-16.79	24.27	**-14.78**
26	PW	0.0185	-0.66	259.54	53.80	30.44	26.09	-14.29	26.51	**-12.90**
27	RW	0.0362	-1.28	318.61	69.90	47.71	49.10	2.90	47.58	**-0.27**
28	VW	0.0185	-1.75	261.56	54.80	29.98	24.46	-18.42	24.50	**-18.27**
29	WD	0.0159	-1.44	298.35	51.40	25.06	21.38	-14.71	22.84	**-8.85**
30	WF	0.0167	-1.36	321.04	69.80	26.83	23.56	-12.20	24.49	**-8.73**
31	WS	0.0190	-0.94	278.54	51.60	28.42	26.12	-8.08	26.93	**-5.22**
32	WW	0.0156	-1.32	348.64	79.60	27.45	22.65	-17.49	23.83	**-13.18**
33	YI	0.0188	-2.31	273.94	51.40	28.88	23.96	-17.05	24.22	**-16.16**
34	YL	0.0188	-1.94	273.94	51.40	27.19	24.49	-9.92	24.88	**-8.50**
35	YV	0.0195	-1.79	264.03	46.80	29.07	25.28	-13.02	25.66	**-11.71**
36	YW	0.0163	-1.36	325.60	71.60	25.66	23.01	-10.30	24.01	**-6.41**
37	YY	0.0170	-1.40	309.31	63.60	19.87	23.47	**18.13**	24.42	22.92

38	FFF	0.0140	-2.10	377.37	90.00	24.03	19.97	-16.90	21.17	**-11.89**
39	FGG	0.0195	-0.30	261.61	32.00	28.75	26.72	-7.07	28.34	**-1.43**
40	$G_LF_LL_L$	0.0173	-1.74	267.61	50.60	27.50	22.69	-17.48	23.24	**-15.47**
41	GFL	0.0173	-1.74	267.61	50.60	27.01	22.69	-16.01	23.24	**-13.97**
42	KYK	0.0492	-1.94	358.71	82.00	56.53	65.70	16.21	62.61	**10.75**
43	MLF	0.0151	-2.77	337.33	72.70	23.17	19.55	-15.62	20.40	**-11.96**
44	WGG	0.0179	-0.26	300.65	41.80	27.81	25.17	-9.50	27.36	**-1.62**
45	WGY	0.0148	-1.16	361.92	72.60	23.96	21.41	-10.65	23.42	**-2.27**
46	WWW	0.0120	-1.98	466.55	119.40	22.71	19.12	-15.83	21.05	**-7.32**
47	YAG	0.0182	-0.50	287.53	38.50	27.36	25.10	-8.27	27.04	**-1.17**
48	YGG	0.0188	-0.30	277.61	33.80	27.69	25.89	-6.50	27.86	**0.61**
49	YPF	0.0147	-1.40	357.97	75.80	22.28	21.20	-4.86	22.87	**2.64**
50	YYL	0.0140	-2.64	385.04	83.20	20.73	18.88	-8.89	20.42	**-1.47**
51	YYY	0.0131	-2.10	413.92	95.40	21.39	19.08	-10.80	20.92	**-2.18**
52	FFFF	0.0116	-2.80	459.96	120.00	19.81	17.44	-11.98	18.95	**-4.33**
53	FGGF	0.0147	-1.00	344.21	62.00	22.16	21.02	-5.14	23.25	**4.93**
54	GGFL	0.0156	-1.54	289.11	51.60	25.44	20.79	-18.28	22.05	**-13.33**
55	GGFM	0.0151	-1.13	298.13	55.10	24.09	20.97	-12.95	22.44	**-6.82**
56	RFDS	0.0268	-2.38	420.40	83.50	38.29	35.97	-6.06	36.79	**-3.93**
57	WGGY	0.0136	-0.96	391.25	73.60	21.69	20.19	-6.90	23.02	**6.13**
58	YGGF	0.0144	-1.00	360.21	63.80	20.93	20.65	**-1.36**	23.19	10.76
59	FFFFF	0.0101	-3.50	533.84	150.00	17.93	15.96	-11.00	17.49	**-2.45**
60	FLEEI	0.0108	-4.79	475.11	101.60	18.16	12.58	-30.75	14.79	**-18.53**
61	RKDVY	0.0355	-3.81	494.02	113.60	46.00	46.78	1.70	46.08	**0.19**
62	TRSAW	0.0252	-2.09	438.85	99.20	36.70	35.13	-4.28	35.83	**-2.37**
63	WGGGY	0.0126	-0.76	419.22	74.60	19.26	19.26	**-0.01**	22.85	18.61
64	YAGFL	0.0121	-2.44	428.78	88.10	20.66	16.94	-18.03	19.50	**-5.64**
65	YGGFM	0.0121	-1.83	426.94	86.90	20.09	17.68	-11.99	20.48	**1.96**
66	YGGWL	0.0118	-2.20	438.39	93.20	19.89	17.10	-14.01	19.74	**-0.77**
67	PPGFSP	0.0117	-0.78	416.48	84.80	20.27	18.61	-8.15	21.71	**7.10**
68	RGPFPI	0.0236	-2.73	481.39	108.70	34.94	32.55	-6.84	33.68	**-3.61**
69	RRPYIL	0.0324	-4.79	566.65	145.20	44.05	43.04	**-2.30**	42.33	-3.90
70	RYLGYL	0.0216	-4.30	550.17	133.90	34.44	28.97	-15.88	30.01	**-12.87**
71	VEPIPY	0.0103	-4.02	478.70	110.60	16.98	13.42	-20.93	15.59	**-8.17**
72	WHWLQL	0.0199	-5.08	616.35	161.70	30.94	27.17	-12.19	28.15	**-9.02**
73	YSGFLT	0.0107	-3.25	488.73	106.00	16.73	14.85	-11.23	17.70	**5.79**
74	DRVYIHP	0.0294	-5.46	580.03	145.90	38.12	38.21	**0.24**	37.90	-0.57
75	MEHFRWG	0.0290	-3.89	615.08	164.00	33.57	40.81	**21.58**	40.94	21.96
76	RPKPQQF	0.0304	-3.18	585.04	151.40	41.66	43.04	**3.30**	43.23	3.77
77	RPPGFSP	0.0221	-1.40	519.91	114.90	34.11	32.84	-3.75	35.15	**3.03**
78	RVYIHPI	0.0305	-6.29	593.20	153.90	42.71	38.76	**-9.25**	37.91	-11.25
79	RVYVHPF	0.0300	-4.86	599.05	159.70	42.53	40.55	**-4.66**	40.08	-5.77
80	YGGFMRF	0.0200	-3.15	560.49	147.00	30.73	29.29	-4.69	30.47	**-0.87**
81	YMEHFRW	0.0270	-4.79	671.41	194.80	39.52	38.58	**-2.38**	38.27	-3.16
82	YPFVEPI	0.0091	-4.72	578.92	140.60	15.57	12.41	-20.29	15.14	**-2.79**

83	ASTTTNYT	0.0092	-3.88	509.54	111.00	14.38	12.27	-14.62	15.28	**6.31**
84	DRVYIHPF	0.0266	-6.16	638.43	175.90	37.06	35.08	**-5.37**	34.60	-6.66
85	NRVYVHPF	0.0278	-5.64	631.18	174.20	39.99	37.24	**-6.88**	36.73	-8.16
86	PPGFSPFR	0.0196	-2.10	540.50	144.90	30.69	30.18	**-1.65**	31.44	2.44
87	RPGFSPFR	0.0291	-2.72	599.57	161.00	41.65	42.46	**1.95**	42.80	2.76
88	WQPPRARI	0.0279	-4.13	658.10	172.40	40.15	39.52	-1.57	40.21	**0.15**
89	RPPGFSPFR	0.0273	-2.72	636.38	175.00	39.75	40.84	**2.75**	41.49	4.38
90	YLEPGPVTA	0.0085	-3.98	613.12	129.10	13.76	12.18	**-11.46**	16.60	20.61
91	IARRHPYFL	0.0345	-6.15	668.84	204.70	46.71	46.87	**0.35**	44.59	-4.54
92	DRVYIHPFHL	0.0315	-8.06	731.63	219.30	40.92	40.96	**0.09**	39.01	-4.68
93	DRVYVHPFHL	0.0317	-7.54	725.36	214.70	41.95	41.76	**-0.45**	40.07	-4.49
94	SYSMEHFRWG	0.0237	-5.15	713.47	219.40	35.60	35.03	**-1.60**	34.65	-2.67
95	YRPPGFSPFR	0.0248	-3.42	719.99	206.80	36.38	38.26	**5.16**	39.16	7.63
96	ELYENKPRRPY	0.0296	-6.52	791.03	243.40	31.97	41.96	31.26	40.84	**27.76**
97	RPKPQQFFGLM	0.0232	-5.75	754.93	225.10	32.61	33.81	**3.68**	33.91	3.97
98	CGYGPKKKRKVGG	0.0471	-4.09	716.60	195.30	53.57	65.86	22.95	64.29	**20.02**
99	ELYENKPRRPFIL	0.0270	-9.37	860.45	280.80	33.52	36.50	8.88	34.28	**2.26**
100	ELYENKPRRPYIL	0.0269	-9.37	865.28	282.60	29.36	36.37	23.89	34.19	**16.44**
101	AGCKNFFWKTFTSC	0.0204	-5.92	778.50	236.60	29.04	30.47	**4.92**	30.78	5.98
102	DRVYIHPFHLVIHN	0.0325	-12.20	883.16	292.20	29.24	40.18	37.41	36.19	**23.78**

