Humboldt-Universität zu Berlin
Institut für Informatik

Studienarbeit

Entwicklung einer Methode zur Vorhersage des zeitlichen Verlaufs chromatographischer Peaks

Benjamin Daeumlich
daeumlic@informatik.hu-berlin.de

17. Dezember 2008

Betreuer: Katja Tham
Andreas Kühn

Inhaltsverzeichnis

Abbildungsverzeichnis

Kapitel 1

Einleitung

In der Proteomforschung wird der Aufbau, die Funktion und das Zusammenspiel von Proteinen und Peptiden untersucht ([Sch03]). Dadurch werden Informationen gewonnen, die für das Leben einer Zelle von entscheidender Bedeutung sind. Diese Informationen sind beispielsweise für pharmazeutische Unternehmen interessant, da damit die Ursachen von Krankheiten ermittelt werden können. In den letzten Jahren haben sich leistungsfähige massenspektrometrische Verfahren zur Proteomanalyse immer stärker durchgesetzt. In Kooperation mit dem Lehrstuhl für Analytische Chemie hat der Lehrstuhl für Datenbanken und Informationssysteme das Projekt "Intelligente datenabhängige Massenspektrometrie" ins Leben gerufen, welches zum Ziel hat, die Analyseverfahren für Proteome zu optimieren.

Bei der Untersuchung von Stoffgemischen (Peptidgemische in der Proteomforschung) werden vorgeschaltete Trennverfahren genutzt, um die einzelnen Stoffe des Gemisches zu separieren. Nach der Trennung können die separierten Stoffe in einem Analysegerät genauer untersucht werden. Durch die Kopplung von Trennsystem und Analysegerät sind einzelne Stoffe nur in einem bestimmten Zeitfenster verfügbar, welches von Stoff zu Stoff variiert. Aufgrund der Eigenschaften des Trennverfahrens sind mehrere Stoffe gleichzeitig in unterschiedlichen Konzentrationen verfügbar. Nach Möglichkeit sollen alle Stoffe genauer untersucht werden. Das Analysegerät kann allerdings nur einen Stoff zu einem Zeitpunkt genauer untersuchen. Somit ist eine Abschätzung der Zeit, in der ein Stoff für Untersuchungen verfügbar ist, notwendig, um eine Untersuchungsreihenfolge zu finden, in der möglichst viele Stoffe

genauer untersucht werden können (im optimalen Fall alle Stoffe). Wenn beispielsweise Stoff A und Stoff B gleichzeitig vorhanden sind und es feststeht, dass Stoff B noch doppelt so lange verfügbar ist wie Stoff A, wird zuerst Stoff A und anschließend Stoff B untersucht. Dieser Sachverhalt tritt auch bei der Massenspektrometrie auf, auf welche sich diese Studienarbeit bezieht. Aus den dabei gewonnenen Messdaten lässt sich für jeden auftretenden Stoff eine Kurve extrahieren, die den zeitlichen Verlauf der Intensität des Stoffes widerspiegelt. Das Problem besteht darin, den zeitlichen Verlauf dieser Intensitätskurve vorherzusagen. So kann entschieden werden, in welcher Reihenfolge die Stoffe untersucht werden müssen. Für die Lösungsfindung spielen somit Funktionen, welche die Intensitätskurve beschreiben, eine wichtige Rolle.

In dieser Studienarbeit wird eine Methode zur Vorhersage des zeitlichen Verlaufs der Intensitätskurve von Stoffen entwickelt. Dies ermöglicht es, die Verweildauer des Stoffes in der Untersuchungsanordnung abzuschätzen. Dazu werden im zweiten Kapitel zuerst Grundlagen geschaffen: Es wird eine Einführung in die Massenspektrometrie gegeben und zudem auf Funktionen eingegangen, welche für die Lösungsfindung benutzt werden. Im darauf folgenden Kapitel werden verschiedene Ansätze zur Vorhersage des zeitlichen Verlaufs der Intensitätskurve untersucht. Weiterhin wird die Umsetzung der Vorhersage und eine Evaluierung dieser erläutert. Außerdem wird im Zuge dieser Arbeit ein Modul entwickelt, welches aus realen Messdaten den zeitlichen Verlauf der Intensitätskurve zur Laufzeit bestimmt. Dieses Modul wird im vierten Kapitel näher beschrieben. Abschließend wird eine Zusammenfassung über die Ergebnisse dieser Arbeit gegeben und es wird auf mögliche weiterführende Arbeiten eingegangen.

Die besondere Herausforderung liegt darin, dass die Vorhersage während der Untersuchung durchgeführt werden muss. Das heißt zum einen, dass sie möglichst kurz nach dem Auftreten eines Stoffes getroffen werden muss und zum anderen, dass notwendige Berechnungen nicht zu viel Zeit in Anspruch nehmen dürfen. Letzteres liegt einerseits daran, dass die Berechnungen für mehrere Stoffe gleichzeitig durchgeführt werden müssen und andererseits werden weitere Untersuchungen gleichzeitig vorgenommen.

Kapitel 2

Grundlagen

In diesem Kapitel wird zuerst eine Einführung in die Massenspektrometrie gegeben, es werden Begrifflichkeiten diesbezüglich definiert und es wird die Struktur und Art der anfallenden Messdaten beschrieben. Anschließend wird auf Funktionen eingegangen, welche für die Beschreibung der in der Einleitung erwähnten Intensitätskurve in Frage kommen. Weiterhin gibt es eine Einführung in das Anpassen von Funktionen an die Punkte der Intensitätskurve, welches für die Vorhersage des zeitlichen Verlaufs benötigt wird.

2.1 Einführung in die Massenspektrometrie

Mit Hilfe der Massenspektrometrie wird die genaue Zusammensetzung von Stoffen aus ihren Einzelbausteinen erforscht (wie z.B. die Aminosäuresequenzen von Proteinen und Peptiden). Ein für die Untersuchung notwendiger Messvorgang setzt sich aus mehreren Messungen (Scans) zusammen, wobei abhängig vom benutzen Messgerät ca. 1 Scan pro Sekunde durchgeführt werden kann. Es wird zwischen Übersichts- und Fragmentierungsscan unterschieden, allerdings kann jeweils nur einer von beiden zu einem Zeitpunkt durchgeführt werden. Ein Übersichtsscan liefert zu einem Zeitpunkt die detektierten Masse-zu-Ladungsverhältnisse (m/z-Verhältnisse) aller auftretenden Stoffe mit entsprechenden Intensitäten.

