

Diplomarbeit

Simulation einer Multikapillarsäule

Elisabeth Böhmer

30. April 2015

Betreuer:

Prof. Dr. Sven Rahmann

Prof. Dr. Jörg Rahnenführer

Fakultät für Informatik

Algorithm Engineering (Ls11)

Technische Universität Dortmund

<http://ls11-www.cs.tu-dortmund.de>

Inhaltsverzeichnis

1	Einleitung	1
1.1	Motivation und Hintergrund	1
1.2	Aufbau der Arbeit	1
1.3	Ziel der Arbeit	2
2	Grundlagen	3
2.1	Chromatographie	3
2.1.1	Der chromatographische Prozess	3
2.1.2	Detektion	3
2.1.3	Technische Daten	4
2.1.4	Peakcharakteristika	5
2.2	Probabilistische Arithmetische Automaten	5
2.2.1	Definition eines PAA	6
2.2.2	Algorithmus zur Berechnung der Wartezeit	6
3	Modell	7
3.1	2-Parameter Modell	7
3.1.1	PAA zu diesem Modell	8
3.1.2	Grenzen des Modells	8
3.2	Weitere Modelle	9
4	Implementierung	11
4.1	Simulation des 2-Parameter-Modells	11
4.1.1	Festlegung der Einheiten	11
4.1.2	Step-by-step	11
4.1.3	By-Event	11
4.2	PAA	11
5	Evaluation	13
6	Diskussion	15

A Weitere Informationen	17
Abbildungsverzeichnis	19
Algorithmenverzeichnis	21
Literaturverzeichnis	23

Kapitel 1

Einleitung

1.1 Motivation und Hintergrund

“Multikapillarsäule”, MCC (engl. Multi Capillary Column)

Trennsäule in der Gaschromatographie

“Simulation”

Keine physikalische Simulation der Moleküle

Keine Interpolation vorhandener Messungen

Kein Überlagern verschiedener Kurven, um Gesamt-Spektrum zu erhalten
sondern: Modell für chromatographischen Prozess

1.2 Aufbau der Arbeit

Im Rahmen der Diplomarbeit soll eine Multikapillarsäule simuliert werden.

Es geht dabei nicht um eine physikalische Simulation auf molekularer Ebene, sondern um die Entwicklung eines abstrakten, probabilistischen Modells, welches mit möglichst wenig Parametern auskommt.

In der Simulation soll jeweils nur ein Stoff, charakterisiert durch diese Parameter, simuliert werden. Das Resultat einer Simulation ist jeweils ein einzelner Peak, welcher durch seine Lage und Form beschrieben werden kann. Die Lage entspricht dabei der Retentionszeit am Maximum des Peaks. Die Form ist durch seine Breite an einer bestimmten Höhe (zum Beispiel Halbwertsbreite) sowie Schiefe gekennzeichnet.

Umgekehrt soll es möglich sein, für eine gegebene Peaklage und Peakform die nötigen Simulationsparameter zu ermitteln, mit denen ein solcher Peak simuliert werden kann.

Zum Vergleich liegen einige MCC-IMS-Messungen von Mischungen sieben bekannter Stoffe vor. Diese Datensätze wurden zur Verfügung gestellt von der Firma B & S Analytik (<http://www.bs-analytik.de/>). Zu den dadurch gegebenen Peaks sollen Parameter für die Simulation ermittelt und diese Peaks simuliert werden.

Gesucht ist letztendlich eine allgemeine Entsprechung der Simulationsparameter zu den Parametern mit denen ein Peak beschrieben werden kann, falls diese existiert. Eine Berechnung des Peaks nur durch die Simulationsparameter und umgekehrt eine Vorhersage der Simulationsparameter zu bekannten Peakdaten soll damit realisiert werden.