Test set

103	AF	0.0218	-0.70	205.90	35.70	30.71	29.34	**-4.43**	29.11	-5.21
104	WA	0.0197	-0.66	267.23	45.50	30.08	27.16	-9.71	28.04	**-6.76**
105	YA	0.0209	-0.70	244.19	37.50	30.77	28.22	-8.29	28.86	**-6.23**
106	DW	0.0159	-1.44	277.51	51.40	22.38	21.38	-4.48	22.38	**0.01**
107	WE	0.0171	-1.28	308.27	56.00	25.78	23.46	-8.98	24.81	**-3.75**
108	FG	0.0227	-0.50	218.27	31.00	32.63	30.57	-6.32	30.81	**-5.58**
109	IF	0.0196	-2.31	247.98	49.60	28.11	24.80	**-11.79**	24.48	-12.92
110	FL	0.0196	-1.94	257.94	49.60	29.43	25.33	-13.91	25.37	**-13.79**
111	VF	0.0202	-1.79	233.96	45.00	29.13	26.20	**-10.06**	25.92	-11.05
112	FW	0.0167	-1.36	309.60	69.80	24.65	23.56	-4.43	24.23	**-1.68**
113	WG	0.0204	-0.46	257.31	40.80	30.32	28.11	-7.29	29.01	**-4.34**
114	YG	0.0217	-0.50	234.27	32.80	30.40	29.32	-3.55	29.95	**-1.48**
115	IY	0.0188	-2.31	259.29	51.40	26.57	23.96	**-9.83**	23.89	-10.09
116	WL	0.0179	-1.90	296.98	59.40	28.02	23.77	-15.17	24.37	**-13.01**
117	LY	0.0188	-1.94	259.29	51.40	22.89	24.49	**7.01**	24.55	7.27
118	WM	0.0173	-1.49	309.74	62.90	25.26	23.70	-6.19	24.63	**-2.51**
119	WP	0.0185	-0.66	285.64	53.80	28.44	26.09	-8.26	27.09	**-4.73**
120	WR	0.0362	-1.28	327.41	69.90	46.96	49.10	4.54	47.78	**1.74**
121	WV	0.0185	-1.75	287.07	54.80	28.68	24.46	-14.73	25.07	**-12.59**
122	VY	0.0195	-1.79	245.27	46.80	28.53	25.28	**-11.39**	25.25	-11.52
123	WY	0.0163	-1.36	332.35	71.60	25.69	23.01	-10.41	24.16	**-5.94**
124	GGF	0.0195	-0.30	223.53	32.00	28.20	26.72	-5.25	27.49	**-2.51**
125	FGFG	0.0147	-1.00	351.18	62.00	23.71	21.02	-11.34	23.41	**-1.27**

MLR 1: using $Q/M^{2/3}$, Es,c and MR

MLR 2: using $Q/M^{2/3}$, Es,c and $\sum \frac{m}{\sqrt{n}}$

Table 2. List of descriptors used in developing QSMR models

Non-"sequence-related" Descriptor	Notation
Number of amino acid residues	N
Molar mass	M
Average residue mass	arm
Peptide net charge	Q
Surface area	S
Hydrophobicity scale	π
Isoelectric point value	PI
Charge to size ratio	$Q/M^{2/3}$
Corrected steric substituent constant	Es,c
Strain parameter	Strain

"Sequence-related" Descriptor[a]	
$\Sigma(m/n^{1/2})$	$\Sigma(w/n^{1/2})$
$\Sigma(m/n)$	$\Sigma(w/n)$
$\Sigma(m/z^{1/2})$	$\Sigma(w/z^{1/2})$
$\Sigma(m/z)$	$\Sigma(w/z)$
$\Sigma(n^{1/2}/m)$	$\Sigma(n^{1/2}/w)$
$\Sigma(z^{1/2}/m)$	$\Sigma(z^{1/2}/w)$
$\Sigma(n/m)$	$\Sigma(z/w)$

[a] when calculating "sequence-related" descriptors: n is the position number of a amino acid in the sequence, m is molar mass of the amino acid at position n, w is mass of the amino acid side chain, z is the number of bonds from the residue to the N-terminal of the peptide

Table3. Specifications of current MLR model

Descriptor	Notation	Coefficient	Mean effect
Charge to size ratio	$Q/M^{2/3}$	1181.96(±31.33)	25.48
Corrected steric substituent constant	$E_{s,c}$	1.79(±0.25)	-4.95
Sequence descriptor	$\Sigma(m/n^{1/2})$	0.02(±0.00)	9.56

Table 4. Correlation matrix of descriptors in current MLR model

	$Q/M^{2/3}$	$E_{s,c}$	$\Sigma(m/n^{1/2})$
$Q/M^{2/3}$	1		
$E_{s,c}$	-0.300	1	
$\Sigma(m/n^{1/2})$	0.273	-0.879	1

Table 5. Comparison of two MLR models

Model	$R^2_{Training}$	R^2_{Test}	$SE_{Training}$	SE_{Test}	F
QSMR 1	0.89	0.90	3.20	1.56	2307
QSMR 2	0.91	0.91	3.01	1.54	3498

QSMR1: use $Q/M^{2/3}$, Es,c and MR
QSMR2: use $Q/M^{2/3}$, Es,c and $\Sigma(m/n^{1/2})$

Table 6. Comparison of experimental and MLR calculated mobility of peptides with different sequences

No.	Peptide	Exp	MLR 1	MLR 2
		$\mu_{ef} \times 10^5$	$\mu_{ef} \times 10^5$	$\mu_{ef} \times 10^5$
103	AF	30.71	29.34	29.11
7	FA	33.13	29.34	29.60
1	AW	30.10	27.16	27.29
104	WA	30.08	27.16	28.04
3	AY	29.58	28.22	28.26
105	YA	30.77	28.22	28.86
106	DW	22.38	21.38	22.38
29	WD	25.06	21.38	22.84
6	EW	26.85	23.46	24.44
107	WE	25.78	23.46	24.81
108	FG	32.63	30.57	30.81
13	GF	31.71	30.57	30.23
10	FI	30.10	24.80	24.70
109	IF	28.11	24.80	24.48
110	FL	29.43	25.33	25.37
22	LF	28.12	25.33	25.15
12	FV	29.66	26.20	26.23
111	VF	29.13	26.20	25.92
112	FW	24.65	23.56	24.23
30	WF	26.83	23.56	24.49
14	GW	30.01	28.11	28.17
113	WG	30.32	28.11	29.01
15	GY	27.06	29.32	29.26
114	YG	30.40	29.32	29.95
115	IY	26.57	23.96	23.89
33	YI	28.88	23.96	24.22
24	LW	27.93	23.77	23.90
116	WL	28.02	23.77	24.37
117	LY	22.89	24.49	24.55
34	YL	27.19	24.49	24.88
25	MW	28.48	23.70	24.27
118	WM	25.26	23.70	24.63
26	PW	30.44	26.09	26.51
119	WP	28.44	26.09	27.09
27	RW	47.71	49.10	47.58
120	WR	46.96	49.10	47.78
28	VW	29.98	24.46	24.50
121	WV	28.68	24.46	25.07

122	VY	28.53	25.28	25.25
35	YV	29.07	25.28	25.66
123	WY	25.69	23.01	24.16
36	YW	25.66	23.01	24.01
39	FGG	28.75	26.72	28.34
124	GGF	28.20	26.72	27.49
125	FGFG	23.71	21.02	23.41
53	FGGF	22.16	21.02	23.25

Table 7. Experimental and MLR predicted retention time for tryptic digest of horse cytochrome C

No.	Peptide sequence	Experimental	MLR 1	MLR 2
		tr(min)	tr(min)	tr(min)
3	GGK	4.39	4.40	4.58
6	CAQCHTVEK	6.27	6.33	6.25
7	TGPNLHGLFGR	6.70	6.79	6.74
8	GDVEK	6.76	6.86	6.90
9	GITWK	6.98	6.95	6.94
10	YIPGTK	7.15	7.20	7.00
11	IFVQK	7.49	7.25	7.32
12	MIFAGIK	7.86	8.00	7.89
13	TGQAPGFTYTDANK	9.44	9.47	9.13
14	EDLIAYLK	9.44	9.87	9.24
15	ATNE	9.51	10.20	9.57
16	EETLMEYLENPK	10.56	10.70	10.21

