Bioinformatik

DOI: 10.1002/ange.200602561

Struktur-Aktivitäts-Beziehungen in der Chromatographie: Retentionsvorhersage für Oligonucleotide mit Supportvektorregression**

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

Für viele physikalisch-chemische und biochemische Prozesse ist die Vorhersage molekularer Eigenschaften unter Vorgabe der Molekülstruktur von großem Interesse.[1] Besonders im Bereich der chromatographischen Trennungen, die auf molekularen Wechselwirkungen zwischen den zu trennenden Molekülen und einem aus zwei Phasen bestehenden Trennsystem beruhen, gibt es Bestrebungen, Retentionszeiten theoretisch vorherzusagen.[2,3]

Lineare-Freie-Enthalpie-Beziehungen modellieren die chromatographische Retention als Summe von energetischen Einzelbeiträgen (Dispersions-, Dipol-Dipol-, π-π-, Protonen-Donor-Acceptor-Wechselwirkungen usw.).[4] Bei Biopolymeren wie Peptiden oder Nucleinsäuren greift man wegen der komplexen Molekülstrukturen dagegen oft auf eine Addition empirisch ermittelter Retentionsbeiträge der individuellen Aminosäuren bzw. Nucleotide zurück und korrigiert diese durch Terme, die die Gesamtstruktur des Moleküls berücksichtigen.[5,6] Für komplexere Moleküle sind diese Vorhersagemodelle jedoch ungenau und die zugehörigen Deskriptoren nur sehr aufwändig zu bestimmen. Sequenzinformation (abgesehen von der Gesamtzusammensetzung) und Information über Sekundärstrukturen werden in diesen Modellen nicht berücksichtigt.

Für Peptide wurden Modelle ausgearbeitet, die die Retention nicht aus den Eigenschaften der Molekülbausteine ableiten, sondern diese aus mit Testanalyten bekannter Struktur gewonnenen Datensätzen erlernen.[7] Die Retentionsdaten von ca. 7000 Peptiden wurden beispielsweise dazu verwendet, ein künstliches neuronales Netzwerk (KNN) für die Vorhersage von Peptid-Retentionszeiten aus den Peptidsequenzen mit einer Genauigkeit von 3 bis 10% zu trainieren.[8] Weitere Verfahren aus dem Bereich des statistischen Lernens, z. B. Supportvektormaschinen (SVMs), lassen sich für Regressionsprobleme anwenden. Neben dem Vorteil, zu

[*] 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 (Deutschland)
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 (Deutschland)

[**] Wir danken LC Packings, Amsterdam, für die Bereitstellung des Kapillar-HPLC-Systems.

exakt einer global optimalen Lösung zu führen (im Unterschied zu KNNs), haben sich Supportvektor-Ansätze auch in praktischen Anwendungen auf chemische Probleme bewährt.[9,10]

Für die Vorhersage der Retention von Oligonucleotiden in der Ionenpaarumkehrphasenchromatographie (IP-RPC) wurde ein Retentionsmodell ausgearbeitet, das auf der Bestimmung und Addition der Retentionsbeiträge der Nucleotide beruhte.[11] Dieses Modell lieferte zufriedenstellende Ergebnisse für relativ hohe Trenntemperaturen (60 °C), bei denen Sekundärstrukturen wenig ausgeprägt sind, während bei niedrigeren Temperaturen der Einfluss von Haarnadeln oder partiellen Doppelsträngen zu einer schlechteren Übereinstimmung zwischen Vorhersage und Experiment führte (eigene Messergebnisse).