Definition 2.1 (m/z-Verhältnis) *Das m/z-Verhältnis macht es möglich, einen Stoff anhand von Masse und Ladung zu charakterisieren. Durch die Ermittlung der*

Ladung kann seine genaue Masse bestimmt werden. Instrumentell bedingt schwankt der Wert des m/z-Verhältnisses für einen Stoff, sodass mehrere m/z-Verhältnisse mit verschiedenen Intensitäten existieren. Das Maximum der daraus resultierenden Verteilungskurve wird als Repräsentant für den Stoff verwendet.

Bei einem Fragmentierungsscan hingegen wird ein detektierter Stoff durch Energiezufuhr derart angeregt, dass es zum Bruch von bestimmten chemischen Bindungen kommt. Somit entstehen mehrere Bruchstücke (Fragmente), aus deren Eigentschaften weitere Informationen über die Zusammensetzung des Stoffes gewonnen werden können.

In Abbildung 2.1 wird der Untersuchungsvorgang schematisch dargestellt.

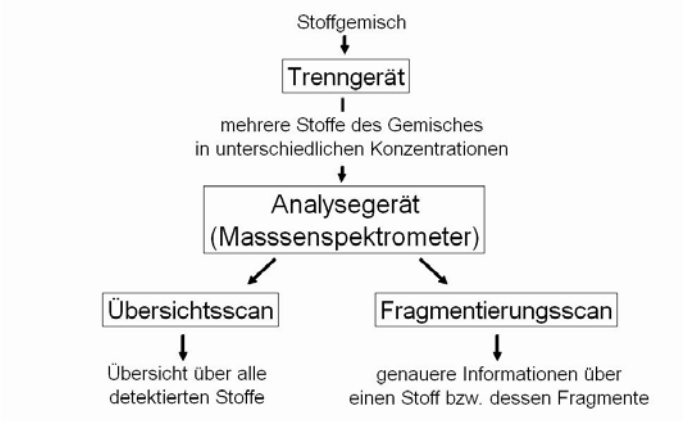


Abbildung 2.1: Schematische Darstellung des Untersuchungsvorgangs

Die bisher gebräuchliche Abfolge von Übersichts- und Fragmentierungsscans ist sehr statisch, das heißt es wird beispielsweise erst ein Übersichtsscan und dann je ein Fragmentierungsscan für die intensivsten Stoffe durchgeführt. Dieser Ablauf wird dann periodisch wiederholt. Durch diese Art der Messabfolge ist allerdings nicht gewährleistet, dass möglichst viele Stoffe genauer untersucht werden können, da der zeitliche Verlauf der auftretenden Stoffe nicht verfolgt und berücksichtigt wird.
Eine Abschätzung für den Zeitpunkt, ab dem ein Stoff aufgrund von zu geringer Intensität nicht mehr für einen Fragmentierungsscan zur Verfügung steht (im Folgenden Schwellwert genannt), ist somit notwendig. Damit kann die Reihenfolge der

Scans dynamisch festgelegt werden, sodass für möglichst alle Stoffe ein bzw. mehrere Fragmentierungsscans durchgeführt werden können.

2.2 Anfallende Messdaten

In jedem Scan werden zu einem Zeitpunkt die auftretenden m/z-Verhältnisse mit ihren zugehörigen Intensitäten gemessen. Somit ergeben sich für einen gesamten Messvorgang drei Dimensionen. Es existieren verschiedene Darstellungen der Messdaten. Dabei handelt es sich um Darstellungen für die instrumentell bedingten Profildaten bzw. centroiden Daten und um das EIC (Extracted Ion Chromatogramm), welches deren zeitlichen Verlauf wiederspiegelt.

Definition 2.2 (Darstellung von Profildaten) *Die Darstellung von Profildaten ist eine zweidimensionale Darstellung, bei der für einen Zeitpunkt jedem auftetenden m/z-Verhhältnis die entsprechende Intensität zugeordnet wird.*

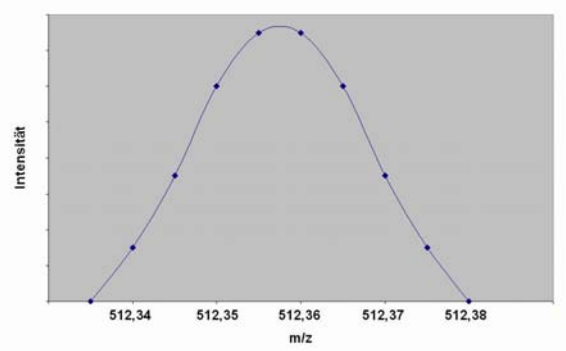


Abbildung 2.2: Darstellung von Profildaten

In Abbildung 2.2 wird ein Beispiel für die Profildatendarstellung gegeben. Es werden dabei Intensitäten für m/z-Verhältnisse zwischen 512,34 und 512,38 gemessen. Der Wertebereich der auftretenden m/z-Verhältnisse ist abhängig von der Auflösung des Messgerätes.

Definition 2.3 (Darstellung centroider Daten) *Centroide Daten sind verarbeitete Profildaten, wobei die Verarbeitung entweder direkt im Messgerät oder nach der*

Messung erfolgt. Es wird für jeden Stoff aus allen zugehörigen m/z-Verhältnissen ein Repräsentant für den jeweiligen Stoff bestimmt. Die Darstellung centroider Daten ist eine zweidimensionale Darstellung, bei der für einen Zeitpunkt jedem m/z-Verhältnis, welches einen Repräsentanten für einen Stoff darstellt, die entsprechende Intensität zugeordnet wird.

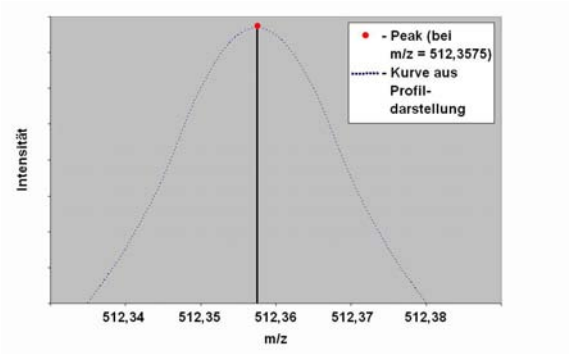


Abbildung 2.3: Darstellung centroider Daten

Definition 2.4 (Peak) *Ein Peak ist Repräsentant für einen Stoff. Es handelt sich dabei um das Intensitätsmaximum der Kurve aus der Profildatendarstellung.*

Abbildung 2.3 zeigt den Peak (roter Punkt), welcher aus der Kurve der Profildatendarstellung (gestrichelte Linie) in Abbildung 2.2 gewonnen wurde. Der Peak befindet sich beim m/z-Verhältnis von 512,3575.

