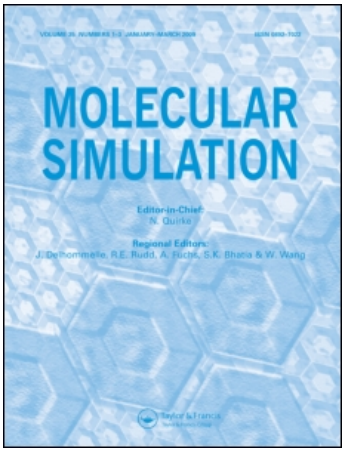


**Molecular Simulation**
Publication details, including instructions for authors and subscription information:
http://www.informaworld.com/smpp/title~content=t713644482

# Quantitative structure-mobility relationship studies of dipeptides in capillary zone electrophoresis using three-dimensional holographic vector of atomic interaction field

J. H. Jing[a]; G. Z. Liang[b]; H. Mei[b]; S. Y. Xiao[a]; Z. N. Xia[a]; Z. L. Li[ab]
[a] College of Chemistry and Chemical Engineering, Chongqing University, Chongqing, P.R. China [b] College of Bioengineering, Chongqing University, Chongqing, P.R. China

**To cite this Article** Jing, J. H. , Liang, G. Z. , Mei, H. , Xiao, S. Y. , Xia, Z. N. and Li, Z. L.(2009) 'Quantitative structure-mobility relationship studies of dipeptides in capillary zone electrophoresis using three-dimensional holographic vector of atomic interaction field', Molecular Simulation, 35: 4, 263 — 269
**To link to this Article: DOI:** 10.1080/08927020802512203
**URL:** http://dx.doi.org/10.1080/08927020802512203

*Molecular Simulation*
Vol. 35, No. 4, April 2009, 263–269

Taylor & Francis
Taylor & Francis Group

# Quantitative structure–mobility relationship studies of dipeptides in capillary zone electrophoresis using three-dimensional holographic vector of atomic interaction field

J.H. Jing[a], G.Z. Liang[b]*, H. Mei[b], S.Y. Xiao[a], Z.N. Xia[a] and Z.L. Li[ab]

[a]*College of Chemistry and Chemical Engineering, Chongqing University, Chongqing, P.R. China;* [b]*College of Bioengineering, Chongqing University, Chongqing, P.R. China*

(*Received 30 June 2008; final version received 29 September 2008*)

To explore the usefulness of empirical models and multivariate analysis techniques in predicting electrophoretic mobilities of dipeptides in capillary zone electrophoresis (CZE), a stepwise multiple regression-multiple linear regression (SMR-MLR) model has been built. The dataset consists of electrophoretic mobilities, measured at pH 2.5, for 53 dipeptides. Among the existing empirical models, the Offord model (i.e., $\mu = Q/M^{2/3}$) gives the best correlation for the dataset. A quantitative structure–mobility relationship (QSMR) has been developed using the Offord's charge-over-mass term as one descriptor combined with the three-dimensional holographic vector of atomic interaction field descriptors to account for the steric, hydrophobic and electrostatic interactions of the amino acid side chains. The MLR results of the dataset show an improvement in the predictive ability of the model over the simple Offord's relationship.

**Keywords:** CZE; electrophoretic mobility; 3D-HoVAIF; QSMR



## Abbreviations and Symbols

| | |
|---|---|
| CZE | capillary zone electrophoresis |
| QSMR | quantitative structure–mobility relationship |
| 3D-HoVAIF | three-dimensional holographic vector of atomic interaction field |
| MSC | molecular structural characterization |
| SMR | stepwise multiple regression |
| MLR | multiple linear regression |
| $\mu_{ef}$ | effective electrophoretic mobility |
| $Q$ | charge |
| $M$ | molecular mass |
| $R^2$ | correlation coefficient |
| $R^2_{CV}$ | cross-validation correlation coefficient |
| SD | standard deviation |
| $F$ | the Fisher's significance ratio |

## 1. Introduction

Capillary zone electrophoresis (CZE) is of much value for peptide separations due to several advantages, such as high efficiency, speed, small sample size, automation and high-throughput capability [1]. However, one of its characteristics is less noticed, that 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 offers a chance for predicting migration patterns from quantitative structure–mobility relationship (QSMR) that could promote optimisation of separations and method development.

In CZE, electrophoretic mobilities of charged compounds are described as Equation (1), which is a simplified equation valid only for small spherical rigid particles (small ions):

$$\mu_{ef} = \frac{q}{6\pi\eta r}, \tag{1}$$

where $\mu_{ef}$ is the effective electrophoretic mobility at a given ionic strength and temperature, $q$ the ion charge, $\eta$ the solution viscosity and $r$ the effective ion radius.

Several empirical models based on Stokes law for motion of ions in an electric field have been developed to explain the dependence of electrophoretic mobility on charge and size [2–14]. Generally, it has been shown that the electrophoretic mobility is proportional to the charge $Q$ and inversely proportional to the molecular mass $M$ as follows:

$$\mu = \frac{aQ}{M^b}, \tag{2}$$

where $a$ and $b$ are constants. Among them, the most common models are $Q/M^{1/3}$, $Q/M^{1/2}$ and $Q/M^{2/3}$.