Unser Modell für die Retention von Oligonucleotiden in der IP-RPC auch bei niedrigeren Trenntemperaturen beruht auf der von Schölkopf et al. vorgeschlagenen ν-Supportvektorregression (SVR).[12] Diese Methode bestimmt ein Modell für einen gegebenen Datensatz, das gleichzeitig den Modellfehler und die Modellkomplexität minimiert. Das Trainieren dieses Modells erfolgt mit einer recht geringen Zahl von 50 bis 100 Oligonucleotiden. Ein Testdatensatz wurde durch Messung der Retentionszeiten von 72 Oligonucleotiden erstellt. Um den Einfluss der Sequenz auf die Retention zu erfassen, wurden 41 der Oligonucleotidsequenzen durch Variation der Sequenz eines 24mers (GTA CTC AGT GTA GCC CAG GAT GCC) generiert. Zur Berücksichtigung möglicher Sekundärstrukturen wurden vier weitere Sequenzen so ausgewählt, dass auch bei höheren Temperaturen stabile Haarnadelstrukturen gebildet werden. Die restlichen Sequenzen wurden schließlich so gewählt, dass sie einen Längenbereich von 15 bis 48 Nucleotiden abdeckten.

Quantitative Struktur-Eigenschafts-Beziehungen (QSPR) codieren die Eingabestrukturen in Form von Merkmalsvektoren. Solche für die vorherzusagende Eigenschaft relevanten Merkmale sind z. B. die Länge, Sequenz und Sekundärstruktur der untersuchten Oligonucleotide. Diese Merkmale werden in numerische Werte übersetzt, die direkt in das Modell einfließen. Zu jeder gemessenen Retentionszeit y_i wird die Sequenz und Struktur des Oligonucleotides durch einen Vektor $\boldsymbol{x}_i$ beschrieben. Merkmal-Werte-Paare ($\boldsymbol{x}_i$, y_i) dienen zur Bestimmung einer optimalen Funktion [Gl. (1)].

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

Diese Funktion ermöglicht die Vorhersage von Retentionszeiten $y = f(\boldsymbol{x})$ für einen beliebigen Merkmalsvektor $\boldsymbol{x}$, also für eine beliebige Sequenz. Für mathematische Details verweisen wir auf die einschlägige Literatur.[13,14]

In unserem Modell besteht der Merkmalsvektor aus elf Werten, die sich aus der Sequenz und der Sekundärstruktur ableiten. Fünf der Werte beschreiben die Sequenz (Länge und relativer Anteil der vier Basen), die restlichen sechs Werte die Schmelzkurve der Sekundärstruktur. Dazu wurde für die Temperaturen 30, 40, 50, 60, 70 und 80 °C die Sekundärstruktur mit der Software Vienna RNA Package Ver. 1.4[15] vorhergesagt und der Prozentsatz der Basen in gepaarten Bereichen errechnet. Aus dem Modell ergaben sich die Gesamtlänge, der Anteil der verschiedenen Basen sowie der Anteil der gepaarten Basen als wichtigste Merkmale eines Oligonucleotids. Andere Merkmalsvektoren, die z. B. die Basenabfolge, Basenstapelung oder andere Codierungen der Sekundärstrukturinformation enthalten, erbrachten keine besseren Ergebnisse.

Das Trainieren des SVR-Modells erfolgte durch Aufteilung des Testdatensatzes in drei Teile (je 24 Datenpunkte). Zwei Drittel der Testdaten wurden für das Trainieren des SVR-Modells verwendet, das restliche Drittel wurde zur Validierung der Vorhersage genutzt. Insgesamt wurde dreifach kreuzvalidiert. Die Werte R^2 und Q^2 geben dabei jeweils die Korrelation des SVR-Modells auf dem Trainingsdatensatz bzw. die Korrelation der vorhergesagten Retentionszeiten mit den experimentellen Zeiten an.

Abbildung 1 zeigt die Ergebnisse für die dreifach kreuzvalidierten Modelle bei 30 und 80 °C. Die geringe Streuung

