technische universität
dortmund

Statistische Analyse und Modellierung von Clusterphänomenen bei Signalproteinen in der Plasmamembran

Dissertation

zur Erlangung des akademischen Grades

Doktor der Naturwissenschaften

im Studiengang Statistik

der Technischen Universität Dortmund

Vorgelegt von: Sabrina Siebert geb. Herrmann

Dortmund, Oktober 2016

Erstgutachterin: Prof. Dr. Katja Ickstadt

Zweitgutachter: Prof. Dr. Jörg Rahnenführer

Zusammenfassung

In der vorliegenden Arbeit wurde sich mit Clusterphänomenen von Signalproteinen beschäftigt. Diese Proteine sind dabei in der Plasmamembran lokalisiert und für die Kommunikation und den Stoffaustausch der Zelle zuständig. Die Daten wurden mit Hilfe von Fluoreszenzmikroskopie am Max-Planck-Institut für molekulare Physiologie in Dortmund in der Arbeitsgruppe von Dr. Peter J. Verveer erhoben.

Um die Clusterphänomene zu untersuchen, können unterschiedliche Blickwinkel und Fragestellungen betrachtet werden. In dieser Arbeit wurde eine zeitliche, eine räumliche und eine zeitlich-räumliche Analyse entsprechender Daten vorgenommen.

In der zeitlichen Analyse wurden Proteinzeitreihen untersucht. Die Proteinzeitreihe ergibt sich aus der Messung der Lichtintensität eines Spots, d.h. eines Proteinclusters, über die Zeit hinweg. Das Ziel ist hier die Segmentierung eben dieser Proteinzeitreihe. Hier wurde ein Bayessches hierarchisches Modell zur Segmentierung genutzt. Dieses lieferte dabei sinnvolle Ergebnisse, wobei jedoch zu beachten war, dass die Anzahl an Segmenten stets als fest angesehen wurde. Um diese Einschränkung aufzuheben, wurde ein Reversible Jump Schritt in das Modell aufgenommen. Mit dieser Erweiterung konnten nun sinnvolle Ergebnisse mit einer höheren Flexibilität für den Anwender erreicht werden.

In der räumlichen Analyse wurde ein Pixelbild aus einer Messung einer lebenden Zelle mit Hilfe von TIRF-Mikroskopie untersucht. Ziel war hier die räumliche Clusterstruktur zu untersuchen, wobei sich auf den Anteil an Proteinen in Clustern beschränkt wurde. Dafür wurden zunächst unterschiedliche Methoden auf einer simulierten Region untersucht. Mit Hilfe dieser Ergebnisse konnte ein Anwendungsschema zur effizienten Kombination eben dieser Methoden aufgestellt werden. Dieses wurde abschließend auf einen experimentellen Datensatz sowie auf eine Dual Colour Simulation angewendet. Es zeigte sich, dass durch das Vorgehen des Schemas die Parameterwahl für einige Methoden vereinfacht wurde und sinnvolle Ergebnisse berechnet werden konnten.

Abschließend wurden in der räumlich-zeitlichen Analyse Proteintracks untersucht. Diese Proteintracks geben den Weg eines Proteins in der Zellmembran über die Zeit hinweg an. Diese Messung wurde simultan für zwei Proteinarten durchgeführt, sodass hier erneut der Dual Colour Fall vorliegt. Ziel war die Bestimmung von Zusammenhängen zweier Proteintracks unterschiedlicher Proteinarten. Um den Zusammenhang bestmöglich berechnen zu können, wurde zunächst diskutiert, welche Eigenschaften einen hohen Zusammenhang repräsentieren. Anschließend wurden diese Eigenschaften zu einem Zusammenhangsmaß zusammen gefügt. Mit diesem Maß wurden zum einen ein simuliertes Beispiel und zum anderen experimentelle Daten analysiert. Es zeigte sich, dass Abhängigkeitsstrukturen durch das Maß gut widergespiegelt wurden und mit Hilfe von Cutoffs eine Auswahl entsprechender Proteintracks erfolgen konnte. Durch diese Auswahl konnten weiter interessante Regionen sowie Cluster identifiziert werden.

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

1. Einleitung

Der menschliche Körper besteht aus vielen verschiedenen Bauteilen und kann somit in viele unterschiedliche Ebenen unterteilt werden. Die menschliche Zelle ist dabei ein wichtiger Grundbaustein. Sie enthält neben vielen verschiedenen Organellen auch die menschliche DNA. Die DNA ist Träger der Erbinformation, der Gene, und somit der Baustein, welcher für das Aussehen und den Aufbau des Menschen bzw. seines Körpers zuständig ist. Somit ist die DNA, und damit auch die Zelle, ein Dreh- und Angelpunkt in der menschlichen Biologie.

Daher ist es nicht verwunderlich, dass Forscher großes Interesse daran haben, die Zelle und die DNA sowie die damit verbundenen Prozesse zu verstehen. Ein wichtiger Prozess ist die Transkription, bei der die DNA übersetzt wird. Das bedeutet, dass die Information aus den Genen nun in verschiedenste andere Bausteine übersetzt wird, wie zum Beispiel Proteine.

Proteine sind dabei ein weiterer wichtiger Bestandteil der menschlichen Zelle und können bis zu 50% des Trockengewichtes einer Zelle ausmachen. Ein Mensch hat dabei *„Zehntausende verschiedene Arten von Proteinen“* (Campbell und Reece, 2006, S. 85, [5]), welche auch eine Vielzahl an Aufgaben im menschlichen Körper übernehmen. Diese Aufgaben reichen dabei vom Transport bestimmter Stoffe bis hin zur Abwehr von Krankheiten.

Ein Beispiel für ein Transportprotein wäre zum Beispiel das Hämoglobin, welches Sauerstoff im Körper transportiert. Antikörper hingegen sind ein gutes Beispiel für Abwehrproteine. Eine weitere wichtige Aufgabe ist die Regulierung von wichtigen biologischen Prozessen, wie z.B. der Kontrolle des Zellwachstums. In vielzelligen Organismen sind dafür Ras-Proteine verantwortlich (Reents et al., 2004, [39]).

Da Proteine über die DNA kodiert werden, wirken sich auch Mutationen auf sie und somit auch auf ihre Funktionalität aus. Je nach Protein und seiner Funktion kann es unterschiedlichste Folgen haben. Im Falle von Ras-Proteinen kann es zu einer unkontrollierten Vermehrung kommen.

Aus diesem Grund ist es sehr bedeutsam Proteine zu untersuchen und zu erforschen. Dies kann – dank moderner Mikroskopie – auch an lebenden Zellen erfolgen, wodurch die Proteine in ihrem natürlichen Ablauf beobachtet und somit auch im zeitlichen Kontext

betrachtet werden können.

In dieser Arbeit werden derartige Proteindaten analysiert. Dabei wird neben einer zeitlichen und einer räumlichen, auch eine Analyse im Hinblick auf die Kombination der vorangegangenen Aspekte erfolgen, d.h. eine räumlich-zeitliche Analyse. Dazu werden für jeden Aspekt passende Methoden gesucht bzw. entwickelt, um sie anschließend für eine Analyse der entsprechenden Daten anzuwenden.

Für eine zeitliche Analyse wird die Lichtintensität eines Proteinspots in einer lebenden Zelle im Verlauf der Zeit gemessen, wodurch eine sogenannte Proteinzeitreihe entsteht. Da mit jedem Protein, welches in einen aktiven oder inaktiven Zustand wechselt, ein Anstieg bzw. ein Abfall der Lichtintensität zu erwarten ist, ist es sinnvoll, eine stückweise konstante Funktion an die Proteinzeitreihe anzupassen. Über diese Segmente kann anschließend die Anzahl der Proteine geschätzt werden. Um nun eine stückweise Funktion anpassen zu können, ist es sinnvoll die Changepoints, d.h. die „Sprünge“ der Lichtintensitäten, zu finden. Dies kann mit Hilfe von Changepoint-Methoden erfolgen, auf welche auch hier zurückgegriffen wird. Genauer wird hier ein Bayessches hierarchisches Modell vorgestellt, welches auch für die vorliegenden Proteindaten adaptiert werden kann.

In der räumlichen Analyse soll das Clusterverhalten bzw. bestimmt Clustercharackteristika in der Zelle untersucht werden. Im Vordergrund steht hier dabei der Anteil an Proteinen, welche in einem Cluster enthalten sind. Dafür werden zunächst verschiedene räumliche (Cluster-)Methoden in einer Simulationsstudie verglichen. Durch diese Untersuchung kann eine effiziente Kombination der unterschiedlichen Methoden herausgearbeitet werden, welche anschließend auf experimentellen Daten sowie auf einer Dual Colour Simulation untersucht wird.

In der räumlich-zeitlichen Analyse werden schließlich Proteine in ihrer Bewegung über die Zeit hinweg beobachtet. Die dadurch entstehenden Proteintracks sollen hinsichtlich ihres Zusammenhangs untersucht werden. Um diesen Zusammenhang möglichst gut beschreiben zu können, wird zunächst betrachtet, welche Charakteristika diesen widerspiegeln. Anschließend wird mit Hilfe dieser Eigenschaften ein entsprechendes Maß entwickelt, welches auf einem simulierten Beispiel evaluiert werden kann. Abschließend wird das Maß auch auf experimentelle Daten angewendet.

Im folgenden Kapitel wird nun zunächst die allgemeine Problemstellung erläutert. An-

schließend folgen zunächst die zeitliche, dann die räumliche und abschließend die räumlich-zeitliche Analyse. Dafür werden in den jeweiligen Kapiteln nochmals die zugehörigen Daten, so wie die Methoden beschrieben. Abschließend erfolgt in jedem Kapitel die Analyse der Daten mit Hilfe der zuvor erläuterten Methoden. In Kapitel 6 erfolgt eine Zusammenfassung der Ergebnisse, sowie eine Diskussion mit möglichen weiteren Ansätzen bzw. Analyseschritten.

2. Biologische und Statistische Problemstellung

In diesem Kapitel soll nun zunächst kurz der biologische Hintergrund vorgestellt werden. Daraus soll anschließend die allgemeine Fragestellung sowie die Formulierung des Zieles dieser Arbeit erarbeitet werden.

2.1. Die Zelle und Proteine

Ein Grundbaustein des menschlichen Körpers ist die Zelle. Sie ist durch eine Zellmembran von ihrer Umgebung abgegrenzt. Neben dem Zytoplasma und den darin enthaltenen Organellen, enthält sie auch den Zellkern (Nukleus). Der Nukleus ist dabei der Ort, an dem unsere Erbinformationen, die DNA *(Desoxyribonukleinsäure)*, lokalisiert ist. Er befindet sich meist im Zellzentrum und ist durch eine Doppelmembran vom Zytoplasma getrennt, wobei stets ein (Informations-)Austausch zwischen dem Nukleus und der Zelle besteht. Dies ist auch auf dem Schaubild in Abbildung 2.1 aus Schäffler und Menche (2000, S. 31, [45]) zu sehen. Die DNA hat eine bedeutende Rolle inne, da sie verschiedenste Bestandteile

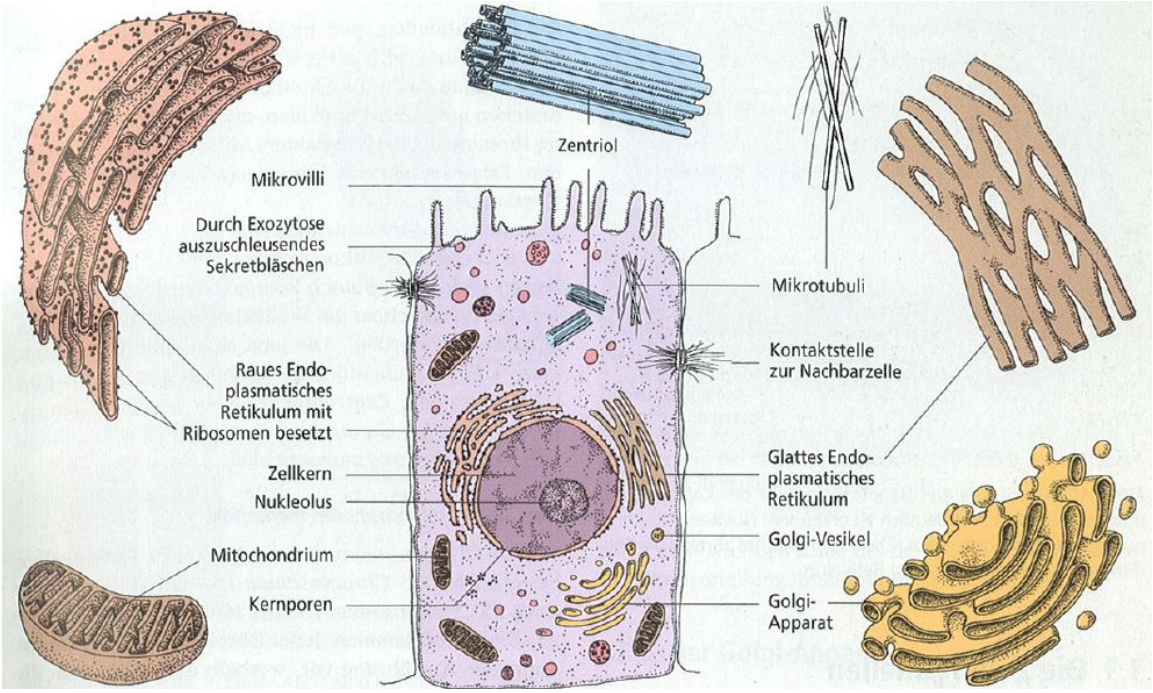

Abbildung 2.1: Schaubild der Zelle, entnommen aus Schäffler und Menche (2000, S. 31, [45]).

der Zelle codiert, aber auch Zellabläufe reguliert. Unter Bestandteile fallen auch Proteine, welche rund die Hälfte des Trockengewichtes einer Zelle ausmachen können (Campbell und Reece, 2006, S. 85, [5]). Sie werden mittels Proteinbiosynthese über die DNA (genauer gesagt über die Ribonukleinsäure, kurz RNA) codiert und in den Ribosomen der Zelle hergestellt. Mutationen innerhalb der DNA können somit unter anderem auch Auswirkungen auf Proteine haben.

Proteine haben dabei (innerhalb) der Zelle viele verschiedene Aufgaben (vom Transport bis hin zur Apoptose, dem programmierten Zelltod) und sind daher ein sehr wichtiger Baustein. Einen Überblick über die möglichen Funktionen gibt dabei Tabelle 5.1 aus Campbell und Reece (2006, S. 86, [5]) oder ein Auszug aus dieser, welcher in Tabelle 2.1 gegeben ist. Die Funktion eines Proteins steht dabei in direktem Zusammenhang zu

Proteintyp	Funktion	Beispiel
Strukturproteine	Formgebung und Halt	Kollagen
Transportproteine	Transport von Stoffen	Hämoglobin
Rezeptorproteine	Reaktion einer Zelle auf chemische Reize	Rezeptoren, z.B. in der Membran einer Nervenzelle
Abwehrproteine	Schutz vor Krankheiten	Antikörper

Tabelle 2.1: Überblick über Proteinfunktionen; Auszug aus Tabelle 5.1. aus Campbell und Reece (2006, S. 86, [5]).

seiner Gestalt, da oft eine Funktion die Bindung eines anderen Moleküls beinhaltet und für diese Bindung eine spezielle Gestalt wichtig ist. Einige Proteine sind auch innerhalb der Zellmembran lokalisiert und funktionieren dort als Transportkanäle für Stoffe oder Rezeptoren, welche Signale aus der Umgebung an das Zellinnere weitergeben können und umgekehrt.

Anhand der verschiedenen Funktionen kann man bereits erkennen, dass Fehlfunktionen (z.B. hervorgerufen durch Mutationen) der Proteine schwerwiegende Folgen haben können. Durch eine Mutation könnte sich zum Beispiel die Gestalt ändern und somit die Bindung des richtigen Moleküls unmöglich werden. Dadurch können bestimmte Vorgänge in einer Zelle nicht mehr oder nur falsch ausgeführt werden.

So kann durch eine Mutation an einer Stelle der DNA, welche ein für den Zellwachstum zuständiges Protein codiert, den normalen Zellwachstum und die Zellteilung empfindlich stören. In diesem Fall kann es nach der Mutation dazu kommen, dass nun das Protein besonders oft exprimiert wird oder aber in einer aktiveren Form in der Zelle vorhanden ist (Campbell und Reece, 2006, S. 431, [5]). Dies beeinflusst den Zellzyklus stark und ebnet den Weg für eine maligne Entartung und somit für Krebszellen.
Im Falle einer Mutation, die sich auf einen Rezeptor auswirkt, würden hingegen falsche Signale an das Zelleinnere weitergegeben werden. Dies betrifft auch die Signalproteine in der Zellmembran. Diese sind für den Stoff- und Informationsaustausch der Zelle mit ihrem Umfeld verantwortlich. Sollte dieser Prozess jedoch durch eine Mutation beeinflusst sein, kann dies schwerwiegende Folgen für den Organismus haben.
Somit wird deutlich, dass die Untersuchung von Proteinen ein wichtiges Forschungsgebiet ist. Weiter bietet sich hier die Möglichkeit für die Medikamentenentwicklung bzw. die Forschung, den Ablauf und den Aufbau von Proteinen zu untersuchen, um einer solchen Mutation bzw. ihren Ausmaßen entgegen zu wirken.
In dieser Arbeit werden im Weiteren vier Proteine betrachtet: Lamin B, KRas, EGFR (Epidermal Growth Factor Receptor) und PTB (Post-Transcriptional Control). Lamina sitzen dabei auf der inneren Zellmembran der Kernhülle. Dort sind sie für die Struktur des Nukleus sowie der Regulation der DNA-Replikation zuständig. Mit ihrer Hilfe kann weiter Chromatin an der Kernmembran befestigt werden, sodass ein kontrollierter Austausch zwischen dem Nukleus und dem Zellplasma gewährleistet ist. Lamina können nun in zwei Unterarten aufgeteilt werden, dem Lamin A und dem Lamin B. Das hier betrachtete Lamin B ist speziell während der Entwicklung der Zelle besonders stark ausgeprägt und wird in dieser Phase eingesetzt.
Im Gegensatz zu dem Laminen hat die Ras-Familie ihre Funktion in der Signaltransduktion. Das KRas Protein spielt dabei z.B. in einer Reihe biochemischer Prozesse eine große Rolle, welche das Wachstum und die Differenzierung von neuen Zellen betrifft. Da diese Rolle sehr wichtig in der Zellentwicklung ist und Mutationen oder Fehler schwerwiegende Folgen haben können, ist es auch nicht verwunderlich, dass eine Reihe von Ras-Onkogenen bekannt ist. Dadurch ist diese Protein-Familie sehr interessant für biologisch-medizinische Untersuchungen.

Das EGFR-Protein ist weiter ein Rezeptor und sitzt in der Zellmembran. Wenn er aktiviert ist, leitet er Signale ans Zellinere, welche das Zellwachstum stimulieren und den Zelltod verhindern können. Somit ist EGFR, wie der Name schon sagt, ein Rezeptor für Wachstumsfaktoren und ein Fehler bzw. eine Mutation können auch hier schwerwiegende Folgen haben. Daher ist es auch nicht verwunderlich, dass EGFR bei vielen Tumorarten hochreguliert oder mutiert vorliegt.
Das letzte Protein, PTB, ist ein RNA-bindendes Protein und kann zwischen dem Nukleus und dem Zytoplasma pendeln. PTB hat seine Funktion beim Splicen der RNA (Ribonukleinsäure), aber auch bei diversen zellulären Prozessen, wie z.B. der Initiierung der Translation.

2.2. Zielsetzung und Literaturüberblick

Wie man durch das vorherige Kapitel vermuten kann, spielen Proteine eine tragende Rolle in der Zelle. Dies liegt nicht nur an der engen Verknüpfung zur DNA, sondern auch an ihren vielseitigen Funktionen. Daher ist es für Wissenschaftler, z.B. Biologen, von Interesse Proteine weiter zu erforschen (z.B. Proteine in Krebszellen). In diesem Fall wäre u.a. relevant, welche Auswirkung die Mutation der DNA hat oder auch, ob die Proteinexpression bei steigender Genexpression ebenfalls ansteigt.
Zum anderen ist aber auch das normale Verhalten in einer Zelle von Interesse. Können die „normalen“ Abläufe in einer Zelle nachvollzogen werden, z.B. welche Vorgänge durch welche chemischen Signale gesteuert werden, so können auch Zusammenhänge zwischen der Proteinaktivität und anderen Vorgängen herstellen. Dabei ist von besonderem Interesse, die Zelle in einer lebenden Form zu betrachten, da hier die Übergänge zwischen verschiedenen Stadien der Zelle betrachtet werden können. Dank der stetigen Weiterentwicklung der Technik ist dies mit Hilfe moderner Lasermikroskopie möglich. Somit kann nun auch das Verhalten eines Proteins über die Zeit hinweg, wie z.B. der zurückgelegte Weg, betrachtet werden.
In dieser Arbeit liegt der Fokus auf der Analyse von Clusterphänomenen in Daten, welche mittels Fluoreszenzmikroskopie an lebenden Zellen erhoben wurden. Die Clusteranalyse ist in der Biologie ein weit verbreitetes Feld, da die Identifikation von ähnlichen Gruppen

in verschiedensten Arten von Daten sehr wichtig ist. Dabei können sowohl Evolutionsdaten (vgl. Mathew et al., 2013, [31]) wie auch Gendaten analysiert werden, wobei in letzterem Fall oft Expressionsdaten genutzt werden (wie in Eisen et al., 1998, [8] oder auch Sunaga, Nievola und Ramos, 2007, [49]). Aber auch das Clusterverhalten von Proteinen ist oft von großem Interesse (vgl. z.B. Arnau, Mars und Marín, 2005, [1]), welches auch hier behandelt wird. Dabei werden die Clusterphänomene aus folgenden drei Blickwinkeln analysiert:

1. zeitlich,
2. räumlich und
3. räumlich-zeitlich.

In der zeitlichen Analyse werden zum einen ChIP-Seq-Daten analysiert, an denen eine Segmentierung vorgenommen werden soll. Dadurch können Gebiete auf dem Chromosom mit auffälliger Konzentration eines bestimmten Proteins identifiziert werden. Zum anderen werden Proteinzeitreihen untersucht, welchen eine stückweise konstante Funktion zu Grunde liegt. Sie bestehen aus der Lichtintensität eines „Spots“ des Mikroskopiebildes, welche über die Zeit hinweg gemessen wurde. Um eine Schätzung der Anzahl an Proteinen in einem Spot ermöglichen zu können, ist hier die Rekonstruktion der stückweise konstanten Funktion das Ziel. Eine Strategie diese stückweise konstante Funktion zu rekonstruieren ist, zunächst die Bruchpunkte/Changepoints zu identifizieren um anschließend zwischen diesen konstante Segmente anzupassen, so dass eine stückweise konstante Funktion entsteht. Daher kann für beide Datensätze das Ziel mit Hilfe von Changepoint-Methoden erreicht werden.
In diesem Zusammenhang existieren eine Vielzahl an Schätz- und Glättungsmethoden. So haben Lai et al. (2005, [28]) in einer Art Übersichtsartikel verschiedene Methoden verglichen. Darunter waren sowohl Schätz- als auch Glättungsmethoden, wie z.B. die Quantilsregression. Mit Hilfe einer Quantilsregression (Eilers und de Menezes, 2005, [7]) können die Daten geglättet werden, wobei das Ergebnis einer stückweisen Funktion entspricht. Somit kann je nach Wahl des Quantils eine Funktion angepasst werden, welche aus kürzeren oder längeren konstanten Segmenten besteht. Eine andere Art der Glättung beschreiben Hupe et al. (2004, [20]) in ihrem Artikel. Dort wird die Glättung mit Hilfe adaptiver Gewichte berechnet. Neben Glättungsmethoden kann weiter aus einem Pool

aus unterschiedlichen Schätzmethoden gewählt werden, welche im Allgemeinen auf einen Changepoint im Mittel testen. So ist der Circular Binary Segmentation Algorithmus von Olshen und Venkatraman (2004, [36]) auf dem Gebiet der arrayCGH-Datenanalyse eine Standardmethode. Bei dieser werden die Changepoints iterativ mittels entsprechendem Testproblem bestimmt. Wang et al. (2005, [51]) haben weiter eine Methode, basierend auf der Theorie des Clusterns entwickelt. Dabei sind zwei wichtige Punkte dieser Methoden, dass zum einen ein bestimmtes Distanzmaß verwendet wird, welches nur für aufeinanderfolgende Beobachtungen definiert ist, zum anderen die Tatsache, dass nur benachbarte Cluster vereinigt werden können. Somit wird die zeitliche bzw. auf dem Chromosom räumliche Abhängigkeitsstruktur beachtet. Aber auch im Bereich der Zeitreihenanalyse sind Changepoint-Methoden zu finden. So haben Fryzlewicz und Subba Rao (2013, [10]) eine Methode entwickelt, welche multiple Changepoints in einem ARCH-Prozess erkennt. Matteson und James (2013, [33]) haben weiter eine Methode entwickelt, mit welcher nicht nur Changepoints im Mittelwert, sondern auch in der Varianz erkannt werden. Dazu nutzen Matteson und James eine retrospektive Schätzung der Lage und der Anzahl an Changepoints mit Hilfe der Clustertheorie und unverzerrten Schätzern. Neben diesen frequentistischen Ansätzen existieren auch unterschiedliche Bayes-Ansätze. So haben Barry und Hartigan bereits 1993 ([3]) eine Methode zur Changepointschätzung entwickelt. Der Vorteil ist hier, dass man neben der Schätzung der stückweise konstanten Funktion auch die a posteriori Wahrscheinlichkeiten der einzelnen Changepoints erhält. Weiter haben Messina et al. (2006, [34]) ein Hidden Markov Modell speziell für die Messung von Proteinen mittels Fluoreszenzmikroskopie entwickelt.

In dieser Arbeit wird ebenfalls Bayesianisch vorgegangen, indem ein Bayessches hierarchisches Modell, zunächst für die ChIP-Seq-Daten, aufgestellt und implementiert wird. Der Vorteil dieser Methode ist, dass die originalen Count-Daten genutzt werden und nicht die $\log_2$-Ratios (im Weiteren kurz: log-Ratio). Es kann anschließend gezeigt werden, dass dieses Modell ebenfalls auf Proteinzeitreihen angewendet werden kann und auch dort sinnvolle Ergebnisse liefert. Ein weiterer Vorteil ist, wie bei der Methode von Barry und Hartigan, dass auch hier Wahrscheinlichkeiten, die marginalen a posteriori Wahrscheinlichkeiten für einen Changepoint innerhalb eines Fensters, zur Verfügung stehen. Mit Hilfe der gefunden Changepoints können anschließend die stückweise konstanten Segmente zwi-

schen den Changepoints definiert werden.

In der räumlichen Analyse liegen Daten in Form eines Pixel-Bildes vor. Dieses entsteht bei der Messung eines Proteins in einer lebenden Zelle mittels Fluoreszenzmikroskopie und enthält helle Punkte, welche Protein(-cluster) repräsentieren. Dieses Messverfahren kann weiter auch auf eine simultane Messung von zwei unterschiedlichen Proteinen erweitert werden, sodass durch die Nutzung von zwei unterschiedlichen Fluoreszenzen für die zwei Proteine ein zweifarbiges Pixelbild entsteht. In diesem Zusammenhang spricht man vom Dual Colour Fall.

Um die Daten nun zu analysieren, können unterschiedliche Strategien verfolgt werden. Zum einen können die Daten als Punktprozess behandelt und dementsprechend modelliert werden. Für diese Art von Analyse existiert eine große Anzahl an Methoden und Strategien. Gelfand et al. (2010, [11], Kapitel 4 und 5) stellen in ihrem Buch verschiedene Methoden zur Analyse und Modellierung solcher Punktprozesse vor. Auch Baddeley und Turner (2005, [2]) haben sich mit der Analyse von Punktprozessen beschäftigt und für die Analyse das R-Paket `spatstat` entwickelt.

Eine weitere Möglichkeit der Analyse ist die Clusteranalyse im Hinblick auf unterschiedliche Clustercharakteristika, z.B. den durchschnittlichen Radius der Cluster oder auch den Anteil an Punkten in Clustern. Um Informationen über den Radius zu erlangen, kann Ripley's K-Funktion genutzt werden (Ripley, 1977, [40]). Die K-Funktion kann dabei an spezielle Punktprozesse durch eine Transformation angepasst werden.

In dieser Arbeit soll jedoch der Fokus auf dem Anteil an Punkten bzw. Proteinen in Clustern liegen. Um diese Proportion zu schätzen, können unterschiedliche räumliche Cluster-Methoden verwendet werden. Jede Methode kann dabei sowohl Vor- als auch Nachteile besitzen. So wird beim hierarchischen Clustern (vgl. z.B. Hartigan, 1975, [17]) die Anzahl an Clustern nicht direkt geschätzt, sodass die geschätzte Proportion abhängig von der gewählten Clusteranzahl ist. Weitere räumliche Clustermethoden, welche neben der Clusterzuordnung auch die Clusteranzahl schätzen, sind z.B. das Extensible Markov Modell (Dunham et al., 2004, [6]) oder auch der DBSCAN Algorithmus (Ester et al., 1996, [9]). Mit diesen Methoden kann nun sowohl die Clusteranzahl, als auch die Clusterzuordnung geschätzt werden, wobei auch eine Einteilung eines Datenpunktes zum Background, d.h. zu keinem Cluster zugehörig, möglich ist. Sofern diese Einteilung erfolgt ist, kann anschlie-

ßend die Proportion der Punkte in Cluster geschätzt werden. Für beide Clustermethoden muss jedoch vor der Analyse (mindestens) ein Parameter gewählt werden. Eine Clustermethode, welche keine informative Parameterwahl benötigt und neben der Schätzung des Anteils auch noch weitere Clustercharakteristika schätzt, ist die Gammics Methode von Schäfer et al. (2015, [44]). Diese modelliert dabei nicht den Punktprozess selber, sondern die Distanz eines Punktes mit seinen Nachbarn mit Hilfe zweier Gamma-Verteilungen. Durch diese Modellierung können neben der Proportion an Punkten in einem Cluster auch der durchschnittliche Radius sowie die durchschnittliche Größe der Cluster geschätzt werden.

