Peptide Retention Time Prediction Yields Improved Tandem Mass Spectrum Identification for Diverse Chromatography Conditions

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

Genome Sciences Department
University of Washington
1705 NE Pacific Street
Seattle, WA 98195-5065
{aklammer,xhyi,maccoss,noble}@gs.washington.edu

Abstract. Most tandem mass spectrum identification algorithms use information only from the final spectrum, ignoring precursor information such as peptide retention time (RT). Efforts to exploit peptide RT for peptide identification can be frustrated by its variability across liquid chromatography analyses. We show that peptide RT can be reliably predicted by training a support vector regressor on a *single* chromatography run. This dynamically trained model outperforms a published statically trained model of peptide RT across diverse chromatography conditions. In addition, the model can be used to filter peptide identifications that produce large discrepancies between observed and predicted RT. After filtering, estimated true positive peptide identifications increase by as much as 50% at a false discovery rate of 3%, with the largest increase for non-specific cleavage with elastase.

Keywords: Mass spectrometry, proteomics, peptide identification, retention time, chromatography, machine learning, support vector regression.

1 Introduction

Full understanding 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 or all of an the cell's proteins. One such technology is shotgun proteomics using liquid chromatography and tandem mass spectrometry (LC-MS/MS)[1] (McCormack et al., 1997; Yates, III, 1998).

[1] Abbreviations used in this manuscript include retention time (RT), liquid chromatography (LC), mass spectometry (MS), tandem mass spectrometry (MS/MS), support vector regressor (SVR), artificial neural network (ANN) and false discovery rate (FDR).

T. Speed and H. Huang (Eds.): RECOMB 2007, LNBI 4453, pp. 459–472, 2007.

In a typical liquid chromatography (LC)-MS/MS experiment (Figure 1A), proteins are digested to peptides, and the peptides are separated by LC on a reverse phase column in order of increasing hydrophobicity. The peptides elute into the mass spectrometer, where tandem mass spectrometry (MS/MS) measures the mass-to-charge ratio of the intact and fragmented peptides, yielding a tandem mass spectrum. One LC-MS/MS experiment 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 automatically deduced by a database search algorithm such as SEQUEST (Eng et al., 1994).

As with any high-throughput technology, shotgun proteomics practioners must constantly battle false positive identifications (Cargile et al., 2004; Qian et al., 2005). The need to reduce false positives has spurred a proliferation of methods for increasing peptide identification confidence. However, most of these methods use information exclusively from the MS and MS/MS stages of analysis, ignoring information from the LC stage, such as retention time (RT). RT is the amount of time that a peptide is retained on the LC column (Figure 1B, top). It has the advantage of being almost entirely independent of the information contained in the MS/MS scan, and can therefore be used to increase peptide identification confidence.

The goal of this paper is to incorporate RT into the peptide identification process to increase peptide identification confidence. Previous efforts along these lines have been hindered by RT variability, even on identical columns or multiple runs of the same sample (Palmblad et al., 2004). Most such methods train a single RT predictor using a limited subset of highly-reproducible chromatography conditions (Krokhin et al., 2004), or perform a normalization that attempts to eliminate variability (Petritis et al., 2006; Strittmatter et al., 2004). In practice, however, researchers use a large number of diverse chromatographic conditions, making a static RT predictor less useful. In this work, we demonstrate how to dynamically train a support vector regressor (SVR) to predict RT for peptides in a given chromatographic analysis, using only data generated during the current run using composition related features (Figures 2 and 1B, bottom).

This approach makes the method portable to new chromatography conditions or sample preparation protocols, adapting to differences in column length, digestion condition, peptide chemistry and MudPIT salt step. Our RT predictions are better correlated with observed RT than those produced by a static predictor trained on different data. Furthermore, by eliminating peptide identifications with an observed retention time that deviates greatly from predicted retention time, our method increases the number of true positive peptide identifications over a range of false discovery rates. For one data set digested with a non-standard enzyme (elastase), we demonstrate an increase of approximately 50% in true positives at a false discovery rate (FDR) of 3%. This result compares favorably with (Strittmatter et al., 2004) (a true positive increase of 15% at 3% FDR, from Table 2), but with much less training data. Thus, the results presented here have implications both for traditional shotgun proteomics research using trypsin, as well as possibly enabling new strategies using non-standard enzymes.