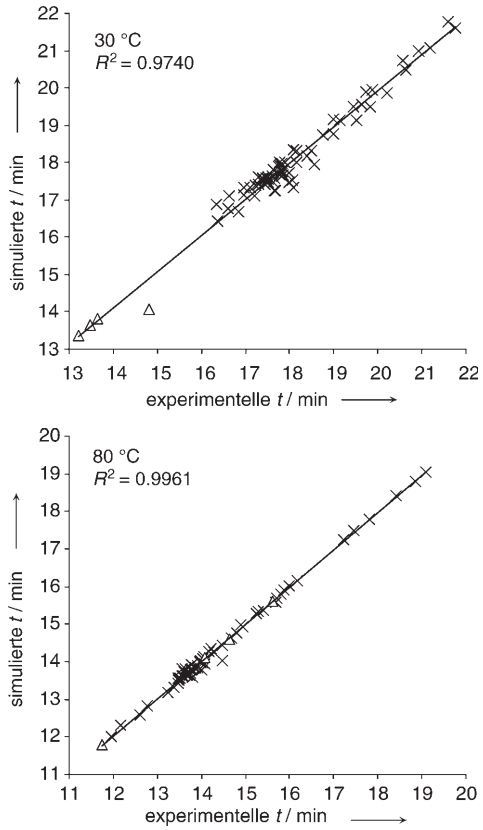

Abbildung 1. Retentionsvorhersage bei 30 und 80 °C; R^2-Werte entsprechen der quadrierten Korrelation zwischen dem Modell und den experimentellen Retentionszeiten. Haarnadelstrukturen sind durch Dreiecke markiert.

der Datenpunkte und das Fehlen von signifikanten Ausreißern belegen die Leistungsfähigkeit der SVR-Modelle zur Vorhersage der Retention über einen großen Temperaturbereich und den gesamten Längenbereich von 15 bis 50 Nucleotiden. Im Unterschied zu anderen Modellen sagt dieses Modell auch die Retention der Haarnadelstrukturen bei beiden Temperaturen mit guter Übereinstimmung vorher.

Beispiele für gemessene und simulierte Chromatogramme sind in Abbildung 2 wiedergegeben. Hier wurden 61 der 72 gemessenen Datenpunkte zum Trainieren des Modells verwendet, um die Retentionszeiten der restlichen elf Oligonucleotide zu simulieren. Tabelle 1 belegt, dass alle Retentionszeiten mit einer Abweichung von maximal 4%, meist weniger als 2%, vorhergesagt wurden. Auch die Retention von Haarnadelstrukturen, die bei erhöhter Temperatur stabil sind, wurde durch unser Modell korrekt wiedergegeben (Abbildung 2e). Interessant ist ebenfalls, dass auch die gemessene Retention für das relativ lange 39mere Oligonucleotid sehr gut mit dem Modell übereinstimmt, obwohl es für diesen Längenbereich nur wenige Trainingsdatenpunkte gab (Abbildung 2d).