On the other hand, Grossman et al. [2] considered peptides as classical linear polymers with $n$ amino acid residues and arrived at an equation that correlated mobility with a function in the form of $\ln(Q+1)/n^{0.43}$. Cifuentes and Poppe [15] 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 et al. [16] have

*Corresponding author. Email: gzliang@cqu.edu.cn

ISSN 0892-7022 print/ISSN 1029-0435 online

DOI: 10.1080/08927020802512203
http://www.informaworld.com

obtained the electrophoretic mobility of 58 peptides ranging in size from 2 to 39 amino acids and charge from 0.65 to 7.82. Heravi et al. [17] tested which of the abovementioned empirical models most accurately reproduces the experimental values of the electrophoretic mobility for the model peptides consisting of 125 peptides, ranging in size from 2 to 14 amino acids and charge from 0.743 to 5.843. Then the charge-over-mass term in the Offord model (i.e., $Q/M^{2/3}$) offered the best fit to their experimental data. But none of the models is able to reproduce adequately the electrophoretic mobility of small peptides. They concluded that although the Offord model gives the best overall mobility, it fails when applied to hydrophobic and highly charged peptides. These researchers also showed that electrophoretic mobilities of peptides cannot be successfully predicted with a reasonable degree of accuracy for all different categories of peptides by relying on two parameter models, namely charge ($Q$) and size dependence ($n$ or $M$) [16]. It seems that some other factors affect the mobility of peptides in CZE.

Recently, several new approaches to the investigation of the structure–mobility relationships of peptides have been developed. Fan et al. [18,19] predicted electrophoretic mobilities of peptides in CZE using the linear heuristic method, a nonlinear radial basis function neural network and support vector machines, in which descriptors were calculated by CODESSA; as a result, the performance of the nonlinear model was closer to the experimental data than that of the linear model. Deiber and co-workers [20,21] explored the evaluation of net charge, hydrodynamic size and shape of peptides through experimental electrophoretic mobilities obtained from CZE, and presented a model for peptide characterisation on the basis of well-established physicochemical equations. Allison et al. [22,23] developed a structure-based methodology based on fundamental electrohydrodynamic theory to characterise the charge and secondary structure of peptides using electrophoretic mobility and bead modelling. Yu and Cheng [24] studied electrophoretic mobility features of 102 peptides with a large range of size, charge and hydrophobicity using machine-learning techniques. These reports mentioned above are mainly focusing on nonlinear modelling methodology.

In the present work, a set of novel molecular descriptors previously developed [25] are further explored for the QSMR model. The main goals of the present work are: (1) to accurately predict the electrophoretic mobilities of bipeptides in CZE; (2) to achieve a better understanding of the physicochemical basis of the motion of a peptide in an electric field. To fulfil these goals, a dataset consisting of 53 dipeptides is chosen. In order to improve the predictive ability of the Offord model, we investigate QSMR models that incorporated other structural descriptors of peptides in addition to charge and size. Thus, in the QSMR models, the ratio of $Q/M^{2/3}$ is used as a hybrid descriptor. And some other suitable descriptors are chosen, which are expressed by three-dimensional (3D) holographic vector of atomic interaction field (3D-HoVAIF)[25], a newly developed molecular structural characterisation (MSC) method.

## 2. Methods and materials

In quantitative structure–activity relationship (QSAR) studies, MSC is the first step, which dedicates to minimise information loss during encoding information on topological or steric characteristics of molecular structure. MSC methods nowadays mainly include two kinds: one is based on molecular two-dimensional (2D) structures and the other on 3D configurations. For the 2D descriptors, since the first topological approach was proposed by Winer [26] in 1947, there have been many other singly parameterised methods based on molecular topological structures (e.g. Hosoya index [27], Randic index [28], Balaban index [29], etc.), achieving predictable performance on organic homologous physicochemical properties [30,31]. Nevertheless, 2D descriptors, incapable of reflecting molecular actual spatial structures and regardless of ligand–receptor interaction mechanisms, are virtually unable to construct effective QSAR models for drug and biological macromolecules. Because of this, 3D approaches appeared in the MSC field. In the 1980s, CoMFA was proposed by Cramer et al. [32] and then a number of similar methods based on molecular spatial structures such as CoMSIA [33], HASL [34], GRID [35] and COMPASS [36] have become the mainstay of current QSAR studies. Combining advantages of both 3D methods in explicit, accurate and explainable representation and 2D methods in easy and quick characterisation, 3D-HoVAIF, proposed by our laboratory, achieved satisfactory results in predicting electrophoretic mobilities of dipeptides in CZE.

### *2.1 Introduction of 3D-HoVAIF*

Atoms in organic molecules commonly include H, C, N, P, O, S, F, Cl, Br and I, they disperse in five families in periodic table of elements respectively, namely group IA, IVA, VA, VIA and VIIA. Based on ‘atoms of similar chemical properties belong to the same species’, atoms under consideration are naturally sorted for five according to their family in periodic table of elements. To further express molecular fine structures, the above-mentioned five atom types are continually classified for 10 in terms of their hybridisation state, which is believed to be the key factor to derive distinct chemical properties, and thus 55 atomic interaction types are ultimately obtained for a molecule (Table 1). Then three common potential energies, i.e. electrostatic, steric and hydrophobic interactions deemed to largely contribute to bioactivities,

Table 1. Ten types of atoms and the resulting 55 types of interactions in 3D-HoVAIF.