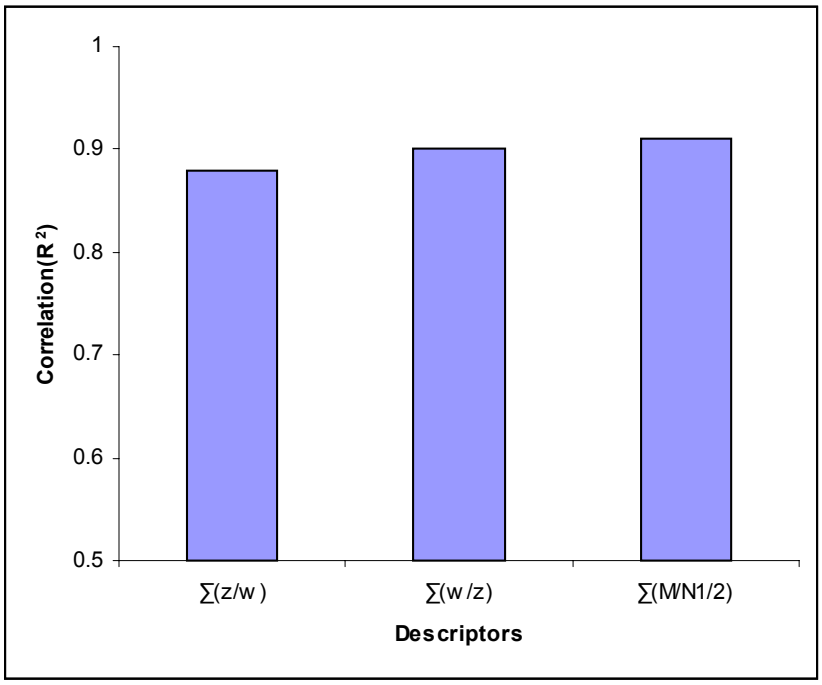

Fig. 1 Correlation between the experimental mobilities and the calculated values using $Q/M^{2/3}$, $E_{s,c}$ and one of the three "sequence-related" descriptors

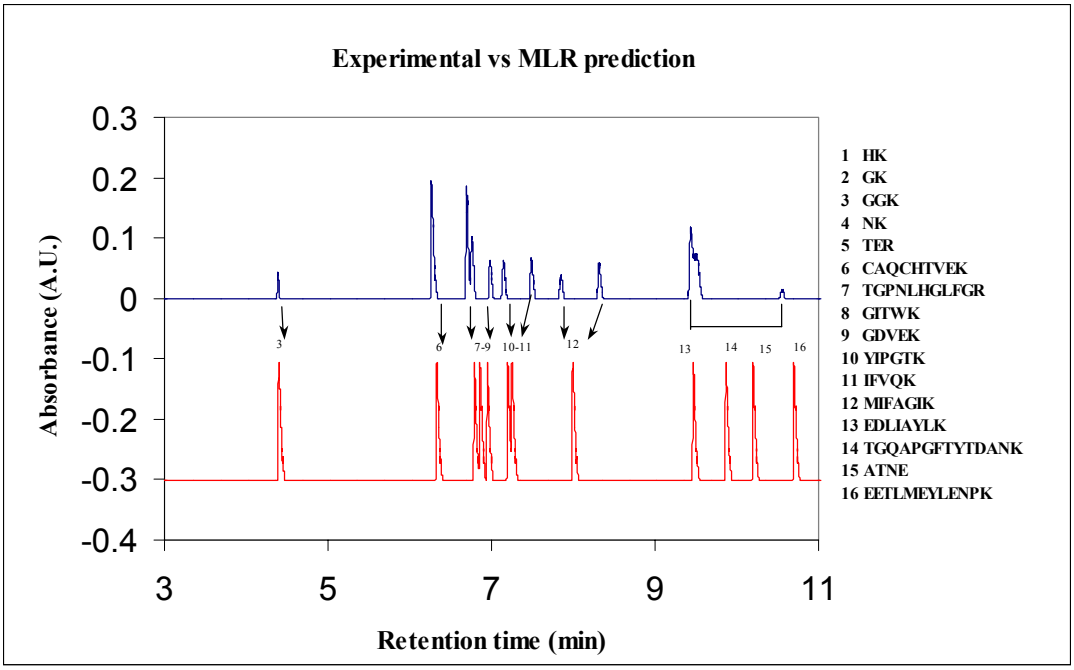

Fig. 2a Comparison of experimental and MLR1 predicted maps of tryptic digest of horse cytochrome C

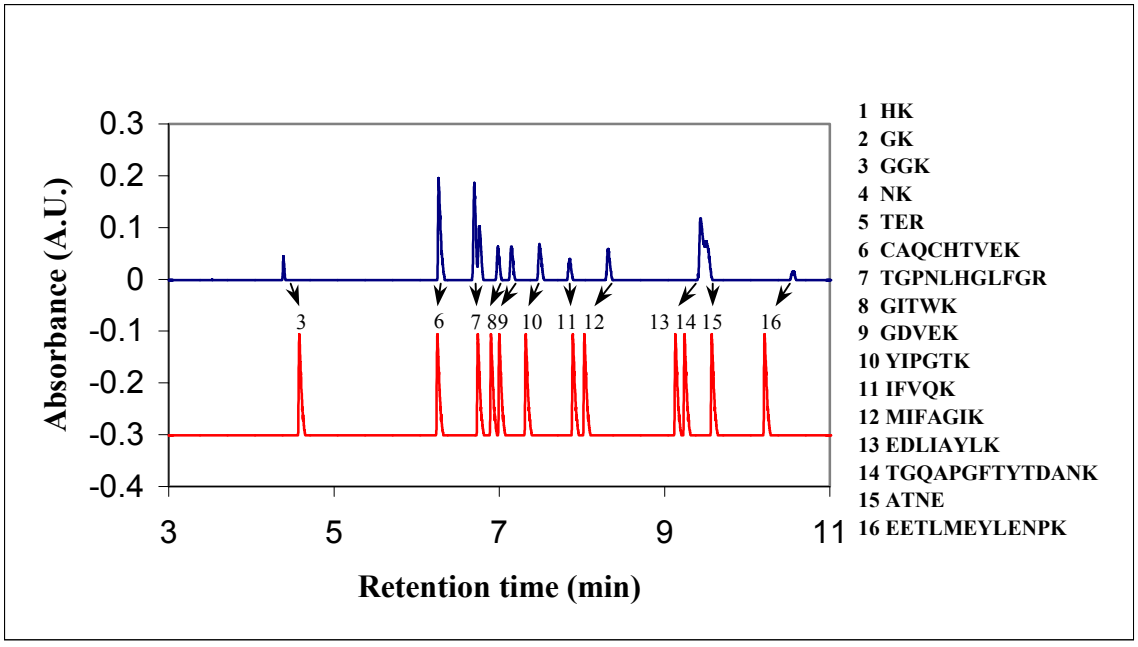

Fig. 2b Comparison of experimental and MLR2 predicted maps of tryptic digest of horse cytochrome C

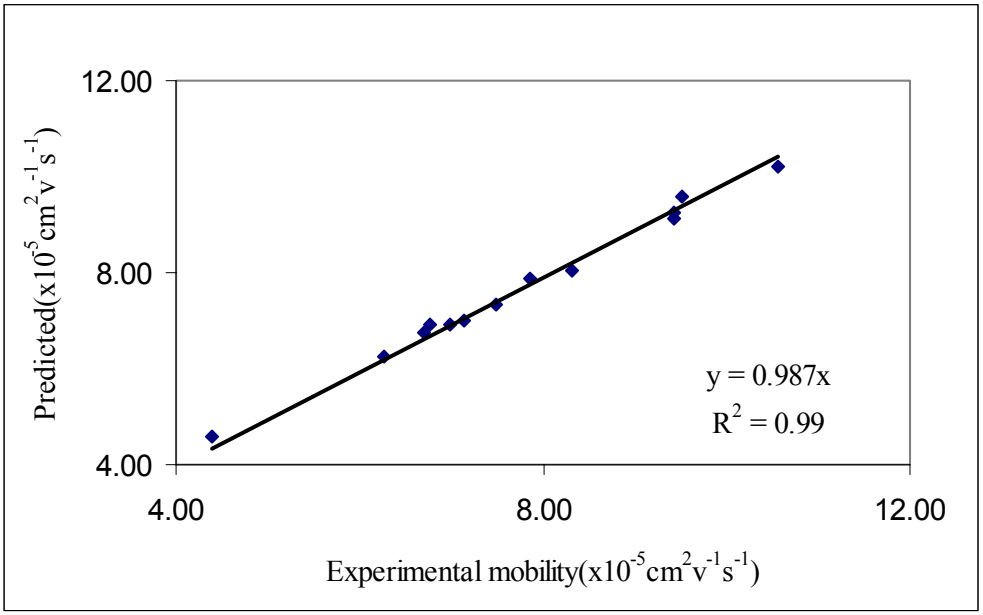

Fig.3 Plot of experimental mobility vs MLR2 predicted mobility of tryptic digest of horse cytochrome C

Chapter 5 Electrophoretic mobility prediction of peptides in Tween20 system using Quantitative Structure-Migration Relationships (QSMR) models

Keywords: Tween20, Peptide, Mobility, MEKC, QSMR

1. Abstract

The aim of this work was to develop a Quantitative Structure-Migration Relationships(QSMR) model for the prediction of peptides' mobilities in Tween20 system. Electrophoretic mobility data of 114 peptides ranging in size from 2 to 13 amino acid residues were collected in this system.

