

A Model for Predicting Elution Time

Raja Loganantharaj

Abstract— The abundance of proteins in a given specimen is inferred by mass spectrometer similar to inferring mRNA expression levels using array technology. It has been shown that prediction accuracy of protein with peptide finger printing will improve when it is coupled with predicted elution time. This has motivated the work on predicting elution time of a peptide. We have developed a model for representing peptide as to be used by a support vector regression (SVR) for predicting elution time. We have shown that our model predicted close to 93% of the observed time.

Index Terms—support machine regression, peptide model, chemical property of peptide, hydrophobicity.

I. INTRODUCTION

While the microarray technology provides the expression levels of mRNA, it is unable to provide an accurate protein abundance level in a specimen. In order to provide a comprehensive view of abundance of proteins of a test specimen, several technologies have been developed recently including a shotgun proteomics using liquid chromatography (LC) and tandem mass spectrometry (MS/MS). In a typical LC MS/MS experiment, proteins are digested by enzymes such as Trypsin, which break a long amino acids sequence by small peptide fragments of length from 3 to over 50 amino acids. In Liquid Chromatography the fragments are physically separated on the basis of hydrophobicity and these peptides elute into mass spectrometer at multiple, unique time points so as to measure the mass to charge ratio. The identities of these corresponding peptides are deduced by finger printing the peptides using programs such as SEQUEST.

Several software programs are developed for identifying peptides/proteins from mass spectra, but they are plagued by higher false positive [1, 2] due to the complex nature of chemical digestive process and the vast number of possible number of peptide sequences. The retention time (RT) or the elution time is the time it takes for a peptide leaving the LC column after it entered into it. It is believed that false positive in identifying peptide/protein will be reduced significantly if the elution time is also taken into account in the identification process.

The early work [3, 4] on predicting retention time was based on retention coefficient approach that is based on summation

of empirically determined amino acid residue retention coefficient. Such approach seems to be working well for small peptides (up to 20 residues), but not suitable for many practical proteomics application where peptide residue may extend to 50 residues.

This paper is organized as following: following introduction, we describe some property of dataset and the factors that influence the prediction model. It is followed by a section on model where we describe the details of the model we developed for predicting the elution time. We describe the machine learning algorithm that builds a regression model so as to predict the elution time. It is followed by a section on results and finally we discuss the results and future work in the section on conclusion.

II. DATASET

Over 107,000 distinct peptides, protein fragments as a result of Trypsin digestion, were collected over different specimens and different runs at the Pacific Northwest Laboratory. The information recorded for each peptide include, mass to charge ratio, cleavage state, signal to noise ratio, cross correlation, and scan time. The observed Similar data sets have been used to predict elution time using feed forward neural network [5, 6].

The elution time of a peptide is influenced by many factors including composition and the chemical properties of the peptide, sample preparation, equipment, digestive enzymes and experimental setups. As expected, we have observed different elution time for each instance of the same peptide. The relative distance between the peaks remains relatively stable across different runs. Either we can align all the peaks across different run and create a data set with relative time, or experiment with a single run with relative time. The problem of aligning peaks itself is an interesting research problem. For the purpose of this paper, we use the instances of a single run to experiment the model and its utility in predicting elution time.

To minimize the noise, we have created a filter with charge state 2 or better, XCorrelation 2.5 or better, cleavage state 2 and signal to noise ratio 25 or better. After applying the filtering we selected few test instances that have over 1500 peptides. The observed elution time of the data set varies from 4 minutes to 96 minutes and the elution time distribution of the data set is shown in Figure 1. The length of the peptides in the data set varies from 4 to 50 and the distribution is shown in Figure 2. The mass/charge distribution of the dataset is shown

in Figure 3.

Figure 1: Distribution of elution time for the data set.

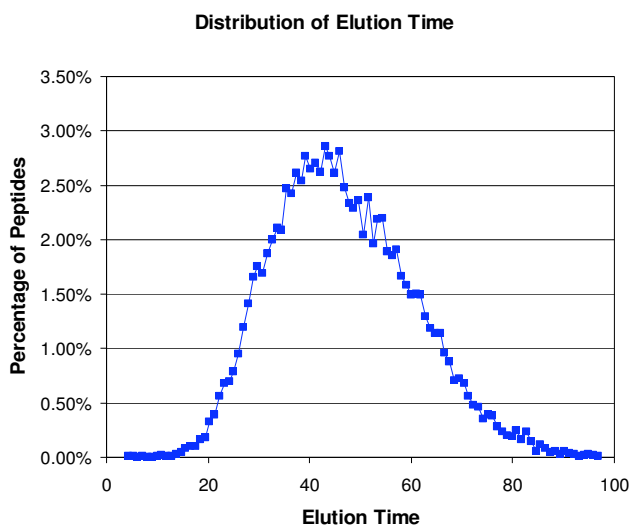


Figure 2: Distribution of the peptide length of the data set.

