BMC Bioinformatics

PROCEEDINGS   **Open Access**

Prediction of peptide drift time in ion mobility mass spectrometry from sequence-based features

Bing Wang[1,2,3*], Jun Zhang[4], Peng Chen[5], Zhiwei Ji[3], Shuping Deng[3], Chi Li[6]

Abstract

Background: Ion mobility-mass spectrometry (IMMS), an analytical technique which combines the features of ion mobility spectrometry (IMS) and mass spectrometry (MS), can rapidly separates ions on a millisecond time-scale. IMMS becomes a powerful tool to analyzing complex mixtures, especially for the analysis of peptides in proteomics. The high-throughput nature of this technique provides a challenge for the identification of peptides in complex biological samples. As an important parameter, peptide drift time can be used for enhancing downstream data analysis in IMMS-based proteomics.

Results: In this paper, a model is presented based on least square support vectors regression (LS-SVR) method to predict peptide ion drift time in IMMS from the sequence-based features of peptide. Four descriptors were extracted from peptide sequence to represent peptide ions by a 34-component vector. The parameters of LS-SVR were selected by a grid searching strategy, and a 10-fold cross-validation approach was employed for the model training and testing. Our proposed method was tested on three datasets with different charge states. The high prediction performance achieve demonstrate the effectiveness and efficiency of the prediction model.

Conclusions: Our proposed LS-SVR model can predict peptide drift time from sequence information in relative high prediction accuracy by a test on a dataset of 595 peptides. This work can enhance the confidence of protein identification by combining with current protein searching techniques.

Background

Ion mobility spectrometry (IMS) has gained significant attentions over the past few decades for rapid, high-resolution separations power, which can separate ions on a millisecond time-scale [1-3]. As a separation technique which based on differences in size and shape of analytes, IMS has proven powerful in the fields of metabolomics, glycomics and proteomics [1,2]. Ion mobililty-mass spectrometry (IMMS), an analytical technique by which IMS coupled with mass spectrometry (MS), have emerged as powerful tools for analyzing biological mixtures, especially for current proteomics studies [4-7]. By combination of the advantages of IMS and MS, IMMS opens up avenues for the detailed structural analysis of large and heterogeneous protein complexes, providing information on the stoichiometry, topology and cross section of their composition [8,9].

A typical proteomics experimental setup using IMMS consists of five components: sample introduction, compound ionization, ion mobility separation, mass separation as well as peptide and protein ion detection [10]. Although these five components all play essential roles in the process, ion mobility separation is crucial for its impact on the consequent mass analysis and peptide ion detection [11]. Ion mobility separation, by which the peptide ions with different cross-sections and molecular charges will be separated, adds a new dimension of separation and makes IMMS an attractive method for analyzing complex proteomics samples. Peptide ion separation can be enhanced by changing different gases, altering electric field strengths, and adopting non-linear electric field gradients, by which peptide identification can be facilitated to achieve high

* Correspondence: wangbing@ustc.edu
[1]The Advanced Research Institute of Intelligent Sensing Network, Tongji University, shanghai, 201804, China
Full list of author information is available at the end of the article

confidence [12]. Even though these efforts improve the separation capability of IMMS, they are still time-consuming, and it is difficult to reproduce under different experimental conditions.

Although IMMS separates peptide ions based on differing cross-sections and molecular charge, the experimental measurement behaves in the way that peptides spend different time through the drift tube. It has been reported that the measurement of peptide ion drift time using IMMS is very reproducible [13-18]. Any two measurements of mobilities (or cross sections) recorded on the same instrument usually agree to within 1% relative uncertainty. Measurements performed by different groups usually agree to within 2%. As a characteristic of different ions, peptide ion drift time can be used to enhance confidence in protein identifications.

There are several efforts which attempt to computationally determine the mobile behaviour of peptide ions in IMS. Valentine et al. predict peptide ion cross sections using intrinsic size parameters (ISPs) and tested it on 271 singly-charged peptides [19]. A quantitative structure-property relationship (QSPR) based approach was proposed for prediction of peptide drift time by Liu et al. and found the structure effect and the charge states of peptide ion contribute a lot to the drift time [20]. Shah et al. employed partial least squares (PLS) and support vector regression (SVR) based approaches to predict the drift time of massive peptide ions with different charge states and demonstrated both techniques significantly outperform the ISPs based calculation by a test on a high confidence database of 8,675 peptide sequences [21]. Zhang et al. presented a quantitative structure-spectrum relationship (QSSR) study to predict peptide drift time and found the sequence-based approach can get better fitting ability and predictive power but worse interpretability than the structure-based approach [22]. Our previous works also attempted to address the same problem by employing artificial neural networks and multiply linear regression models [23-25]. Although these studies contributed the drift time prediction of peptide ions a lot, ISP based calculations did not show the high performance in peptides with high charged states, and structure-based methods have to construct and optimize the geometrical structures of peptides which will bring inevitable errors into prediction models.

In this paper, a least square-support vectors regression (LS-SVR) model is presented to predict peptide ion drift time in IMMS just from the sequence-based features of peptide. The sequence pattern of each peptide was represented as a 36-component vector, which was consisted of for descriptors, i.e., molecular weight, sequence length, amino acid composition and pseudo amino acid composition. In construction of the LS-SVR regression, a 10-fold cross-validation strategy was employed to determine the

optimized values of the regression parameters. Our proposed LS-SVR method was applied into three peptide ions datasets with different charge states, i.e., +1, +2, +3.