A small number of peptides with different amino acid residues were used to determine individual contribution of each amino acid residue and N-, C-terminal to the electrophoretic mobility of the whole peptide using group contribution approach(GCA). The sum of the individual contribution was calculated for other peptides and used as a new descriptor in predicting their mobilities in 10 mM Tween20 system. A multi-linear regression(MLR) model was developed using this "contribution" descriptor ΣC_{AA}, S_{np}(nonpolar surface area) and "sequence-related" descriptor $\sum \frac{m}{\sqrt{n}}$. The values of these three descriptors were also used as inputs in developing artificial neural networks(ANN). Both MLR model and ANN model can accurately predict the electrophoretic mobilities for other peptides in the data set of 10mM Tween20 system. ANN did not bring big improvement of the prediction accuracy , which indicates that nonlinearity is not significant in predicting peptide's mobility in 10mM Tween20 system.

Same method and descriptors were used on the peptide mobility data in 50mM Tween20 system. It was observed that the same MLR model was not the best for 50mM Tween20 system. New descriptors need to be introduced into the MLR model to improve the prediction accuracy. In this system, ANN did not improve the prediction accuracy dramatically either.

2. Introduction

Micellar electrokinetic chromatography(MEKC) is a CE technique that can separate solutes based on their different partitioning into the micelle pseudostationary phase. It can be viewed as the hybrid of reveres-phase liquid chromatography(RPLC) and capillary zone electrophoresis(CZE)[1]. MEKC has been widely applied to peptide separations due to its unique features such as high efficiency, short analysis time, small sample size and little solvent consumption. Another big advantage of MEKC over conventional chromatographic techniques is the flexibility and ease of changing the chemical composition of the pseudostationary phase. Old pseudostationary phase can be easily replaced by simply rinsing the capillary with the new pseudostationary phase. It will greatly facilitate method development for the separation of complex mixtures.

Tween20 is a nonionic surfactant that can offer different selectivity from anionic surfactants like SDS. Since it does not increase the conductivity of the buffer, it can be used at high concentrations. Unlike in capillary zone electrophoresis(CZE), solutes will be separated based on their different partitioning into micelles in Tween20 system.

Many research groups have studied the migration of peptides in capillary zone electrophoresis and various theoretical or semi-empirical models have been developed to predict peptides' electrophoretic mobilities in CZE[2-8]. Most of these models relate peptide mobility to its charge(Q), molar mass(M), number of residues in the sequence(N) or average residue mass(m). In our previous work, a QSMR model was developed for the same purpose using artificial neural network(ANN) technique[9]. The values of three descriptors namely Offord's charge over mass term($Q/M^{2/3}$), corrected steric substituent

constant($E_{s,c}$) and molar refractivity(MR) were used as inputs for the neural network. With our QSMR model, accurate prediction of peptides' electrophoretic moilities in CZE was achieved.

To the best of our knowledge, no electrophoretic mobility prediction model has ever been developed for peptides in MEKC systems. In this work, experimental data of 114 standard peptides were collected in 10mM and 50mM Tween20 systems and electrophoretic mobility of each peptide was calculated. All the 114 peptides were randomly divided into three groups: training set(72 peptides), test set(18 peptides) and validation set(14 peptides). All the peptides, their mobilities and descriptor values were listed in Table 1. Data of peptides in the training set were used to generate the multi-linear regression(MLR) models. Many descriptors reported by other research groups were tried in developing the best MLR model including number of amino acid residues, molar mass, average residue mass, peptide net charge, surface area, hydrophobicity scale, isoelectric point value, corrected steric substituent constant, strain parameter, Z-scales, alpha-helix content etc. The descriptors in the best MLR model were used as inputs in developing an artificial neural network(ANN) to predict peptide mobility. Data in the test set and validation set were used to test the robustness of our prediction model. Good results have been achieved in 10mM Tween20 system.

3. Experimental

3.1 Chemicals and Materials.

A total of 114 standard peptides were purchased from either Sigma Chemical Co. (St. Louis, MO) or Bachem Co. (Torrance, CA). Tween20, decanophenone, sodium

hydroxide, phosphoric acid, were obtained from Sigma Chemical Co. (St. Louis, MO). 50mM phosphate buffer solution was prepared by adjusting the pH of 50mM phosphoric acid solution to 2.5 using 0.1M NaOH. Tween20 solutions were then prepared in 50mM phosphate buffer(pH2.5). All solutions were filtered through 0.2 μm Acrodisc filter(STRL, Eatontown, NJ) before use.

3.2 Instrument

A Beckman PACE5000 CE instrument with a UV detector was used in this work. The detection wavelength was set at 214 nm and 37 °C was used as the separation temperature. The voltage was set at 20 kV. Bare silica capillary was used in this study. The total length of the capillary was 37cm and the length from inlet to the detection point was 30cm. Before any injection, the bare-silica capillary was conditioned by the following rinsing procedure using high pressure(20 PSI): MilliQ water for 20 minutes**,** NaOH/Methanol solution for 10 minutes, MilliQ water for 20 minutes and Tween20 solution for 15 minutes. Between injections, the capillary was pressure- rinsed with Tween20 solution for 5 minutes.

4. Results and Discussion

4.1 Determination of peptides' electrophoretic mobilities

In this work, electrophoretic mobilities of peptides were determined by Eq.1

$$\mu_{ef} = \frac{L_t L_d}{V}\left(\frac{1}{t_r} - \frac{1}{t_{eo}}\right) \qquad Eq.1$$

where L_t is the total length of the capillary, L_d is the length from the capillary inlet to the detection point, V is the applied voltage, t_r is the peptide retention time and t_{eo} is the retention time of a solute that co-elute with the Tween20 micelle(decanophenone).

For 10mM Tween20 system, the electrophoretic mobilities of peptides range from 1.306 to 37.872 x $10^{-5}cm^2v^{-1}s^{-1}$, while electrophoretic mobility values range from 2.082 to 28.406 x $10^{-5}cm^2v^{-1}s^{-1}$ in 50mM Tween20 system.

4.2 Individual contribution of each amino acid residue and N-,C-terminal to the mobility of the whole peptide

Group contribution approach(GCA) was used in this work to determine the individual contribution of each amino acid residue and N-, C-terminal to the electrophoretic mobility of the whole peptide in Tween20 systems. Electrophoretic mobility of a peptide was considered as the sum of individual contribution of each amino acid residue in the sequence and contribution from the N-, C-terminal of the peptide. A small number of peptides in the data set were used for this purpose. Data of these peptides in each Tween20 system were used to determine individual contribution to peptide mobility from each amino acid residue and the N-, C- terminal in that system. Multi-linear regression(MLR) function of software Minitab[10] was used to determine the individual contributions. In this case, electrophoretic mobilities were treated as response, the numbers of each amino acid residue were used as independent descriptors and the intercept of the MLR equation was considered the collective contribution of both N- and C-terminal.

The individual contributions for both 10mM Tween20 and 50mM Tween20 systems were shown in Table 2. It was noticed that only for Arginine(R), Histidine(H) and Lysine(K), the individual contribution values were big positive values for both Tween20 systems. It is because at pH 2.5, only these three residues will contribute one positive charge and greatly raise the charge-size ratio. They will increase peptide

mobility although the bulkiness of the peptide is also increased. The collective contribution of both N- and C-terminal of a peptide is also a big positive number in both systems since the net charge of these two terminals is 0.83 at pH2.5. For the rest of the amino acid residues, their contributions are either close to zero or a negative value. These residues carry zero or negative charge at pH2.5, and they will decrease the charge-size ratio of the whole peptide in Tween20 system.

For both Tween20 systems, charge-size ratio is important in determining the electrophoretic mobility of a peptide, which is similar to the situation of capillary zone electrophoresis(CZE). But Tween20 systems are MEKC systems and the retention mechanisms are different from CZE system. It is noticed that individual contribution of some amino acid residues changes relatively big between these two Tween20 systems which indicates that even the retention mechanisms of these two Tween20 systems are different.