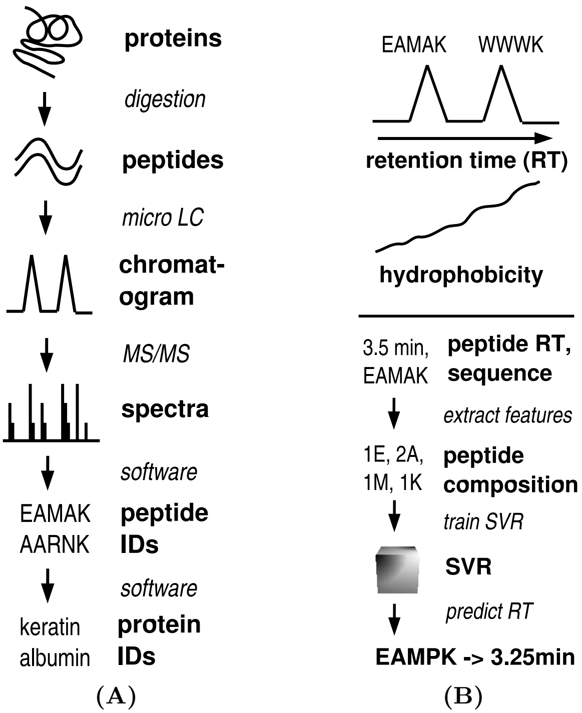

Fig. 1. Experimental overview. (A) In mass spectrometry, proteins are digested to peptides, which are then separated using liquid chromatography and analyzed in a tandem mass spectrometer. The experimental procedure yields a chromatogram, MS and MS/MS spectra, and (ideally) peptide and protein identifications. (B, top) For reverse phase chromatography, each peak in a chromatogram corresponds to a peptide retained on the column for an amount of time that depends on its hydrophobicity. (B, bottom) We train a support vector regressor with composition-related features to predict RT for unknown peptides from the *same* chromatography run.

1.1 Related Work

Understanding and predicting peptide RT has a long history. For reverse phase chromatography, peptide RT is roughly proportional to peptide hydrophobicity (Frenz et al., 1990). Many models assume that peptide RT is a linear function of peptide amino acid composition (Meek, 1980; Browne et al., 1982; Guo et al., 1987; Hearn et al., 1988; Bihan et al., 2004). More recent models augment the compositional approach with parameters for peptide length or mass (Mant et al., 1989), or terms for residue context (Mant and Hodges, 2006) or positional effects such as the identity of the N-term residue (Krokhin et al., 2004). Still more sophisticated models include parameters for structural features or measured chemical properties (Bączek et al., 2005; Petritis et al., 2006).

The most accurate and sophisticated peptide RT predictor is that of Petritis et al. (2006), first presented in simpler form in Petritis et al. (2003),

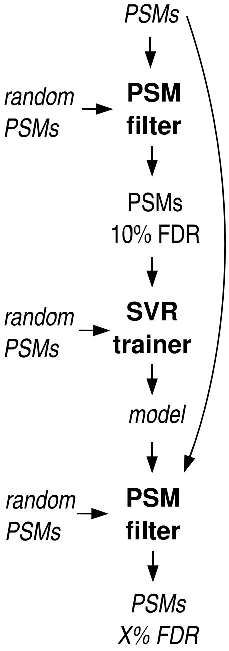

Fig. 2. Overview of data flow for peptide identification improvement. For each LC-MS/MS experiment, we start with a collection of peptide-spectrum matches (PSMs, top). We use PSMs to a random proteome to filter the original PSMs, producing high-confidence PSMs at a 10% false discovery rate. These filtered PSMs, along with a second set of random 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 random PSMs are used to select the final set of real PSMs at a desired FDR. Not shown is the five-fold cross-validation used to validate this method.

which uses an artificial neural network (ANN) to predict a normalized form of RT. The large amount of data required to train the ANN (for Petritis et al. (2006) 345,000 nonredundant peptides) makes retraining for new chromatography conditions impractical. Although one could in theory transform predicted RT values for different conditions, it is not clear how to handle changes in peptide elution order. A more recent, but less complicated, ANN has since been published (Shinoda et al., 2006).