| No. | Atom styles | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | H | 1–1 | 1–2 | 1–3 | 1–4 | 1–5 | 1–6 | 1–7 | 1–8 | 1–9 | 1–10 |
| 2 | C($sp^3$) | | 2–2 | 2–3 | 2–4 | 2–5 | 2–6 | 2–7 | 2–8 | 2–9 | 2–10 |
| 3 | C($sp^2$) | | | 3–3 | 3–4 | 3–5 | 3–6 | 3–7 | 3–8 | 3–9 | 3–10 |
| 4 | C(sp) | | | | 4–4 | 4–5 | 4–6 | 4–7 | 4–8 | 4–9 | 4–10 |
| 5 | N($sp^3$), P($sp^3$) | | | | | 5–5 | 5–6 | 5–7 | 5–8 | 5–9 | 5–10 |
| 6 | N($sp^2$), P($sp^2$) | | | | | | 6–6 | 6–7 | 6–8 | 6–9 | 6–10 |
| 7 | N(sp), P(sp) | | | | | | | 7–7 | 7–8 | 7–9 | 7–10 |
| 8 | O($sp^3$), S($sp^3$) | | | | | | | | 8–8 | 8–9 | 8–10 |
| 9 | O($sp^2$), S($sp^2$) | | | | | | | | | 9–9 | 9–10 |
| 10 | F, Cl, Br, I | | | | | | | | | | 10–10 |

are involved in 3D-HoVAIF descriptors, ultimately giving rise to $55 \times 3 = 165$ interactions for an organic molecule.

Electrostatic interaction is expressed by Coulomb's law:

$$E_{mn}(E) = \sum_{i\in m, j\in n} \frac{e^2}{4\pi\varepsilon_0}\frac{Z_i Z_j}{r_{ij}} \quad (1 \le m \le 10, m \le n \le 10). \tag{3}$$

In Equation (3), $r_{ij}$ denotes interatomic Euclid distance, with the unit of Å ($1\,\text{Å} = 10^{-10}\,\text{m} = 0.1\,\text{nm}$); $e$ is the elementary charge ($1.6021892 \times 10^{-19}\,\text{C}$); $\varepsilon_0$ represents permittivity $8.85418782 \times 10^{-12}\,\text{C}^2/\text{J m}$ in vacuum; $Z$ represents amounts of Mülliken partial charges for atoms; $m$ and $n$ are atomic attributes. Electrostatic interactions among all the atoms included in a molecule could be given out by this equation, then accumulating them together into each of the 55 interaction items according to their atom-pair attributes.

Steric interaction describes interatomic spatial non-dipole–dipole or dipole-induced interactions, in 3D-HoVAIF, it is given by Lennard-Jones (Equation (4)). Among, $\varepsilon_{ij} = (\varepsilon_{ii}\varepsilon_{jj})^{1/2}$ is a potential well of atomic pairs, with its value taken from [37]; $D$ is corrected constant for interatomic interactions with the value of 0.01 [38]; $R^*_{ij} = (C_h R^*_{ii} + C_h R^*_{jj})/2$, is van der Waals' radius for modified atom-pair, with corrected factor $C_h$ of 1.00 in case of $sp^3$ hybridisation, 0.95 $sp^2$ hybridisation [39] and 0.90 sp hybridisation [38].

$$E_{mn}(S) = \sum_{i\in m, j\in n} \varepsilon_{ij} D\left[\left(\frac{R^*_{ij}}{r_{ij}}\right)^{12} - 2\left(\frac{R^*_{ij}}{r_{ij}}\right)^{6}\right] \quad (1 \le m \le 10, m \le n \le 10). \tag{4}$$

Hydrophobic interaction here is described as Equation (5), which is simply defined in Hint proposed by Kellogg et al. [40,41]. Amongst, $S$ is the solvent accessible surface area for atoms, indicating information on surface area when water molecule probes rolling its sphere at the atomic surface; $a$ is the atomic hydrophobic constant [42]; $T$ is the discriminant function [40], denoting entropic changing orientation in case of diverse interatomic interactions taking place.

$$E_{mn}(H) = \sum_{i\in m, j\in n} S_i a_i S_j a_j \mathrm{e}^{-r_{ij}} T_{ij} \quad (1 \le m \le 10, m \le n \le 10). \tag{5}$$

### 2.2 *Parameterisation of molecular structures*

Molecular steric structures of 53 dipeptides were firstly auto-constructed by Chem3D (e.g. Figure 1), and then optimised at AM1 level by MOPAC semi-experience quantum chemistry software in Chem3D. Then net electric charge of atoms was calculated in single-point form by Mulliken analysis. After the above two items were input, respectively, into the forms of Descartes coordinates and net electric charge amounts, 3D-HoVAIF descriptors were produced by applying 3D-HoVAIF.EXE, an applied programme written by our laboratory. Removing all the empty items, there were ultimately 78 3D-HoVAIF descriptors corresponding to a molecule.

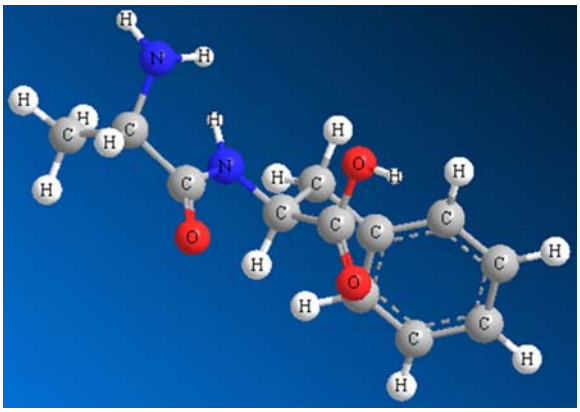

Figure 1. The steric structure of the 1st sample (AF).

