A machine learning approach for the prediction of DNA and peptide HPLC retention times (extended abstract)

Marc Sturm[1], Oliver Kohlbacher[1], Sascha Quinten[2] and Christian G. Huber[2]

[1] Eberhard Karls University Tübingen, Simulation of Biological Systems
72076, Tübingen, Sand 14, Germany
{sturm|kohlbacher}@informatik.uni-tuebingen.de
[2] Saarland University, Instrumental Analysis and Bioanalysis
66041, Saarbrücken, Postfach 15 11 50, Germany
{sascha.quinten|c.huber}@mx.uni-saarland.de

Abstract. Here we present a method for prediction of HPLC retention times based on support vector regression. In contrast to existing prediction methods for DNA, our method takes the secondary structure of DNA into account. The method is also well suited for retention time prediction of peptides.

Keywords. high performance liquid chromatography, mass spectrometry, retention time, prediction, peptide, DNA, support vector regression

1 Introduction

High performance liquid chromatography (HPLC) has become one of the most efficient methods for the separation of biomolecules. It is an important tool in DNA purification after synthesis as well as DNA quantification. In both cases the separability of different oligonucleotides is essential. The prediction of oligonucleotide retention times prior to the experiment may detect superimposed nucleotides and thereby help to avoid futile experiments. In 2002 Gilar et al. [1] proposed a simple mathematical model for the prediction of DNA retention times, that reliably works at high temperatures only (at least 70°C). To cover a wider temperature rang we incorporated DNA secondary structure information in addition to base composition and length. We used support vector regression (SVR) [2] for the model generation and retention time prediction.

A similar problem arises in shotgun proteomics. Here HPLC coupled to a mass spectrometer (MS) is used to analyze complex peptide mixtures (thousands of peptides). Predicting peptide retention times can be used to validate tandem-MS peptide identifications made by search engines like SEQUEST [3]. Recently several methods including multiple linear regression [4] and artificial neural networks [5] were proposed, but SVR has not been used so far.

Dagstuhl Seminar Proceedings 05471
Computational Proteomics
http://drops.dagstuhl.de/opus/volltexte/2006/548

2 Materials and Methods

2.1 DNA dataset

Our dataset consists of 72 nucleotides of length 15 up to 48 bases. To study the influence of secondary structure on the retention time at different temperatures, the dataset consists of nucleotides that contain little to no secondary structure and others where the whole sequence forms one hairpin. Fig. 1 shows the fraction of bases involved in a secondary structure at different temperatures. Of

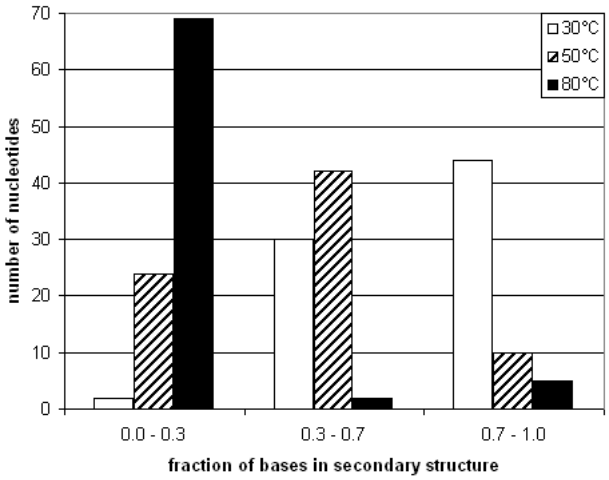

Fig. 1. Fraction of secondary structure at different temperatures.

these 72 nucleotides the retention time was determined at 30, 50 and 80°C on a monolythic PS-DVB column. The homo-oligonucleotides T_{14} and T_{26} were used as internal standard in order to normalize the retention times between different measurements.

DNA secondary structure prediction The secondary structure of the DNA at different temperatures was predicted using the RNAFold tool of the Vienna RNA Package [6].

2.2 Peptides dataset

For the peptide retention time prediction, a training dataset of 1835 peptides and corresponding retention times was generated from a large amount of SEQUEST [3] MS/MS peptide identifications. Only the high quality identifications were included in our dataset (criteria are described in [5]). The retention times were normalized to the interval [0,1] relative to the gradient length as the dataset contained spectra recorded with several different HPLC gradients.

2.3 Support vector regression

SVR is a machine learning technique that uses a training dataset to derive a model for the prediction of some property of the data points. In this case the sought-for property is the retention time. Each data point in the training set consists of a vector of numbers, the so called features. Features for the retention time prediction might be overall length of the sequence, composition of the sequence, etc. The mathematical details of SVR have been described by Vapnik [2].
In this study the libSVM implementation [7] was used with the Radial Basis Function kernel. The kernel parameters C and g have been optimized using a grid search and 3-fold cross validation.

3 Results and discussion

3.1 DNA

For the prediction of DNA retention times the length, the base composition and the fraction of bases in secondary structure were used as features for the SVR.