Recently, a handful of RT predictors have found practical application in enhancing confidence of peptide identifications. Palmblad et al. (2002) predict RT for each peptide using least-squares regression to determine amino acid weights, and then use a χ^2 test to rank candidate peptides based on deviation from expected mass and predicted RT. As the authors admit, their RT prediction is poor compared to other competing efforts, and improvement in protein identification is modest.

Kawakami et al. (2005) use the sum of residue retention coefficients to predict RT for peptides and phosphopeptides, but they make no clear distinction between training and test data. When predictions are made on data not used in training, the correlation between predicted and observed RT deteriorates, a sign of overfitting.

By far, Strittmatter et al. (2004) is the most successful example of using RT prediction to increase peptide identification confidence. They use the ANN of Petritis et al. (2003) to exclude all SEQUEST peptide identifications with predicted normalized RT that deviates more than 10% from observed normalized RT. The result is roughly a 50% decrease in estimated false positives, although true positives also decrease, frustrating a straightforward interpretation.

The methods presented in this paper yield a reduction in false positives as great as that in Strittmatter et al. (2004), but require reduced training data, produced from a single LC-MS/MS run.

2 Methods

2.1 Data Sets

We analyze eight separate data sets (Table 1), chosen to represent a diverse set of chromatography conditions. Exact sample preparation protocols are given in the supplement (`noble.gs.washington.edu/proj/rt`); here we give only brief descriptions to highlight major differences. All data sets are from the yeast *Saccharomyces Cerevisiae*. The first three data sets are taken from the middle and end of a 12-hour, 6-step, 2-phase strong cation exchange and reverse phase multi-dimensional protein identification technology (MudPIT) analysis (Washburn et al., 2001) of a tryptic digest of the soluble *S. Cerevisiae* proteome. The MudPIT was performed with C18 beads, while all subsequent analyses are with C12 beads. The number after the C refers to the length of the carbon chain on the beads to which the peptides bind. Different length chains interact with the peptides differently. The next three data sets are reverse phase analyses of a tryptic digest of the soluble yeast proteome, each with a different length column of 20cm, 40cm and 60cm. A fourth identically prepared yeast sample was analyzed with the ion-pairing agent trifluoroacetic acid (TFA). The two final data sets are from yeast samples digested with the non-specific enzymes chymotrypsin or elastase. Chymotrypsin cleaves after aromatic residues F, W, and Y, and elastase cleaves after small hydrophobic residues A, L, I and V. Summary statisics for the data sets, and for the training and testing data sets extracted from them, are shown in Table 1.

2.2 Training and Testing Set Extraction

A high-confidence set of training and testing data is extracted from each of the eight data sets. The spectra are first searched against both the real and shuffled versions of the *S. cerevisiae* proteome with SEQUEST (Eng et al., 1994) and then identifications are filtered using the following criteria: charge state of +2, peptide sequence ending in K or R (except for the chymotrypsin and elastin data sets), and allowing any number of missed tryptic cleavages.

We use the number of matches in the search against a shuffled proteome as an estimate of false positive matches in a search against the real proteome. Both searches use identical search criteria. High-confidence spectra identifications are selected by setting an Xcorr threshold so that the number of matches to the shuffled proteome above this threshold is 10% of the number of matches to the real proteome; this is equivalent to a 10% FDR. If the number of real matches at a 10% FDR is less than 200, then the top 200 spectra are used; this is because regression performance deteriorated with less than 200 spectra. When multiple spectra matched a single peptide according to these criteria, the spectrum with the highest Xcorr is selected, to avoid bias in the regression towards common peptides. The resulting set of peptides and retention times is split to form a 3:1 ratio between the training and testing data sets (Table 1) for each chromatography run, which are then used to train and test the SVR. No peptides are allowed to occur in both the training and testing data sets.

