Anal. Chem. **2007,** *79,* 6111−6118

Improving Tandem Mass Spectrum Identification Using Peptide Retention Time Prediction across Diverse Chromatography Conditions

Aaron A. Klammer,† Xianhua Yi,† Michael J. MacCoss,† and William Stafford Noble*,†,‡

Department of Genome Sciences, and Department of Computer Science and Engineering, University of Washington, Seattle, Washington 98195

Most algorithms for identifying peptides from tandem mass spectra use information only from the final spectrum, ignoring non-mass-based information acquired routinely in liquid chromatography tandem mass spectrometry analyses. One physiochemical property that is always obtained but rarely exploited is peptide chromatographic retention time. Efforts to use chromatographic retention time to improve peptide identification are complicated because of the variability of retention time in different experimental conditions—making retention time calculations nongeneralizable. We show that peptide retention time can be reliably predicted by training and testing a support vector regressor on a small collection of data from a *single* liquid chromatography run. This model can be used to filter peptide identifications with observed retention time that deviates from predicted retention time. After filtering, positive peptide identifications increase by as much as 50% at a false discovery rate of 3%. We demonstrate that our dynamically trained model generalizes well across diverse chromatography conditions and methods for generating peptides, in particular improving peptide identification using nonspecific proteases.

Understanding the major functions of the cell requires accurate measurement and characterization of its main biochemical actors, proteins. While much can be learned from the study of individual proteins, in vivo a protein invariably acts in concert with other biomolecules. These interactions differ according to cell type, the state of the cell, and its response to external stimuli. Several technologies have the potential to provide a comprehensive view of many of the cell's proteins in a single experiment. One widely used technology is shotgun proteomics using liquid chromatography (LC)-tandem mass spectrometry (MS/MS).[1,2]

In a typical LC−MS/MS experiment, proteins are enzymatically digested to peptides, which are then separated by microcapillary reversed-phase chromatography. The eluting peptides are emitted into a mass spectrometer using electrospray ionization. The mass spectrometer automatically measures the mass-to-charge (m/z) ratio of the intact and fragmented peptides, yielding a tandem mass spectrum. One LC−MS/MS experiment typically yields tens of thousands of MS/MS spectra. The identity of the peptides that produced the spectra, and thus the identity of the original proteins, can be obtained by database search algorithms such as SEQUEST.[3]

As with any high-throughput technology, shotgun proteomics experiments must manage the tradeoff between maximizing true positive identifications and minimizing false positive identifications.[4,5] The need to reduce false positives has spurred the development of methods for increasing the sensitivity and specificity of peptide identification algorithms. However, most of these methods use information exclusively from the MS and MS/MS stages of analysis, ignoring information from the chromatographic separation, such as retention time. The chromatographic retention time is the amount of time that a peptide is retained on the column and is closely related to the peptide's molecular structure, polarity, and hydrophobicity.[6] It has the advantage of being independent of the information contained in the MS/MS spectrum and can therefore be used in conjunction with the information in the MS/MS spectrum to increase peptide identification confidence.

Understanding and predicting peptide retention time has a long history. For reversed-phase chromatography, peptide retention time increases with increasing peptide hydrophobicity.[6] Many models assume that peptide retention time is a function of peptide amino acid composition.[7−11] However, it is clear from experimental

* To whom correspondence should be addressed. E-mail: noble@gs.washington.edu.

† Department of Genome Sciences.

‡ Department of Computer Science and Engineering.

(1) McCormack, A. L.; Schieltz, D. M.; Goode, B.; Yang, S.; Barnes, G.; Drubin, D.; Yates, J. R., III. *Anal. Chem.* **1997**, *69* (4), 767−776.

(2) Yates, J. R., III. *Anal. Chem.* **1998**, *33*, 1−19.

(3) Eng, J. K.; McCormack, A. L.; Yates, J. R., III. *J. Am. Soc. Mass Spectrom.* **1994**, *5*, 976−989.

(4) Cargile, B. J.; Bundy, J. L.; Freeman, T. W.; Stephenson, J. L., Jr. *J. Proteome Res.* **2004**, *3*, 1082−1085.

(5) Qian, W. J.; Liu, T.; Monroe, M. E.; Strittmatter, E. F.; Jacobs, J. M.; Kangas, L. J.; Petritis, K.; Camp, D. G., II; Smith, R. D. *J. Proteome Res.* **2005**, *4* (1), 53−62.

(6) Frenz, J.; Hancock, W. S.; Henzel, W. J.; Horva'th, C. In *HPLC of Biological Macromolecules: Methods and Applications*; Gooding, K. M., Regnier, F. E., Eds.; Marcel Dekker: New York, 1990.

(7) Meek, J. L. *Proc. Natl. Acad. Sci. U.S.A.* **1980**, *77*, 1632−1636.

(8) Browne, C. A.; Bennett, H. P. J.; Solomon, S. *Anal. Biochem.* **1982**, *124*, 201−208.

(9) Guo, D.; Mant, C. T.; Taneja, A. K.; Parker, J. M.; Hodges, R. S. J. *J. Chromatogr.* **1987**, *386*, 205−222.

(10) Hearn, M. T.; Aguilar, M. I.; Mant, C. T.; Hodges, R. S. *J. Chromatogr.* **1988**, *438*, 197−210.

(11) Le Bihan, T.; Robinson, M. D.; Stewart, I. I.; Figeys, D. J. *J. Proteome Res.* **2004**, *3*, 1138−1148.