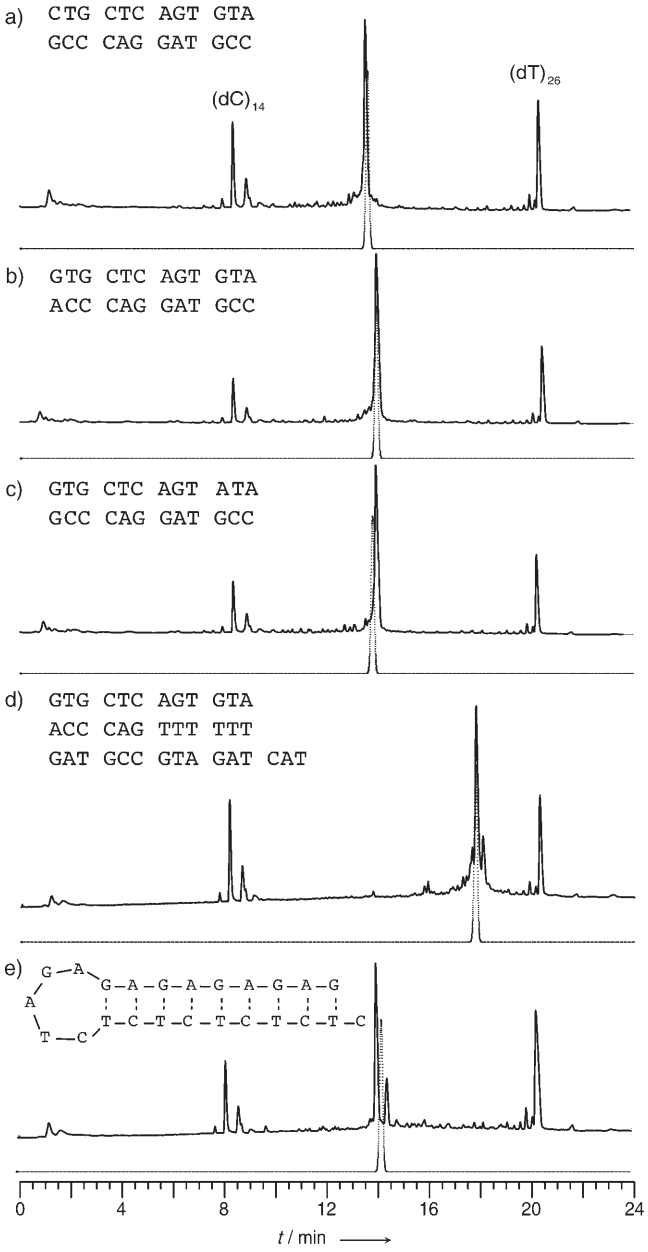

Abbildung 2. Vergleich der gemessenen und simulierten Chromatogramme von Oligonucleotiden bei 80°C. Für die Peakprofile in den Simulationen wurden die experimentell bestimmte mittlere Peak-Standardabweichung von 0.0575 min und eine Gauß-Funktion verwendet.

Die Anwendung unseres Verfahrens auf eine reale Oligonucleotidmischung ist in Abbildung 3 gezeigt. Es handelt sich um eine Mischung von Oligonucleotid-Primern, die für eine multiplexe Amplifikation durch Polymerasekettenreaktion verwendet wurde. Die Stärke des Modells zeigt sich darin, dass nicht nur die Retentionszeiten der relativ ungewöhnlichen, A,T-reichen Sequenzen, sondern auch die Retentionsreihenfolge der kurzen Oligonucleotide mit Ausnahme einer Inversion der Reihenfolge der Oligonucleotide 5 und 6 korrekt und mit einer durchschnittlichen bzw. maximalen Abweichung von 1.6% bzw. 3.2% vorhergesagt werden konnten. Die gute Übereinstimmung zwischen Theorie und Experiment ist Indiz dafür, dass die im Modell integrierten Strukturparameter die Wechselwirkungen zwischen Oligonucleotid und stationärer Phase in der IP-RPC sehr gut beschreiben.

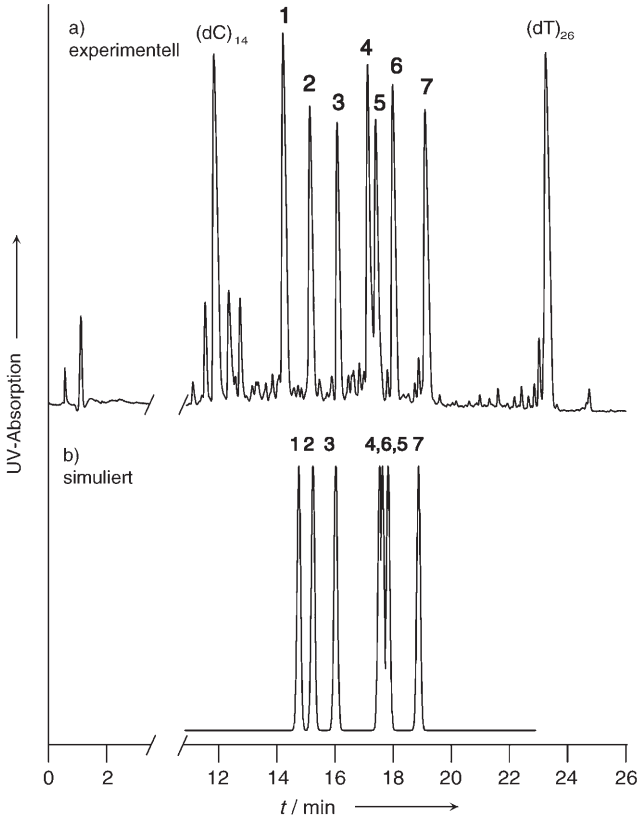

Abbildung 3. Vorhersage einer Trennung von sieben Oligonucleotid-Primern mit Längen von 14 bis 18 Nucleotiden bei 50°C. **1** = GTA GAG GTA GGT TGG, **2** = GAA GAG TTT TTG GA, **3** = GGG TTA ATT TGA GGT, **4** = TTT AGT TAG AAA AAG TT, **5** = GTT TAA ATA GGA AAT TT, **6** = GTT GGG ATT TTT GTA TTG, **7** = TAA GTT TTT TTT TGT TGT.