Results and discussion

In this work, all the raw data generated from the IMMS were processed using MassLynx V4.1, an instrument control software, to obtain the drift time for each peptide ion peak. MassLynx is a powerful software for analyzing and processing the data acquired from mass spectrometers which are developed Waters Corporation. The peptides generated from tryptic digestion of 20 pure proteins were used for our model development and testing in this study. Peptide charge status was manually assigned based on the m/z spacing between isotopic peaks. As a result, the total of 595 peptides assigned ions which came from the 20 proteins became the dataset for this work. Within this dataset, 212 peptides were singly charged, 306 were doubly and 77 were triply charged. More details can be found in our previous work [12,26].

IMS separate ions based on the fact ions with different shapes and charge states travel though the drift tube at different velocities. In the drift tube, the ions were pulled by a weak electric field and opposed by the inset buffer gas. The charge state is a very important factor for the drift time. Therefore, we developed the SVR models for singly-, doubly- and triply-charged peptides, respectively. In this work we denotes dataset of singly-charged peptides as DataS, doubly-charged peptides as DataD, and triply-charged as DataT.

Table 1 shows the distributions of peptide molecular weight, sequence length and drift time in each of the three datasets. It can be seen that the smallest peptide just formed by 3 amino acids with singly-charge state, and the largest one have 34 amino acids from DataD and DataT, which indicate that peptides with large molecular weight and long amino acid sequences, tend to have high charge states. The peptide ion drift time is also significantly related to the overall ion charge state. The mean value of peptide drift time for the singly-charged peptides is 7.48 s while that of the doubly-charged and the triply-charged peptides are 3.07 s and 2.28 s, respectively. The peptides with high charge states drift through the cell in

Table 1 Distribution of peptide molecular weight, sequence length and drift time in original datasets with different charge states

	Molecular weight (Da)		Sequence length		Drift time (s)	
	range	mean	range	mean	range	Mean
DataS	374.28-2088.9	900.14	3-19	7.9	2.17-24.5	7.48
DataD	605.35-3412.7	1470.39	5-34	13.2	1.08-9.39	3.07
DataT	981.56-3503.7	2046.30	8-34	18.3	1.08-3.97	2.28

a relative high velocity. Another reason is the higher charge states the peptide is, the higher probability that they form a 3-dimensional spatial structure will be.

Prediction performance evaluation

In this study, we developed the LS-SVR models for predicting peptide drift time for the singly-, doubly-, and triply-charged peptides, respectively. A 10-fold cross-validation strategy was employed in the training and test process of the regression models, by which all observations in each datasets are used for both training and validation. This cross-validation can provide reliable learning of our model from the original data.

The purpose of this work is to predict ion drift time of peptides by elucidating the relationship between the dependent variable, i.e., peptide drift time, and the sequence-based peptide features we used, i.e., peptide molecular weight, sequence length, AAC and PseAAC. For regression analysis, there are many criteria by which they can be evaluated and compared. The root mean square error ($RMSE$) and coefficient of determination (R^2) are selected in this work to evaluate the predictive performance of our LS-SVR models.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n} (dt_i - dt'_i)^2}{n}} \qquad (1)$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (dt_i - dt'_i)^2}{\sum_{i=1}^{n} (dt'_i - \overline{dt_i})^2} \qquad (2)$$

where n is the number of peptide in the dataset, dt is the experimentally observed peptide ion drift time, dt the predicted drift time by LS-SVR models, $\overline{dt}$ is the overall average value of peptide drift time. R^2 takes any value between 0 and 1, with a value closer to 1 indicating the regression model is of better performance.

Furthermore, in order to assess the prediction accuracy of LS-SVR models, a prediction variation threshold, η_t, was defined by the relative variation of the predicted drift time from the experimentally observed values. If the relative variation between observed and predicted drift time is smaller than η_t, the prediction will be seen as reliable, otherwise, unreliable.

$$\eta_t = \frac{|dt - dt'|}{dt} \qquad (3)$$

Where η is the prediction variation, dt' is the predicted peptide ion drift time and dt is the experimentally observed peptide ion drift time.
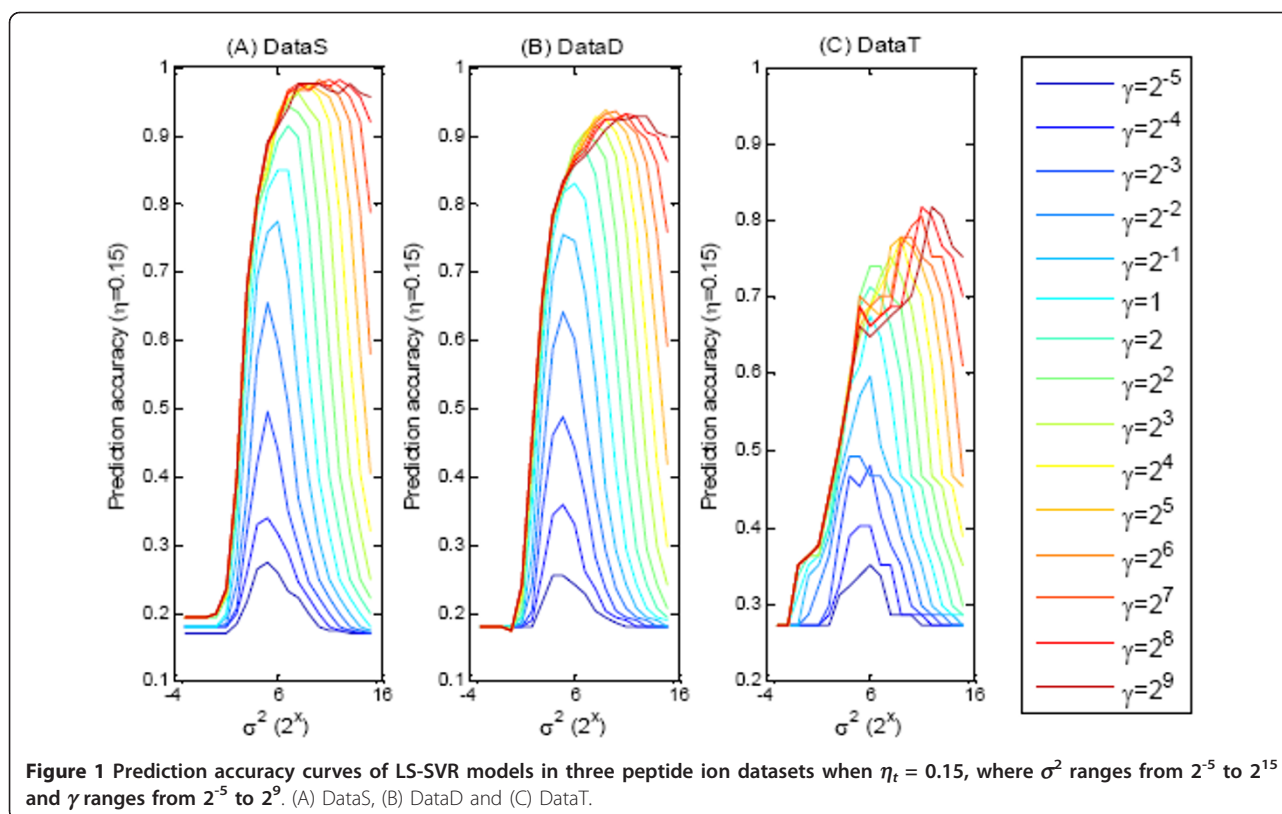
Parameters selection