4.3 Reproducibility of experiment.

Every 5-6 peptides were mixed together(200 μl each) to form a peptide mixture solution. 10 μl of diluted decanophenone was added into each peptide mixture to form a peptide sample. Identity of each peptide peak was confirmed by matching the retention time with the individual retention time of the peptides in the mixture. Each peptide sample was injected three times and excellent reproducibility was achieved. The RSD of the electrophoretic mobilities of peptides ranges from 0 to 1.07%, with an average of 0.26% for the whole data set.

4.4 Multi-linear regression(MLR) study

The sum of the individual contribution from each amino acid residue and both termini was calculated for each peptide in the data set for both 10mM Tween20 system and 50mM Tween20 system. It was used as one descriptor(ΣC_{AA}) in developing the best MLR model for the prediction of peptide mobility in respective Tween20 system. All previous reported descriptors by our group and other research groups were used together with contribution descriptor ΣC_{AA} as inputs for MLR analysis. The stepwise function of software Minitab[10] was used for MLR model generation. This procedure was repeated until all descriptors were included into the model. The correlation function of Minitab was then used to develop the correlation matrix and remove the descriptors that are highly correlated with other descriptors but contribute less to peptide mobility. The best MLR model was chosen based on correlation coefficient, F test and standard error. The best MLR model for 10mM Tween20 system was consisted of three descriptors: contribution descriptor ΣC_{AA}, nonpolar surface area S_{np}[11] and sequence-related descriptor $\sum \frac{m}{\sqrt{n}}$ [12] as in Eq.2:

$$\mu = a\,\Sigma C_{AA} + b\,S_{np} + c\,\sum \frac{m}{\sqrt{n}} \qquad \text{Eq. 2}$$

where n is the position number of a amino acid residue in the sequence and m is the molar mass of the amino acid at position n.

The specifications for this MLR model and the mean effect of each descriptor are shown in Table 3. The correlation matrix of these three descriptors is shown in Table 4. The MLR calculated electrophoretic mobilities are listed in Table 1 for all peptides in the data set.

ΣC_{AA} shows the largeset mean effect of 18.18, while the mean effect of S_{np} and $\sum \frac{m}{\sqrt{n}}$ are relatively small(-1.06 and 0.62 respectively). The electrophoretic mobility decreases as the nonpolar surface area increases. The bigger the nonpolar surface area, the stronger the peptide will interact with the Tween20 micelles, which will reduce its electrophoretic mobility. The mean effect of $\sum \frac{m}{\sqrt{n}}$ agrees well with our previous study[12] while its value is much smaller than the corresponding value in CZE. It also indicates that it will be more difficult to separate peptides with same amino acid composition but different sequence in Tween20 system.

With this MLR model, R^2 of 0.953 is achieved for the training set. To explore the nonlinear characteristic of peptide's electrophoretic mobility in Tween20 system, the values of these three descriptors are used as inputs to develop artificial neural network in this work.

4.5 Artificial neural network analysis

Artificial neural networks are developed and optimized in this work to explore the nonlinear characteristic of peptide's electrophoretic mobility in Tween20 system. The electrophoretic mobilities are used as target values and the values of the three descriptors in MLR model are used as inputs. The specifications of the final ANN are listed in Table 5. The comparison of this ANN model with the best MLR model is given in Table 6 and Fig 1&2. It's indicated that ANN brings only a small improvement to the mobility prediction accuracy for 10mM Tween20 system. The electrophoretic mobilities of peptides do not have obvious nonlinear characteristic in this system.

4.6 Analysis of 50mM Tween20 system

Same descriptors and methods are used to analyze peptide mobility data of 50mM Tween20 system. It's noticed that the best MLR model for 10mM Tween20 system is not the best for 50mM Tween20 system. With this MLR model, R^2 of only 0.822 is achieved for the same training set used in 10mM Tween20 system. Even with optimized ANN model, there's still no big improvement in prediction accuracy(R^2 = 0.828).

Through the data analysis of 50mM Tween20 system, it shows that the current MLR model can't accurately predict the electrophoretic mobilities of peptides in 50mM Tween20 system. New descriptors need to be used in developing a good MLR model for the accurate mobility prediction in this system. The electrophoretic mobilities of peptides do not have obvious nonlinear characteristic in this system either.

5. Conclusions

In this work, a new Quantitative Structure-Migration Relationships(QSMR) model is developed for peptide mobility prediction in 10mM Tween20 system using a data set of 114 standard peptides. A new contribution descriptor ΣC_{AA} is first developed using the mobility data of a small number of peptides. ΣC_{AA}, together with nonpolar surface area S_{np} and sequence-related descriptor $\sum \frac{m}{\sqrt{n}}$ forms the best MLR model in predicting peptide mobility in 10mM Tween20 system. An artificial neural network is also developed to explore the nonlinear characteristic of peptide mobility in this system. With this ANN model, there's no big improvement in prediction accuracy. No obvious nonlinear characteristic of peptide mobility is observed in this Tween20 system.

Same method is used to analyze peptide mobility data in 50mM Tween20 system. The MLR model can't accurately predict peptide mobility in this system. New descriptor needs to be used in developing a good MLR model for this system. The optimized ANN model doesn't show big improvement in prediction accuracy. Peptide mobility does not have obvious nonlinear characteristic in this system either.

It's concluded that there's some change in the mechanism of peptides' retention between these two Tween20 systems.

References.

[1] Khaledi, M.G., High performance capillary electrophoresis – theory, techniques and applications, Wiley-interscience. 1998.

[2] Grossman, P. D., Colburn, J. C., Lauer, H. H., Anal. Biochem. 1989, 179(1), 28-33.

[3] Offord, R. E., Nature (London) 1966, 211, 591-593.

[4] Compton, B. J., J. Chromatogr. 1991, 559, 357-366.

[5] Adamson, N. J., Reynolds, E. C., J. Chromatogr. B 1997, 699, 133-147.

[6] Messana, I., Rossetti, D. V., Cassiano, L., Misiti, F., Giardina, B., Castagnola, M., J. Chromatogr. B 1997, 699, 149-171.

[7] Kasicka, V., Electrophoresis 1999, 20, 3084-3105.

[8] Cifuentes, A., Poppe, H., Electrophoresis 1997, 18, 2362-2376.

[9] Jalali-Heravi, M., Shen, Y., Hassanisadi, M. and Khaledi, M.G., Manuscript submitted to Electrophoresis.

[10] Minitab Release 12, http://www.minitab.com

[11] http://roselab.jhu.edu/utils/unfolded.html

[12] Shen, Y., Jalali-Heravi, M. and Khaledi, M.G., Manuscript in preparation

Table 1. The experimental, MLR and ANN calculated values of electrophoretic mobility ($cm^2\ v^{-1}s^{-1}$) of standard peptides in 10mM Tween20 system together with the values of descriptors appearing in the models

