research articles Journal of proteome research

Training, Selection, and Robust Calibration of Retention Time Models for Targeted Proteomics

Luminita Moruz,[†] Daniela Tomazela,[‡] and Lukas Käll*[,†,¶]

Center for Biomembrane Research, Department of Biochemistry and Biophysics, and Stockholm Bioinformatics Center, Stockholm University, Stockholm, Sweden, and Department of Genome Sciences, University of Washington, Seattle, Washington 98195

Received May 20, 2010

Accurate predictions of peptide retention times (RT) in liquid chromatography have many applications in mass spectrometry-based proteomics. Most notably such predictions are used to weed out incorrect peptide−spectrum matches, and to design targeted proteomics experiments. In this study, we describe a RT predictor, ELUDE, which can be employed in both applications. ELUDE's predictions are based on 60 features derived from the peptide's amino acid composition and optimally combined using kernel regression. When sufficient data is available, ELUDE derives a retention time index for the condition at hand making it fully portable to new chromatographic conditions. In cases when little training data is available, as often is the case in targeted proteomics experiments, ELUDE selects and calibrates a model from a library of pretrained predictors. Both model selection and calibration are carried out via robust statistical methods and thus ELUDE can handle situations where the calibration data contains erroneous data points. We benchmarked our method against two state-of-the-art predictors and showed that ELUDE outperforms these methods and tracked up to 34% more peptides in a theoretical SRM method creation experiment. ELUDE is freely available under Apache License from http://per-colator.com.

Keywords: retention time prediction • support vector regression • targeted proteomics • peptide identification • bioinformatics

Introduction

Shotgun proteomics is an attractive and accurate technique to analyze the protein content of any biological mixture. For this, the proteins are digested (typically by trypsin) into peptides, and subjected to reversed-phase liquid chromatography (LC) and mass spectrometry (MS). By analyzing the spectra from the MS, we may identify the peptides and infer the proteins in our original sample. An effective way to increase the selectivity of shotgun proteomics is to use retention time (RT) predictors[1−3] to estimate a peptide's chromatographic retention time from its amino acid sequence. One may use such predictions to increase the number of peptide identifications by weeding out matches with a large deviation between observed and predicted retention time.[1,4−7]

For a long time, RT was predicted as a linear combination of the counts of a peptide's amino acids weighted by a "hydrophobicity" scale.[5,8] Krokhin et al. increased performance over such approaches by including features that capture more detailed properties of peptides, such as positional effects, propensity to form helical structures, and the peptides' hydrophobic moments.[1] Petritis et al. designed an even more detailed model using position-specific amino acid information as input to an artificial neural network.[3] Even though this approach reported an impressive prediction accuracy, it required an extensive training set, and hence was hard to retrain for new chromatographic conditions. As the chromatographic conditions affect the observed retention time of peptides, Klammer et al. proposed a scheme where the retention time model was calibrated for each data set at hand.[6] A similar approach was implemented in the OpenMS package using an oligo-kernel support vector regressor.[9]

Traditionally, we use fragmentation spectra from so-called data-dependent acquisition to identify peptides injected in a mass spectrometer.[10] In such assays, the on-board software of the mass spectrometer makes autonomous decisions on which peptides to select for fragmentation. The choice is based on the abundance of the peptide species at each point in time. Data-dependent acquisition has the inadvertent property of rendering better coverage for high-abundance peptides, leading to poor identification of the low-abundance peptides.[11,12] As a consequence, the field of proteomics has taken interest in targeted approaches, where the mass spectrometers are programmed to track individual peptides. If we know from the start when a peptide present in the sample will elute from the chromatographic column, we may specifically instruct the mass spectrometer to fragment the mass to charge ratio corresponding to the peptide of interest. Particularly, a technique known as selected reaction monitoring (SRM) has become of interest.[13,14]

* To whom correspondence should be addressed: E-mail: lukas.kall@cbr.su.se.
[†] Center for Biomembrane Research, Department of Biochemistry and Biophysics, Stockholm University.
[‡] University of Washington.
[¶] Stockholm Bioinformatics Center, Stockholm University.

10.1021/pr1005058

Published on Web 08/24/2010

Table 1. Investigated Data Sets[a]

data set	running time	unique peptides	non-tryptic peptides	in-source fragments	training peptides	test peptides
TFA Tryptic	230 min	4033	34	19	1995	1985
Yeast 20cm	104 min	3332	56	22	1623	1631
Yeast 40cm	108 min	3767	49	19	1850	1849
Yeast 60cm	112 min	3924	46	13	1935	1930
Jup 60min	74 min	2426	59	74	1144	1149
Jup 90min	100 min	2567	29	33	1248	1257
Jup 120min	132 min	2948	47	50	1419	1432
Jup 180min	188 min	3113	49	43	1507	1514
Jup 240min	219 min	3526	62	107	1674	1683
Luna 60min	72 min	2462	46	76	1167	1173
Luna 90min	99 min	2596	44	92	1235	1225
Luna 120min	138 min	2806	40	67	1350	1349
Luna 180min	197 min	3231	75	103	1522	1531
Luna 240min	216 min	3292	78	123	1542	1549

[a] We have evaluated ELUDE on 14 data sets corresponding to different chromatographic conditions. The total running time, the number of unique peptides in each data set, as well as the number of non-tryptic, in-source fragments, and final size of the training and test sets are displayed.

In SRM analysis, we specifically monitor for the mass to charge ratios of certain peptides and one or more of their fragments.

From a bioinformatics point of view, targeted proteomics is challenging as it reverses the relationship between predictions and experiments compared to data-dependent techniques. Instead of predicting peptides from measured fragmentation spectra in a data-dependent acquisition experiment, predicted peptides are validated in target proteomics assays.[15] So far most SRM-based studies have focused on validating peptides that were previously observed by traditional MS/MS, and hence had known retention times and fragmentation spectra.[13,14,16] While this remains the most accurate approach to predict retention time, it limits considerably the potential of SRM experiments as we may only use priorly observed peptide species. To truly increase the coverage of a proteome and to create SRM experiments[17,18] including peptides that were not priorly detected by MS/MS would require better bioinformatics methods to predict proteotypic peptides and their retention times. In this study, we focus on the latter problem: retention time prediction.