As what state in Methods part, LS-SVR models with Gaussian kernel was adopted to predict peptides drift time. There are two important parameters for this kind of regression model, i.e., the width of Gaussian kernel parameter σ, and the regularization factor γ. The correct setting of these two parameters of the LS-SVR models is of critical importance in enabling us to achieve good regression performances. In this work, the grid-searching scheme is used to determine these two parameters based on cross validation strategy. Specifically, the σ^2 and γ were tuned simultaneously in a grid ranging from $2^{-5}, 2^{-4}, ..., 2^{15}$ for σ^2 and from $2^{-5}, 2^{-4}, ..., 2^{9}$ for γ. The prediction accuracy of LS-SVR models for each peptide dataset was seen as the objective function to determine the optimum combination of σ^2 and γ, where the value of η_t was set as 0.15.

The accuracy curves for different combination of the σ^2 and γ in the three peptide datasets were shown in the Figure 1. It can be seen that the regression performance of LS-SVR models are heavily depend on the selection of the parameters σ^2 and γ. When γ is fixed, the prediction accuracy goes up with the increase of σ^2 to an apex and then goes down. For DataS, the top 5 prediction accuracy values correspond to the combinations [σ^2, γ] of [2^{10}, 2^6], [2^{11}, 2^7], [2^{12}, 2^8], [2^{13}, 2^9], and [2^9, 2^5]. The top 5 LS-SVR models for DataD have the combination parameters of [2^9, 2^5], [2^{10}, 2^6], [2^{11}, 2^7], [2^{11}, 2^8], and [2^9, 2^6]. For the peptide dataset with triply-charge, DataT, the top 5 combinations are [2^{11}, 2^8], [2^{12}, 2^9], [2^{10}, 2^7], [2^{11}, 2^8], and [2^{12}, 2^9]. Overall the three datasets, the value [2^{11}, 2^8] can achieve the best prediction accuracy for the LS-SVR models when η_t = 0.15. Therefore, the σ^2 of 2^{11} and γ of 2^8 were selected for the subsequent analysis in this work.

Prediction performance

A 10-fold cross validation was implemented in the construction of LS-SVR models, by which the different separation of the original dataset will bring the changes of predicted drift time for each peptide. For evaluating the uncertainty in the regression performance of our model which come from the randomness of the dataset separations, the regression procedure was repeated for ten times. The mean of the prediction drift times for each peptide from these ten times experiments were taken as the finally predicted value. Also the variation of the ten times was studied to exam the stability of our proposed LS-SVR models.

The prediction performance was shown in Table 2. It can be seen that our models ca achieved very good prediction ability for different peptide dataset, i.e., 0.9811 for DataS, 0.9379 for DataD, and 0.8312 for DataT.

Figure 1 Prediction accuracy curves of LS-SVR models in three peptide ion datasets when $\eta_t = 0.15$, **where** σ^2 **ranges from** 2^{-5} **to** 2^{15} **and** γ **ranges from** 2^{-5} **to** 2^9. (A) DataS, (B) DataD and (C) DataT.

