Bioinformatics

DOI: 10.1002/anie.200602561

Structure–Activity Relationships in Chromatography: Retention Prediction of Oligonucleotides with Support Vector Regression**

*Oliver Kohlbacher, Sascha Quinten, Marc Sturm, Bettina M. Mayr, and Christian G. Huber**

The prediction of molecular properties for a given molecular structure is of considerable interest for many physicochemical and biochemical processes.[1] Efforts are particularly being made to theoretically predict retention times in the area of chromatographic separations that are based on molecular interactions between the molecules partitioning in separating systems comprised of two phases.[2,3]

Linear free-enthalpy relationships model chromatographic retention as a sum of individual energy contributions (dispersion, dipole–dipole, π–π, proton donor–acceptor interactions, etc.).[4] However, owing to their complex molecular structures, as in the case of biopolymers such as peptides or nucleic acids, use is frequently made of the addition of empirically calculated retention contributions of the individual amino acids or nucleotides that are then corrected by terms that take account of the total structure of the molecule.[5,6] However, for more complex molecules, this prediction model is imprecise and the relevant descriptors are very complex to determine. Sequence information (apart from the total composition) and information on secondary structures are not considered in these models.

Models have been developed for peptides that do not derive retention from the properties of the molecular building blocks, but learn them from data sets obtained with test analyses of known structures.[7] The retention data of approximately 7000 peptides have been used, for example, to train an artificial neural network (ANN) for the prediction of peptide retention times from peptide sequences with an accuracy of 3–10%.[8] Other methods from the field of statistical learning, for example, support vector machines (SVMs), may be used for regression problems. In addition to the advantage of leading exactly to a globally optimal solution (unlike ANNs), support vector approaches have also proved themselves for practical use with chemical problems.[9,10]

A model based on the determination and addition of the retention contributions of the nucleotides has been developed to predict the retention of oligonucleotides in ion-pair reversed-phase chromatography (IP-RPC).[11] This model provided satisfactory results at relatively high separating temperatures (60 °C) for cases in which secondary structures are less pronounced, whereas at lower temperatures, the influence of hairpin or partial double strands led to a poor correlation between the prediction and the experiment (own measurement results).

Our model for the retention of oligonucleotides in IP-RPC, even at low temperatures, is based on ν-support vector regression (SVR) as proposed by Schölkopf et al.[12] This method determines a model for a given data set that at the same time minimizes the model error and the model complexity. The training of this model is accomplished with a low number of 50–100 oligonucleotides. A test data set was created by the measurement of the retention times of 72 oligonucleotides. To record the influence of the sequence on the retention, 41 of the oligonucleotide sequences were generated by variation of a sequence of a 24 mer (GTA CTC AGT GTA GCC CAG GAT GCC). To take into account other possible secondary structures, four further sequences were selected that form stable hairpin structures even at higher temperatures. The remaining sequences were finally selected so that they covered a length range of 15–48 nucleotides.

Quantitative structure–property relationships (QSPR) code the input structures in the form of characteristic vectors. Relevant characteristics for the property to be predicted are,

[*] Dipl.-Chem. S. Quinten, Dr. B. M. Mayr, Prof. Dr. C. G. Huber
Fachbereich Chemie
Instrumentelle Analytik und Bioanalytik
Universität des Saarlandes
Gebäude B2.2, 66123 Saarbrücken (Germany)
Fax: (+49) 681-302-2433
E-mail: christian.huber@mx.uni-saarland.de
Prof. Dr. O. Kohlbacher, Dipl.-Inf. M. Sturm
Abteilung Simulation biologischer Systeme
Eberhard-Karls-Universität Tübingen
Sand 14, 72076 Tübingen (Germany)

[**] We thank LC Packings, Amsterdam, for the provision of the capillary HPLC system.

for example, length, sequence, and secondary structure of the oligonucleotides under investigation. These characteristics are translated into numerical values that flow directly into the model. The sequence and structure of the oligonucleotide are described by a vector $\boldsymbol{x}_i$ at each measured retention time y_i. Characteristic-value pairs ($\boldsymbol{x}_i$, y_i) serve to determine an optimal function [Eq. (1)].