Um mögliche Vor- und Nachteile einzelner Methoden zu ermitteln, soll in dieser Arbeit ein Vergleich eben dieser erfolgen. Weiter soll ein Schema entwickelt werden, wie die entsprechenden Methoden effizient kombiniert werden können.

Im letzten Fall, der räumlich-zeitlichen Analyse, werden zwei Proteine in einer lebenden Zelle über die Zeit hinweg beobachtet. Dabei werden die Wege gemessen, welche die Proteine zurück gelegt haben. Diese Wege werden auch Tracks genannt.

Um Daten dieser Art analysieren zu können, können unterschiedlichste Strategien verfolgt und sich Methoden verschiedener Anwendungsgebiete bedient werden. Eines dieser Anwendungsgebiete, in dem ebenfalls Wege beobachtet werden, ist das Animal Tracking. Hier werden Tiere oder auch Tiergruppen beobachtet um das Verhalten eben dieser zu modellieren. So haben Kranstuber und Smolla ein `R`-Paket entwickelt, um Animal Tracking Daten anschaulich darzustellen und zu analysieren (`move`, 2015, [27]). Mit diesem Paket kann z.B. die „utilization density“ mit Hilfe eines dynamischen Bewegungs-Modells, basierend auf Brownschen Brücken, berechnet werden (vgl. Kranstuber et al., 2012, [26]). Dadurch kann das Gebiet berechnet werden, in welchem sich das Tier mit einer bestimmten Wahrscheinlichkeit aufhält. Weiter entwickelte Gurarie (2014, [16]) das `R`-Paket `bcpa`, welches mit Hilfe einer Likelihood-basierten Methode Veränderungen/Changepoints in bestimmten Bewegungsparametern erkennt.

In dieser Arbeit soll jedoch der Fokus auf mögliche Zusammenhänge einzelner Objekte und somit den Tracks gelegt werden. Zwei mögliche Korrelationsmaße, welche für räumlich-zeitliche Daten geeignet sind, sind zum einen Moran's I und zum anderen die Korrelation nach Syrjala. Moran führte 1950 ([35]) eine Teststatistik zur Untersuchung der räumli-

chen Autokorrelation benachbarter Lokationen ein. Diese verwendet zunächst, unter der Annahme eines konstanten Mittelwerts, die Differenzen der Beobachtungen und eben diesem Mittel. Dies dient der Trendbereinigung. Moran's I setzt sich dann aus der Summe aller paarweisen Produkte der jeweiligen trendbereinigten Beobachtung zusammen, welche durch einen weiteren Term standardisiert werden. Dabei werden die Produkte nicht benachbarter Lokationen durch eine Identikatorfunktion mit Null multipliziert und haben somit keinen Einfluss. Syrjala (1996, [50]) vergleicht hingegen zur Berechnung seiner Korrelation zwei räumliche kumulative Verteilungsfunktionen auf einem Gitter. Dafür werden zwei Variablen jeweils an den Gitterpunkten gemessen und anschließend die entsprechenden Verteilungsfunktionen auf Gleichheit getestet.
In dieser Arbeit soll speziell der Zusammenhang zweier Tracks betrachtet werden, d.h. der Fokus liegt nicht auf bestimmten Gebieten oder Lokationen, sondern auf dem räumlich-zeitlichen Verlauf zweier Proteine. Dafür wird im weiteren Verlauf ein Zusammenhangsmaß hergeleitet, welches für zwei Tracks neben der räumlichen Nähe pro Zeitpunkt auch betrachtet, ob diese je Zeitpunkt einem gemeinsamen Cluster zugeteilt werden würden.
Eine genauere Darstellung der entsprechenden Daten und auch der spezifischen Problemstellung erfolgt in den entsprechenden Kapiteln 3.1, 4.1 und 5.1.

3. Ein Bayessches hierarchisches Modell zur Segmentierung

In diesem Kapitel wird die zeitliche Analyse von Proteindaten behandelt. Dafür werden zunächst die in diesem Kapitel verwendeten Daten sowie das spezifische Ziel beschrieben. Anschließend wird der Bayessche hierarchische Segmentierungsalgorithmus erläutert sowie eine weitere zum Vergleich verwendete Methode. Abschließend folgt die Analyse der Daten. Alle Methoden werden mit Hilfe der Statistik-Software `R` (2015, [38]) berechnet. Weiter wird das R-Pakete `DNAcopy` (Seshan und Olshen, [47]) genutzt.

3.1. Daten- und Problembeschreibung

Im Folgenden werden nun zwei Datensätze, welche in der zeitlichen Analyse verwendet werden, beschrieben, wobei der erste Datensatz ChIP-Seq-Daten, der zweite Proteinzeitreihen enthält. Weiter wird darauf aufbauend, das spezifische Analyseziel erläutert.

3.1.1. ChIP-Seq Daten

Mit Hilfe von ChIP-Seq (vom englischen „Chromatin-Immunoprecipitation Sequencing“), einem experimentellen Versuch in der Genetik, können Zusammenhänge zwischen der DNA und Proteinen (genauer: Transkriptionsfaktoren) untersucht werden (Park, 2009, [37]). Dadurch können bestimmte Abläufe innerhalb der Zelle genauer analysiert werden. Oft sind bei ChIP-Seq Versuchen bestimmte Zellen von Interesse, z.B. Krebszellen.
Um die ChIP-Seq-Daten zu erheben, wird die DNA in kurze Fragmente zerteilt und isoliert. Diese Fragmente können nun mit Hilfe einer (high-throughput) Sequenzierung ihren spezifischen Stellen des Chromosoms zugeordnet werden. Um zufällige Zusammenhänge zwischen der DNA und Proteinen erkennen zu können, wird stets eine Referenzprobe in den Versuch mit aufgenommen. So können zufällige Anomalien gut identifiziert und von systematischen Veränderungen unterschieden werden.
In dieser Arbeit werden ChIP-Seq-Daten analysiert, welche Aufschluss über den Zusam-

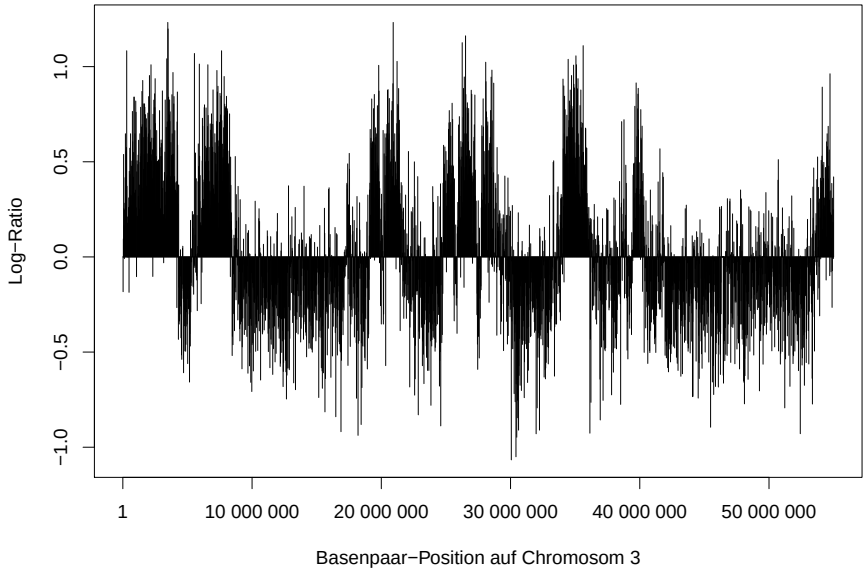

Abbildung 3.1: Beobachtete log-Ratios des Lamin Bs gegenüber der Referenz auf Chromosom 3 (Basenpaare 0 bis 55 000 000); dabei gibt das log-Ratio das Verhältnis von Lamin B gegenüber der Referenz zum log_2 wieder.

menhang zwischen der DNA und Lamin B geben sollen. Lamine sind dabei Moleküle, welche in der menschlichen Zelle auf der inneren Zellmembran des Zellkerns (Nukleus) lokalisiert sind (vgl. Abbildung 2.1). Da dieser die DNA enthält, haben Lamine eine räumliche Nähe zur DNA (vgl. auch Kapitel 2.1). Dies lässt bereits darauf schließen, dass ihre Funktion in Zusammenhang mit der DNA steht, was auch der Fall ist: Lamine sind für die Struktur des Nukleus sowie der Regulation der DNA-Replikation zuständig. Weiter kann mit Hilfe der Lamine Chromatin an der Kernmembran befestigt werden, sodass ein kontrollierter Austausch zwischen Nukleus und Zellplasma gewährleistet ist.
Weiter können Lamine in zwei Gruppen unterteilt werden: Lamin A und Lamin B. Diese zwei Gruppen unterscheiden sich neben der Struktur auch in ihrem Verhalten während der Mitose (der Zellteilung). Lamin B ist während der Entwicklung der Zelle besonders stark ausgeprägt und wird in dieser Phase eingesetzt (Gruenbaum et al., 2005, [14]).
Da bei der Entstehung vieler Krankheiten die Differenzierung (Entwicklung einer Zelle) eine wichtige Rolle spielt, wie z.B. bei einer Krebserkrankung, ist es wichtig diesen Vorgang zu untersuchen. Versteht man die Zusammenhänge und findet Möglichkeiten diesen

Prozess bzgl. bestimmter Faktoren zu regulieren, so könnte dies ein Schlüssel für die Behandlung einiger Krankheiten sein. Somit ist ein Ziel, DNA-Segmente zu finden, welche

Chromosom	Start [bp]	Ende [bp]	c_i	b_i
⋮	⋮	⋮	⋮	⋮
3	0	19 999	0	0
3	20 000	39 999	9	5
3	40 000	59 999	45	20
3	60 000	79 999	64	20
3	80 000	99 999	67	28
3	100 000	119 999	78	32
3	120 000	139 999	66	25
3	140 000	159 999	75	40
3	160 000	179 999	73	33
3	180 000	199 999	75	21
⋮	⋮	⋮	⋮	⋮

Tabelle 3.1: Auszug aus dem ChIP-Seq-Datensatz: hier sind die Signal Counts c_i und die Background Counts b_i für die ersten 10 Fenster auf Chromosom 3 gegeben.

in engem Zusammenhang mit Lamin B stehen (Guelen et al., 2008, [15]).
Die hier vorliegenden Daten sind in Fenstern der Länger 20 000 Basenpaare (bp) eingeteilt. Für jedes Fenster i, $i = 1, \ldots, N$, enthält der Datensatz neben der Angabe des Chromosoms, die Start- und Endposition der Fenster sowie die Anzahl der „Counts“ c_i für Lamin B bzw. für die Referenz b_i. Beispielhaft sind die ersten zehn Beobachtungen des Datensatzes für Chromosom 3 in Tabelle 3.1 dargestellt sowie auch grafisch in Abbildung 3.1 (Chromosom 3 wird auch weiterhin stets als Beispiel genutzt). In der ersten Spalte ist dabei die Angabe des Chromosom, dann folgen der Start- und Endpunkt des Fensters in den Spalten zwei und drei und abschließend sind in Spalte vier und fünf die Anzahlen der Counts für Lamin B und die Referenz eingetragen. Weiter kann schnell die Gesamtanzahl der Counts über $n_i = c_i + b_i$ für Fenster i hergeleitet werden, $i = 1, \ldots, N$.
Die vorliegenden Daten enthalten keine fehlenden Werte, d.h. in jedem Fenster sind sowohl

Angaben für die Signal als auch die Background Counts gegeben.

3.1.2. Proteinzeitreihen

Ein weiterer Datensatz, welcher für die zeitliche Analyse verwendet wird, besteht aus Proteinzeitreihen. Diese wurden am Max-Planck-Institut für molekulare Physiologie Dortmund 2011 von Dipl.-Biol. Franziska Thorwirth (aus der Arbeitsgruppe von Dr. Peter J. Verveer) erhoben.
Bei der Datenerhebung wurden fixierte Präparate von stabilen Standardzellen (hier 16HBE-40 human bronchial epithelial cells) verwendet. Diese wurden vorher mit Hilfe eines genetisch modifiziertem BAC *(bacterial artificial chromosome)* Vektors hergestellt. Die BAC Vektoren sind dabei genetisch so modifiziert, dass sie den genetischen Lokus von humanem KRas enthalten und ein Fusionsprotein aus KRas und einem Fluoreszenzprotein (hier mCitrine, eine Variante des gelb fluoreszierenden Proteins (YFP)) exprimiert wird. Dabei ist KRas ein Protein, welches für die Regulierung des Zellwachstums und der Differenzierung verantwortlich ist. Eine Mutation kann somit zum Kontrollverlust des Zellwachstums führen und für bestimmte Krankheiten (wie z.B. Krebs) förderlich sein. Dies macht KRas zu einem interessanten Forschungsgebiet bei der Entwicklung von (Krebs-)Medikamenten. Die für diesen Datensatz interessierenden Proteine sitzen in der Plasmamembran und können „alleine“ an einer Stelle lokalisiert, aber auch in einem Cluster enthalten sein. Dank der Präparation im Vorfeld und der TIRF-Mikroskopie *(total internal fluorescence microscopy)* können nun die KRas-Proteine in der Membran sichtbar gemacht werden. Durch die TIRF-Mikroskopie kann hier sichergestellt werden, dass nur wenig Rauschen aus dem Zellinneren aufgenommen wird, da bei dieser Mikroskopie lediglich eine schmale Schicht direkt auf dem Objektträger beobachtet wird. Eine schematische Darstellung des experimentellen Aufbaus einer TIRF-Messung ist in Abbildung 1 in Anhang C zu finden. Dadurch kann sich auf die Proteine in der Zellmembran, welche direkt auf dem Objektträger liegt, konzentriert werden.
In der Mikroskopie werden die Fluoreszenzproteine mit Licht einer bestimmten Wellenlänge angeregt, welches entsprechend dem Anregungsspektrums des Proteins gewählt wird. Als Ergebnis der Mikroskopie entsteht ein schwarzes Pixel-Bild mit hellen Punkten. Diese

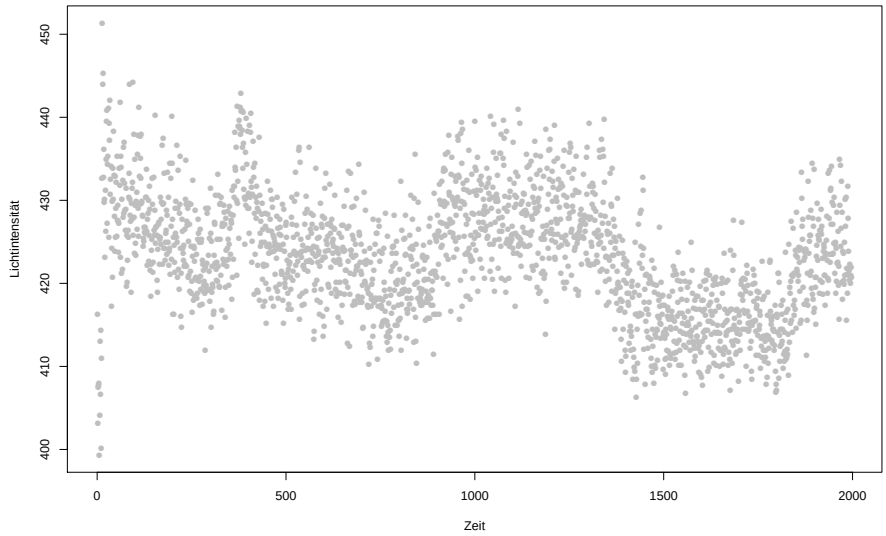

Abbildung 3.2: Beispielhafte Proteinzeitreihe aus dem Experiment mit 100 Watt und 50 ms.

sogenannten Spots erscheinen an den Stellen, an denen mindestens ein Protein aktiv ist. Das Licht aus den Spots, die sogenannte Emission, besitzt dabei eine andere Wellenlänge als das Licht, mit dem die Fluoreszenzproteine angeregt wurden. Die Lichtintensität eines solchen Spots wird nun über die Zeit hinweg gemessen und ergibt somit eine Proteinzeitreihe. Die Helligkeit hängt dabei stark mit der Anzahl an enthaltener Proteine in dem Spot zusammen: Viele (aktive) Proteine in einem Spot ergeben eine hohe Lichtintensität, wenige Proteine eine geringe Lichtintensität. Ein Beispiel einer Proteinzeitreihe ist in Abbildung 3.2 zu sehen.

Es ist gut zu erkennen, dass die Proteinzeitreihe eine stückweise konstante Gestalt inne hat. Dies resultiert daraus, dass die Proteine nach einer gewissen Zeit vom aktiven in den inaktiven Zustand wechseln. Wäre nur der Weg von aktiv nach inaktiv möglich, so würde die Proteinzeitreihe eine monoton fallende stückweise konstante Funktion ergeben. Wie man an Abbildung 3.2 jedoch erkennen kann, ist dies nicht der Fall. Die erneuten Anstiege resultieren daher, dass die Proteine zum einen *bleachen* und zum anderen *blinken* können. Bleachen bedeutet für ein Protein den unumkehrbaren Wechsel von aktiv zu inaktiv und somit kann das Protein nicht mehr detektiert werden. Blinken hingegen bedeutet den temporären Wechsel von aktiv zu inaktiv oder umgekehrt. Im Falle des Blinken kann somit

ein Protein zwischen den Zuständen wechseln, wohingegen beim Bleachen das Protein in einem inaktiven Endzustand gewechselt ist.
Unter der Annahme, dass die Proteinzeitreihe so lange beobachtet wurde, bis alle Proteine gebleached sind, kann nun eine Schätzung für die Anzahl an Proteine im Spot berechnet werden. Dies kann über die Anzahl an stückweise konstanten Segmente erfolgen. Dabei

Experiment Nr.	Einstellung des Lasers [in Watt]	Zeitdauer der Aufnahme [in ms]	Anzahl gemessener Proteinzeitreihen	Anzahl Messzeitpunkte
1	100	50	936	1998
2	150	100	937	1998
3	150	100	1930	1998
4	200	100	245	998

Tabelle 3.2: Einstellungen der vier Experimente (dabei sind die Einstellungen für Experiment 2 und 3 identisch, jedoch ist zu erkennen, dass für Experiment 3 deutlich mehr Proteinzeitreihen gemessen wurden).

muss stets beachtet werden, dass an einer Treppenstufe mehr als ein Protein den Zustand wechseln kann.
Bei der Fluoreszenzmikroskopie können nun zusätzlich verschiedene Einstellungen vorgenommen werden. Zum einen kann die Watt-Zahl des Lasers eingestellt werden, zum anderen die Zeitdauer der Aufnahme. Ist die Zeitdauer zum Beispiel auf 50 ms eingestellt, so ergibt sich die Intensität des Spots dafür einen Zeitpunkt als Summe der Intensitäten, welche in den 50 ms gemessen wurden. Die Einstellungen der vier Experimente, welche die vorliegenden Daten ergeben, sind in Tabelle 3.2 zusammengefasst. Weiter enthalten die vorliegenden Proteinzeitreihen keine fehlenden Werte.

3.1.3. Ziel in der zeitlichen Analyse

Das Ziel der Analyse der ChIP-Seq-Daten ist die Identifizierung der Segmente, welche stark mit Lamin B assoziiert sind. Dies bedeutet in diesem Fall, dass man genau die Abschnitte auf dem Chromosom sucht, an denen die Lamin B Counts, c_i, höher als die

des Backgrounds/der Referenz, b_i, sind bzw. wo das log-Ratio deutlich erhöht ist. Somit sucht man Segmente, welche über Strukturbrüche/Changepoints (CPs) definiert werden können.

Für die Proteinzeitreihen ist die Identifizierung der stückweise konstanten Segmente innerhalb der Proteinzeitreihe das Ziel. Über diese Segmente könnte anschließend eine Schätzung der Anzahl enthaltener Proteine in einem Spot berechnet werden. Zum Beispiel kann für jede „Stufe abwärts“ ein Protein vermutet werden, welches in einen inaktiven Zustand wechselt (codiert mit -1). Für eine „Stufe aufwärts“ kann hingegen ein Protein vermutet werden, welches in den aktiven Zustand wechselt (codiert mit +1). Somit kann am Ende durch Summieren der codierten Werte ein Wert der enthaltenen Proteine angegeben werden. Die genaue Schätzung der Anzahl soll in dieser Arbeit jedoch nicht erfolgen. Lediglich die Identifikation der Segmente ist von Interesse.

Da hier bei beiden Daten-Arten Segmente gesucht werden, welche sich durch einen Strukturbruch bzw. durch einen Changepoint (CP) zu ihren umliegenden Beobachtungen abgrenzen, ist es naheliegend beide Ziele mit Hilfe von Changepointmethoden zu analysieren. Dabei kann das Changepoint-Problem im einfachsten Fall wie folgt beschrieben werden: Seien X_i Zufallsvariablen mit $i = 1, \ldots, N$. Der Zeitpunkt bzw. die Stelle t ist ein CP, wenn die Zufallsvariablen X_i mit $i = 1, \ldots, t$ und die Zufallsvariablen X_j mit $j = t + 1, \ldots, N$ verschiedenen Verteilungen folgen. Auf die zwei Datensätze bezogen wäre ein Changepoint die Stelle bzw. der Zeitpunkt, an dem die Counts c_i von einer erhöhten/niedrigen Konzentration in eine niedrige/erhöhte wechseln bzw. an dem ein konstantes Segment endet bzw. ein neues beginnt.

Für diese Analyse wird im Folgenden ein hierarchisches Modell für die ChIP-Seq-Daten aufgestellt und implementiert. Die Ergebnisse werden dabei stets mit denen des *Circular Binary Segmentation (CBS) Algorithmus* verglichen. Anschließend wird gezeigt, dass dieses Modell auch auf die Proteinzeitreihen angewendet werden kann.

3.2. Das Bayessche hierarchische Modell

In diesem Unterkapitel wird nun der Segmentierungsalgorithmus genauer erläutert. Dabei wird zunächst auf das zugrunde liegende Bayessche hierarchische Modell sowie das dar-

aus folgende a posteriori Modell eingegangen. Abschließend folgt eine kurze Erläuterung der Implementierung. Das Modell wurde in Zusammenarbeit mit Katja Ickstadt, Peter Müller und Holger Schwender entwickelt. Die Ausführungen dieses Kapitels, sowie ein Teil der Analyse der ChIP-Seq-Daten, sind in dem Artikel Herrmann et al. (2014, [19]) veröffentlicht worden.

3.2.1. Das hierarchische Modell

Um das zugrunde liegende hierarchische Modell zu definieren, wird zunächst die grundlegende Notation eingeführt:
Im Folgenden sei zum einen der Vektor der Signal Counts, hier die Counts des Lamin B, durch $c = (c_1, \ldots, c_N)$ sowie der Vektor der Backgroundcounts durch $b = (b_1, \ldots, b_N)$ gegeben. Weiter sei ähnlich wie in Kapitel 3.1.1 ein Vektor der Gesamtanzahlen der Counts durch $n = c + b$ gegeben. Diese sind bereits alle bekannt und den Daten zu entnehmen.
Die Anzahl der Segmente sei weiter mit K, die (Erfolgs-)Wahrscheinlichkeiten für erhöhte Lamin B Counts in jedem Segment mit $p = (p_1, \ldots, p_K)$ sowie die Segmentgrenzen (d.h. die entsprechenden CPs) mit $t_1 = 1, t_2, \ldots, t_K, t_{K+1} = N$ benannt. Dabei ist t_1 stets die erste Beobachtung und t_{K+1} die letzte Beobachtung der untersuchten Region. Dabei wird hier die Anzahl an Segmenten, K, als fest angesehen und muss somit vor Beginn der Analyse festgelegt werden. Die weiteren Parameter, p und t, sind hingegen unbekannt und werden in der Bayesschen Analyse mittels a priori Verteilungen modelliert.
Um nun das Modell aufzustellen, kann man zunächst überlegen, dass für die Lamin B Counts c_i eine Binomial-Verteilung mit segmentspezifischer Wahrscheinlichkeit p_k, $k = 1, \ldots, K$, sinnvoll ist. Diese Annahme führt zu folgender Formulierung der Likelihood $p(c \mid n, p, t)$:

$$p(c \mid n, p, t) = \prod_{k=1}^{K} \prod_{t_{k-1}}^{t_k - 1} \mathrm{Bin}(c_k | n_k, p_k). \tag{1}$$

Dabei muss beachtet werden, dass hier eine bedingte Likelihood für c_i gegeben n_i formuliert wurde (die Verteilung der $n_i = c_i + b_i$ wird hier nicht weiter modelliert).
Im nächsten Schritt werden nun die a priori Verteilungen festgelegt. Dabei werden zunächst die Segmentgrenzen t_k und die Wahrscheinlichkeiten p_k betrachtet, $k = 1, \ldots, K$. Für die Segmentgrenzen wird als a priori Verteilung das Produkt zweier Gleichvertei-

lungen verwendet. Die Gleichverteilung über dem Intervall (a, b) sei im Folgenden mit $\text{Unif}(a, b)$ bezeichnet. Somit ergibt sich als a priori Verteilung:

$$\begin{aligned} t_1 &= 1, \\ t_k &\sim \text{Unif}(2, K) \cdot \text{Unif}(t_{k-1}, t_{k+1}), \quad k = 2, \ldots, K, \\ t_{K+1} &= N. \end{aligned}$$

Dabei sind t_1 und t_{K+1}, wie oben bereits erwähnt, stets als erste bzw. letzte Beobachtung festgelegt. Für t_k, $k = 2, \ldots, K$, beschreibt das Produkt der Gleichverteilungen, dass zunächst zufällig ein CP t_k aus einer Gleichverteilung $\text{Unif}(2, K)$ gezogen wird. Anschließend wird dann der (neue) Wert des CPs aus einer $\text{Unif}(t_{k-1}, t_{k+1})$ gezogen.
Die a priori Verteilung der Wahrscheinlichkeiten p_k, $k = 1, \ldots, K$, ist hingegen ein wenig komplizierter, da die Verteilung davon abhängen soll, ob das Segment mit einer erhöhten Konzentration des Lamin B assoziiert ist, oder nicht. Daher wird als Hilfe ein zusätzlicher Vektor $w = (w_1, \ldots, w_K)$ mit latenten Indikatoren eingeführt. Diese Indikatoren spiegeln die Konzentration des Lamin B wieder und nehmen dabei die Werte 0 und 1 an, wobei $w_k = 1$ eine erhöhte Konzentration des Lamin B in Segment k und $w_k = 0$ keine erhöhte Konzentration des Lamin B in Segment k bedeutet, $k = 1, \ldots, K$. Die a priori Verteilung der p_k, $k = 1, \ldots, K$, kann nun über folgende Beta-Verteilungen (im Folgenden mit Be) beschrieben werden:

$$p(p_k \mid w, \mu_0, \mu_1, \rho) = \begin{cases} \text{Be}(\mu_0, \rho) & \text{falls } w_k = 0 \\ \text{Be}(\mu_1, \rho) & \text{falls } w_k = 1 \end{cases}.$$

Dabei sei an dieser Stelle darauf hingewiesen, dass hier eine Beta-Verteilung $\text{Be}(\mu_{w_k}, \rho)$ mit $0 < \mu_{w_k} < 1$, wie in Robert and Rousseau (2002, [42]) beschrieben, verwendet wird. Dies hat den Vorteil, dass bei dieser Beta-Verteilung $0 < \mu_{w_k} < 1$ ein Lageparameter (der Erwartungswert) und ρ ein Streuungsmaß (der Variationskoeffizient) ist. Somit ist die Interpretation der Parameter leicht. Weiter muss eine in dieser Beta-Verteilung nach Robert und Rousseau beinhaltete Einschränkung für ρ bei der weiteren Modellierung beachtet werden.
Wie man schon an den a priori Verteilungen erkennen kann, ist die Einführung des Vektors der latenten Variablen w ein Schlüsselpunkt in der Analyse der (nicht) erhöhten Konzentration des Lamin B. Bei der Aufnahme dieser Parameter ins Modell verschiebt sich somit

das Interesse von t und p auf t und w, d.h. auf die marginale a posteriori Verteilung von t und w gegeben den Daten, $p(t, w \mid \text{Daten})$.
Um das Modell nun zu komplettieren, fehlen noch die Verteilungen der Parameter der Beta-Verteilungen μ_0, μ_1 und ρ sowie die Verteilung des Vektors w (die sogenannten Hyperprior-Verteilungen).
Da die Segmente mit einer erhöhten bzw. normalen Konzentration an Lamin B stets iterieren sollen, kann zur Vereinfachung der Wahl einer Verteilung für den Vektor w, eine nützliche Eigenschaft des Vektors w mit beliebigen w_1 festgelegt werden:

$$w_{2l+1} = w_1 \quad \text{und} \quad w_{2l} = 1 - w_1,$$

wobei l eine positive ganze Zahl ist. Somit ist es ausreichend im Folgenden w_1 zu schreiben, da der Vektor w mit w_1 automatisch definiert ist. Diese Annahme ist dabei intuitiv und dadurch unbedenklich, da zwei gleiche aufeinanderfolgende Einträge im Vektor w bedeuten würden, dass zwei Segmente mit erhöhtem oder fehlendem Lamin B nebeneinander liegen. In diesen Fall würden diese beiden Segmente zu einem vereint werden, sodass wieder die iterative Folge von 0 und 1 im Vektor w gewährleistet ist.
Aus dieser Annahme resultiert nun, dass der Vektor w alleine durch w_1 definiert ist. Somit muss lediglich w_1 modelliert werden. Hier wird eine Bernoulli-Verteilung mit zufälliger Wahrscheinlichkeit π, Bernoulli(π), sowie eine nicht-informative Gleichverteilung Unif$(0, 1)$ für die Erfolgswahrscheinlichkeit der Bernoulli-Verteilung gewählt.
An dieser Stelle müssen nun noch die Verteilungen der Parameter der Beta-Verteilungen festgelegt werden, um das Modell zu komplettieren. Für (μ_0, μ_1) fällt hier die Wahl auf zwei Gleichverteilungen, wobei sich die Intervalle zwar überschneiden, dennoch leicht verschoben sind: $\mu_0 \sim \text{Unif}(0, 0.75)$ und $\mu_1 \sim \text{Unif}(0.25, 0.9)$, mit der Nebenbedingung $\mu_1 - \mu_0 > 0.25$. Dabei wird für μ_1 als obere Grenze 0.9 statt 1 gewählt, um die Streuung nach oben zu begrenzen. Für ρ wird eine Gammaverteilung mit Parametern 2 und 1

gewählt, Gamma$(2, 1)$. Zusammenfassend ergeben sich somit folgende Verteilungen:

$$\begin{aligned} w_1 \mid \pi &\sim \text{Bernoulli}(\pi), \\ \pi &\sim \text{Unif}(0, 1), \\ \mu_0 &\sim \text{Unif}(0, 0.75), \\ \mu_1 &\sim \text{Unif}(0.25, 0.9), \\ \rho &\sim \text{Gamma}(2, 1). \end{aligned} \tag{2}$$

Somit kann nun mit Hilfe von (2) die a priori Verteilung mit Hilfe eines Parametervektors $\theta = (p_1, \ldots, p_K, t_0, \ldots, t_K, w_1, \mu_0, \mu_1, \rho, \pi)$ wie folgt formuliert werden:

$$p(\theta) \propto p(t \mid K) \cdot \prod_{k=1}^{K} p(p_k \mid w, \mu_0, \mu_1, \rho) \cdot p(w_1 \mid \pi) \cdot p(\pi) \cdot p(\rho) \cdot p(\mu_0, \mu_1) \cdot p(K). \tag{3}$$

An dieser Stelle sei noch einmal auf die Annahme hingewiesen, dass die Anzahl an Segmenten (wie oben beschrieben) festgelegt ist. Daraus folgt, dass sich $p(K)$ in (3) zu einer Einpunktverteilung ergibt.