Comparing to DataS and DataD, the prediction accuracy of the triply-charge peptide ions in DataT is a little bit poor. One reason for this situation is that the dataset's size is small, i.e., 77 peptide in DataT, which can not provide sufficient information in the model training. Another reason, we believe, is that the charge state of DataT is higher than that of DataS and DataD, which usually cause the peptide longer. The mean length of peptides in DataT is 18.3, which is 1.4 times of that in DataD, and 2.3 times in DataS. The longer of the peptide length is, the more chance the peptide form the secondary structure will be. Obviously, the changes in space conformation will contribute the peptide's velocity in drift cell and therefore, affect the peptide ion's drift time.
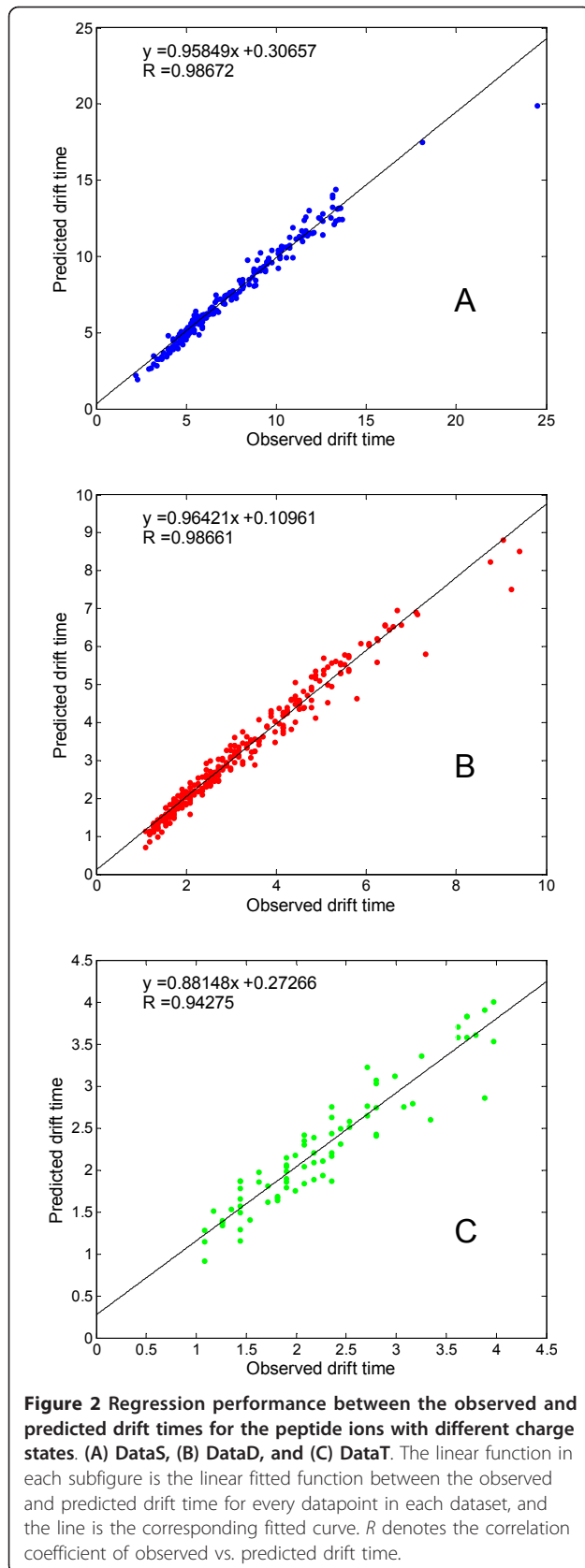
Table 2 Prediction performance of LS-SVR models under a variation threshold of 15% in three peptide ion's datasets

	Prediction accuracy[a]		RMSE	R^2
DataS	0.9811	(0.9736±0.081)	0.5202	0.9718
DataD	0.9379	(0.9340±0.061)	0.2602	0.9721
DataT	0.8312	(0.7883±0.025)	0.2637	0.8727

a. The prediction accuracy for each dataset was shown as the format of A(B ±C), where A denotes the prediction accuracy from the mean of predicted drift times, B the mean prediction accuracy of the ten repeat times, and C the standard deviation of the accuracy of the ten repeat times.

It can be found from Table 2 that the prediction accuracy from the mean of the predicted drift times is better than the mean accuracy of the ten repeat experiments. It can get 0.0075, 0.0039 and 0.0479 for DataS, DataD, and DataT, respectively, which indicated that the combination regression model will improve the predictive power of predictors. From Table 2, it can also be seen that the standard deviation of the prediction accuracy of the ten repeat experiments is very small, i.e., 0.081, 0.061 and 0.025 for the three datasets. It demonstrate our LS-SVR models are stable and statistically valid because a small change in the data, such as the different split of the training and test dataset, may lead to large changes of the prediction performance.

The relative small $RMSE$ and R^2 shown in Table 2 also indicted the powerful regression performance of LS-SVR models in prediction of peptide ion's drift times in IMMS. We got very small $RMSE$ values for DataD and DataT, and a little higher value, 0.52, for DataS, which is reasonable for the big range of the original drift time, from 2.17 s to 24.5 s. The R^2 values of around 0.97 for DataS and DataD, 0.87 for DataT are shown high correlation between the predicted and experimental observed peptide drift times. More details about the regression results can be found in Figure 2, where the line showed the linear fitting between the predicted and observed drift time in a least-squares sense. The high correlation

Figure 2 Regression performance between the observed and predicted drift times for the peptide ions with different charge states. **(A) DataS, (B) DataD, and (C) DataT**. The linear function in each subfigure is the linear fitted function between the observed and predicted drift time for every datapoint in each dataset, and the line is the corresponding fitted curve. *R* denotes the correlation coefficient of observed vs. predicted drift time.

coefficients, i.e., 0.987 for both DataS and DataD, and 0.943 for DataT, signifies the LS-SVR model we proposed here can capture the general properties by which different peptides fly through drift cell in different velocities.

After the LS-SVR models had finished the regression analysis for the three datasets with different charge states ions, the variation threshold η_t will decide which peptide can be predicted correctly. Figure 3 displays the relation between the fraction of peptide ions whose drift time are predicted correctly and the accuracy threshold η_t. It can be seen that our proposed method can get best prediction performance in the DataS. The reason we believe is the peptides in DataS are small and have higher probability they adopt elongated conformations in order to minimize coulomb repulsion, while the peptides in DataT usually are large and have higher probability to form secondary structure when they go through the drift cell in IMMS instrument. It can be found even the variation threshold is set as 0.10, there are more than 90% peptides can be predicted correctly, by which the prediction performance of our LS-SVR model can be demonstrated. If the conformation information can be added into the regression model, the predictive power for doubly- and triply-charge peptides will be increased undoubtedly.