$$f(\boldsymbol{x}) = \boldsymbol{w} \cdot \boldsymbol{x} + b \text{ with } \boldsymbol{w}, \boldsymbol{x} \in \Re^n,\ b \in \Re \quad (1)$$

This function allows the prediction of retention times $y = f(\boldsymbol{x})$ for an arbitrary characteristic vector $\boldsymbol{x}$, that is, for an arbitrary sequence. We refer to comprehensive literature for mathematical details.[13,14]

In our model, the characteristic vector comprises eleven values that are derived from the sequence and secondary structure. Five of the values describe the sequence (length and relative proportion of the four bases), whereas the remaining six values represent the melting curve of the secondary structure. For this purpose, the secondary structure was predicted with Vienna RNA Package Version 1.4[15] for the temperatures 30, 40, 50, 60, 70, and 80 °C, and the percentage fraction of the bases in the paired regions was then calculated. The model showed the total length, the fraction of the different bases, and the fraction of paired bases to be the most important characteristics of an oligonucleotide. Other characteristic vectors that contained, for example, the base sequence, base stacking, or other codes containing secondary-structure information, did not produce better results.

Training of the SVR model was carried out by dividing the test data set into three parts (each of 24 data points). Two thirds of the test data were used for training the SVR model, and the remaining third was used for validation of the prediction. Triple cross-validation was carried out throughout. In each case, the values R^2 und Q^2 represent the correlation of the SVR model with the training data set and the correlation of the predicted retention times with the experimental times, respectively.

Figure 1 shows the results for the triple cross-validated model at 30 and 80 °C. The low scattering of the data points and the absence of significant outliers confirm the capabilities of the SVR model for the prediction of retention over a large temperature range and the total length region of 15–50 nucleotides. Unlike other models, the retention of hairpin structures was also predicted with the aid of SVR, with good agreement shown at both temperatures.

Examples of measured and simulated chromatograms are reproduced in Figure 2. In this case 61 of the 72 measured data points were used for training the model to simulate the retention times of the remaining eleven oligonucleotides. Table 1 confirms that all retention times were predicted with a maximum deviation of 4%, but with most less than 2%. The retention of hairpin structures that are stable at elevated temperature was also correctly reproduced by our model (Figure 2e). It is also interesting that the retention measured for the relatively long 39-mer oligonucleotide also agreed very well with the model even though there were few training data points for this length region (Figure 2d)

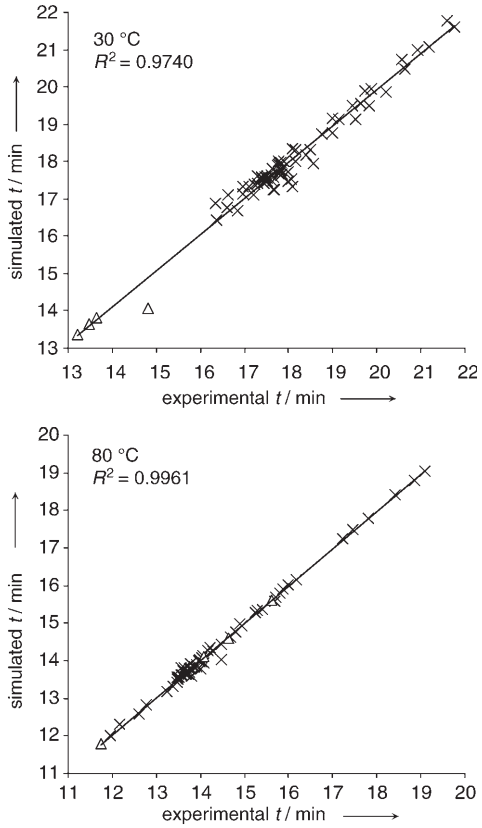

Figure 1. Retention prediction at 30 and 80 °C; R^2 values correspond to the squared correlation between the model and the experimental retention times. Hairpin structures are indicated by triangles.