3.2.2. Das a posteriori Modell

Mit Hilfe der oben definierten Likelihood und den a priori Verteilungen kann nun das a posteriori Modell aufgestellt werden. Dafür soll nun die allgemein bekannte Beta-Verteilung mit Parametern α und β verwendet werden. Daher muss zunächst eine Umrechnung der Parameter von μ_0, μ_1 und ρ in $\alpha_0, \alpha_1, \beta_0$ und β_1 vorgenommen werden. Diese ergibt sich zu

$$\alpha_l = \frac{1 - \mu_l}{\rho^2} - \mu_l \quad \text{und} \quad \beta_l = \frac{(1 - \mu_l)^2}{\mu_l \rho^2} - (1 - \mu_l), \quad \text{mit } l \in \{0, 1\}. \tag{4}$$

Eine Herleitung für diese Umformungen in (4) sind im Anhang A unter A.1 zu finden.
Kombiniert man nun die Likelihood aus (1) und die a priori Verteilung aus (3), so erhält

man die gemeinsame a posteriori Verteilung:

$$\begin{aligned} p(p, w_1, \mu_0, \mu_1, \rho, \pi, t|c, N) \propto & \prod_{k=1}^{K} \prod_{i=t_{k-1}}^{t_k-1} \binom{n_i}{c_i} p_k^{c_i} (1-p_k)^{n_i-c_i} \frac{1}{K-2} \frac{1}{t_{k+1}-t_{k-1}} \cdot \\ & \cdot \prod_{k=1}^{K} \left[\left(\frac{\Gamma(\alpha_0+\beta_0)}{\Gamma(\alpha_0)\Gamma(\beta_0)} p_k^{\alpha_0-1} (1-p_k)^{\beta_0-1} \right)^{1-w_k} \cdot \right. \\ & \left. \cdot \left(\frac{\Gamma(\alpha_1+\beta_1)}{\Gamma(\alpha_1)\Gamma(\beta_1)} p_k^{\alpha_1-1} (1-p_k)^{\beta_1-1} \right)^{w_k} \right] \cdot \\ & \cdot \pi^{w_1} (1-\pi)^{1-w_1} \frac{1}{0.75} \frac{1}{0.9-0.25} \frac{1}{\Gamma(2)} \rho \exp(\rho) \frac{\Gamma(2)}{\Gamma(1)\Gamma(1)}, \end{aligned} \tag{5}$$

mit $w_k \in \{0, 1\}$ und Γ der Gamma-Funktion.

Durch Integration können in dieser a posteriori Verteilung nun die segmentspezifischen Wahrscheinlichkeiten $p_k, k = 1, \ldots, K$, marginalisiert werden. Dadurch ergibt sich:

$$\begin{aligned} p(w_1, \mu_0, \mu_1, \rho, \pi, t \mid c, N) \propto & \prod_{k=1}^{K} \frac{1}{K-2} \frac{1}{t_{k+1}-t_{k-1}} \frac{\Gamma(\alpha_{w_k}+\beta_{w_k})}{\Gamma(\alpha_{w_k})\Gamma(\beta_{w_k})} \cdot \\ & \cdot \pi^{w_1} (1-\pi)^{1-w_1} \rho \exp \rho \cdot \frac{\Gamma(\sum_i c_i + \alpha_{w_k})\Gamma(\sum_i b_i + \beta_{w_k})}{\Gamma(\sum_i n_i + \alpha_{w_k} + \beta_{w_k})}, \end{aligned} \tag{6}$$

wobei hier $\sum_i$ für $\sum_{i=t_k}^{t_{k+1}-1}$ steht.

Das Integral in Bezug auf die p_k gleicht dabei einem Beta-Integral, d.h. einem Integral über den Kern einer Beta-Verteilung mit Parametern $\sum_i c_i + \alpha_{w_k}$ und $\sum_i (n_i - c_i) + \beta_{w_k}$. Eine detailliertere Herleitung dieser Aussage ist im Anhang unter A.2 zu finden.

3.2.3. Implementierung

Um das Modell aus Kapitel 3.2.1 nun zu implementieren und folglich Daten zu analysieren, werden Markov Chain Monte Carlo (MCMC) a posteriori Simulationen verwendet (vgl. Robert and Casella, 2004, [41]). Dafür wird hier ein Metropolis Hastings Algorithmus mit Random Walk Proposal genutzt (vgl. Roberts et al., 1997, [43]).

Bei der Implementierung muss zunächst eine beliebige Anfangssegmentation sowie für alle weiteren Parameter ein Startwert festgelegt werden. In den dann folgenden Iterationen wird anschließend jeweils ein Vorschlag für die Parameter, die sogenannten *Proposals*, generiert. Für das hier vorgestellte Bayessche hierarchische Modell werden symmetrische Random Walk Proposals für μ_0, μ_1, ρ und π verwendet. Diese werden mit Hilfe einer Normalverteilung generiert, d.h. der Vorschlag $\mu_0^\star$ für μ_0 wird mittels $\mu_0^\star \sim N(\mu_0, 0.001)$

bestimmt. Dies geschieht analog für die anderen Parameter. Dabei wird hier eine Varianz von 0.001 gewählt, um eine sinnvolle Akzeptanzwahrscheinlichkeit zu erhalten.
Sollte der Fall eintreten, dass die Proposal außerhalb des von der a priori Verteilung unterstützen Bereiches liegt, so wird die a posteriori Verteilung des Vorschlags auf 0 gesetzt und die Proposal wird somit abgelehnt. Der Parameter ρ ist dabei durch folgende Ungleichung beschränkt:

$$\sqrt{\frac{(1-\mu_1)^2}{2\mu_1 - \mu_1^2}} > \rho.$$

Die Herleitung für diese Ungleichung ist Kapitel A.3 in Anhang Anhang A zu entnehmen. Dabei ergibt sich die Restriktion aus der Wahl der zwei unterschiedlichen Beta-Verteilungen, da die Shapeparameter α und β der in `R` implementierten Beta-Verteilung Beta(α, β) positiv sein müssen. Weiter nehmen wir noch $\beta > 1$ an, um zu verhindern, dass bei extremen Werten die Wahrscheinlichkeit p_k sich nicht stets zu 1 berechnet.
Als Proposal für w_1 wird hingegen eine Zufallszahl aus der vollständig bedingten a posteriori Verteilung in (6) gezogen. Dies ist ähnlich zu einem Gibbs Sampling Schritt. Da nicht jeder Faktor der a posteriori Verteilung von w_1 abhängt, ist leicht zu erkennen, dass

$$p(w_1 \mid \pi, \mu_0, \mu_1, \rho, c, b) \propto \pi^{w_1}(1-\pi)^{1-w_1} \cdot \\ \cdot \prod_{k=1}^{K} \left(\frac{\Gamma(\alpha_{w_k} + \beta_{w_k})}{\Gamma(\alpha_{w_k})\Gamma(\beta_{w_k})} \frac{\Gamma(\sum_i c_i + \alpha_{w_k})\Gamma(\sum_i b_i + \beta_{w_k})}{\Gamma(\sum_i n_i + \alpha_{w_k} + \beta_{w_k})} \right),$$

mit $w_{2\ell+1} = w_1$ und $w_{2\ell} = 1 - w_1$.
Anschließend wird nun mit der Aktualisierung der CPs t_k, $k = 2, \ldots, K$, begonnen. Dabei wird zufällig ein CP t_k gezogen und verschoben. Der vorgeschlagene neue CP kann dabei jede mögliche Stelle außer der anderen CPs $t_1, \ldots, t_{k-1}, t_{k+1}, \ldots t_N$ sein.
Somit kann hier die entsprechende Metropolis Hastings Akzeptanzwahrscheinlichkeit als einfacher Quotient der a posteriori Verteilung mit den neuen vorgeschlagenen Parametern und den aktuellen Parametern berechnet werden. Dabei wird der Ausdruck in Formel (6) für die gemeinsame a posteriori Verteilung genutzt.

3.2.4. Reversible Jump Erweiterung

Eine Möglichkeit, die Anzahl der Segmente K variabel zu modellieren, wäre die Einbettung eines Reversible Jump in den Metropolis Hastings Algorithmus (vgl. Green, 1995,

[13]). Dies wurde in einem ersten Ansatz wie folgt angewendet.

In jeder Iteration wird nun mit Hilfe von vorher festgelegten Wahrscheinlichkeiten bestimmt, ob in der jeweiligen Iteration ein CP verschoben, ein neues Segment hinzugefügt (*Split*) oder ein Segment gelöscht (*Merge*) werden soll. Diese Wahrscheinlichkeiten seien im folgenden durch $p_{Step} = (p_{MH}, p_{Split}, p_{Merge})$ bezeichnet und müssen sich stets zu 1 addieren. Durch die letzten zwei Optionen verändert sich automatisch die Anzahl der Segmente von K zu $\tilde{K}$, da ein neues Segment hinzugefügt oder zwei Segmente verbunden werden.

Für den normalen Metropolis-Hasting-Schritt, in dem ein CP verschoben wird, bleibt das Vorgehen wie zuvor beschrieben. Für den Merge- und den Split-Schritt hingegen müssen das Vorgehen sowie die Akzeptanzwahrscheinlichkeiten angepasst werden. Hier werden Parameter aus der Iteration zuvor übernommen, sodass lediglich das Löschen bzw. Hinzufügen eines Segmentes beurteilt wird und nicht die Wahl der Parameter. Im Merge-/Split-Schritt wird zunächst ein Segment k (mit CPs t_k und t_{k+1}) ausgewählt, welches gelöscht (Merge) bzw. in welches ein neues Segment integriert wird. Anschließend werden im Merge-Schritt die zugehörigen CPs gelöscht und der Vektor w entsprechend angepasst. Im Split-Schritt werden in dem ausgewählten Segment zwei neue CPs zufällig gezogen, sodass ein neues Segment (innerhalb des alten Segments) in dem ausgewählten Segment entsteht und anschließend ebenfalls der Vektor w angepasst.

Nun muss folglich auch die Akzeptanzwahrscheinlichkeit im Split- bzw. Merge-Schritt angepasst werden. Dafür sei zunächst Λ_k die Anzahl aller möglichen Segmente innerhalb des ausgewählten Segment k, welche sich wie folgt berechnet:

$$\begin{aligned}
\Lambda_k &= (t_{k+1} - 1) - (t_k + 2) + \ldots + 1 \\
&\overset{\star}{=} \frac{((t_{k+1} - 1) - (t_k + 2))((t_{k+1} - 1) - (t_k + 2) + 1)}{2} \\
&= \frac{(N_k - 3)(N_k - 2)}{2},
\end{aligned}$$

wobei $N_K = t_{k+1} - t_k$ ist und $\star$ gilt, weil $\sum_{i=1}^{N} i = \frac{n(n+1)}{2}$. Weiter sei $\theta = (t_0, \ldots, t_K, w_1, \mu_0, \mu_1, \rho, \pi)$ wie bereits in Kapitel 3.2.1 definiert und beinhalte die Parameter aus der vorherigen Iteration. Für die neue Segmentierung mit $\tilde{K}$ sei mit $\tilde{\theta} = (t_0, \ldots, t_{\tilde{K}}, w_1, \mu_0, \mu_1, \rho, \pi)$ ebenfalls ein Parametervektor analog zu θ definiert. Somit ergibt sich die Akzeptanzwahrscheinlichkeit $r_s(\theta, \tilde{\theta})$ für den Split-Schritt bzw. $r_m(\theta, \tilde{\theta})$ für den Merge-Schritt nun

zu

$$r_s(\theta, \tilde{\theta}) = \frac{p(\tilde{\theta})}{p(\theta)} \frac{\tilde{K}^{-1}}{K^{-1}\Lambda_k^{-1}} \frac{p_{Merge}}{p_{Split}} \frac{p(\tilde{K})}{p(K)} \quad \text{und}$$

$$r_m(\theta, \tilde{\theta}) = \frac{p(\tilde{\theta})}{p(\theta)} \frac{\tilde{K}^{-1}\Lambda_k^{-1}}{K^{-1}} \frac{p_{Split}}{p_{Merge}} \frac{p(\tilde{K})}{p(K)}.$$

Der erste Quotient ist dabei stets wie im Metropolis-Hastings-Schritt der Quotient der a posteriori Verteilung für die zwei Parameter-Settings. Diese werden analog wie zuvor berechnet. Lediglich die a priori Verteilung für K bzw. $\tilde{K}$ muss nun wieder beachtet werden. Die letzten drei Quotienten sind speziell auf den Merge- bzw. Split-Schritt zugeschnitten. Hier fließt zum einen die Anzahl an Segmenten mit der Anzahl aller möglichen Segmente innerhalb des ausgewählten Segments ein sowie die vorher festgelegten Wahrscheinlichkeiten für einen Split- bzw. Merge-Schritt. Der letzte Quotient entspricht dem Ratio der Wahrscheinlichkeiten von $\tilde{K}$ und K. Hier wird zur Berechnung eine geometrische Verteilung mit Dichte $p(x) = p(1-p)^{x-1}$, $p \in [0,1]$, angenommen, sodass sich dieser Quotient wie folgt berechnet:

$$\frac{p(\tilde{K})}{p(K)} = \frac{p(1-p)^{\tilde{K}-1}}{p(1-p)^{K-1}} = (1-p)^{\tilde{K}-K},$$

wobei hier $p = 0.5$ gewählt wurde.

3.3. Der Circular Binary Segmentation Algorithmus als Vergleichsmethode

Der Circular Binary Segmentation (kurz CBS) Algorithmus ist eine für Microarray-Daten entwickelte Changepoint-Methode. Die Daten bestehen dabei meist aus den log-Ratios der Testprobe (z.B. krankem Gewebe) gegen eine Referenz (z.B. gesundem Gewebe) sowie ihren Lokationen auf dem Chromosom. Das Ziel des CBS Algorithmus ist die Identifikation der Changepoints und somit auch der Segmente, welche ein erhöhtes oder verringertes log-Ratio besitzen (vgl. auch Kapitel 3.1.3).

Der CBS Algorithmus baut dabei auf den Binary Segmentation Algorithmus auf. Hier wird zunächst nur ein CP gesucht bzw. identifiziert. Das Vorgehen kann dabei wie folgt beschrieben werden:

Seien $X_1, \ldots, X_N$ Zufallsvariablen mit Realisierungen $x_1, \ldots, x_N$, welche die log-Ratios

(oder im Falle einer Proteinzeitreihe die Lichtintensitäten) der N Lokationen repräsentieren. Weiter seien $\gamma_i = x_1 + \ldots + x_i$ die partiellen Summen, $i = 1, \ldots, N$. Zur Identifikation eines CPs an einer bestimmten Stelle i, kann unter der Annahme einer Normalverteilung mit bekannter Varianz eine Likelihood-Ratio-Teststatistik mit dem Hypothesenpaar H_0*: Es liegt kein Changepoint vor* gegen H_1*: Es liegt genau ein Changepoint an einer unbekannten Stelle i vor* verwendet werden, welche wie folgt definiert ist:

$$Z_B = \max_{1 \leq i \leq N} |Z_i| \quad \text{mit} \quad Z_i = \left\{ \frac{1}{i} + \frac{1}{N-i} \right\}^{-1/2} \left\{ \frac{\gamma_i}{i} - \frac{\gamma_N - \gamma_i}{N-i} \right\}.$$

Die Nullhypothese, dass kein CP vorhanden ist, kann genau dann verworfen werden, wenn die Teststatistik Z_B größer als das obere α-Quantil der Verteilung von Z_B unter H_0 ist. Dieses α-Quantil kann dabei mit Hilfe von MC-Simulationen und unter Nutzung der Approximation der „tail probabilities“ (Siegmund, 1986, [48]) berechnet werden. Wird die Nullhypothese verworfen und kann man somit davon ausgehen, dass ein CP vorhanden ist, so kann die Stelle des CPs durch den Index i geschätzt werden, für den die Teststatistik maximal wird.

Da im allgemeinen Fall die Varianz nicht als bekannt vorausgesetzt werden kann, wird diese durch eine Schätzung aus den Daten ersetzt. Dadurch kann nun die Teststatistik Z_B nicht verwendet werden. An ihrer Stelle wird folgende Teststatistik T genutzt:

$$T_{ij} = \frac{\overline{x}_{ij} - \overline{z}_{ij}}{s_{ij}\{(j-i)^{-1} + (n-j+i)^{-1}\}^{1/2}} \quad \text{und} \quad T = \max_{1 \leq i < j \leq N} |T_{ij}|,$$

wobei $\overline{x}_{ij} = \frac{x_{i+1}+\ldots+x_j}{j-i}$, $\overline{z}_{ij} = \frac{x_1+\ldots+x_i+x_{j+1}+\ldots+x_N}{N-j+i}$ sowie s_{ij}^2 der zugehörige mittlere quadratische Fehler (MSE; als Schätzer der Varianz aus den Daten) ist. An der Formel für T_{ij} ist gut zu erkennen, dass das Mittel der Beobachtungen von $i+1$ bis j mit dem Mittel der restlichen Beobachtungen verglichen wird. Dies resultiert aus der Vorstellung des CBS, dass das Segment als Kreis angesehen wird, d.h. das betrachtete Segment ist an der ersten und letzten Beobachtung „verbunden“, sodass ein Kreis entsteht. Somit kann das Testproblem auch als Test auf zwei Kreisbögen mit unterschiedlichen Mittelwerten aufgefasst werden. Testentscheidung kann zum einen wie zuvor über MC-Simulationen und das α-Quantil oder über den p-Wert erfolgen, d.h. ist der p-Wert kleiner als das vorgegeben Niveau α, so kann die Nullhypothese, dass kein CP existiert, verworfen werden. Anschließend (im Falle einer Ablehnung der Nullhypothese) werden die CPs über die Indizes geschätzt, welche die Teststatistik maximieren. Wenn somit $j < N$ ist, so wird an

dieser Stelle nicht nur ein CP, sondern gleich 2 CPs geschätzt.
Dieses Vorgehen kann für bekannte Varianz auch wieder auf die Likelihood-Ratio-Teststatistik angewendet werden. Für nähere Informationen sei an dieser Stelle jedoch auf Olshen und Venkatraman (2004, [36]) verwiesen.
Bei dem obigen Vorgehen kann nun der sogenannte *Edge Effekt* auftreten. Dieser beschreibt den Fall, dass bei der Schätzung der CPs entweder i „nahe“ an 1, oder j „nahe“ an N ist. Man würde also von 2 CPs ausgehen, obwohl nur ein CP vorhanden ist. Um dieses Problem zu behandeln wird in diesen Fällen ein CP entfernt, sofern er nicht nochmals durch die Daten bestätigt wird.
Ein Vorteil des CBS Algorithmus ist, dass zwar stets von einer Normalverteilung der Daten ausgegangen wurde, dies aber mittels eines Permutationsansatzes umgangen werden kann. Nähere Informationen zu dem Permutationsansatz sowie zu zwei Modifikationen für große Datenmengen (eine Stopping-Rule sowie die Hybrid-p-Wert-Methode) können Olshen und Venkatraman (2004, [36]) entnommen werden.

3.4. Analyse der Daten und Diskussion der Ergebnisse

Zunächst werden **ChIP-Seq-Daten** betrachtet, wobei sich hier, wie bereits in Kapitel 3.1.1 erwähnt, auf die Basenpaare 0 bis 55 000 000 auf Chromosom 3 beschränkt wird und sind in Abbildung 3.1 zu sehen. Anschließend wird gezeigt, dass auch eine Anwendung auf **Proteinzeitreihen** möglich und auch sinnvoll ist. Abschließend folgt eine Diskussion der Ergebnisse sowie eine Erweiterung mittels Reversible Jump.

3.4.1. Analyse der ChIP-Seq-Daten

Die Anwendung des Bayesschen hierarchischen Modells für ChIP-Seq-Daten kann mit einem einfachen Beispiel, dem Beta-Binomial-Modell, motiviert werden. Dafür werden zunächst die Daten in Fenster mit einer jeweiligen Breite von 50 Beobachtungen unterteilt. Hier wird, wie oben beschrieben, Chromosom 3 als Beispiel genutzt. Anschließend wird nun für jedes Fenster ein Binomialtest mit der Nullhypothese $H_0 : p = 0.5$ gegen die Alternative $H_1 : p \neq 0.5$ durchgeführt. Hier ist für alle Fenster der p-Wert kleiner als

0.05 ($p < 0.05$), abgesehen von zwei Fenstern am Ende von Chromosom 3. In diesen zwei Fenstern sind allerdings sowohl die Lamin B Counts als auch die Backgroundcounts gleich Null und somit ist der Test hier nicht definiert. Das Ergebnis ist grafisch in Abbildung 3.3 dargestellt. Mit Hilfe dieses Beispiels wird deutlich, dass mit dem Binomial-Test zu

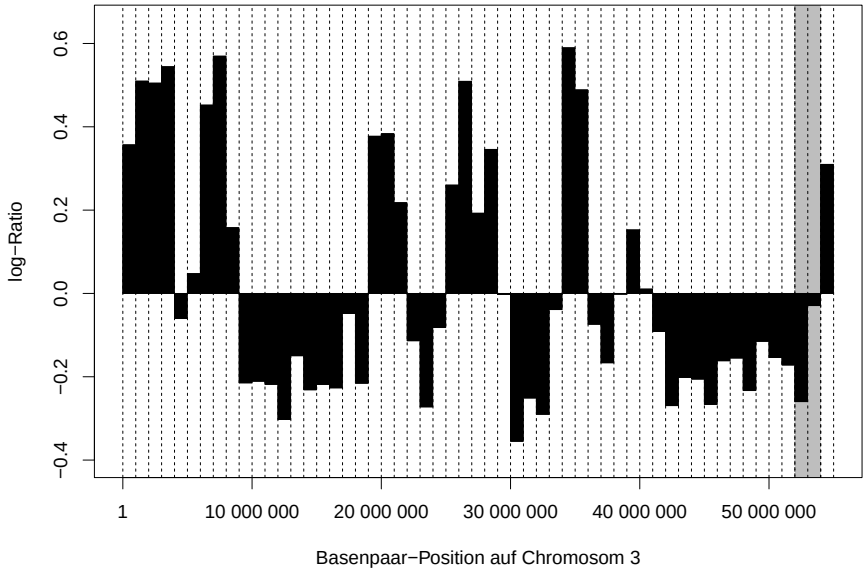

Abbildung 3.3: Grafische Darstellung der ChIP-Seq-Daten, welche in Fenster mit jeweils 50 Beobachtungen unterteilt sind. Dabei sind die grau hinterlegten Flächen, die zwei Fenster, in welchen der Binomial-Test nicht definiert ist.

viele CPs identifiziert werden. Diese Tatsache und die fehlende Möglichkeit mit fehlenden Beobachtungen, d.h. Counts gleich Null, umgehen zu können, ist die Motivation ein anderes Modell aufzustellen und zu nutzen. Das Bayessche hierarchische Modell, welches in Kapitel 3.2.1 beschrieben wurde, kann an dieser Stelle mit beiden Problemen umgehen.
Beim Bayesschen hierarchischen Modell wird von einer festen Anzahl an Segmenten, K, ausgegangen. Als Startsequenzen werden im Folgenden zwei verschiedene Sequenzen für die Implementierung aus Kapitel 3.2.3 verwendet. Die erste Start-Segmentierung stammt aus der Berechnung des CBS Algorithmus, die andere Startsequenz ist hingegen eine zufällig gewählte Segmentierung der Daten. Für das MCMC-Verfahren werden anschließend 50 000 Iterationen und ein Burn-In von 1 000 Iterationen gewählt.
Die Ergebnisse des hierarchischen Modells werden mit denen des CBS Algorithmus vergli-

chen. Da der CBS Algorithmus $K = 17$ Segmente geschätzt hat, werden wir im Weiteren auch stets von einer festen Segmentanzahl von 17 ausgehen.
Das Resultat der Anwendung des Bayesschen hierarchischen Modells mit der CBS-Startsequenz ist in Abbildung 3.4 zu finden. In Abbildung 3.4 a) ist dabei die logarithmierte

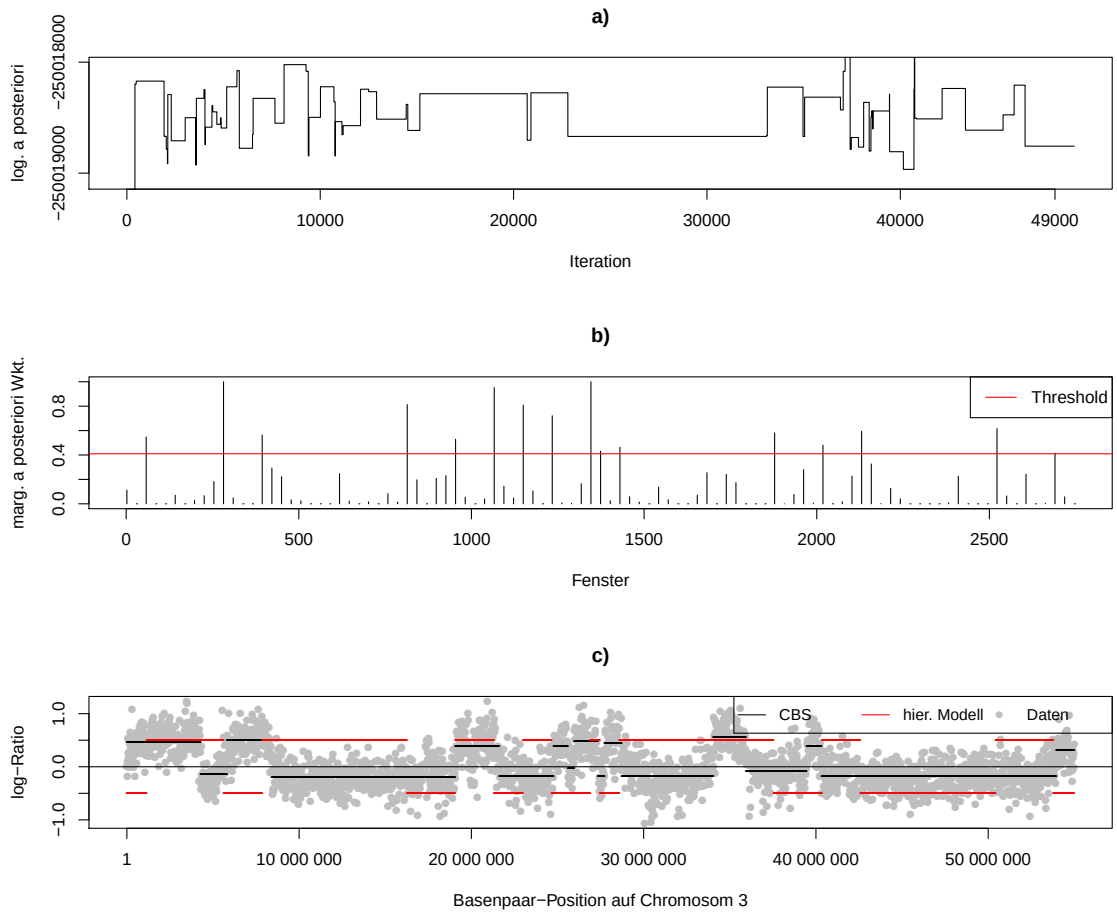

Abbildung 3.4: Ergebnis der 49 000 MCMC-Iterationen (50 000 Iterationen insgesamt, jedoch davon 1 000 Iterationen Burn-In) des Random Walk Metropolis Hastings Algorithmus mit der CBS-Startsequenz: a) logarithmierte a posteriori, b) marginale a posteriori Wahrscheinlichkeit für mindestens einen CP auf einem Gitter mit 100 Fenstern und c) die resultierende Segmentierung, basierend auf den 16 CPs mit höchster marginaler a posteriori Wahrscheinlichkeit aus b) im Vergleich zur Segmentierung des CBS Algorithmus.

a posteriori in Abhängigkeit der Iterationen zu sehen. Dabei ist gut zu erkennen, dass das Verhalten der Markov Kette gut ist und nach einem kurzen Burn-In die stationäre