An impediment in predicting the retention time of a peptide in the context of SRM analysis is the availability of data to train a RT model. In many cases, we may not be able, as Klammer et al., to employ a set of reliable identified peptides to train a RT model, since the retention time information is required prior to running the experiment. In this study, we propose an alternative workflow, where we require the user to run under representative chromatographic conditions a small control sample of about 50−100 peptides. We use the observed RTs of these peptides to select a model from a library of pretrained predictors and calibrate this model for the chromatographic condition at hand. The selected and calibrated model can be subsequently used to predict retention times for any peptides of interest. This workflow, together with the standard approach of training a new predictor for each condition at hand, was implemented in a software package called ELUDE.

Methods

Sample Preparation and LC−MS/MS Analysis. To generate the MS/MS data, a soluble protein sample from *Caenorhabditis elegans* lysate was reduced, carbamidomethylated, and digested with trypsin in the presence of an acid-labile detergent (RapiGest, Waters Corp., Milford, MA). The resulting peptides were analyzed by μLC−MS/MS on a Thermo LTQ FT Ultra instrument. We used a linear gradient containing a decreasing proportion of buffer A, consisting of 95% water, 5% acetonitrile, and 0.1% formic acid and an increasing proportion of buffer B, consisting of 80% acetonitrile, 20% water, and 0.1% formic acid. The procedure was repeated 10 times exhausting all permutations of a 60, 90, 120, 180, and 240 min gradient time, and two types of columns, a Jupiter Proteo 90 Å and a Luna (both from Phenomenex, Inc., Torrance, CA).

Data Processing. The 10 collected sets of fragmentation spectra were searched using a 10 ppm mass window with SEQUEST[19] against WormDB as well as a decoy database generated by reversing the sequences in the WormDB. To ensure that the retention times of the apexes of the peptide ions were used for each PSM, the retention times were allocated using Bullseye.[20] In addition to these 10 sets, we downloaded four sets of spectra described in Klammer et al.[6] The resulting 14 data sets were postprocessed using Percolator v1.14,[21] using the-unique-peptides switch and all identifications having a posterior error probability[22] higher than 1% were rejected. The remaining peptides were divided in two equal-sized sets, one used to train the predictor and the other to evaluate its performance. ELUDE removed all non-tryptic peptides from both training and test sets. The 14 data sets investigated, including the total running time, number of unique peptides, number of non-tryptic peptides, and final size for the training and test sets, are displayed in Table 1.

Detection of In-Source Fragmentation. Quite frequently, peptides fragment during the electrospray ionization step of the MS/MS experiment, due to so-called in-source fragmentation. These peptide fragments are challenging in the context of retention time prediction as the mass spectrometer only detects a moiety of the peptide that actually has gone through the chromatography column. In this work, we say that a detected peptide, *child*, is an in-source fragment of the peptide *parent* when all of the following statements are true: (i) the sequence of *child* is contained in *parent*; (ii) the difference in observed retention time between *parent* and *child* is less than 5% of the observed retention time of *parent*; (iii) *parent* is tryptic and *child* non-tryptic or the difference in hydrophobicity between *parent* and *child* weighted according to the Kyte and Doolittle index is greater than 5. This heuristic was implemented in our retention time prediction package, ELUDE.

The amount of in-source fragmentation detected in our data is indicated in Table 1. Since in-source fragments have a

Table 2. Features Implemented in ELUDE[a]

number	feature name[b]
1−2	Sum of amino acids' retention coefficients*
3−4	Average of amino acids' retention coefficients*
5−6	N-terminal amino acid's retention coefficient*
7−8	C-terminal amino acid's retention coefficient*
9−10	Sum of squared difference in retention coefficient between neighboring amino acids*
11−12	Sum of retention coefficient of amino acids next to K or R residues*
13−14	Sum of retention coefficient of amino acids next to D or E residues*
15−18	Sum of the retention coefficient of the 2 consecutive amino acids with most and least retention coefficient.*
19−22	Sum of the retention coefficient of the 5 consecutive amino acids with most and least retention coefficient.*
23−30	Highest and lowest values of the "hydrophobic moment" for α helices and β sheets*[c]
31−34	Highest and lowest normalized α helical retention coefficient*
35	Peptide's length
36	Sum of amino acids' bulkiness coefficients[26]
37	Number of occurrences of the polar residues R, K, D, E, N, Q
38	Number of consecutive occurrences of the polar residues R, K, D, E, N, Q
39	Number of occurrences of the hydrophobic residues A, I, L, M, F, W, Y, V, C
40	Number of consecutive occurrences of the hydrophobic residues A, I, L, M, F, W, Y, V, C
41−60	Number of occurrences of each amino acid

[a] Each peptide was described by a set of 60 features, out of which 17 required the use of a retention scale. Since ELUDE implements two such scales, the Kyte and Doolittle index and a SVR-trained retention index, each of these features was described by two values. [b] (*): Features implemented for each of the two retention indices. [c] The hydrophobic moment was computed for windows of 11 amino acids.

negative impact on both training and testing the predictor, we removed all such fragments from both the training and test sets. By default, ELUDE removes all in-source fragmented peptides from the training data, and command line options are available to remove these fragments from the test set or store them for further analysis.

Training Procedure for the Retention Time Model. The training procedure in ELUDE consists of three steps: (1) we derived a set of linear retention time coefficients forming a retention time index, (2) we calculated the features describing each peptide in the training set, and (3) we trained a retention time model using support vector regression. These steps are described below.

