A statistical learning approach to the modeling of chromatographic retention of oligonucleotides incorporating sequence and secondary structure data

Marc Sturm1*, Sascha Quinten2, Christian G. Huber2 and Oliver Kohlbacher1

1Simulation of Biological Systems, Eberhard Karls University Tübingen and 2Department of Chemistry, Instrumental Analysis and Bioanalysis, Saarland University, Saarbrücken, Germany

ABSTRACT

We propose a new model for predicting the retention time of oligonucleotides. The model is based on support vector regression using features derived from base sequence and predicted secondary structure of oligonucleotides. Because of the secondary structure information, the model is applicable even at relatively low temperatures where the secondary structure is not suppressed by thermal denaturing. This makes the prediction of oligonucleotide retention time for arbitrary temperatures possible, provided that the target temperature lies within the temperature range of the training data.

We describe different possibilities of feature calculation from base sequence and secondary structure, present the results and compare our model to existing models.

INTRODUCTION

The structure of biological macromolecules like polypeptides and nucleic acids comprises linear polymers of a limited number of small building blocks such as amino acids or nucleotides. Even so, because of their large size and a certain degree of flexibility of the polymer chains, biopolymers are usually folded, forming a distinct 3D structure which is generally essential for the biological function of the molecules. Studies of the 3D structure of nucleic acids go back to the seminal work of Watson and Crick, who discovered the DNA double helix. Since then, a large number of 3D structures of single- and double-stranded nucleic acids have been solved by X-ray crystallography and nuclear magnetic resonance spectroscopy, which allow a deep insight into the structure–function relationships of nucleic acids.

In spite of the broad availability of structural data, investigations into the relationship between the 3D structure of nucleic acids and their interaction with stationary phases in chromatographic separation systems have been quite limited (1,2), mainly because of the highly dynamic nature of the process, and the involvement of both a liquid and a solid phase in the phase transfer. Since DNA and RNA molecules are employed in many biochemical applications, such as polymerase chain reaction, genotyping (3), DNA sequencing (4), hybridization assays and gene therapy (5), understanding these phase transfer processes would not only help to improve our selection and optimization of chromatographic separation systems in order to purify and analyze them, but also to gain valuable information about the behavior of nucleic acid molecules in multi-phase systems and at interphases.

Commonly used models describing the behavior of solute molecules at the liquid–solid interface, such as linear free enthalpy relationships (6), are based on the fundamental Hammet equation and knowledge of principal structural descriptors, such as intrinsic molar volume, Hildebrand-solubility parameters, polarizability and proton-donor/acceptor parameters. However, these structural descriptors are difficult and tedious to determine experimentally, especially for biological macromolecules. Other models rely on learning structure–retention relationships from a set of standard analytes. By doing so, Gilar et al. (7) have developed a model for the retention of oligonucleotides in ion-pair reversed-phase chromatography by adding up the retention contributions of the individual nucleotides, which have been determined experimentally from analyzing homo-oligonucleotides. Since the input structural parameters for the model are only oligonucleotide length and base composition, it is only applicable at relatively high temperatures of 60–80°C where secondary structures are usually suppressed because of thermal denaturing. A similar model was used by...
Otvos et al. (8) for peptide nucleic acids. The model was based upon retention contribution of monomer building units, which were studied by Meek (9) and Guo et al. (10).

We propose a new model for retention time prediction which differs in two ways from the Gilar approach. First, we incorporate several additional features of the oligonucleotides into the model, the most important being predicted secondary structure information. The temperature dependency of the secondary structure is captured by adding information similar to a melting curve. The second difference lies in model creation. Support vector regression (SVR) is used instead of simple linear or logarithmic models. SVR can model non-linear relationships while optimizing both the model performance and the model complexity. These improvements lead to a reliable model with a significantly increased prediction performance.