10.1021/ac070262k CCC: $37.00
Published on Web 07/11/2007

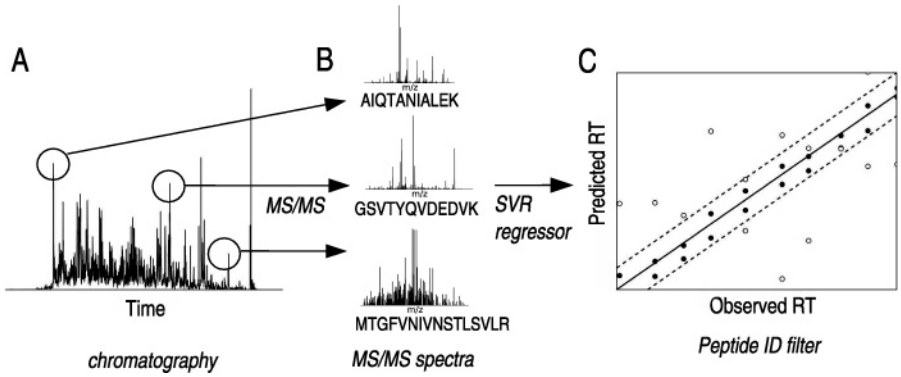

Figure 1. Overview of chromatography analysis. (A) Chromatography separates peptides over time. (B) Spectra acquired throughout the chromatographic separation are identified as peptides. (C) An SVR learns to predict retention time (solid black line) from high-confidence peptide identifications (black circles) from a *single* chromatography run. False positive peptide identifications (white circles) can be eliminated if their observed retention time differs greatly from their predicted retention time (white circles outside dashed lines).

data that two peptides with identical amino acid compositions can be chromatographically separated from one another.[12] More recent models augment the compositional approach with parameters for peptide length or mass,[13] diresidue or triresidue composition,[14] or positional effects such as the identity of the N-term residue.[15] Still more sophisticated models include parameters for secondary structure or peptide hydrophobic moment.[16,21]

The most accurate and sophisticated peptide retention time predictor is that of Petritis et al.[12] first presented in simpler form 3 years earlier,[17] which uses an artifcial neural network (ANN) to predict a normalized form of retention time. The large amount of data required to train the ANN (for Petritis et al.,[12] 345 000 nonredundant peptides) makes retraining for new chromatography conditions impractical. A more recent, but less complicated, ANN has since been published,[18] but does not outperform the method presented in Petritis et al. In addition, Petritis et al. also demonstrated that spiked peptides can be used to align data from different chromatograms. However, this technique requires an additional experimental step and does not account for changes in relative peptide elution order.

Recently, a handful of retention time predictors have been used to enhance the confidence of peptide identifications. Palmblad et al.[19] predicted retention time for each peptide using least-squares regression to determine amino acid weights and then used a χ^2 test to rank candidate peptides based on deviation from expected mass and predicted retention time. As the authors acknowledged, their retention time prediction is poor compared to prior work and improvement in protein identification is modest. Kawakami et al.[20] used the sum of residue retention coefficients to predict retention time for peptides and phosphopeptides, but they made no clear distinction between training and test data. When predictions are made on data not used in training, the correlation between predicted and observed retention time deteriorates, a sign of overfitting.

The goal of this paper is to use chromatographic retention time to increase the confidence of peptide identifications by tandem mass spectrometry. Previous efforts to increase identification confidence using retention time have been limited to conditions (e.g., column, mobile phase, gradient) identical to those used to train the retention time predictor[21] or require additional experimental steps.[12] Most such methods train a single retention time predictor using a limited subset of highly reproducible chromatography conditions[15] or perform a normalization that attempts to eliminate variability.[12,22] In practice, however, researchers use a large number of diverse chromatographic conditions, making a static retention time predictor not particularly useful. In this work, we demonstrate the application of a dynamically trained support vector regressor (SVR) to predict retention time for peptides in a given LC−MS/MS analysis (Figure 1), using only data generated during the current run. We use features similar to Krokhin et al.,[15] and see our work as extending that approach to allow application across diverse data sets and conditions. Our model may not outperform other models trained and tested under highly similar conditions, as in Krokhin et al.[15] or Petritis et al.[12] However, our approach is portable to new chromatography conditions or sample preparation protocols, adapting to differences in column length, protease, ion-pairing agent, and MudPIT salt step for each individual LC−MS/MS analysis. Furthermore, by eliminating peptide identifications with an observed retention time that deviates greatly from their predicted retention time, our method increases the number of true positive peptide identifications over a range of false discovery rates. The results presented here have implications for traditional shotgun proteomics research using trypsin in addition to some less frequently used enzymes.