Retention Time Index. There are many linear retention time indices described in the literature.[8,23−25] However, as we wanted an index representative for each experimental condition, we derived a new linear retention index for each modeled data set. Hence, before calculating our feature set, we trained a linear support vector regressor (SVR) where we modeled a peptide's RT as a linear combination of the number of each type of amino acids of the peptide. The two parameters ε (tolerance margin) and C (slack penalty) were calibrated via a grid search using 3-fold cross-validation, with $C \in \{2^i, i \in \{-6, -5, ..., 5, 6\}\}$ and $\varepsilon \in \{10^i, i \in \{-3, -2, -1\}\}$. The weights of the trained SVR provided a linear retention coefficient for each amino acid for the conditions at hand.

Feature Calculation. We described each peptide by a set of 60 features outlined in Table 2. The features that required a retention time index are marked with an asterisk (*) in Table 2. These features were calculated with both our derived linear retention time index and the Kyte-Doolittle hydrophobicity scale.[25]

Some of these features require a more detailed description. The "hydrophobic moment" was proposed by Eisenberg et al. as a measure of the amphiphilicity of a protein segment.[27] For a given sequence segment $(a_k, ..., a_j)$, the "hydrophobic moment" for the angle δ was defined as $\mu_{k,j}(\delta) = [(\sum_{i=k}^{j} h(a_i) \cos(i\delta))^2 + (\sum_{i=k}^{j} h(a_i) \sin(i\delta))^2]^{1/2}$, where $h(a_i)$ denoted the retention coefficient of the amino acid a_i and δ represented the angle of rotation per residue (100° for α helices and 180° for β sheets). To determine potential amphipathic regions of a peptide, we slid a window of 11 amino acids along the sequence and calculated the "hydrophobic moment" at each step for the α helices and β sheets. The highest and lowest such values were included as features.

We have also computed a feature indicative of strong and weak "hydrophobic" sides of α helices. Our method was based on a scoring function s, defined for each amino acid $a_k, k > 4$, as $s(k) = h(a_k) + \cos(300°)(h(a_{k+3}) + h(a_{k-3})) + \cos(400°)(h(a_{k-4}) + h(a_{k+4}))$, where $h(a_i)$ denoted the retention coefficient of the amino acid a_i. Large score values indicated a highly "hydrophobic" side of an α helix. We implemented this feature by sliding a window of 9 amino acids along each peptide and computing the s score illustrated above. We included as features the highest and lowest such scores.

While some of these features did not seem to have a major influence on the prediction accuracy for our data sets, we have nevertheless chosen to keep them since they may be of value when training for other types of chromatographic conditions.

Support Vector Regression. The retention time predictor in ELUDE uses the ε support vector regression (ε-SVR)[28] functionality as implemented in the libSVM[29] package. We used a Gaussian kernel and the values of the parameters C (slack penalty), ε (tolerance margin), and σ (the Gaussian radius) were calibrated via 3-fold internal cross-validation within the training set, with $C \in \{2^i | i \in \{-2, -1, ..., 7\}\}$, $\varepsilon \in \{10^i, i \in \{-3, -2, -1\}\}$, $\sigma \in \{2^i | i \in \{-8, -7, ..., 2\}\}$. The values of the features describing the peptides (Table 2) were scaled to values between 0 and 1. For each of the data sets illustrated in Table 1, the training peptides were used to train the ε-SVR, while the test peptides were employed to evaluate its performance.

Robust Selection and Calibration of Models. If the amount of training data is insufficient for training an accurate predictor, ELUDE implements a procedure using the training data to adopt previously trained models instead. The procedure consists of two steps. First, we select a retention time model from a library of previously trained retention models, and second, we calibrate the selected model.

The selection of the most suitable model is carried out as follows: ELUDE predicts retention times for the peptides in the training set using all the available models in the library of previously trained models; in each case, the correlation between predicted and observed retention times is computed, and the predictor yielding the highest correlation is chosen. Since Pearson's correlation coefficient, r, is easily affected by outliers, we employed a more robust selection measure—Spearman's rank correlation coefficient ρ.

As the goal is to predict absolute retention time under the current conditions (and not the retention time of the conditions for which the model was trained), an additional calibration step is required. ELUDE implements a robust regression method,

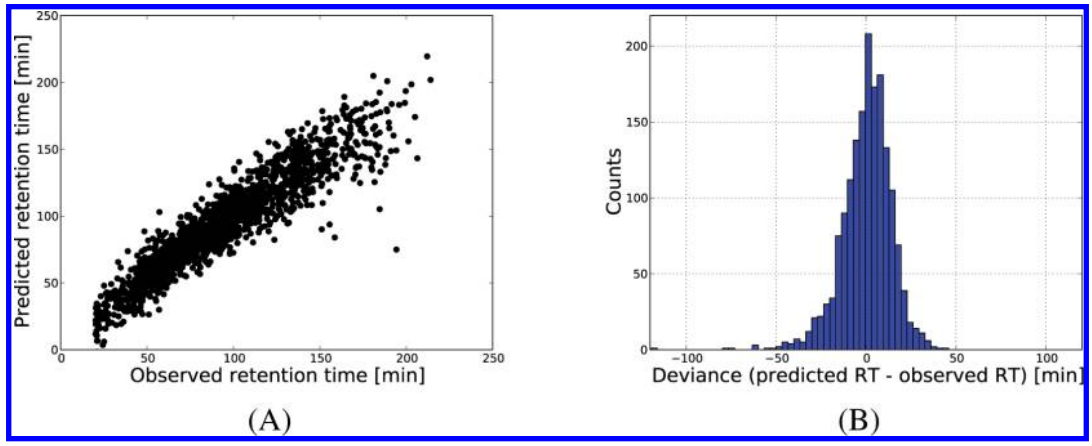

Figure 1. Prediction accuracy of ELUDE. We predicted retention times for a set of 1683 peptides, and plotted (A) their predicted retention times as a function of the observed retention times and (B) a histogram of the deviance between predicted and observed retention times.

the FAST-least trimmed squares (FAST-LTS) regression proposed by Rousseeuw and Van Driessen.[30] The initial version of LTS regression was based on the subset of h data points that yielded the smallest sum of squared residuals.[31] Nevertheless, this approach proved to be infeasible in practice since the computational time increased too fast with the size of the data set. This limitation was circumvented in FAST-LTS by using a heuristic to select a subset of h "good" data points, that is, data points that follow the main trend in the data. We implemented this method in ELUDE and set h to 95% of the available training peptides. This implies that our calibration procedure can accommodate up to 5% outliers in the training data. Depending on the quality of the data, the user can set a different threshold via command line options.

