BIOINFORMATICS ORIGINAL PAPER
Vol. 26 no. 13 2010, pages 1601–1607
doi:10.1093/bioinformatics/btq245

Sequence analysis
Advance Access publication May 21, 2010

Machine learning based prediction for peptide drift times in ion mobility spectrometry

Anuj R. Shah[1], Khushbu Agarwal[1], Erin S. Baker[1], Mudita Singhal[1], Anoop M. Mayampurath[2], Yehia M. Ibrahim[1], Lars J. Kangas[3], Matthew E. Monroe[1], Rui Zhao[1], Mikhail E. Belov[1], Gordon A. Anderson[1] and Richard D. Smith[1,*]

[1]Fundamental and Computational Sciences Directorate, Pacific Northwest National Laboratory, 999 Battelle Boulevard, Richland, WA 99352, [2]School of Informatics, Indiana University, Bloomington, IN 47408 and [3]National Security Directorate, Pacific Northwest National Laboratory, 999 Battelle Boulevard, Richland, WA 99352, USA

Associate Editor: John Quackenbush

ABSTRACT

Motivation: Ion mobility spectrometry (IMS) has gained significant traction over the past few years for rapid, high-resolution separations of analytes based upon gas-phase ion structure, with significant potential impacts in the field of proteomic analysis. IMS coupled with mass spectrometry (MS) affords multiple improvements over traditional proteomics techniques, such as in the elucidation of secondary structure information, identification of post-translational modifications, as well as higher identification rates with reduced experiment times. The high throughput nature of this technique benefits from accurate calculation of cross sections, mobilities and associated drift times of peptides, thereby enhancing downstream data analysis. Here, we present a model that uses physicochemical properties of peptides to accurately predict a peptide's drift time directly from its amino acid sequence. This model is used in conjunction with two mathematical techniques, a partial least squares regression and a support vector regression setting.

Results: When tested on an experimentally created high confidence database of 8675 peptide sequences with measured drift times, both techniques statistically significantly outperform the intrinsic size parameters-based calculations, the currently held practice in the field, on all charge states (+2, +3 and +4).

Availability: The software executable, imPredict, is available for download from http://omics.pnl.gov/software/imPredict.php

Contact: rds@pnl.gov

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on January 6, 2010; revised on April 18, 2010; accepted on May 2, 2010

1 INTRODUCTION

The capability of ion mobility spectrometry coupled with mass spectrometry (IMS-MS) to characterize biological mixtures has been extensively reported (Valentine, Counterman *et al.*, 1998; Henderson, Valentine *et al.*, 1999). IMS offers fast separation times (Baker *et al.*, 2009) as well as additional information on ion structure that can assist e.g. identification of post-translational modifications (Ruotolo, Gillig *et al.*, 2004) as well as gas-phase collision cross sections of peptides (Mason, McDaniel 1988). Determination of collision cross section can aid structural studies and provide clues to understanding protein function. Calculating the collision cross section also facilitates peptide identification and allows the reduction in false discovery rates (FDRs).

Valentine *et al.* (1999a) investigated the use of intrinsic size parameters (ISPs) for prediction of peptide ion cross sections in IMS. While the results reported for this technique were promising, the application of such a technique to high throughput data analysis is challenging due to modeling the geometry of the ion which is required to solve a series of transport equations to determine the potential of interaction between the buffer gas atoms and the ion. One of the first efforts for computationally calculating collision cross sections and ion mobility drift times was based on determination of the ISPs of amino acid residues using calculations that were based on the stoichiometry of each residue (Shvartsburg *et al.*, 2001). The ISP technique is based on the assumption that all peptides are closely packed without major internal cavities and are near spherical in shape. As a result, the volume (and thus the cross section) of a peptide is directly proportional to the sum of the areas of the constituent atoms divided by the total mass. The use of the ISP technique facilitates computational calculation of collision cross sections and hence allows determination of ion mobilities and peptide drift times. The prediction of peptide drift times has practical application in peptide identification and may assist in the reduction of FDRs in high throughput IMS experiments. The ISP method was tested on 271 singly protonated peptides from the original 660 peptides published in Valentine *et al.* (1999b). While the results of this calculation are accurate for this dataset, the technique has not been tested on datasets containing higher charge states peptides. A recent meeting abstract (Wang *et al.*, 2009) uses an artificial neural network to predict peptide drift times directly without calculating cross sections and claims prediction performance of over 90% on charges +1 (212 peptides) and +2 (306 peptides) while reporting a 75% performance on charge +3 data (77 peptides). Prediction performance in a regression setting is calculated by selecting an error threshold and counting predictions within the error threshold as true predictions and those outside the threshold as false predictions. A more recent method proposed by Liu *et al.* (2009) developed quantitative structure–property relationships (QSPR) to predict the drift times for a set of 1481 peptides using information derived from molecular structures. Additionally, they

*To whom correspondence should be addressed.