(12) 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., 2nd; Smith, R. D. *Anal. Chem.* **2006**, *78* (14), 5026−5039.
(13) Mant, C. T.; Zhou, N. E.; Hodges, R. S. *J. Chromatogr., A* **1989**, *476*, 363−375.
(14) Mant, C. T.; Hodges, R. S. *J. Chromatogr., A* **2006**, *1125*, 211−219.
(15) Krokhin, O. V.; Craig, R.; Spicer, V.; Ens, W.; Standing, K. G.; Beavis, R. C.; Wilkins, J. A. *Mol. Cell. Proteomics* **2004**, *3*, 908−919.
(16) Bączek, T.; Wiczling, P.; Marszałł M.; Heyden, Y. V.; Kaliszan, R. *J. Proteome Res.* **2005**, *4*, 555−563.
(17) Petritis, K.; Kangas, L. J.; Ferguson, P. L.; Anderson, G. A.; Pasa-Tolic, L.; Lipton, M. S.; Auberry, K. J.; Strittmatter, E. F.; Shen, Y.; Zhao, R.; Smith, R. D. *Anal. Chem.* **2003**, *75* (5), 1039−1048.
(18) Shinoda, K.; Sugimoto, M.; Yachie, N.; Sugiyama, N.; Masuda, T.; Robert, M.; Soga, T.; Tomita, M. *J. Proteome Res.* **2006**, *5*, 3312−3317.
(19) Palmblad, M.; Ramstrom, M.; Markides, K. E.; Hakansson, P.; Bergquist, J. *Anal. Chem.* **2002**, *74*, 5826−5830.
(20) Kawakami, T.; Tateishi, K.; Yamano, Y.; Ishikawa, T.; Kuroki, K.; Nishimura, T. *Proteomics* **2005**, *5*, 856−64.
(21) Palmblad, M.; Ramstrom, M.; Bailey, G. B.; McCutchen-Maloney, S. L.; Bergquist, J.; Zeller, L. C. *J. Chromatogr., B* **2004**, *803*, 131−135.
(22) Strittmatter, E. F.; Kangas, L. J.; Petritis, K.; Mottaz, H. M.; Anderson, G. A.; Shen, Y.; Jacobs, J. M.; Camp, D. G., 2nd; Smith, R. D. *J. Proteome Res.* **2004**, *3* (4), 760−769.

Table 1. Twelve Data Sets Used To Train and Test the Support Vector Regressor

data set	total[a]	confident[b]	train[c]	test[d]
20 cm	6929	2073	1554	519
40 cm	7220	2409	1806	603
60 cm	7459	2774	2080	694
TFA	11977	3179	2384	795
chymotrypsin	2191	200[e]	150	50
elastase	4377	200[e]	150	50
MudPIT-1	11339	863	647	216
MudPIT-2	9550	1553	1164	389
MudPIT-3	8488	1320	990	330
MudPIT-4	7871	1096	822	274
MudPIT-5	7373	888	666	222
MudPIT-6	6439	757	567	190

[a] Number of +2 spectra associated with unique peptides that satisfy that data set's enzyme specificity requirement. [b] Number of high-confidence spectra selected at a 10% FDR. [c] Number of spectra in regressor training subset. [d] Number of spectra in regressor testing subsets. [e] These data sets had less than 200 examples at 10% FDR; as a result, the top 200 PSMs were used, regardless of FDR.

METHODS

Data Sets. We analyze 12 separate chromatography data sets (Table 1), chosen to represent a diverse set of chromatography and sample preparation conditions. All data sets are from a complex mixture of peptides from the digestion of a largely unfractionated lysate of the yeast *Saccharomyces cerevisiae*. The first three data sets are reversed-phase analyses of a tryptic digest of the soluble yeast proteome, each with a different length column of 20, 40, and 60 cm. A fourth identically prepared yeast sample is analyzed with the ion-pairing agent, trifluoroacetic acid (TFA). Two additional data sets are from the same yeast sample digested with the enzymes chymotrypsin or elastase. The final six data sets are taken from the six steps of a 12-h multidimensional separation using strong cation-exchange and reversed-phase LC (i.e., multidimensional protein identification technology, or MudPIT[23]) for the analysis of a tryptic digest of the soluble yeast proteome. Summary statisics for the data sets, and for the training and testing data sets extracted from them, are shown in Table 1. Details of the exact methods for producing the data sets follow.

Sample Preparation. A complex yeast lysate was prepared by growing strain S288c in 500 mL of yeast extract/peptone/dextrose medium, harvested at OD 1.2, and lysed in a BeadBeater (BioSpec Products, Inc. Bartelsville, OK) in 50 mM ammonium bicarbonate at pH 7.8. Unbroken cells and cell debris were removed by centrifugation at 5000*g* for 10 min. A 45-μL aliquot of the supernatant was mixed with 5 μL of 1% PPS (Protein Discovery, Knoxville, TN), heated at 90 °C for 2 min, and then treated serially with dithiothreitol and iodoacetic acid for 30 min each as described previously.[24] The reduced and alkylated protein mixture was digested to peptides with the addition of either trypsin, chymotrypsin, or elastase at a 1:50 enzyme/substrate ratio. The mixture was incubated at 37 °C for 4 h and quenched by acidification with HCl. The digest was centrifuged at 14 000 rpm at 4 °C in a microcentrifuge, and the supernatant stored at −80 °C until analyzed by mass spectrometry.

LC−MS/MS. The samples were analyzed by data-dependent tandem mass spectrometry using the following LC−MS/MS analysis. A 75-μm-i.d. fused-silica capillary (Polymicro Tech, Phoenix, AZ) was pulled to a tip using a CO_2 laser puller (Sutter Instruments) and slurry packed with 4-μm, 90-Å-pore size Jupiter Proteo reversed-phase material (Phenomenex, Ventura, CA) using a pressure bomb. In all data sets, the column was 40 cm long, except for the two data sets in which we varied the column length, to 20 and 60 cm. The column was placed inline with an Agilent 1100 Binary HPLC and autosampler (Palo Alto, CA). The flow was split precolumn to create a flow rate of ~500 nL/min through the column, as described previously.[24]

As peptides eluted from the microcapillary columns, they were emitted into an LTQ mass spectrometer (ThermoFisher Scientific, San Jose, CA) with the application of a 2.4-kV spray voltage applied distal to the solvent split. A cycle of one full-scan mass spectrum (400−1400 m/z) followed by five data-dependent MS/MS spectra at a 35% normalized collision energy was repeated continuously throughout each analysis. Application of mass spectrometer scan functions and HPLC solvent gradients were controlled by the Xcalibur data system (ThermoFisher Scientific). Inorganic buffer was 95% water/5% acetonitrile/0.1% formic acid (buffer A), and organic buffer was 5% water/95% acetonitrile/0.1% formic acid (buffer B), except in the TFA data set. In the TFA data set, 0.01% TFA was used in place of 0.1% formic acid.