Benchmarked Methods. We benchmarked ELUDE against two commonly used predictors, SSRCalc v3.0[1] and the retention time predictor implemented in The OpenMS Proteomics Pipeline (TOPP).[32,9] All predictors were evaluated on the 14 data sets displayed in Table 1 as detailed below.

SSRCalc. We applied SSRCalc via the Web prediction server available at http://hs2.proteome.ca/SSRCalc/SSRCalc.html. To provide absolute predictions, the tool requires two user calculated regression coefficients a and b. We predicted the uncalibrated retention time of the training peptides using SSRCalc, and employed these predictions together with the observed retention times to derive the coefficients a and b via least-squares regression. We utilized the resulting coefficients to predict retention times for the test peptides. Each data set was run on both conditions provided by SSRCalc, 100 and 300 Å. The 100 Å condition yielded better results for all data sets, and therefore, we only displayed this data.

TOPP. TOPP is a pipeline including several small applications for the analysis of HPLC/MS data. Retention time prediction in TOPP is carried out in two steps. First, a predictor is trained by using an application called RTModel; in a second step, a tool called RTPredict is used to predict retention times by making use of the previously generated predictor. For all the investigated data sets, we employed the training peptides to build a model using RTModel. We calibrated the parameters via grid search in a similar fashion to the one reported by the authors (5-fold cross-validation for C, ν, σ). In a subsequent stage, we predicted retention times for the test peptides using the RTPredict utility.

Results

Performance of ELUDE. We evaluated the performance of our retention time predictor, ELUDE, by applying it to a data set consisting of 3526 unique peptides (Jup 240min). We divided our data into two equal sized bins, and selected one as training set and the other as test set. All non-tryptic and in-source fragmented peptides were removed from both sets, leading to a final training set of 1674 peptides and a final test set of 1683 peptides. We trained and tested ELUDE with the data and plotted the relationship between predicted and observed retention time (Figure 1). The Pearson's correlation between observed and predicted retention time for the 1683 test peptides was $r = 0.94$.

We further assessed the performance of ELUDE by applying it on 13 additional data sets corresponding to different chromatographic conditions (Table 1). Just as before, we divided the data in equally sized training and test sets and removed non-tryptic and in-source fragmented peptides. As a comparison, we predicted retention time for the same data sets using SSRCalc and the RT prediction tools implemented in the TOPP pipeline. For all predictors, the training peptides were used to train/calibrate a model which was then employed to predict retention times for the test peptides. The performance of the predictors was evaluated in terms of Pearson's correlation r between predicted and observed retention times for the test peptides, as well as the minimal time window including the deviations between observed and predicted retention times for 95% of the peptides ($\Delta t_{95\%}$). In addition, we have generated histograms displaying the deviance between predicted and observed retention times for SSRCalc and RTPredict (Supplementary Figure S1), as well as scatter plots illustrating the prediction errors of ELUDE and SSRCalc (Supplementary Figure S2A) and ELUDE and RTPredict (Supplementary Figure S2B).

ELUDE yielded better correlations and time windows for all data sets. The Pearson correlations yielded by ELUDE ranged between 0.92 and 0.97, compared to 0.89−0.94 and 0.89−0.96 by SSRCalc and RTPredict, respectively (Supplementary Table S1). In addition, ELUDE had to cover windows corresponding to 18.9−26.6% of the total running time for the different test sets, compared to 24.0−33.1% of SSRCalc and 21.3−32.3% of RTPredict (Table 3). The average difference between the time windows yielded by ELUDE and SSRCalc was 5.7% of the total

Table 3. Comparison between the Performances of ELUDE, SSRCalc, and RTPredict[a]

	$\Delta t_{95\%}$			ELUDE's increase in capacity	
data set	ELUDE	SSRCalc	RTPredict	SSRCalc	RTPredict
TFA tryptic	43 min (18.9%)	55 min (24.0%)	49 min (21.3%)	27.3%	12.8%
Yeast 20cm	23 min (21.7%)	30 min (28.9%)	27 min (25.5%)	33.2%	17.7%
Yeast 40cm	23 min (21.2%)	30 min (27.6%)	27 min (24.6%)	30.5%	16.6%
Yeast 60cm	22 min (19.4%)	29 min (26.0%)	24 min (21.8%)	33.7%	13.0%
Jup 60min	16 min (21.6%)	21 min (27.7%)	18 min (24.9%)	28.5%	15.1%
Jup 90min	23 min (22.6%)	26 min (26.5%)	25 min (25.4%)	17.2%	11.6%
Jup 120min	28 min (21.4%)	35 min (26.8%)	33 min (25.2%)	25.5%	18.0%
Jup 180min	43 min (23.0%)	54 min (28.7%)	50 min (26.9%)	24.8%	17.0%
Jup 240min	54 min (24.6%)	65 min (29.6%)	67 min (30.7%)	20.2%	24.8%
Luna 60min	16 min (23.0%)	20 min (27.8%)	18 min (24.9%)	20.8%	8.0%
Luna 90min	22 min (22.3%)	27 min (27.4%)	26 min (25.8%)	22.6%	15.1%
Luna 120min	29 min (21.2%)	37 min (26.8%)	34 min (24.6%)	26.8%	15.3%
Luna 180min	42 min (21.4%)	53 min (27.1%)	50 min (25.5%)	26.6%	17.6%
Luna 240min	57 min (26.6%)	71 min (33.1%)	70 min (32.3%)	24.2%	26.1%