		Descriptors			Mobility				
No.	Peptide sequence	ΣC_{AA}	S_{np}	$\sum \frac{m}{\sqrt{n}}$	Exp	MLR		ANN	
Training set					μ_{ef} x 10^5	μ_{ef} x 10^5	% Dev	μ_{ef} x 10^5	% Dev
1	AF	19.645	84.60	205.90	19.094	19.545	2.36	18.926	-0.88
2	A_LY_L	19.182	84.90	217.21	19.077	19.081	0.02	18.418	-3.45
3	AY	19.182	84.90	217.21	18.122	19.081	5.30	18.477	1.96
4	D_LF_L	14.801	38.40	249.91	13.362	15.252	14.15	14.119	5.67
5	DW	12.348	37.40	277.51	11.670	12.776	9.47	11.646	-0.21
6	EW	16.125	37.40	291.54	15.043	16.712	11.10	16.158	7.41
7	F_LF_L	15.991	76.80	282.00	16.170	16.007	-1.01	14.917	-7.75
8	FF	15.991	76.80	282.00	16.413	16.007	-2.47	14.922	-9.08
9	FG	19.596	38.40	218.27	20.881	20.160	-3.45	19.492	-6.65
10	FI	18.482	69.30	257.94	18.351	18.647	1.61	17.676	-3.68
11	FL	17.602	73.70	257.94	18.233	17.675	-3.06	16.874	-7.45
12	FM	18.253	77.00	270.70	17.918	18.326	2.28	17.980	0.35
13	FV	19.445	74.50	248.03	19.252	19.554	1.57	19.345	0.48
14	FW	13.538	75.80	309.60	12.155	13.531	11.32	13.205	8.64
15	GF	19.596	38.40	191.88	19.309	20.113	4.16	20.336	5.32
16	GW	17.143	37.40	219.48	17.587	17.636	0.28	17.604	0.10
17	HW	26.171	37.40	299.57	28.789	27.129	-5.77	29.872	3.76
18	IF	18.482	69.30	247.98	17.541	18.629	6.21	18.562	5.82
19	IW	16.029	68.30	275.58	14.996	16.153	7.72	15.968	6.48
20	KW	26.474	37.40	290.60	31.604	27.426	-13.22	30.129	-4.67
21	KF	28.927	38.40	263.00	33.477	29.902	-10.68	33.111	-1.09
22	LF	17.602	73.70	247.98	17.924	17.657	-1.49	17.311	-3.42
23	LW	15.149	72.70	275.58	14.642	15.181	3.68	14.680	0.26
24	LY	17.139	74.00	259.29	17.042	17.194	0.89	16.889	-0.90
25	MW	15.799	76.00	293.62	15.089	15.840	4.98	15.715	4.15
26	PW	15.528	37.40	259.54	17.202	16.036	-6.78	15.941	-7.33
27	RW	26.948	37.40	318.61	30.515	27.967	-8.35	30.824	1.01
28	VW	16.992	73.50	261.56	16.409	17.053	3.92	16.804	2.40
29	VY	18.982	74.80	245.27	17.595	19.066	8.36	19.004	8.01
30	WA	17.192	83.60	267.23	18.590	17.129	-7.86	17.267	-7.12
31	WD	12.348	37.40	298.35	14.851	12.814	-13.72	12.805	-13.78
32	WE	16.125	37.40	308.27	15.620	16.742	7.19	17.293	10.71
33	WF	13.538	75.80	321.04	12.467	13.551	8.69	13.438	7.79
34	WG	17.143	37.40	257.31	18.637	17.705	-5.00	17.872	-4.11
35	WM	15.799	76.00	309.74	15.431	15.869	2.84	15.942	3.32
36	WP	15.528	37.40	285.64	18.890	16.083	-14.86	16.335	-13.53
37	WV	16.992	73.50	287.07	18.136	17.099	-5.72	17.130	-5.55
38	WY	13.075	76.10	332.35	12.496	13.088	4.74	13.223	5.82

39	YA	19.182	84.90	244.19	19.539	19.130	-2.10	19.863	1.66
40	YG	19.133	38.70	234.27	19.192	19.705	2.67	20.412	6.35
41	YI	18.019	69.60	273.94	17.852	18.193	1.91	18.491	3.58
42	YL	17.139	74.00	273.94	17.776	17.220	-3.13	17.480	-1.66
43	YV	18.982	74.80	264.03	18.374	19.099	3.95	19.467	5.95
44	YW	13.075	76.10	325.60	11.918	13.076	9.72	13.159	10.41
45	YY	15.065	77.40	309.31	14.718	15.089	2.52	15.261	3.69
46	FGG	18.777	38.40	261.61	18.415	19.390	5.30	19.802	7.53
47	$G_LF_LL_L$	16.783	73.70	267.61	16.830	16.844	0.09	16.444	-2.29
48	WGG	16.324	37.40	300.65	17.352	16.935	-2.40	17.065	-1.65
49	WGY	12.256	76.10	361.92	12.168	12.293	1.03	12.060	-0.89
50	WWW	4.208	112.20	466.55	1.306	3.646	179.08	3.420	161.82
51	YAG	18.363	84.90	287.53	17.319	18.360	6.01	18.041	4.17
52	YYL	12.251	112.70	385.04	12.571	11.820	-5.97	13.289	5.72
53	YYY	10.177	116.10	413.92	11.065	9.677	-12.55	10.577	-4.41
54	FGGF	14.353	76.80	344.21	14.055	14.423	2.62	13.614	-3.14
55	GGFL	15.964	73.70	289.11	15.745	16.035	1.84	14.676	-6.79
56	GGFM	16.615	77.00	298.13	15.846	16.679	5.26	15.452	-2.48
57	WGGY	11.437	76.10	391.25	11.905	11.498	-3.41	10.883	-8.59
58	FFFFF	2.719	192.00	533.84	1.810	1.113	-38.52	1.869	3.28
59	FLEEI	11.994	104.60	475.11	11.161	11.829	5.98	12.976	16.26
60	TRSAW	22.928	83.60	438.85	22.928	23.378	1.96	21.776	-5.02
61	WGGGY	10.618	76.10	419.22	11.406	10.701	-6.18	10.448	-8.40
62	PPGFSP	13.322	38.40	416.48	13.028	14.021	7.62	14.291	9.69
63	RRPYIL	30.744	104.90	566.65	30.067	31.404	4.45	30.731	2.21
64	VEPIPY	10.344	105.70	478.70	10.640	10.111	-4.97	11.097	4.29
65	DRVYIHP	26.197	105.70	580.03	25.244	26.709	5.80	24.528	-2.83
66	RVYIHPI	29.878	136.60	593.20	28.727	30.113	4.82	28.686	-0.14
67	YPFVEPI	5.920	144.10	578.92	9.741	5.176	-46.86	8.438	-13.37
68	DRVYIHPF	21.773	144.10	638.43	23.836	21.698	-8.97	22.143	-7.10
69	RPPGFSPFR	26.870	76.80	636.38	26.968	27.910	3.49	27.195	0.84
70	DRVYIHPFHL	27.169	179.40	731.63	28.018	26.961	-3.77	28.727	2.53
71	SYSMEHFRWG	23.082	153.10	713.47	22.216	23.063	3.82	22.967	3.38
72	CGYGPKKKRKVGG	37.871	74.60	716.60	37.872	39.476	4.24	37.108	-2.02
Test set									
73	AW	17.192	83.60	233.50	16.972	17.068	0.57	15.964	-5.94
74	DF	14.801	38.40	249.91	12.993	15.252	17.39	14.632	12.62
75	FA	19.645	84.60	228.19	20.978	19.585	-6.64	18.902	-9.90
76	GY	19.133	38.70	203.19	18.009	19.649	9.11	19.501	8.28
77	HY	28.161	38.70	283.28	27.167	29.142	7.27	28.598	5.26
78	IY	18.019	69.60	259.29	17.164	18.166	5.84	17.635	2.74
79	L_LW_L	15.149	72.70	275.58	14.633	15.181	3.74	14.189	-3.03
80	VF	19.445	74.50	233.96	17.975	19.529	8.64	19.015	5.78
81	WL	15.149	72.70	296.98	15.451	15.219	-1.50	14.372	-6.98
82	WW	11.085	74.80	348.64	6.393	11.075	73.24	10.110	58.15
83	FFFF	7.143	153.60	459.96	4.793	6.096	27.17	5.787	20.75
84	FGFG	14.353	76.80	351.18	14.795	14.436	-2.43	13.725	-7.23
85	RFDS	24.815	38.40	420.40	25.042	25.928	3.54	26.858	7.25
86	YAGFL	11.126	158.60	428.78	11.754	10.094	-14.13	8.576	-27.04
87	RGPFPI	21.782	69.30	481.39	23.377	22.467	-3.89	23.719	1.46

88	PPGFSPFR	17.884	76.80	540.50	20.294	18.433	-9.17	19.621	-3.32
89	IARRHPYFL	33.759	189.50	668.84	33.022	33.531	1.54	30.993	-6.15
90	DRVYVHPFHL	28.132	184.60	725.36	27.367	27.875	1.85	28.065	2.55
Validation set									
91	K_LF_L	28.927	38.40	263.00	31.084	29.902	-3.80	32.664	5.08
92	WR	26.948	37.40	327.41	31.279	27.983	-10.54	30.157	-3.59
93	WS	18.989	37.40	278.54	17.585	19.654	11.77	19.117	8.71
94	FFF	11.567	115.20	377.37	10.651	11.063	3.87	12.154	14.11
95	GFL	16.783	73.70	267.61	16.812	16.844	0.19	15.597	-7.23
96	GGF	18.777	38.40	223.53	18.701	19.322	3.32	18.421	-1.50
97	YGG	18.314	38.70	277.61	17.694	18.935	7.01	18.295	3.39
98	YGGF	13.890	77.10	360.21	12.860	13.968	8.62	13.438	4.50
99	YGGFM	11.728	115.70	426.94	12.160	11.312	-6.97	13.151	8.15
100	YGGWL	8.624	111.40	438.39	9.206	8.178	-11.16	8.382	-8.95
101	RPPGFSP	22.308	38.40	519.91	23.002	23.512	2.22	23.640	2.78
102	RVYVHPF	28.350	149.30	599.05	28.200	28.365	0.58	28.116	-0.30
103	YGGFMRF	16.290	154.10	560.49	18.905	15.741	-16.74	23.695	25.34
104	NRVYVHPF	26.801	149.30	631.18	26.801	26.819	0.07	26.333	-1.74

Table 2. Individual contribution of each amino acid residue and N-, C-terminal to the electrophoretic mobility of peptide in Tween20 systems