Conclusions

To enhance the confidence of peptide identification, a LS-SVR model was developed in this study to predict peptide ion drift time for IMMS measurements. In LS-SVR, there are two parameters, i.e., the width of Gaussian kernel parameter σ, and the regularization factor γ, have to be selected for their influence on the regression accuracy. A grid searching strategy was employed to optimize the selection of these two parameters. Based on the peptide sequence, a 34-component vector was extracted as representation to construct our LS-SVR models on three peptide ion datasets with different charge states. With the prediction accuracy threshold η was set to 0.15, we achieved very high performance, i.e., 0.9811 and 0.9379, for the peptide ions with singly- and doubly-charge, which indicated the prediction capability of the LS-SVR models. It is reasonable that there is a relative lower prediction accuracy of 0.8312 for DataT, for the peptides with higher charge states have a higher probability that they can form a secondary structure. This kind of situation will be improved if the structure information can be added into our proposed LS-SVR models; even more computational cost will be requested.

Methods
Peptide dataset
The total of 595 peptides of 20 pure proteins used in this work was reported in our previous work [12]. The
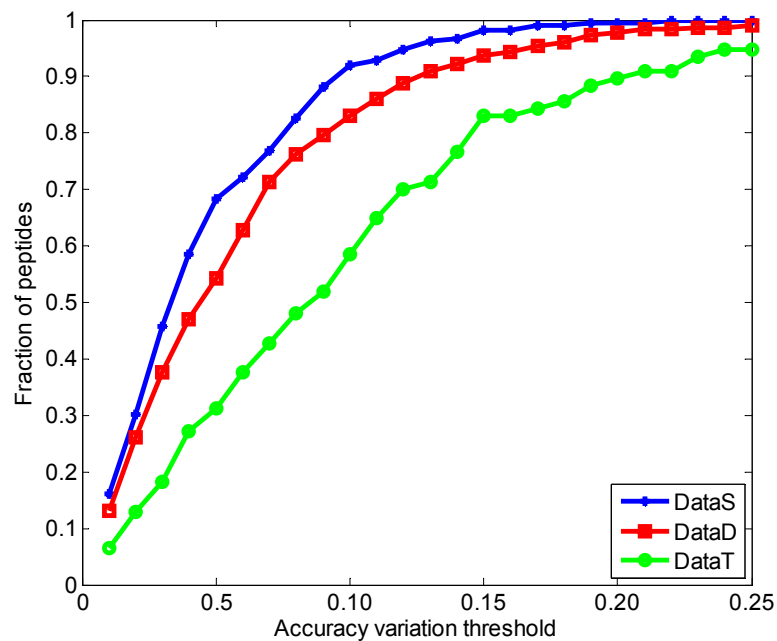
Figure 3 Fraction of peptide ions correctly predicted at different accuracy variation levels. A higher curve indicates a larger number of peptides for a given threshold value.

proteins were purchased from Sigma Aldrich and used without further purification. The peptide fragments were produced from the pure proteins according to the details of the sample preparation section in the report, and then were analyzed by direct electrospray into the Synapt HDMS instrument (Waters). Peptide ion assignments were obtained from a peptide mass fingerprint for each tryptic digest. As a result, in the dataset with 595 peptide ions, there are 212 peptides were singly charged, 306 were doubly charged and 77 were triply charged. More details about the experimental processing of samples can be obtained from the work [12,26].

Support vector regression

Support vector machines, a specific class of machine learning algorithms which was firstly proposed by Vapnik and his co-workers in 1995 [12], have proven very effective for solving pattern classification problems, even for the dataset in small size. For a binary classification problem, the main idea of SVM is to select a hyper-plane that separates the positive from negative samples while maximizing the minimum margin. Currently, SVM has been became one of the most popular machine learning methods, which has been applied to various domains of interest, such as bioinformatics, cheminformatics, image processing, data mining, knowledge discovery, and etc. In many applications, SVM can achieve excellent performance for the character that the capacity of the SVM system is controlled by parameters that do not depend on the dimensionality of feature space [27-32].

In the same way as with classification task, SVM can also be applied to the case of regression which is called support vector regression (SVR). In statistics, regression analysis is a statistical technique for estimating the relationships among variables. All the regression tasks can be formulated as to seek an estimation function which can approximate the observations within an acceptable error range. In this study, least square support vector regression (LS-SVR), a version of SVR which can reduce the complexity of optimization processes, was adopted for the drift time prediction[33].

Given a training dataset $D = \{x_i, y_i\}(i = 1, 2, ..., n)$, $x \mathbf{R} \in {}^n$, $y \in \mathbf{R}$, where x_i is the input vector, y_i is its corresponding target vector and n is the size of the dataset, SVR can construct regression model by using nonlinear mapping function $\varphi(\cdot)$ as follows:

$$y(x) = w^T\phi(x) + b, w \in \mathbf{x}, b \in \mathbf{R} \qquad (4)$$

where w is the vector of coefficients and b a constant. Usually, w and b are obtained by minimizing the upper bound of generalization error. Accordingly, the regression problem in LS-SVR can be transformed into the following optimization problem[34]:

$$\min \quad 1/2 w^T w + 1/2\gamma \sum_{i=1}^{l} e_i^2 \qquad (5)$$

$$s.t. \quad y_i = w^T\phi(x_i) + b + e_i(i = 1, 2, \dots, l)$$

where γ is the regularization parameter, is applied to control the minimization of estimation error and the function smoothness, and e_i is the error between actual output and predictive output of the i-th input data. The high value of γ denotes the good fitting of the training data points is stressed, and in the case of noisy data a smaller γ value should be taken to avoid overfitting. In order to solve the optimization problem, the Lagrangian function is formulated as following:

$$L(w, b, e, \alpha) = 1/2w^T w + 1/2\gamma \sum_{i=1}^{n} e_i^2 - \sum_{i=1}^{n} \alpha_i [w^T \phi(x_i) + b + e_i - y_i] \quad (6)$$

where $\alpha = (\alpha_1, \alpha_2, ..., \alpha_l)$ is the Lagrange multiplier. The KKT conditions are used for optimality by differentiating L with the variable w,b,e,α, which is shown as follows.