[a] The performances of ELUDE, SSRCalc, and RTPredict were evaluated in terms of the time window around the predicted time that would include the observed retention times of the peptide in 95% of the cases ($\Delta t_{95\%}$). In parentheses we display the percentage that $\Delta t_{95\%}$ represents out of the total running time. ELUDE yielded lower time windows for all the investigated data sets, with an average window of 22.06% from the total length of the chromatographic run compared to 27.79% for SSRCalc and 25.68% for RTPredict. The last two columns display the increase in the capacity to track peptides in a theoretical SRM creation experiment when using ELUDE.

running time, while the same difference between ELUDE and RTPredict amounted to 3.6%.

The full impact of these results becomes clear when we express them in terms of peptide tracking capacity in a theoretical SRM method creation experiment. Low $\Delta t_{95\%}$ windows imply that the mass spectrometer has to cover smaller time intervals for each peptide; hence, it can monitor a larger number of peptides. The capacity to track peptides is therefore inversely proportional to the retention time window of the method.[33] When computing the increase in capacity achieved by using ELUDE (Table 3, last two columns), we found that on average ELUDE tracks 25.9% more peptides than SSRCalc and 16.3% more than RTPredict in a theoretical SRM method creation experiment. In addition, training a model in ELUDE was much faster than in RTPredict: for the Jup 240min data set, ELUDE built a model in only 4 min while RTPredict performed the same task in almost 19 h.

Since the quality of the predictions yielded by ELUDE is dependent on the amount of training data available, we investigated the connection between the number of training examples and the $\Delta t_{95\%}$ expressed as percentage of the chromatography running time. Figure 2 displays this relationship for the Jup 240min data set. Following this, we recommend to employ at least 200 unique peptides to train a SVR model.

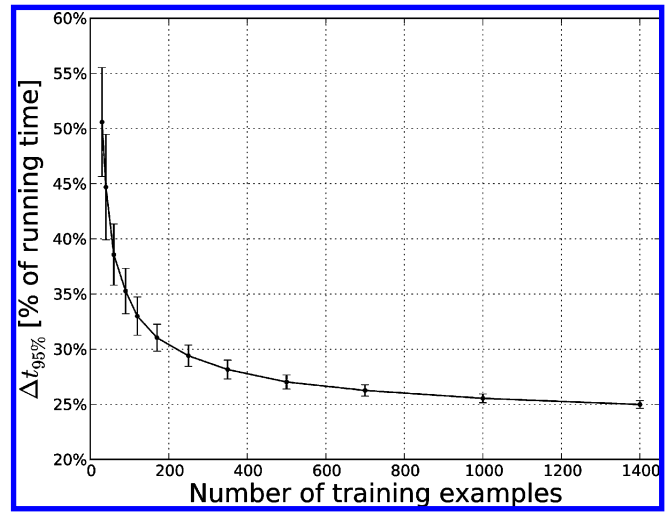

Figure 2. The minimal time window including 95% of the deviations between observed and predicted retention time, $\Delta t_{95\%}$, as a function of the number of training examples. For the Jup 240min data set, we computed the $\Delta t_{95\%}$ value when different amounts of training data were used to build the SVR model. Each point represents the average value of 100 runs, while the bars illustrate the standard deviation.

Selection and Calibration of Retention Models. As previously shown, ELUDE yields the best accuracy when we have training sets consisting of more than 1000 peptides. However, this may often be a too stringent requirement for targeted proteomics experiments where the preferred workflow involves running only a simple mixture of peptides prior to experiments. To overcome this limitation, our predictor includes a library of models corresponding to the experimental conditions of the 14 analyzed data sets. The proposed workflow is to use a small number of peptides to select the model that best fits the data, calibrate this model for the given conditions, and finally apply it to predict the retention times of the peptides of interest. In addition, users can add their own models to the library, such that models corresponding to a large number of chromatographic conditions can be easily developed and shared within the MS community. Two points are essential when implementing this workflow: (1) The method for selecting the most suitable model; (2) How to calibrate the chosen model. The solutions for these problems as implemented in ELUDE are discussed below.

Automatic Model Selection. Given a small set of training data, the goal is to select the model that gives the best predictions for peptides analyzed under the same chromatographic conditions as the training ones. This is accomplished in ELUDE by predicting the retention times of the training peptides using all the available models. The predictor that gives the best correlation between observed and predicted retention times is selected.

We checked whether a model trained on a set of peptides can accurately predict retention times for peptides generated under different chromatographic conditions by inspecting the correlations obtained when retention times of the test peptides were predicted using models built with training peptides from

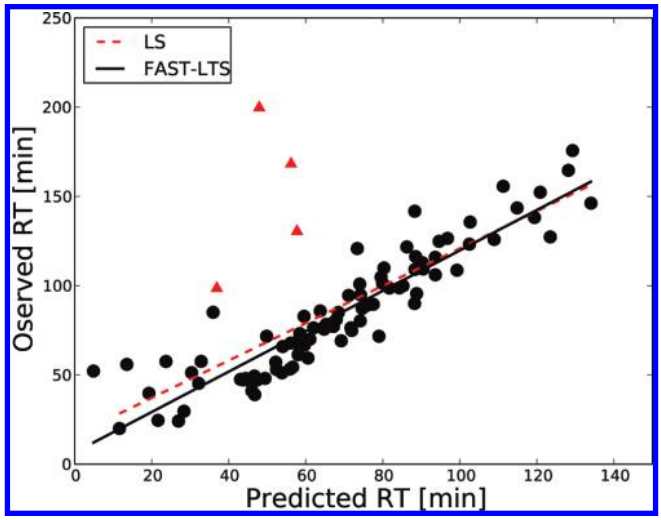

Figure 3. Behavior of FAST-least trimmed squares and least-squares regression on imperfect data. Each point in the figure corresponds to a training peptide. The predicted RT originates from the best matching model as selected by ELUDE. The LTS regression line is displayed with black color, while the LS model is illustrated by a red dashed line. The peptides represented by red triangles are discarded by LTS when building the regression model, while LS uses all training examples. As expected, the LS model is affected by the presence of a few outlier points, while LTS is robust to such data.