also investigated the use of multiple mathematical techniques such as partial least squares regression (PLSR), least-squares support vector machines (LS-SVM) (Suykens and Vandewalle, 1999) and a Gaussian Process (GP) to predict drift times. They concluded that diversified properties (particularly structural topological information and charge distribution) contribute to the peptide drift time and an LS-SVM and GP coupled with a genetic algorithm for feature selection was capable of capturing the linear and non-linear relationships between the QSPRs. In this article, we present a model consisting of a set of amino acid properties that helps accurately predict the ion mobility drift times for a large experimentally derived database of peptides with varying lengths and charge states, using partial least squares (PLS) and support vector regression (SVR) based methods. The novelty of our method is the determination of ion mobility drift times from physicochemical properties that can be directly calculated from a peptide amino acid sequence. Additionally, our method is computationally efficient and requires no manual determination of molecular structures, thereby making it suitable for high throughput data analysis pipelines. Both the PLS and SVR techniques provide significantly higher prediction performance when compared to the ISP-based technique proposed previously (Shvartsburg *et al.*, 2001) when tested on a database of 8675 peptides generated at the proteomics facility at Pacific Northwest National Laboratory.

2 METHODS

2.1 IMS-MS measurements for experimental dataset creation

IMS provides a powerful tool to separate ions based on their 3D shape. IMS distinguishes ions based on the fact that different ion shapes and charge states travel at different velocities when pulled by a weak electric field through a drift cell filled with an inert buffer gas. In order to quantitatively measure the drift time, mobility and collision cross section of an ion, experiments were conducted using an IMS-TOF MS instrument (Baker, Clowers *et al.*, 2007). The ions were analyzed by passing them in packets into a 98-cm long drift cell filled with ∼4 Torr of ultrapure nitrogen buffer gas. Once in the drift cell, a uniform electric field gently pulled the ions through the buffer gas where they quickly reached equilibrium between the forward acceleration force imposed by the electric field and the frictional drag force from the buffer gas. This causes the ions to drift at constant velocity, v, proportional to the applied field Eas shown in Equation (1) (Mason, McDaniel 1988):

$$v_d = K \cdot E \tag{1}$$

where the proportionality constant, K (in cm^2/V·s), is termed the mobility of the ions. Because K is dependent on the number density of the buffer gas (the number of objects per specified volume), it is usually standardized with respect to pressure and temperature and termed reduced mobility, K_o. The reduced mobilities were determined by collecting arrival time distributions (ATDs) at four different electric field voltages. The drift time, t_D, of a particular ion was extracted from the center of the ATD peak and can be written as:

$$t_D = \frac{l^2}{K_o} \cdot \frac{273.16}{760T} \cdot \frac{p}{V} + t_o \tag{2}$$

where l is the length of the drift cell, V is the voltage drop across the cell, p is the pressure and t_o is the time the ion spends between the exit of the drift cell and the MS detector. The expression for t_D in Equation (2) has a linear dependence on p/V where the slope of the line is inversely proportional to K_o and the y-intercept is equal to t_o. To acquire K_o, t_D is extracted from the center of the ATD peak and plotted against p/V. The slope of this linear fit is used to calculate K_o and from the reduced mobility, the collision cross section can be determined. The relationship between the mobility of an ion and its collision cross section has been derived in detail using kinetic theory (McDaniel and Mason, 1973) and is given by:

Table 1. Distribution of peptide counts based on charge states

Charge state	+1	+2	+3	+4	+5	+6
Number of peptides	14	3933	3916	717	90	5

$$K_o = \frac{3\mathrm{q}}{16N_o} \cdot \left(\frac{2\pi}{\mu k_b T}\right)^{1/2} \cdot \frac{1}{\Omega} \tag{3}$$

where q is the ion charge, N_o is the buffer gas density at standard temperature and pressure (STP), μ is the reduced mass of the collision partners, k_b is Boltzmann's constant and Ω is the momentum transfer collision integral also termed the collision cross section.

Using the linear dependence of Equation (2) on p/V, four different measurements were conducted on individual samples from human plasma, mouse plasma and Shewenella Oneidensis MR-1 at 4T pressure and voltages varying from 1.6, 1.8, 2.0 and 2.2 kV. Individual raw data files were then de-isotoped using Decon2LS (Jaitly *et al.*, 2009), and the identified isotopic distributions were clustered using VIPER (Monroe *et al.*, 2007), aligned (Jaitly *et al.*, 2006) and peak matched to existing Accurate Mass and Time (AMT) (Pasa-Tolic *et al.*, 2004) tag databases for the respective organisms to identify peptides using default parameters for all software and only the mass and elution time dimensions. The drift time for each LC-MS feature was determined by selecting the maximum abundance IMS scan across the entire LC scan range for a feature. Using a consensus approach, only peptides that were observed in at least three out of the four experiments (for each sample) were considered to be part of the final dataset. An additional condition was placed on the experimentally measured drift times for these peptides where the drift times were expected to be linear with p/V across different experiments to the tune of a cross-correlation coefficient value of 0.9999. Lastly, all drift times reported in the database were standardized to a pressure of 4.0T and a voltage of 1800 V.