Verteilung erreicht hat.
In Teil b) derselben Abbildung ist hingegen die marginale a posteriori Wahrscheinlichkeit für einen CP in den jeweiligen Fenstern zu sehen, wobei insgesamt 100 Fenster mit einheitlicher Breite gewählt wurden. Anschließend wird ein Grenzwert, der sogenannte *Threshold*, so gewählt, dass die besten 16 (variablen) CPs gewählt wurden. Dies bedeutet es werden die 16 Fenster gewählt, für welche die marginale a posteriori Wahrscheinlichkeit am höchsten war. Dabei sollte beachtet werden, dass hier 16 CPs gewählt wurden, da $K = 17$ Segmente als fest angenommen worden waren, was eine CP-Anzahl von 18 ergibt. Jedoch sind der erste und der letzte CP fest (erste bzw. letzte Beobachtung), sodass 16 variable CPs eine Segmentanzahl von 17 ergibt.
Die Segmentierung, welche aus der Wahl der CPs mit Hilfe eines Thresholds für die marginale a posteriori Wahrscheinlichkeit resultiert, ist schließlich in Abbildung 3.4 c) zu sehen. Dabei wurden die errechneten Segmente des hierarchischen Modells auf +/- 0.5 gesetzt. Man könnte an dieser Stelle auch das Segmentmittel nutzen (wie im CBS Algorithmus), jedoch ist hier lediglich „Lamin B erhöht“ oder „Lamin B nicht erhöht“ von Bedeutung. Daher können die Segmente in Abhängigkeit von w_k, $k = 1, \ldots, K$, auf eine positive (wenn $w_k = 1$) oder negative (wenn $w_k = 0$) Konstante gesetzt werden, hier +/-0.5.
Zum Vergleich ist in Abbildung 3.4 c) auch das Ergebnis der Segmentierung des CBS Algorithmus abgetragen. Es ist deutlich zu erkennen, dass im Vergleich zur Segmentierung des CBS Algorithmus, einige CPs beibehalten oder nur leicht verschoben wurden. Andere hingegen wurden entfernt und an anderer Stelle wieder hinzugefügt. Daraus folgt, dass an manchen Abschnitten von Chromosom 3 nun mehr bzw. weniger Segmente vorhanden sind im Vergleich zur Segmentierung des CBS Algorithmus. Dies ist zum Beispiel an dem zusätzlichen Segment in Abschnitt von 20 000 000 bis 25 000 000 Basenpaaren (bps) zu erkennen. Aufgrund der festgelegten Segmentanzahl fehlt an anderer Stelle ein CP, hier z.B. im Abschnitt zwischen 35 000 000 und 40 000 000 bps. Eine weitere Folge ist, dass auch die Zuteilung durch den Vektor $w = (w_1, \ldots, w_K)'$ für die nachfolgenden Segmente verschoben sein kann.
Analog wurden die Daten mit einer zufälligen Startsequenz analysiert. Die Ergebnisse dazu sind in Abbildung 3.5 dargestellt. In Abbildung 3.5 a) ist erneut die logarithmierte a posteriori Verteilung abgetragen. Auch in diesem Fall erreicht sie nach einem kurzen

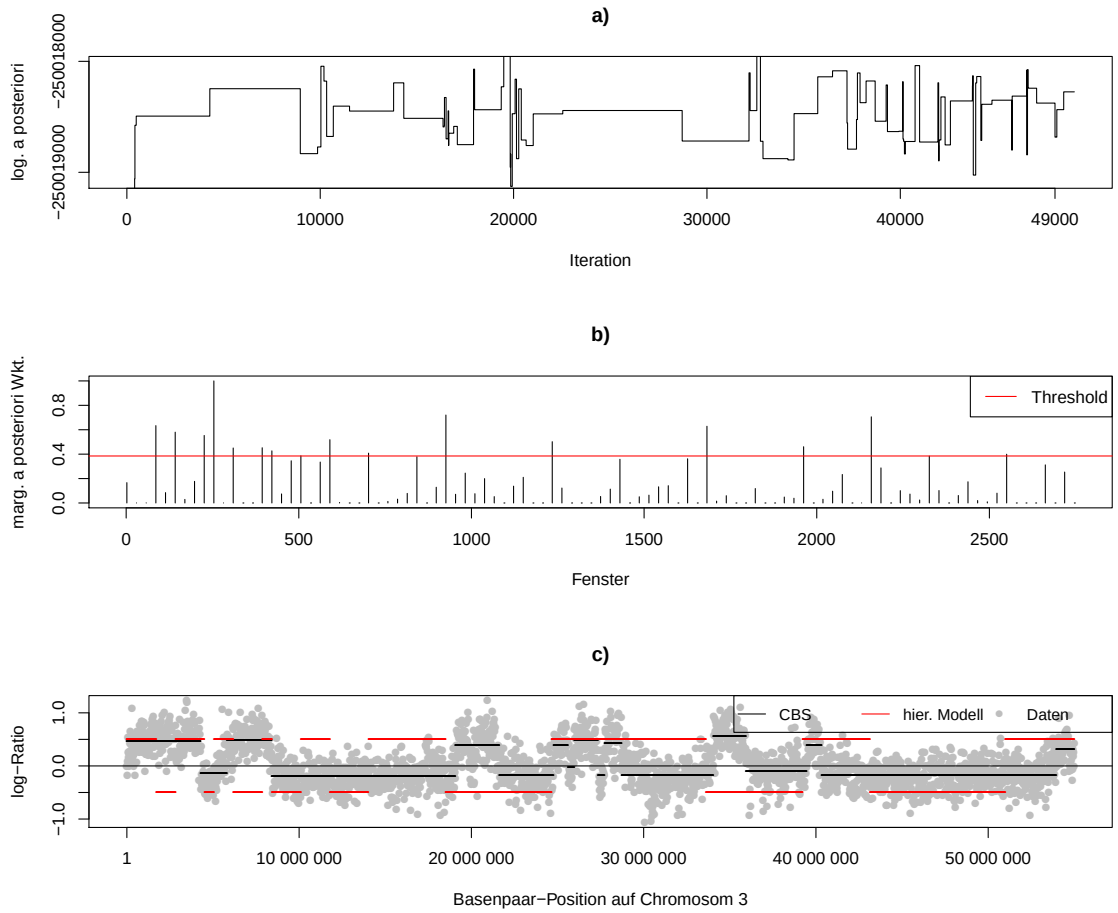

Abbildung 3.5: Ergebnis der 49 000 MCMC-Iterationen (50 000 Iterationen insgesamt, jedoch davon 1 000 Iterationen Burn-In) des Random Walk Metropolis Hastings Algorithmus mit einer zufälligen Startsequenz: a) logarithmierte a posteriori, b) marginale a posteriori Wahrscheinlichkeit für mindestens einen CP auf einem Gitter mit 100 Fenstern und c) die resultierende Segmentierung, basierend auf den 16 CPs mit höchster marginaler a posteriori Wahrscheinlichkeit aus b) im Vergleich zur Segmentierung des CBS Algorithmus.

Burn-In von 1 000 Iterationen ihre stationäre Verteilung und die Markov Kette zeigt ein gutes Verhalten. Die marginale a posteriori Wahrscheinlichkeit für die einzelnen Fenster ist in Abbildung 3.5 b) zu sehen. Hier wurde wie zuvor ein Threshold gewählt, sodass die besten 16 variablen CPs ausgewählt wurden. Die aus diesen CPs resultierende Segmentierung ist abschließend in Abbildung 3.5 c) im Vergleich zur Segmentierung des CBS

Algorithmus zu sehen. Auch hier zeigt sich erneut, dass die CPs aus dem Bayesschen hierarchischen Modell und dem CBS Algorithmus in Teilen der Daten sehr ähnlich sind. Dennoch ist auch hier (speziell am Anfang der Daten) zu erkennen, dass zusätzliche Segmente eingefügt wurden, was wiederum bei fester Anzahl an Segmenten bedeutet, dass diese Segmente an anderer Stelle fehlen.

Auch wenn hier ersichtlich ist, dass das MCMC-Mixing nicht ideal arbeitet, so beeinflusst die Wahl der Startsequenz die resultierende Segmentierung nicht. Der Algorithmus arbeitet für beide Startsequenzen gut und führt zu sinnvollen Segmentierungen der Daten.

Da bis jetzt nur das Verhalten der logarithmierten a posteriori Verteilung betrachtet wurde, wird nun noch das Konvergenzverhalten der Markov Kette für vereinzelte Parameter betrachtet. Dafür wurden zum einen Geweke's Diagnostik (Geweke, 1992, [12]) für die Parameter μ_0, μ_1 und ρ berechnet sowie die Diagnostik aus Heidelberger und Welch (1983, [18]). Eine kurze Vorstellung der Grundidee beider Diagnostiken ist in Kapitel A.4 in Anhang A zu finden. Die Ergebnisse zur Konvergenzdiagnostik sind in Tabelle 3.3 zu finden. Dabei ist kein Anzeichen für eine fehlende Konvergenz vorhanden. Einzig für den Parameter ρ konnte mit Hilfe der Z-Statistik der Geweke Diagnostik für die CBS-Startsequenz keine Konvergenz zum Niveau von 5% ($\alpha = 0.05$) nachgewiesen werden. Der Cramer van Mises Test der Heidelberger und Welch Diagnostik spricht hingegen bei keinem Parameter und keiner Startsequenz gegen eine Konvergenz.

Zusammenfassend konnte bis hier hin gezeigt werden, dass sowohl der CBS Algorithmus als auch das Bayessche hierarchische Modell flexibel sind und bzgl. der Daten zu sinnvollen Ergebnissen führen. Im Gegensatz zum CBS Algorithmus liefert das Bayessche hierarchische Modell nicht nur Punktschätzer, sondern ein Wahrscheinlichkeitsmodell der unbekannten Segmentierung und erlaubt daher eine variable Anzahl an Segmenten, auch wenn für einen besseren Vergleich zum CBS Algorithmus hier bislang nur die Punktschätzer betrachtet wurden.

Um eine variable Anzahl an Segmenten zu betrachten und dem Problem zu entgehen, dass Segmente in Teilen der Daten fehlen, da sie an anderer Stelle hinzugefügt wurden, wird die Analyse nun für verschiedene Wahlen von K wiederholt, hier $K \in \{15, \ldots, 20\}$. Die Analyse erfolgte dabei für jedes K analog zu der bisherigen Analyse für $K = 17$.

Eine Zusammenfassung der Lage- und Streuungsmaße der a posteriori Verteilung ist in

Geweke's Diagnostik

Aufteilung im ersten Fenster = 0.1, Aufteilung im zweiten Fenster = 0.5

CBS-Startsequenz		zufällige Startsequenz	
Parameter	Z-Statistik	Parameter	Z-Statistik
μ_0	-1.715	μ_0	-3.249
μ_1	-1.652	μ_1	-3.195
ρ	1.561	ρ	2.810

Heidelberger und Welch Diagnostik

Stationaritäts-Test (Cramer van Mises Statistik)

CBS-Startsequenz		zufällige Startsequenz	
Parameter	Cramer van Mises Test (p-Wert)	Parameter	Cramer van Mises Test (p-Wert)
μ_0	passed (0.1038)	μ_0	passed (4.13e-01)
μ_1	passed (0.1329)	μ_1	passed (3.68e-01)
ρ	passed (0.1291)	ρ	passed (2.96e-01)

Tabelle 3.3: Konvergenzdiagnostiken der Parameter μ_0, μ_1 and ρ für die Analyse der ChIP-Seq-Daten in Abhängigkeit der gewählten Startsequenz (dabei steht passed für „bestanden“).

Tabelle 3.4 zu finden. Dabei ist die Varianz für $K = 18$ minimal, der Median und auch der Mittelwert ist hingegen für eine Segmentanzahl von $K = 19$ maximal. Man kann daraus schließen, dass eine leichte Präferenz für höhere Werte als beim CBS Algorithmus existiert. Die Ergebnisse für $K = 19$ sind in Abbildung 3.6 grafisch dargestellt. Auch hier ist ersichtlich, dass die logarithmierte a posteriori in Abbildung 3.6 a) ein gutes Verhalten zeigt. In Abbildung 3.6 b) ist erneut die marginale a posteriori Wahrscheinlichkeit abgetragen sowie der Threshold, welcher sich für $K = 19$ Segmente (18 variable CPs) ergibt. In Teil c) der Abbildung ist abschließend wieder die resultierende Segmentierung im Vergleich zur Segmentierung des CBS Algorithmus eingezeichnet. Dabei fällt auf, dass der Hauptunterschied am Ende von Chromosom 3 liegt, bei 45 000 000 bis 50 000 000 bps. Hier

K	Minimum	Median	Mittelwert	Maximum	Varianz
15	-250 019 264	-250 018 560	-250 018 489	-250 017 328	80 807
16	-250 019 582	-250 018 488	-250 018 478	-250 017 562	67 407
17	-250 019 558	-250 018 477	-250 018 446	-250 018 044	40 924
18	-250 018 796	-250 018 446	-250 018 439	-250 017 982	29 326
19	-250 019 195	-250 018 383	-250 018 414	-250 017 871	33 680
20	-250 019 802	-250 018 387	-250 018 458	-250 017 910	58 230

Tabelle 3.4: Zusammenfassende Lage- und Streuungsmaße der logarithmierten a posteriori (50 000 Iterationen, inkl. einem Burn-In von 1 000 Iterationen) des Random Walk Metropolis Hastings Algorithmus mit zufälliger Startsequenz für unterschiedliche Wahlen von K.

ist ein zusätzliches Segment durch das Bayessche hierarchische Modell eingefügt worden. Die zwei CPs, welche dieses Segment umgeben, haben dabei eine auffällig hohe marginale a posteriori Wahrscheinlichkeit (siehe Abbildung 3.6 b)). Dieses zusätzliche Segment am Ende von Chromosom 3 kann hier auch durch die Daten bestätigt werden. Die zugrundeliegenden Daten zwischen 45 000 000 und 55 000 000 bps haben eine oszillierende Struktur. An der Stelle des neuen Segmentes kann eine leichte Welle nach oben erkannt werden. Die Vermutung liegt daher nahe, dass das Bayessche hierarchische Modell besser mit oszillierenden Strukturen in Segmenten von variabler Länge (kurz oder lang) umgehen kann als der CBS Algorithmus, sofern diese Oszillation nicht um Null schwankt. Dieses Ergebnis untermauert den Vorteil einer Modellierung kurzer Segmente und die variable Wahl von K, wie es hier gezeigt wurde. Beides kann als Vorteil gegenüber dem CBS Algorithmus angesehen werden.

An dieser Stelle sei noch einmal darauf hingewiesen, dass wir hier beispielhaft nur Chromosom 3 analysiert haben. Der Algorithmus kann natürlich auf alle Chromosome erweitert sowie auf andere Arten von Daten angewendet werden. Letzteres erfolgt im folgenden Kapitel durch eine Analyse von Proteinzeitreihen mit Hilfe des Bayesschen hierarchischen Modells.

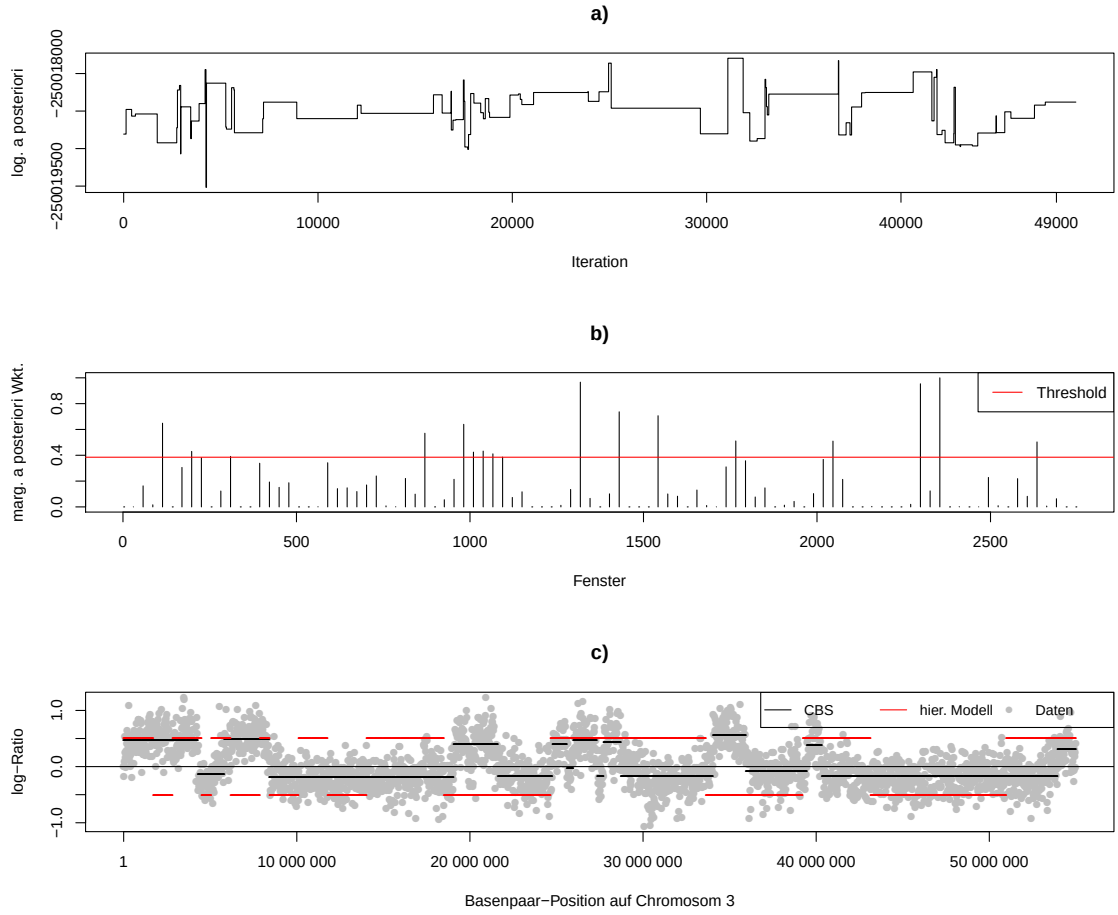

Abbildung 3.6: Ergebnis der 49 000 MCMC-Iterationen (50 000 Iterationen insgesamt, jedoch davon 1 000 Iterationen Burn-In) des Random Walk Metropolis Hastings Algorithmus mit einer zufälligen Startsequenz und $K = 19$: a) logarithmierte a posteriori, b) marginale a posteriori Wahrscheinlichkeit für mindestens einen CP auf einem Gitter mit 100 Fenstern und c) die resultierende Segmentierung, basierend auf den 18 CPs mit höchster marginaler a posteriori Wahrscheinlichkeit aus b) im Vergleich zur Segmentierung des CBS Algorithmus.

3.4.2. Analyse der Proteinzeitreihen

Wie bereits in Kapitel 3.1.3 beschrieben, ist das Ziel der Proteinzeitreihenanalyse ebenfalls die Identifikation von CPs. In verschiedenen Abschlussarbeiten wurden diverse Methoden, wie u.a. der CBS Algorithmus, angewendet sowie verglichen. Dabei wurde gezeigt, dass

die Methoden nicht nur technisch „anwendbar“ sind, sondern auch gute Ergebnisse liefern. Daher liegt es nahe, das obige hierarchische Modell auf die Proteinzeitreihen anzuwenden. Dies soll nun exemplarisch für je eine Zeitreihe pro Experimenteneinstellung (vgl. Tabelle 3.2) durchgeführt werden.

Zunächst muss das Problem der Dateneingabe betrachtet werden. Das Modell wurde für Countdaten aufgestellt, sodass es Signal Counts und Backgroundcounts modelliert. Im Falle einer Proteinzeitreihe ist jedoch nur ein Wert pro Zeitpunkt gemessen worden, die Lichtintensität im Spot zu diesem Zeitpunkt. Diese Angabe kann somit als Signal betrachtet werden und als c_i im Modell eingegeben werden. Nun fehlt noch der Background b_i. Hier ist es naheliegend eine Art Background passend zur Proteinzeitreihe zu konstruieren. Im Falle der Proteinmessung ist der Background das Signal, welches vorhanden ist, wenn kein Protein aktiv ist. Somit kann der Background als minimale Lichtintensität der Proteinzeitreihe plus einen zufälligen Faktor aufgefasst werden. Dies geschieht für die Analyse der Proteinzeitreihe mit Hilfe einer Exponentialverteilung:

$$b_i = \min_j c_j + \epsilon_i,$$

wobei ϵ_i eine Zufallszahl aus einer Exponentialverteilung mit Parameter $0.5 + \frac{1}{\max_i c_i - \min_i c_i}$ ist, d.h. $\epsilon_i \sim \text{Exp}(0.5 + \frac{1}{\max_i c_i - \min_i c_i}), i = 1, \ldots, N$. An dieser Stelle wird die Exponentialverteilung genutzt, da auf Grund physikalischer Begebenheiten die Emission durch eine Exponentialverteilung beschrieben werden kann.

Weiter sollte noch beachtet werden, dass sich, auf Grund der Anzahl an Beobachtungspunkten, nicht immer exakt 100 Fenster, wie bei den ChIP-Seq-Daten, bei der Berechnung der marginalen a posteriori Wahrscheinlichkeit realisieren lassen. Daher wird im Folgenden stets „ca. 100 Fenster“ geschrieben.

Somit kann nun das hierarchische Modell angewendet werden. Auch hier wird der CBS Algorithmus für einen Vergleich herangezogen, d.h. als Segmentanzahl K wird auch hier das Ergebnis des CBS Algorithmus verwendet. Weiter werden 100 000 Iterationen durchgeführt sowie ein Burn-In von 1 000 Iterationen genutzt.

Das Ergebnis für eine Proteinzeitreihe aus Experiment Nr. 1 (100 Watt und 50 ms) ist in Abbildung 3.7 zu sehen. Dabei ist in a) die logarithmierte a posteriori dargestellt. Es kann festgestellt werden, dass das Verhalten der Markov Kette gut ist und eine stationäre

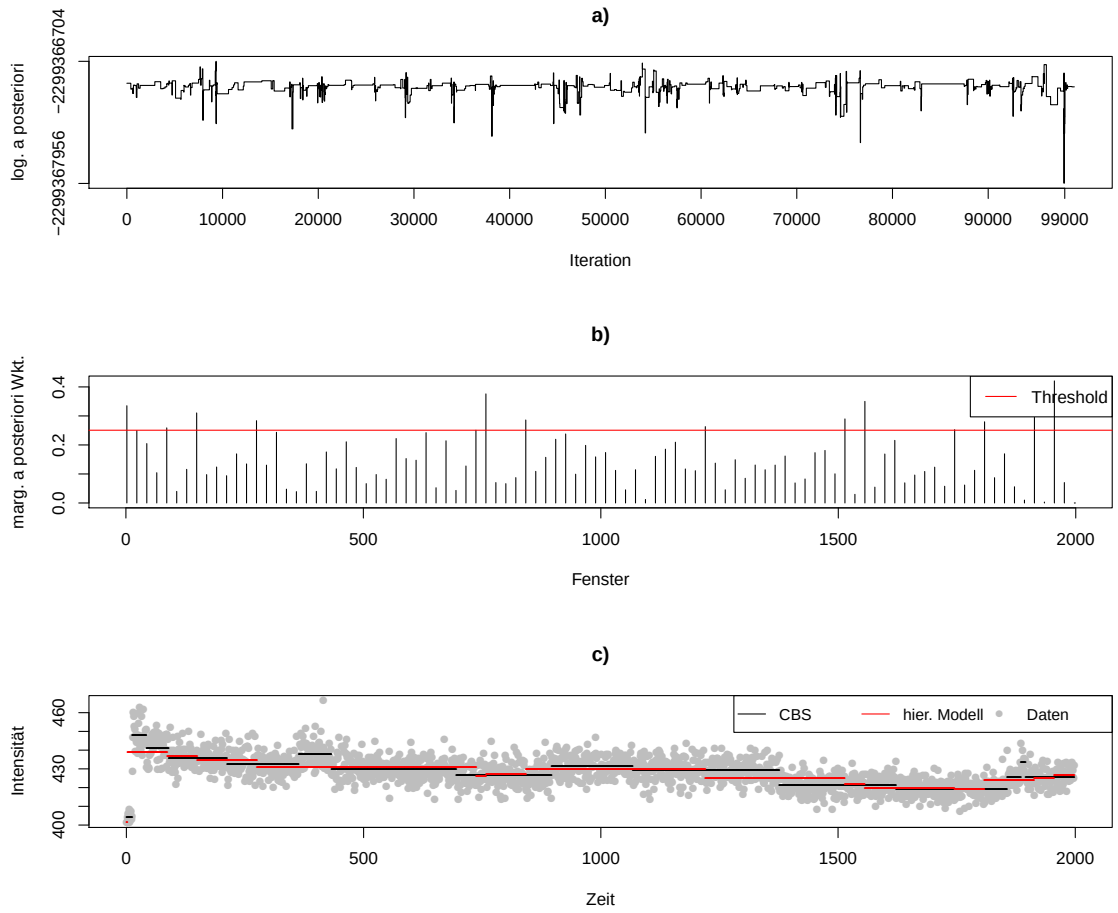

Abbildung 3.7: Ergebnis der 99 000 MCMC-Iterationen (100 000 Iterationen insgesamt, jedoch davon 1 000 Iterationen Burn-In) des Random Walk Metropolis Hastings Algorithmus mit einer zufälligen Startsequenz für eine exemplarische Proteinzeitreihe aus Experiment 1: a) logarithmierte a posteriori, b) marginale a posteriori Wahrscheinlichkeit für mindestens einen CP auf einem Gitter mit ca. 100 Fenstern und c) die resultierende Segmentierung, basierend auf den 18 CPs mit höchster marginaler a posteriori Wahrscheinlichkeit aus b) im Vergleich zur Segmentierung des CBS Algorithmus.

Verteilung angenommen wurde. In Abbildung 3.7 b) ist die marginale a posteriori Wahrscheinlichkeit für einen CP in den jeweiligen Fenstern abgetragen. Auch hier ist analog zu den Abbildungen vorher der Grenzwert (Threshold) in rot eingezeichnet, durch welchen die CPs ausgewählt und die endgültige Segmentierung festgelegt werden. Diese ist in Abbildung 3.7 c) im Vergleich zur Segmentierung durch den CBS Algorithmus abgetra-

gen. Hier zeigt sich ein ähnliches Bild zu den Ergebnissen der ChIP-Seq-Daten-Analyse. Es werden einige CPs des CBS Algorithmus nur leicht verschoben, wohin gegen andere entfernt bzw. in manchen Segmenten zusätzliche CPs hinzugefügt werden. Diese Tatsache führt dazu, dass an einigen Stellen der Zeitreihe eine feinere Segmentierung vorliegt, wohingegen an anderen Stellen die Segmentierung (zwangsläufig) gröber ist. Dies ist zum Beispiel zwischen Zeitpunkt 1 500 und 1 800 zu sehen. Hier ist die Segmentierung etwas feiner. Dadurch ist an anderer Stelle eine gröbere Segmentierung zu erkennen, z.B. zwischen Zeitpunkt 250 und 750.
Die Ergebnisse für jeweils eine Proteinzeitreihe aus den anderen drei Experimenten sind im Anhang C in den Abbildungen 2, 3 und 4 zu finden.
An dieser Stelle kann überlegt werden, den Threshold für die Analyse von Proteinzeitreihen so zu verändern, dass sich eine höhere und vom CBS Algorithmus unabhängige Anzahl an CPs ergibt. Bis jetzt wurde der Threshold so bestimmt, indem die CPs mit höchster marginaler a posteriori Wahrscheinlichkeit gewählt wurden, um die gleiche Anzahl an Segmenten zu erhalten, welche sich aus dem CBS Algorithmus ergeben haben. Eine weitere Möglichkeit ist, den Threshold als halbe Differenz der maximalen und minimalen marginalen a posteriori Wahrscheinlichkeit zu wählen. Somit ist die resultierende Anzahl der Segmente nicht auf die Anzahl der Segmente aus dem CBS Algorithmus beschränkt (i.A. wird sie größer sein).
In den Abbildungen 3.8, 3.9 sowie 5 und 6 in Anhang C sind nun die Ergebnisse auf Basis des neuen Threshold im Vergleich zu denen des alten Grenzwertes für die vier exemplarischen Proteinzeitreihen zu sehen. Dabei ist in Teil a) von Abbildung 3.8 die logarithmierte a posteriori zu sehen, in Teil b) die marginale a posteriori Wahrscheinlichkeit für einen CP in den jeweiligen Fenstern und in c) die aus den CPs, basierend auf dem Threshold, resultierende Segmentierung. Hier ist mit Threshold 1 und der roten Linie der ursprüngliche Threshold gemeint, welcher die gleiche Anzahl an Segmenten ergibt die der CBS Algorithmus geschätzt hat. Mit Threshold 2 und der blauen Kennzeichnung ist hingegen der neue Threshold bezeichnet.
In Abbildung 3.8 b) ist zu erkennen, dass der neue Threshold (Threshold 2) etwas kleiner ist als der Threshold zuvor. Dadurch werden mehr CPs erfasst und somit erhöht sich die Anzahl an Segmenten. Diese sind in der Segmentierung in Abbildung 3.8 c) dargestellt.

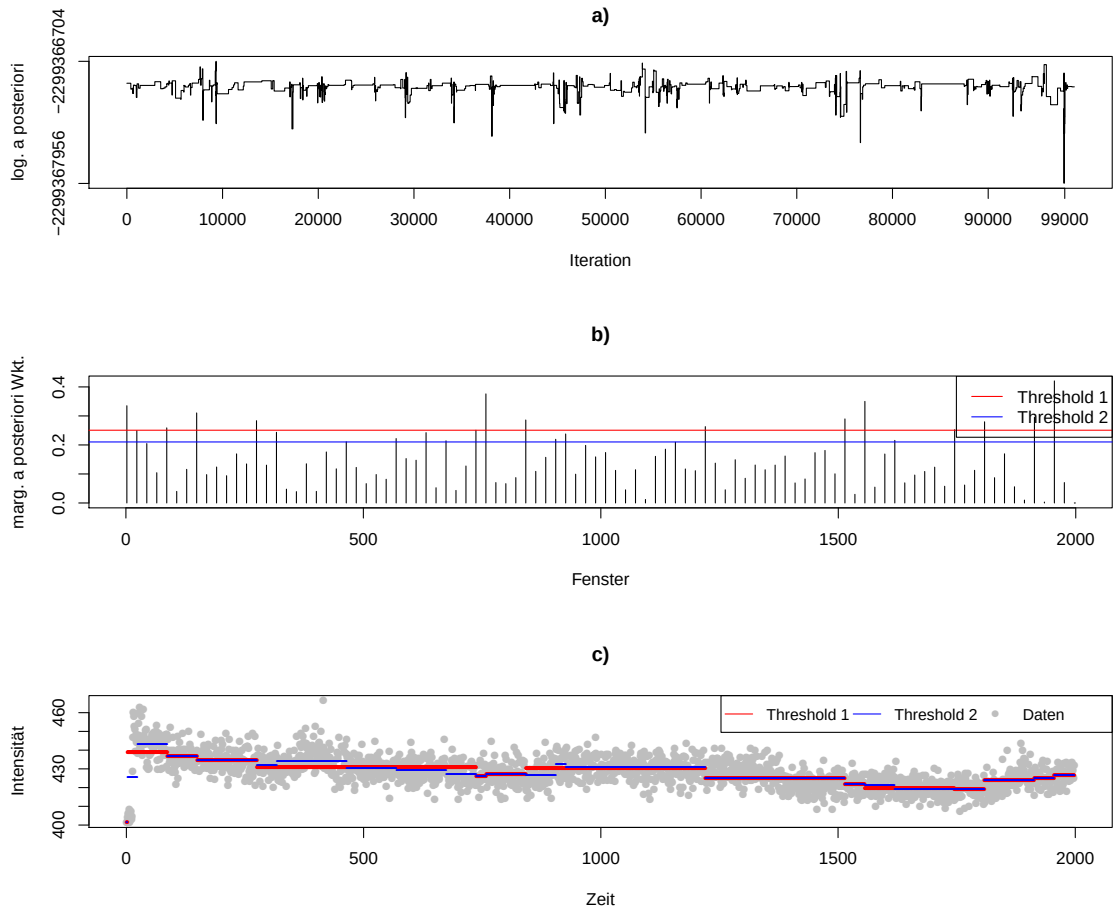

Abbildung 3.8: Ergebnis der 99 000 MCMC-Iterationen (100 000 Iterationen insgesamt, jedoch davon 1 000 Iterationen Burn-In) des Random Walk Metropolis Hastings Algorithmus mit einer zufälligen Startsequenz und beiden Thresholds für eine exemplarische Proteinzeitreihe aus Experiment 1: a) logarithmierte a posteriori, b) marginale posteriori Wahrscheinlichkeit für mindestens einen CP auf einem Gitter mit ca. 100 Fenstern und c) die resultierende Segmentierung, basierend auf den CPs mit höchster marginaler a posteriori Wahrscheinlichkeit aus b) im Vergleich zur Segmentierung des CBS Algorithmus.

