Confidence Estimation for Peptide Retention-Time Prediction

Corresponding Author [1,*], Co-Author [2] and Co-Author [2*]

[1]Department of XXXXXXX, Address XXXX etc.
[2]Department of XXXXXXXX, Address XXXX etc.

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Associate Editor: XXXXXXX

ABSTRACT

Motivation: Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text.

Results: Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text

Availability: Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text

Contact: name@bio.com

1 INTRODUCTION

The task of retention time prediction, focuses on determining the retention time of a peptide given its amino acid sequence. Accurate peptide retention-time prediction in protein mass spectrometry can highly increase the efficiency of peptides spectrum matches in data-independent acquisition (DIA). Similar to many other machine learning methods, this prediction is done by training a model on a training set and using it to evaluate the unseen peptides. In previous studies, different models such as artificial neural networks (ANN) **?** and support vector regression **?** (SVR) have been applied to this task. These models have been originally designed to solve classifications tasks and they show certain limitations when they are applied to regression tasks.

When using machine learning to predict an entity, we are always interested in the confidence of this prediction. Ideally, this confidence to reflect certain properties of the observed data with respect to the trained model. As an example, this property can reflect how close is the observation to samples that the model is trained on. Such a confidence measure can be very useful at the evaluation stage, due to the fact that most machine learning methods can only generalize near the data points they have observed during the training.

2 APPROACH

The task of retention time prediction, focuses on determining the retention time of a peptide given its amino acid sequence. To obtain a robust model for solving this task, one needs to address the following problems. Most machine learning frameworks require the inputs to come from a vector space. To map the peptide into a vector space by extracting biologically meaningful features **?**, collecting general statistical entities collected from the sequences **?** or using different kernels such as spectrum kernel **?** or string kernel **?**. Once the feature vectors are calculated, a machine learning framework should be selected for solving the prediction problem. The choice this method can highly affect the quality of the prediction. For example in **?**, the authors have chosen Support Vector Regression (SVR) framework **?** to solve the regression problem.

In this paper, we throughly analyze the retention time prediction problem from the machine learning perspective. First, we will look at different methods that can be used for mapping the peptide sequences into vector spaces. In this process, our aim is to determine the pros and cons of different feature extraction techniques in association with large data analysis. Second, we will look at Gaussian Process (GP) **?** as more sophisticated framework for solving the prediction task and compare its performance with the widely used SVR framework. We will also analyze how varying the size of the training set can affect the performance of both models.

3 METHODS

4 RESULTS

- GPs are performing better than epsilon-SVRs (and RVMs) * Plot of deltaRT95 as a function of training set size

-The Elude features give better performance (, but take longer time to calculate) * Compare with Elude-RBF, BOW-RBF (1,2 and 3-mers), and spectrum kernel * Plot of deltaRT95 as a function of training set size

- GPs are predicting confidence * Plot a actual obs-pred RT error as a function of stdv.

- Our Method ... outperforms the state-of-the-art, (Elude and SSRC) * Plot observed vs predicted RT in colormaps in 3 different subplots * Alternatively (and probably better), plot 3 histograms of obs-pred RT for the three methods, and overlay them using transparent colors (e.g. set their alpha-value)

- Our method can predict RT of PTMs * Plot histogram of error (obs-pred RT) for a set of phospho peptides.

- GP confidence, observed vs predicted * Compare the test and train set

[*]to whom correspondence should be addressed

Table 1. This is table caption

head1	head2	head3	head4
row1	row1	row1	row1
row2	row2	row2	row2
row3	row3	row3	row3
row4	row4	row4	row4

This is a footnote

Fig. 1. Caption, caption.

Fig. 2. Caption, caption.

- GP confidence, Find a list of peptides and sort them according to their similarity to the training set * Plot their confidence

5 DISCUSSION

6 CONCLUSION

1. this is item, use enumerate
2. this is item, use enumerate
3. this is item, use enumerate

Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text. Figure 2 shows that the above method Text Text Text Text Text Text Text Text Text Text Text Text. Bofelli *et al.*, 2000 might want to know about text text text text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text. Figure 2 shows that the above method Text Text Text Text Text Text Text Text Text Text Text Text. Bofelli *et al.*, 2000 might want to know about text text text text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text.

Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text. Figure 2 shows that the above method Text Text Text Text

ACKNOWLEDGEMENT

Text Text Text Text Text Text Text Text. Bofelli *et al.*, 2000 might want to know about text text text text

Funding: Text Text Text Text Text Text Text Text.

REFERENCES

Bofelli,F., Name2, Name3 (2003) Article title, *Journal Name*, **199**, 133-154.

Bag,M., Name2, Name3 (2001) Article title, *Journal Name*, **99**, 33-54.

Yoo,M.S. *et al.* (2003) Oxidative stress regulated genes in nigral dopaminergic neurnol cell: correlation with the known pathology in Parkinson's disease. *Brain Res. Mol. Brain Res.*, **110**(Suppl. 1), 76–84.

Lehmann,E.L. (1986) Chapter title. *Book Title*. Vol. 1, 2nd edn. Springer-Verlag, New York.

Crenshaw, B.,III, and Jones, W.B.,Jr (2003) The future of clinical cancer management: one tumor, one chip. *Bioinformatics*, doi:10.1093/bioinformatics/btn000.

Auhtor,A.B. *et al.* (2000) Chapter title. In Smith, A.C. (ed.), *Book Title*, 2nd edn. Publisher, Location, Vol. 1, pp. ???–???.

Bardet, G. (1920) Sur un syndrome d'obesite infantile avec polydactylie et retinite pigmentaire (contribution a l'etude des formes cliniques de l'obesite hypophysaire). PhD Thesis, name of institution, Paris, France.