2.3 Support Vector Regression

As with other forms of regression, an SVR learns a function that relates a dependent variable (in this case, RT) to a set of independent variables. An SVR builds a regressor out of a subset of the training examples, known as support vectors. Training examples that are within a tolerance value ϵ of the model prediction are ignored (Vapnik, 1995). To generate the independent variables, each peptide from the training and test sets is represented as a 63-element vector comprised of the following: 20 elements to represent the total number of each amino acid residue in the peptide; 40 binary elements to represent the identity of the extreme N-terminal (N-term) and penultimate C-terminal (C-term) residues, respectively; and three additional elements to represent the identity of the last C-term residue (either K or R), and the peptide length and mass. For the non-specific enzymes, the ultimate C-terminal residue is used instead of the penultimate, and the K or R term is set to zero.

An SVR is trained on each high-quality training set and tested by measuring the R value between predicted and observed RT on a held-out test set. 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 dataset x. It is important to note that a separate SVR is trained for each data set in Table 1.

The 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 (Section 3.2); 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 chosen by three-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 radically 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. C was initially allowed to range over ten 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} and 10^{-8}. R values are reported after hyperparameter selection on the appropriate held-out test set (Table 1).

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`.

Table 1. Eight data sets used to train and test the support vector regressor. Each column lists the total number of +2 spectra associated with peptides that satisfy that data set's trypticity requirements (Total), the number of high-confidence spectra selected at a 10% FDR (Confident), and the subsets of the high-confidence spectra used to train and test the performance of the regressor.

Data set	Total	Confident	Train	Test
Y-20CM	6929	2073	1554	519
Y-40CM	7220	2409	1806	603
Y-60CM	7459	2774	2080	694
Y-TFA	11977	3179	2384	795
Y-CHYMO	2191	200	150	50
Y-ELAST	4377	200	150	50
Y-MUDPIT-1	2227	280	210	70
Y-MUDPIT-2	3035	485	363	122

3 Results

3.1 Support Vector Regression

We first evaluate our dynamically trained regressor by comparing it to a published, fixed-parameter regressor from Krokhin et al. (2004). We measure performance by comparing correlation (measured by R value) between observed and predicted RT for our SVR with the correlation between observed and predicted relative hydrophobicity from the fixed-parameter regression. One of the kernels (either Gaussian or linear kernel) outperforms the fixed parameter regression across all data sets (Table 2 and Figure 3). Furthermore, the performance of the fixed and learned regressors are qualitatively the same: data sets that had relatively poor correlation for one method had similarly poor correlation for the other. In general, the regression performs best on data sets with a large number of high-confidence identifications (Table 1).

Table 2. R values for a fixed regression compared to a learned regression using the Gaussian or linear kernels. Correlation for eight data sets for fixed parameters described in Krokhin et al. (2004) (Fixed) and parameters learned for each dataset with a Gaussian or linear kernel. The Fixed values differ in the first and third columns because they are evaluated on slightly different randomly selected subsets of the high-confidence PSMs.

Data set	Fixed	Gaussian	Fixed	Linear
20CM	0.881	0.908	0.877	0.892
40CM	0.892	0.897	0.894	0.891
60CM	0.914	0.926	0.889	0.892
CHYMO	0.871	0.865	0.761	0.792
ELAST	0.823	0.850	0.843	0.856
TFA	0.818	0.842	0.882	0.905
MUDPIT-1	0.743	0.783	0.797	0.850
MUDPIT-2	0.806	0.803	0.791	0.828

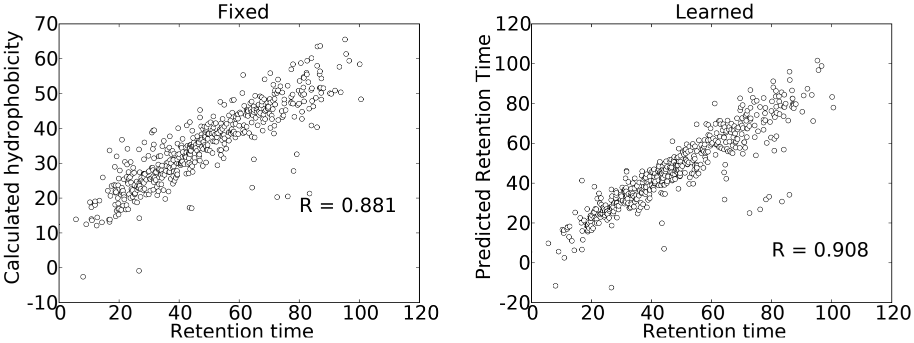

Fig. 3. Example of retention time prediction. Predictions of hydrophobicity, a proxy for retention time (RT), made by a fixed parameter linear regression from Krokhin et al. (2004) (left) are less accurate than RT predictions by a support vector regression that is trained and tested on subsets of data from the same chromatography run (right).