Auffällig ist hier, dass am Ende der Beobachtungsreihe die letzten drei Segmente gleich geblieben sind, wohingegen gerade im Bereich zwischen Zeitpunkt 250 und 750 eine deutlich feinere Segmentierung vorzufinden ist. Dadurch ist eine bessere Anpassung an das zirkulierende Verhalten der Proteinzeitreihe gegeben.

Für Experiment 4 ist das Verhalten ähnlich zu Experiment 1. Für Experiment 2 und 3 ist das Verhalten anders, da durch den CBS Algorithmus bereits eine große Anzahl an Segmenten vorgegeben wird (jeweils über 20 Segmente). In Abbildung 3.9 ist beispielhaft das Ergebnis von Experiment 2 zu sehen. Es ist deutlich zu erkennen, dass in diesem Fall

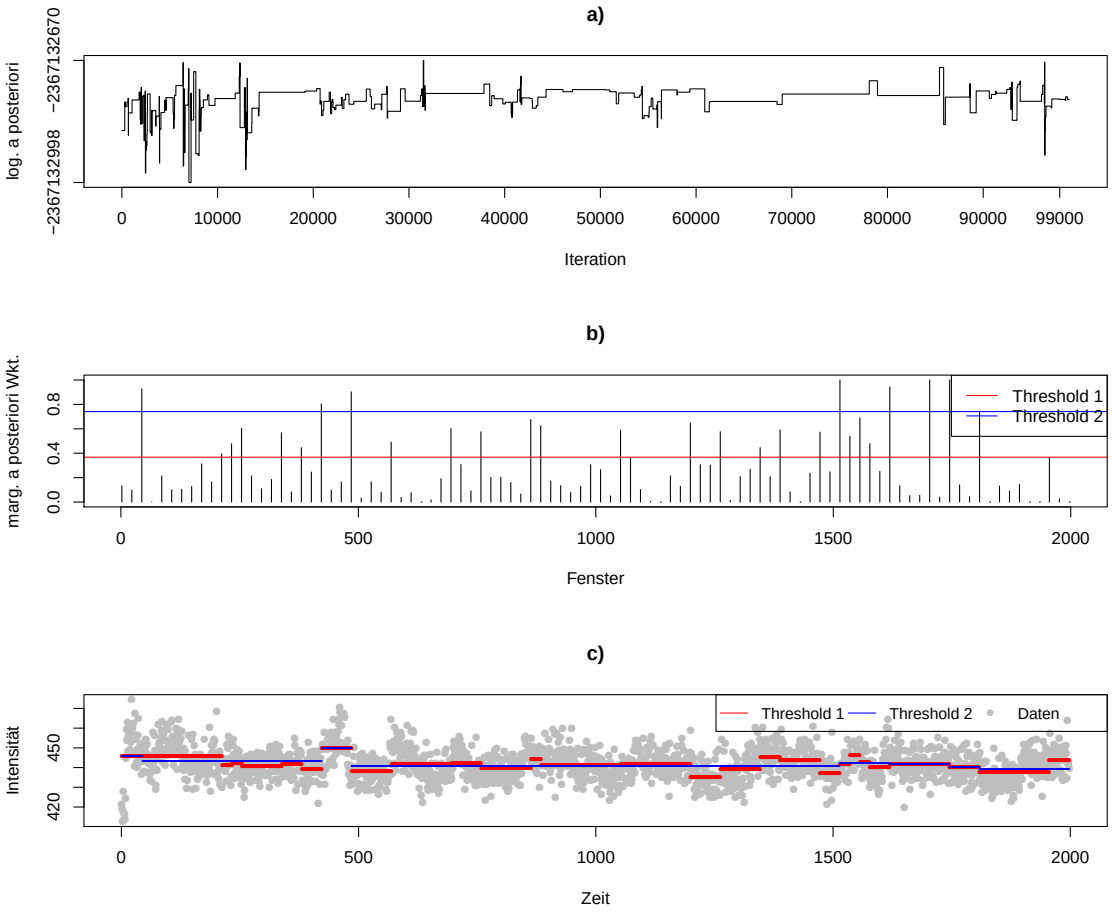

Abbildung 3.9: Ergebnis der 99 000 MCMC-Iterationen (100 000 Iterationen insgesamt, jedoch davon 1 000 Iterationen Burn-In) des Random Walk Metropolis Hastings Algorithmus mit einer zufälligen Startsequenz und beiden Thresholds für eine exemplarische Proteinzeitreihe aus Experiment 2: a) logarithmierte a posteriori, b) marginale a posteriori Wahrscheinlichkeit für mindestens einen CP auf einem Gitter mit ca. 100 Fenstern und c) die resultierende Segmentierung, basierend auf den CPs mit höchster marginaler a posteriori Wahrscheinlichkeit aus b) in Abhängigkeit vom Threshold.

der neue Threshold 2 größer als Threshold 1 ist. Dadurch ist die Segmentierung mit dem

alten Threshold feiner als mit dem neuen Threshold.

3.4.3. Erweiterung durch einen Reversible Jump Schritt

Eine weitere Möglichkeit, die Variabilität von K zu modellieren, ist die Erweiterung durch einen Reversible Jump Schritt, welcher in Kapitel 3.2.4 beschrieben wurde. Dieser erste Ansatz wird exemplarisch für die ChIP-Seq-Daten sowie für die Beispielproteinzeitreihe aus Experiment 1 mit $p_{Step} = (0.8, 0.1, 0.1)$ angewendet. Dabei werden für beide Daten jeweils 100 000 Iterationen mit einem Burn-In von 1 000 Iterationen genutzt. Als Startsequenz werden für die ChIP-Seq-Daten sowohl eine zufällige als auch die CBS-Startsequenz genutzt. Für die Proteinzeitreihe wird nur die zufällige Sequenz verwendet sowie eine Segmentanzahl von $K = 20$ angenommen.
Die Ergebnisse des Reversible Jump für die ChIP-Seq-Daten sind in den Abbildungen 3.10 sowie 7 in Anhang C zu finden. Das Ergebnis für die Proteindaten ist in Abbildung 8 in Anhang C zu finden.
Da nun innerhalb der Iterationen erlaubt ist, die Anzahl der Segmente zu verändern (durch Split/Merge), muss nach der letzten Iterationen die Anzahl der Segmente nicht mehr zwangsläufig identisch zu der Anzahl beim Start der Iterationen sein. Dadurch ist die Wahl des Thresholds wie zuvor nicht mehr sinnvoll. Alternativ können nun Thresholds wie der Median oder andere beliebige Quantile genutzt werden.
Dies kann in den oben genannten Abbildungen gesehen werden, z.B. 3.10. Hier wurde zum einen über die marginale a posteriori Wahrscheinlichkeit für einen CP die Segmentierung berechnet, wodurch die Anzahl an Segmenten K die gleiche Anzahl ist, wie vorher angenommen. Zum Anderen wurde sowohl der Median als auch das 75%-Quantil als Threshold zur Berechnung einer Segmentierung genutzt. Dabei kann festgestellt werden, dass die Segmentierung mit dem Median als Threshold die feinste Segmentierung erzeugt. Am Anfang sowie am Ende von Chromosom 3 sind recht kurze Segmente zu finden und somit eine sehr feine Segmentierung, wohingegen in der Mitte längere Segmente resultierten. Für die anderen zwei Segmentierungen ist am Anfang und Ende von Chromosom 3 ein ähnliches Bild vorzufinden, jedoch ist auffällig, dass in der Mitte des Chromosomenabschnitts ein langes Segment entstanden ist. Dieses lange Segment ist in Bezug auf die Daten jedoch

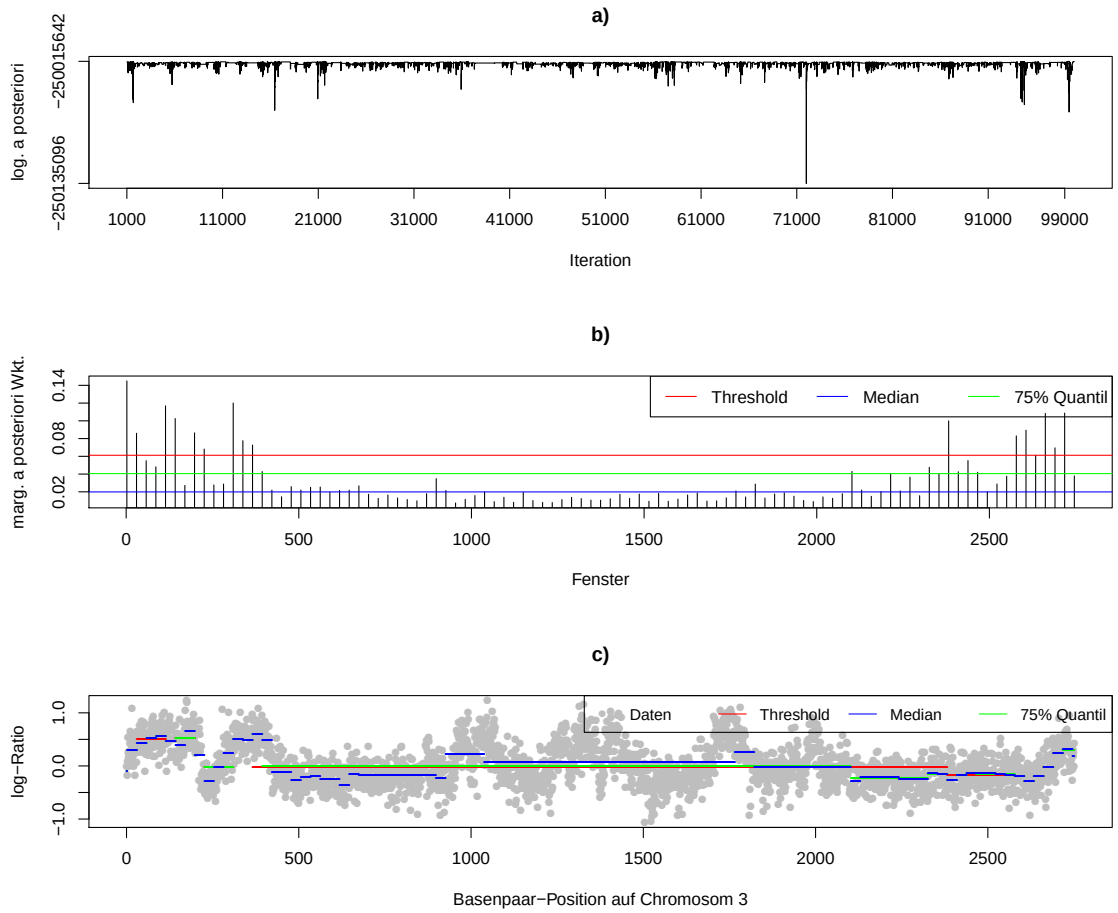

Abbildung 3.10: Ergebnis der 99 000 MCMC-Iterationen (100 000 Iterationen insgesamt, jedoch davon 1 000 Iterationen Burn-In) des Random Walk Metropolis Hastings Algorithmus mit einer zufälligen Startsequenz und dem Reversible Jump für die ChIP-Seq-Daten: a) logarithmierte a posteriori, b) marginale a posteriori Wahrscheinlichkeit für mindestens einen CP auf einem Gitter mit ca. 100 Fenstern und c) die resultierende Segmentierung, basierend auf den CPs mit höchster marginaler a posteriori Wahrscheinlichkeit aus b) im Vergleich zur Segmentierung des CBS Algorithmus.

keine gute Anpassung. Somit ist hier die Segmentierung über den Median die best angepasste Segmentierung.

Ein ähnliches Bild ist für den Reversible Jump der ChIP-Seq Daten mit der CBS-Startsequenz sowie für die den Reversible Jump der Proteindaten zu sehen. Bei Letzterem ist hingegen noch sehr auffällig, dass der Threshold sowie der Median und das 75%-Quantil

sehr nahe beieinander liegen. Dennoch bringt auch hier die Wahl des Medians die beste Anpassung und die feinste Segmentierung. Weiter ist in Tabelle 1 in Anhang B zu erkennen, dass sich für eine Startsegmentanzahl von 10 bzw. 20 nach einer kurzen Iterationenanzahl die geschätzte Anzahl an Segmenten kaum bzw. nicht mehr ändert. Darüber hinaus wird deutlich, dass je größer das Quantil gewählt wird, desto gröber ist die Segmentierung.

3.4.4. Diskussion der Ergebnisse

Durch die zuvor durchgeführten Analysen konnte gezeigt werden, dass das Bayessche hierarchische Modell sowohl auf ChIP-Seq-Daten, aber auch auf Proteinzeitreihen angewendet werden kann. Es muss lediglich beachtet werden, dass bei einer Anwendung auf Proteinzeitreihen die Backgroundcounts durch die minimale Lichtintensität und einen kleinen Zufallsfehler erzeugt werden muss.
Die Analysen haben hier für beide Arten von Daten gute Ergebnisse geliefert. Dabei musste jedoch stets eine Anzahl an Segmenten vorgegeben werden sowie eine Startsequenz der CPs. In den in dieser Arbeit durchgeführten Analysen wurde hier die Segmentanzahl verwendet, welche sich aus dem CBS Algorithmus ergab. Als Startsequenzen wurden bei der Analyse der ChIP-Seq-Daten zum einen die aus dem CBS Algorithmus resultierenden CPs genutzt sowie eine zufällige Startsequenz. Bei den Proteinzeitreihen wurde sich lediglich auf eine zufällige Startsequenz beschränkt, da diese bei den ChIP-Seq-Daten ebenso gute Ergebnisse lieferte wie die CBS-Startsequenz.
Um die Anzahl an Segmenten flexibler zu behandeln wurden drei Strategien verfolgt: die Variation der Startsegmentanzahl, eine Modifikation des Thresholds sowie die Einbettung eines Reversible Jump Schritts.
Im Fall der ChIP-Seq-Daten wurden, neben der „Ausgangs-Anzahl“ an Segmenten, noch weitere mögliche Anzahlen für K betrachtet. Für die Proteinzeitreihen hingegen wurde eine weitere mögliche Berechnung des Threshold betrachtet. Durch die Analysen konnte gezeigt werden, dass sowohl die Wahl einer anderen Anzahl K als auch der neue Threshold zu einer feineren Segmentierung führten (sofern im Falle des neuen Threshold das verwendete K nicht bereits sehr hoch war). Durch die feinere Segmentierung wurde die Anpassung an die Daten genauer und fehlende Segmente konnten ausgeglichen werden.

Weiter wurde das Bayessche hierarchische Modell um einen Reversible Jump Schritt erweitert, sodass die Anzahl der Segmente K innerhalb der Iterationen variieren kann. Dies macht den Algorithmus nochmals flexibler. Dabei kann durch die Gewichtung der Reversible Jump Schritte (Split und Merge) vom Anwender die Flexibilität des Bayesschen hierarchischen Modells gesteuert werden. Somit kann die Variabilität der Anzahl an Segmenten K vom Anwender eingeschränkt werden.
Bei der Anwendung des erweiterten Bayesschen hierarchischen Modells auf die bereits zuvor analysierten Datensätze zeigte sich, dass der Reversible Jump Schritt gute Ergebnisse lieferte. Es musste jedoch beachtet werden, dass nun die Definition eines Thresholds zur Findung der finiten Segmentierung zwar wie zuvor durchgeführt werden kann, jedoch nicht immer sinnvoll ist. Hier hat sich gezeigt, dass ein über Quantile definierter Threshold sinnvoll ist und die daraus resultierende Segmentierung die Daten gut anpasst.
Dabei zeigte sich, dass der Median als Threshold die feinste und beste Segmentierung hervorbringt. Jedoch fiel auf, dass am Anfang und Ende der analysierten Daten (Anfang und Ende von Chromosom 3 bzw. Start und Ende der Proteinmessung) die Segmentierung feiner war als in der Mitte der Daten. Durch einen entsprechend klein gewählten Threshold (i.Allg. dem Median) konnte man jedoch stets einem zu langen Segment in der Mitte der Proteinzeitreihe entgehen.

4. Entwicklung eines Analyseschemas für räumliche Daten mit Clusterstruktur

Wie im vorherigen Kapitel sollen auch in diesem Kapitel Proteindaten analysiert werden. Dieses Mal jedoch in einem räumlichen Kontext. Die experimentellen Daten, welche erneut an Ras Proteinen erhoben wurden, sowie eine Simulationsstudie, welche sowohl für den Single Colour als auch für den Dual Colour Fall verwendet wurde, werden im folgenden Unterkapitel näher erläutert. Weiter erfolgt die Herleitung des Zieles der räumlichen Analyse sowie die Methodenbeschreibung. Anschließend werden die Daten mit den zuvor beschriebenen Methoden analysiert und die Ergebnisse zusammenfassend diskutiert.

4.1. Daten- und Problembeschreibung

Im Folgenden werden nun eine Simulationsstudie sowie die experimentellen Single Colour Daten vorgestellt. Anschließend wird das hier spezifische Problem genauer erläutert.

4.1.1. Simulationsstudie

In dieser Arbeit werden sowohl Single Colour, als auch Dual Colour Daten simuliert. Dabei bedeutet Single Colour, dass lediglich ein Protein untersucht und daher nach der Fluoreszenzmikroskopie nur eine Farbe auf dem Pixelbild zu sehen ist. Bei Dual Colour Daten werden hingegen zwei Proteine simultan betrachtet. Dafür werden die zwei Proteine durch unterschiedliche Farben gekennzeichnet, meistens rot und grün, und werden anschließend mit Hilfe von Fluoreszenzmikroskopie gemessen. Das dadurch entstandene Bild enthält somit sowohl rote als auch grüne Punkte (von den jeweiligen Proteinen), daher der Name Dual Colour. Dieses Dual Colour Bild kann jedoch nach der Messung in seine zwei Farben aufgespalten werden, sodass wieder zwei Einzelbilder entstehen.
In dieser Arbeit wird eine Simulation aus Schäfer et al. (2014, [44]) genutzt und für den Dual Colour Fall weiter angepasst. Die Simulation basiert auf einer Verallgemeinerung eines Matérn Cluster Prozesses (Matérn, 1986, [30]) und kann wie folgt beschrieben werden:

Zunächst werden Elternpunkte mit Hilfe eines Poisson-Prozess im zweidimensionalen Raum gezogen. In einem Matérn Cluster Prozess würden nun alle Elternpunkte durch ein Cluster mit Radius r ersetzt werden. In der hier genutzten Verallgemeinerung werden jedoch nicht alle Elternpunkte, sondern lediglich ein vorher festgelegter Prozentsatz durch Cluster ersetzt. Durch diesen Schritt wird garantiert, dass es Punkte im Hintergrund gibt, welche nicht zu einem Cluster gehören, die sogenannten Monomere oder auch Singletons. Somit wird mit dieser Simulation ein zweidimensionales Bild erzeugt, in dem einige Punkte in Clustern und einige Punkte Singletons sind.
Für eine Single Colour Simulation wird nun lediglich eines dieser Bilder erzeugt (mit Hilfe des verallgemeinerten Matérn Cluster Prozesses). Dabei werden die Parametereinstellungen aus Tabelle 4.1 verwendet. Ein Beispiel für eine solche Simulation ist in Abbildung

Parameter		Werte
Proportion der Punkte in Clustern (insgesamt)	p	0.4, 0.8
durchschn. Clustergröße	μ	4,8
durchschn. Clusterradius [nm]	r	15,30
Punktedichte (insgesamt)	λ	125
Detektionsfehler [nm]	σ	20

Tabelle 4.1: Übersicht der verwendeten Simulationsparameter, sowohl für die Single als auch die Dual Colour Simulation.

4.1 zu sehen.
Für eine Dual Colour Simulation müssen hingegen zwei Bilder simuliert werden. Das erste simulierte Bild repräsentiert dabei die „grünen“ Proteinpunkte, das zweite die „roten“ Proteinpunkte. Für die Dual Colour Simulation werden dabei hier folgende drei Szenarien S 1, S 2 und S 3 unterschieden.

S 1 Die zwei Proteine verhalten sich unabhängig voneinander, d.h. die zwei Datensätze können separat erzeugt werden.

S 2 Die Clusterzentren der zwei Proteine sind korreliert, d.h. die Clusterzentren des ersten Proteins werden für das zweite Protein nur leicht verschoben in eine beliebige Richtung und anschließend die Tochterpunkte erzeugt. Die Monomere werden dabei

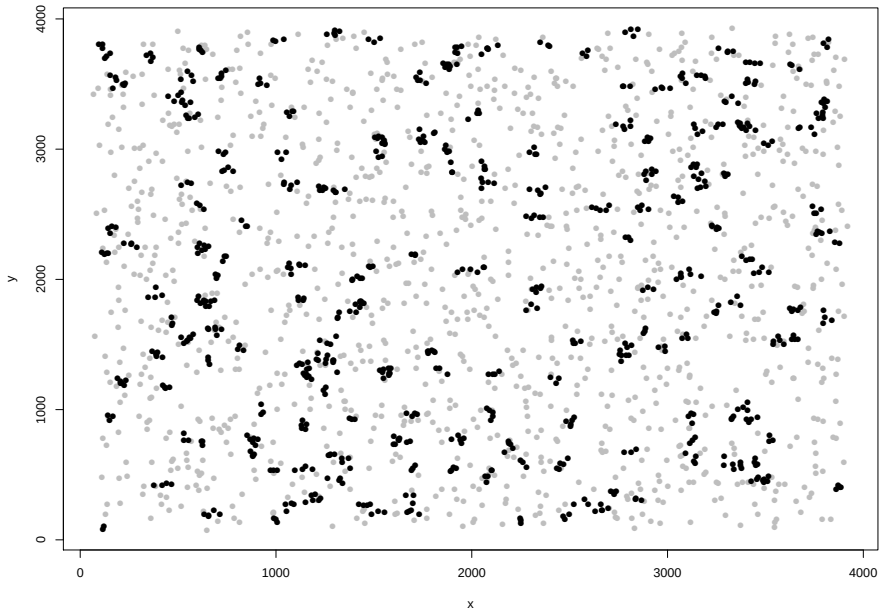

Abbildung 4.1: Simuliertes Datenbeispiel mit insgesamt 40% an Punkten in Clustern, einer durchschnittlichen Clustergröße von 4 Punkten, einem durchschnittlichen Clusterradius von 15 nm, einer totalen Punktedichte von 125 Punkten/μm^2 und einem Detektionsfehler von 20 nm; dabei repräsentieren die schwarzen Punkte, die Punkte in Clustern.

separat erneut erzeugt.

S 3 Die Proteine in den Clustern sind korreliert, d.h. die Punkte in den Clustern des ersten Proteins werden leicht versetzt und als Cluster von Protein zwei genutzt. Die Monomere werden auch hier wieder separat erzeugt.

Die so erzeugten Daten können nun zu einem Dual Colour Bild zusammengefügt werden. Da bei experimentellen Daten die beiden Proteine jedoch in ihre Einzelbilder getrennt würden, werden auch hier die beiden Einzelbilder separat analysiert.

4.1.2. Experimentelle Daten

Zur Erhebung der experimentellen Daten wurden Ras Proteinen in der basalen Plasmamembran zunächst mit mEos2 fusioniert (Ras-mEos2). Anschließend wurden sie mit

Hilfe der TIRF-Mikroskopie (vgl. Kapitel 3.1.2) gemessen. Die Datenerhebung erfolgte dabei wie zuvor im zeitlichen Kontext am Max-Planck-Institut für molekulare Physiologie Dortmund in der Arbeitsgruppe um Dr. Peter J. Verveer. Die genauen Lokationen der Ras-mEos2 Proteine wurden dabei mit Hilfe von *photo activated localisation microscopy (PALM)* berechnet. PALM ist dabei eine Bildmethode, welche wie folgt beschrieben werden kann:
Die Fluoreszenz mEos2 ist in ihrem ursprünglichen Stadium grün, kann jedoch durch UV Licht zu einer roten Form umgewandelt werden. Wenn nun ein UV-Laser mit entsprechender Wellenlänge verwendet wird, so wandelt sich eine kleine Teilmenge der mEos2 Luminophore in rotes Licht um. Diese Teilmenge kann anschließend mit Hilfe der TIRF-Mikroskopie gemessen werden, wobei individuelle Spots sichtbar werden. Diese Spots können nun bis auf 20 nm genau bestimmt werden, indem ein zweidimensionales Gaußprofil für jeden Spot des Bildes angepasst wird. Die andauernde Messung der Zelle kann bei den Proteinen nun zum Bleachen (vgl. 3.1.2) der roten Form des mEos2 führen. Dadurch sind die Spots nicht mehr sichtbar. In diesem Fall wird eine neue Teilmenge der Luminophore angeregt, sodass auch sie sich in die rote Form des mEos2 umwandeln und auf dieselbe Art lokalisiert werden können. Dieses Vorgehen wird so lange wiederholt, bis alle Proteine gemessen wurden.
Durch die Kombination der TIRF-Mikroskopie und PALM kann zum einen die Messung auf die Zellmembran konzentriert werden und zum anderen eine hohe Genauigkeit (20 nm) gewährleistet werden. Die gemessenen Daten enthalten abschließend die Koordinaten der einzelnen Proteine.

4.1.3. Ziel der räumlichen Analyse

Säugetierzellen werden, wie bereits in Kapitel 2.1 erwähnt, durch ihre Zellmembran von ihrer Umgebung abgegrenzt. Die Zelle kommuniziert dabei mit ihrer Umgebung mit Hilfe von Proteinen, welche genau in oder an dieser Zellmembran sitzen. Daher ist naheliegend zu vermuten, dass die räumliche Verteilung dieser Proteine eine große Rolle in der Signalübertragung der Zelle spielen. Somit ist das Ziel, die räumliche Clusterstruktur von Proteinen in der Zellmembran zu untersuchen. Dadurch könnten Rückschlüsse auf das

Verhalten von Proteinen gezogen werden, z.B. ob sich Proteine in bestimmten Situationen „suchen“ und zusammenfinden. Dies kann sowohl in verschiedenen Stadien der Zelle (z.B. in den verschiedenen Phasen der Mitose) oder unter unterschiedlichen Einflüssen (z.B. der Anwesenheit oder dem Fehlen eines bestimmten Hormons), aber auch zwischen zwei verschiedenen Proteinen von Interesse sein. Durch die Möglichkeit auch lebende Zellen mit Hilfe von Fluoreszenzmikroskopie zu beobachten sowie der Dual Colour Messung (d.h. es werden zwei Proteine gleichzeitig betrachtet), können Daten für beide Ansätze erhoben werden.
Bei der Analyse dieser Daten können nun vielfältige Charakteristika Aufschluss über die Clusterstruktur bringen, z.B. die räumliche Lage, aber auch die Anzahl an Proteinen, welche sich in Clustern befinden oder auch die Größe der Cluster. In dieser Arbeit wird dabei der Fokus auf die Schätzung des Anteils an Proteinen in Clustern gelegt.
Um solche Charakteristika schätzen zu können, werden in dieser Analyse verschiedene statistische Methoden miteinander verglichen, darunter sowohl räumliche als auch modellbasierte Methoden. Diese Methoden werden dabei genauer im folgenden Unterkapitel beschrieben. Weiter soll ein Schema entworfen werden, wie die Methoden effizient kombiniert werden können.

4.2. Räumliche Cluster- sowie grafische Methoden

Im folgenden Unterkapitel werden nun die statistischen Methoden erläutert, mit welchen anschließend die zuvor beschriebene Simulation sowie die Daten analysiert werden sollen.

4.2.1. Hierarchisches Clustern

Das Ziel einer Clusteranalyse besteht darin, eine Menge an Objekten $o_1, \ldots, o_N$ in $h < N$ homogene Gruppen, den sogenannten Clustern $\{G_1, \ldots, G_h\}$, einzuteilen. Es existieren verschiedene Strategien und Methoden, um dieses Ziel zu erreichen, z.B. das hierarchische Clustern (z.B. Johnson, 1967, [23]), k-Means (Kanungo et al., 2002, [24]) oder auch Partitioning Around Medoids (kurz: PAM; Kaufman und Rousseeuw, 2008, [25]). Bei den letzten zwei Verfahren muss die Anzahl der Cluster, im Gegensatz zum hierarchischen

Clustern, vorher festgelegt werden. Im Folgenden wird sich hier auf das hierarchische Clustern beschränkt.