The application of our method to a real oligonucleotide mixture is shown in Figure 3. It shows the mixture of

Table 1: Comparison of measured and predicted retention times of oligonucleotides at 30 and 80 °C.[a]

Sequence	Retention at 30 °C [min]			Retention at 80 °C [min]		
	measured	calcd	% error	measured	calcd	% error
GTG CTC AGT GTA ACC CAG GAT GCC	17.64	17.76	−0.64	13.99	13.93	0.42
GTG CTC AGT **A**TA **G**CC CAG GAT GCC	17.50	17.59	−0.51	14.00	13.78	1.59
ATG CTC AGT GTA **G**CC CAG GAT GCC	17.77	17.64	0.73	14.08	13.87	1.49
CTG CTC AGT GTA **G**CC CAG GAT GCC	17.23	17.49	−1.56	13.47	13.59	−0.83
GTG CTC AGT GTA **G**CC CAG GAT GC**G**	17.40	17.52	−0.69	13.67	13.58	0.64
GTG CTC AGT GTA **G**CC CAG GAT GC**A**	17.82	17.58	1.37	13.99	13.88	0.83
GTG CTC AGT GTA **G**CC CAG **A**AT GCC	17.29	17.65	−2.08	13.62	13.81	−1.39
GTG CTC AGT GTA **G**CC CAG GAT **A**CC	17.46	17.65	−1.06	13.85	13.81	0.34
GTG CTC AGT GTA **G**CC CAG GAT G**A**C	17.61	17.63	−0.12	13.88	13.82	0.46
GAG AGA GAG AGA TCT CTC TCT CTC	13.22	13.76	−4.11	14.08	14.26	−1.30
GTG CTC AGT GTA ACC CAG TTT TTT GAT GCC GTA GAT CAT	20.93	20.98	−0.28	17.80	17.76	0.24
	Q^2:	0.989			0.987	

[a] Bold letters represent modifications with respect to the first 24-mer sequence.

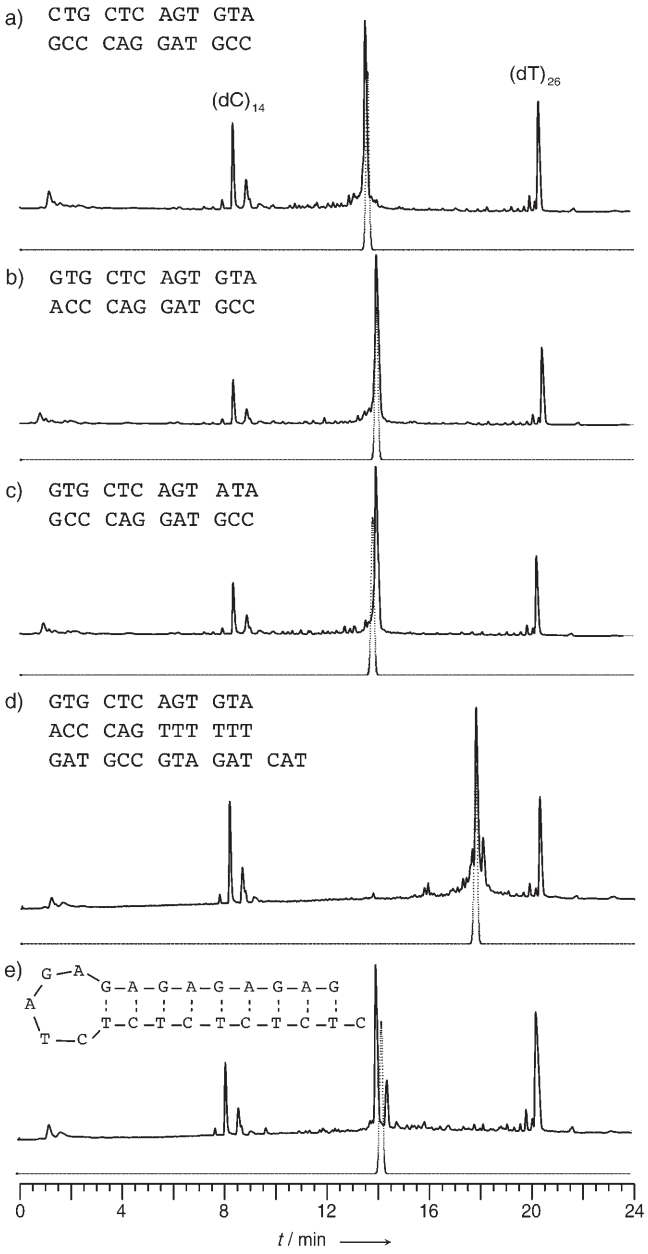

Figure 2. Comparison of the measured and simulated chromatograms of the oligonucleotides at 80 °C. The experimentally determined, mean peak standard deviation of 0.0575 min and a Gauss function were used for the peak profiles in the simulation.