**MATERIALS AND METHODS**

**Chemicals and oligonucleotides**

Acetonitrile (far UV HPLC grade) and acetic acid (analytical reagent grade) were obtained from Riedel-de-Haën (Seelze, Germany). Triethylamine (analytical reagent grade) and ethylenediamine tetraacetate (EDTA, analytical reagent grade) were purchased from Fluka (Buchs, Switzerland). Triethylammonium acetate was prepared by mixing equal amounts of triethylamine and acetic acid. Water was purified by means of a purification system (Purelab Ultra Genetic, Elga, Siershahn, Germany). Oligonucleotides were synthesized by MWG-Biotech AG (Ebersberg, Germany) or Biospring (Frankfurt, Germany). The sequences of the 72 different oligonucleotides were used to create training and evaluation datasets for the retention model are collected in Table S1 of the Supplementary Data.

**Ion-pair reversed-phase high-performance liquid chromatography**

Ion-pair reversed-phase high-performance liquid chromatography (IP-RP-HPLC) was performed with a fully automated capillary/nano HPLC system (Model Ultimate 3000, Dionex, Amsterdam, The Netherlands) equipped with a low-pressure gradient micropump (Model LPG-3600) with a vacuum degasser (Model SRD-3600), an integrated column oven, a microcolumn switching unit and flow-splitting device (Model FLM-3100), a microautosampler (Model WPS-3000) and a UV-detector (Model UVD 3000) with a 3-nL Z-shaped detector cell (Model UVD 3000). The system was controlled by a personal computer with Chromelion Version 6.60 SP2 (Dionex). The 60 × 0.20 mm i.d. poly-(styrene/divinylbenzene) monolithic column was prepared according to the published protocol (11) and is available from Dionex.

**Generation of experimental retention time datasets**

One to five nanograms of oligonucleotides, dissolved in 1 μl water, were injected and chromatographed with a 30-min linear gradient from 0–16% acetonitrile in 100 mmol/l triethylammonium acetate, 0.5 mmol/l EDTA, pH 7.0. The flow rate was adjusted to 2 μl/min. The gradient delay volume of the used LC system was 5.94 μl. Although retention times of oligonucleotides were highly reproducible (0.26% relative SD in absolute retention time from three repetitive injections), (dC)_{14} and (dT)_{26} homo-oligonucleotides were coinjected as internal standards to normalize retention. Normalization was performed using Equation (1) (t and 7 representing the retention time and average retention time, respectively, of the oligonucleotides)

\[ t_{\text{normalized}} = (t_{\text{measured}} - t_{(dC)_{14}}) \times \frac{t_{(dT)_{26}} - t_{(dC)_{14}}}{t_{(dT)_{26}} - t_{(dC)_{14}}} + t_{(dC)_{14}} \]

and resulted in retention time measurement errors of <0.04% (0.6 s). A representative separation of an oligonucleotide including the two standards is illustrated in Figure 1. The 72 different oligonucleotides were chromatographed at column temperatures of 30, 40, 50, 60 and 80°C and the measured normalized retention times are summarized in Table S1 of the Supplementary Data.

**Support vector regression**

A number of methods can be used to derive models from training data. In this study, SVR is used for model generation and predicting retention times. Squared Pearson correlation coefficients were used to evaluate the quality of the models (R²) and the predictions (Q²).

SVR is a machine-learning technique that uses a training dataset to derive a model for the prediction of a quantitative response. Just like in linear regression methods, the training data consists of m pairs of feature vector x and quantitative response y:

\[(x_1, y_1), \ldots, (x_m, y_m) \in \mathbb{R}^p \times \mathbb{R} \]

Features of oligonucleotides for retention time prediction might be: overall length, number of adenines, number of thymines, etc. The quantitative response is the retention time.
For these training data, $\varepsilon$-SVR (12) determines the optimal parameters $w$ and $b$ of the function

$$y = f(x) = (w \cdot x) + b \quad w, x \in \mathbb{R}^n, b \in \mathbb{R}$$

by minimizing the objective function

$$\frac{1}{2}||w||^2 + \frac{C}{m} \sum_{i=1}^{m} |y_i - f(x_i)|_\varepsilon$$

In this equation, the first addend minimizes the model complexity while the second minimizes the $\varepsilon$-insensitive training error — i.e. training errors below a fixed $\varepsilon$ are not penalized. $C$ is a constant which defines the trade-off between these two objectives. In 2001, Schölkopf et al. (13) proposed $\nu$-SVR, a modified version of $\varepsilon$-SVR which minimizes $\varepsilon$ along with the model complexity and the training error. This simplifies the use of SVR, as $\varepsilon$ no longer has to be chosen a priori.