Beim hierarchischen Clustern werden die Objekte iterativ über ihre Ähnlichkeit zu Clustern zusammengefügt. Die Ähnlichkeit von zwei Objekten wird dabei (meist) über die Distanz zwischen den zwei Objekten gemessen, d.h. ist die Distanz klein, so scheinen sich die Objekte zu ähneln, liegen sie hingegen weit entfernt, so ist dies nicht der Fall. Als Distanzmaß kann aus einem großen Pool an Distanzmaßen gewählt werden, z.B. die euklidische Distanz oder die Manhattan-/Cityblock-Distanz. Um nun die Objekte iterativ zu Clustern zu vereinigen, gibt es beim hierarchischen Clustern zwei Verfahren:

Bei einem agglomerativen Vorgehen, wird mit der feinsten Partition begonnen. Dies bedeutet, dass zunächst jedes Objekt ein eigenes Cluster bildet. Anschließend werden iterativ immer die zwei Cluster vereinigt, welche sich am nähsten bzw. ähnlichsten sind. Dies wird solange fortgeführt, bis alle Objekte ein großes Cluster bilden. Ein divisives Verfahren startet hingegen mit der gröbsten Partition, d.h. alle Objekte bilden ein Cluster. Anschließend werden iterativ das (bzw. in weiteren Schritten die) Cluster in sich unähnliche Gruppen gespalten. Dies wird solange fortgeführt, bis das Verfahren am Ende bei der feinsten Partition angelangt ist.

Um die Distanz zwischen zwei Clustern während der Iterationen bestimmen zu können, existieren drei Verfahren: das Single-, das Complete- und das Average-Linkage (vgl. auch Hartigan, 1975, [17]). Für diese drei Verfahren kann die Distanz zwischen zwei Clustern G_i und G_j, $i \neq j$ und einem beliebigen Distanzmaß $dist$ wie folgt beschrieben werden:

$$\begin{aligned}
\text{Single-Linkage:} \quad & dist_S(G_i, G_j) = \min_{\phi \in G_i, \varphi \in G_j} dist(o_\phi, o_\varphi), \\
\text{Complete-Linkage:} \quad & dist_C(G_i, G_j) = \max_{\phi \in G_i, \varphi \in G_j} dist(o_\phi, o_\varphi), \\
\text{Average-Linkage:} \quad & dist_A(G_i, G_j) = \frac{1}{N_i, N_J} \sum_{\phi \in G_i} \sum_{\varphi \in G_j} dist(o_\phi, o_\varphi).
\end{aligned}$$

Dabei ist N_i die Anzahl an Objekten in Cluster G_i und N_j die Anzahl an Objekten in Cluster G_j.

Die iterative Zusammenführung (bzw. Aufspaltung) der Cluster wird oft grafisch in einem sogenannten Dendogramm dargestellt. Dabei werden auf der x-Achse die Objekte angeordnet. Durch senkrechte und vertikale Linien wird dann verdeutlicht, in welcher Reihenfolge die Objekte zu Clustern zugeordnet wurden. In Abbildung 4.2 b) ist beispielhaft ein Den-

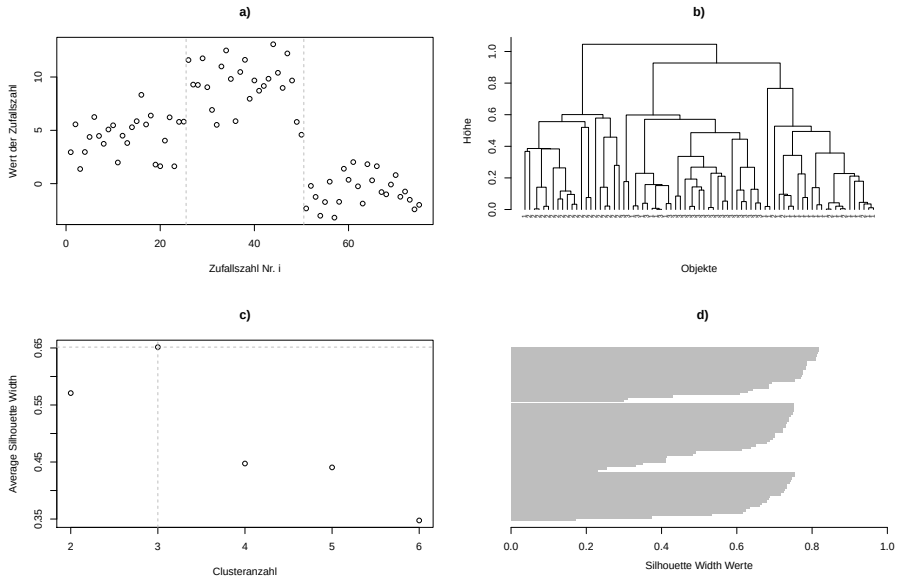

Abbildung 4.2: Ergebnis des hierarchischen Clustern für ein Beispiel von je 25 Stichprobenwerten aus $\mathcal{N}(\mu = 5, \sigma^2 = 3)$, $\mathcal{N}(\mu = 10, \sigma^2 = 5)$ und $\mathcal{N}(\mu = 0, \sigma^2 = 2)$; dabei sind in a) die Zufallszahlen abgetragen, in b) das Dendogramm, in c) die durchschnittlichen Silhouette Width Werte sowie in d) ein Silhouette Plot für eine Clusteranzahl von 3.

dogramm für eine Ziehung von jeweils 25 Stichprobenwerten aus 3 Normalverteilungen ($\mathcal{N}(\mu = 5, \sigma^2 = 3)$, $\mathcal{N}(\mu = 10, \sigma^2 = 5)$ und $\mathcal{N}(\mu = 0, \sigma^2 = 2)$) abgebildet (vgl. Abbildung 4.2 a)). Um die optimale Anzahl an Clustern in diesem Fall zu finden, kann die Silhouette Width für verschiedene Anzahlen angewendet werden. Für eine feste Anzahl an Clustern repräsentiert diese für jedes Objekt die Güte seiner Anpassung in dem zugeordneten Cluster. Der Silhouette Width Wert s_i für eine Beobachtung o_i ist definiert als

$$s_i = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}, \tag{7}$$

wobei $a(i)$ die durchschnittliche Distanz zwischen o_i und allen anderen Objekten im gleichen Cluster ist sowie $b(i)$ das Minimum aller Distanzen von o_i und allen anderen Clustern. Somit nimmt s_i Werte zwischen -1 und 1 an. Ist die Silhouette Width nahe 1, so ist das Objekt sehr gut durch sein Cluster repräsentiert. Für Objekte mit $s_i < 0$ gilt hingegen, dass sie nicht gut durch das ihnen zugeteilte Cluster repräsentiert werden. Dies kann auch in einem Silhouette Plot wie in Abbildung 4.2 d) veranschaulicht werden. Um die optima-

le Clusteranzahl zu bestimmen, wird die Average Silhouette Width (für eine bestimmte Anzahl an Clustern), also der Mittelwert aller Silhouette Width Werte, zu Rate gezogen. Anschließend wird die Anzahl an Clustern gewählt, welche die maximale Average Silhouette Width erreicht. Dies ist für das oben genannte Beispiel in Abbildung 4.2 c) veranschaulicht worden.

4.2.2. Average Shifted Histogram

Das Average Shifted Historgram (ASH) kann als eine grafische Methode zur Dichteschätzung von räumlichen Daten aufgefasst werden. Eine zentrale Annahme des ASH ist, dass die Daten in $k^\star$ Intervalle, den sogenannten Bins, eingeteilt sind. Das k-te Bin ist in diesem Zusammenhang definiert als $B_k = [\tau_k, \tau_{k+1})$ mit ξ_k der Anzahl an Beobachtungen, die in B_k fallen, $k = 0, \ldots, k^\star - 1$. Die Dichteschätzung ist gegeben durch $\hat{f}_k(x) = \frac{\xi_k}{N(\tau_{k+1} - \tau_k)}$, wobei N die Gesamtanzahl an Beobachtungen ist.
Häufig wird weiter angenommen, dass $\omega = \tau_{k+1} - \tau_k, \forall k$, und $\tau_0 = 0$, d.h. alle Intervalle die gleiche Breite besitzen. Durch diese Annahme vereinfacht sich die geschätzte Dichte zu $\hat{f}_k(x) = \frac{\xi_k}{N\omega}$. In diesem Fall ist der einzige unbekannte Parameter ω, weshalb diese Methode oftmals als „nicht-parametrisch“ bezeichnet wird (Scott und Sain, 2005, [46]).
Ist hingegen τ_0 unbekannt, so existieren zwei unbekannte Parameter: ω und τ_0. Dabei kann τ_0 als eine Art Störparameter aufgefasst werden, sodass sich das Problem des Störparameters durch m verschobene (shifted) Histogramme umgangen werden kann. Diese Histogramme sollen jeweils um einen Faktor $\delta = \frac{\omega}{m}$ vom vorherigen Histogramm verschoben sein.
Für äquidistante Intervalle $B_k^\star = [\tau_0 + k\delta, \tau_0 + (k+1)\delta)$ mit Breite δ und den zugehörigen Anzahlen an Beobachtungen ξ_k, ist das *Average Shifted Histogram (ASH)* somit konstant über alle Intervalle und die Dichteschätzung ergibt sich zu

$$\hat{f}_{ASH} = \frac{1}{N\omega} \sum_{i=1-m}^{m-1} \left(1 - \frac{|i|}{m}\right) \xi_{k+i}, \quad \text{für } x \in B_k^\star. \tag{8}$$

In obiger Formel wurden gleiche Gewichte für alle verschobenen Histogramme verwendet. Diese können auch durch Intervall-spezifische Gewichte ersetzt werden (z.B. einen Kern $ker(x)$), sodass sich obige Formel verallgemeinern lässt zu: $\hat{f}_{ASH} = \frac{1}{N\omega} \sum_{i=1-m}^{m-1} w_m(i)\xi_{k+i}$,

für $x \in B_k^\star$. Da für $m \to \infty$ ASH den Kern-Dichte-Schätzer approximiert, gilt weiter

$$\hat{f}_{ker} = \frac{1}{N\omega} \sum_{i=1}^{N} ker\left(\frac{x - x_i}{\omega}\right) = \frac{1}{N} \sum_{i=1}^{N} ker_\omega(x - x_i), \tag{9}$$

wobei $ker_\omega(x) = \frac{1}{\omega} ker(\frac{x}{\omega})$.
Für weitere Informationen zum ASH oder der Erweiterung für vektorwärtige Daten sei an dieser Stelle auf Scott und Sain (2005, [46]) verwiesen.

4.2.3. Extensible Markov Modelle

Das *Extensible Markov Modell (EMM)* ist eine Markov Kette, welche sich über die Zeit hinweg verändern kann. Dabei spiegelt die Markov Kette die Clusterstruktur wieder und kann so als gerichteter Graph (mit fester Struktur) interpretiert werden, wobei die Cluster die Knoten sind. Dabei kann mit jedem Zeitpunkt, zu dem neue Daten erhoben werden, die Clusterstruktur und somit der gerichtete Graph angepasst werden. Das Vorgehen des EMM kann wie folgt beschrieben werden:
Das EMM besteht aus einer Markov Kette mit K_t Knoten zu jedem Zeitpunkt t, $t = 1, \ldots, T$, und einem Algorithmus, welcher zwischen drei alternativen Schritten wählen kann: EMMCluster, EMMIncrement und EMMDecrement.
EMMCluster teilt ein Objekt o_{it}, $i = 1, \ldots, N_t$, einem Cluster G_i, $i = 1, \ldots, K_t$ an einem gegeben Zeitpunkt t zu, $t = 1, \ldots, T$, wobei die Cluster durch die Knoten repräsentiert werden. Ein Objekt wird einem bestehenden Cluster zugeordnet, falls es „nah genug“ an diesem liegt. Ist dies nicht der Fall, so wird ein neues Cluster geschaffen, welches nur dieses eine Objekt enthält. Die Entscheidung, ob ein Objekt „nah genug“ zu einem Cluster liegt, wird mit Hilfe eines vorher vom Anwender festgelegten Parameter, dem Threshold ζ, bestimmt. Dieser Schritt kann somit als eine Art „Nächste-Nachbar-Algorithmus“ angesehen werden.
EMMIncrement berechnet hingegen die Übergangswahrscheinlichkeiten der Markov Kette zwischen den Knoten. Dies erfolgt durch das Speichern der Anzahl an Objekten, welche als „Gast“ an dem entsprechenden Knoten waren. Diese Anzahlen werden dann als Indikator genutzt um die entsprechenden Übergangswahrscheinlichkeiten zu berechnen.
Der letzte mögliche Schritt ist EMMDecrement, welcher die Größe der Markov Kette

und somit die des Graphen reduzieren kann. Durch diesen Schritt kann einer zu großen Anzahl an Knoten, d.h. an Clustern, vorgebeugt werden und somit gleichzeitig einer zu feinen Partition der Daten.

Algorithmus 1 : Veranschaulichung des Vorgehen des EMM in algorithmischer Schreibweise.

Data : Objekte $o_{it}, i = 1, \ldots, \quad N_t, t = 1, \ldots, T$;
Threshold ζ;

for *alle Zeitpunkte* $t = 1, \ldots, T$ **do**

Ordne Objekte o_{it}, $i = 1 \ldots, N_t$ ihrem nächsten Knoten/Cluster zu;
Dafür nutze die drei Schritte EMMCluster, EMMIncrement und EMMDecrement;
EMMCluster: Ordne Objekt o_{it}, $i = 1, \ldots, N_t$, Cluster G_i, $i = 1, \ldots, K_t$ zu, sofern bei beliebigem Distanzmaß gilt: $dist(o_i t, G_i) < \zeta$; Ist diese Ungleichung für kein Cluster G_i erfüllt, soll o_{it} ein neues Cluster und damit einen neuen Knoten darstellen;
EMMIncrement: Berechne nach der Zuordnung des Objekts die Übergangswahrscheinlichkeiten für die Markov-Kette neu;
EMMDecrement: Falls der Graph und damit die Markov Kette zu groß wird, so wird mit EMMDecrement ein Knoten, d.h. ein Cluster, aus dem Grafen entfernt;

end

Result : Markov-Kette mit K_T Knoten und entsprechender Einteilung der Daten in die Cluster;

Das Vorgehen des EMM ist auch in Algorithmus 1 dargestellt. Weiter lernt das EMM durch seinen Aufbau unaufhörlich während der Anwendung und ist dadurch „*generic incremented model whose nodes can have any kind of representative*“ (Dunham, Meng und Huang, 2004, [6]). Daher sind Online-Daten ein gutes Anwendungsgebiet für das EMM.

Für nähere Details sei an dieser Stelle auf Dunham, Meng und Huang (2004, [6]) verwiesen.

4.2.4. DBSCAN

Der DBSCAN (Density-Based Spatial Clustering of Applications with Noise) Algorithmus ist eine weit verbreitete räumliche Clustermethode, welche homogene Gruppen innerhalb eines Datensatzes $\{x_1, \ldots, x_N\}$ sucht. Ob zwei Datenpunkte ähnlich sind und demselben Cluster zugeteilt werden sollen, wird dabei mit Hilfe von Distanzen zwischen den entsprechenden Punkten definiert. In diesem Zusammenhang wird von Ester et al. (1996, [9]) die Theorie der „density-reachability“ eingeführt und kann mit Hilfe der zwei Parameter ϵ und $MinPts$ wie folgt definiert werden:
Ein Punkt x_i ist **direkt density-reachable** für x_j, falls die Distanz der zwei Punkte höchstens ϵ beträgt, d.h. $dist(x_i, x_j) \leq \epsilon$, sowie mindestens $MinPts$ Punkte in der ϵ-Nachbarschaft von x_j vorhanden sind, $i, j \in \{1, \ldots, N\}, i \neq j$. Somit wird nicht nur die Distanz als Indikator für die Clusterzugehörigkeit genutzt, sondern auch, ob das Cluster genug Punkte enthalten würde. Sind zwei Punkte x_i und x_j nicht direkt density-reachable, existiert jedoch eine Reihe an Punkten $x_i = x_1^\star, \ldots, x_m^\star = x_j$ in der alle aufeinanderfolgenden Punkte direkt density-reachable sind, so sind x_i und x_j **density-reachable**.
Neben den oben genannten Beziehungen, führen Ester et al. einen weiteren Begriff in ihrem Artikel ein, welcher die symmetrische Beziehung zweier Punkte beschreibt, die **density-connectivity.** Dabei sind zwei Punkte x_i und x_j density-connected, falls ein Punkt x_k existiert, der sowohl density-reachable von x_i als auch von x_j ist, $i, j, k \in \{1, \ldots, N\}, i \neq j \neq k$.
Die Cluster werden im DBSCAN Algorithmus mit Hilfe der zuvor beschriebenen Eigenschaften gebildet. Demnach besteht ein Cluster G aus einer nichtleeren Teilmenge des Datensatzes, wenn für diese Teilmenge gilt, dass

1. für alle Punkte x_i, x_j gilt, dass $x_i \in G$ und x_j ist density-reachable von x_i (bzgl. der Parameter ϵ und $MinPts$) und
2. für alle Punkte x_i, x_j gilt, dass x_i density-connected zu x_j ist, $i, j, k \in \{1, \ldots, N\}$, $i \neq j \neq k$.

Um die Clusterzuordnung vorzunehmen, müssen jedoch zunächst die Parameter ϵ und $MinPts$ definiert bzw. gewählt werden. Sind diese Parameter gegeben (oder im Ideal-

fall bekannt), so kann der DBSCAN Algorithmus mit einem zufälligen Punkt $x_\star$ starten und für diesen density-reachable Punkte suchen (bzgl. der zwei Parameter). Ist $x_\star$ ein Datenpunkt aus dem Inneren eines Clusters, so findet der DBSCAN Algorithmus dieses Cluster bzgl. der Parametereinstellungen von ϵ und $MinPts$. Ist $x_\star$ hingegen ein Punkt am Rand des Clusters, so sind dort keine density-reachable Punkte für $x_\star$ und der Algorithmus wählt einen anderen Punkt um fortzufahren. Für diesen wird erneut nach denisty-reachable Punkten gesucht. Diese Prozedur wird iterativ weitergeführt bis alle Cluster bestimmt wurden.

4.2.5. Die Gammics Methode

Die von Schäfer et al. (2015, [44]) entwickelte Gammics Methode ist eine weitere räumliche Clustermethode. Sie sucht mit Hilfe eines Bayesschen hierarchischen Modell Cluster in Punktprozessen, wobei ein Cluster mindestens zwei Datenpunkte enthalten muss. Ein Vorteil dieser Methode ist, dass sie drei Clustercharakteristika simultan schätzt: den durchschnittlichen Anteil an Proteinen in Clustern, die durchschnittliche Größe, d.h. wie viele Objekte durchschnittlich in einem Cluster enthalten sind, sowie den durchschnittlichen Radius der Cluster.
Um dies zu erreichen, wird die Distanz zwischen einem Punkt und seinem κ-ten Nachbarn modelliert und nicht der Punktprozess selber (im Folgenden $\kappa = 2$). Der Clusterradius und die Größe werden nach der Modellierung in einem zweiten Schritt algorithmisch geschätzt.
Seien nun die Daten Realisationen von Zufallsvariablen $X_i, i = 1, \ldots, N$, wobei diese Zufallsvariablen für zufällige Punktkoordinaten in einer Region $P \subset \mathbb{R}^2$ stehen. Weiter sei mit D_i die Distanzen zwischen Punkt X_i und seinem nähsten Nachbarn X_j mit Realisation $d_i = dist(x_i, x_j)$ beschrieben. Die Gammics Methode modelliert nun eine Funktion der Distanzen D_i^2, da über die Distanzen entschieden werden kann, ob ein Punkt X_i einem Cluster angehört. Die Funktion ist dabei gegeben durch:

$$Y_i(D_i^2) = \begin{cases} 1, & X_i \text{ ist Teil eines Clusters} \\ 0, & X_i \text{ ist kein Teil eines Clusters} \end{cases} . \tag{10}$$

Für D_i^2 werden nun zwei Gamma-Verteilungen, mit Dichte $f(x) = \frac{1}{\Gamma(a_\star)\beta_\star^{\alpha_\star}} x^{\alpha_\star - 1} e^{-\frac{\alpha_\star}{\beta_\star}}$, angepasst. Eine dieser zwei Gamma-Verteilungen repräsentiert dabei die Punkte in Clustern, die andere Verteilung die Punkte außerhalb der Cluster. Somit gilt

$$D_i^2 | Y_i = k \sim \text{Gamma}(\alpha_k, \beta_k), \quad k = 0, 1, \tag{11}$$

$$\alpha_k | Y_i = k \sim \text{Gamma}(a_k, b_k), \quad k = 0, 1 \quad \text{und}$$

$$\frac{1}{\beta_k} | Y_i = k \sim \text{Gamma}(c_k, d_k), \quad k = 0, 1. \tag{12}$$

Die implizite Verteilung der Punkte zu einer der Gamma-Verteilungen wird über die Mittelwerte von $Y = (Y_1, \ldots, Y_N)'$ ermittelt. Dafür wird angenommen, dass Y_i Bernoulli verteilt ist, $i = 1, \ldots, N$. Dadurch ergibt sich

$$Y_i \sim \text{Bernoulli}(p_c), \tag{13}$$

$$p_c \sim \text{Beta}(\alpha, \beta). \tag{14}$$

Das Mischungsmodell der Gammics Methode ergibt sich somit aus den Formeln (10) - (14). Die Hyperparameter $a_0, a_1, b_0, b_1, c_0, c_1, d_0, d_1, \alpha$ und β müssen vorher vom Anwender definiert werden. Im Weiteren seien $a_0 = 3, a_1 = 2, b_0 = 1, b_1 = 1, c_0 = 1, c_1 = 4, d_0 = 0.5, d_1 = 1$, und $\alpha = \beta = 1$ (sofern nicht anders definiert).

Die Schätzung des durchschnittlichen Clusterradius sowie der Clustergröße hängt nun sowohl von den Verteilungen, als auch von der Zuordnung geclustert bzw. nicht geclustert ab. Diese Clusterchrackteristika können nun über die zwei Gamma-Verteilungen algorithmisch ermittelt werden. Dafür wird zunächst die Überschneidung der zwei Gamma-Verteilungen betrachtet. Diese kann wie folgt formuliert werden:

$$L_c = \{x | p_c \cdot f(x | \alpha_{y,1}, \beta_{y,1}) = (1 - p_c) \cdot f(x | \alpha_{y,0}, \beta_{y,0})\}, \tag{15}$$

wobei f erneut die Dichte der Gamma-Verteilung ist. In einem zweiten Schritt kann nun mit Hilfe von L_c die Distanz D_q eines Punktes X_q zu seinem nähsten Nachbarn über $q = \text{argmin}_{i: D_i^2 \geq L_c} D_i^2$ bestimmt werden. Mit Hilfe dieser Distanzen können nun der Clusterradius sowie die Clustergröße ermittelt werden.

Das vollständige Modell kann abschließend durch einen Gibbs Sampling Markov Chain Monte Carlo (MCMC) Ansatz, inklusive einem Metropolis Schritt zur Aktualisierung der Shape Parameter der Gamma-Verteilungen implementiert und berechnet werden (MATLAB (7.10.0, 2010, [32])).

4.3. Analyse und Vergleich der Methoden

In diesem Kapitel werden nun die Methoden aus Kapitel 4.2 angewendet und evaluiert. Dafür wird zunächst ihr Verhalten auf einer kleinen Region einer simulierten Single Colour Messung untersucht. Anschließend wird die entsprechende Schätzung des Anteils an Proteinen in Clustern für verschiedene Simulationsbeispiele betrachtet, sodass aus dieser Untersuchung ein Vorgehensschema abgeleitet werden kann. Abschließend erfolgt zum einen eine Analyse experimenteller Single Colour Daten und zum anderen simulierter Dual Colour Daten. Die Analysen werden mit Hilfe der Programmiersoftware R (2015, [38]) und den zugehörigen Paketen `cluster`, `ash`, `rEMM`, `spatstat` und `fpc` sowie der Programmiersoftware MATLAB (7.10.0, 2010, [32]) durchgeführt.

4.3.1. Untersuchung der Methoden auf einer ersten Single Colour Simulation

Im Folgenden soll nun zunächst auf einer einzelnen simulierten *Region Of Interest (ROI)* einer Single Colour Simulation das Verhalten der vorgestellten Methoden untersucht werden. Dafür wurde ein Datensatz mit einer totalen Proportion an Punkten in Clustern von 40%, einer mittleren Clustergröße von 4 Proteinen, einem mittleren Clusterradius von 15 nm, einer totalen Punktdichte von 125 points/μm^2 und einem Detektionsfehler von 20 nm erzeugt. Die simulierte Single Colour ROI war bereits in Abbildung 4.1 in Kapitel 4.1.1 dargestellt. Diese simulierte ROI enthält 195 Cluster und insgesamt 1976 Proteine. Somit ergibt sich eine realisierte Proportion von ca. 39.2% Proteinen in Clustern.
Zunächst wird das hierarchische Clustern mit Average Linkage zur Analyse der simulierten ROI herangezogen. Um in diesem Fall die optimale Anzahl an Cluster bestimmen zu können, wird die Average Silhouette Width genutzt. In Abbildung 4.3 ist zum einen das sich aus dem hierarchischen Clustern ergebende Dendogramm (links) und zum anderen die Kurve der Average Silhouette Width Werte (rechts) abgetragen. Es ist deutlich zu erkennen, dass die optimale Anzahl an Cluster, gemessen an der Average Silhouette Width, 451 beträgt. Wenn man darauf basierend das Dendogramm „abschneiden“ würde, dass genau 451 Cluster entstehen, so erhält man lediglich 34 Singletons, d.h. Cluster mit nur einem Protein und somit einen geschätzten Anteil an Proteinen in Clustern von 98.3%. Die Proportion an Proteinen in Clustern wird folglich deutlich überschätzt. Nimmt man

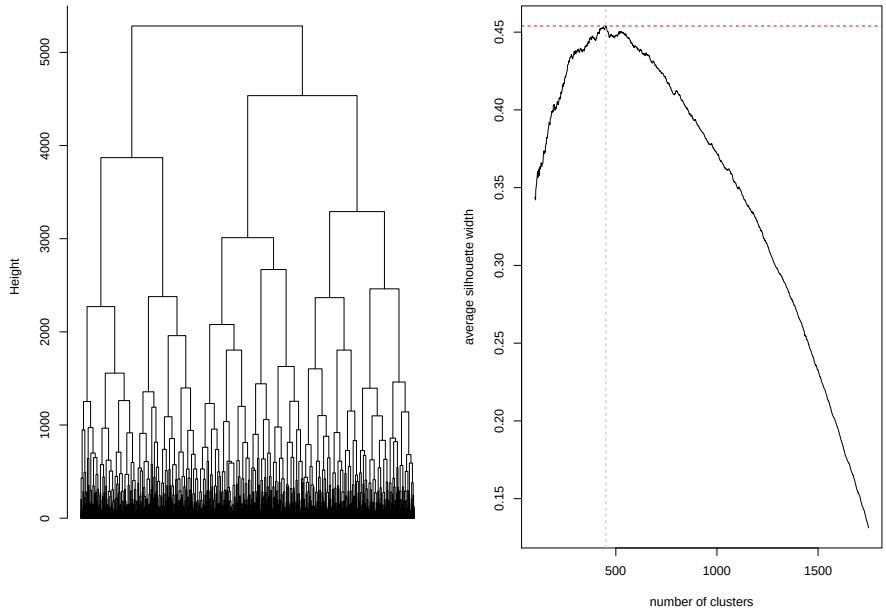

Abbildung 4.3: Dendogramm des hierarch. Clusterns und Kurve der Average Silhouette Width Werte für eine erste simulierte Single Colour ROI; links: Dendogramm des hierarchischen Clustern mit Average Linkage; rechts: Kurve der Average Silhouette Width Werte.

nun weiter an, dass alle Cluster mit maximal 5 Objekten zum Background gehören, so erhält man 331 Singletons und eine geschätzte Proportion von 83.2%. Auch dieses Mal wird der Anteil deutlich zu hoch geschätzt. Um dieses Problem zu lösen, wurden daher a priori Informationen über die Anzahl an Cluster mit Hilfe des Average Shifted Histograms (ASH) gewonnen. Die sich aus der Anwendung des ASH ergebende Grafik ist in Abbildung 4.4 zu sehen. Auf Grund dieses Contour Plots kann eine grobe Schätzung der Anzahl an Cluster abgeben werden, welche hier ungefähr bei 150 Clustern liegt. Wenn nun weiter angenommen wird (durch Expertenwissen oder eine vorherige Analyse), dass in jedem Cluster durchschnittlich 5 Proteine enthalten sind, so ergibt sich eine geschätzte Anzahl an Background-Proteinen von 1976 - $(150 \cdot 5) = 1226$. Mit dieser a priori Information ergibt sich der geschätzte Anteil an Proteinen in Clustern zu ca. 39.7%. Die aus dieser Anzahl an Clustern resultierende Zuordnung (mittels hierarchischem Clustern mit Average Linkage) ist in Abbildung 4.5 abgetragen. Somit funktioniert die Schätzung mit Hilfe von ASH und a priori Informationen deutlich besser. Es sei jedoch zu beachten,

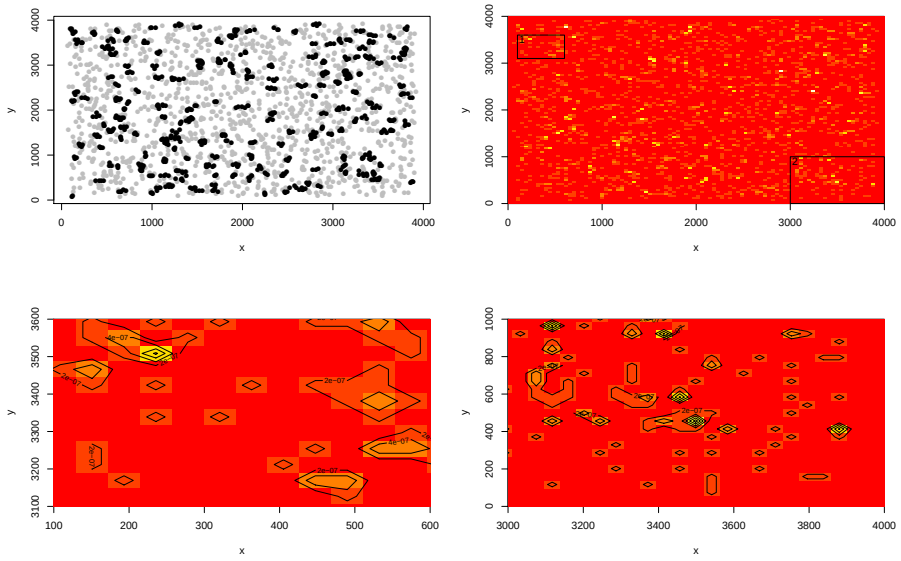

Abbildung 4.4: oben links: simulierte ROI (vgl. auch Abb. 4.1); oben rechts: Contour Plot als Ergebnis des Average Shifted Histogram der simulierten Daten; unten links: Contour Plot des Average Shifted Histogram für Part 1; unten rechts: Contour Plot des Average Shifted Histogram für Part 2.

dass nun das hierarchische Clustern ausschließlich für die Zuordnung der Proteine zu den Clustern bzw. dem Status „Singleton“ genutzt wird.