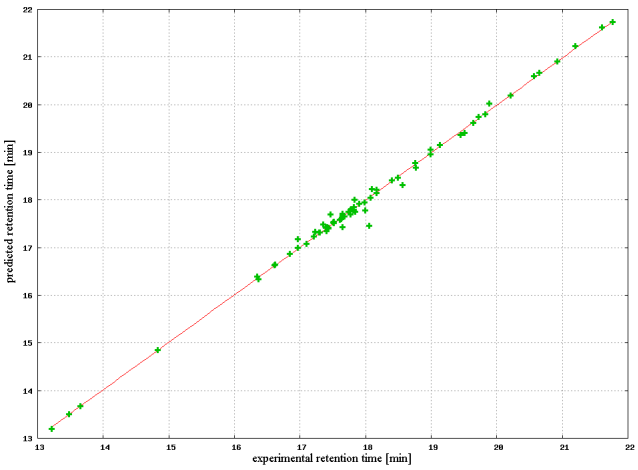

Fig. 2. Correlation of experimental and predicted DNA retention times.

Fig. 2 shows a comparison of experimental and predicted retention times for 80°C. This figure shows only the model accuracy as the model was trained on the same data it had to predict afterwards. Due to over-fitting of the model the real prediction performance might be not as good as shown here. However, the squared correlation coefficient in 3-fold cross-validation (Q^2) is higher than 0.95 for all temperatures as Table 1 shows.
The comparison of our model to the Gilar model reveals the shortcomings of the

Table 1. Comparison of prediction methods for DNA retention times.

Temperature [°C]	SVR[Q^2]	Gilar model[Q^2]
80	0.974	0.970
50	0.962	0.727
30	0.951	0.538

latter at lower temperatures. Both models perform very well for 80°C. However the performance of the Gilar model dramatically drops with the temperature, while our model performs well for all temperatures.
These results show that the secondary structure has a large effect on DNA retention time. At high temperatures no secondary structure is present, so the model performance is similar to the Gilar model.

3.2 Peptides

For peptide retention time prediction only the length and amino acid composition were used as features for the SVR. In Fig. 3 one can see that our model achieved a good correlation between experimental and predicted retention times. The squared correlation coefficient is 0.959. However several outliers, especially

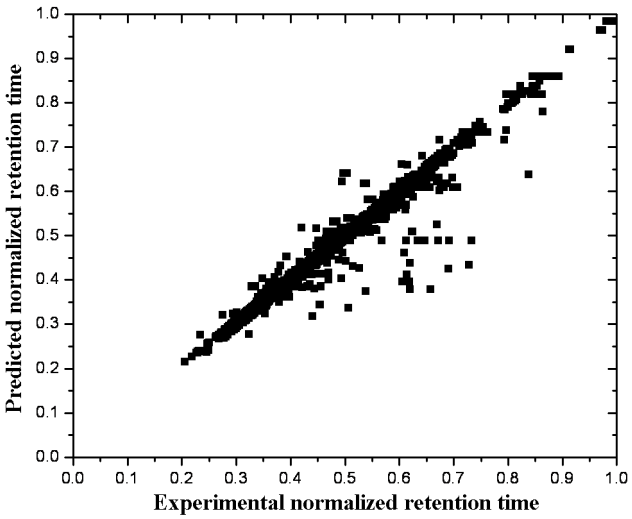

Fig. 3. Correlation of experimental and predicted peptide retention times.

below the diagonal, are present. Many of these outliers have the same amino acid sequence but very different experimental retention times, which is probably caused by incorrect SEQUEST identifications.

The comparison of our model with the artificial neural networks approach by

Table 2. Comparison of the peptide retention time models.

method	prediction MSE	model MSE
SVR	2.5%	0.07%
ANN	-	4.8%

Petritis et al. [5] is difficult as two completely different datasets were used. Thus it is not clear in how far the results are comparable at all. Looking at the mere numbers our model seems to work a bit better, as the mean squared error (MSE) of the prediction is lower than the MSE of the Petritis model (See Table 2). For a more accurate evaluation of the model performance better datasets without contradicting data have to be created. A second interesting point for the future would be to improve the model by adding physico-chemical properties such as hydrophobicity and charge distribution.

References

1. Gilar, M., Fountain, K., Budman, Y., Neue, U., Yardley, K., Rainville, P., Russell, R.n., Gebler, J.: Ion-pair reversed-phase high-performance liquid chromatography analysis of oligonucleotides: retention prediction. J Chromatogr A. **958** (2002) 167–82
2. Vapnik, V.: The Nature of Statistical Learning Theory. Wiley, New York, USA (1999)
3. Sadygov, R.G., Yates, J.R.: A hypergeometric probability model for protein identification and validation using tandem mass spectral data and protein sequence databases. Anal Chem (2003) 3792–8
4. Baczek, T., Wiczling, P., Marszall, M., Heyden, Y.V., Kaliszan, R.: Prediction of peptide retention at different hplc conditions from multiple linear regression models. J Proteome Res (2004) 555–63
5. Petritis, K., Kangas, L.J., Ferguson, P.L., Anderson, G.A., Pasa-Tolić, L., Lipton, M.S., Auberry, K.J., Strittmatter, E.F., Shen, Y., Zhao, R., Smith, R.D.: Use of artificial neural networks for the accurate prediction of peptide liquid chromatography elution times in proteome analyses. Anal Chem (2003) 1039–48
6. Hofacker, I., Fontana, W., Stadler, P., Bonhoeffer, L., Tacker, M., Schuster, P.: Fast folding and comparison of rna secondary structures. Monatsh.Chem. **125** (1994) 167–188
7. Chang, C.C., Lin, C.J.: LIBSVM: a library for support vector machines. (2001) Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.