### *2.3 Dataset and modelling*

The dataset obtained from [17] is listed in Table 2. After getting the 3D-HoVAIF descriptors, together with the ratio of $Q/M^{2/3}$, stepwise multiple regression-multiple linear regression (SMR-MLR) statistical analysis was used to generate the QSAR models. In all models, $R^2$ is

Table 2. The calculated electrophoretic mobilities, experimental values and errors.

| No.[a] | Peptide sequence | Exp $\mu_{ef} \times 10$ [$m^2 v^{-1} s^{-1}$] | Cal.[b] | Err.[b] | Cal.[c] | Err.[c] | Cal.[d] | Err.[d] |
|---|---|---|---|---|---|---|---|---|
| 1 | AF | 30.71 | 31.340 | 0.630 | 30.672 | −0.038 | 30.15 | −0.56 |
| 2 | AY | 29.58 | 29.035 | −0.545 | 29.787 | 0.207 | 31.01 | 1.43 |
| 3 | DF | 23.22 | 23.244 | 0.024 | 26.200 | 2.980 | 23.29 | 0.07 |
| 4 | DW | 22.38 | 22.477 | 0.097 | 25.073 | 2.693 | 21.89 | −0.49 |
| 5 | FA | 33.14 | 31.576 | −1.564 | 30.672 | −2.468 | 30.15 | −2.99 |
| 6 | FF | 27.91 | 27.720 | −0.190 | 27.640 | −0.270 | 25.35 | −2.56 |
| 7 | FG | 32.63 | 32.243 | −0.387 | 31.449 | −1.181 | 31.43 | −1.2 |
| 8 | FI | 30.1 | 29.064 | −1.036 | 28.852 | −1.248 | 25.47 | −4.63 |
| 9 | FL | 29.43 | 29.128 | −0.302 | 28.852 | −0.578 | 26.00 | −3.43 |
| 10 | FM | 27.8 | 28.359 | 0.559 | 28.138 | 0.338 | 25.69 | −2.11 |
| 11 | FV | 29.67 | 29.698 | 0.028 | 29.321 | −0.349 | 26.91 | −2.76 |
| 12 | FW | 24.65 | 26.753 | 2.103 | 26.513 | 1.863 | 24.06 | −0.59 |
| 13 | GW | 30.01 | 30.391 | 0.381 | 29.389 | −0.621 | 28.84 | −1.17 |
| 14 | GY | 27.06 | 29.860 | 2.800 | 30.472 | 3.412 | 30.13 | 3.07 |
| 15 | HW | 45.71 | 45.710 | 0.000 | 47.884 | 2.174 | 51.82 | 6.11 |
| 16 | IF | 28.11 | 29.128 | 1.018 | 28.852 | 0.742 | 25.47 | −2.64 |
| 17 | IW | 27.85 | 27.875 | 0.025 | 27.414 | −0.436 | 23.80 | −4.05 |
| 18 | LF | 28.12 | 29.170 | 1.050 | 28.852 | 0.732 | 26.00 | −2.12 |
| 19 | PW | 30.44 | 28.410 | −2.030 | 27.849 | −2.591 | 26.71 | −3.73 |
| 20 | RW | 47.71 | 47.244 | −0.466 | 46.745 | −0.965 | 50.40 | 2.69 |
| 21 | VF | 29.14 | 29.729 | 0.589 | 29.321 | 0.181 | 26.91 | −2.23 |
| 22 | VY | 28.54 | 27.647 | −0.893 | 28.655 | 0.115 | 25.96 | −2.58 |
| 23 | WA | 30.08 | 29.778 | −0.302 | 28.819 | −1.261 | 27.85 | −2.23 |
| 24 | WE | 25.78 | 26.950 | 1.170 | 26.470 | 0.690 | 24.02 | −1.76 |
| 25 | WF | 26.83 | 26.612 | −0.218 | 26.513 | −0.317 | 24.06 | −2.77 |
| 26 | WG | 30.32 | 30.243 | −0.077 | 29.389 | −0.931 | 28.84 | −1.48 |
| 27 | WL | 28.02 | 27.720 | −0.300 | 27.413 | −0.607 | 24.34 | −3.68 |
| 28 | WM | 25.26 | 27.126 | 1.866 | 26.907 | 1.647 | 24.24 | −1.02 |
| 29 | WP | 28.44 | 28.335 | −0.105 | 27.849 | −0.591 | 26.71 | −1.73 |
| 30 | WR | 46.96 | 47.520 | 0.560 | 46.745 | −0.215 | 50.40 | 3.44 |
| 31 | WS | 28.42 | 28.277 | −0.143 | 28.295 | −0.125 | 26.76 | −1.66 |
| 32 | WV | 28.68 | 28.293 | −0.387 | 27.883 | −0.797 | 25.07 | −3.61 |
| 33 | WW | 27.45 | 25.937 | −1.513 | 25.698 | −1.752 | 23.07 | −4.38 |
| 34 | WY | 25.69 | 24.990 | −0.700 | 26.158 | 0.468 | 23.49 | −2.2 |
| 35 | YA | 30.77 | 29.421 | −1.349 | 29.798 | −0.972 | 28.98 | −1.79 |
| 36 | YG | 30.4 | 30.007 | −0.393 | 30.472 | 0.072 | 30.13 | −0.27 |
| 37 | YI | 28.88 | 27.021 | −1.859 | 28.082 | −0.798 | 24.59 | −4.29 |
| 38 | YL | 27.19 | 27.085 | −0.105 | 28.081 | 0.891 | 25.13 | −2.06 |
| 39 | YV | 29.07 | 27.755 | −1.315 | 28.655 | −0.415 | 25.96 | −3.11 |
| 40 | YW | 25.66 | 25.267 | −0.393 | 26.158 | 0.498 | 23.49 | −2.17 |
| 41 | YY | 19.87 | 23.986 | 4.116 | 26.619 | 6.749 | 24.00 | 4.13 |
| 42 | EW | 26.85 | 26.265 | −0.585 | 26.470 | −0.38 | 24.02 | −2.83 |
| 43 | HY | 43.27 | 44.841 | 1.571 | 49.486 | 6.216 | 53.78 | 10.51 |
| 44 | KF | 49.97 | 51.111 | 1.141 | 51.752 | 1.782 | 56.78 | 6.81 |
| 45 | LW | 27.93 | 27.917 | −0.013 | 27.413 | −0.517 | 24.34 | −3.59 |
| 46 | VW | 29.98 | 28.475 | −1.505 | 27.883 | −2.097 | 25.07 | −4.91 |
| 47 | WD | 25.06 | 23.890 | −1.17 | 25.073 | 0.013 | 21.89 | −3.17 |
| 48 | AW | 30.1 | 29.689 | −0.411 | 28.819 | −1.281 | 27.85 | −2.25 |
| 49 | GF | 31.71 | 32.235 | 0.525 | 31.449 | −0.261 | 31.43 | −0.28 |
| 50 | IY | 26.57 | 26.948 | 0.378 | 28.082 | 1.512 | 24.59 | −1.98 |
| 51 | KW | 48.8 | 48.283 | −0.517 | 48.654 | −0.146 | 52.85 | 4.05 |
| 52 | LY | 22.89 | 26.986 | 4.096 | 28.081 | 5.191 | 25.13 | 2.24 |
| 53 | MW | 28.48 | 27.237 | −1.243 | 26.907 | −1.573 | 24.24 | −4.24 |