The final dataset, downloadable as Supplementary Material 1, consists of 8675 peptide sequences of varying lengths (6–42 residues) populated with measured collision cross sections and drift times resulting in an overall FDR of 3.2% (refer Supplementary Material 2 for details on FDR estimation for this dataset). The FDR calculated above is a measure of confidence associated with each peptide in the final dataset. Table 1 shows a distribution of the number of peptides present for each charge state. Robust models of prediction were built only for +2, +3 and +4 charge states. Charge states +1 and +5 or higher are very sparse and we believe the peptides present in these datasets do not exhibit sufficient sequence variability to develop comprehensive drift time prediction models that would perform adequately on novel peptides. Additional experiments on complex biological mixtures containing higher charged species will result in adequate training data for such models.

2.2 PLSR

The PLS approach is widely applied in the field of chemometrics (Wold *et al.*, 2001), in sensory evaluation (Ortiz *et al.*, 2006) and more recently to predict drift times in IMS (Liu *et al.*, 2009). The PLS technique is a method for relating two matrices, X (the independent variable) and Y (the response variable) by a linear multivariate model, but goes beyond traditional regression in that it models also the structure of X and Y. The general underlying model of multivariate PLS is given by

$$\begin{aligned} X &= \mathrm{TP}^T + E \\ Y &= \mathrm{TQ}^T + F \end{aligned} \tag{4}$$

where X is an $n \times n$ matrix of predictors, Y is an $n \times p$ matrix of responses, T is a $n \times l$ matrix (the *score*, *component* or *factor* matrix), P and Q are, respectively, $m \times l$ and $p \times l$ loading matrices and matrices E and F are the error terms, assumed to be i.i.d. normal.

2.3 SVR

The application of support vector machines (SVMs) in classification problems is well studied within the wider bioinformatics community. Numerous examples exist in literature where SVMs are used as a classification technique with state-of-the-art prediction accuracies and as such a detailed discussion here would be redundant. The application of SVMs as a technique for high-dimensional regression provides the ability to predict real values with high precision and is relatively less common in literature. We select ε-SVR as our intention is to not penalize prediction errors that are below the resolution of the instrument setup when measuring drift times. Additionally, the use of an SVR method provides a predictive probability interval based on the framework suggested by Lin and Weng (2004). As formulated in Vapnik (1998), the ε-SVR problem can be explained as follows. Given a set of training data available to us $\Gamma = \{x_i, y_i\}_{i=1}^{N}$ where $y_i = 1, \ldots, N$ are continuous output values, the goal of ε-SVR is to approximate a linear function of the form

$$f(x) = <w, x> + b \tag{5}$$

with $w \in R^N$ and $b \in R$ that has at most ε deviation from all possible targets y_i and $<.,.>$ denotes a dot product in R^N. To ensure that we do not accept any deviations larger than ε, one can minimize the norm,

$$\text{i.e.} \, \|w\|^2 = <w, w>$$

As outlined in Smola and Scholkopf (2004), one can write this as a convex optimization problem with the introduction of slack variables ξ, ξ_i^* to cope with otherwise infeasible problems to minimize

$$\begin{aligned} & \tfrac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}(\xi_i + \xi_i^*) \\ \text{subject to} \quad & y_i - \langle w, x_i \rangle - b \le \varepsilon + \xi_i \\ & \langle w, x_i \rangle + b - y_i \le \varepsilon + \xi_i^* \\ & \xi_i, \xi_i^* >= 0 \end{aligned} \tag{6}$$

The constant C determines the trade-off between the flatness of f and the amount up to which errors larger than ε are tolerated. The selection of C and ε has significant effect on the performance of the SVR. The value of C is derived from the training data using the following equations as suggested by Cherkassky and Ma (2004)

$$C = \max(|\bar{y} + 3\sigma_y|, |\bar{y} - 3\sigma_y|) \tag{7}$$

where $\bar{y}$ and σ_y are the mean and standard deviation of the drift times from the entire dataset. The value of ε is selected based on the resolution of the instrument setup for drift times. When compared to a PLSR technique, the SVR is a more robust method that is capable of modeling both linear as well as non-linear relationships between the feature vector and the target values. The robustness comes at a computational cost when training the algorithm as it involves solving a quadratic programming problem. The testing phase of the algorithm or prediction for new peptides is a linear calculation and a fast operation. Using a least squares non-linear SVM technique (Suykens and Vandewalle, 1999), also referred commonly as ridge regression (Hoerl, 1962), could avoid much of the computation as it involves solving a linear programming problem as opposed to a quadratic one.