Definition 2.5 (EIC) *Im EIC (Extracted Ion Chromatogramm) wird die Intensität eines Peaks über die Zeit hinweg verfolgt. Die entstehende Kurve wird auch als chromatographischer Peak bezeichnet. Es ist eine Darstellung des zeitlichen Verlaufs der Profildaten bzw. der centroiden Daten.*

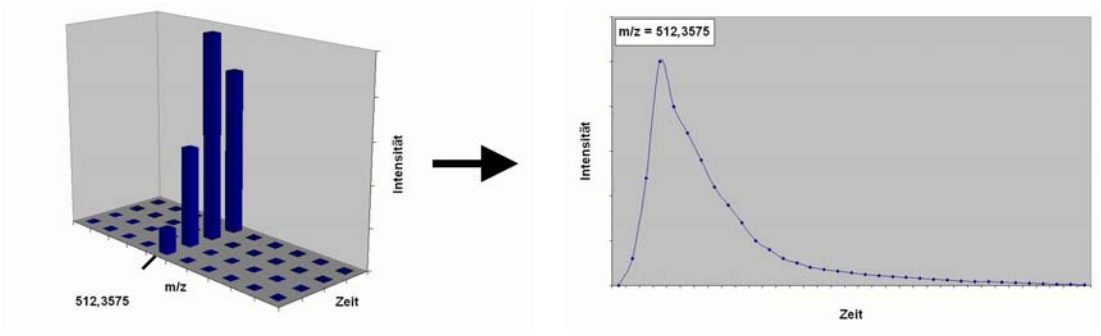


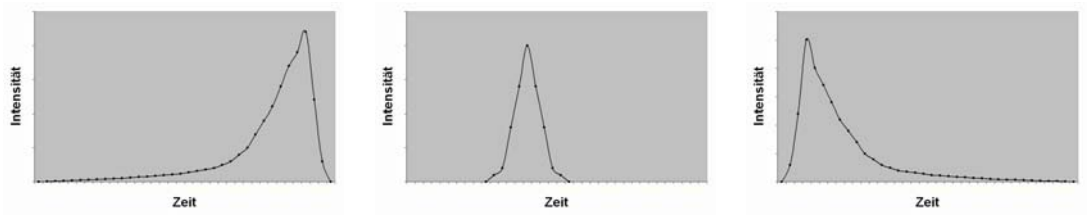
Abbildung 2.4: Extracted Ion Chromatogramm (chromatographischer Peak)

Abbildung 2.4 zeigt das Extracted Ion Chromatogramm bzw. den chromatographischen Peak für das m/z-Verhältnis von 512,3575.

2.3 Funktionen zur Beschreibung chromatographischer Peaks

In der Literatur ([TLP01],[ZP02],[Li02],[VBDM01]) werden eine Anzahl von Funktionen genannt, welche den Verlauf von chromatographischen Peaks beschreiben (von nun an Peakfunktionen genannt). Im Folgenden wird eine Klassifizierung der Peakfunktionen vorgenommen und es wird insbesondere auf deren Parameter eingegangen.

In [SVR00] werden Peakfunktionen beschrieben und klassifiziert. Nahezu alle Peakfunktionen basieren auf Gauss- und Lorentz-Funktionen. Es wird zwischen symmetrischen und asymmetrischen Peakfunktionen unterschieden. Asymmetrische Funktionen können auch Peaks mit langem Anstieg und kurzem Abfall ("Fronting") bzw. langem Abfall und kurzem Anstieg ("Tailing") beschreiben. Die verschiedenen Arten von Peakfunktionen werden in Abbildung 2.5 skizziert.



(a) asymmetrisch - "Fronting" (b) symmetrisch (c) asymmetrisch - "Tailing"

Abbildung 2.5: Klassifikation von Peakfunktionen

Die in [TLP01] und [VBDM01] angegebenen Peakfunktionen haben folgende Parameter:

- Maximum - Intensitätsmaximum des chromatographischen Peaks;
- Zeitpunkt des Maximums - Zeitpunkt, an dem das Intensitätsmaximum auftritt;
- Symmetrieparameter - Parameter, welcher die Symmetrieeigenschaften eines Peaks beschreibt, positiv für Peaks mit langem Anstieg, 0 für symmetrische Peaks, negativ für Peaks mit langem Abfall;
- Peakbreite - Parameter, welcher die Breite eines Peaks beschreibt, je größer er ist, desto breiter ist der Peak.

Eine in der Literatur sehr häufig beschriebene Funktion zur Beschreibung chromatographischer Peaks ist die Funktion "Exponential Modified Gauss", welche in [VBDM01] durch folgende Gleichung beschrieben wird:

$$y(t) = \frac{h}{a} exp\left[\frac{1}{2}\left(\frac{w}{a}\right)^2 - \frac{t-t_c}{a}\right] \int_{-\infty}^{\frac{t-t_c}{w}-\frac{w}{a}} \frac{1}{\sqrt{2\pi}} exp\left[\frac{y^2}{2}\right] \, \mathrm{d}y \tag{2.1}$$

Der Parameter h steht dabei für das Maximum des chromatographischen Peaks, t_c für den Zeitpunkt des Maximums, a ist der Symmetrieparameter und w die Peakbreite. Somit enthält diese Funktion alle oben aufgeführten Parameter. Es existieren

Verfahren, eine Funktion an vorhandene Datenpunkte (in unserem speziellen Fall die Punkte des chromatographischen Peaks) anzupassen.

Beim Anpassen einer Funktion an vorhandene Punkte werden die Parameter der Funktion solange variiert, bis der Abstand zwischen den Punkten und der Funktionskurve möglichst gering ist. Diese Methode wird im Folgenden Funktionsfitting genannt. Es gibt verschiedene Programme, die das Funktionsfitting nach Eingabe der Messpunkte und Auswahl einer Funktion automatisch durchführen. Abbildung 2.6 zeigt ein mit dem Programm "OriginPro 7.5" von OriginLab durchgeführtes Funktionsfitting an Realdaten. Als Funktion wurde "Exponential Modified Gauss" verwendet.