[a] The 1st–36th samples compose training set and the 37th–53rd samples compose test set. [b] $QM$ + 3D-HoVAIF, the 5-descriptor model. [c] Calculated results using the data from [17]. [d] Results are selected from [17].

correlation coefficient, $R^2_{CV}$ is cross-validation correlation coefficient, SD is the standard deviation of the estimates and $F$ is the Fisher's significance ratio.

## 3. Results and discussion

Comprising seven atom types as H, C($sp^3$), C($sp^2$), N($sp^3$), N($sp^2$), O($sp^2$), O($sp^3$) and S($sp^3$) in the 53 dipeptides, 87 empty items are found in the above-mentioned 165 3D-HoVAIF descriptors. Removing all the empty items, there are ultimately 78 3D-HoVAIF descriptors corresponding to a molecule. Combined with Offord's charge-over-mass term ($Q/M^{2/3}$) as one descriptor, there are in total 79 descriptors (see Supplementary Materials, available online) for one molecule. In this work, two partition methods are adopted on this sample set: one is the whole 53 molecules (Table 2) served as sample set and the other, the top 36 compounds as training set and the remaining as test set to construct models.

All models are built by technique of MLR. And MLR introduces variables in turn according to the values of Fisher prominent test by SMR, and in combination with leave-one-out cross-validation, the optimum variable number is determined in case the correlation coefficient is major and SD is lesser. Firstly, the whole 53 molecules are utilised to construct models. Table 3 lists the results of SMR analysis of variables. In Table 3, $R^2$ and $R^2_{CV}$ increase gradually with the increase of number of descriptors, meanwhile, SD and $SD_{CV}$ decrease. It is shown that for model 5, the values of $R^2$ and $R^2_{CV}$ are the maximal and the values of SD and $SD_{CV}$ are the minimum, which are very near to those of model 4. To determine which model can provide more satisfactory estimation stability and predictive ability, the sample set is divided into two parts: training set includes 36 samples and test set the remaining. The results are as follows:

The 4-descriptor model: $R^2 = 0.966$, SD = 1.097, $F = 224.675$;
The 5-descriptor model: $R^2 = 0.968$, SD = 1.083, $F = 184.662$.

From above analysis, the 5-descriptor model is regarded as the best model, which is expressed as Equation (6) ($R^2 = 0.967$, $R^2_{CV} = 0.959$, SD = 1.261, $SD_{CV} = 1.405$, $F = 277.409$):

$$\mu_{ef} = 12.647 + 945.590\,(0.918)Q/M^{2/3} - 0.04\,(-0.267)V_{63} + 23.465\,(0.156)V_{38} + 0.005(0.077)V_{163} - 185.974(-0.075)V_{43}. \quad (6)$$

In the above equation, the standardised regression coefficients are placed within brackets. Transforming the independent variables to standardised form makes the coefficients more comparable, since they are all in the same units of measure. With the biggest standardised regression coefficient 0.918, $Q/M^{2/3}$ is considered the most relative to $\mu_{ef}$. 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. Then $V_{63}$, a 3D-HoVAIF descriptor, which describes the steric interaction between the atoms of H and O($sp^3$) or S($sp^3$), provides negative contribution to $\mu_{ef}$. The next one $V_{38}$, standing for electrostatic interaction between the atoms of N($sp^3$) and O($sp^3$) or S($sp^3$), gives positive contribution to $\mu_{ef}$. The last two are $V_{163}$ and $V_{43}$, which express hydrophobic interaction between the atoms of O $sp^2$ hybridisation, electrostatic interaction between the N($sp^2$) and O($sp^3$) or S($sp^3$), respectively.

Figure 2 plots the calculated against observed values for training set and test set of the 5-descriptor model. Amongst, all samples are uniformly dispersed around an origin passed line forming an angle of 45°. Figure 3 is the residue distribution scatter, wherein errors of most samples are dispersed within a range of ± 2SD, suggesting good stabilities and predictabilities of the model. The calculated electrophoretic mobilities, the experimental values and errors for all peptides studied in this work (training, test sets) are listed in Table 2.