2.3.1 Peptide vectorization A common characteristic between both the PLS and the SVR technique is the modeling of the independent variable matrix X that is a certain dimensions in length and is commonly referred to as a feature vector. This requirement dictates that each peptide be represented as a fixed-length n-dimensional vector of properties. Each dimension in this fixed-length vector plays a role in the final determination of the ion mobility drift times for these peptides. Our feature vector is a combination of the reversed phase high performance liquid chromatography (LC) peptide elution time (normalized on a scale of zero to unity), the physicochemical properties of a peptide and a structural representation of its amino acid composition. The normalized elution time (NET) of a peptide represents the time a peptide is retained within a high-performance reverse phase liquid chromatography separations column. We use a software utility that predicts the NET directly from the amino acid composition of a peptide. Details of this algorithm are further described in Petritis *et al.* (2003) while the software is freely available from our web site: http://omics.pnl.gov/software/NETPredictionUtility.php. The physicochemical properties of a peptide/protein sequence have been successfully used to predict subcellular localization of protein sequences (Garg *et al.*, 2009; Tantoso and Li, 2008), protein–protein interactions (Agrawal *et al.*, 2005; Bock and Gough, 2001; Nanni and Lumini, 2006), promoter regions (Uren *et al.*, 2006), long disordered regions (Hirose *et al.*, 2007) and sequence homology (Yang *et al.*, 2008) among others. Our selection of these properties is a variant of a model used for predicting proteotypic peptides (peptides that are likely to be observed in MS-based experiments) (Webb-Robertson *et al.*, 2008). In order to encode the structural composition of amino acid residues we have tried multiple encodings from using the output of predictive algorithms such as PredictProtein (Rost *et al.*, 2004) to profile-based string kernel representations (Kuang *et al.*, 2004). However, a simplistic encoding scheme, that divides the peptides into fifths and accommodates the tail end, achieves the highest levels of accuracy. Each chunk is encoded as a 20-dimensional vector (a dimension for every amino acid residue) resulting in 100 dimensions for the chunks and 20 dimensions for the tail (the left-over residues after dividing by 5), encoding a histogram of amino acid distributions. The granularity of this representation (number of chunks and chunk size) can be increased and we have experimented with different chunk sizes, with the best performance reported below. We also investigated the use of the ISP parameter as a component of the feature vector. The information content within the ISP is captured by the mass, length and the peptide chunks and as a result did not see any significant improvement in the model performance. Table 2 outlines the final set of features selected that model diverse peptide properties directly derived from a peptide sequence and provide the best prediction performance. The individual features are on different scales and hence the feature vectors are normalized to have zero mean and unit variance. The Supplementary Material 3 contain the details on the scales of the different features.

Table 2. Peptide features for drift time prediction

Index	Feature description
1	Molecular weight
2	Normalized elution time (Petritis *et al.*, 2006)
3	Peptide length
4	Gas phase basicity (Zhang, 2004)
5	Number of non-polar hydrophobic residues
6	Number of uncharged polar hydrophilic residues
7	Number of positively charged polar hydrophilic residues
8	Number of negatively charged polar hydrophilic residues
9	Hydrophobicity—Eisenberg scale (Eisenberg et al., 1984)
10	Hydrophilicity—Hopp–Woods Scale (Hopp and Woods, 1983)
11	Hydropathicity—Roseman scale (Roseman, 1988)
12	Polarity—Zimmerman polarity (Zimmerman *et al.*, 1968)
13	Bulkiness (Zimmerman *et al.*, 1968)
14–134	120 dimensional encoding for structure

3 RESULTS AND DISCUSSION

We have developed a computational method, imPredict, to predict the ion mobility drift times for polypeptide ions of modest size (e.g. ≤50-mer) directly from their amino acids compositions for +2, +3 and +4 charge states using two mathematical techniques, PLSR and SVR. We compare the prediction performance of both these techniques with the ISP (Shvartsburg *et al.*, 2001) technique, on an experimentally derived dataset.

3.1 Predicting ion mobility drift times

We compared the results of imPredict using both SVR and PLS with the ISP technique for three charge states (+2 to +4). The results reported herein are using a 5-fold cross validation (CV) technique for the PLS and SVR prediction models while the ISP technique was re-implemented in software. The ISP for a residue is directly related to its density (average mass/size ratio of its constituent atoms) and is easily calculated by the following formula as suggested by (Shvartsburg *et al.*, 2001).

where r_i are the radii of the atoms and m_i are their masses. The ISP is related to the cross section of the peptide sequence by a linear equation presented in Valentine *et al.* (1999b). The drift time for the peptide can then be calculated using the theory presented in Section 2.1. The results of the ISP technique have only been published thus far on a singly protonated dataset containing 263 peptides. This is the first report of that technique on higher charge state data.

The n-fold CV technique, as suggested by Salzberg (1997), assures an un-biased evaluation for learning algorithms by randomly dividing the entire dataset in n ($n=5$ in our case) distinct subparts. Of the n subparts, a single subpart is retained as the validation data for testing the model, and the remaining $n-1$ parts are used as the training data. The CV process is then repeated n times, with each of the parts used exactly once as the validation data. We use the plsregress method, available as part of the statistics toolbox in Matlab R2008a, to test the PLSR method. Using minimum mean squared error (MSE) as the determining factor, the optimal number of components was determined to be 30, 44 and 9 for the respective charge states. For the ε-SVR models, the libsvm software library (Chang and Lin, 2001) was used with a linear kernel. The choice of epsilon SVR as opposed to a Huber loss function is motivated by a rigorous evaluation described here (Cherkassky and Ma, 2004). Using Equation (7), the C values for the respective charge state models (+2, +3 and +4) are 34.75, 33.5 and 33.36 while the ε values were constant at 0.5 based on the drift time resolution of the instrument setup. Additionally, we used the probability estimates as produced by the libSVM software to generate predictive intervals for each target drift time for downstream data analysis.