In Kapitel 2 wird zu Einem die Chromatographie, insbesondere die Gaschromatographie vorgestellt und der chromatographische Prozess erklärt. Zum Anderen PAA

1.3 Ziel der Arbeit

Ziel dieser Diplomarbeit ist es, eine Entsprechung von Peakcharakteristika, wie sie in Kapitel 2 beschrieben werden, zu den Simulationsparametern, die für das endgültige Modell verwendet werden, zu finden, falls es eine solche Entsprechung gibt. Das heißt es soll sowohl möglich sein, zu einem gegebenen Peak die Parameter bestimmen, mit denen eben jener Peak simuliert werden kann. Umgekehrt soll zu gegebenen Simulationsparametern, der dadurch erzeugte Peak vorhergesagt werden können.

Um dieses Ziel zu erreichen, wird zunächst mit einem sehr einfachen Modell begonnen, welches nach und nach angepasst werden muss, bis die durch die Referenzdatensätze vorgegebenen Peaks ausreichend angenähert werden können.

1. Start mit einfachem 2-Parameter-Modell
2. Experimentelle Arbeit: Simulation mit verschiedenen Parametern
3. Überprüfung, ob Referenzpeaks angenähert werden können
4. Verfeinerung/Erweiterung des Modells
5. Wiederholung von 2-4 bis gegebene Peaks ausreichend angenähert

Kapitel 2

Grundlagen

2.1 Chromatographie

Die Chromatographie ist ein Verfahren zur Auftrennung von Stoffgemischen. Die Auftrennung erfolgt dabei zwischen zwei sogenannten Phasen, der stationären und der mobilen Phase, welche sich in unterschiedlichen Aggregatzuständen befinden und untereinander nicht mischen.

Es existieren zum einen die Flüssigchromatographie (LC, engl. Liquid Chromatography), bei der die mobile Phase eine Flüssigkeit ist und die stationäre Phase ein Feststoff. Bei der Gaschromatographie (GC) ist die mobile Phase ein Gas und es wird zusätzlich nach der stationären Phase unterschieden. Ist diese ein Feststoff, so spricht man von gepackten Säulen. Bei der Kapillartechnik hingegen werden die Trennsäulen innen mit einem Flüssigkeitsfilm als stationäre Phase beschichtet.

Beispielsweise kann die GC in einer Multikapillarsäule (MCC, engl. Multi Capillary Colum) stattfinden. Sie besteht aus ca. 1000 einzelnen Kapillaren. Jede davon ist innen mit der sog. stationären Phase beschichtet. Außerdem kommt ein Trägergas, die sog. mobile Phase, zum Einsatz, welches die Analyte durch die Säule transportiert.

2.1.1 Der chromatographische Prozess

Die Substanzen unterscheiden sich vor allem durch ihre Wechselwirkungen mit der stationären Phase. Während dieser Wechselwirkungen haften die Teilchen an der stationären Phase, bewegen sich also nicht fort. Finden wenig Wechselwirkungen statt, passieren die Teilchen die Säule schneller, als wenn viele Wechselwirkungen stattfinden. Dies beeinflusst die Retentionszeit, also die Zeit, die zum Durchlaufen der Säule gebraucht wird.

2.1.2 Detektion

Nach Durchlaufen der Säule wird detektiert, welche Menge an Substanzen austreten. Dabei wird nicht die Art des Analyts festgestellt, sondern nur die Menge der zum jeweiligen Zeit-