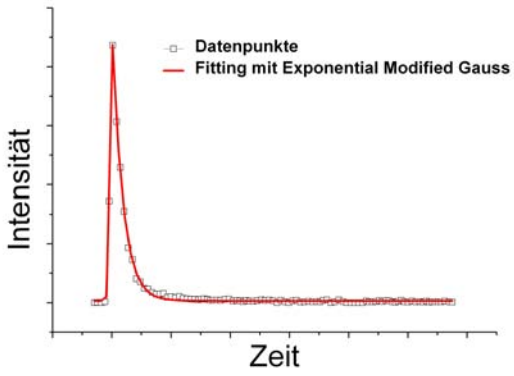


Abbildung 2.6: Funktionsfitting eines EIC mit "Exponential Modified Gauss"

Wie Abbildung 2.6 zeigt, beschreibt die gefundene Funktion den chromatographischen Peak sehr gut (maximal $0,5\%$ Abweichung bezüglich des Wertes des Intensitätsmaximums), allerdings sind die initialen Werte der Parameter im Vorfeld nicht gut abzuschätzen und somit werden zuerst alle Messdaten benötigt, um ein optimales Funktionsfitting durchführen zu können.

Kapitel 3

Vorhersage des Verlaufs

In diesem Kapitel wird eine Methode zur Vorhersage des Verlaufs von chromatografischen Peaks entwickelt. Es wird auf verschiedene Lösungsansätze eingegangen. Anschließend wird die Vorhersage anhand von realen Testdaten bewertet.

3.1 Ansätze

Bei der Suche nach einer Methode zur Vorhersage des Verlaufs von chromatographischen Peaks wurden verschiedene Ansätze verfolgt. Die ersten beiden Ansätzen hatten zum Ziel, den Zeitpunkt, an dem der Schwellwert erreicht wird (im Folgenden Peakende genannt), abzuschätzen. Da es letztendlich nur von Bedeutung ist, wie lange ein Stoff untersucht werden kann, würde diese Art der Vorhersage eine ausreichende Lösung des Problems darstellen.
Die Ansätze werden im Folgenden erläutert.

3.1.1 Ansatz über Winkel

Die Idee des ersten Ansatzes ist es, eine Vorhersage ohne komplexe Peakfunktionen zu treffen. Es wurde versucht, über die Höhe des Anstieges eines chromatographischen Peaks, welcher aus wenigstens zwei Messpunkten ermittelt werden kann, einen Winkel für einen linearen Abfall abzuleiten. Dies würde es gestatten, das Peakende abzuschätzen. Der Vorteil dieser Art der Vorhersage wäre es, dass sie zum einen

bereits im Anstieg getroffen wird und dass zum anderen keine komplexen Peakfunktionen sondern nur lineare Funktionen benötigt werden.
Theoretische Betrachtungen haben allerdings gezeigt, dass alleine über die Höhe des Anstieges keine Vorhersage durchgeführt werden kann, da keine eindeutige Abbildung zwischen Höhe des Anstieges und Peakende existiert.

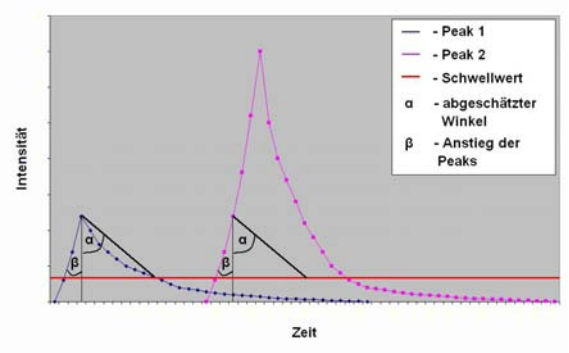


Abbildung 3.1: Winkelansatz

In Abbildung 3.1 wird deutlich, dass trotz des gleichen Anstieges von Peak 1 und Peak 2 in den ersten drei Punkten (Winkel β) der Verlauf sehr unterschiedlich sein kann. Wird nun anhand von Peak 1 ein Winkel für einen linearen Abfall festgelegt (Winkel α), sodass das Peakende (der Schwellwert ist die rote Linie) genau vorhergesagt wird, lässt sich dieser Winkel nicht auf Peak 2 übertragen.
Dieser Ansatz lässt sich also aufgrund der Eigenschaften der Anstiege chromatographischer Peaks nicht zur Problemlösung verwenden. Es wird deutlich, dass alleine aus den Anstiegen keine Vorhersage durchgeführt werden kann. Somit wird das Intensitätsmaximum als weitere Größe benötigt.

3.1.2 Ansatz über einfache Funktionen

In diesem Ansatz wurden einfache symmetrische Funktionen zur Vorhersage benutzt. Die Idee ist es, einen Zusammenhang zwischen dem Intensitätsmaximum des Peaks und dem Peakende zu finden. Somit kann die Vorhersage erst nach dem Erreichen des Maximums durchgeführt werden. Dies ist im Rahmen der Problematik realisierbar, da zum einen das Maximum bereits nach sehr kurzer Zeit erreicht ist und da zum anderen eine Vorhersage des Verlaufs ohne Kenntnis des Maximums nicht gegeben

ist. Im Gegensatz zum ersten Ansatz würde die Vorhersage erst später, das heißt nach Erreichen des Maximums, durchgeführt werden, aber es wird auch hier auf komplexe Peakfunktionen verzichtet. In Abbildung 3.2 wird dieser Ansatz verdeutlicht.

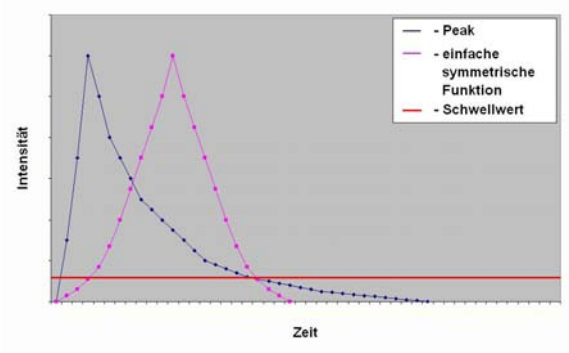


Abbildung 3.2: Ansatz über einfache Funktionen

Untersuchungen an einfachen symmetrischen Funktionen haben allerdings gezeigt, dass kein einfacher Zusammenhang (Proportionalität) zwischen Intensitätsmaximum des Peaks und dem Peakende existiert. Es wurden beispielsweise für verschiedene Peaks die Parameter der Gaußschen Glockenkurve bestimmt, es ließ sich aber kein Zusammenhang finden. Somit wird deutlich, dass komplexere Funktionen verwendet werden müssen.