Figure 1 displays a comparative analysis of the fraction of peptides that can be correctly predicted based on various percentage error threshold levels using all techniques. The error residuals are calculated as a fraction of the difference in measured and predicted value to the measured value. When multiplied by hundred, they give the percentage error for each prediction. Methods for charge +2 carry a dot as marker, for charge +3 carry a triangle as marker and for charge +4 carry a square. The ISP parameter technique is represented by a dotted line, the PLS based method with a long dash and dash line while the SVR based method is denoted with a dash and dot line. The figure resembles a cumulative receiver operating characteristic curve and aims to depict the number of peptides that can be accurately predicted at a given error threshold.

A higher curve (larger y intercept) indicates a larger number of peptides with predicted drift times that have errors that are smaller than the supplied threshold (x-axis) value. As can be seen from the figure, the imPredict curves yield more accurate results at all threshold levels when compared to the ISP method. Additionally, one can observe that the curves for the imPredict methods using PLS and SVR closely follow each other suggesting that the two techniques yield comparable performance. To further distinguish between the three methods, we report other measures of accuracy such as the squared value of the Pearson's correlation coefficient (Rodgers and Nicewander, 1988), generally represented as R^2 and the MSE. The Pearson's coefficient is a standard method of estimating the degree to which two series are correlated while the MSE is another way to quantify the difference between the actual value and the predicted value for an estimator.

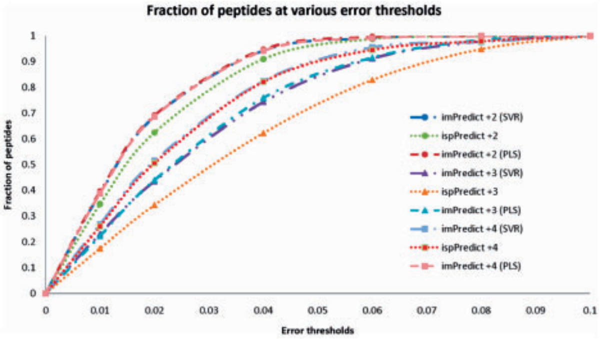

Fig. 1. Fraction of peptide ions correctly predicted at different threshold values. Methods for charge +2 carry a dot as marker, for charge +3 carry a triangle as marker and for charge +4 carry a square. The ISP parameter technique is represented by a dotted line, the PLS-based method with a long dash and dash line while the SVR-based method is denoted with a dash and dot line. A higher curve indicates a larger number of peptides for a given threshold value.

Table 3. Pearson's correlation coefficient and MSE for different prediction techniques

Charge state	+2	+3	+4
imPredict (PLS)	96.20 (0.290)	90.62 (0.813)	93.08 (0.556)
imPredict (SVR)	96.00 (0.308)	90.16 (0.869)	89.63 (0.722)
ISP	95.03 (0.396)	87.05 (1.289)	89.06 (0.866)

Table 3 presents the Pearson's correlation coefficient (R^2) and the MSE (in brackets) for the predicted drift times when compared to the experimentally measured drift times. A higher value on the Pearson's co-efficient and a lower value of the MSE indicate a better predictor. In all cases, the PLS-based method yields the best value, the SVR is a close second and the ISP technique is the worst performer. Figure 2 plots the histograms for percentage error when predicting drift times for peptides with charges +2 (a), +3 (b) and +4 (c). A higher and narrower curve indicates a larger number of peptides with smaller percentage errors and is reflective of better performance and higher accuracy levels. From Figure 2a and Table 3, one can conclude that all three methods predict the charge +2 peptides to high levels of accuracy. The PLS method outperforms the other two methods and the near-Gaussian nature of the error distributions suggests a strong linear dependence of drift times on the model. The charges +3 and +4, on the other hand, are difficult to predict and the MSEs on these peptides are higher for all methods. For the charge +3 peptides,