Für das EMM muss zunächst ein Threshold ζ gewählt werden, wobei hier zwei verschiedene Werte verwendet werden: 25 und 55. Die geschätzte Proportion an Proteinen in Clustern kann nun über die Singletons berechnet werden, welche durch alle Cluster mit nur einem Objekt repräsentiert werden. Der geschätzte Anteil ergibt sich dann aus $1 - \frac{\#\text{Singletons}}{\#\text{aller Proteine}}$. Für $\zeta = 25$ ergibt sich eine geschätzte Proportion von geclusterten Proteinen von ca. 35.2%, für einen Threshold $\zeta = 55$ ergibt sich eine geschätzte Proportion von ca. 70.9%. Hier wird deutlich, dass in diesem Fall der Threshold von 25 besser passt. Die resultierende Clusterzuordnung der Proteine ist in Abbildung 4.6 zu sehen. An dieser Abbildung ist gut zu erkennen, dass die Zuordnung der Proteine in die Gruppen „Clusterprotein“ bzw. „Singleton“ schwieriger ist. Die Schätzung des Anteils an Proteinen in Clustern funktioniert hingegen, bei geeigneter Wahl des Threshold, gut. Für diese simulierte ROI war ein Threshold von 25 eine gute Wahl, was ungefähr dem Durchmesser entspricht. Man kann daher vermuten, dass man mit a priori Informationen über den

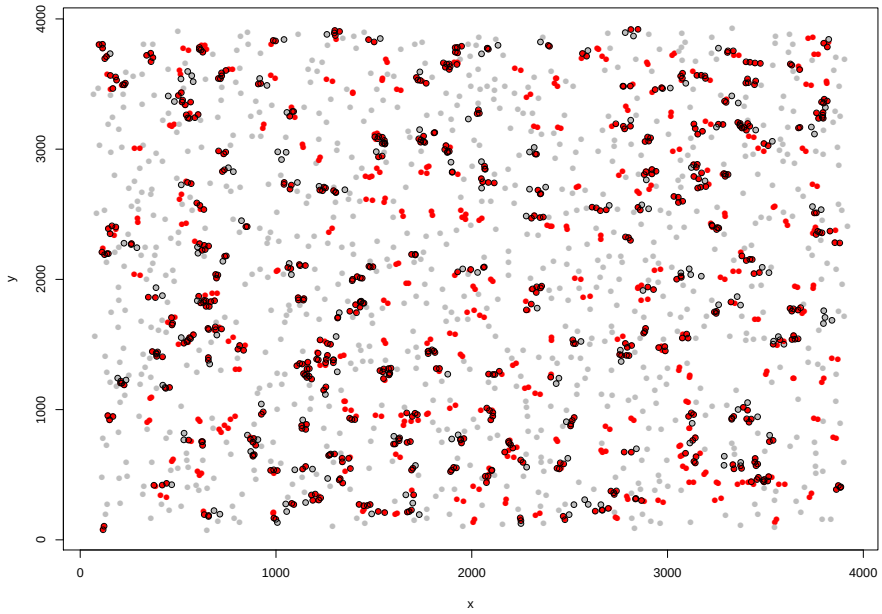

Abbildung 4.5: Resultierende Clusterzuordnung unter Verwendung von ASH und hierarchischem Clustern (die roten Punkte entsprechen hier den geclusterten Punkten, die grauen den Singletons und die Punkte mit einem schwarzen Rand den Punkten, welche tatsächlich als geclustert simuliert wurden).

Radius auch eine gute Wahl des Thresholds für das EMM treffen kann und somit eine gute Schätzung für den Anteil an Proteinen erhält.

Um diese a priori Informationen zu erhalten, könnte Ripley's K-Funktion bzw. die entsprechende Transformation genutzt werden. Da sie aber für diese Anwendung den Radius deutlich überschätzt, wird an dieser Stelle nicht weiter auf die K-Funktion eingegangen.

Eine weitere Clustermethode, die bei räumlichen Daten Anwendung findet, ist der DBSCAN Algorithmus. Auch hier muss ebenfalls ein Parameter geschickt gewählt werden, hier ϵ (vgl. auch 4.2.4; der Parameter $MinPts$ wird im weiteren als Default-Einstellung $MinPts = 5$ gewählt und nicht weiter betrachtet). Für die Analyse der simulierten ROI werden hier vier Werte für ϵ angenommen: 25, 50, 75 und 100. Die aus diesen Parametereinstellungen resultierenden Clusterzuordnungen sind in Abbildung 4.7 dargestellt. Auf Basis von Abbildung 4.7 und Tabelle 4.2, welche die geschätzten Proportionen an Proteinen in Clustern enthält, wird deutlich, dass $\epsilon = 50$ die beste Wahl für diese simulierte ROI ist. Der DBSCAN Algorithmus mit $\epsilon = 25$ hingegen unterschätzt die Proportion deutlich.

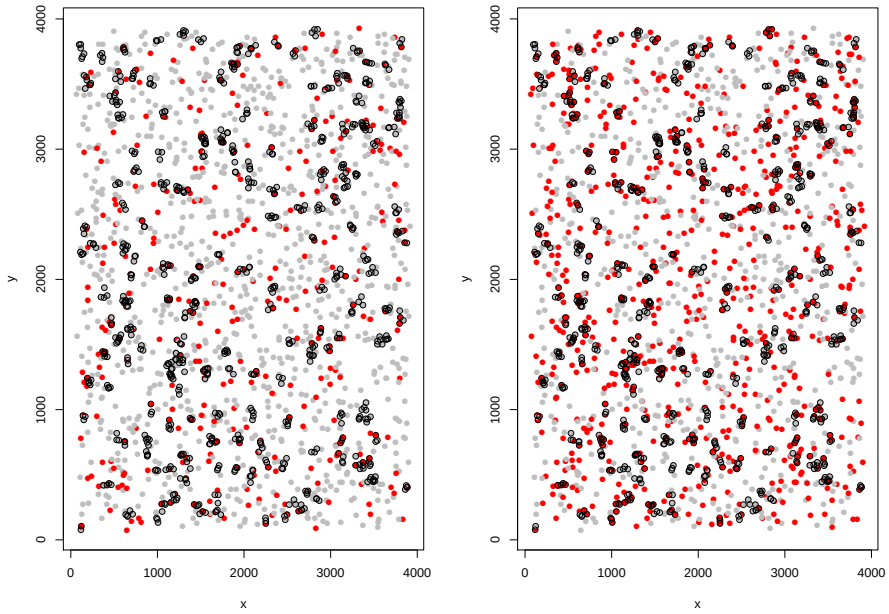

Abbildung 4.6: Clusterergebnis des EMM mit einem Threshold von 25 (links) und einem Threshold von 55 (rechts): die roten Punkte entsprechen hier den geclusterten Punkten, die grauen den Singletons und die Punkte mit einem schwarzen Rand den Punkten, welche tatsächlich als geclustert simuliert wurden.

Wahl von ϵ	geschätzter Anteil von Proteine in Clustern
25	2.6%
50	30.6%
75	60.9%
100	83.8%

Tabelle 4.2: Geschätzte Proportion von Proteinen in Clustern mit dem DBSCAN Algorithmus in Abhängigkeit von der Wahl für ϵ.

Abschließend soll nun noch die Gammics Methode auf die simulierte ROI angewendet werden. Die Gammics Methode schätzt bei der Anwendung gleich 3 Charakteristika: den Anteil an Proteinen in Clustern, den mittleren Radius der Cluster sowie die mittlere Größe

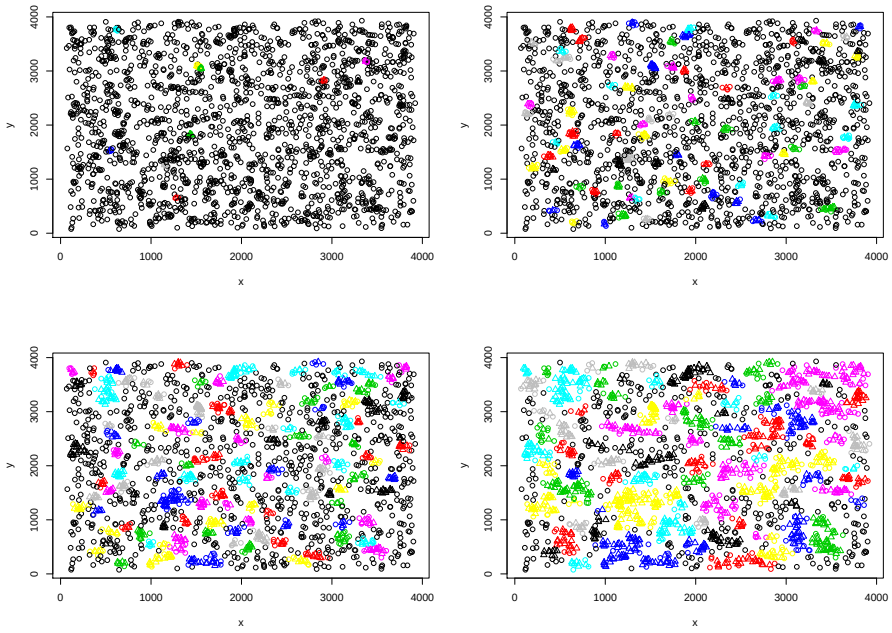

Abbildung 4.7: Ergebnis der Clusterzuweisung nach der Analyse mittels DBSCAN Algorithmus mit folgenden Parametereinstellungen: oben links: $\epsilon = 25$, oben rechts: $\epsilon = 50$, unten links: $\epsilon = 75$, unten rechts: $\epsilon = 100$; die schwarzen Punkte entsprechen dabei den Datenpunkte, welche dem Background zugeordnet wurden, d.h. den Singletons; die farbig markierten Punkte entsprechen hingegen den jeweiligen Punkten in einem Cluster.

der Cluster. Für dieses Beispiel wird eine geschätzte Proportion von 41.5%, ein mittlerer Radius von 9.9 nm und eine mittlere Größe von 3.7 Proteinen geschätzt. Diese Werte sind sehr nahe an den Vorgaben der Simulation und die Gammics Methode liefert somit gute Ergebnisse.

4.3.2. Schätzung der Proportion an Proteinen in Clustern

Im Weiteren werden nun alle Methoden (mit entsprechenden Parametern), welche im vorherigen Kapitel gute Ergebnisse lieferten, auf weitere Simulationseinstellungen im Single Colour Fall angewendet. Für die simulierten Szenarien wurden dabei lediglich drei Parameter variiert (der Anteil an Proteinen in Clustern p, die Clustergröße μ und der Clusterradius r) und jeweils ein Bild einer ROI erzeugt. Für jede Parametereinstellung

erfolgt dies vier mal. Die geschätzten Anteile für Proteine in Clustern für die entsprechenden Methoden mit den zugehörigen Parametern sind für jede einzelne Simulation in Tabelle 2 im Anhang B abgetragen. Die für jede Parametereinstellung der Simulation gemittelten Schätzungen sind hingegen in Tabelle 4.3 abgetragen. Die weiteren Spalten

Simulationsparameter	$\hat{p}_{prior}$	$\hat{p}_{ash}$	$\hat{p}_{EMM25}$	$\hat{p}_{EMM55}$	$\hat{p}_{db50}$	$\hat{p}_{db75}$	$\hat{p}_{db100}$	$\hat{p}_{gammics}$
p=0.4, μ=4, r=15	0.37	0.35	0.37	0.74	0.34	0.64	0.86	0.41
p=0.4, μ=4, r=30	0.37	0.32	0.37	0.73	0.32	0.64	0.85	0.47
p=0.4, μ=8, r=15	0.37	0.23	0.43	0.73	0.47	0.62	0.81	0.40
p=0.4, μ=8, r=30	0.37	0.21	0.42	0.72	0.47	0.63	0.82	0.42
p=0.8, μ=4, r=15	0.73	0.27	0.53	0.85	0.53	0.79	0.89	0.79
p=0.8, μ=4, r=30	0.72	0.31	0.48	0.84	0.50	0.77	0.91	0.80
p=0.8, μ=8, r=15	0.71	0.29	0.67	0.86	0.83	0.88	0.92	0.81
p=0.8, μ=8, r=30	0.72	0.34	0.62	0.84	0.78	0.87	0.92	0.80

Tabelle 4.3: Durchschnittliche geschätzte Proportion in Abhängigkeit der Simulationsparameter sowie der verwendeten Methoden mit zugehörigen Parametern; $\hat{p}_{prior}$: der Schätzer der Proportion an Proteinen in Clustern, bei Annahme von 150 Clustern für $p = 0.4$ und 300 für $p = 0.8$; $\hat{p}_{ash}$: der Schätzer für den Anteil an Proteinen in Clustern bei Verwendung des hierarchischen Clustern und ASH zur Bestimmung der Anzahl an Clustern (spezifisch für jedes Simulationsszenario); $\hat{p}_{EMM25}$ bzw. $\hat{p}_{EMM55}$: die geschätzte Proportion bei Verwendung des EMM mit Threshold 25 bzw. 55; $\hat{p}_{db50}$, $\hat{p}_{db75}$ und $\hat{p}_{db100}$: der geschätzte Anteil an Proteinen in Clustern unter Verwendung des DBSCAN Algorithmus mit Parameter $\epsilon = 50$, 75 und 100 und $\hat{p}_{gammics}$: die geschätzte Proportion unter Verwendung der Gammics Methode.

enthalten die Schätzungen der Proportion für die jeweiligen Methoden mit den zugehörigen Parametern in den verschiedenen Szenarien: $\hat{p}_{prior}$: der Schätzer der Proportion an Proteinen in Clustern, bei Annahme von 150 Clustern für $p = 0.4$ bzw. 300 für $p = 0.8$; $\hat{p}_{ash}$: der Schätzer für den Anteil an Proteinen in Clustern bei Verwendung des hierarchischen Clustern und ASH zur Bestimmung der Anzahl an Clustern; $\hat{p}_{EMM25}$ bzw. $\hat{p}_{EMM55}$: die geschätzte Proportion bei Verwendung des EMM mit Threshold 25 bzw. 55; $\hat{p}_{db50}$,

$\hat{p}_{db75}$ und $\hat{p}_{db100}$: der geschätzte Anteil an Proteinen in Clustern unter Verwendung des DBSCAN Algorithmus mit Parameter $\epsilon = 50$, 75 und 100 und $\hat{p}_{gammics}$: die geschätzte Proportion unter Verwendung der Gammics Methode.

Die Schätzung mit a priori Informationen $\hat{p}_{prior}$ ist nah an der wahren Proportion, wenn man als a priori Information 150 bzw. 300 Cluster (je nach Proportion p) annimmt. Weiter ist auch die Schätzung $\hat{p}_{ash}$, in der die Anzahl an Clustern mit Hilfe von ASH ermittelt wird (vgl. Tabelle 3 in Anhang B), für eine Proportion von ca. 40% gut, wohingegen $\hat{p}_{ash}$ für $p = 0.8$ leicht unterschätzt. Ein möglicher Grund ist in Abbildung 4.8 zu sehen. Es ist deutlich zu erkennen, dass für $p = 0.8$ (rechts) weniger „helle Punkte“ zu erkennen sind und somit weniger Cluster geschätzt würden als für $p = 0.4$ (links). Dies kann zu einer

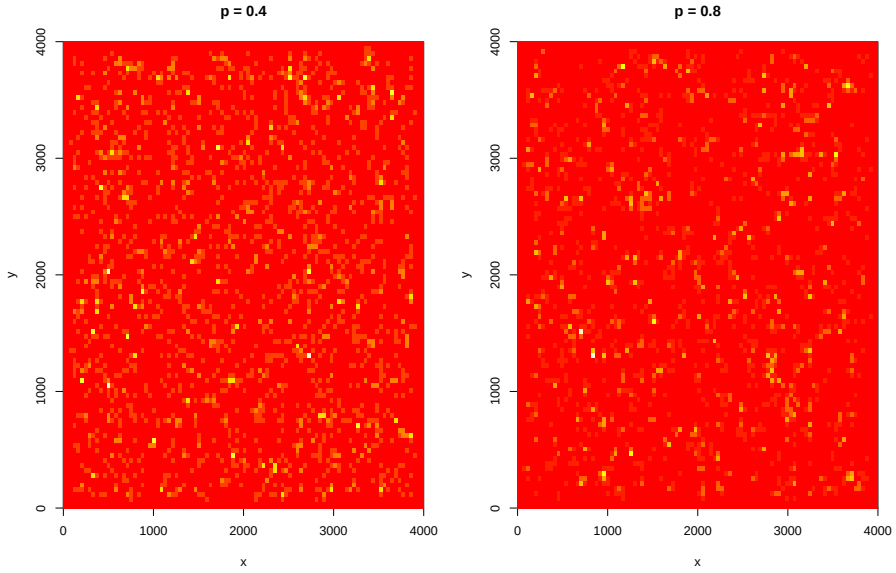

Abbildung 4.8: Ergebnis der Analyse mittels ASH zweier simulierter ROIs, wobei links p=0.4 und rechts p=0.8 simuliert wurde.

leichten Unterschätzung für eine hohe Proportion an Punkten in Clustern führen.

Im Falle des EMM sind die geschätzten Werte, für einen passenden Threshold, nah an den wahren Proportionen. Dabei liefert, für eine Proportion von $p = 0.4$, das EMM mit einem Threshold von 25 gute Ergebnisse, für eine Proportion von $p = 0.8$ hingegen ein Threshold von 55.

Der DBSCAN Algorithmus liefert ebenfalls gute Schätzungen ist aber, wie das EMM, von der Wahl des Parameters abhängig. Hier ist zu erkennen, dass $\epsilon = 50$ für Settings mit

$p = 0.4$ eine gute Parameterwahl ist, mit Ausnahme des Settings p=0.4, μ=4, r=30. Für simulierte Szenarien mit $p = 0.8$ ist hingegen $\epsilon = 75$ eine gute Wahl (sowie $\epsilon = 50$ für das Settings mit $p = 0.8$ and $\mu = 8$). Weiter ist auffällig, dass für den DBSCAN Algorithmus ein Unterschied in den Schätzungen in Abhängigkeit von μ vorliegt, wenn $\epsilon = 50$ gewählt wurde.
Die Gammics Methode liefert ohne nötige Parameterwahl gute Schätzungen für die Proportion geclusterter Proteine. Weiter sind auch der geschätzte mittlere Radius sowie die geschätzte mittlere Größe der Cluster nah an den Simulationsparametern (vgl. Tabelle 4 in Anhang B).

4.3.3. Diskussion erster Ergebnisse

Durch die Analyse der simulierten Daten wurde deutlich, dass ASH, das EMM und der DBSCAN Algorithmus eine gute Wahl zur Schätzung der Proportion der Proteine in Clustern sind, sofern a priori Informationen vorhanden sind oder aber der (Tuning-)Parameter der entsprechenden Methode gut gewählt wurde. Die Gammics Methode hingegen liefert eine gute Schätzung ohne eine nötige Parameterwahl. Tabelle 4.4 gibt somit einen Über-

Methode	Rechenzeit	Anforderungen an den Anwender	Anzahl der Schritte
hier. Clust.	niedrig	Anwender muss optimale Anzahl an Cluster definieren → z.B. über Sil. Width oder ASH	zwei Schritte: hier. Clustern + Sil. Width/ASH
ASH	niedrig	Anwender muss Grad der Glättung bestimmen	ein Schritt
EMM	niedrig - mittel	Anwender muss Threshold definieren (Threshold ≈ Durchmesser)	ein - zwei Schritte: ggf. Schätzung des Radius
DBSCAN	niedrig	Anwender muss ϵ wählen	ein Schritt
Gammics	hoch	keine Vorkenntnisse notwendig	ein Schritt

Tabelle 4.4: Eigenschaften sowie Vor- und Nachteile der verwendeten Methoden.

blick über die Methoden in Bezug auf die Rechenzeit, Anforderungen an den Anwender und die Anzahl an Schritten.

Es hat sich gezeigt, dass ASH eine nützliche Methode ist, um Vorinformationen zu erlangen, z.B. für eine Schätzung der Anzahl an Clustern oder aber um ROIs zu identifizieren. Das EMM, der DBSCAN Algorithmus und die Gammics Methode eignen sich hingegen gut zur Schätzung des Anteils an Proteinen in Clustern. Dabei sei zu beachten, dass sowohl für das EMM als auch für den DBSCAN Algorithmus ein Parameter gewählt werden muss, diese aber weniger rechenintensiv sind. Für die Gammics Methode sind keine weiteren Vorkenntnisse notwendig, auch wenn diese hilfreich sein könnten. Der Nachteil liegt jedoch in der Rechenzeit.

4.4. Herleitung und Anwendung eines Analyseschemas

In diesem Kapitel wird nun ein Analyseschema zur effizienten Nutzung der oben verwendeten Methoden hergeleitet. Anschließend wird dieses Analyseschema zum einen auf experimentelle Single Colour Daten, zum anderen auf simulierte Dual Colour Daten angewendet. Zur Analyse werden hier ebenfalls die zuvor genannte Software und ihre Pakete genutzt.

4.4.1. Analyseschema zur effizienten Kombination bekannter Methoden

Wie in Kapitel 4.3.3 beschrieben wurde, haben die verschiedenen Methoden jeweils Vor- und auch Nachteile. Um nun alle Vorteile der unterschiedlichen Methoden nutzen zu können und die Nachteile zu beheben, wurde ein Analyseschema zur effizienten Kombination der obigen Methoden entwickelt. Dieses ist in Abbildung 4.9 zu sehen. Demnach sollten zunächst mit Hilfe von ASH (kleine) ROIs gefunden - sofern sie nicht durch Expertenwissen bereits gewählt wurden - und anschließend mit der Gammics Methode analysiert werden um a priori Informationen zu erhalten. Diese können anschließend für weitere Methoden genutzt und somit größerer ROIs oder die ganze Zelle analysiert werden. Nach dieser Analyse besteht die Möglichkeit einer Feedback-Analyse, d.h. einer erneuten Analyse bestimmter ROIs mit der Gammics Methode. Dadurch können nochmals spezifischere Informationen ermittelt werden, z.B. durch eine bessere Wahl der ROIs. Abschließend

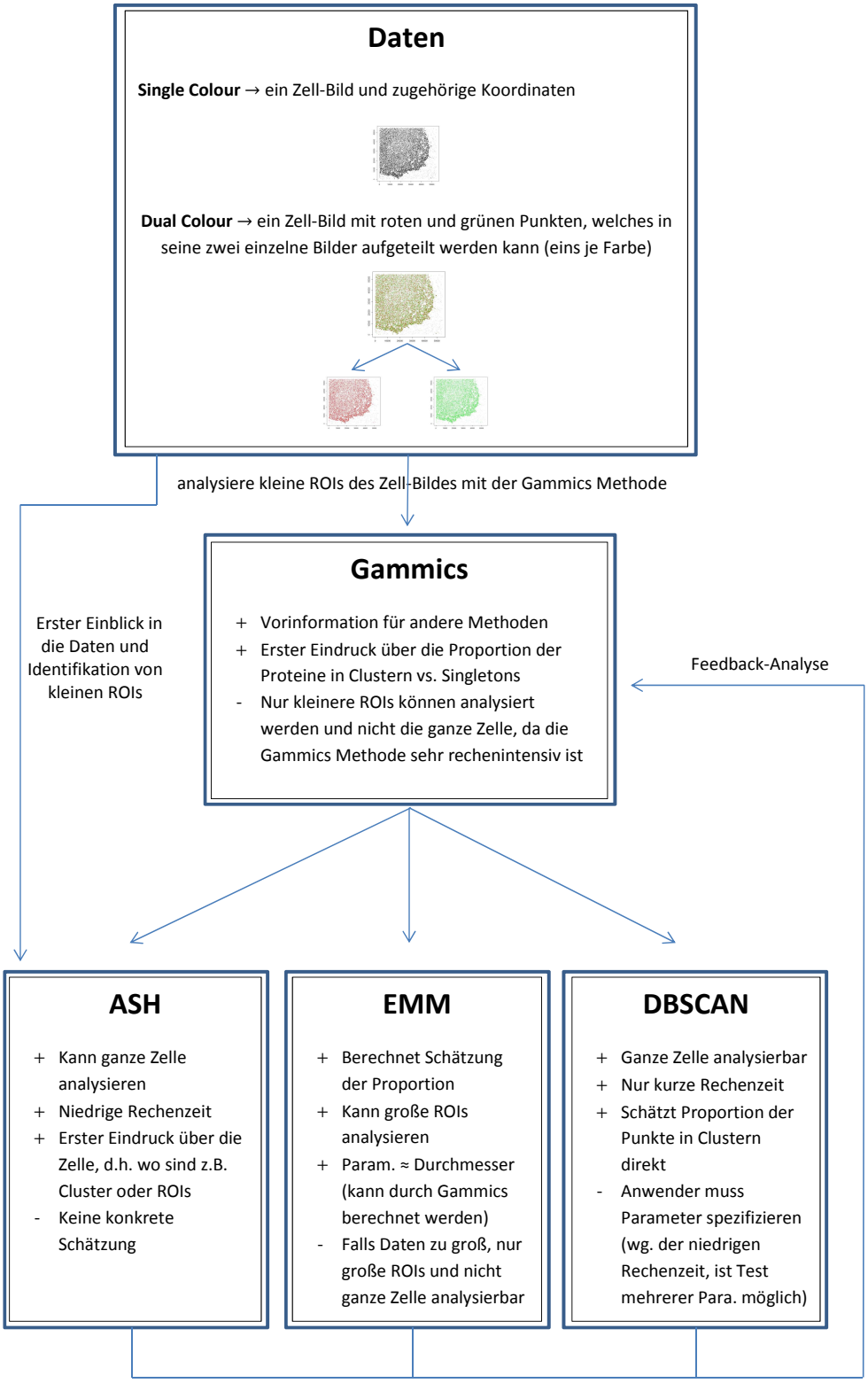

Abbildung 4.9: Schema zur effizienten Kombination der Analysemethoden inklusive einer Feedback Analyse für Single Colour als auch Dual Colour Daten.

kann mit den angepassten Vorinformationen erneut die ganze Zelle oder größere Teile analysiert werden.

4.4.2. Analyse experimenteller Single Colour Daten mit Hilfe des Analyseschemas

Zunächst werden nun experimentelle Single Colour Daten analysiert. Diese sind in Abbildung 4.10 a) zu sehen und wurden bereits in Kapitel 4.1.2 sowie in Schäfer et al. (2014, [44]) eingeführt. Um nun einen ersten Eindruck über die Verteilung der Cluster bzw.

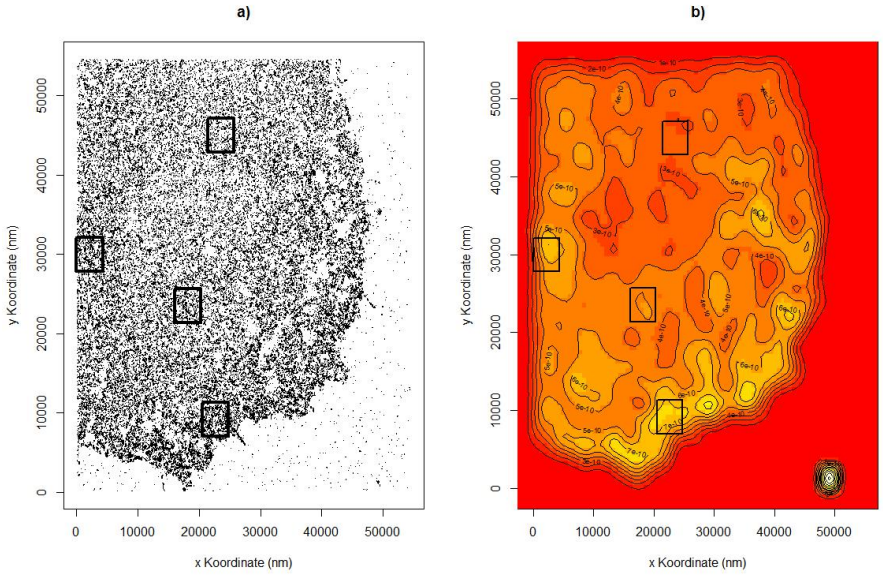

Abbildung 4.10: a) Experimentelle Single Colour Daten einer lebenden Zelle mit vier markierten ROIs, wobei die Ras-Lokalisierung mit Hilfe von PALM und einer Vorverarbeitung erfolgte; b) Ergebnis der Analyse der experimentellen Daten mit Hilfe von ASH.

der Proteine zu erhalten, wurde zunächst ASH angewendet. Der resultierende Plot ist in Abbildung 4.10 b) zu sehen. Dort wird deutlich, dass die ROIs gut gewählt sind und verschiedene Proteindichten abdecken. Anschließend wird auf die gekennzeichneten ROIs die Gammics Methode angewendet, um erste Charakteristika der Cluster zu erhalten. Die Ergebnisse sind in Tabelle 4.5 abgetragen, wobei die durchschnittlichen Schätzer durch einen Querstrich ($\bar{}$), die medianen Schätzer durch eine Tilde ($\tilde{}$) gekennzeichnet sind. Die Schätzungen des durchschnittlichen Radius und der durchschnittlichen Größe können in

ROI	$\hat{\bar{p}}$ ($\hat{\tilde{p}}$)	$\hat{\bar{\mu}}$ ($\hat{\tilde{\mu}}$)	$\hat{\bar{r}}$ ($\hat{\tilde{r}}$)
1	0.7528 (0.7573)	4.6297 (4.8022)	21.9653 (23.3396)
2	0.6631 (0.6621)	4.4658 (4.3965)	19.6638 (19.2250)
3	0.6583 (0.6580)	5.1815 (5.4357)	17.4426 (18.7699)
4	0.5944 (0.5923)	4.3179 (4.4116)	16.4657 (18.0482)
Mittelwert	0.6672 (0.6674)	4.6487 (4.7615)	18.8844 (19.8457)

Tabelle 4.5: Durchschnittliche (mit $\bar{}$ gekennzeichnet) und mediane (mit $\tilde{}$ gekennzeichneten) Schätzer der Analyse der experimentellen Single Colour Daten mit der Gammics Methode.

der weiteren Analyse für andere Methoden als a priori Informationen verwendet werden. Für das EMM wurde in Kapitel 4.3.1 bereits gezeigt, dass eine gute Wahl des Thresholds ungefähr dem Durchmesser entsprach. Wie in Tabelle 4.5 zu sehen ist, wurde der durchschnittliche Radius in den vier ROIs mit der Gammics Methode auf ca. 16 nm bis 22 nm geschätzt. Daher wird das EMM nun mit den Thresholds 30, 35 und 40 angewendet. Da die hier analysierte Zelle viele Proteine beinhaltet und somit die Datenmenge sehr hoch ist, sowie das EMM auf Markov Chain Monte Carlo Methoden basiert, muss an dieser Stelle die Zelle in vier Teile aufgeteilt werden, welche anschließend separat analysiert werden. Die Aufteilung kann Abbildung 9 in Anhang C entnommen werden. Die mit dem EMM geschätzten Proportionen in Abhängigkeit von der Wahl des Parameters sind in Tabelle 4.6 abgetragen. Man kann deutlich erkennen, dass die Schätzungen innerhalb

Teil der Zelle	$\hat{p_{30}}$	$\hat{p_{35}}$	$\hat{p_{40}}$
a	0.8001	0.8369	0.8636
b	0.8131	0.8481	0.8742
c	0.8232	0.8561	0.8808
d	0.8056	0.8417	0.8694

Tabelle 4.6: Geschätzte Proportionen des EMM in Abhängigkeit von der Wahl des Thresholds und dem Teil der Zelle.

eines Parameters ähnlich sind, jedoch alle zwischen 0.8 und 0.89, somit deutlich über den

geschätzten Werten der Gammics Methode für die vier ROIs, liegen.
Abschließend wurde noch der DBSCAN Algorithmus auf die experimentellen Single Colour Daten angewendet mit $\epsilon \in \{25, 50, 75, 100\}$. Dabei ergaben sich folgende Schätzungen:

$$\hat{p}_{25} = 0.5888, \quad \hat{p}_{50} = 0.8155, \quad \hat{p}_{75} = 0.8825 \quad \text{und} \quad \hat{p}_{100} = 0.9190.$$

Hier ist gut zu erkennen, dass die Schätzungen des DBSCAN mit $\epsilon \in \{50, 75\}$ ähnlich zu denen des EMM sind. Dies deckt sich mit den Ergebnissen aus der Simulation.