$$\begin{cases} \dfrac{\partial L}{\partial w} = 0 \rightarrow w = \sum_{i=1}^{n} \alpha_i \phi(x_i) \\ \dfrac{\partial L}{\partial b} = 0 \rightarrow \sum_{i=1}^{n} \alpha_i = 0 \\ \dfrac{\partial L}{\partial e_i} = 0 \rightarrow \alpha_i = \gamma e_i, \ i = 1, \dots, n \\ \dfrac{\partial L}{\partial \alpha_i} = 0 \rightarrow w^T \phi(x_i) + b + e_i - y_i = 0, \ i = 1, \dots, n \end{cases} \quad (7)$$

By solving the upper linear system, the final solution of the primal problem can be represented in the following form.

$$f(x) = \sum_{i=1}^{n} w_i K(x, x_i) + b \quad (8)$$

where $K(\bullet)$ is kernel function which can satisfy Mercer's condition corresponds to a dot product ion some feature spaces [34]. The most used kernel functions include the Gaussian RBF $K(x, x_i) = \exp(||x - x_i|| / 2\sigma^2)$ with a width of σ, sigmoid and the polynomial kernel $K(x, x_i) = (a_1 x x_i + a_2)^d$ with an order of and constants a_1 and a_2. Gaussian RBF kernel is employed in this study, and the kernel parameter $\sigma 2$ and γ, therefore, should be determined firstly. Currently, many approaches have been applied in parameter optimization of SVR, such as experience [27], grid searching [35], particle swarm optimization(PSO) [36], genetic algorithm(GA) [37], simulated annealing algorithm [38]. Considering computing complexity, cross-validation grid searching, the most used method, is selected to determine the parameters $\sigma 2$ and γ in LSSVR model.

Peptide representation
To implement LS-SVR model to predict peptide drift time in IMMS, each peptide have be represented as a vector with specific peptide features. Because each peptide is not consistent in the length, and the shape is affected by the charge state of the peptides, only features were extracted from the peptide sequence, therefore, are used to represent the peptide in this work.

Peptide molecular weight
In IMMS, the ions are pulled by a uniform electric field through the buffer gas in the drift cell. Therefore, the molecular weight of peptide is one of the most important parameters which can affect ion mobility. Karasek et al. found there is a linear relationship between the reduced mobility of alkylamines and molecular weight under a specific experimental setting [39]. Also, other researches reported that the reduced mobility is inversely proportional to ion mass [40]. For a peptide P whose sequence is consisted of N amino acid residues as follows:

$$P = R_1 R_2 \cdots R_i \cdots R_N \quad (9)$$

Where R_i denote the i-th amino acid in the peptide. The molecular weight of P can be calculated as:

$$MW_P = \sum_{i=1}^{N} mw_i + (N - 1) \times 18 \quad (10)$$

where mw_i is the molecular weight of i-th amino acid in the peptide sequence.

Sequence length
The sequence length (SL) of peptide, N, plays an important role in the formation of peptide's structure. The longer of the peptide sequence is, the more chance the peptide folds into a secondary or tertiary structure. Except charge states, IMS distinguishes ions based on the ion shapes which is affected by the sequence length. The previous work indicated that peptides only with primary structure will have smaller ion mobility than that with secondary structure, and smaller more than that with tertiary structure.

Amino acid composition
All the peptide information is contained in its complete amino acid sequence. Therefore, it is the best choice for representing each peptide by its complete sequence. Amino acid composition (AAC) is one of the popular approaches to address protein or peptide representation problem because it is simple, yet powerful feature in prediction of protein structure, interaction, and functional sites. Generally, there are only twenty standard amino acid residues are considered in AAC. Therefore, AAC is a 20-components vector, where each component shows the occurrence number of an amino acid type in

the peptide sequence (in many works, ACC is expressed by the occurrence frequencies, not numbers). For peptide P, ACC can be expressed by

$$ACC_P = (a_1 \ a_2 \ ... \ a_{20})^T \qquad (11)$$

Where a_i denotes the normalized frequency of i-th type of amino acid in peptide P.

Pseudo-amino acid composition

Though AAC can represent peptides in a very simply way, it ignores all the information of amino acid sequence-order effects, which decide the local environment of each amino acid in the peptide. Therefore, Pseudo amino acid composition (PseAAC) was originally introduced by Kuo-Chen for representing proteins and had demonstrated its effectiveness in improving protein subcellular localization prediction, membrane protein type prediction and other works [41]. For peptide P, PseAAC could be formulated as

$$PseAAC_P = (p_1, \ p_2, \ \cdots, \ p_{20}, \ p_{20+1}, \ \cdots, \ p_{20+\lambda})^T, (\lambda < N) \quad (12)$$

Where $p_1, p_2, ..., p_{20}$, are associated with the conventional amino acid composition of P, which already represented by sequence length and ACC in above, and $p_{20+1}, p_{20+2}, \cdots, p_{20+\lambda}$ are the λ correlation factors that reflect the 1st tier, 2nd tier, ..., and the λ-th tier sequence order correlation patterns. Therefore, only $p_{20+1}, p_{20+2}, \cdots, p_{20+\lambda}$ in $PseAAC_P$ have been adopted for representing peptides. In this work, six characters of 20 amino acid, i.e., hydrophobicity, hydophilicity, mass, pK1(alpha-COOH), pK2(NH3) and pI(at 25 °), have been used for calculated $PseAAC_P$, and λ is set up to 2.