3.2 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:

$$\hat{w} = \sum_i \alpha_i \hat{x}_i \tag{1}$$

where $\hat{w}$ is the feature weight vector, $\hat{x}_i$ is the ith training example (in this case, the 63-element vector representing a peptide), and α_i is the weight associated with the ith training example by the SVR. Weights correspond to the 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

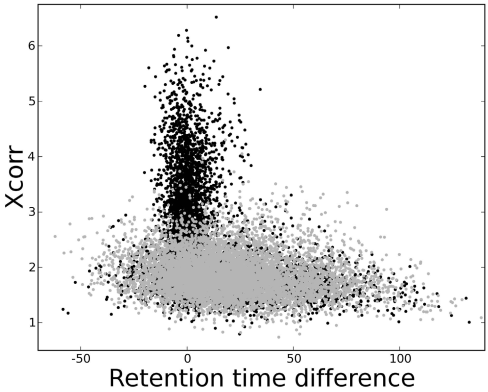

Fig. 4. Predicted retention time difference for real and random peptide-spectrum matches. Shown are the Xcorr values and difference between observed RT and RT predicted by the Gaussian kernel for matches to the real yeast proteome (black) and the shuffled yeast proteome (gray) for the 20CM data set.

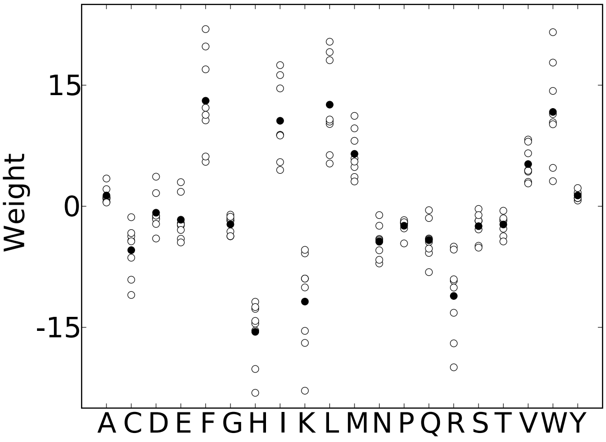

Fig. 5. Contributions to retention time (RT). Shown are the support vector regression weights for the linear kernel for the 20 features corresponding to peptide amino acid composition; higher values indicate a positive contribution to RT. White circles indicate the individual weights for each of the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

Figure 5. We observe several expected trends: hydrophobic residues such F and W have higher weights, and hydrophilic residues such as K and R show lower weights. While the SVR weights are largely consistent across chromatography conditions, there are some notable differences, such as the relative weight of K and R. Weights for different length columns (20CM, 40CM, 60CM) are qualitatively similar, but differ in magnitude. The largest weights are associated with the non-specific cleavages, while the smallest are associated with the MudPIT analysis (supplement).

3.3 Improved Peptide Identification

In addition to measuring the R value of predicted RT on the test set, each trained SVR is also tested for its ability to eliminate false positive peptide identifications from its respective chromatography run. We assess confidence of peptide identifications by searching the spectra from each data set against a shuffled version of the appropriate organism's proteome sequence database; any hits to this database above a particular Xcorr threshold are considered an estimate of the number of false positives FP against the *real* database. Then, if P is the number of positive hits to the real database, FDR can be calculated using: $FDR = FP/P$. To reduce FDR, we eliminate identifications with observed RT that deviate from the predicted RT by a constant amount of time, and then measure whether this filtering step improves the number of true positives over a range of FDR thresholds compared to identifications without filtering. An example of the deviation of predicted and observed retention time for matches to the real and random proteomes is shown in Figure 4.