Residue/Terminal	**Individual contribution**	
	10mM Tween20	50mM Tween20
A	-0.77	-0.64
R	8.99	7.05
N	-1.55	-4.41
D	-5.61	-5.14
C	-1.84	-2.22
Q	-1.92	-1.73
E	-1.84	-3.34
G	-0.82	-2.02
H	8.21	5.99
I	-1.93	-2.27
L	-2.81	-2.76
K	8.51	6.12
M	-2.16	-2.81
F	-4.42	-2.67
P	-2.43	-1.53
S	1.03	0.53
T	-4.28	0.16
W	-6.88	-4.69
Y	-4.89	-3.84
V	-0.97	-0.21
N- and C-terminal	24.84	17.66

Table 3. Specifications of the best MLR model for 10mM Tween20 system

Descriptor	Notation	Coefficient	Mean effect
Sum of individual contribution	ΣC_{AA}	1.035(±0.031)	18.184
Nonpolar surface area	S_{np}	-0.014(±0.007)	-1.061
Sequence-related descriptor	$\sum \frac{m}{\sqrt{n}}$	0.002(±0.001)	0.619
Constant		0.012(±0.007)	0.012

Table 4. Correlation matrix of the descriptors of the best MLR model

R^2	ΣC_{AA}	S_{np}	$\sum \frac{m}{\sqrt{n}}$
ΣC_{AA}	1		
S_{np}	-0.099	1	
$\sum \frac{m}{\sqrt{n}}$	0.239	0.697	1

Table 5. Specifications of the optimized ANN model for 10mM Tween20 system

Number of nodes in the input layer	3
Number of nodes in the hidden layer	6
Number of nodes in the output layer	1
Number of iterations	35500
Learning rate	0.3
Momentum	0.9

Table 6. Comparison of the best MLR model and the ANN model

Model	$R^2_{Training}$	R^2_{Test}	$R^2_{Validation}$	$SE_{Training}$	SE_{Test}	$SE_{Validation}$	F
MLR	0.955	0.944	0.957	1.343	1.586	1.526	1470
ANN	0.976	0.943	0.957	0.989	1.721	1.597	2785

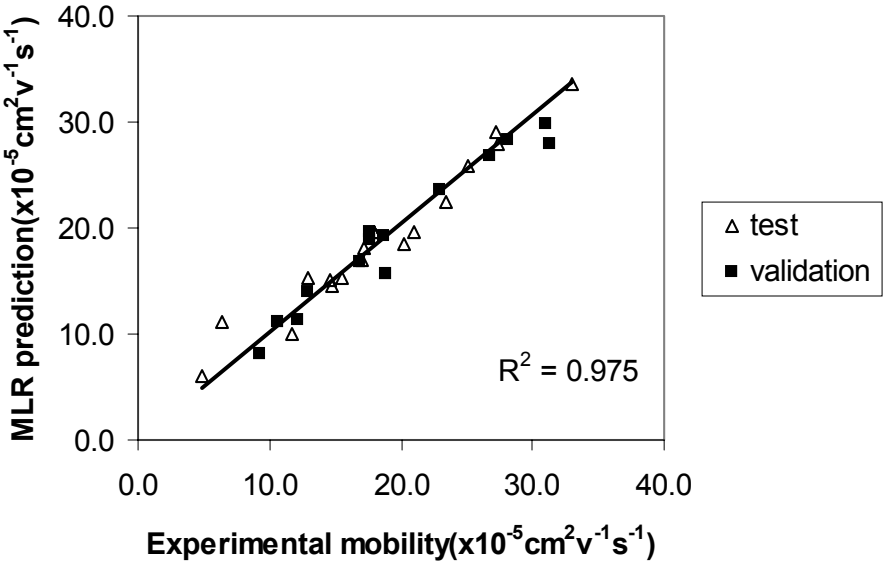

Fig 1. Plot of MLR calculated electrophoretic mobilities against experimental values for the test and validation sets in 10mM Tween20 system

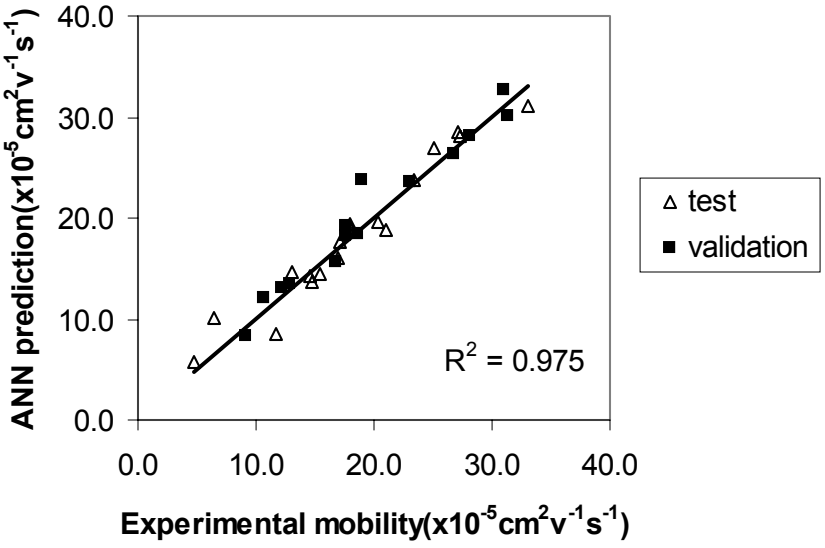

Fig 2. Plot of ANN calculated electrophoretic mobilities against experimental values for the test and validation sets in 10mM Tween20 system

Chapter 6 Future Trends

1. QSMR model development

In chapter 2-4, Quantitative Structure-Migration Relationships(QSMR) models have been developed and validated for the prediction of peptide mobilities in CZE based on the data of 125 peptides[1,2,3]. With these QSMR models and physicochemical parameters of new peptides, their electrophoretic mobilities in CZE can be predicted and subsequently converted into migration times in a certain CZE setup. This will greatly facilitate method development and reduce the time and effort that may take to optimize the separation of a complex peptide mixture.

Relatively accurate predictions had been achieved with our QSMR models. To improve the accuracy of these QSMR models, efforts need to be made in the following three aspects:

1. find a more accurate way to calculate the net charge of peptide

In our study, the net charge of a peptide was calculated as the sum of net charge of each amino acid and end group calculated by Henderson-Hassellbach equation[4]. If there are neighboring charged groups in a peptide, they will influence each other and change the net charge of the whole peptide. At pH higher than 2.5, there's even no acceptable method to calculate the net charge of a peptide. Quantum mechanics and commercial software may be needed to solve this problem.

2. further study on sequence descriptors

In chapter 4, a "sequence-related" descriptor was developed and used in QSMR model. It improves the prediction accuracy of peptides with same amino acid composition but different sequence. And it also improves the prediction accuracy when applied to predicting peptide map of tryptic digest of horse cytochrome C in CZE[3].

This is our first attempt to introduce a sequence descriptor into the peptide QSMR models. Better sequence descriptors that take into consideration of peptides' 3-D structures will be developed in our future study. With those descriptors, we might better understand how peptide sequence influences its migration in CZE.

3. Introduce a suitable hydrophobicity descriptor into QSMR model

As is widely believed, hydrophobicity also plays a role in peptides' migration in CZE. In our previous study, side-chain hydrophobicity descriptor π[5] was tried in developing QSMR models. The regression coefficients of this descriptor in developed QSMR models were very small. It means hydrophobicity was not an important factor for peptides' migration in CZE which is not true. This could be caused by the factor that hydrophobic peptides are less represented in our dataset.

In future studies, more hydrophobic peptides will be included into the data set. At the same time, other hydrophobicity descriptors will be tried in developing QSMR models.

2. QSPR model development

Some preliminary results have been achieved in developing a QSPR model to predict the micelle-water partition coefficients of peptides in SDS/Hexanol systems[7]. With out QSPR models, we can calculate peptides' partition coefficients and their retention time in our MEKC systems. Due to the limited source of descriptors and lack of commercial descriptor-calculation software, the accuracy of our current QSPR models is not as good as our QSMR models.

To improve the performance of the QSPR models, more descriptors need to be found or developed. Also, more peptides need to be added into our dataset.

Data of peptides in LiPFOS system had been collected, and these data will be used to develop QSPR models in LiPFOS system.

3. Artificial Neural Networks

Artificial neural networks(ANNs) have self-learning capacity and they are able to model complex data without the need for a detailed understanding of the underlying phenomena[8-10]. They will improve the accuracy of the QSMR and QSPR models by incorporating some nonlinear characteristics of peptides' migration/partition in CE systems.

ANNs have been successfully applied to peptide mobility prediction in CZE[1,2]. In future study, ANNs will also be applied to QSPR models to improve prediction accuracy.