SVR, as so far described here, cannot model non-linear relationships between input features and the quantitative response. Non-linearity is achieved by transforming the input features into a higher-dimensional space using so-called kernel functions. A linear model in the transformed feature space corresponds to a non-linear model in the original feature space. Details about kernel functions can be found in (12).

In this study, the libSVM implementation (http://www.csie.ntu.edu.tw/~cjlin/libsvm) of $\nu$-SVR was used with radial basis function kernel. The kernel parameter gamma and the trade-off $C$ have been optimized using grid search and 3-fold cross-validation.

Oligonucleotide sequence selection

Our dataset consists of retention times measured for 72 oligonucleotides ranging from 15 to 48 bases. As one focus of this study is the influence of secondary structure on the retention time, the dataset consists of oligonucleotides that contain little to no secondary structure and others where nearly all bases form a hairpin. Four sequences that form stable hairpin structures even at elevated temperatures were included. Figure 2 shows the average fraction of paired bases plotted against the measurement temperature.

A second point of interest is the influence of the base sequence on the retention time. That is why the dataset contains two groups of oligonucleotides that have the same overall base composition, but slight differences in the base sequence. Both groups are derived from the 24mer GTGTCAGTGTAGCAGAGTGC. In the first group, one guanine is exchanged for an adenine; in the second, one guanine is exchanged for a cytosine.

Features and models

Secondary structure prediction. All secondary structures used in our models are not experimental but predicted. For this purpose, the tool RNAFold of the Vienna RNA Package (14) version 1.4 was used.

Input features. The selection of input features for performing SVR is a critical step, as the features determine the performance of the prediction. Leaving out essential features or adding unnecessary features both lead to a drop in prediction accuracy. We thus tested several different models, i.e. input feature sets. Each model consists of several model components which group closely related features together. Besides these model components, each model implicitly contains the temperature and the length of the oligonucleotide. The model components can be grouped into composition components, structural components and energy components. Table 1 contains an overview of all components.

| Name       | Features | Description                                           |
|------------|----------|-------------------------------------------------------|
| COUNT      | 4        | Base frequencies (#A, #C, #G and #T in the sequence). |
| CONTACT    | 16       | Dinucleotide frequencies (#CG, #CA, #CT, #CC, ...).  |
| SCONTACT   | 10       | Dinucleotide frequencies, independent of direction (#CG+#CG, #CA+#AC, #CC, ...). |
| PAIRED     | 4        | Fraction of A, C, G and T inside stems.               |
| UNPAIRED   | 4        | Fraction of A, C, G and T outside stems.              |
| STRUCTURE  | 12       | Fraction of A, C, G and T in stems, in loops, or unpaired. |
| MULTISTRUC | 6        | Fraction of bases in stems.                           |
| MULTITW    | 12       | Fraction of bases that are unpaired, and in stems or loops. |
| MULTIDETAIL| 36       | Fraction of bases in stems, and loops. Fraction of unpaired A, C, G and T. |
| SESUM      | 1        | Sum of the stacking energies of adjacent bases.       |
| STACKING   | 2        | Sum of the enthalpic ($\Delta H$) and entropic ($\Delta S$) contributions to the free energy. |

Figure 2. Average fraction of bases in stems and loops for the 72 oligonucleotides. The secondary structures were predicted using RNAFold (Secondary Structure Prediction section).
Sequence components are calculated from the oligonucleotide sequence and describe the base composition or sequence details of the oligonucleotide. COUNT reflects the base composition, while CONTACT and SCONTACT contain dinucleotide frequencies, which contain information on the adjacency of bases. The key idea here is that stacking bases will influence the secondary structure and the interaction with the stationary phase.