Table 4 lists the reference reports and our results for the 53 molecules, showing the 3D-HoVAIF model greatly gains by comparison, with its $R^2$ remarkably superior to others. The resulting $R^2$ of 0.967 shows some improvements compared with the $R^2$ of 0.919 for the Offord model. The reason may be like that, to accurately determine the peptide charges, an accurate knowledge

Table 3. Stepwise multiple regression analysis of variables.

| Model | Constant | $QM$ | $V_{63}$ | $V_{38}$ | $V_{163}$ | $V_{43}$ | $R^2$ | $R^2_{CV}$ | SD | $SD_{CV}$ | $F$ |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 9.379 | 987.275 (0.958) | | | | | 0.919 | 0.908 | 1.908 | 2.024 | 575.677 |
| 2 | 13.190 | 951.208 (0.923) | −0.027 (−0.179) | | | | 0.949 | 0.939 | 1.521 | 1.663 | 468.028 |
| 3 | 11.666 | 916.263 (0.890) | −0.043 (−0.286) | 21.105 (0.140) | | | 0.958 | 0.949 | 1.402 | 1.541 | 370.630 |
| 4 | 13.682 | 897.497 (0.871) | −0.044 (−0.291) | 25.594 (0.170) | 0.005 (0.081) | | 0.963 | 0.955 | 1.318 | 1.468 | 316.265 |
| 5 | 12.647 | 945.590 (0.918) | −0.040 (−0.267) | 23.465 (0.156) | 0.005 (0.077) | −185.974 (−0.075) | 0.967 | 0.959 | 1.261 | 1.405 | 277.409 |

Standardised regression coefficients are placed within brackets. Transforming the independent variables to standardised form makes the coefficients more comparable since they are all in the same units of measure.

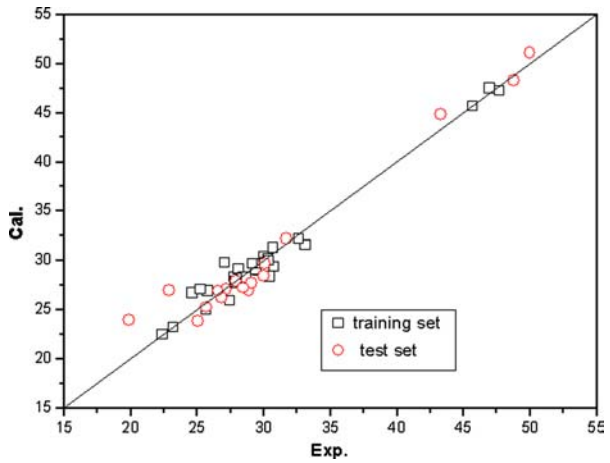


Figure 2. Calculated against observed values for training set and test set of the 5-descriptor model.

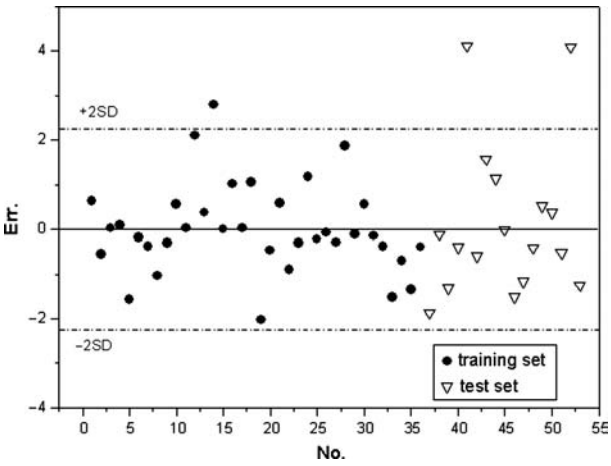


Figure 3. The residue distribution scatter of the 5-descriptor model.

Table 4. Comparison between QSAR models.

| No. | Model | $R^2$ | $R^2_{CV}$ | SD | $F$ |
|---|---|---|---|---|---|
| 1 | $Q/M^{2/3}$ | 0.919 | 0.908 | 1.908 | 575.677 |
| 2 | 3D-HoVAIF[a] | 0.967 | 0.959 | 1.261 | 277.409 |
| 3 | 3D-HoVAIF[b] | 0.968 | 0.722 | 1.083 | 184.662 |
| 4 | [17][c] | 0.895 | – | – | – |
| 5 | [17][d] | 0.941 | 0.839 | 1.441 | 168.762 |

[a] $Q/M^{2/3}$ + 3D-HoVAIF($V_{63}$,$V_{38}$,$V_{163}$,$V_{43}$), 53 samples. [b] $Q/M^{2/3}$ + 3D-HoVAIF($V_{63}$,$V_{38}$,$V_{163}$,$V_{43}$), 36 samples. [c] ($Q/M^{2/3}$) + corrected steric substituent constant ($E_{s,c}$) + molar refractivity ($MR$), 90 samples. [d] ($Q/M^{2/3}$) + corrected steric substituent constant ($E_{s,c}$) + molar refractivity ($MR$), 36 samples.

of the ionisation constants of the amino acid residues in peptides is required. However, the available p$K_a$ values for the ionisable functional groups in the amino acids are used for all peptides, neglecting their composition and sequence. This could lead to erroneous determination of peptide charges as the ionisation of certain amino acid residues in a sequence could be affected by electrostatic and steric interactions with the nearest neighbouring residues [43]. The steric and electrostatic interactions occur in accordance with the composition and sequence of amino acid residues in peptides, which in turn could be a source of error in the determination of p$K_a$ and subsequently peptide charges depending on the nature of peptides used as the training and/or test sets. In a word, using the p$K_a$ values of free amino acids instead of real p$K_a$ values of ionogenic groups in peptides may lead to very significant errors in estimation of peptide charge. In our models, the deficiency is redeemed by 3D-HoVAIF descriptors. It is suggested by Tables 2 and 4 that the superiorities to other reference reports indicate that the 5-descriptor model here is capable of extending into external sample sets.