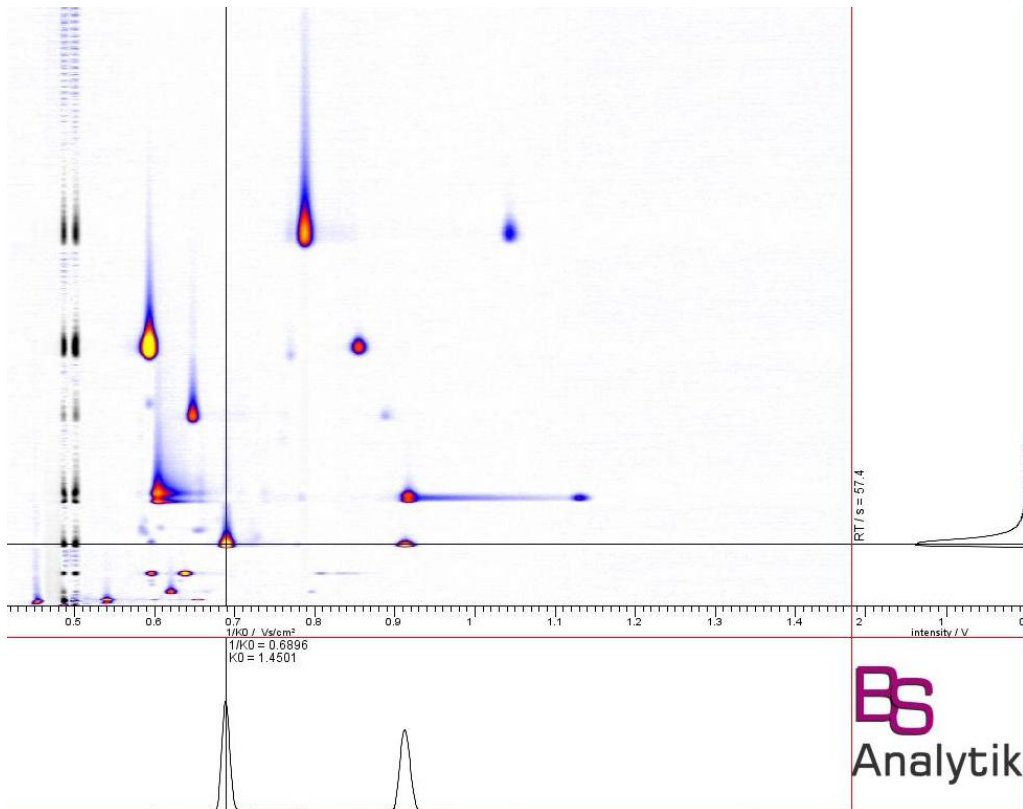


Abbildung 2.1: Spektrum einer MCC-IMS-Messung

punkt austretenden Stoffe. Dabei zeichnet der Detektor ein Chromatogramm auf, welches beispielhaft in zu sehen ist

Alternativ kann die Gaschromatographie auch als Vorverarbeitung für Verfahren wie Massenspektrometrie (MS) oder Ionen-Mobilitäts-Spektrometrie (IMS) dienen. In diesen Fällen werden die aus der MCC austretenden Moleküle direkt ionisiert und in den entsprechenden Geräten weiter analysiert.

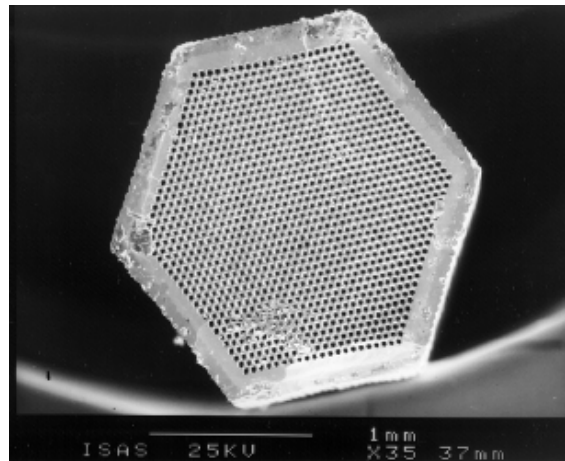
Die für diese Arbeit vorliegenden Referenzdatensätze stammen aus einer solchen MCC-IMS-Kopplung. Ein Beispiel dafür ist in 2.1 zu sehen. Zu beobachten ist, dass schnelle Teilchen Peaks zu frühen Zeitpunkten erzeugen, die eine relativ geringe Varianz aufweisen, hingegen spätere Peaks tendenziell breiter werden. Ideale Peaks haben die Form einer Gaußkurve.