Since retention time behavior strongly depends on the secondary structure, sequence-based features alone are suitable only for high temperatures. At lower temperatures, the secondary structure requires the inclusion of structure-based components in the prediction. These components can be subdivided into two groups. The first group considers only the secondary structure of the temperature at which the measurement was performed, whereas the second group contains information on secondary structures for temperatures of 30, 40, 50, 60, 70 and 80°C.

Figure 3 shows the predicted secondary structure of the oligonucleotide GTGCTCAGTGTAGCCCAGGATGGG at 40°C. Examples of features calculated from the predicted structure are listed in Table 2. For the simple structural components, only the measurement temperature is considered. For the multi-temperature structural components, the secondary structure and the resulting features are calculated for different temperatures.

The single temperature components PAIRED and UNPAIRED have four features each reflecting the fraction of A, T, C and G inside stems and outside stems, respectively. The STRUCTURE component contains more detailed information, as it considers the fraction of A, T, C and G inside stems and inside loops and unpaired nucleotides.

The multi-temperature structural components reflect secondary structure information as well. In contrast to the single temperature components, they do not only consider the measurement temperature but a range of temperatures (30, 40, 50, 60, 70 and 80°C). Adding this gradient of secondary structure information for different temperatures provides information similar to a melting curve.

The simplest component in this group is MULTISTRUCT, which, for each of the six temperatures, contains the fraction of all bases that are inside stems. MULTITWO contains the fraction of bases in stems and loops as well as the fraction of unpaired bases. The most detailed representation of the secondary structure is MULTIDETAIL. It contains the fraction of bases in stems, the fraction of bases in loops, and the fractions of A, C, G and T that are unpaired.

Energy components describe the ring stacking effects between adjacent bases. These energies might influence retention time as they have to be overcome in order for the aromatic ring system of the base to interact with the column packing. The EMBOSS suite (15) was the source of the stacking energies of base pairs. The thermodynamic parameters used for the STACKING component were taken from (16).

**Single temperature and combined temperatures models.** The standard approach in predicting retention time is to create temperature-specific models, i.e. each model is based on data for a single temperature. Thus, if predictions for several temperatures are needed, one model has to be created for each temperature.

The model components that consider multiple temperatures allow for a more general approach. The data points of one dataset may differ in temperature, which leads to a model that can predict retention times over the whole temperature range of the training data. The advantage of this approach is that all available data can be integrated into a single, more general model. This reduces the amount of data needed for the individual temperatures.

Combining datasets of all temperatures leads to a dataset with 432 data points (72 data points from six temperatures). To avoid overfitting, we included data from only one randomly chosen temperature for each oligonucleotide. The resulting model was then used to predict the retention times of the remaining measurements.

**RESULTS**

**Model creation**

From the 11 model components presented in Table 1, we created models containing one or more of these components. As the number of models that can be built out of 11 components is huge, we focused on models built out of two or three components, which mostly contain one composition component and one structural component. These models will subsequently be referred to by the names of the components they contain (e.g. ‘count_sesum’ for a model that consists of the COUNT component and the SESUM component).

For each temperature, the models were trained and the results evaluated. Additionally, the models were trained on the combined temperature dataset as described in Single Temperature and Combined Temperatures Models section.
Table 3. Prediction performance ($Q^2$) of selected models in cross-validation