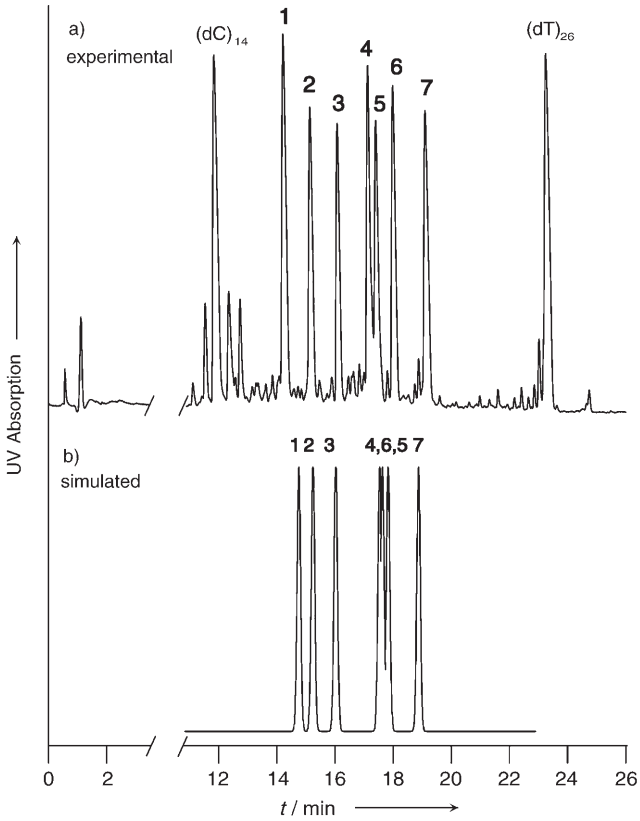

Figure 3. Prediction of a separation of seven oligonucleotide primers with lengths of 14 to 18 nucleotides at 50 °C. **1** = GTA GAG GTA GGT TGG, **2** = GAA GAG TTT TTG GA, **3** = GGG TTA ATT TGA GGT, **4** = TTT AGT TAG AAA AAG TT, **5** = GTT TAA ATA GGA AAT TT, **6** = GTT GGG ATT TTT GTA TTG, **7** = TAA GTT TTT TTT TGT TGT.

oligonucleotide primers that was used for a multiplex amplification by PCR. The strength of the model is demonstrated in that not only were the retention times of the relatively unusual A,T-rich sequences correctly predicted, but also, with the exception of the inversion of the sequence of oligonucleotides **5** and **6**, the retention sequence of the short oligonucleotides had an average and maximum deviation of 1.6% and 3.2%, respectively. The good agreement between the theory and experiment is an indication that the structural parameters integrated into the model describe the interactions between oligonucleotides and the stationary phase in IP-RPC very well.

In summary, a model has been developed that, for the first time, allows prediction of the retention of oligonucleotides with an accuracy of better than 3% over the whole temperature range of 30–80 °C. The method is characterized by training of the model by a restricted set of 50–100 oligonucleotides, and the influence of secondary structure on retention is taken into consideration. With the possibility of the inclusion and evaluation of many arbitrary structural parameters, it is possible to determine quickly and simply which structural properties are pivotal for the interactions of a molecule in a chromatographic separation system. In future work, we would like to exploit the method for the improved prediction of peptide retention times.