2.1.3 Technische Daten

Eine MCC besteht aus ca. 1000 Kapillaren mit je

- 20 – 80 μm Durchmesser
- Stationäre Phase ist Flüssigkeitsfilm, ca. 0,1 – 0,8 μm dick

→ MCC etwa 2 – 6 mm dick und 20 cm lang

Abbildung 2.2: Querschnitt einer MCC ¹

2.1.4 Peakcharakteristika

Um später die simulierten Peaks mit denen aus den Referenzdatensätzen gut vergleichen zu können, sollen nun einige Eigenschaften aufgeführt werden, die einen Peak beschreiben.

Die offensichtlichste Eigenschaft ist die Lage des Peaks, genauer gesagt, die Lage des Maximums des Peaks.

Außerdem können die Peaks verschiedene Formen haben. Im theoretischen Idealfall haben sie die Form einer Gaußkurve, jedoch tritt oft ein Tailing auf, welches verschiedene Ursachen haben kann. Unter Anderen seien hier genannt: zusätzliche Adsorptionseffekte, die beim Altern einer Säule auftreten [1], technische Ursachen wie kleinste Hohlräume zwischen der Säule und dem Gaseinlass bzw. -Austritt, sowie einige Stoffe, die generell zu Tailing neigen. Beim Tailing steigt die Kurve zunächst stark an, sinkt jedoch nach Erreichen des Maximalwerts deutlich langsamer ab, es entsteht ein Schwanz (engl. Tail). Auch der umgekehrte Fall, Fronting genannt, kann auftreten, ist jedoch in den vorliegenden Datensätzen nicht zu beobachten.

Als drittes können Peaks unterschiedlich breit sein. Da jedoch ein Peak mit größerer Intensität, also möglicherweise größerer Stoffmenge, automatisch auch einen breiteren Peak erzeugt, betrachten wir die Halbwertsbreite als Maß für die Peakbreite. Die Halbwertsbreite wird auf halber Maximalhöhe des Peaks gemessen. Im Fall von auftretendem Tailing oder Fronting ist es sinnvoll, je einen Wert für rechts und links des Maximalwerts zu berechnen.

2.2 Probabilistische Arithmetische Automaten

Ein Probabilistischer Arithmetischer Automat (PAA) nach [2] ist ein Modell, mit dem eine Folge zufälliger Operationen beschrieben werden kann. Für PAA existieren Algorithmen,

¹<http://yas.yanaco.co.jp/products/import-gc-ims.html>

welche eine gemeinsame Verteilung von Zuständen und Werten oder auch die Verteilung der Wartezeit für einen Wert berechnen. Wie in Kapitel 3 beschrieben wird, kann das Modell zur Simulation einer Multikapillarsäule auch als PAA formuliert werden. Mit dieser Formulierung ist die Zeit, die zum Durchlaufen einer Säule gebraucht wird, dann die Wartezeit für den Wert, welcher der Länge der Säule entspricht. Deshalb kann ein PAA nützlich sein, um neben der eigentlichen Simulation auch noch eine erwartete Verteilung der Ankunftszeiten der Teilchen zu berechnen.

Zunächst sei hier eine Definition für den PAA gegeben, anschließend wird der Algorithmus zur Berechnung der Wartezeit beschrieben.

2.2.1 Definition eines PAA

2.2.1 Definition (PAA). Ein Probabilistischer Arithmetischer Automat (PAA) ist ein Tupel $\mathcal{P} = (\mathcal{Q}, q_0, T, \mathcal{V}, v_0, \mathcal{E}, (e_q)_{q \in \mathcal{Q}}, (\theta_q)_{q \in \mathcal{Q}})$, dabei ist:

- \mathcal{Q} eine endliche Menge von Zuständen
- $q_0 \in \mathcal{Q}$ der Startzustand
- $T : \mathcal{Q} \times \mathcal{Q} \rightarrow [0,1]$ eine Übergangsfunktion mit $\sum_{q' \in \mathcal{Q}} T(q, q') = 1$ das heißt $(T(q, q'))_{q, q' \in \mathcal{Q}}$ ist eine stochastische Matrix
- \mathcal{E} eine endliche Menge von Emissionen
- $e_q : \mathcal{E} \rightarrow [0,1]$ eine Wahrscheinlichkeitsverteilung der Emissionen für jeden Zustand
- \mathcal{V} eine Menge von Werten
- v_0 der Startwert
- $\theta_q : \mathcal{V} \times \mathcal{E} \rightarrow \mathcal{V}$ eine Operation für jeden Zustand

Dabei entspricht $\mathcal{P} = (\mathcal{Q}, q_0, T)$ einer Markovkette und $\mathcal{P} = (\mathcal{Q}, q_0, T, \mathcal{E}, (e_q)_{q \in \mathcal{Q}})$ einem Hidden Markov Model.