3.1.3 Ansatz über Peakfunktionen

Da der Ansatz über einfache Funktionen für unser Problem keine genügende Lösung bereitstellt, werden nun komplexere Funktionen betrachtet, welche den gesamten chromatographischen Peak gut annähern. Um solch eine Funktion zu finden, können die in [TLP01] und [ZP02] beschriebenen Peakfunktionen an reale Messpunkte gefittet werden. Die Vorhersage des zeitlichen Verlaufs soll möglichst sofort nach Auftreten eines Stoffes getroffen werden. Da das Funktionsfitting allerdings zuerste alle Messpunkte benötigt um ein optimales Ergebnis zu erzielen, ist es nur bedingt nutzbar. Wird aber eine Peakfunktion gefunden, welche die chromatographischen Peaks möglichst gut annähert, muss eine Korrelation zwischen den vorhandenen Messdaten und den Parametern der Peakunktion gefunden werden. Für das Funktionsfitting von chromatographischen Peaks lässt sich beispielsweise das Programm "OriginPro

7.5" von OriginLab verwenden. Dort lassen sich neue Funktionen sehr gut einpflegen, sodass bei der Suche nach einer Funktion, welche die chromatographischen Peaks der Massenspektroskopie beschreibt, sehr viele verschiedene Funktionen getestet werden können.
Dieser Ansatz liefert über das Funktionsfitting ein sehr gutes Ergebnis, falls im Vorfeld alle Messdaten bekannt sind. Die Herausforderung liegt nun darin, eine Peakfunktion zu finden, welche die chromatographischen Peaks gut annähert und deren Parameter im Vorfeld festgelegt werden können.

3.1.4 Fazit

Die ersten beiden Ansätze kommen ohne komplexe Peakfunktionen aus. Leider lassen sich diese Ansätze nicht für unsere Problemstellung verwenden. Durch den ersten Ansatz wird zudem deutlich, dass alleine der Anstieg keine Vorhersage des Peakendes erlaubt, sondern dass erst das Intensitätsmaximum abgewartet werden muss. Der dritte Ansatz hingegen lässt sich anwenden, es muss jedoch eine Peakfunktion gefunden werden, welche die chormatographischen Peaks gut annähert und deren Parameter im Vorfeld festgelegt werden können. Somit wird im Folgenden der dritte Ansatz benutzt, um die Vorhersage durchzuführen.

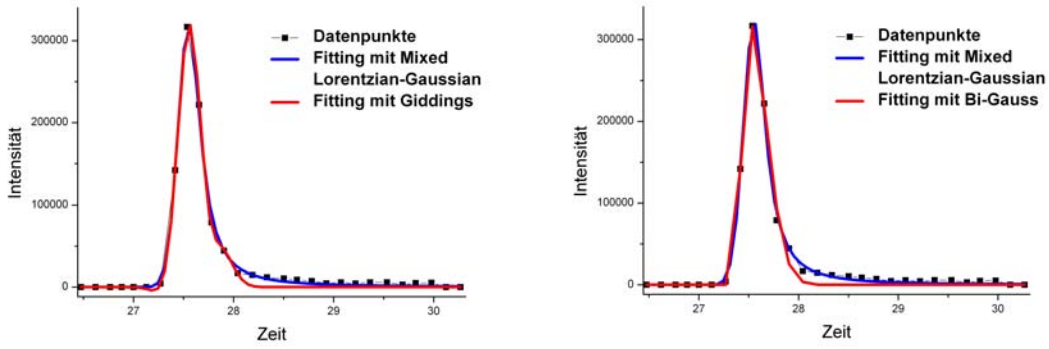
3.2 Umsetzung des gewählten Ansatzes

In diesem Abschnitt wird der gewählte Ansatz über Peakfunktionen erläutert. Dazu muss eine Peakfunktion gefunden werden, welche die oben genannten Anforderungen erfüllt.
Es hat sich herausgestellt, dass die im verwendeten System gemessenen chromatographischen Peaks asymmetrisch sind und einen im Vergleich zum Anstieg sehr langen Abfall haben. Durch diese Charakterisierung der chromatographischen Peaks lassen sich von den wesentlichen 90 Peakfunktionen, welche in [TLP01] und [ZP02] beschrieben werden, bereits 47 ausschließen.
Nach Gegenüberstellung der restlichen in Frage kommenden Funktionen wird deutlich, dass alle das Intensitätsmaximum als Parameter enthalten. Dies ist ein weiteres Argument für die Festlegung, eine Abschätzung erst nach Erreichen des Maximums

durchzuführen.
Es wurden Peakfunktionen gefunden, welche gut an die gemessenen chromatographischen Peaks angepasst werden können. Die Funktion "Exponential Modified Gauss" ist wie bereits in Kapitel 2 erwähnt solch eine Funktion. Um eine Vorhersage möglichst schnell treffen zu können, muss es nun eine Korellation zwischen den Messdaten und den Parametern der Funktion geben. Da solch eine Korellation für diese Funktion nicht gegeben ist, musste eine Alternative gefunden werden.
In [TLP01] werden weiterhin 12 Peakfunktionen beschrieben, welche chromatographische Peaks simulieren. Diese sind durch eine geringe Anzahl von Parametern charakterisiert.
Nach der Untersuchung dieser Funktionen wurde schließlich die Funktion "Mixed Lorentzian-Gaussian" ausgewählt, welche die chromatographischen Peaks am Besten annähert. Die Funktion "Mixed Lorentzian-Gaussian" wird in [VBDM01] durch folgende Gleichung beschrieben:

$$y(t) = \begin{cases} h * exp\left[-\frac{(t-t_c)^2}{2w_1^2}\right] & \text{falls } t < t_c \\ h * \left[1 + \left(\frac{t-t_c}{w_2}\right)^2\right]^{-1} & \text{falls } t \geq t_c \end{cases} \tag{3.1}$$

Der Parameter h steht dabei für die Intensität des Maximums, der Parameter t_c für den Zeitpunkt des Erreichens des Maximums und die Parameter w_1 und w_2 stehen jeweils für die Breite der beiden Peakteile (Anstieg bzw. Abfall).
Die anderen 11 Funktionen aus [TLP01], welche chromatographische Peaks simulieren, beschreiben den im Vergleich zum Anstieg sehr langen Abfall sehr schlecht, das heißt sie fallen zu schnell ab. Dies wird in den folgenden Abbildungen 3.3 am Beispiel der Funktionen "Giddings" und "Bi-Gauss" deutlich.