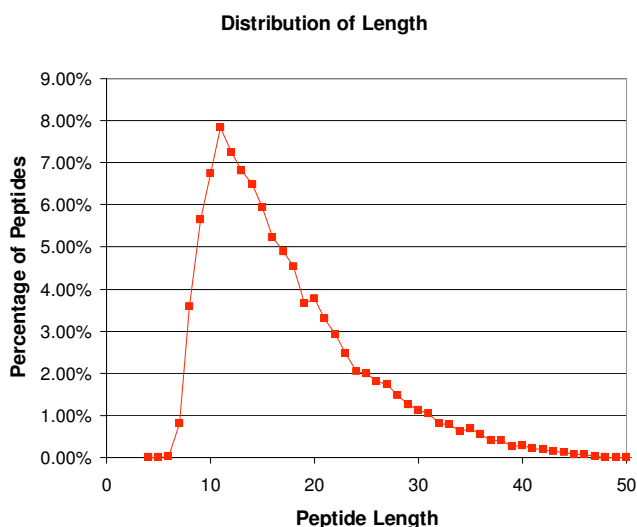
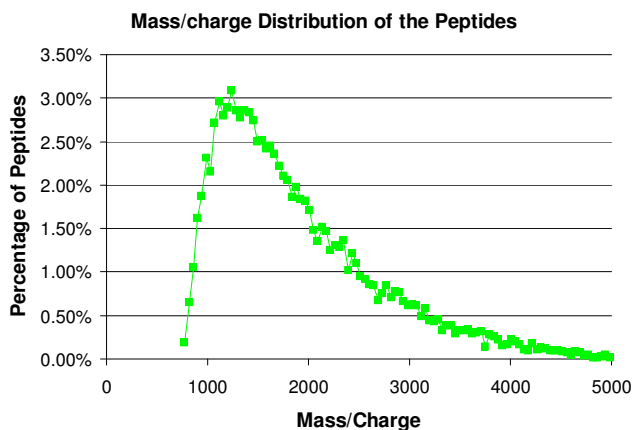


Figure 3: Distribution of mass to charge ratio of the data set.



III. MODEL

There are several chemical properties that affect the elution time of a peptide and we have considered several properties including mass to charge ratio, hydrophobicity, peptide length, secondary structure of the peptide such as helix, sheet and coil, polarity, charge distribution, aromatic property. The correlation coefficient between hydrophobicity and the elution time is 0.72 and their scattered plot is shown in Figure 4. We also examined correlation between the elution time and other parameters. The length of the peptide has a weaker correlation, 0.6, with the elution time and their scattered plot is given in Figure 5.

Figure 4: Relationship between elution time and hydrophobicity of peptide ($\rho = 0.72$)

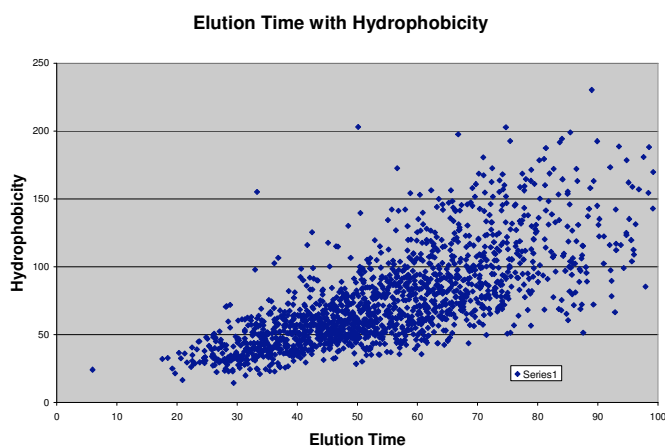
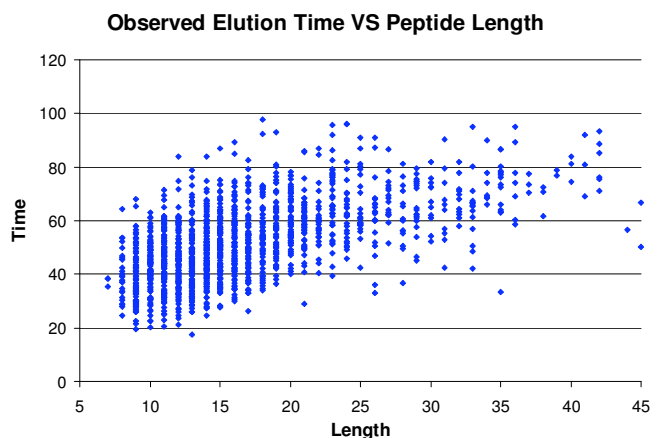