## 4. Conclusions

A simple computer-assisted model has been developed for accurate prediction of the electrophoretic mobilities of dipeptides. At first, the abilities of the existing empirical models in reproducing $\mu_{ef}$ are examined. Among different models, the Offord model shows a reasonable correlation. However, this model does not account for several factors that affect peptide mobility. Our results confirm the conclusion that $\mu_{ef}$ cannot be successfully predicted for all categories of peptides relying only on the two parameters of charge and size. Then, a MLR model is generated by considering different parameters that may affect $\mu_{ef}$, generated by 3D-HoVAIF. Appearance of the 3D-HoVAIF descriptors and Offord charge-to-mass parameter in this model indicates the importance of the steric, hydrophobic and electrostatic interactions in the electromigration mechanism.

## Acknowledgements

The authors thank the Foundations of National High Technology (863) Programme (No. 2006AA02Z312), Innovative Group Program for Graduates of Chongqing University, Science and Innovation Fund (No. 200711C1A0010260).

## References

[1] V. Kašička, *Recent developments in CE and CEC of peptides*, Electrophoresis 29 (2008), pp. 179–206.
[2] P.D. Grossman, J.C. Colburn, and H.H. Lauer, *A semiempirical model for the electrophoretic mobilities of peptides in free-solution capillary electrophoresis*, Anal. Biochem. 179 (1989), pp. 28–33.
[3] R.E. Offord, *Electrophoretic mobilities of peptides on paper and their use in the determination of amide groups*, Nature (London) 211 (1966), pp. 591–593.
[4] B.J. Compton, *Electrophoretic mobility modeling of proteins in free zone capillary electrophoresis and its application to monoclonal antibody microheterogeneity analysis*, J. Chromatogr. 559 (1991), pp. 357–366.
[5] N.J. Adamson and E.C. Reynolds, *Rules relating electrophoretic mobility, charge and molecular size of peptides and proteins*, J. Chromatogr. B 699 (1997), pp. 133–147.

[6] I. Messana, D.V. Rossetti, L. Cassiano, F. Misiti, B. Giardina, and M. Castagnola, *Peptide analysis by capillary (zone) electrophoresis*, J. Chromatogr. B 699 (1997), pp. 149–171.

[7] V. Kašička, *Capillary electrophoresis of peptides*, Electrophoresis 20 (1999), pp. 3084–3105.

[8] A. Cifuentes and H. Poppe, *Behavior of peptides in capillary electrophoresis: Effect of peptide charge, mass and structure*, Electrophoresis 18 (1997), pp. 2362–2376.

[9] M.A. Survay, D.M. Goodall, S.A.C. Wren, and R.C. Rowe, *Self-consistent framework for standardising mobilities in free solution capillary electrophoresis: Applications to oligoglycines and oligoalanines*, J. Chromatogr. A 741 (1996), pp. 99–113.

[10] R.F. Cross and M.G. Wong, *Objective testing for the dependence of electrophoretic mobilities upon size in capillary zone electrophoresis*, Chromatographia 53 (2001), pp. 431–436.

[11] V. Solinova, V. Kašička, D. Koval, and J. Hlavacek, *Separation and investigation of structure-mobility relationships of insect oostatic peptides by capillary zone electrophoresis*, Electrophoresis 25 (2004), pp. 2299–2308.

[12] R. Plasson and H. Cottet, *Determination of homopolypeptide conformational changes by the modeling of electrophoretic mobilities*, Anal. Chem. 77 (2005), pp. 6047–6054.

[13] V. Solinova, V. Kašička, P. Sazelova, T. Barth, and I. Miksik, *Separation and investigation of structure-mobility relationship of gonadotropin-releasing hormones by capillary zone electrophoresis in conventional and isoelectric acidic background electrolytes*, J. Chromatogr. A 1155 (2007), pp. 146–153.

[14] S. Mittermayr, M. Olajos, T. Chovan, G.K. Bonn, and A. Guttman, *Mobility modeling of peptides in capillary electrophoresis*, Trends Anal. Chem. 27 (2008), pp. 407–417.

[15] A. Cifuentes and H. Poppe, *Simulation and optimization of peptide separation by capillary electrophoresis*, J. Chromatogr. A 680 (1994), pp. 321–340.

[16] G.M. Janini, C.J. Metral, H.J. Issaq, and G.M. Muschik, *Peptide mobility and peptide mapping in capillary zone electrophoresis. Experimental determination and theoretical simulation*, J. Chromatogr. A 848 (1999), pp. 417–433.

[17] M.J. Heravi, Y. Shen, M. Hassanisadi, and M.G. Khaledi, *Prediction of electrophoretic mobilities of peptides in capillary zone electrophoresis by quantitative structure-mobility relationships using the offord model and artificial neural networks*, Electrophoresis 26 (2005), pp. 1874–1885.

[18] H.X. Liu, X.J. Yao, C.X. Xue, R.S. Zhang, M.C. Liu, Z.D. Hu, and B.T. Fan, *Study of quantitative structure-mobility relationship of the peptides based on the structural descriptors and support vector machines*, Anal. Chim. Acta 542 (2005), pp. 249–259.