| Model                                      | 30°C | 40°C | 50°C | 60°C | 80°C | ALL | Average |
|---------------------------------------------|------|------|------|------|------|-----|---------|
| count_scontact_multistruct                  | 0.978| 0.978| 0.950| 0.965| 0.975| 0.954| 0.967   |
| count_multistruct_stack                    | 0.956| 0.977| 0.960| 0.957| 0.983| 0.953| 0.964   |
| count_multistruct                          | 0.951| 0.976| 0.962| 0.959| 0.974| 0.954| 0.963   |
| count_multitwo_stack                      | 0.950| 0.975| 0.964| 0.959| 0.985| 0.936| 0.961   |
| scontact_multistruct_stack                 | 0.971| 0.972| 0.953| 0.954| 0.963| 0.954| 0.961   |
| scontact_multitwo                          | 0.972| 0.968| 0.949| 0.957| 0.955| 0.955| 0.959   |
| count_multitwo                            | 0.938| 0.974| 0.964| 0.963| 0.981| 0.934| 0.959   |
| paired_unpaired_stack                     | 0.947| 0.953| 0.956| 0.955| 0.983| 0.934| 0.955   |
| scontact_multitwo_stack                   | 0.958| 0.965| 0.946| 0.950| 0.964| 0.941| 0.954   |
| paired_unpaired                           | 0.942| 0.949| 0.954| 0.962| 0.963| 0.929| 0.950   |
| scontact                                  | 0.912| 0.897| 0.897| 0.952| 0.973| 0.931| 0.927   |
| count_scontact                            | 0.902| 0.866| 0.843| 0.961| 0.988| 0.924| 0.914   |
| count                                     | 0.583| 0.621| 0.762| 0.918| 0.988| 0.873| 0.791   |
| Average                                    | 0.920| 0.929| 0.928| 0.955| 0.975| 0.936|         |

Models are sorted according to average performance.

Prediction accuracy in cross-validation

For each model, the prediction correlation $Q^2$ in 3-fold cross-validation was determined for all temperatures. Table 3 shows the prediction performance for a selection of models:

Among the composition components SCONTACT and CONTACT did not show much difference, so only results for SCONTACT are shown. In the group of single temperature secondary structure components, the combination of PAIRED and UNPAIRED was the only one that performed at a similar level to the multi-temperature structural components, where MULTISTRUCT and MULTITWO were the best components. Out of the energy components, STACKING always performed better than SESUM, so the models containing SESUM are not shown. The results of the omitted models can be found in Table S2 of the Supplementary Data.

Single temperatures datasets. All models show a good prediction correlation on the 80°C dataset but the prediction performance averaged over all models drops with the temperature. The models that are exclusively based on sequence information (‘count’, ‘scountact’ and ‘count_scontact’) cannot compete with those models that use secondary structure information at lower temperatures.

Combined temperatures dataset. The models with multi-temperature structural components outperform all other models on the combined temperature dataset. Out of these components, ‘multistruct’ performs best.

For all further evaluations, we chose the model ‘count_multistruct_stack’. It was preferred over the model ‘count_scontact_multistruct’, which shows a slightly better average performance, because it has far fewer features and thus has less tendency to overfit.

Homology reduction. In order to rule out the possibility that the good prediction performance stems from overfitting caused by sequence homology present in the dataset, three homology-reduced datasets were created.

Table 4. Prediction performance on homology-reduced datasets at 30 and 80°C

| Dataset               | Oligonucleotides | 30°C | 80°C |
|-----------------------|------------------|------|------|
| Unreduced             |                  | 0.957| 0.983|
| Two bases differing   |                  | 0.954| 0.986|
| Three bases differing |                  | 0.957| 0.965|
| Five bases differing  |                  | 0.911| 0.963|

We simply removed oligonucleotides from the original dataset until the remaining sequences differed in at least two, three or five bases, respectively. Table 4 shows that there is almost no drop in the prediction performance for homology-reduced datasets. Only the dataset with oligonucleotides that differ at least in five bases actually performs a little worse than the rest, which is probably caused by the small number of training data points.

Comparison to the Gilar model

In 2002, Gilar et al. (7) proposed a simple mathematical model for predicting the retention time of oligonucleotides. The model takes only the overall length and the base composition of oligonucleotides into account. As Equation (5)

$$rt = \sum_{i \in \{A, T, C, G\}} a_i N_i + b_i N_i \ln(N)$$

where $N_i$ is the number of occurrences of base $i$ in the oligonucleotide and $N$ is the overall length of the oligonucleotide (i.e. $N = N_A + N_T + N_C + N_G$). There are two coefficients $a_i$ and $b_i$ for each base type $i$ which have to be determined experimentally. Therefore, the retention times of homo-oligonucleotides of different lengths were measured for each base type. The individual contributions of each base can then be fitted to the measured times in order to determine $a_i$ and $b_i$. We compared the Gilar model to our model ‘count_multistruct_stack’. The determination of the Gilar model
coefficients for our dataset will be described elsewhere (manuscript in preparation).