The 20-, 40-, and 60-cm data sets used a 2-h gradient, consisting of 16 min of 5% B, followed by an 84-min gradient to 32% B, followed by a 2-min pulse of 80% B, followed by an 18-min equilibration with 2% B. The TFA, chymotrypsin, and elastase data sets used a 4-h gradient, consisting of 27-min 5% B, a 193-min gradient from 0% B to 32% B, a 5-min step to 80% B, and a final 15-min equilibration at 5% B.

MudPIT LC/LC−MS/MS. The trypsin digest of yeast proteins was also analyzed using a six-step MudPIT analysis. The MudPIT analysis was performed in a fashion similar to that described previously.[25] Briefly, a triphasic column was constructed by packing (tip first) a 100-μm-i.d. capillary pulled to a tip with 7 cm of reversed-phase material (Jupiter, Proteo Phenomonex), 3.5 cm of strong cation-exchange material (Whatman, Partisphere SCX 5 μm), and an addition 4 cm of reversed-phase material. The protein digest was pressure-loaded directly onto the rear end of a triphasic chromatography column. Once loaded with the protein digest, the column was placed inline with an Agilent 1100 Binary HPLC and analyzed using a six-step multidimensional separation as described previously.[25] The salt step elutions were provided by injecting 50 μL of ammonium acetate buffer at concentrations of 0, 100, 200, 500, and 800 mM and 5 M, using an autosampler inline between the HPLC and the column.

Training and Testing Set Extraction. A high-confidence set of training and testing data was extracted from each of the 12 data sets. First, the spectra were searched against both target and decoy versions of a fasta file containing the translated predicted yeast open-reading frames (from Apr-02, 2004) using SEQUEST with no enzyme specificity.[3] The decoy database was produced by randomly shuffling the sequences in the target yeast protein sequence database. Identifications were filtered using the following

(23) Washburn, M. P.; Wolters, D.; Yates, J. R., III. *Nat. Biotechnol.* **2001**, *19*, 242−247.
(24) Klammer, A. A.; MacCoss, M. J. *J. Proteome Res.* **2006**, *5* (3), 695−700.
(25) McDonald, W. H.; Ohi, R.; Miyamoto, D. T.; Mitchison, T. J.; Yates, J. R., III. *Int. J. Mass Spectrom.* **2002**, *219*, 245−251.

criteria: charge state of +2, peptide sequence ending in K or R and preceded by K or R (except for peptides at the beginning or end of a protein, or for the chymotrypsin and elastase data sets), and allowing any number of missed cleavages sites.

We used the number of matches in a search against the decoy database as a proxy for false positive matches in the search against the target database. High-confidence spectrum identifications were selected by setting an Xcorr threshold so that the number of matches to the decoy database above this threshold is 10% of the number of matches to the target database; this yields ~10% false discovery rate (FDR) in the matches to the target database. If the number of target matches at a 10% FDR was less than 200, then the top 200 scoring spectra were used. We required a minimum of 200 identifications because regression performance deteriorated with less than 200 identifications (see Supporting Information). Regardless of the number of spectra, when multiple spectra matched a single peptide sequence, the spectrum with the highest Xcorr was selected to avoid bias in the regression toward common peptides. The resulting set of peptides and retention times was split to form a 3:1 ratio between the training and testing data sets (Table 1) for each chromatography run, which were then used to train and test the SVR. No peptides were allowed to occur in both the training and testing data sets.

Support Vector Regression. As with other forms of regression, a support vector regressor learns a function that relates a dependent variable (in this case, retention time) to a set of independent variables. An SVR builds a regressor out of a subset of the training examples, known as support vectors.[26] Training examples that are within a tolerance value ϵ of the model prediction are ignored. To generate the independent variables, each peptide from the training and test sets is represented as a 63-element vector composed of the following: 20 elements represent the total number of each amino acid residue in the peptide; 40 binary elements represent the identity of the extreme N-terminal (N-term) and penultimate C-terminal (C-term) residues, respectively; and 3 additional elements represent the identity of the last C-term residue (either K or R) and the peptide length and mass. For the data sets generated with the nontryptic enzyme elastase and chymotrypsin, the ultimate C-terminal residue is used instead of the penultimate, and the K or R term is set to zero.

A separate SVR is trained for each data set in Table 1. The SVR is tested by measuring the R value between predicted and observed retention time on a subset test set not used in training. R value is a statistical measure of the correlation between two data sets. The R value for two data sets x and y of length n is given by $r = Cov(x,y)/\sigma_x\sigma_y$, where $Cov(x,y) = n\sum xy - \sum x\sum y$, the covariance of data sets x and y, and $\sigma_x = \sqrt{n\sum x^2 - (\sum x)^2}$, the standard deviation of data set x. An SVR is trained and tested twice using two kinds of kernels: a linear kernel, because it allows ready interpretion of the weight it assigns to each feature (see Results); and a Gaussian kernel (also known as a radial-basis function kernel), because it allows maximum flexibility in the functions that it can successfully regress.