(a) Giddings (b) Bi-Gauss

Abbildung 3.3: Funktionsfitting mit Giddings, Bi-Gauss und Mixed Lorentzian-Gaussian

Die rote Kurve beschreibt das Funktionsfitting mit der Peakfunktion "Giddings" (Abbildung 3.3(a)) bzw. "Bi-Gauss" (Abbildung 3.3(b)). Im Vergleich dazu wird jeweils das Funktionsfitting mit der Peakfunktion "Mixed Lorentzian-Gaussian" durch die blaue Kurve beschrieben. Es wird deutlich, dass die roten Kurven zu schnell abfallen und somit der im Vergleich zum Anstieg sehr lange Abfall nicht gut beschrieben wird. Die blaue Kurve hingegen überlagert die Datenpunkte sehr gut.
Im Folgenden werden die Gründe für die Auswahl der Funktion "Mixed Lorentzian-Gaussian" beschrieben:

- Die Funktion kann gut an die gemessenen chromatographischen Peaks angepasst werden (siehe Anhang A.2). Viele der Funktionen in [TLP01] und [ZP02] beschreiben insbesondere den im Vergleich zum Anstieg sehr langen Abfall sehr schlecht, da sie zu schnell abklingen.

- Beide Teile (Anstieg und Abfall) des chromatographischen Peaks werden in dieser Funktion völlig separat voneinander beschrieben. Da eine Abschätzung erst nach dem Erreichen des Maximums durchgeführt wird, kann der Anstieg vernachlässigt werden. Dies ist ein Vorteil im Vergleich zur Funktion "Exponential Modified Gauss", in der Anstieg und Abfall gemeinsam beschrieben werden.

- Die Funktion hat neben dem Intensitätsmaximum und dessen Zeitpunkt nur einen Parameter für die Peakbreite. Ein Symmetrieparameter, welcher beispielsweise in der Gleichung der Funktion "Exponential Modified Gauss" zu finden ist, ist nicht enthalten.

- Die Funktionsgleichung enthält kein Integral wie beispielsweise die Gleichung der Funktion "Exponential Modified Gauss". Somit ist der Aufwand für Berechnungen geringer.

- Es lässt sich eine einfache Korrellation zwischen den Parametern der Funktion und den Messdaten finden. Es exisiteren zwar auch andere Peakfunktionen, die wie die ausgewählte gut an die chromatographischen Peaks angepasst werden können (beispielsweise "Exponential Modified Gauss"), allerdings ist dabei keine Korellation erkennbar.

Nachdem eine Funktion ausgewählt wurde, musste im nächsten Schritt eine Korrelation zwischen den Parametern der Funktion und den Messdaten gefunden werden, sodass sich der zeitliche Verlauf des chromatographischen Peaks möglichst frühzeitig und genau vorhersagen lässt. Dazu wurden reale Messdaten von 20 verschiedenen Peaks betrachtet. Da in der ausgewählten Funktion Anstieg und Abfall separat behandelt werden und da der genaue Verlauf des Anstieges aufgrund der Abschätzung nach Erreichen des Intensitätsmaximums keine Rolle spielt, muss er für die Vorhersage nicht betrachtet werden. Weitere Untersuchungen an den realen Messdaten haben gezeigt, dass sich der Parameter w_2 einer ersten Abschätzung zufolge lediglich im Bereich zwischen 0.13 und 0.34 bewegt (siehe Anhang A.1). Somit wird für eine zuerst noch ungenaue Vorhersage des Verlaufs nach Erreichen des Maximums der Wert für den Parameter w_2 auf 0.235 gesetzt, was genau dem Wert in der Mitte zwischen den beiden Grenzwerten entspricht. Nach dem Erhalt von mehr Messpunkten im zeitlichen Verlauf lässt sich dieser Wert weiter anpassen, sodass die Vorhersage besser wird, umso mehr Punkte zur Verfügung stehen.
Die Auswirkungen des Parameters w_2 auf die Form der Funktionskurve wird in der folgenden Abbildung 3.4 deutlich.

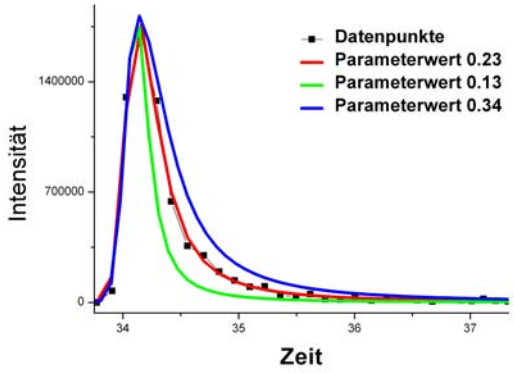


Abbildung 3.4: Peakfunktionen für verschiedene Werte des Parameters w_2

Es wird deutlich, dass ein kleinerer Wert des Parameters einen stärkeren Abfall bewirkt. Je größer der Wert ist, desto später fällt die Funktionskurve ab.

3.3 Evaluierung

Zur Evaluierung der entwickelten Vorhersage werden beispielhaft drei Peaks betrachtet. Die Abschätzung wurde mit dem im Zuge dieser Arbeit entwickelten Modul durchgeführt, welches in Kapitel 4 näher beschrieben wird. Es wurden drei unterschiedliche Peaks ausgewählt, das heißt ein Peak mit geringer Intensität (Peak 1), ein Peak mit hoher Intensität und langer Dauer (Peak 2) und ein Peak mit kurzer Dauer (Peak 3). In Tabelle 3.1 sind allgemeine Informationen über den jeweiligen Peak und die Ergebnisse der Vorhersage zu finden, wobei zwei Abschätzung durchgeführt wurden. Die erste Abschätzung wurde ein Datenpunkt und die zweite vier Datenpunkte nach Erreichen des Intensitätsmaximums durchgeführt.

	Peak 1	**Peak 2**	**Peak 3**
Allgemeine Informationen über den Peak			
m/z-Verhältnis	758.26	681.84	439.23
Intensität des Maximums	80000	1670000	104000
Dauer des Peaks	106s	241s	65s
Anzahl der Datenpunkte	10	34	10
Abschätzung ein Datenpunkt nach Erreichen des Maximums			
abgeschätztes Peakende	85s	259s	85s
Wert von w_2	0.465	0.325	0.465
Abweichung vom realen Peakende	21s	18s	20s
Abschätzung vier Datenpunkte nach Erreichen des Maximums			
abgeschätztes Peakende	97s	224s	63s
Wert von w_2	0.54	0.28	0.345
Abweichung vom realen Peakende	9s	17s	2s

Tabelle 3.1: Evaluierung an verschiedenen Peaks

Es wird deutlich, dass die Abweichung des abgeschätzten Peakendes vom realen Peakende für alle drei betrachteten Peaks stets unter 30 Sekunden ist. Weiterhin verbessert sich die Abschätzung, wenn sie nicht direkt nach Erreichen des Intensitätsmaximums sondern erst drei Datenpunkte später durchgeführt wird.
Somit ist die entwickelte Methode zur Vorhersage des zeitlichen Verlaufs chromatographischer Peaks für die Praxis geeignet, da die Abweichung des abgeschätzeten Peakendes vom realen Peakende im Vergleich zur Dauer des Peaks sehr gering ist (weniger als 10%) .