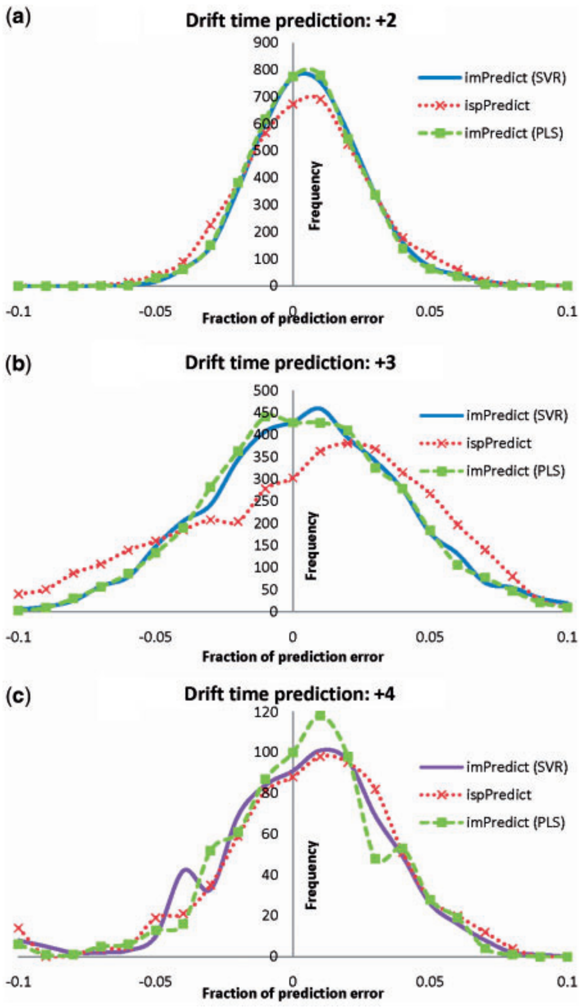

Fig. 2. Fraction of prediction error histograms for all techniques on different charge state peptides. The red dotted line with crosses represents the ISP algorithm, the dashed line with squares is for the the PLS technique and the solid line represents the SVR technique. Different charge states are plotted on different graphs for clarity. The x-axis values when multiplied by 100 gives the percentage error for those prediction models.

Figure 2b, the ISP and the SVR methods tend to overestimate the drift times while the PLS method slightly underestimates them. As far as the charge +4 peptides are concerned, all three methods overestimate the drift times (Fig. 2c) and once again the PLS method is the best performer.

3.2 Feature selection

The quest for an optimal set of properties modeling the drift times is an ongoing research effort and a never ending one as well. However, of the given set of properties described above, it is important to discern which features play a critical role in the modeling of ion mobility drift times. Variable selection on PLS models is a computationally intensive exercise where each feature can be left out and a measure of performance prediction (like MSE) computed to determine whether the feature is an important attribute or otherwise. A similar approach can be used with the SVR model where the MSE and CV R^2 values can be used to determine the contribution of each feature. We use the F-score technique as suggested by Guyon *et al.* (2005) in conjunction with SVR for determination of the importance of each feature. Table 4 presents the five most important features for the three charge states along with their F-score values while Supplementary Material 4 contains the entire F-score tables. The blanks in the table represent a component from the structural encoding. As expected in all three cases, the molecular weight, NET and the length of the peptide are indicated as important attributes.

Table 4. F-score for the top five features for each charge state

Order	Charge +2	Charge +3	Charge +4
1	MW (1342.78)	MW (401.60)	–
2	Length (115.91)	NET (290.37)	–
3	NET (41.15)	Length (121.5)	MW (75820.08)
4	–	Non-polar hydrophobic residues (69.09)	NET (1483.56)
5	Gas phase basicity (26.33)	–	Length (582)

MW, molecular weight.

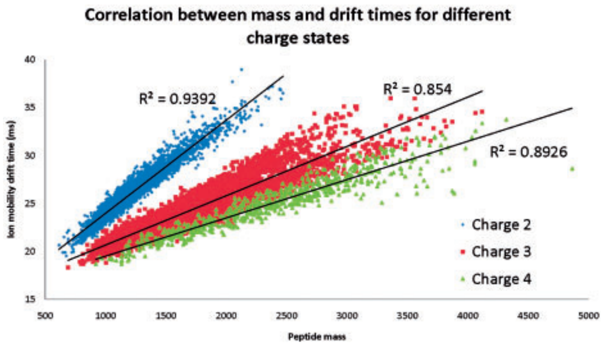

Fig. 3. Correlation between mass and drift times for different peptide charge states. Using peptide mass alone a prediction model can achieve high accuracy (93% on charge 2, 85% on charge 3 and 89% on charge 4) predictions.

Higher charge state models have greater dependence on the other features while models for lower charge states (specifically +2) are highly dependent on the mass of the ions. Figure 3 illustrates the correlation (93% on charge +2, 85% on charge +3 and 89% on charge +4) that can be achieved by a prediction model using only individual peptide masses. The introduction of each additional feature within the machine learning framework incrementally improves the prediction accuracy. Results for +1 and +5 charge state are not shown as there are only 14 and 90 high confidence drift time measurements and they do not represent adequate peptide variability.

3.3 Discussion

Given the increased application of IMS to biological separations and proteomics research, there is a need for a high throughput fast computational method to predict ion mobility drift times.

For one, such a computational utility can be an indispensable tool for extending existing AMT tag databases, built with extensive tandem MS experiments, with the added dimension of ion mobility at no additional cost. High throughput IMS-MS measurements can then be used and a 3D peptide signature (mass, elution time and drift time) can be compared against extended AMT tag databases. The added dimension of separation is expected to reduce false positive matches and help biological inference. The practical challenge with this approach would be the high degrees of prediction accuracies desired from such a prediction utility. While the data for the charge +2 peptides suggest the feasibility of the imPredict approach, charges +3 and +4 present opportunities to improve the performance. As the current imPredict model uses the same set of features to represent a peptide for all charge states, it is possible to expand on this research and include charge-specific properties in the feature set. The superior performance of the PLSR-based method suggests that the drift times have a strong linear dependence on the set of properties. The introduction of more properties depicting peptide structure such as the presence of certain sequence-structural motifs (Bystroff and Baker, 1998) or dipeptide combinations of residues could help improve the prediction accuracies. We believe that the currently generated training data does not exhibit sufficient variability in amino acid residues (especially for charge +4) to investigate sequence-order based feature vectors. To develop a comprehensive utility such as the NET prediction tool (Petritis *et al.*, 2003), one would have to continually evolve the model as more peptides are available for training.