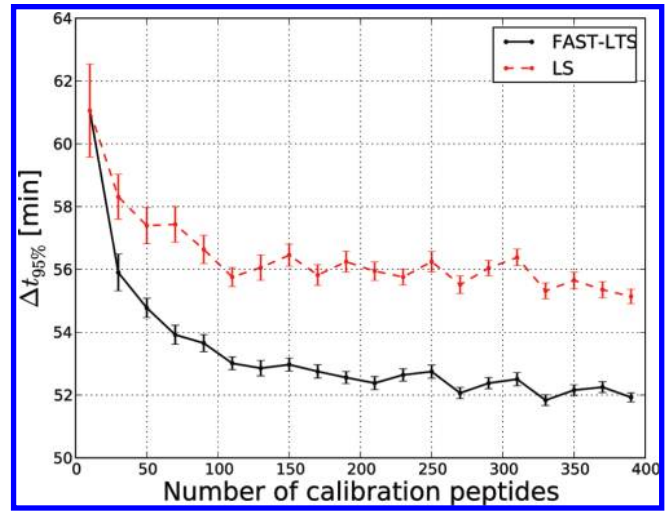

Figure 4. Accuracy of LS and FAST-LTS when calibrating a model trained on data from different chromatographic conditions. For the Jup 240min data set, we investigated the values of the $\Delta t_{95\%}$ window when different amounts of training peptides were employed to calibrate the model selected by ELUDE. We plotted $\Delta t_{95\%}$ as a function of the number of calibration peptides when using both LS (red dashed line) and FAST-LTS (black line) for calibration. Each point in the figure is the average of 100 runs and the bars illustrate the standard error. FAST-LTS yields better results regardless of the size of the calibration set.

different data sets. Given their similarity, only one of the Luna and Jupiter data sets was investigated (Luna 120min and Jup 120min). A high correlation between predicted and observed RT ($\rho > 0.94$) was observed for all the combinations of models and test sets (Supplementary Table S2). This suggests that, following a careful calibration, there is virtually no loss in accuracy when employing a precompiled model.

Calibration of Models. As the classical least-squares regression (LS) method is sensitive to outliers, we chose a robust method for model calibration, the FAST-least trimmed squares (FAST-LTS) regression proposed by Rousseeuw and Van Driessen.[30] Figure 3 depicts the essence of this method: while LS uses all available data to build the regression line, LTS employs only a subset of "good" data points, discarding peptides that are suspected to be outliers.

To assess the potential of this approach, we predicted retention time for the Jup 240min test peptides using a model trained on the training peptides from one of the other 13 data sets. For calibration, we used peptides with a false discovery rate >0.05, as this is a cutoff widely used in mass spectrometry experiments. Subsets ranging between 10 and 300 calibration peptides were employed, and in each run, the model was selected automatically by ELUDE. As a comparison, we performed the same analysis using least-squares regression for data calibration. We evaluated the outcome in terms of $\Delta t_{95\%}$, that is, the size of the time window around the predicted time that includes the true retention time for 95% of the peptides. The results of these analyses (Figure 4) show that employing FAST-LTS to calibrate a retention model leads to considerably better predictions than the standard LS method ($\Delta RT_{95\%}$ with ~3 min lower regardless of the number of calibrating peptides). Furthermore, these data indicate that using a set of about 100 calibration peptides gives predictions comparable to the ones obtained when a full model is trained.

Discussion

We have designed a new package, ELUDE, which accurately predicts peptide retention times in reversed-phase liquid chromatography. The package includes a heuristic to detect in-source fragmentation, and allows the user to both train a predictor for the condition at hand or calibrate an existing model from a library. ELUDE is particularly useful for targeted proteomics assays, where usually only small sets of training peptides are available prior to initializing the experiments. We show that our predictor yields excellent results for such data, outperforming existing predictors both in terms of accuracy and speed. In addition, by virtue of its design, ELUDE is expected to handle data generated under different types of gradient.

Our calibration procedure is based on a robust regression method known as FAST-LTS. With perfect data, the procedure would yield results comparable to standard least-squares regression. However, when dealing with noisy data, as often is the case in practice, robust regression will provide more accurate predictions.

It is frequently claimed that a peptide's retention time is proportional to its hydrophobicity. However, hydrophobicity in itself is an ill-defined property.[34] Furthermore, if we look closer into the thermodynamics of the chromatographic system, we find that the property governing the concentration of organic solvent required to release a peptide from the stationary phase is its ability to lower the free energy between the stationary and mobile phase. This ability depends on a large number of factors such as the three-dimensional structure of the peptide and detailed physicochemical properties. ELUDE implements a number of features that capture these properties. However, to fully model the process would require structural prediction of the peptides under varying conditions of the stationary and mobile phase. While it is currently not feasible