Hyperparameters for each kernel are selected using 3-fold cross-validation on the training set. For both kernels, the SVR is trained with an ϵ insensitive-loss hyperparameter of 0.1. Other values of ϵ did not yield significantly different results. Another hyperparameter used in the regression is the soft-margin penalty C, which can be thought of as a bound on the weight that can be given to each training example. The value of C is initially allowed to range over 10 orders of magnitude from 10^{-3} to 10^{7}. For the final cross-validation, to decrease processing time, C is constrained to be 10^{-1}, 10^{0}, or 10^{1} for the linear kernel, and 10^{5}, 10^{6}, or 10^{7} for the Gaussian kernel. The Gaussian kernel has an additional hyperparameter σ, which corresponds to the width of the Gaussians used; it is set to 10^{-6}, 10^{-7}, or 10^{-8}. R values are reported on a held-out test set.

The SVR is implemented using the publicly available software package PyML (pyml.sourceforge.net). Source code for producing the results presented here can be found at http://noble.gs.washington.edu/proj/rt. Model construction and application took on average ~20 min.

(26) Vapnik, V. *The Nature of Statistical Learning Theory*; Springer: New York, 1995.

Table 2. *R* Values for a Fixed Regression Compared to a Learned Regression Using the Gaussian or Linear Kernels[a]

data set	fixed	Gaussian[b]	fixed[c]	linear[d]
20 cm	0.881	0.908	0.877	0.892
40 cm	0.892	0.897	0.894	0.891
60 cm	0.914	0.926	0.889	0.892
chymotrypsin	0.871	0.865	0.787	0.790
elastase	0.823	0.850	0.822	0.851
MudPIT-1	0.697	0.766	0.764	0.684
MudPIT-2	0.851	0.898	0.871	0.894
MudPIT-3	0.856	0.878	0.876	0.891
MudPIT-4	0.853	0.894	0.853	0.859
MudPIT-5	0.867	0.911	0.813	0.812
MudPIT-6	0.839	0.918	0.836	0.894
TFA	0.818	0.842	0.863	0.886

[a] Correlation for 12 data sets for fixed parameters described in Krokhin et al.[15] (fixed) and parameters learned for each data set with a Gaussian or linear kernel. [b] Significant with p-value of ≤ 0.01. [c] There are two fixed regressions because the Gaussian and linear kernels were evaluated on slightly different testing subsets. [d] Not significant with p-value of ≤ 0.01.

Table 3. Analysis of Six-Step MudPIT

data set	unfiltered[a]	filtered[b]	unfiltered (cumulative)[c]	filtered (cumulative)[d]
MudPIT-1	723	723	723	723
MudPIT-2	1331	1418	1940	2029
MudPIT-3	1170	1185	2695	2802
MudPIT-4	918	1047	3165	3377
MudPIT-5	763	823	3531	3786
MudPIT-6	653	653	3956	4214

[a] Number of unique peptides at 5% FDR before filtering spectra with observed retention time that deviates significantly from predicted retention time. [b] Number of unique peptides at 5% FDR after filtering spectra with observed retention time that deviates significantly from predicted retention time. [c] Cumulative number of unique peptides for the entire MudPIT up to the given step before filtering. [d] Cumulative number of unique peptides for the entire MudPIT up to the given step after filtering.

RESULTS

Support Vector Regression. We first evaluate our dynamically trained regressor by comparing it to the published, fixed-parameter regressor from Krokhin et al.[15] We measure performance by comparing correlation (measured by R value) between

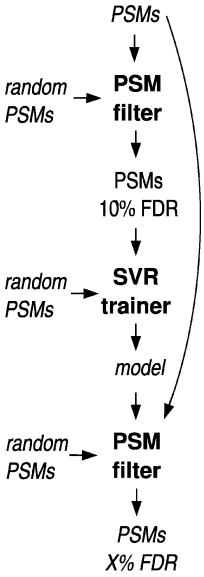

Figure 2. Overview of data flow. For each LC−MS/MS experiment, we start with a collection of peptide-spectrum matches (PSMs). We use PSMs to a decoy sequence database to filter these PSMs to produce high-confidence PSMs at a 10% FDR. These PSMs, along with a second set of decoy PSMs, are used to train a support vector regressor and to select a threshold for filtering PSMs based on their retention time, yielding a trained model. The model and a third set of decoy PSMs are used to select the final set of target PSMs at a desired FDR. Not shown is the 5-fold cross-validation used to validate this method.