4 CONCLUSIONS

The highly reproducible nature of ion mobility drift times for peptide ions makes IMS-MS platforms highly attractive for high throughput peptide analyses. However, the lack of a purely computational technique to determine these drift times limits downstream high throughput analyses and peptide identifications. Here we present imPredict, a computational technique that provides distinct models for charge states +2, +3 and +4, and produces highly accurate drift times that are statistically significantly better than the ones obtained with the ISP-based technique, the only other computational technique previously reported for accurate calculation of drift times, for all considered charge states.

The imPredict algorithm illustrates the feasibility of a purely computational technique for accurately predicting ion mobility drift times. The high prediction performance attained by the PLSR method suggests required improvement in modeling the drift times for charge +3 and +4.

ACKNOWLEDGEMENTS

The research was performed using EMSL, a national scientific user facility sponsored by the Department of Energy's Office of Biological and Environmental Research and located at Pacific Northwest National Laboratory.

Funding: National Institutes of Health National Center for Research Resources (RR 18522); National Institutes of Health National Cancer Institute (R21 CA12619-01); Washington State Life Sciences Discovery Fund; Laboratory Directed Research and Development Program at Pacific Northwest National Laboratory.

Conflict of Interest: none declared.

REFERENCES

Agrawal,R.K. *et al.* (2005) Predict protein-protein interaction using heuristic approaches. In *3rd International Conference on Intelligent Sensing and Information Processing*. IEEE Computer Society, pp. 93–98.

Baker,E.S. *et al.* (2007) Ion mobility spectrometry-mass spectrometry performance using electrodynamic ion funnels and elevated drift gas pressures. *J. Amer. Soc. Mass Spectrom.*, **18**, 1176–1187.

Baker,E.S. *et al.* (2009) An LC-IMS-MS platform providing increased dynamic range for high-throughput proteomic studies. *J. Proteome Res.*, **9**, 997–1006.

Bock,J.R. and Gough,D.A. (2001) Predicting protein-protein interactions from primary structure. *Bioinformatics*, **17**, 455–460.

Bystroff,C. and Baker,D. (1998) Prediction of local structure in proteins using a library of sequence-structure motifs. *J. Mol. Biol.*, **281**, 565–577.

Chang,C.-C. and Lin,C.-J. (2001) LIBSVM: a library for support vector machines. Available at http://www.csie.ntu.edu.tw/~cjlin/libsvm (last accessed date May 14, 2010).

Cherkassky,V. and Ma,Y. (2004) Practical selection of SVM parameters and noise estimation for SVM regression. *Neural Netw.*, **17**, 113–126.

Eisenberg,D. *et al.* (1984) The hydrophobic moment detects periodicity in protein hydrophobicity. *Proc. Natl Acad. Sci. USA*, **81**, 140–144.

Garg,P. *et al.* (2009) SubCellProt: predicting subcellular localization using machine learning approaches. *In Silico Biol.*, **9**, 35–44.

Guyon,I. *et al.* (eds) (2005) *Combining SVMs with Various Feature Selection Strategies*. Springer, Berlin/Heidelberg.

Hirose,S. *et al.* (2007) POODLE-L: a two-level SVM prediction system for reliably predicting long disordered regions. *Bioinformatics*, **23**, 2046–2053.

Henderson,S.C. *et al.* (1998) ESI/Ion Trap/Ion Mobility/Time-of-Flight mass spectrometry for rapid and sensitive analysis of biomolecular mixtures. *Anal. Chem.*, **71**, 291–301.

Hoerl,A.E. (1962) Application of ridge analysis to regression problems. *Chem. Eng. Prog.*, **58**, 54–59.

Hopp,T.P. and Woods,K.R. (1983) A computer program for predicting protein antigenic determinants. *Mol. Immunol.*, **20**, 483–489.

Jaitly,N. *et al.* (2009) Decon2LS: an open source software package for automated processing and visualization of high resolution mass spectrometry data. *BMC Bioinformatics*, **10**, 87.

Jaitly,N. *et al.* (2006) Robust algorithm for alignment of liquid chromatography-mass spectrometry analyses in an accurate mass and time tag data analysis pipeline. *Anal. Chem.*, **78**, 7397–7409.

Kuang,R. *et al.* (2004) Profile-based string kernels for remote homology detection and motif extraction. In *Computational Systems Bioinformatics Conference (CSB'04)*. IEEE Computer Society, Stanford, CA, pp. 152–160.

Lin,C.-J. and Weng,R.C. (2004) Simple probabilistic predictions for support vector regression. *Technical Report*, Department of Computer Science, National Taiwan University.