Experimental Section

The retention times of the oligonucleotides (in each case 2.5 ng was injected) were determined on a monolithic poly-(styrene/divinylbenzene) column ($60 \times 0.20\ mm^2$ diameter) with the aid of a 30 min (0–16%) gradient of acetonitrile in aqueous triethylamine acetate ($100\ mmol\,L^{-1}$)/ethylenediaminetetraacetic acid solution ($0.5\ mmol\,L^{-1}$), detection at 254 nm, and column temperature of 30–80 °C. The flow rate was held constant at $2.0\ \mu L\,min^{-1}$ with the aid of a HPLC system with an active split (U3000, LC Packings, Amsterdam). The standard deviation of the measured retention times for $(dT)_{18}$ from 20 repeat measurements lay between 1.3 s (0.097% relative standard deviation (RSD)) at 40 °C and 3.1 s (0.28%

RSD) at 70 °C. The RSDs of the measured retention times fell to 0.028 % (40 °C) and 0.072 % (70 °C) by correction with internal standards (in each case 1 ng of $(dC)_{14}$ and $(dT)_{26}$). The standardized net retention time t' was calculated from Equation (2).

$$t' = (t - t_C)\frac{\bar{t}_T - \bar{t}_C}{t_T - t_C} + \bar{t}_C \quad (2)$$

t is the experimental retention time of the oligonucleotide, t_C and t_T are the retention times of $(dC)_{14}$ and $(dT)_{26}$, respectively, and $\bar{t}_C$ and $\bar{t}_T$ are the average retention times of $(dC)_{14}$ and $(dT)_{26}$, respectively, in all experiments. The SVR model was generated with the software package libSVM, Version 2.8[16].

Received: June 27, 2006
Published online: September 29, 2006

Keywords: bioinformatics · liquid chromatography · oligonucleotides · secondary structures · structure–activity relationships

[1] M. H. Abraham, J. Le, W. E. Acree, Jr., P. W. Carr, A. J. Dallas, *Chemosphere* **2001**, *44*, 855 – 863.
[2] R. M. Smith, *Retention and Selectivity in Chromatography*, Elsevier, Amsterdam, **1995**.
[3] P. Jandera, *Adv. Chromatogr.* **2005**, *43*, 1 – 108.
[4] L. C. Tan, P. W. Carr, M. H. Abraham, *J. Chromatogr. A* **1996**, *752*, 1 – 18.
[5] M. Palmblad, M. Ramstrom, K. E. Markides, P. Hakansson, J. Bergquist, *Anal. Chem.* **2002**, *74*, 5826 – 5830.
[6] T. Baczek, P. Wiczling, M. Marszall, Y. Vander Heyden, R. Kaliszan, *J. Proteome Res.* **2005**, *4*, 555 – 563.
[7] R. Kaliszan, T. Baczek, A. Bucinski, B. Buszewski, M. Sztupecka, *J. Sep. Sci.* **2003**, *26*, 271 – 282.
[8] K. Petritis, L. J. Kangas, P. L. Ferguson, G. A. Anderson, L. Pasa-Tolic, M. S. Lipton, K. J. Auberry, E. F. Strittmaier, Y. Shen, R. Zhao, R. D. Smith, *Anal. Chem.* **2003**, *75*, 1039 – 1048.
[9] K. R. Müller, G. Ratsch, S. Sonnenburg, S. Mika, M. Grimm, N. Heinrich, *J. Chem. Inf. Model.* **2005**, *45* 249 – 253.
[10] E. Byvatov, U. Fechner, J. Sadowski, G. Schneider, *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1882 – 1889.
[11] M. Gilar, K. J. Fountain, Y. Budman, U. D. Neue, K. R. Yardley, P. D. Rainville, R. J. Russell II, J. C. Gebler, *J. Chromatogr. A* **2002**, *958*, 167 – 182.
[12] B. Schölkopf, R. Bartlett, A. Smola, R. Williamson in *Proceedings of the 8th International Conference on Artificial Neural Networks* (Ed.: L. Niklasson, M. Boden, T. Ziemke), **1998**, p. 111 – 116.
[13] V. N. Vapnik, *The Nature of Statistical Learning Theory*, Wiley, New York, **1999**.
[14] T. Hastie, R. Tibshirani, J. Friedman, *The Elements of Statistical Learning*, Springer, New York, **2001**.
[15] I. Hofacker, *Vienna RNA Package, RNA Secondary Structure Prediction and Comparison*, http://www.tbi.univie.ac.at/~ivo/RNA (2006).
[16] C.-C. Chang, C.-J. Lin, *LIBSVM: a library for support vector machines*, http://www.csie.ntu.edu.tw/~cjlin/libsvm (2001).