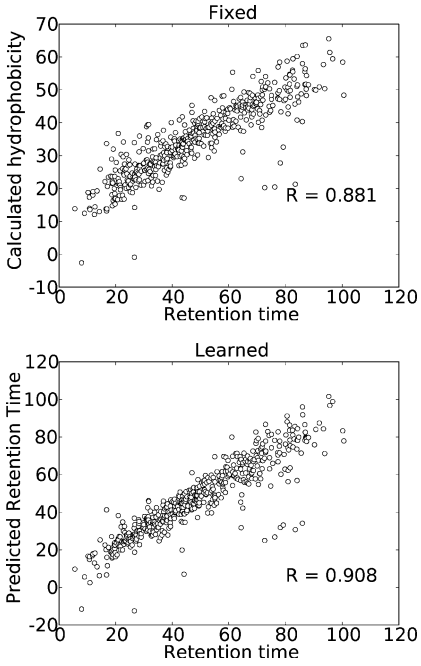

Figure 3. Example of retention time prediction. Predictions of hydrophobicity, a proxy for retention time, made by a fixed parameter linear regression Krokhin et al.[15] (top) are less accurate than retention time predictions by a support vector regression that is trained and tested on subsets of data from the same chromatography run (bottom).

observed and predicted retention time for our SVR with the correlation between observed and predicted relative hydrophobicity from the fixed-parameter regression. One of the kernels (either the Gaussian or linear kernel) outperforms the fixed parameter regression across all data sets (Table 2 and Figure 3). When tested with the two-sided Wilcoxon sign-ranked test, the correlations of the Gaussian kernel predictions with retention time is greater than the Krokhin et al.[15] correlations with a p-value of <0.01. The performance of the fixed and learned regressors are nonetheless qualititatively similar: data sets that had relatively poor correlation for one method had similarly poor correlation for Krokhin et al.[15] In general, the regression performs best on the data sets with a large number of high confidence identifications (Table 1).

Residue Weights. An advantage of using a linear kernel for the SVR is that it allows calculation of the weights for each feature, using the following formula:

$$w = \sum_i \alpha_i x_i \qquad (1)$$

where w is the feature weight vector, x_i is the ith training example, and α_i is the weight associated with the ith training example by the SVR. Weights correspond to each feature's relative contribution to retention time. After performing the regression on each data set, we calculate the weights given to each residue for peptide composition, shown in Figure 5 and the Supporting Information. We observe several expected trends: hydrophobic residues such F and W have higher weights; hydrophilic residues such as K and R show lower weights. Again, as expected, weights for different length columns (20, 40, 60 cm) are similar and apparently differ only by a scaling factor. While the SVR weights are largely consistent across chromatography conditions, there are some notable differences, such as the weight of H in the MudPIT-1 data set versus the other data sets.

Improved Peptide Identification. In addition to measuring the R value of predicted retention time on the test set, each trained SVR is also tested for its ability to eliminate false positive peptide identifications from its respective chromatography run. It is important to note that even had the Krokhin et al.[15] hydrophobicity predictions been better correlated with retention time than the SVR predictions, the method would still be less useful for improving peptide identification. This method cannot be used for filtering out identifications with unexpected retention time, because hydrophobicity gives only *relative* rather than absolute retention time. Thus, we see our method as a natural extension of the method presented in Krokhin et al.[15] to peptide identification improvement. We assess confidence of peptide identifications by searching the spectra from each data set against a shuffled version of the appropriate protein sequence database known as the decoy database; any hits to this database above a particular threshold are considered an estimate of the number of false positives (FP) against the target fasta database. Then, if P is the number of positive hits to the target database, FDR can be calculated using FDR = FP/P. To reduce the FDR, we eliminate identifications with observed retention time that deviates from the predicted retention time by a constant amount of time (Figure 1). We then measure whether this filtering step improves the number of positives over a range of FDR thresholds compared to number of positives without filtering. An example of the deviation of predicted and observed retention time for matches to the target and decoy databases is shown in Figure 4.

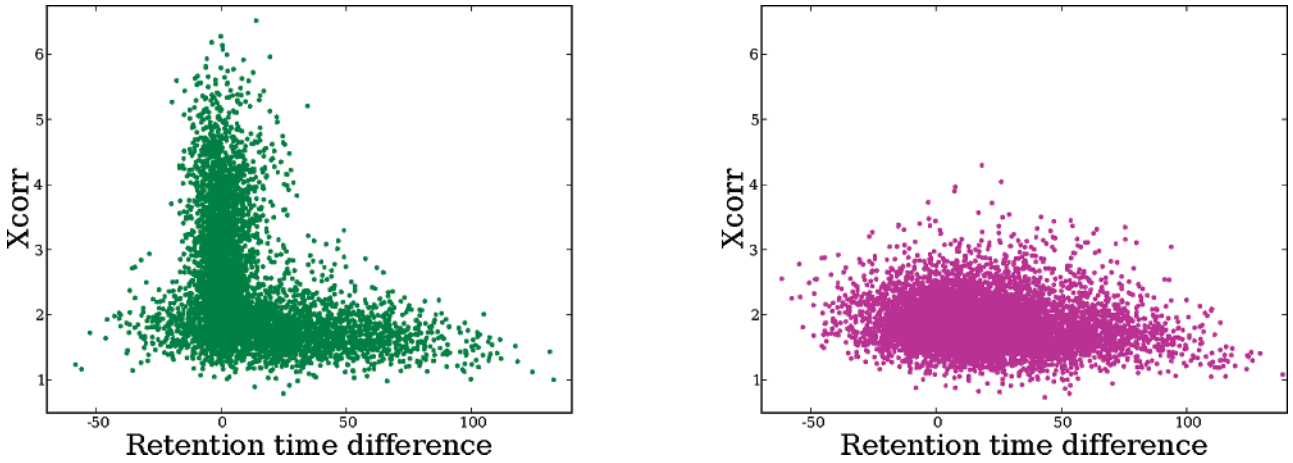

Figure 4. Predicted retention time difference for target and decoy peptide−spectrum matches. Shown are the Xcorr values and difference between observed retention time and retention time predicted by the Gaussian kernel for matches to the target database (green) and the decoy database (magenta) for the 20-cm data set.