4. Multi-dimensional separation setup

In our group, a home-made 2-D separation system is under development. The first dimension of this system is a MKEC system. Sample fractions separated in the first dimension will be transfer to the second dimension(CZE) through a microvalve for further separation. With this 2-D system, we can generate 2-D peptide maps to test and validate our prediction models. And this 2-D system can be interfaced with Mass Spectrometry to form a 3-D separation system. Peak identities can be easily determined, and credibility of our prediction models will be greatly improved.

References

[1] Jalali-Heravi, M., Shen, Y., and Khaledi, M.G., To be published in Electrophoresis

[2] Jalali-Heravi, M., Shen, Y., and Khaledi, M.G., Manuscript in preparation.

[3] Shen, Y., Jalali-Heravi, M., and Khaledi, M.G., Manuscript in preparation.

[4] Skoog, B., Wichman, A., Trends. Anal. Chem. 1986, 5(4), 82-83

[5] Fauchere, J. L., and Pliska, V., Eur. J. Med. Chem. – Chim. Ther. 1983, 18, 369-375

[6] Jalali-Heravi, M., Shen, Y., and Khaledi, M.G., Manuscript in preparation.

[7] Jalali-Heravi, M., Shen, Y., and Khaledi, M.G., Manuscript in preparation.

[8] Haykin, S., Neural Network, Prentice Hall, Englewood Cliffs, NJ, 1994.

[9] Zupan, J., Gasteiger, J., Neural Networks in Chemistry and Drug Design, Wiley-VCH, Weinheim, 1999.

[10] Bose, N. K., Liang, P., Neural Network, Fundamentals, McGraw-Hill, New York, 1996.

Appendix 1: Matlab program for the calculation of "sequence-related" descriptors

```
function cal
clc
clear all
Gr = 1.07;        % for Gly
Ar = 15.09;       % for Ala
Vr = 43.15;       % for Val
Lr = 57.17;       % for Leu
Ir = 57.17;       % for Ile
Fr = 91.19;       % for Phe
Yr = 107.19;      % for Tyr
Wr = 130.23;      % for Trp
Mr = 75.21;       % for Met
Sr = 31.09;       % for Ser
Rr = 100.20;      % for Arg
Kr = 72.19;       % for Lys
Hr = 81.16;       % for His
Dr = 59.10;       % for Asp
Er = 73.13;       % for Glu
Cr = 47.15;       % for Cys
Qr = 72.15;       % for Gln
Nr = 58.12;       % for Asn
Pr = 41.13;       % for Pro
Tr = 45.12;       % for Thr

Gm = 75.07;       % for Gly
Am = 89.09;       % for Ala
Vm = 117.15;      % for Val
Lm = 131.17;      % for Leu
Im = 131.17;      % for Ile
Fm = 165.19;      % for Phe
Ym = 181.19;      % for Tyr
Wm = 204.23;      % for Trp
Mm = 149.21;      % for Met
Sm = 105.09;      % for Ser
Rm = 174.20;      % for Arg
Km = 146.19;      % for Lys
Hm = 155.16;      % for His
Dm = 133.10;      % for Asp
Em = 147.13;      % for Glu
Cm = 121.15;      % for Cys
Qm = 146.15;      % for Gln
Nm = 132.12;      % for Asn
```

```
Pm = 115.13;      % for Pro
Tm = 119.12;      % for Thr

S =input('sequence ');
Slength = length(S);
tmp1=0;
tmp2=0;
y1=0;
y2=0;
y3=0;
y4=0;
y5=0;
y6=0;
y7=0;
y8=0;

for is=1:Slength
  switch S(is)
      case 'A',tmp1=Ar,tmp2=Am;
      case 'G',tmp1=Gr,tmp2=Gm;
      case 'V',tmp1=Vr,tmp2=Vm;
      case 'L',tmp1=Lr,tmp2=Lm;
      case 'I',tmp1=Ir,tmp2=Im;
      case 'F',tmp1=Fr,tmp2=Fm;
      case 'Y',tmp1=Yr,tmp2=Ym;
      case 'W',tmp1=Wr,tmp2=Wm;
      case 'M',tmp1=Mr,tmp2=Mm;
      case 'S',tmp1=Sr,tmp2=Sm;
      case 'R',tmp1=Rr,tmp2=Rm;
      case 'K',tmp1=Kr,tmp2=Km;
      case 'H',tmp1=Hr,tmp2=Hm;
      case 'D',tmp1=Dr,tmp2=Dm;
      case 'E',tmp1=Er,tmp2=Em;
      case 'C',tmp1=Cr,tmp2=Cm;
      case 'Q',tmp1=Qr,tmp2=Qm;
      case 'N',tmp1=Nr,tmp2=Nm;
      case 'P',tmp1=Pr,tmp2=Pm;
      case 'T',tmp1=Tr,tmp2=Tm;

  end

  y1=y1+tmp1/sqrt(is);
  y2=y2+tmp1/(is);
  y3=y3+tmp1/sqrt(3*is-1);
  y4=y4+tmp1/(3*is-1);
```

```
   y5=y5+tmp2/sqrt(is);
   y6=y6+tmp2/(is);
   y7=y7+tmp2/sqrt(3*is-1);
   y8=y8+tmp2/(3*is-1);
end


[y1,
y2,
y3,
y4,
y5,
y6,
y7,
y8]



function sequence
clc
clear all
Gr = 1.07;          % for Gly
Ar = 15.09;         % for Ala
Vr = 43.15;         % for Val
Lr = 57.17;         % for Leu
Ir = 57.17;         % for Ile
Fr = 91.19;         % for Phe
Yr = 107.19;        % for Tyr
Wr = 130.23;        % for Trp
Mr = 75.21;         % for Met
Sr = 31.09;         % for Ser
Rr = 100.20;        % for Arg
Kr = 72.19;         % for Lys
Hr = 81.16;         % for His
Dr = 59.10;         % for Asp
Er = 73.13;         % for Glu
Cr = 47.15;         % for Cys
Qr = 72.15;         % for Gln
Nr = 58.12;         % for Asn
Pr = 41.13;         % for Pro
Tr = 45.12;         % for Thr

Gm = 75.07;         % for Gly
Am = 89.09;         % for Ala
Vm = 117.15;        % for Val
Lm = 131.17;        % for Leu
```

```
Im = 131.17;       % for Ile
Fm = 165.19;      % for Phe
Ym = 181.19;      % for Tyr
Wm = 204.23;     % for Trp
Mm = 149.21;     % for Met
Sm = 105.09;      % for Ser
Rm = 174.20;      % for Arg
Km = 146.19;      % for Lys
Hm = 155.16;      % for His
Dm = 133.10;      % for Asp
Em = 147.13;      % for Glu
Cm = 121.15;      % for Cys
Qm = 146.15;      % for Gln
Nm = 132.12;      % for Asn
Pm = 115.13;      % for Pro
Tm = 119.12;      % for Thr

S =input('sequence ');
Slength = length(S);
tmp1=0;
tmp2=0;
y1=0;
y2=0;
y3=0;
y4=0;
y5=0;
y6=0;
y7=0;
y8=0;

for is=1:Slength
  switch S(is)
      case 'A',tmp1=Ar,tmp2=Am;
      case 'G',tmp1=Gr,tmp2=Gm;
      case 'V',tmp1=Vr,tmp2=Vm;
      case 'L',tmp1=Lr,tmp2=Lm;
      case 'I',tmp1=Ir,tmp2=Im;
      case 'F',tmp1=Fr,tmp2=Fm;
      case 'Y',tmp1=Yr,tmp2=Ym;
      case 'W',tmp1=Wr,tmp2=Wm;
      case 'M',tmp1=Mr,tmp2=Mm;
      case 'S',tmp1=Sr,tmp2=Sm;
      case 'R',tmp1=Rr,tmp2=Rm;
      case 'K',tmp1=Kr,tmp2=Km;
      case 'H',tmp1=Hr,tmp2=Hm;
      case 'D',tmp1=Dr,tmp2=Dm;
```

```
        case 'E',tmp1=Er,tmp2=Em;
        case 'C',tmp1=Cr,tmp2=Cm;
        case 'Q',tmp1=Qr,tmp2=Qm;
        case 'N',tmp1=Nr,tmp2=Nm;
        case 'P',tmp1=Pr,tmp2=Pm;
        case 'T',tmp1=Tr,tmp2=Tm;

    end

    y1=y1+1/(tmp1/sqrt(is));
    y2=y2+1/(tmp1/(is));
    y3=y3+1/(tmp1/sqrt(3*is-1));
    y4=y4+1/(tmp1/(3*is-1));

    y5=y5+1/(tmp2/sqrt(is));
    y6=y6+1/(tmp2/(is));
    y7=y7+1/(tmp2/sqrt(3*is-1));
    y8=y8+1/(tmp2/(3*is-1));
end


[y1,
y2,
y3,
y4,
y5,
y6,
y7,
y8]
```