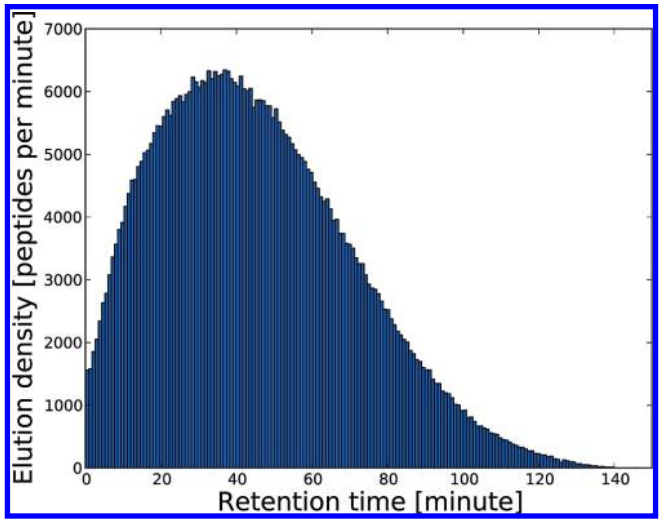

Figure 5. Histogram of the predicted retention times of the peptides in an *in silico* digest of the human proteome. We used the predictor trained on our Yeast 60cm data set ($\Delta t_{95\%} = 22$ min) to predict the retention times of a tryptic *in silico* digest of the human proteome (Human IPI database ver. 3.71). Most peptides are predicted to elute in the first half of the run.

to predict peptide stuctures at a large scale, the shortcut to add more sophisticated features remains the most attractive solution.

The core predictor of ELUDE is based on a radial basis function (RBF) kernel. Although RBF provides good results, it is a standard choice and one could imagine that a more customized kernel could improve the quality of the predictions. One such example is the oligo-border kernel function developed by Pfeifer et al.,[9] which may be combined with our kernel to potentially yield even better predictions. Furthermore, we implement a two-step procedure where we first use linear regression to obtain a retention time index which we subsequently employ in training the final predictor. Another approach would be to use multitask learning procedures[35] to train both predictors at the same time.

In the investigated data sets, we have observed a variable amount of in-source fragmentation. In the yeast data, the in-source fragments represented up to 0.7% of the peptides, whereas in the Jupiter and Luna data sets, these constituted up to 3.7% of the peptides. We do not know whether these variations are related to the instrumentation employed in the experiments or they are rather a consequence of a different tuning of the ion source parameters. Regardless of how fragmentation occurs, their presence negatively impacts both the training and the evaluation of the performance of any retention time predictor. To deal with this problem, we implemented a procedure in ELUDE to detect and remove such fragments from the training and test data.

The availability of an accurate predictor facilitates the study of different characteristics of the chromatographic process. As an example, we have used ELUDE to predict retention times for all of the unique peptides that are created following an *in silico* digestion of the human proteome. The results (Figure 5) show that, if such an experiment would be feasible, most of the peptides would elute in the first part of the chromatographic run. This suggests that the current design of liquid-chromatography may not be optimal for protein identification in highly complex biological mixtures. One way to overcome this problem would be to use a nonlinear gradient of organic solvent when running the chromatography; such an assay will give a more equal spread of the peptides during the chromatography run, and therefore may lead to better peptide identification.

ELUDE can be downloaded under Apache License at http://per-colator.com.

Acknowledgment. This work was supported by grants from the Swedish Research Council and the Carl Trygger Foundation. The authors would like to thank Sander Pronk and Robert Daniels (both at Center for Biomembrane Research, Stockholm University) for helpful comments.

Supporting Information Available: Prediction accuracy of SSRCalc and RTPredict; comparison of prediction errors between ELUDE, SSRCalc, and RTPredict; comparison between the performances of ELUDE, SSRCalc, and RTPredict; ELUDE's sensitivity to differences in chromatographic conditions. This material is available free of charge via the Internet at http://pubs.acs.org.

References

(1) Krokhin, O. V.; Craig, R.; Spicer, V.; Ens, W.; Standing, K. G.; Beavis, R. C.; Wilkins, J. A. An improved model for prediction of retention times of tryptic peptides in ion pair reversed-phase HPLC: its application to protein peptide mapping by off-line HPLC-MALDI MS. *Mol. Cell. Proteomics* **2004**, *3*, 908–919.

(2) Krokhin, O. V. Sequence-specific retention calculator. Algorithm for peptide retention prediction in ion-pair RP-HPLC: application to 300- and 100-A pore size C18 sorbents. *Anal. Chem.* **2006**, *78*, 7785–7795.

(3) Petritis, K.; Kangas, L. J.; Yan, B.; Monroe, M. E.; Strittmatter, E. F.; Qian, W.-J.; Adkins, J. N.; Moore, R. J.; Xu, Y.; Lipton, M. S.; Camp, D. G.; Smith, R. D. Improved peptide elution time prediction for reversed-phase liquid chromatography-MS by incorporating peptide sequence information. *Anal. Chem.* **2006**, *78*, 5026–5039.

(4) Strittmatter, E. F.; Kangas, L. J.; Petritis, K.; Mottaz, H. M.; Anderson, G. A.; Shen, Y.; Jacobs, J. M.; Camp, D. G.; Smith, R. D. Application of peptide LC retention time information in a discriminant function for peptide identification by tandem mass spectrometry. *J. Proteome Res.* **2004**, *3*, 760–769.

(5) Palmblad, M.; Ramstrom, M.; Markides, K.; Hakansson, P.; Bergquist, J. Prediction of chromatographic retention and protein identification in liquid chromatography/mass spectrometry. *Anal. Chem.* **2002**, *74*, 5826–5830.

(6) Klammer, A. A.; Yi, X.; MacCoss, M. J.; Noble, W. S. Improving tandem mass spectrum identification using peptide retention time prediction across diverse chromatography conditions. *Anal. Chem.* **2007**, *79*, 6111–6118.

(7) Pfeifer, N.; Leinenbach, A.; Huber, C.; Kohlbacher, O. Improving Peptide Identification in Proteome Analysis by a Two-Dimensional Retention Time Filtering Approach. *J. Proteome Res.* **2009**, *8*, 4109–4115.