[19] W.P. Ma, F. Luan, H.X. Zhang, X.Y. Zhang, M.C. Liu, Z.D. Hu, and B.T. Fan, *Accurate quantitative structure-property relationship model of mobilities of peptides in capillary zone electrophoresis*, Analyst 131 (2006), pp. 1254–1260.

[20] M.V. Piaggio, M.B. Peirotti, and J.A. Deiber, *Exploring the evaluation of net charge, hydrodynamic size and shape of peptides through experimental electrophoretic mobilities obtained from CZE*, Electrophoresis 27 (2006), pp. 4631–4647.

[21] M.B. Peirotti, M.V. Piagg, and J.A. Deiber, *Hydration, charge, size, and shape characteristics of peptides from their CZE analyses*, J. Sep. Sci. 31 (2008), pp. 548–554.

[22] Y. Xin, H. Mitchell, H. Cameron, and S.A. Allison, *Modeling the electrophoretic mobility and diffusion of weakly charged peptides*, J. Phys. Chem. B 110 (2006), pp. 1038–1045.

[23] H.X. Pei, Y. Xin, and S.A. Allison, *Using electrophoretic mobility and bead modeling to characterize the charge and secondary structure of peptides*, J. Sep. Sci. 31 (2008), pp. 555–564.

[24] K. Yu and Y.Y. Cheng, *Machine learning techniques for the prediction of the peptide mobility in capillary zone electrophoresis*, Talanta 71 (2007), pp. 676–682.

[25] P. Zhou, F.F. Tian, and Z.L. Li, *Three dimensional holographic vector of atomic interaction field (3D-HoVAIF)*, Chemometr. Intell. Lab. Syst. 87 (2007), pp. 88–94.

[26] H. Winer, *Structural determination of paraffin boiling point*, J. Am. Chem. Soc. 69 (1947), pp. 2636–2641.

[27] H. Hosoya, *A new proposed quantity characterizing the topological nature of structural isomers of saturated hydrocarbons*, Bull. Chem. Soc. 44 (1971), pp. 2332–2339.

[28] M. Randic, *On characterization of molecular branching*, J. Am. Chem. Soc. 97 (1975), pp. 6609–6615.

[29] A.T. Balaban, *High discrimination distance-based topological index*, Chem. Phys. Lett. 89 (1982), pp. 399–404.

[30] A.R. Katritzky, V.S. Lobanov, and M. Karelson, *QSPR: The correlation and quantitative prediction of chemical and physical properties from structure*, Chem. Soc. Rev. 24 (1995), pp. 279–287.

[31] A.R. Katritzky, U. Maran, V.S. Lobanov, and M. Karelson, *Structurally diverse quantitative structure-property relationship correlations of technologically relevant physical properties*, J. Chem. Inf. Comput. Sci. 40 (2000), pp. 1–18.

[32] R.D. Cramer, D.E. Patterso, and J.D. Bunce, *Comparative molecular field analysis (CoMFA). 1. Effect of shape on binding of steroids to carrier proteins*, J. Am. Chem. Soc. 110 (1988), pp. 5959–5967.

[33] G. Klebe, U. Abraham, and T. Mietzner, *Molecular similarity indices in a comparative analysis (CoMSIA) of drug molecules to correlate and predict their biological activity*, J. Med. Chem. 37 (1994), pp. 4130–4146.

[34] A.M. Doweyko, *The hypothetical active site lattice-an approach to modelling active sites from data on inhibitor molecules*, J. Med. Chem. 31 (1988), pp. 1396–1406.

[35] P.J. Goodford, *A computational procedure for determining energetically favorable binding sites on biologically important molecules*, J. Med. Chem. 28 (1985), pp. 849–857.

[36] A.N. Jain, T.G. Dietterich, R.H. Lathrop, D. Chapman, R.E. Critchlow, T.A. Webster, and T. Lozaoperez, *COMPASS – a shape-based machine learning tool for drug design*, J. Comput. Aided Mol. Des. 8 (1994), pp. 635–652.

[37] M. Levitt, *Protein folding by restrained energy minimization and molecular dynamics*, J. Mol. Biol. 170 (1983), pp. 723–764.

[38] G.E. Kellogg, J.C. Burnett, and D.J. Abraham, *Very empirical treatment of solvation and entropy: A force field derived from Log Po/w*, J. Comput. Aided Mol. Des. 15 (2001), pp. 381–393.

[39] M. Hahn, *Receptor surface models. 1. Definition and construction*, J. Med. Chem. 38(12) (1995), pp. 2080–2090.

[40] F.C. Wireko, G.E. Kellogg, and D.J. Abraham, *Allosteric modifiers of hemoglobin. Crystallographically determined binding sites and hydrophobic binding/interaction analysis of novel hemoglobin oxygen effectors*, J. Med. Chem. 34 (1991), pp. 758–767.

[41] G.E. Kellogg, G.S. Joshi, and D.J. Abraham, *New tools for modeling and understanding hydrophobicity and hydrophobic interactions*, Med. Chem. Res. 1 (1992), pp. 444–453.

[42] J. Pei, Q. Wang, J. Zhou, and L. Lai, *Estimating protein-ligand binding free energy: Atomic solvation parameters for partition coefficient and solvation free energy calculation*, Proteins 57 (2004), pp. 651–664.

[43] E.C. Rickard, M.M. Strohl, and R.G. Nielsen, *Correlation of electrophoretic mobilities from capillary electrophoresis with physicochemical properties of proteins and peptides*, Anal. Biochem. 197 (1991), pp. 197–207.