The retention time threshold used to filter identifications is identified in the following manner, as outlined in schematic in Figure 2. In addition to the PSMs from the real yeast proteome, we use PSMs from three shuffled proteomes. The first shuffled proteome is used to select identifications at 10% FDR, as described in Section 2.2. The second shuffled proteome is used to calculate the true 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 minutes (in 10 minute 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 RT deviation threshold. We then determine the performance of that threshold by calculating true positives across the same range of FDR values using the third shuffled proteome. We repeat this procedure five times, and report an average of the true positives obtained on each of the five iterations. This is compared to an average of true 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 shuffled proteomes (Huttlin et al., 2006).

The results, shown in Figure 6, 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 (Figure 6, top), digestion conditions (Figure 6, middle) and MudPIT salt step (Figure 6, bottom). The improvement in peptide identification is largest with the non-specific digest elastase. Increases in true positives tend to be largest in the 2% to 3% FDR range, and the Gaussian kernel outperforms the linear kernel in most cases, except for the 60CM column. At a 3% FDR, the largest relative increase in true positive peptide identifications is 52% for the Gaussian kernel on the ELAST data set, from 509 to 772 identifications; the smallest increase is 15% for the 60CM data set, from 1967 to 2270 identifications.

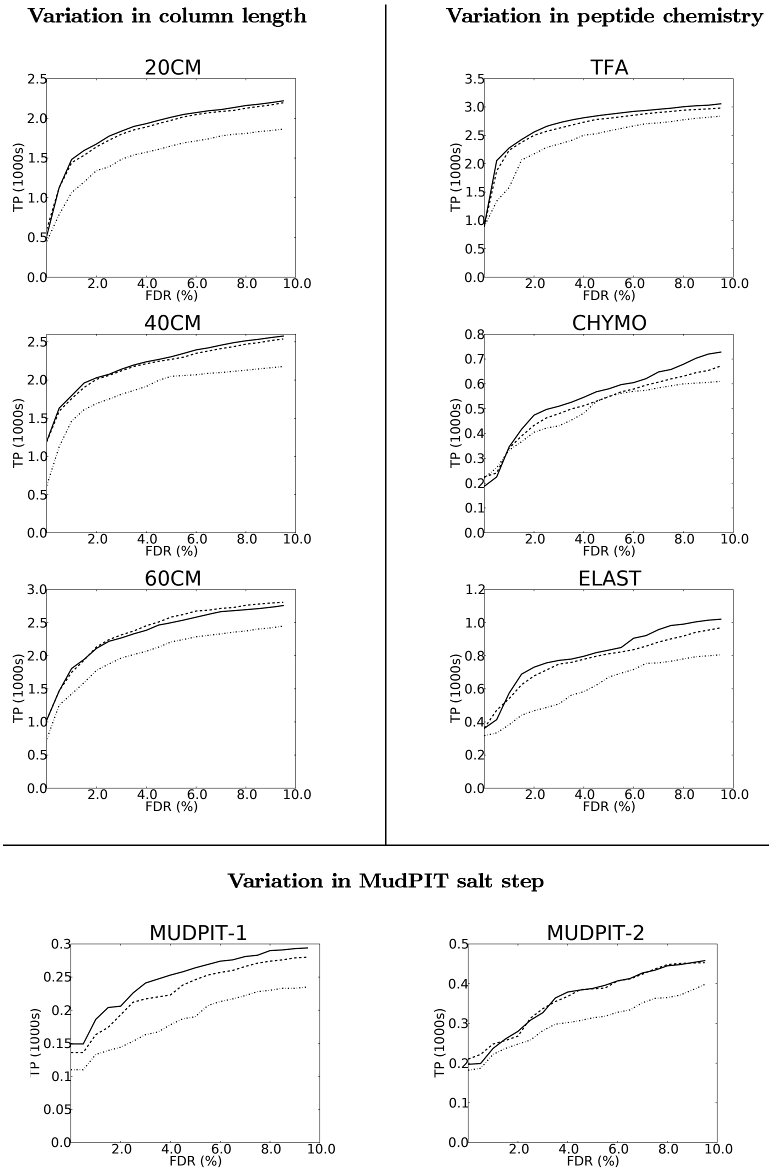

Fig. 6. Improved peptide identification over varying conditions. The dynamically trained SVR is able to cope with chromatographic differences due to variations in column length (left), peptide chemistry (right) and MudPIT salt step (bottom). Spectra from diverse chromatography conditions are searched against the appropriate proteome to yield positive IDs and a shuffled proteome to yield an estimate of false positive IDs. Shown are plots of false discovery rate vs. true positives. The solid curve (Gaussian) and heavy dotted curve (Linear) 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 light dotted curve (Unfiltered) is without any filtering.

4 Discussion

We have demonstrated that a dynamically trained support vector regressor is capable of learning to predict peptide RT from a single LC-MS/MS run across a variety of chromatographic conditions, adapting to variation in column length, digestion conditions, peptide chemistry, and MudPIT salt step. Furthermore, using the SVR to filter peptide identifications results in an increase in true positive identifications across almost all false discovery rates and data sets. Of special interest is the improvement in identifications for samples with non-specific enzyme cleavage, a form of analysis typically plagued by false positive identifications.