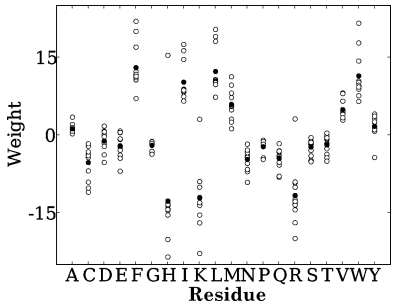

Figure 5. Contributions to retention time. Shown are the support vector regression weights for the linear kernel for the 20 features corresponding to peptide amino acid composition; higher values indicate a positive contribution to retention time. White circles indicate the individual weights for each of the 12 data sets; black circles indicate the means for all data sets.

Figure 2 illustrates how peptide identifications are filtered using retention time. In addition to the peptide−spectrum matches (PSMs) from the target yeast sequence database, we use PSMs from three decoy databases. The first decoy database is used to select identifications at 10% FDR, as described in the Methods section. The second decoy database is used to calculate the positives across a range of FDR values between 0.5 and 10% (in 0.5% increments) for a range of retention time thresholds between 0 and 240 min (in 10-min increments). The retention time threshold that produces the highest number of true positives across the largest number of FDR values is selected as the optimal maximum retention time deviation threshold. We then determine the performance of that threshold by calculating positives across the same range of FDR values using the third decoy database. We repeat this procedure five times and report an average of the positives obtained on each of the five iterations. This average of positives is compared to an average of positive performance without any retention time filtration across the same five iterations. The multiple iterations are made necessary by the high variance associated with false positive estimates from decoy databases.[27]

The results, in Figure 6 and Table 3, show a consistent decrease in false positive peptide identifications across all the data sets and most FDR thresholds. The dynamically trained SVR effectively adapts to variation in column length, MudPIT salt step, ion-pairing agent, and protease (Figure 6). The improvement in peptide identification is largest with the nontryptic digest elastase. Increases in positives tend to be largest in the 2−3% FDR range, and the Gaussian kernel outperforms the linear kernel in most cases, except for the 20- and 60-cm data sets. At a 3% FDR, the largest relative increase in positive peptide identifications is 52% for the Gaussian kernel on the elastase data set, from 509 to 772 identifications; the smallest increase is 15% for the 60-cm data set, from 1967 to 2270 identifications. Unique peptide identifications increase when the MudPIT steps are analyzed individually and as a group (Table 3).

Training and Testing on Different Data Sets. A primary hypothesis motivating our work is that a static regressor will not generalize well across different conditions. To demonstrate that it is necessary to train a new model on each data set, we trained an SVR on the MudPIT-2 data set and tested it on the TFA data set. Filtering for retention time in this case degrades performance versus not filtering, as shown in Figure 7. For example, at an FDR of 5%, before filtering the TFA data set has 2550 unique positive peptide identifications, compared to 2287 after filtering. Both analyses show reduction in performance compared to an SVR both trained and tested on the TFA data set, which has 2868 identifications.

DISCUSSION

We have demonstrated that a dynamically trained support vector regressor is capable of learning to predict peptide retention time from a small collection of data from a single LC−MS/MS run across a variety of conditions. Our approach is capable of adapting to varying chromatography conditions. Our goal is not to predict retention time per se, but rather to use retention time as a means to improve peptide identfication. Using the SVR to filter peptide identifications results in an increase in positive identifications at most false discovery rate thresholds and data sets. Of special interest is the improvement in identifications for samples with nontryptic enzyme cleavage, analysis of which is usually complicated by an inability to constrain identifications with knowledge of enzyme cleavage specificity. It is important to note that filtering identifications in this manner is impossible with other

(27) Huttlin, E. L.; Hegeman, A. D.; Harms, A. C.; Sussman, M. R. *J. Proteome Res.* **2006**.