Kapitel 4

Beschreibung des Moduls

Im Zuge dieser Arbeit wurde ein Modul entwickelt, welches die Vorhersage des zeitlichen Verlaufs chormatographischer Peaks für reale Messdaten durchführt. Dieses Modul wurde in Java implementiert. In diesem Kapitel wird kurz dessen Funktionsweise beschrieben.

Das Modul liest im zeitlichen Verlauf Messdaten ein, verarbeitet diese und erzeugt eine Ausgabe. Der Ablauf wird in Abbildung 4.1 skizziert.

Abbildung 4.1: Arbeitsweise des Moduls

Die drei Schritte werden im Folgenden näher beschrieben.

Schritt 1: Einlesen der Messdaten

Bei jedem durchgeführten Scan eines Messvorgangs wird je eine Datei mit Datenpunkten (m/z-Verhältnis, Intensität) von undefinierten bzw. definierten Peaks

erzeugt. Diese Datenpunkte werden in Listen eingelesen und anschließend weiter verarbeitet.

Schritt 2: Verarbeitung der Daten

Bei der Verarbeitung der zuvor eingelesenen Datenpunkte werden mehrere Aktionen durchgeführt.
Zum einen werden Datenpunkte zu vorhandenen Peaks zugeordnet bzw. es werden neue Peaks angelegt. Gehört ein Datenpunkt aus der Liste der undefinierten bzw. definierten Peaks zu einem vorhandenen Peak, wird er entsprechend hinzugefügt. Sobald ein Datenpunkt in der Liste der definierten Peaks gefunden wird, der zu keinem bereits vorhandenen Peak gehört, wird ein neuer Peak erzeugt. In diesem Fall wird in den Listen der undefinierten Peaks der vorangegangenen Scans nach weiteren Punkten dieses Peaks gesucht.
Zum anderen wird der Parameter gegebenenfalls optimiert und es wird das Peakende bestimmt. Der Parameter wird genau dann optimiert, wenn ein Datenpunkt zu einem Peak hinzugefügt wird und wenn das Maximum bereits erreicht ist. Dazu wird die Summe der Quadrate der Abweichungen (LRS-Summe) der Intensität der Datenpunkte von der Intensität der Peakfunktion minimiert. Es gibt zwei verschiedene Methoden für die Minimierung: Die erste Methode berechnet die LRS-Summe für alle möglichen Parameterwerte und bestimmt daraus das Minimum, die zweite Methode hingegen verändert den letzten bestimmten Parameterwert in eine Richtung solange, bis sich der Wert der LRS-Summe nicht mehr verringert. Ist der Parameter ermittelt, wird das Peakende berechnet.
Weiterhin wird ein Peak gegebenenfalls abgeschlossen. Dies ist der Fall, falls für eine bestimmte Anzahl von Scans kein Datenpunkt zu einem Peak zugeordnet werden kann. Wenn zu einem späteren Zeitpunkt das gleiche m/z-Verhältnis auftaucht, wird eine neuer Peak angelegt.

Schritt 3: Erzeugung der Ausgabe

In jedem Scan wird in einem nach der aktuellen Scannummer benannten Unterverzeichnis für jeden nicht beendeten Peak eine Ausgabedatei erzeugt, welche

das m/z-Verhältnis des Peaks und alle Datenpunkte (Scannummer, Zeit, Intensität) inklusive Abschätzung (aktuelle Parameter, Peakende, LRS-Summe) enthält. Am Ende des Messvorgangs wird eine Zusammenfassung generiert, in der für jeden gültigen Peak das m/z-Verhältnis und der Bereich der Scannummern, in denen der Peak auftritt, zu finden sind. Zudem wird für jeden gültigen Peak eine Datei mit allen Datenpunkten generiert. Die Kritierien für die Gültigkeit können vom Benutzer festgelegt werden. Es handelt sich dabei zum Beispiel um eine minimale Anzahl von Punkten bzw. um eine minimale Intensität des Maximums.

Das Modul wurde variabel implementiert, das heißt es können viele Werte (zum Beispiel der Schwellwert) vom Benutzer festgelegt werden. Auch die Verwendung von anderen Funktionen ist ohne Weiteres durch das Anpassen von wenigen Methoden möglich.

Eine kurze Beschreibung der Klassen des Moduls ist in Anhang B zu finden.

Kapitel 5

Zusammenfassung

In diesem Kapitel werden die Ergebnisse der Arbeit zusammengefasst, es wird deren Nutzen für die Praxis erläutert und es wird ein Ausblick auf weiterführende Arbeiten gegeben, in denen die Ergebnisse dieser Arbeit verwendet werden können.

In dieser Studienarbeit wurde eine Methode zur Vorhersage des zeitlichen Verlaufs chromatographischer Peaks entwickelt. Die ersten Ansätze dazu verwendeten einfache Funktionen, mit deren Hilfe nicht der genauen zeitliche Verlauf, sondern lediglich das Peakende abgeschätzt werden sollte. Diese Ansätze mussten allerdings wieder verworfen werden. Es stellte sich zudem heraus, dass eine Abschätzung erst nach Erreichen des Maximums des chromatographischen Peaks durchgeführt werden kann. Anschließend wurden komplexere Peakfunktionen untersucht. Über das Anpassen dieser Funktionen an die Messdaten wurden Peakfunktionen gefunden, welche die chromatographischen Peaks gut beschreiben. Allerdings lassen sich die initialen Werte derer Parameter nicht im Vorfeld bestimmen, sodass eine optimale Vorhersage erst möglich ist, nachdem alle Messdaten vorhanden sind. Dies stellt aber aufgrund der Notwendigkeit einer frühzeitigen Abschätzung keine Lösung der Problemstellung dar. Letztendlich wurde eine Peakfunktion gefunden, welche die chromatographischen Peaks gut beschreibt und für die sich die initialen Werte der Parameter im Vorfeld gut abschätzen lassen. Mit Hilfe dieser Funktion kann eine Vorhersage bereits kurz nach Erreichen des Maximums durchgeführt werden.
Weiterhin wurde ein Modul entwickelt, welches aus realen Messdaten die chromatographischen Peaks extrahiert und deren zeitlichen Verlauf direkt nach Erreichen des

Maximums abschätzt. Die Abschätzung wird nach dem Erhalt mehrerer Messpunkte weiter verfeinert.
Eine Evaluation der Methode an realen Messdaten hat gezeigt, dass die Qualität der Abschätzung für die Praxis ausreichend ist. Die Abweichung des abgeschätzten Peakendes vom realen Peakende ist sehr gering (unter 30 Sekunden für die betrachteten Peaks). Somit bleibt nach Erreichen des Intensitätsmaximums genug Zeit, um zu entscheiden, ob dieser Stoff in der nächsten Zeit untersucht werden muss.
Die Ergebnisse dieser Studienarbeit können in weiterführenden Arbeiten zur Bestimmung einer möglichst optimalen Messabfolge genutzt werden, sodass möglichst alle detektierten Stoffe mit Fragmentierungsscans genauer untersucht werden können.