2.2.2 Algorithmus zur Berechnung der Wartezeit

Kapitel 3

Modell

3.1 2-Parameter Modell

Als Grundlage für die Entwicklung eines Simulationsmodells dient die Beobachtung des chromatographischen Prozesses. Dieser ist wie in Kapitel 2 beschrieben gekennzeichnet durch viele Teilchen, welche häufig zwischen der stationären und mobilen Phase wechseln. Es liegt also nahe, dass im Modell ebenfalls Teilchen simuliert werden, die zwischen zwei Zuständen, welche die beiden Phasen repräsentieren, wechseln. Der Phasenwechsel geschieht jeweils mit einer bestimmten Wahrscheinlichkeit. Dabei kann es möglich sein, dass die Wahrscheinlichkeiten für den Wechsel in die eine oder andere Richtung voneinander abhängig sind. Da dies jedoch nicht sein muss, wird für das Modell der allgemeinere Fall von unabhängigen Wahrscheinlichkeiten angenommen. Es sei also p_s die Wahrscheinlichkeit, dass ein Teilchen, welches sich bereits in der stationären Phase befindet, auch stationär bleibt und $1 - p_s$ die Wahrscheinlichkeit, dass es in die mobile Phase übergeht. Analog seien p_m und $1 - p_m$ die Wahrscheinlichkeiten, dass ein mobiles Teilchen in der mobilen Phase bleibt bzw. zur stationären Phase wechselt.

Damit ergibt sich als Modell ein einfacher Automat mit zwei Zuständen $\mathcal{Q} = \{s, m\}$ und den oben beschriebenen Transitionen. Eine graphische Darstellung des Modells zeigt 3.1

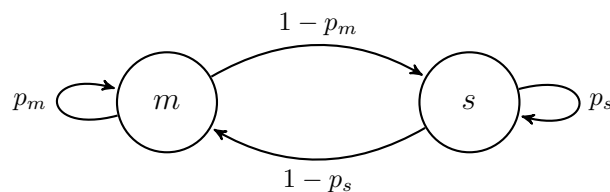


Abbildung 3.1: Graphische Darstellung des 2-Parameter Modells

Für die Simulation müssen viele Teilchen, ausgehend vom mobilen Zustand, den Automaten durchlaufen. Dabei wird zusätzlich zum Zustand der Teilchen auch der Ort, an

dem sie sich befinden, verwaltet. Wenn sich ein Teilchen im mobilen Zustand befindet, wird dieser Ortszähler erhöht. Außerdem werden die insgesamt benötigten Schritte gezählt. Dadurch leiten sich dann die Ankunftszeiten der Teilchen bei Erreichen eines bestimmten Wertes des Ortszählers ab.

Der genaue Ablauf der Simulation wird in Kapitel 4 beschrieben.

3.1.1 PAA zu diesem Modell

Blabla...

Der PAA ist formal definiert durch: $\mathcal{Q} = \{s, m\}$, $q_0 = m$, $T = \begin{pmatrix} p_m & 1 - p_m \\ 1 - p_s & p_s \end{pmatrix}$, $\mathcal{E} = \{0, 1\}$, $e_s(0) = 1, e_s(1) = 0, e_m(0) = 0, e_m(1) = 1$, $\mathcal{V} = [0, \dots, l]$, $v_0 = 0$, $\theta_s = \theta_m = +$