Feature normalization

From the above section, it can be found that four types of sequence-based features were applied to represent peptides. However, these four features are of different physical dimension of quantity and different value ranges. The imbalanced expression level of different features will result in a variation in contribution of each of them to the drift time predictor. To remove the bias of expression level, all of the feature values have to be normalized to equally reflect (as much as possible) the influence of each feature. In this work, all values of each feature always fall within a fixed interval [-1, 1] by

$$f_{normalized} = 2 \times (f - f_{\min})/(f_{\max} - f_{\min}) - 1 \qquad (13)$$

where f is the raw value of feature, $f_{normalized}$ denotes the normalized value of this feature, $f_{\min}$ and $f_{\max}$ are the minimum and maximum values of the corresponding feature category.

Regression model construction

In our experiment, regression predictor is designed using LS-SVR model to solve drift time prediction from peptide sequence-based features. Based on the description of peptide representation, the LS-SVR model for predicting peptide drift time are constructed on a vector consisted of four sequence-based features, of which MW is of with 1 dimension, SL with 1 dimension and AAC with 20 dimensions. For PseAAC, the dimension is 12 for we employed 2-tier sequence correlation pattern with 6 amino acid characters. As a result, each peptide is represented in the predictor by a 34-component vector. For the peptide datasets, i.e., DataS, DataD and DataT, we construct three LS-SVR model for each dataset because the determinative effect of charge state to ion mobility.

Cross-validation

To evaluate the prediction performance of each regression model, a 10-fold cross-validation strategy was adopted for regression analysis. Specifically for singly-charged peptides, DataS is randomly partitioned into 10 sub-datasets, of which a single sub-dataset is retained as the validation data for testing the model, and the remaining 9 sub-datasets are used as training data. After training processes were finished, the LS-SVR model can be applied to the prediction task. This process is then repeated 9 times with each of the ten sub-datasets used exactly once as the testing data. The 10 results from the folds are combined to evaluate the prediction performance.

Author details
[1]The Advanced Research Institute of Intelligent Sensing Network, Tongji University, shanghai, 201804, China. [2]The Key Laboratory of Embedded System and Service Computing, Ministry of Education, Tongji University, Shanghai 201804, China. [3]School of Electronics and Information Engineering, Tongji University, Shanghai 201804, China. [4]School of Electronic Engineering & Automation, Anhui University, Hefei, Anhui 230601, China. [5]Computer, Electrical and Mathematical Sciences and Engineering Division, King Abdullah University of Science and Technology, Thuwal, 23955-6900, Saudi Arabia. [6]Department of Medicine, University of Louisville, Louisville, KY 40202, USA.