(8) Meek, J. L. Prediction of peptide retention times in high-pressure liquid chromatography on the basis of amino acid composition. *Proc. Natl. Acad. Sci. U.S.A.* **1980**, *77*, 1632–1636.

(9) Pfeifer, N.; Leinenbach, A.; Huber, C.; Kohlbacher, O. Statistical learning of peptide retention behavior in chromatographic separations: a new kernel-based approach for computational proteomics. *BMC Bioinf.* **2007**, *8*, 468.

(10) Ducret, A.; Van Oostveen, I.; Eng, J. K.; Yates, J. R., III; Aebersold, R. A. High throughput protein characterization by automated reverse-phase chromatography/electrospray tandem mass spectrometry. *Protein Sci.* **1998**, *7*, 706–719.

(11) Domon, B.; Broder, S. Implications of new proteomics strategies for biology and medicine. *J. Proteome Res.* **2004**, *3*, 253–260.

(12) Gstaiger, M.; Aebersold, R. Applying mass spectrometry-based proteomics to genetics, genomics and network biology. *Nat. Rev. Genet.* **2009**, *10*, 617–627.

(13) Addona, T. A.; et al. Multi-site assessment of the precision and reproducibility of multiple reaction monitoring-based measurements of proteins in plasma. *Nat. Biotechnol.* **2009**, *27*, 633–641.

(14) Picotti, P.; Bodenmiller, B.; Mueller, L.; Domon, B.; Aebersold, R. Full dynamic range proteome analysis of S. cerevisiae by targeted proteomics. *Cell* **2009**, *138*, 795–806.

(15) Domon, B.; Aebersold, R. Mass spectrometry and protein analysis. *Science* **2006**, *312*, 212.
(16) Yocum, A. K.; Chinnaiyan, A. M. Current affairs in quantitative targeted proteomics: multiple reaction monitoring-mass spectrometry. *Briefings Funct. Genomics Proteomics* **2009**, *8*, 145–157.
(17) Prakash, A.; Tomazela, D.; Frewen, B.; MacLean, B.; Merrihew, G.; Peterman, S.; MacCoss, M. Expediting the development of targeted SRM assays: using data from shotgun proteomics to automate method development. *J. Proteome Res* **2009**, *8*, 2733–2739.
(18) MacLean, B.; Tomazela, D.; Shulman, N.; Chambers, M.; Finney, G.; Frewen, B.; Kern, R.; Tabb, D.; Liebler, D.; MacCoss, M. Skyline: An open source document editor for creating and analyzing targeted proteomics experiments. *Bioinformatics* **2010**, *26*, 966–968.
(19) Eng, J.; McCormack, A.; Yates, J.; Eng, J. K.; McCormack, A. L.; Yates, J. R. and others An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *J. Am. Soc. Mass Spectrom.* **1994**, *5*, 976–989.
(20) Hsieh, E.; Hoopmann, M.; MacLean, B.; MacCoss, M. Comparison of database search strategies for high precursor mass accuracy MS/MS data. *J. Proteome Res* **2010**, *9*, 1138–1143.
(21) Käll, L.; Canterbury, J. D.; Weston, J.; Noble, W. S.; MacCoss, M. J. Semi-supervised learning for peptide identification from shotgun proteomics datasets. *Nat. Methods* **2007**, *4*, 923–925.
(22) Käll, L.; Storey, J.; Noble, W. Non-parametric estimation of posterior error probabilities associated with peptides identified by tandem mass spectrometry. *Bioinformatics* **2008**, *24*, i42.
(23) Browne, C. A.; Bennett, H. P. J.; Solomon, S. The isolation of peptides by high-performance liquid chromatography using predicted elution positions. *Anal. Biochem.* **1982**, *124*, 201–208.
(24) Yoshida, T. Calculation of peptide retention coefficients in normal-phase liquid chromatography. *J. Chromatogr., A* **1998**, *808*, 105–112.
(25) Kyte, J.; Doolittle, R. F. A simple method for displaying the hydropathic character of a protein. *J. Mol. Biol.* **1982**, *157*, 105–132.
(26) Zimmerman, J. M.; Eliezer, N.; Simha, R. The characterization of amino acid sequences in proteins by statistical methods. *J. Theor. Biol.* **1968**, *21*, 170–201.
(27) Eisenberg, D.; Weiss, R. M.; Terwilliger, T. C. The hydrophobic moment detects periodicity in protein hydrophobicity. *Proc. Natl. Acad. Sci. USA* **1984**, *81*, 140–144.
(28) Vapnik, V. N. *The Nature of Statistical Learning Theory*; Springer-Verlag New York, Inc.: New York, NY, 1995.
(29) Chang, C.-C.; Lin, C.-J. *LIBSVM: a library for support vector machines*; 2001; software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.
(30) Rousseeuw, P. J.; Driessen, K. Computing LTS regression for large data sets. *Data Min. Knowledge Discovery* **2006**, *12*, 29–45.
(31) Rousseeuw, P. J. Least median of squares regression. *J. Am. Stat. Assoc.* **1984**, *79*, 871–880.
(32) Kohlbacher, O.; Reinert, K.; Gropl, C.; Lange, E.; Pfeifer, N.; Schulz-Trieglaff, O.; Sturm, M. TOPP—the OpenMS proteomics pipeline. *Bioinformatics* **2007**, *23*, e191−197.
(33) Bertsch, A.; Jung, S.; Zerck, A.; Pfeifer, N.; Nahnsen, S.; Henneges, C.; Nordheim, A.; Kohlbacher, O. Optimal de novo design of MRM experiments for rapid assay development in targeted proteomics. *J. Proteome Res.* **2010**, *9*, 2696–2704.
(34) Chandler, D. Interfaces and the driving force of hydrophobic assembly. *Nature* **2005**, *437*, 640–647.
(35) Caruana, R. Multitask learning. *Mach. Learn.* **1997**, *28*, 41–75.

PR1005058