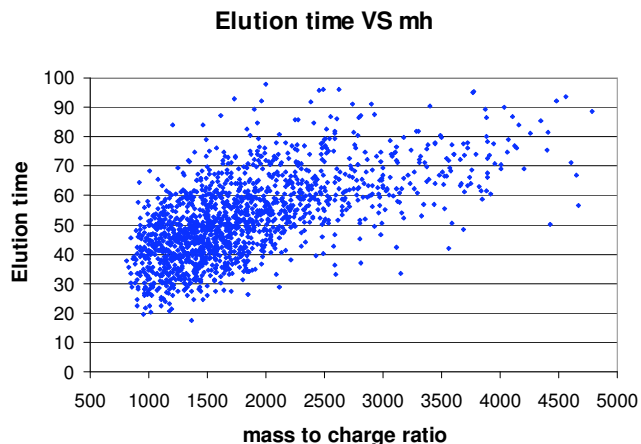
Figure 5: Relationship between the elution time and peptide length ($\rho = 0.6$)



The mass to charge ratio is closely related to the length of the peptide ($\rho = 0.987$), and hence we examined the relationship between the observed elution time with mass to charge ratio. They correlate with coefficient 0.626 and the scatted plot of the relationship is shown in Figure 6.

Figure 6: Relationship between Elution time and mass to

charge ratio ($\rho = 0.626$)



A peptide is a sequence of amino acids. Each amino acid is associated with some degree of affinity to hydrophobicity and it is quantified by a scale as has been described by Kyte and Doolittle [7] in Journal of Molecular Biology. For example, Isoleucine has the highest affinity (9) to hydrophobicity while Arginine has the lowest affinity (0). For a given peptide the summation of the hydrophobicity of all the amino acid will represent the hydrophobicity of the peptide.

We are using an approximate model for inferring secondary structure of a peptide. For the purpose of modeling, we do not have to predict the secondary structure, but represent the affinity of the peptide to each one of the secondary structures namely helices, sheets and coils.

We have examined how elution time is related to predicted affinity values of the following secondary structures of a peptide: helices, sheets and coils. The correlation coefficients of elution time with other parameters are shown in table 1.

Table 1: Correlation coefficient

Affinity to secondary structure	Elution Time
Helices	0.66
Sheets	0.68
Coils	0.56

As we have illustrated previously, the following parameters of a peptide influence the predictability of elution time with varying degree as evident by the correlation coefficient of the observed elution time with each of the parameters: mass to charge ratio, length of the peptide, hydrophobicity, and affinity to secondary structures.

We will describe about a model based on these factors to predict elution time. We provide a pseudo code in Table 2 for building a model from a given peptide.

Table 2: Pseudo code for creating a model of a peptide

```

Procedure build_model{
  Sequences ← set of peptide sequences
  For each seq in sequences do{
    Length ← get_length(seq)
    Freq_dist ← get_frequency_dist(seq)
    Hydrophobicity ← get_hydrophobicity(seq)
    Hydrophobic_dist ← get_hydrophobic_dist(seq)
    Propensity_helix ← get_helix_prop(seq)
    Propensity_helix_dist ← get_helix_prop_dist(seq)
    Propensity_sheet ← get_sheet_prop(seq)
    Propensity_sheet_dist ← get_sheet_prop_dist(seq)
    Propensity_coil ← get_coil_prop(seq)
    Propensity_coil_dist ← get_coil_prop_dist(seq)
    Charge_dist ← get_polarity_dist(seq)
    Aromatic_dist ← get_aromatic_dist(seq)
    Aromatic_dist ← get_aromatic_dist(seq)
    Hydrophilicity ← get_hydrophilicity(seq)
    Hydrophilic_dist ← get_hydrophilic_dist(seq)
  }
}