Anhang A

Tabellen mit Testdaten

Dieser Bereich des Anhangs enthält Tabellen mit Testdaten.

A.1 Wert des Parameters w_2

In diese Tabelle sind die durchs Fitten ermittelten Werte des Parameters w_2 für 20 verschiedene Peaks zu finden. Die Werte wurden mit dem Werkzeug "OriginPro 7.5" von OriginLab ermittelt.

Peaknr.	1	2	3	4	5	6	7	8	9	10
w_2	0.323	0.280	0.243	0.283	0.340	0.139	0.203	0.214	0.159	0.158

	11	12	13	14	15	16	17	18	19	20
	0.150	0.181	0.253	0.229	0.217	0.130	0.208	0.259	0.296	0.309

Tabelle A.1: Wert des Parameters w_2

Es wird deutlich, dass der Wert von w_2 zwischen den Werten 0.13 und 0.34 liegt.

A.2 Abweichung des Funktionsfittings

In der folgenden Tabelle ist die Abweichung des durch das Funktionsfitting ermittelten Peakendes vom realen Peakende zu finden und es wird die Verweilzeit des Peaks angegeben. Als Schwellwert wird dabei der Intensitätswert 10000 benutzt, da dieser in der Praxis als realistisch angesehen wird. Die Zeiten sind in Minuten angegeben.

Peaknr.	1	2	3	4	5	6	7
Dauer	6.87	0.68	3.42	8.12	7.45	1.13	1.91
Ende Fit	40.46	29.91	37.34	50.22	57.20	28.34	26.62
Ende Real	40.37	29.93	37.32	50.36	57.60	28.35	26.65
Abweichung	0.09	0.02	0.02	0.14	0.40	0.03	0.03

	8	9	10	11	12	13	14
	3.55	1.56	0.56	0.52	0.92	0.10	4.45
	33.84	33.41	51.21	51.15	51.16	50.92	44.87
	33.67	33.47	51.25	51.16	51.24	50.92	44.96
	0.14	0.06	0.04	0.01	0.08	0.00	0.09

	15	16	17	18	19	20	
	4.69	3.49	2.77	14.92	12.8	13.66	
	44.71	43.65	43.87	40.22	40.23	40.63	
	44.99	44.04	43.79	41.20	41.03	41.26	
	0.18	0.39	0.08	0.98	0.80	0.63	

Tabelle A.2: Abweichung des Funktionsfittings

Es wird deutlich, dass sich die Abweichung für kurzweilige Peaks, die weniger als 3 Minuten andauern, im Sekundenbereich bewegt. Für länger andauernde Peaks (mehr als 10 Minuten) kann die Abweichung auch 30 Sekunden betragen, allerdings ist dieser Sachverhalt aufgrund der längeren Verweildauer nicht relevant, da trotzdem genug Zeit bleibt, den zugehörigen Stoff genauer zu untersuchen.

Anhang B

Kurzbeschreibung der Klassen des Moduls

Im Folgenden gibt es eine kurze Beschreibung der Klassen des Moduls. Es wurde in der Programmiersprache Java implementiert.

- **Mainprog.java:** Dies ist die Hauptklasse des Moduls. Es werden die Dateien mit den definierten und undefinierten Peaks nacheinander eingelesen. Dabei werden die Punkte den entsprechenden Peaks zugeordnet und es wird die Ausgabe generiert.
- **PeakReader.java:** Diese Klasse ist für das einlesen der Daten aus den Datendateien verantwortlich.
- **DataPoint.java:** Diese Klasse repräsentiert einen Datenpunkt, welcher aus m/z-Verhältnis, Zeit, Intensität und Scannummer besteht.
- **Peak.java:** Diese Klasse repräsentiert einen Peak. Es werden Methoden zur Verfügung gestellt, welche prüfen, ob ein Datenpunkt zum aktuellen Peak gehört. Ist dies der Fall, kann der Datenpunkt zum Peak hinzugefügt werden. Weiterhin wird die Anpassung des Parameters vorgenommen und es wird entschieden, ob ein Peak beendet wird bzw. ob ein Peak gültig ist.
- **Constants.java:** In dieser Klasse sind alle Konstanten enthalten, welche vom Benutzer angepasst werden können.

Literaturverzeichnis

[Li02] Jianwei Li. Comparison of the capability of peak functions in describing real chromatographic peaks. *Journal of Chromatography A*, 952:63–70, 2002.

[Sch03] Dr. Karsten Schürrle. Proteomforschung - Die Werkzeuge des Lebens nutzen, Bundesministerium für Bildung und Forschung, BMBF Publik. Juli 2003.

[SVR00] A. G. Stromberg S. V. Romanenko. Classification of mathematical models of peak-shaped analytical signals. *Journal of Analytical Chemistry*, 55(11):1024–1028, 2000.

[TLP01] Zs. Papai T. L. Pap. Application of a new mathematical function for describing chromatograhic peaks. *Journal of Chromatography A*, 930:53–60, 2001.

[VBDM01] G. Giorgio Bombi Valerio B. Di Marco. Mathematical functions for the representation of chromatographic peaks. *Journal of Chromatography A*, 931:1–30, 2001.

[ZP02] T. L. Pap Zs. Papai. Analysis of peak asymmetry in chromatography. *Journal of Chromatography A*, 953:31–38, 2002.

Selbständigkeitserklärung

Hiermit erkläre ich, dass ich die vorliegende Arbeit mit dem Titel *"Entwicklung einer Methode zur Vorhersage des zeitlichen Verlaufs chromatographischer Peaks"* selbständig und ohne unerlaubte Hilfe verfasst habe.

Berlin, den 17. Dezember 2008

Benjamin Daeumlich

Einverständniserklärung

Ich bin damit einverstanden, dass die vorliegende Arbeit mit dem Titel *"Entwicklung einer Methode zur Vorhersage des zeitlichen Verlaufs chromatographischer Peaks"* in der Bibliothek ausgelegt wird.

Berlin, den 17. Dezember 2008

Benjamin Daeumlich