It is important to note that filtering identifications in this manner is not possible with other methods of predicting RT (such as calculating relative hydrophobicity, as in (Krokhin et al., 2004)), since these methods only predict relative, not absolute retention time. This is highlighted by the difference in scales between the learned and fixed retention time regression in Figure 3. Converting relative to absolute retention time would require methods similar to those outlined here.

Our SVR method does not come without limitations. In particular, data sets of low complexity would probably not produce a diverse enough set of peptides 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. Analysis of such data sets 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.

Acknowledgments. The authors would like to thank the NSF for funding through grant T32 HG00035 and the NIH for funding through grant P41 RR11823, as well as anonymous reviewers for providing several useful suggested improvements.

Bibliography

Bączek, T., P. Wiczling, M. Marszałł, Y. V. Heyden, and R. Kaliszan (2005). Prediction of peptide retention at different HPLC conditions from multiple linear regression models. *Journal of Proteome Research 4*, 555–563.

Bihan, T. L., M. D. Robinson, I. I. Stewart, and D. J. Figeys (2004). Definition and characterization of a “trypsinosome” from specific peptide characteristics by nano-HPLC-MS/MS and *in silico* analysis of complex protein mixtures. *Journal of Proteome Research 3*, 1138–1148.

Browne, C. A., H. P. J. Bennett, and S. Solomon (1982). The isolation of peptides by high-performance liquid chromatography using predicted elution positions. *Analytical Biochemistry 124*, 201–208.

Cargile, B. J., J. L. Bundy, T. W. Freeman, and J. J. L. Stephenson (2004). Potential for false positive identifications from large databases through tandem mass spectrometry. *Journal of Proteome Research 3*, 1082–1085.

Eng, J. K., A. L. McCormack, and J. R. Yates, III (1994). An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *Journal of the American Society for Mass Spectrometry 5*, 976–989.

Frenz, J., W. S. Hancock, W. J. Henzel, and C. Horva'th (1990). *HPLC of Biological Macromolecules: Methods and Applications*. Marcel Dekker.

Guo, D., C. T. Mant, A. K. Taneja, J. M. Parker, and R. S. J. Hodges (1987). Effects of ion-pairing reagents on the prediction of peptide retention in reversed-phase high-performance liquid chromatography. *Journal of Chromatography 386*, 205–222.

Hearn, M. T., M. I. Aguilar, C. T. Mant, and R. S. Hodges (1988). High-performance liquid chromatography of amino acids, peptides and proteins. LXXXV. evaluation of the use of hydrophobicity coefficients for the prediction of peptide elution profiles. *Journal of Chromatography 438*, 197–210.

Huttlin, E. L., A. D. Hegeman, A. C. Harms, and M. R. Sussman (2006). Prediction of error associated with false-positive rate determination for peptide identification in large-scale proteomics experiments using a combined reverse and forward peptide sequence database strategy. *Journal of Proteome Research*.