Zusammenfassend wurde erstmals ein Modell ausgearbeitet, das eine Vorhersage der Retention von Oligonucleotiden mit einer Genauigkeit von besser als 3% über den gesamten Temperaturbereich von 30 bis 80°C ermöglicht. Das Verfahren zeichnet sich dadurch aus, dass sich das Trai-

Tabelle 1: Vergleich gemessener und vorhergesagter Retentionszeiten von Oligonucleotiden bei 30 und 80 °C.[a]

Sequenz	Retention bei 30 °C [min]		rel. Fehler	Retention bei 80 °C [min]		rel. Fehler[%]
	gem.	ber.	[%]	gem.	ber.	[%]
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] Fettgedruckte Buchstaben kennzeichnen Modifikationen gegenüber der ersten 24mer-Sequenz.

nieren des Modells auf 50 bis 100 Oligonucleotide beschränkt und der Einfluss von Sekundärstrukturen auf die Retention berücksichtigt wird. Durch die Möglichkeit zur Aufnahme und Evaluation beliebig vieler Strukturparameter lässt sich schnell und einfach feststellen, welche Struktureigenschaften für die Wechselwirkungen eines Moleküls in einem chromatographischen Trennsystem ausschlaggebend sind. In Folgearbeiten möchten wir das Verfahren auch zur verbesserten Vorhersage von Peptid-Retentionszeiten nutzen.

Experimentelles

Die Retentionszeiten der Oligonucleotide (jeweils 2.5 ng injiziert) wurden in einer monolithischen Poly(styrol/divinylbenzol)-Säule (60 × 0.20 mm Innendurchmesser) mithilfe eines 30-minütigen Gradienten von 0–16 % Acetonitril in einer wässrigen Lösung von 100 mmol L^{-1} Triethylammoniumacetat und 0.5 mmol L^{-1} Ethylendiamintetraessigsäure mit Detektion bei 254 nm und Säulentemperaturen von 30–80 °C bestimmt. Die Flussgeschwindigkeit wurde mithilfe eines HPLC-Systems mit aktivem Split (U3000, LC Packings, Amsterdam) bei 2.0 µL min^{-1} konstant gehalten. Die Standardabweichung der gemessenen Retentionszeiten von $(dT)_{18}$ aus 20 Wiederholungsmessungen lag zwischen 1.3 s (RSD 0.097 %) bei 40 °C und 3.1 s (RSD 0.28 %) bei 70 °C. Durch Korrektur über interne Standards (jeweils 1 ng $(dC)_{14}$ und $(dT)_{26}$) reduzierten sich die relativen Standardabweichungen (RSDs) der gemessenen Retentionszeiten auf 0.028 % (40 °C) bzw. 0.072 % (70 °C). Die normierte Nettoretentionszeit t' errechnete sich aus Gleichung (2). t ist die experimentelle

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

Retentionszeit des Oligonucleotids, t_C und t_T sind die Retentionszeiten von $(dC)_{14}$ bzw. $(dT)_{26}$, $\bar{t}_C$ und $\bar{t}_T$ sind die durchschnittlichen Retentionszeiten von $(dC)_{14}$ bzw. $(dT)_{26}$ über alle Experimente. Das SVR-Modell wurde mit dem Softwarepaket libSVM in der Version 2.8[16] erstellt.

Eingegangen am 27. Juni 2006
Online veröffentlicht am 29. September 2006

Stichwörter: Bioinformatik · Flüssigkeitschromatographie · Oligonucleotide · Sekundärstrukturen · Struktur-Aktivitäts-Beziehungen

[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* (Hrsg.: L. Niklasson, M. Boden, T. Ziemke), **1998**, S. 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).