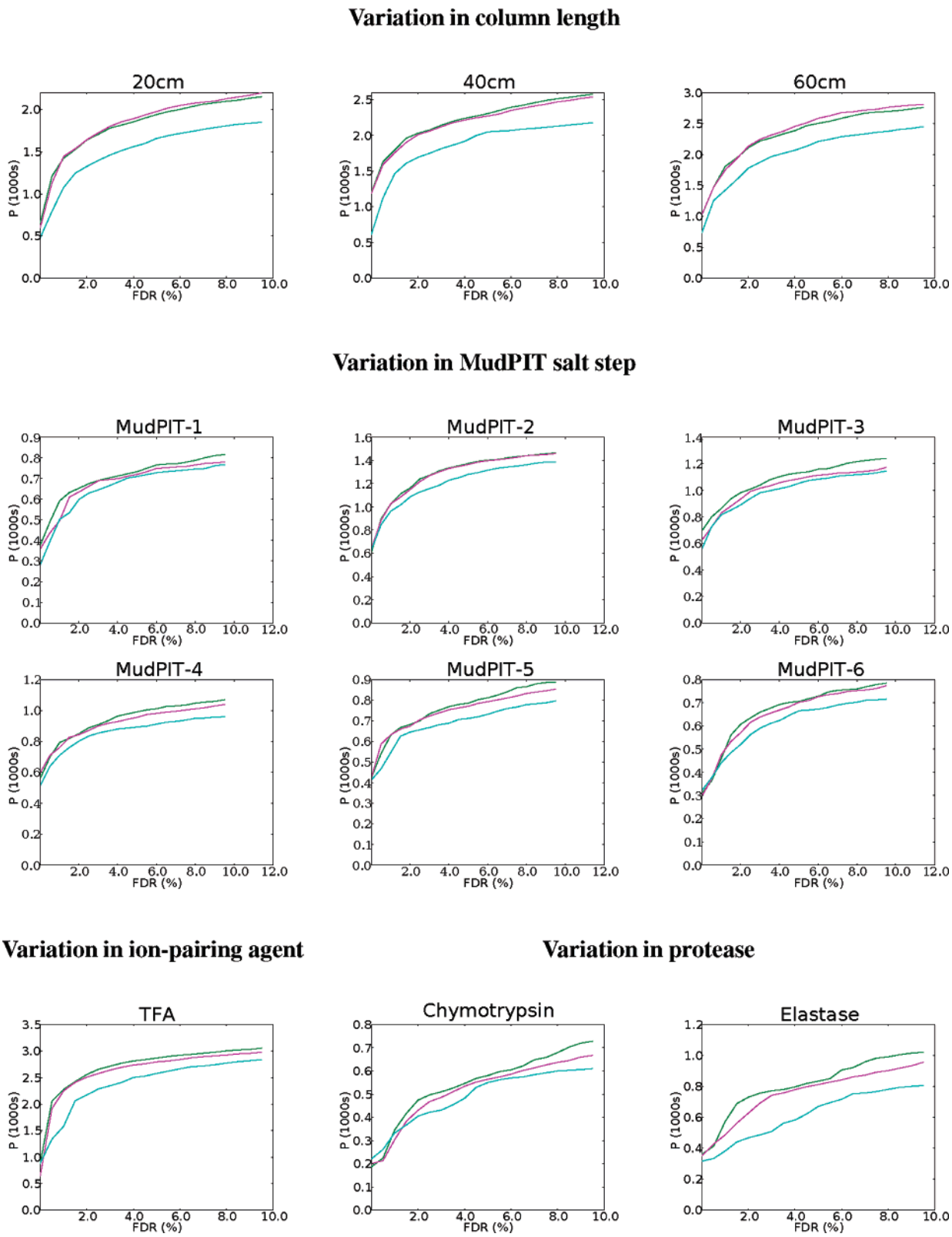

Figure 6. Improved peptide identification over varying conditions**.** The dynamically trained SVR is able to cope with differences due to variations in column length (top), MudPIT salt step (middle), ion-pairing agent (bottom, left) and protease (bottom, right). Spectra from diverse conditions are searched against the appropriate sequence database to yield positive IDs and a decoy database to yield an estimate of false positive IDs. Shown are plots of FDR vs positives (*P*). The green and magenta curves are for the test data set after filtering with the best classifier found on the training data using the Gaussian and linear kernels, respectively, while the cyan curve is without any filtering.

methods of predicting retention time (such as calculating relative hydrophobicity) because these methods predict only relative and not absolute retention time.

Our SVR method is not without limitations. In particular, data sets of low complexity, such as two-dimensional gel spots or purified proteins, will not produce a diverse enough set of peptides

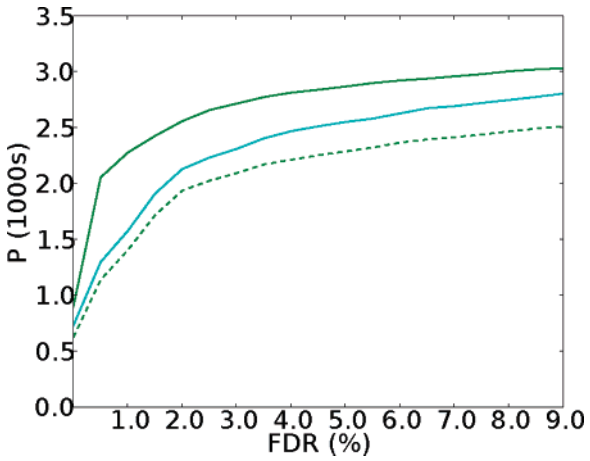

Figure 7. Performance deterioration when applying an SVR trained on data from one kind of chromatography to data from a different kind of chromatography**.** Positives vs FDR for a Gaussian kernel SVR trained on a data set with a typical number of identifications (MudPIT-2) and tested on another (TFA) before (cyan) and after filtering (dashed green) peptide identifications with unexpected retention time. Also shown for comparison is the curve after both training and testing with peptides from the TFA data set (solid green).

to allow for accurate regression. In addition, poor-quality data sets with few identifications (less than 100 above the 10% FDR) will also fail to yield good regressions (see Supporting Information). Analysis of data sets from relatively simple mixtures could benefit from improved selection of high-confidence identifications or from an approach that combines data from the poor-quality data set with data from higher quality data sets.

Further enhancements to our algorithm are possible. In practice, the optimal retention time filtering threshold may vary throughout the chromatography run. Therefore, future work includes modulating the retention time threshold used to filter peptide identifications to account for variability in chromatography quality across a run.

ACKNOWLEDGMENT

The authors thank the anonymous reviewers for their positive feedback and constructive criticism. This research was supported in part from funding acquired from the National Institutes of Health Grants T32 HG00035, P41 RR11823, R01 DK069386, and R01 EB007057.

SUPPORTING INFORMATION AVAILABLE

Additional information as noted in text. This material is available free of charge via the Internet at http://pubs.acs.org.

Received for review February 7, 2007. Accepted May 24, 2007.

AC070262K