```

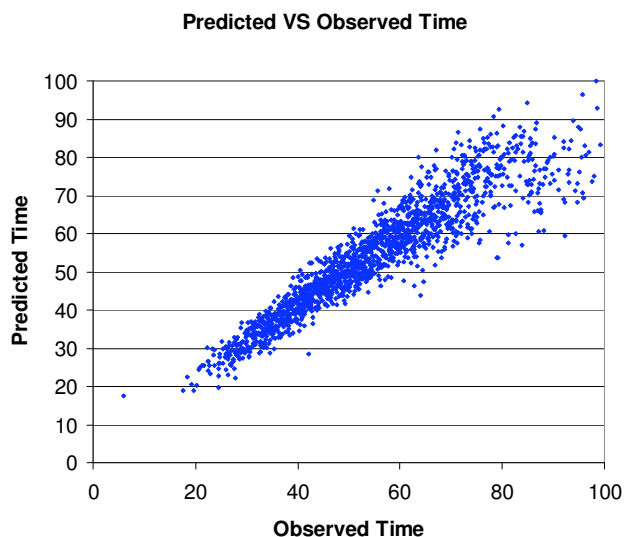
To obtain frequency distribution of a sequence, the number of each amino acid count is divided by the length of the sequence. The hydrophobicity of a sequence is obtained by summing up the hydrophobicity associated with each amino acid in the sequence. Further we will take the virtual center of gravity of a feature in the sequence by taking the summation of first moment of feature divided by the summation of the feature. In this manner we have created the model for each of the features that have some influence on the elution time prediction.

IV. MACHINE LEARNING

We are using support vector regression (SVR) of LibSVM to train and predict the elution time with the peptide model. The model is first scaled. By randomly sampling the scaled model, we have created 10 samples that are used for 10 fold cross validation. Nine samples are joined together and being trained with SVR and the model is applied to predict the elution time of the other left out sample, the test set. We have combined the 10 different predictions and compared with the observed time. The predicted elution time of the model correlates with the observed time with the coefficient of 0.93 and the scattered plot of the comparisons is shown in Figure 7.

Alternatively, the prediction accuracy is also measured with respect the proximity of the predicted time with the observed time. On the average, over 87% of the prediction falls within 5% of the known elution time. Overall, the model performs quite well in predicting the elution time.

Figure 7: Predicted time VS observed elution time ($\rho = 0.93$)



V. SUMMARY AND CONCLUSION

We have developed a peptide model using composition and chemical properties of a peptide for predicting its elution time. This model was trained with regressive support vector machine from LibSVM package. The results we have obtained is very encouraging and is comparable with other prediction methods of elution time, but it is not possible to have a direct comparison with other systems since we do not have any access to their system or the test data.

The support vector machine was very sensitive to parameter selection and by choosing appropriate combination of parameters; the performance of the prediction can be further improved. In the model we have not taken any parameters relevant to experiment set up or to any chemicals that are used for digestive purpose. The model can be also improved by capturing such information. Further, we would like to take the entire data set with different test runs and repeat the study and examine the accuracy of prediction and the robustness of the model.

ACKNOWLEDGMENT

The author would like to acknowledge the opportunity and the support from the Pacific Northwestern National Laboratory, particularly Dr. Joshua Adkins and his group members.

REFERENCES

- [1] B. J. Cargile, J. L. Bundy, and J. L. Stephenson, Jr., "Potential for false positive identifications from large databases through tandem mass spectrometry," *J Proteome Res*, vol. 3, pp. 1082-5, 2004.
- [2] K. Cottingham, "Name that peptide," *Anal Chem*, vol. 76, pp. 94A-97A, 2004.

- [3] J. L. Meek, "Prediction of peptide retention times in high-pressure liquid chromatography on the basis of amino acid composition," *Proc Natl Acad Sci U S A*, vol. 77, pp. 1632-6, 1980.
- [4] M. C. Wilce, M. I. Aguilar, and M. T. Hearn, "High-performance liquid chromatography of amino acids, peptides and proteins. CXXII. Application of experimentally derived retention coefficients to the prediction of peptide retention times: studies with myohemerythrin," *J Chromatogr*, vol. 632, pp. 11-8, 1993.
- [5] K. Petritis, 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, "Use of artificial neural networks for the accurate prediction of peptide liquid chromatography elution times in proteome analyses," *Anal Chem*, vol. 75, pp. 1039-48, 2003.
- [6] K. Petritis, L. J. 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. Camp, 2nd, and R. D. Smith, "Improved peptide elution time prediction for reversed-phase liquid chromatography-MS by incorporating peptide sequence information," *Anal Chem*, vol. 78, pp. 5026-39, 2006.
- [7] J. Kyte and R. F. Doolittle, "A simple method for displaying the hydrophatic character of a protein," *J Mol Biol*, vol. 157, pp. 105-32, 1982.