References

1. Henderson SC, Valentine SJ, Counterman AE, Clemmer DE: **ESI/ion trap/ion mobility/time-of-flight mass spectrometry for rapid and sensitive analysis of biomolecular mixtures.** *Anal Chem* 1999, **71**(2):291-301.
2. Hoaglund CS, Valentine SJ, Sporleder CR, Reilly JP, Clemmer DE: **Three-dimensional ion mobility/TOFMS analysis of electrosprayed biomolecules.** *Anal Chem* 1998, **70**(11):2236-2242.
3. Kanu AB, Wu C, Hill HH Jr: **Rapid preseparation of interferences for ion mobility spectrometry.** *Anal Chim Acta* 2008, **610**(1):125-134.
4. Harry EL, Weston DJ, Bristow AW, Wilson ID, Creaser CS: **An approach to enhancing coverage of the urinary metabonome using liquid chromatography-ion mobility mass spectrometry.** *J Chromatogr B Analyt Technol Biomed Life Sci* 2008, **871**(2):357-361.
5. Budimir N, Weston DJ, Creaser CS: **Analysis of pharmaceutical formulations using atmospheric pressure ion mobility spectrometry combined with liquid chromatography and nano-electrospray ionisation.** *Analyst* 2007, **132**(1):34-40.
6. Li H, Giles K, Bendiak B, Kaplan K, Siems WF, Hill HH Jr: **Resolving structural isomers of monosaccharide methyl glycosides using drift tube and traveling wave ion mobility mass spectrometry.** *Anal Chem* 2012, **84**(7):3231-3239.
7. Mechref Y, Hu Y, Garcia A, Hussein A: **Identifying cancer biomarkers by mass spectrometry-based glycomics.** *Electrophoresis* 33(12):1755-1767.
8. Zinnel NF, Pai PJ, Russell DH: **Ion mobility-mass spectrometry (IM-MS) for top-down proteomics: increased dynamic range affords increased sequence coverage.** *Anal Chem* 2012, **84**(7):3390-3397.
9. Zhong Y, Hyung SJ, Ruotolo BT: **Ion mobility-mass spectrometry for structural proteomics.** *Expert Rev Proteomics* 2012, **9**(1):47-58.
10. Uetrecht C, Rose RJ, van Duijn E, Lorenzen K, Heck AJ: **Ion mobility mass spectrometry of proteins and protein assemblies.** *Chem Soc Rev* **39**(5):1633-1655.
11. Kanu AB, Dwivedi P, Tam M, Matz L, Hill HH Jr: **Ion mobility-mass spectrometry.** *J Mass Spectrom* 2008, **43**(1):1-22.
12. Wang B, Valentine S, Plasencia M, Raghuraman S, Zhang X: **Artificial neural networks for the prediction of peptide drift time in ion mobility mass spectrometry.** *BMC Bioinformatics* 2010, **11**:182.
13. van Duijn E, Barendregt A, Synowsky S, Versluis C, Heck AJ: **Chaperonin complexes monitored by ion mobility mass spectrometry.** *J Am Chem Soc* 2009, **131**(4):1452-1459.
14. Thalassinos K, Grabenauer M, Slade SE, Hilton GR, Bowers MT, Scrivens JH: **Characterization of phosphorylated peptides using traveling wave-based and drift cell ion mobility mass spectrometry.** *Anal Chem* 2009, **81**(1):248-254.
15. Venne K, Bonneil E, Eng K, Thibault P: **Improvement in peptide detection for proteomics analyses using NanoLC-MS and high-field asymmetry waveform ion mobility mass spectrometry.** *Anal Chem* 2005, **77**(7):2176-2186.
16. Williams JP, Scrivens JH: **Coupling desorption electrospray ionisation and neutral desorption/extractive electrospray ionisation with a travelling-wave based ion mobility mass spectrometer for the analysis of drugs.** *Rapid Commun Mass Spectrom* 2008, **22**(2):187-196.
17. Verbeck GF, Ruotolo BT, Gillig KJ, Russell DH: **Resolution equations for high-field ion mobility.** *J Am Soc Mass Spectrom* 2004, **15**(9):1320-1324.
18. Baker ES, Tang K, Danielson WF, Prior DC, Smith RD: **Simultaneous fragmentation of multiple ions using IMS drift time dependent collision energies.** *J Am Soc Mass Spectrom* 2008, **19**(3):411-419.
19. Hoaglund-Hyzer CS, Counterman AE, Clemmer DE: **Anhydrous protein ions.** *Chem Rev* 1999, **99**(10):3037-3080.
20. Valentine SJ, Counterman AE, Clemmer DE: **A database of 660 peptide ion cross sections: use of intrinsic size parameters for bona fide predictions of cross sections.** *J Am Soc Mass Spectrom* 1999, **10**(11):1188-1211.
21. Liu XH, Liang J, Fan JC, Shang ZC: **Prediction of Ion Drift Times for a Proteome-Wide Peptide Set Using Partial Least Squares Regression, Least-Squares Support Vector Machine and Gaussian Process.** *Qsar & Combinatorial Science* 2009, **28**(11-12):1386-1393.
22. Shah AR, Agarwal K, Baker ES, Singhal M, Mayampurath AM, Ibrahim YM, Kangas LJ, Monroe ME, Zhao R, Belov ME, *et al*: **Machine learning based prediction for peptide drift times in ion mobility spectrometry.** *Bioinformatics* **26**(13):1601-1607.
23. Zhang Y, Jin Q, Wang S, Ren R: **Modeling and prediction of peptide drift times in ion mobility spectrometry using sequence-based and structure-based approaches.** *Comput Biol Med* **41**(5):272-277.
24. Wang B, Valentine S, Raghuraman S, Plasencia M, Zhang X: **Prediction of peptide drift time in ion mobility-mass spectrometry.** *BMC Bioinformatics* 2009, **10**.
25. Wang B, Valentine S, Plasencia M, Zhang XA: **Prediction of Drift Time in Ion Mobility-Mass Spectrometry Based on Peptide Molecular Weight.** *Protein and Peptide Letters* 2010, **17**(9):1143-1147.
26. Wang B, Valentine S, Plasencia M, Zhang X: **Prediction of drift time in ion mobility-mass spectrometry based on Peptide molecular weight.** *Protein Pept Lett* 2010, **17**(9):1143-1147.
27. Vapnik VN: **The nature of statistical learning theory.** New York: Springer; 1995.
28. Wang B, Chen P, Huang DS, Li JJ, Lok TM, Lyu MR: **Predicting protein interaction sites from residue spatial sequence profile and evolution rate.** *Febs Letters* 2006, **580**(2):380-384.
29. Wang B, Wong HS, Huang DS: **Inferring protein-protein interacting sites using residue conservation and evolutionary information.** *Protein and Peptide Letters* 2006, **13**(10):999-1005.
30. Chen P, Wang B, Wong HS, Huang DS: **Prediction of protein B-factors using multi-class bounded SVM.** *Protein Pept Lett* 2007, **14**(2):185-190.
31. Huang DS, Zheng CH: **Independent component analysis-based penalized discriminant method for tumor classification using gene expression data.** *Bioinformatics* 2006, **22**(15):1855-1862.
32. Zheng CH, Zhang L, Ng VT, Shiu SC, Huang DS: **Molecular pattern discovery based on penalized matrix decomposition.** *IEEE/ACM Trans Comput Biol Bioinform* **8**(6):1592-1603.
33. Chen P, Liu C, Burge L, Li J, Mohammad M, Southerland W, Gloster C, Wang B: **DomSVR: domain boundary prediction with support vector regression from sequence information alone.** *Amino Acids* **39**(3):713-726.
34. Suykens JAK: **Least squares support vector machines.** River Edge, NJ: World Scientific; 2002.
35. Khemchandani R, Chandra S: **Regularized least squares fuzzy support vector regression for financial time series forecasting.** *Expert Systems with Applications* 2009, **36**(1):132-138.
36. Kavaklioglu K: **Modeling and prediction of Turkey's electricity consumption using Support Vector Regression.** *Applied Energy* 2011, **88**(1):368-375.
37. Hong WC: **Chaotic particle swarm optimization algorithm in a support vector regression electric load forecasting model.** *Energy Conversion and Management* 2009, **50**(1):105-117.
38. Fernandez M, Miranda-Saavedra D: **Genome-wide enhancer prediction from epigenetic signatures using genetic algorithm-optimized support vector machines.** *Nucleic Acids Res* **40**(10):e77.
39. Zhang QL, Shan GL, Duan XS, Zhang ZN: **Parameters Optimization of Support Vector Machine based on Simulated Annealing and Genetic Algorithm.** *2009 Ieee International Conference on Robotics and Biomimetics (Robio 2009), Vols 1-4* 2009, 1302-1306.
40. Karasek FW, Hill HH Jr, Kim SH: **Plasma chromatography of heroin and cocaine with mass-identified mobility spectra.** *J Chromatogr* 1976, **117**(2):327-336.
41. Tuovinen K, Paakkanen H, Hänninen O: **Detection of pesticides from liquid matrices by ion mobility spectrometry.** *Analytica Chimica Acta* 2000, **404**(1):7.