Kawakami, T., K. Tateishi, Y. Yamano, T. Ishikawa, K. Kuroki, and T. Nishimura (2005). Protein identification from product ion spectra of peptides validated by correlation between measured and predicted elution times in liquid chromatography/mass spectrometry. *Proteomics 5*, 856–64.

Krokhin, O. V., R. Craig, V. Spicer, W. Ens, K. G. Standing, R. C. Beavis, and J. A. Wilkins (2004). An improved model for prediction of retention times of tryptic peptides in ion pair reversed-phase hplc. *Molecular & Cellular Proteomics 3*, 908–919.

Mant, C. T. and R. S. Hodges (2006). Context-dependent effects on the hydrophilicity/hydrophobicity of side-chains during reversed-phase high-performance liquid chromatography: Implications for prediction of peptide retention behaviour. *Journal of Chromatography A 1125*, 211–219.

Mant, C. T., N. E. Zhou, and R. S. Hodges (1989). Correlation of protein retention times in reversed-phase chromatography with polypeptide chain length and hydrophobicity. *Journal of Chromatography A 476*, 363–375.

McCormack, A. L., D. M. Schieltz, B. Goode, S. Yang, G. Barnes, D. Drubin, and J. R. Yates, III (1997). Direct analysis and identification of proteins in mixtures by LC-MS/MS and database searching at the low-femtomole level. *Analytical Chemistry 69*(4), 767–776.

Meek, J. L. (1980). Prediction of peptide retention times in high-pressure liquid chromatographic on the basis of amino acid composition. *Proceedings of the National Academy of Sciences of the United States of America 77*, 1632–1636.

Palmblad, M., M. Ramstrom, G. B. Bailey, S. L. McCutchen-Maloney, J. Bergquist, and L. C. Zeller (2004). Protein identification by liquid chromatography-mass spectrometry using retention time prediction. *Journal of Chromatography, B 803*, 131–135.

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

Petritis, K., L. Kangas, B. Yan, M. E. Monroe, E. F. Strittmatter, W. J. Qian, J. N. Adkins, R. J. Moore, Y. Xu, M. S. Lipton, D. G. C. 2nd, and R. D. Smith (2006). Improved peptide elution time prediction for reversed-phase liquid chromatography-MS by incorporating peptide sequence information. *Analytical Chemistry 78*(14), 5026–5039.

Petritis, K., L. J. Kangas, P. L. Ferguson, G. A. Anderson, L. Pasa-Tolic, M. S. Lipton, K. J. Auberry, E. F. Strittmatter, Y. Shen, R. Zhao, and R. D. Smith (2003). Use of artificial neural networks for the accurate prediction of peptide liquid chromatography elution times in proteome analyses. *Analytical Chemistry 75*(5), 1039–1048.

Qian, W. J., T. Liu, M. E. Monroe, E. F. Strittmatter, J. M. Jacobs, L. J. Kangas, K. Petritis, D. G. C. II, and R. D. Smith (2005). Probability-based evaluation of peptide and protein identifications from tandem mass spectrometry and SEQUEST analysis: The human proteome. *Journal of Proteome Research 4*(1), 53–62.

Shinoda, K., M. Sugimoto, N. Yachie, N. Sugiyama, T. Masuda, M. Robert, T. Soga, and M. Tomita (2006). Prediction of liquid chromatographic retention times of peptides generated by protease digestion of the escherichia coli proteome using artificial neural networks. *Journal of Proteome Research 5*, 3312–3317.

Strittmatter, E. F., L. J. Kangas, K. Petritis, H. M. Mottaz, G. A. Anderson, Y. Shen, J. M. Jacobs, D. G. C. 2nd, and R. D. Smith (2004). Application of peptide LC retention time information in a discriminant function for peptide identification by tandem mass spectrometry. *Journal of Proteome Research 3*(4), 760–769.

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

Washburn, M. P., D. Wolters, and J. R. Yates, III (2001). Large-scale analysis of the yeast proteome by multidimensional protein identification technology. *Nature Biotechnology 19*, 242–247.

Yates, III, J. R. (1998). Mass spectrometry and the age of the proteome. *Analytical Chemistry 33*, 1–19.