Liu,X. *et al.* (2009) 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 Comb. Sci.*, **28**, 1386–1393.

Mason,E.A. and McDaniel,E.W. (1988) *Transport Properties of Ions in Gases*. John Wiley and Sons, New York, p. 560.

McDaniel,E.W. and Mason,E.A. (eds) (1973) *The Mobility and Diffusion of Ions in Gases*. Wiley, New York.

Monroe,M.E. *et al.* (2007) VIPER: an advanced software package to support high-throughput LC-MS peptide identification. *Bioinformatics*, **23**, 2021–2023.

Nanni,L. and Lumini,A. (2006) An ensemble of K-local hyperplanes for predicting protein-protein interactions. *Bioinformatics*, **22**, 1207–1210.

Ortiz,M.C. *et al.* (2006) Sensitivity and specificity of PLS-class modelling for five sensory characteristics of dry-cured ham using visible and near infrared spectroscopy. *Anal. Chim. Acta*, **558**, 125–131.

Pasa-Tolic,L. *et al.* (2004) Proteomic analyses using an accurate mass and time tag strategy. *Biotechniques*, **37**, 621–624, 626–633, 636 passim.

Petritis,K. *et al.* (2003) Use of artificial neural networks for the accurate prediction of peptide liquid chromatography elution times in proteome analyses. *Anal. Chem.*, **75**, 1039–1048, Medium: X.

Petritis,K. *et al.* (2006) Improved peptide elution time prediction for reversed-phase liquid chromatography-MS by incorporating peptide sequence information. *Anal. Chem.*, **78**, 5026–5039.

Rodgers,J.L. and Nicewander,W.A. (1988) Thirteen ways to look at the correlation coefficient. *Am. Statist.*, **42**, 59–66.

Roseman,M.A. (1988) Hydrophobicity of the peptide C=O...H-N hydrogen-bonded group. *J. Mol. Biol.*, **201**, 621–623.

Rost,B. *et al.* (2004) The PredictProtein server. *Nucleic Acids Res.*, **32**, W321–W326.

Ruotolo,B.T. *et al*. (2002) Analysis of protein mixtures by matrix-assisted laser desorption ionization-ion mobility-orthogonal-time-of-flight mass spectrometry. *Int. J. Mass Spectrom.*, **219**, 253–267.

Salzberg,S. (1997) On comparing classifiers: pitfalls to avoid and a recommended approach. *Data Min Knowl. Discov.*, **1**, 317–328.

Shvartsburg,A.A. *et al.* (2001) Prediction of peptide ion mobilities via a priori calculations from intrinsic size parameters of amino acid residues. *J. Am. Soc. Mass Spectrom.*, **12**, 885–888.

Smola,A.J. and Scholkopf,B. (2004) A tutorial on support vector regression. *Stat. Comput.*, **14**, 199–222.

Suykens,J.A.K. and Vandewalle,J. (1999) Least squares support vector machine classifiers. *Neural Proc. Lett.*, **9**, 293–300.

Tantoso,E. and Li,K.-B. (2008) AAIndexLoc: predicting subcellular localization of proteins based on a new representation of sequences using amino acid indices. *Amino Acids*, **35**, 345–353.

Uren,P. *et al.* (2006) Promoter prediction using physico-chemical properties of DNA. In *Computational Life Sciences II*. Vol. 4216 of *Lecture Notes in Bioinformatics*, Lecture Notes in Computer Science, Springer, Berlin, New York, pp. 21–31.

Valentine,S.J. *et al*. (1998) Gas-phase separations of protease digests. *J. Amer. Soc. Mass Spectrom.*, **9**, 1213–1216.

Valentine,S.J. *et al.* (1999a) 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.*, **10**, 1188–1211.

Valentine,S.J. *et al.* (1999b) Intrinsic amino acid size parameters from a series of 113 lysine-terminated tryptic digest peptide ions. *J. Phys. Chem. B*, **103**, 1203–1207.

Vapnik,V. (1998) *The Nature of Statistical Learning*. Springer, New York.

Wang,B. *et al.* (2009) Prediction of peptide drift time in ion mobility-mass spectrometry. *BMC Bioinformatics*, **10**, A1.

Webb-Robertson,B.-J.M. *et al.* (2008) A support vector machine model for the prediction of proteotypic peptides for accurate mass and time proteomics. *Bioinformatics*, **24**, 1503–1509.

Wold,S. *et al.* (2001) PLS-regression: a basic tool of chemometrics. *Chemometrics Intell. Lab. Syst.*, **58**, 109–130.

Yang,Y. *et al.* (2008) Remote protein homology detection using recurrence quantification analysis and amino acid physicochemical properties. *J. Theor. Biol.*, **252**, 145–154.

Zhang,Z. (2004) Prediction of low-energy collision-induced dissociation spectra of peptides. *Anal. Chem.*, **76**, 3908–3922.

Zimmerman,J.M. *et al.* (1968) The characterization of amino acid sequences in proteins by statistical methods. *J. Theor. Biol.*, **21**, 170–201.