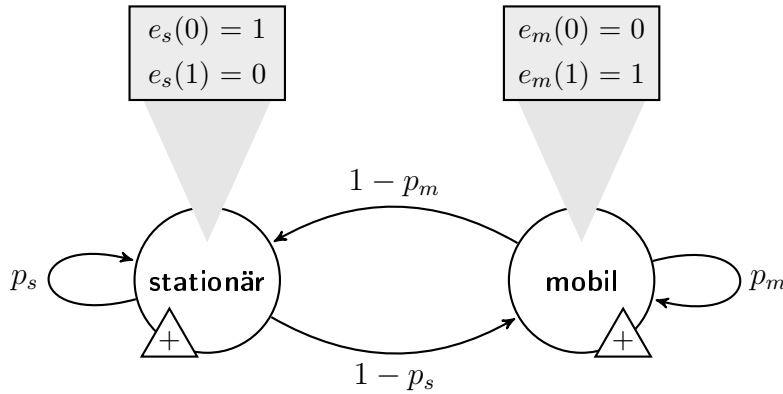


Abbildung 3.2: PAA für das 2-Parameter Modell

Und noch mehr Text, der sich auf 3.2 bezieht

3.1.2 Grenzen des Modells

Eine genaue Analyse der Peaks, die mit dem 2-Parameter Modell erzeugt werden können, findet sich in Kapitel 5. An dieser Stelle sei nur vorweg genommen, dass es zwei Hauptprobleme mit dem Modell zu geben scheint.

Zum Einen scheint es in den vorliegenden Datensätzen einige Peaks zu geben, deren kombinierte Lage und Breite nicht mit dem Modell simulierbar sind. Es ist zwar sowohl in den echten wie auch in den simulierten Daten die klare Tendenz zu erkennen, dass Peaks, die zu einer höheren Retentionszeit auftreten, auch breiter sind, als frühe Peaks. Es gibt also keine sehr schmalen Peaks am Ende des Spektrums. Allerdings existieren einige Peaks, deren Breite geringer ist, als alle mit dem Modell simulierbaren Peaks zu dieser Retentionszeit.

Zum Anderen weisen fast keine der simulierten Peaks ein Tailing auf. Lediglich ein sehr stark eingeschränkter Parameterbereich erzeugt ein Tailing. Das führt dazu, dass Tailing

nur in einem sehr kleinen Retentionszeitbereich auftritt, allerdings nicht, wie es in den Referenzdaten zu beobachten ist, zu den eher späten Zeiten.

3.2 Weitere Modelle

Bisher fand keine Unterscheidung zwischen Adsorption und Lösung der Teilchen an bzw. in der stationären Phase statt. In der Realität können beide Wechselwirkungen parallel zueinander statt finden. Außerdem ist es möglich, dass sich die Wahrscheinlichkeiten, in einen der beiden Zustände überzugehen oder in die mobile Phase zurückzukehren, unterscheiden. Daher liegt es nahe, einen neuen Zustand einzuführen, sodass die Adsorption und die Lösung voneinander getrennt behandelt werden. Die Tatsache, dass Tailing, wie bereits anfangs erwähnt, unter anderem durch zusätzlich Adsorptionseffekte verursacht sein kann, lässt vermuten, dass durch eine Anpassung der Parameter ein Tailing in der Simulation verursacht werden kann.