The Gilar model fits very well at 80°C but drops to $R^2 = 0.58$ at 30°C (Figure 4). In contrast, the ‘count_multistruct_stack’ model maintains a performance of $R^2 \geq 0.95$ for all temperatures. Figure 5 shows the retention times predicted by the Gilar model at 30 and 80°C. At 80°C, the prediction performance is equal to the performance of our models. However, at 30°C, the hairpin structures (marked with circles) cannot be predicted correctly any more and generally, prediction errors increase significantly (Figure 4).

Influence of secondary structure

In order to demonstrate the effect of adding secondary structure information to a sequence-based model, we compared the model error of the models ‘count’ and ‘count_multistruct_stack’. Table 5 shows the average relative model error of oligonucleotides with a similar fraction of bases involved in their secondary structure. The ‘count_multistruct_stack’ model can compensate for the influence of the secondary structure. However, the relative model error of the ‘count’ model rises up to nearly 38 and 23% for several oligonucleotides at 30 and 50°C, respectively. It is simply impossible for the ‘count’ model to derive a reasonable model from the training data at lower temperatures as, it does not incorporate secondary structure information. At 80°C, the ‘count_multistruct_stack’ model still has lower relative model errors than the ‘count’ model.

The amount of data required for training

For practical use, the number of training data points that are required for a reasonable model is very important. So we determined an average prediction performance for the range of 10–60 training data points. We used the following procedure:

(i) Repeat the following steps 200 times:

(a) Randomly choose 10 data points as a test set.

Table 5. Average relative model error of oligonucleotides with a similar fraction of bases involved in secondary structure

| Temp. | Model                     | Average relative model error |
|-------|---------------------------|------------------------------|
|       | Fraction of bases in secondary structure | 0.0 – 0.2 | 0.2 – 0.4 | 0.4 – 0.6 | 0.6 – 0.8 | 0.8 – 1.0 |
| 30°C  | count                     | 0.98%                       | 1.12%                       | 1.41%                       | 15.47%           | 37.62% |
| 30°C  | count_multistruct_stack   | 0.65%                       | 1.01%                       | 1.21%                       | 2.35%            | 1.19%  |
| 50°C  | count                     | 0.78%                       | 0.80%                       | 1.11%                       | 17.93%           | 22.70% |
| 50°C  | count_multistruct_stack   | 0.38%                       | 0.39%                       | 0.43%                       | 0.27%            | 0.26%  |
| 80°C  | count                     | 0.57%                       | 2.12%                       | na                          | 1.94%            | 0.40%  |
| 80°C  | count_multistruct_stack   | 0.44%                       | 0.34%                       | na                          | 0.28%            | 0.09%  |
DISCUSSION

The methods, experiments and results presented here clearly demonstrate that predicting oligonucleotide retention time can be significantly improved using secondary structure information and SVR. The use of secondary structure information improves the prediction performance, especially at low temperatures and for oligonucleotides that form highly stable secondary structures. The second point of interest in this study was the effect that the base sequence had on the retention time. The fact that our best model contains no explicit sequence information shows that it plays only a subordinate role in this process. Rather, it is the influence of the base sequence on the secondary structure that seems to be the base sequences’ main contribution. Explicitly modeling the base sequence is therefore not necessary. However, a closer examination of the influence of the base sequence remains to be done.

The second performance boost came from the use of SVR, probably because of its ability to model non-linear relations and because it handles outliers better. The essential step, and limiting factor, when working with a SVR model is the selection of the training data. A good prediction performance for oligonucleotides that lie far outside the training feature space is very unlikely. Thus, it is crucial to ensure a broad coverage of feature space, both in terms of secondary structure and base composition.