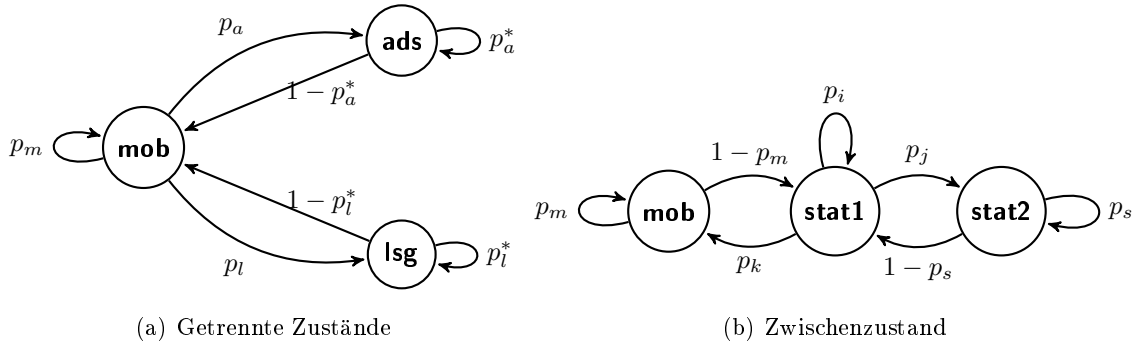


Abbildung 3.3: Mögliche weitere Modelle

In 3.3 sind zwei Möglichkeiten gezeigt, dem bisherigen Modell einen neuen Zustand hinzuzufügen.

Eine weitere Möglichkeit, das Modell zu verändern, besteht darin, das Gleichgewicht, welches sich zwischen den beiden Phasen aufbaut, zu berücksichtigen. Damit müssten die Wahrscheinlichkeiten, den Zustand zu wechseln, nicht mehr fest vorgegeben sein, sondern sich dynamisch während der Simulation aus der aktuellen Verteilung der Teilchen an einem bestimmten Ort auf die Zustände berechnen.

Außerdem existieren Idee, noch eine Sättigung der freien Plätze zur Adsorption einzubauen

Kapitel 4

Implementierung

4.1 Simulation des 2-Parameter-Modells

4.1.1 Festlegung der Einheiten

20 cm entspricht 200000 Schritten, in die MCC unterteilt wird. 1 Schritt ist damit 1 μm Trärgas hat nach etwa 0,1 Sekunden die Säule durchlaufen. Eine Zeiteinheit entspricht demnach ca. $0.1/200000$ Sekunden 0.0000005s 0.0005ms 0,5 μs

4.1.2 Step-by-step

Es kann entweder jeder Zeitschritt der Chromatographie simuliert und dabei für jedes Teilchen entschieden werden, ob es sich bewegt. Dabei angenommen, dass die Wahrscheinlichkeit der Wechselwirkung unabhängig vom aktuellen Ort oder dem Zeitpunkt des letzten Phasenwechsels ist. Es wird für jedes Teilchen festgehalten, wo es sich befindet und ähnlich wie bei der echten Chromatographie beobachtet, wann wie viele Teilchen die Säule durchquert haben.

4.1.3 By-Event

Bei Übergangswahrscheinlichkeiten, die sehr groß oder klein sind, kommt es selten zu Zustandsänderungen, sodass es in diesem Fall effizienter ist, eine Wartezeitmethode zu nutzen: Es wird für jedes Teilchen entschieden, wann es den Zustand wechselt und, falls es mobil ist, wie weit es bis dahin weiter wandert. Es werden also nur die tatsächlichen Ereignisse simuliert, Schritte für Zeitpunkte, zu denen nichts passiert, entfallen.

4.2 PAA

Die Berechnung des PAA in Java, dabei Referenz auf mosdi (Zitiere Paper, in dem das vorgestellt wird und sage, dass ich das verwende)

Kapitel 5

Evaluation

Kapitel 6

Diskussion

Anhang A

Weitere Informationen

Abbildungsverzeichnis

2.1	Spektrum einer MCC-IMS-Messung	4
2.2	Querschnitt einer MCC	5
3.1	Graphische Darstellung des 2-Parameter Modells	7
3.2	PAA für das 2-Parameter Modell	8
3.3	Mögliche weitere Modelle	9

Algorithmenverzeichnis

Literaturverzeichnis

- [1] KOLB, BRUNO: *Gaschromatographie in Bildern [eine Einführung]*. Wiley-VCH, Weinheim, 2., überarbeitete und erweiterte Auflage, 2003.
- [2] MARSCHALL, TOBIAS, INKE HERMS, HANS-MICHAEL KALTENBACH und SVEN RAHMANN: *Probabilistic Arithmetic Automata and Their Applications*. IEEE/ACM Trans. Comput. Biol. Bioinformatics, 9(6):1737–1750, November 2012.