We tested 11 model components that cover base composition, base sequence, secondary structure and base stacking. Although our model performed very well, there is still room for improvements. New features that model additional chemical properties can be easily added to our model by appending them to the feature vector. We expect that this kind of model extensibility can further help to improve the oligonucleotide retention time prediction model in the future.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

Conflict of interest statement. None declared.

REFERENCES

1. Gelhaus, S., LaCourse, W.R., Hagan, N.A., Amarasinghe, G.K. and Fabris, D. (2003) Rapid purification of RNA secondary structures. *Nucleic Acids Res.*, 31, e135.
2. Huber, C.G. and Berti, G.N. (1996) Detection of partial denaturation in AT-rich DNA fragments by ion-pair reversed-phase chromatography. *Anal. Chem.*, 68, 2959–2965.
3. Braun, A., Little, D.F. and Koster, H. (1997) Detecting CFTR gene mutations by using primer oligo base extension and mass spectrometry. *Clin. Chem.*, 43, 1151–1158.
4. Sanger, F., Nicklen, S. and Coulson, A.R. (1977) DNA sequencing with chain-terminating inhibitors. *Proc. Natl Acad. Sci. USA*, 74, 5463–5467.
5. Bumcrot, D., Manoharan, M., Koteliansky, V. and Sah, D.W.Y. (2006) RNAi therapeutics: a potential new class of pharmaceutical drugs. *Nat. Chem. Biol.*, 2, 711–719.
6. Tan, L.C., Carr, P.W. and Abraham, M.H. (1996) Study of Retention in reversed-phase liquid chromatography using linear solvation energy relationships. I. The Stationary Phase. *J. Chromatogr. A*, 752, 1–18.
7. Gilar, M., Fountain, K.J., Badman, Y., Yeue, U.D., Yardley, K.R., Rainville, P.D., Russell, R.J. and Gebler, J.C. (2002) Ion-pair reversed-phase high-performance liquid chromatography analysis of oligonucleotides: retention prediction. *J. Chromatogr. A*, 958, 167–82.
8. Hoffmann, R., Bril, G., Yoneyama, M. and Otvos, L. (1998) Prediction of retention times of peptide nucleic acids during reversed-phase high-performance liquid chromatography. *J. Chromatogr. A*, 814, 111–119.
9. Meek, J.L. (1980) Prediction of peptide retention times in high-pressure liquid chromatography on the basis of amino acid composition. *Proc. Natl Acad. Sci. USA*, 77, 1632–1636.
10. Guo, D., Mant, C.T., Taneja, A.K., Parker, J.M.R. and Hodges, R.S. (1986) Prediction of peptide retention times in reversed phase HPLC: determination of retention coefficients of amino acid residues of model synthetic peptides. *J. Chromatogr.*, 359, 499–518.
11. Huber, C.G., Premstaller, A. and Oberacher, H. (2001) Method and apparatus for separating polynucleotides using monolithic capillary columns. Patent application US, 09/770410.
12. Vapnik, V.N. (2001) *The Nature of Statistical Learning Theory.* Springer-Verlag New York Inc., New York, USA.
13. Schölkopf, B., Smola, A.J., Williamson, R.C. and Bartlett, P.L. (2000) New support vector algorithms. *Neural Computation*, 12, 1207–1245.

14. Hofacker, I.L., Fontana, W., Stadler, P.F., Bonhoeffer, L.S., Tacker, M. and Schuster, P. (1994) Fast folding and comparison of RNA secondary structures. *Monatsh. Chem.*, 125, 167–188.

15. Rice, P., Longden, I. and Bleasby, A. (2000) EMBOSS: the European Molecular Biology Open Software Suite. *Trends Genet.*, 16, 276–277.

16. Sugimoto, N., Nakano, S., Yoneyama, M. and Honda, K. (1996) Improved thermodynamic parameters and helix initiation factor to predict stability of DNA duplexes. *Nucleic Acids Res.*, 24, 4501–4505.