4.4.3. Anwendung des Analyseschemas auf eine Dual Colour Simulationsstudie

Ein weiteres Anwendungsgebiet ist, wie bereits erwähnt, der Dual Colour Fall. Dafür wurden in diesem Fall verschiedene Datensätze simuliert, welche jeweils ein „grünes“ so wie ein „rotes“ Protein enthalten. Die „roten“ Proteine werden, wie in Kapitel 4.1.1 beschrieben, in Abhängigkeit von dem „grünen“ Protein erzeugt. Anschließend können die simulierten Daten in je zwei Einzelbilder aufgeteilt werden. Diese Einzelbilder können nun analog zu den Single Colour Daten analysiert werden.
Um einen Eindruck über die verschiedenen Settings zu bekommen, wurde zunächst ASH angewendet. Einen beispielhaften Plot der verschiedenen Einzelbilder (je nach Setting) für eine Parametereinstellung der Simulation ist in Abbildung 4.11 zu finden. Es ist gut zu erkennen, dass zwischen den verschiedenen Szenarien Unterschiede in der Verteilung der Punkte vorhanden sind: in a) und b) sind Regionen mit einer hohen Punktdichte nicht korreliert, d.h. es existieren keine systematischen Ähnlikeiten zwischen Stellen mit einer hohen Punktedichte. Somit wird hier das Szenario S 1 (Unabhängigkeit) gut widergespiegelt. Für c) und d) hingegen sollten die Stellen mit einer hohen Punktdichte ähnlich zu denen aus a) sein, da eine Abhängigkeit in den Szenarien besteht. Dies ist in Abbildung 4.11 gut zu erkennen. Somit spiegeln die simulierten Daten auch die Szenarien S2 und S3 gut wieder.
Die Ergebnisse der Gammics Methode sind in Tabelle 4 in Anhang B abgetragen. Auch hier liefert die Gammics Methode, wie schon für die Single Colour Simulation, gute Ergebnisse, welche nahe an den Simulationsparametern liegen. Lediglich für Szenario S 2 wird der durchschnittliche Radius der roten Proteine für alle Simulationen unterschätzt. Auch für Szenario S 3 besteht die Tendenz den durchschnittlichen Radius im Falle von $r = 15$ zu

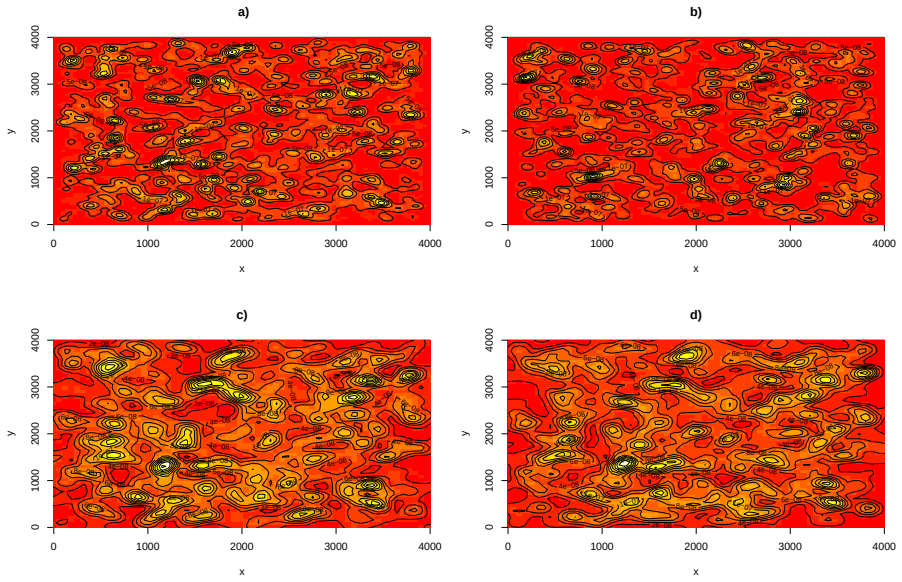

Abbildung 4.11: Ergebnis der Analyse einer Dual Colour Simulation mit Parametern $p = 0.4, \mu = 4, r = 15$ und den unterschiedlichen drei Settings mit Hilfe von ASH: a) grünes Protein, b) rotes Protein nach Setting 1, c) rotes Protein nach Setting 2, d) rotes Protein nach Setting 3.

über- bzw. von $r = 30$ zu unterschätzen. Dennoch kann auch hier die Gammics Methode

Simulationsparameter	$\hat{p}_{prior}$	$\hat{p}_{EMM25}$	$\hat{p}_{EMM55}$	$\hat{p}_{db50}$	$\hat{p}_{db75}$	$\hat{p}_{db100}$
p=0.4, μ=4, r=15	0.37	0.34	0.67	0.30	0.54	0.74
p=0.4, μ=4, r=30	0.37	0.31	0.65	0.26	0.54	0.73
p=0.4, μ=8, r=15	0.37	0.39	0.65	0.44	0.55	0.71
p=0.4, μ=8, r=30	0.37	0.38	0.67	0.43	0.57	0.72
p=0.8, μ=4, r=15	0.73	0.50	0.82	0.48	0.74	0.87
p=0.8, μ=4, r=30	0.72	0.45	0.80	0.44	0.74	0.87
p=0.8, μ=8, r=15	0.71	0.64	0.84	0.80	0.87	0.89
p=0.8, μ=8, r=30	0.72	0.59	0.83	0.75	0.85	0.90

Tabelle 4.7: Durchschnittlich geschätzte Proportion der Proteine in Clustern für die roten Proteine aus dem Dual Colour Datensatz für Szenario 3 (Korrelierte Proteine, welche in Clustern liegen).

zur Gewinnung von a priori Informationen genutzt werden, welche nun anschließend für

die Parameterwahl für beispielsweise das EMM genutzt werden kann.
Die durch die weiteren Methoden geschätzten Proportionen für Proteine in Clustern sind in den Tabellen 5 und 6 in Anhang B sowie 6 zu entnehmen. Durch die erste Anwendung von Gammics konnten auch hier die Parameter für das EMM und auch DBSCAN wie zuvor im Single Colour Fall eingegrenzt werden. Somit konnte Rechenzeit eingespart werden. Aus den Tabellen wird weiter deutlich, dass auch hier das EMM und der DBSCAN Algorithmus mit den entsprechend geeignet gewählten Parametern gute Ergebnisse liefert haben. So sind hier die geschätzten Proportionen nahe an den Simulationseinstellungen.
Zusammenfassend zeigte sich, dass auch im Dual Colour Fall die Vorgehensweise aus Abbildung 4.9 zu empfehlen ist (vgl. Kapitel 4.3.3).

4.5. Diskussion der Ergebnisse

Durch die erste Vergleichsanalyse auf einer simulierten Single Colour Simulation konnte gezeigt werden, dass die untersuchten Methoden sinnvolle Ergebnisse liefern. Es zeigten sich dabei für jede Methode Vor- und Nachteile. Diese konnten die Rechenzeit oder aber eine geschickte Parameterwahl betreffen.
Um diese Vorteile nutzen zu können, wurde ein Analyseschema zur effizienten Kombination der verwendeten Methoden aufgestellt (vgl. Abbildung 4.9). Mit Hilfe dieses Schemas wurden anschließend sowohl experimentelle Single Colour Daten als auch auf eine Dual Colour Simulation analysiert. Es zeigte sich, dass durch die geschickte Kombination der Methoden eine effizientere Analyse möglich wahr.
Durch die Hintereinanderschaltung der Methoden konnte der Nachteil der Parameterwahl umgangen werden. So konnte durch die erste Anwendung der Gammics Methode auf einer kleinen ROI a priori Informationen für die Parameterwahl des EMM gewonnen werden.
Weiter konnte auch gezeigt werden, dass das Schema sowohl für Single als auch für Dual Colour Daten anwendbar ist und somit in beiden Fällen die Analyse vereinfacht.

5. Analyse des Zusammenhangs räumlich-zeitlicher Proteindaten

In diesem Kapitel werden nun die zwei vorherigen Aspekte kombiniert und eine räumlich-zeitliche Analyse von Proteindaten durchgeführt. Die in diesem Kontext gemessenen Daten bestehen aus sogenannten Tracks von Proteinen, d.h. den Wegen der Proteine, welche sie innerhalb der Zelle zurückgelegt haben. Da hier, wie zuvor in der räumlichen Analyse, zwei unterschiedliche Proteine betrachtet werden, liegt erneut ein Dual Colour Problem vor.
Im Folgenden werden nun ein simuliertes Beispiel sowie die Daten genauer vorgestellt. Weiter erfolgt die Formulierung des Analyseziels. Anschließend wird ein Zusammenhangsmaß hergeleitet, welches zunächst auf das simulierte Beispiel und dann auf die experimentellen Daten angewendet wird.

5.1. Daten- und Problembeschreibung

In diesem Unterkapitel werden zum einen ein simuliertes Beispiel, zum anderen die experimentellen Daten vorgestellt. Anschließend wird das Ziel der räumlich-zeitlichen Analyse veranschaulicht.

5.1.1. Simuliertes Trackingbeispiel

In dieser Arbeit wird neben den experimentellen Trackingdaten auch ein simuliertes Beispiel zur Validierung herangezogen. Das Ziel ist (Protein-)Tracks zu simulieren, d.h. die Wege eines Objektes (z.B. eines Proteins, welches sich in bzw. an der Zellmembran bewegt). Dies erfolgt erneut im Dual Colour Fall, sodass von zwei Objekten jeweils Tracks simuliert werden. Dabei sollen möglichst viele interessante und wichtige Fälle in dem simulierten Trackingbeispiel enthalten sein.
Das hier verwendete simulierte Beispiel beinhaltet von zwei Objekten je acht Tracks, welche über zehn Frames, d.h. Zeitpunkte, hinweg verfolgt wird. In Abbildung 5.1 sind die 16 Tracks zu sehen, wobei sie entsprechend dem zugehörigen Objekt farblich gekennzeichnet

sind (z.B. rot = Protein 1, grün = Protein 2). Eine dreidimensionale räumliche Darstel-

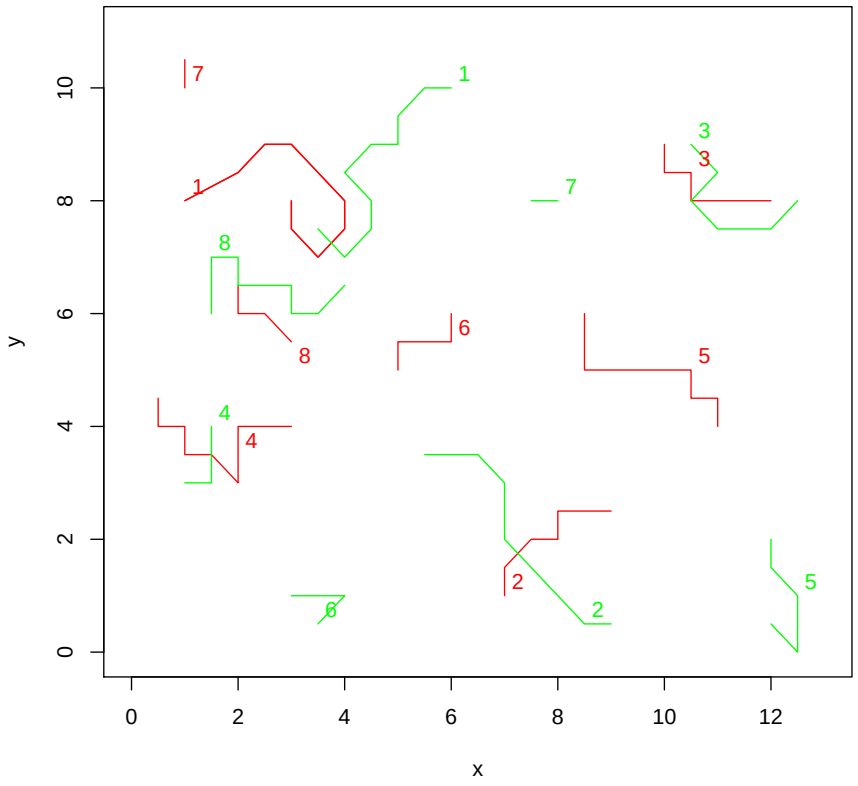

Abbildung 5.1: Grafische Darstellung der 16 simulierten Tracks (die farbliche Kennzeichnung entspricht dabei der Zugehörigkeit zu Objekt 1 (rot) bzw. 2 (grün), die Zahl an den entsprechenden Tracks dient einer möglichen Zuordnung).

lung ist weiter in Abbildung 10 in Anhang C zu finden. Es ist gut zu erkennen, dass die Simulation verschiedene Fälle bzgl. der Interaktion zwischen zwei unterschiedlichen Objekten beinhaltet, z.B.

- zwei Objekte, die sich aufeinander zu bewegen und anschließend nahe beieinander verlaufen (Track 1 grün und Track 1 rot),
- Objekte die nahe beieinander verlaufen (z.B. Track 3 grün und Track 3 rot),
- sich kreuzende Tracks (Track 2 grün und Track 2 rot bzw. Track 4 grün und Track 4 rot) oder

- ein kurzer Track, welcher in der (aktiven) Zeit nahe einem anderen Track verläuft (Track 8 grün und Track 8 rot).

Des Weiteren sind sowohl kürzere als auch längere sowie räumlich entfernte Trackpaarungen in dem simulierten Beispiel zu finden.
Da alle wichtigen und interessanten Fälle in der Simulation enthalten sind, kann mit ihrer Hilfe sowohl evaluiert als auch validiert werden.

5.1.2. Experimentelle Trackingdaten

Die hier zu analysierenden Protein-Trackingdaten wurden am Max-Planck-Institut für molekulare Physiologie Dortmund in der Arbeitsgruppe um Dr. Peter J. Verveer erhoben und vorverarbeitet. Im Folgenden wird der Prozess der Datenerhebung kurz erläutert. Eine genauere Beschreibung ist Ibach et al. (2015, [21]) zu entnehmen.
Die Trackingdaten wurden mit Hilfe von *Dual Colour Single Particle Tracking* an lebenden Zellen gemessen. Hier wurden die Tracks des Wachstum-Proteins EGFR sowie des Proteins PTB erhoben. Dabei wurden die Proteine zunächst mit Fluophoren versehen, wodurch die Fusionen Cy3-SNAP-EGFR und EGFP-PTB entstanden sind. Dabei ist Cy3 ein organisches Fluophor, welches rot exprimiert, wohingegen EGFP ein grün exprimierendes Fluoreszenz-Protein ist. Anschließend wurde die Zelle mit EGF stimuliert und mit Hilfe von *Dual Colour TIRF* Mikroskopie beobachtet. Durch die Dual Colour TIRF-Mikroskopie wurde dabei - wie in den Datenerhebungen zuvor - gewährleistet, dass lediglich Proteine in oder an der Zellmembran gemessen werden. Wie durch die Fluophore zu vermuten, wird das EGFR Protein auf dem Mikroskopiebild durch rote Spots, das PTB Protein durch grüne Spots repräsentiert.
Die Experimente wurden in diesem Fall für vier Zeitpunkte nach der Stimulation mit EGF (0, 2, 5 und 10 Minuten) gemessen. Dies erfolgte für jeweils 15 Zellen. Bei jedem Experiment wurden dann die jeweiligen Zellen über 150 Zeitpunkte hinweg gemessen. In dieser Arbeit wird lediglich eine Zelle aus den einzelnen Messungen 0, 2, 5 und 10 Minuten nach der Stimulation betrachtet.
Die Messung ergibt nun für eine Reihe an Zeitpunkten (hier 150) jeweils ein 2D Bild pro Zeitpunkt, die sogenannten Frames. Eine schematische Darstellung ist in Abbildung

5.2 zu sehen. Aus diesen zweidimensionalen Bildern werden anschließend die einzelnen

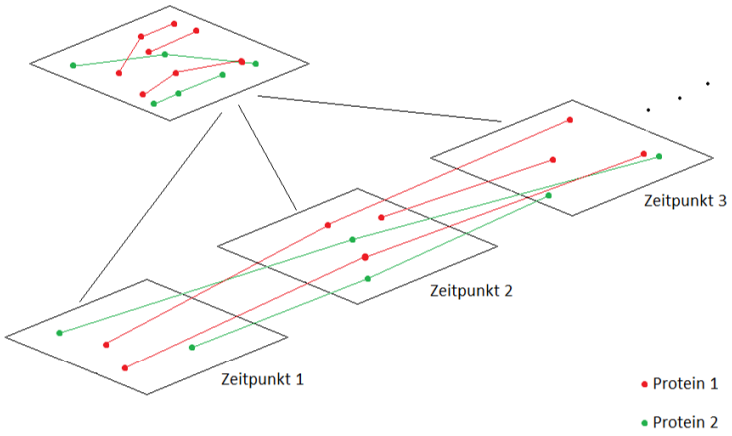

Abbildung 5.2: Vereinfachte schematische Darstellung der Datenerhebung bzw. anschließenden Rekonstruktion der Tracks für den Dual Colour Fall.

Proteintracks rekonstruiert. Dafür wurden in diesem Fall zwei MATLAB Pakete genutzt: `u-track` und `vbSPT`. Die Rekonstruktion der einzelnen Tracks, d.h. die Zuordnung zwischen den Frames, erfolgt dabei mit Hilfe von `u-track` (Jaqaman et al., 2008, [22]). Bei dieser Rekonstruktion werden für die jeweiligen Tracks die räumlichen (x/y)-Koordinaten pro Frame ermittelt. Anschließend kann mit `vbSPT` sowohl ein Diffusionsparameter als auch eine Klassifikation des Proteins bzgl. seiner Mobilität erfolgen: sehr beweglich, beweglich, immobil. Der Diffusionsparameter sowie die drei Stadien werden in dieser Arbeit jedoch nicht weiter betrachtet.
Einen Überblick über die Beschaffenheit der Daten ist beispielhaft für die Messung 0 Minuten nach Stimulation in Abbildung 5.3 bzw. in Abbildung 11 in Anhang C zu finden. Da ein Protein nur selten über alle 150 Frames (durchgängig) gemessen wurde, können die Tracks fehlende Werte enthalten. Diese können dabei in zwei Fälle eingeteilt werden:

1. der Wert fehlt, da es eine Unterbrechung des Tracks aufgrund von Blinken gab und
2. der Wert fehlt, da das Protein in dem Frame nicht aktiv war und daher nicht gemessen wurde.

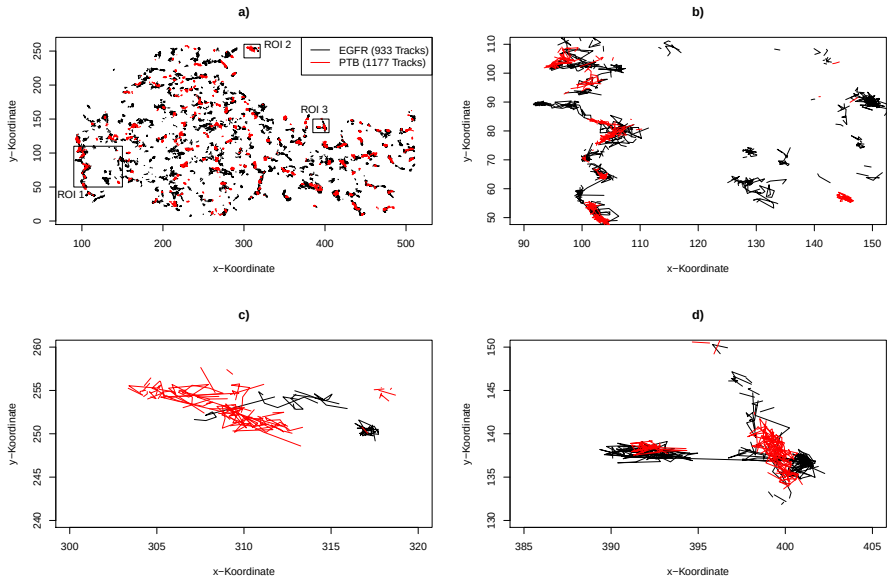

Abbildung 5.3: Darstellung aller EGFR- und PTB-Tracks aus dem Experiment 0 Minuten nach Stimulation für Zelle 1; dabei ist in a) die vollständige Zelle zu sehen, in b) der Ausschnitt, welcher in a) mit „ROI 1“, in c) der Ausschnitt, welcher mit „ROI 2“ und in d) der Ausschnitt, welcher in a) mit „ROI 3“ bezeichnet ist, abgebildet.

Um diese Unterscheidung der fehlenden Werte nochmals zu verdeutlichen, ist in Tabelle 5.1 ein kurzes Beispiel gegeben. So wurde Track 1 in diesem Beispiel von Frame 1 bis Frame 8 gemessen. In den Frames 9 und 10 war das Protein nicht mehr aktiv (gebleached) und konnte es nicht mehr gemessen werden. Somit sind an diesen Stellen in den Daten `NA`s enthalten. Diese fehlenden Werte würden dem zweiten Fall entsprechen. Analog dazu haben auch Track 2 und Track 3 fehlende Werte dieser Art: Track 2 für die Frames 1, 2, 3 und 10 sowie Track 3 für die Frames 1, 2, 8, 9, und 10. Weiter hat Track 2 in Frame 7 ebenfalls einen nicht gemessenen Wert. Dieser entspricht einem kurzzeitigen Blinken, sodass nur eine kurze Unterbrechung entsteht. Dieser durch Blinkent enstandene fehlende Wert entspricht Fall 1. Im vorliegenden Datensatz sind fehlende Werte aus Fall 2 häufiger enthalten.

Einen Überblick über die Daten der Messungen der ersten Zelle zu den weiteren Zeitpunkten nach Stimulation sind in den Abbildungen 12, 13 und 14 in Anhang C zu finden. Der Aufbau der Daten zu den anderen Messzeitpunkten nach Stimulation sowie die Qualität

Frame	Track 1	Track 2	Track 3
1	gemessen	-	-
2	gemessen	-	-
3	gemessen	-	gemessen
4	gemessen	gemessen	gemessen
5	gemessen	gemessen	gemessen
6	gemessen	gemessen	gemessen
7	gemessen	-	gemessen
8	gemessen	gemessen	-
9	-	gemessen	-
10	-	-	-

Tabelle 5.1: Veranschaulichung der Entstehung fehlender Werte.

dieser Daten ist wie oben beschrieben. Lediglich die Anzahl an gemessenen Tracks pro Protein ist unterschiedlich. Diese können ebenfalls den Abbildungen im Anhang entnommen werden.

5.1.3. Ziel der räumlich-zeitlichen Analyse

Wie bereits in Kapitel 2.1 erwähnt wurde, besteht die Zelle aus vielen verschiedenen Bestandteilen, u.a. vielen unterschiedlichen Proteinen. Diese können sich in ihrer Funktion ähneln, aber auch voneinander abhängen, d.h. es gibt interagierende Proteine. Diese Interaktionen können zu bestimmten Abläufen in der Zelle führen, aber auch zu deren Hemmung. Um nun die Funktionsweise einer Zelle verstehen zu können, müssen auch Interaktionen betrachtet werden.
Die räumliche Interaktion, d.h. die räumliche Clusteranalyse, wurde bereits in Kapitel 4 im Single wie im Dual Colour Fall untersucht. In diesem Kapitel sollen nun mit Hilfe von Dual Colour Single Particle Tracking erhobene Daten untersucht werden, wobei die Interaktion zwischen zwei Proteinen im räumlich-zeitlichen Kontext im Fokus steht. Die so gemessenen Tracks sollen somit Aufschluss über eine mögliche Abhängigkeit geben.
Als Interaktion von Proteinen kann dabei ein räumlicher Zusammenhang von zwei Pro-

teintracks im Verlauf der Zeit aufgefasst werden. Dabei sollte Trackpaaren, welche eine räumliche Nähe haben oder aber aufeinander zu laufen, ein hohes Maß an Zusammenhang zugeordnet werden.
Ziel der räumlich-zeitlichen Analyse stellt somit das Auffinden von Trackpaaren dar, welche einen hohen Zusammenhang aufweisen. Da hier der Dual Colour Fall betrachtet wird, besteht ein Trackpaar jeweils aus einem Track pro Protein, d.h. ein EGFR-Track sowie ein PTB-Track.
Folglich ist das statistische Ziel, ein Maß zu finden bzw. zu entwickeln, welches genau diesen Zusammenhang zwischen zwei Tracks repräsentiert. Dies erfolgt in Anlehnung an bereits existierende räumliche Korrelationsmaße. Dabei wird dieses Maß hier bewusst nicht Korrelationsmaß, sondern Zusammenhangsmaß genannt, da es unterschiedliche Eigenschaften, die einen Zusammenhang in unterschiedlichen Arten, mit einbeziehen wird.

5.2. Empirisches Zusammenhangsmaß für zwei Proteintracks

Im Folgenden soll nun ein Zusammenhangsmaß für zwei (Protein-)Tracks hergeleitet werden. Dafür wird zunächst eine Motivation über die möglichen Bestandteile des Maßes erfolgen um anschließend das Zusammenhangsmaß zu formulieren. Dabei sei im Folgenden ein Track von Objekt/Protein $i \in \{1,2\}$ durch $\mathcal{P}_i = \left(\begin{pmatrix} x_{i,1} \\ y_{i,1} \end{pmatrix}, \ldots, \begin{pmatrix} x_{i,N_i} \\ y_{i,N_i} \end{pmatrix} \right)$ gegeben, wobei $x_{i,j}$ die x- und $y_{i,j}$ die y-Koordinate von Protein i zum Zeitpunkt j beschreibt, $i \in \{1,2\}, j = 1, \ldots, N_i$. N_i gibt hingegen die Anzahl an **insgesamt** gemessenen Frames von Objekt i an, d.h. nicht die trackspezifische Anzahl an Frames, sondern die für das Experiment insgesamt gemessene Anzahl an Frames (meist $N_1 = N_2$), $j = 1, \ldots, N_i$, $i = 1, 2$. Demnach kann ein Track für Frames, in denen es nicht gemessen wurde, fehlende Werte enthalten (im Folgenden mit `NA` bezeichnet).

5.2.1. Motivation

Um ein Maß für den Zusammenhang zweier Tracks herzuleiten, muss zunächst überlegt werden, wodurch ein Zusammenhang zwischen zwei Tracks entstehen kann. Im Fall von Proteintracks in einer Zelle sollte ein Zusammenhang durch eine kleine Distanz repräsen-

tiert werden. Dabei kann dies eine grundsätzliche Nähe im räumlichen Sinn bedeuten, aber auch eine Bewegung aufeinander zu. Somit sollte die Distanz zwischen zwei Tracks eine Rolle spielen. Sei nun die euklidische Distanz für zwei Proteintracks $\mathcal{P}_1$ und $\mathcal{P}_2$ in Frame $j \in \{1, \ldots, N = \max\{N_1, N_2\}\}$ wie folgt definiert:

$$dist_j(\mathcal{P}_1, \mathcal{P}_2) = \begin{cases} \sqrt{(x_{1,j} - x_{2,j})^2 + (y_{1,j} - y_{2,j})^2} & , \mathcal{P}_1 \text{ \& } \mathcal{P}_2 \text{ in Frame } j \text{ gemessen,} \\ \texttt{NA} & , \text{sonst.} \end{cases}$$

Somit ergibt sich die Distanz für zwei (Protein-)Tracks über alle Frames hinweg zu einem Vektor: $\mathcal{D}(\mathcal{P}_1, \mathcal{P}_2) = (dist_j(\mathcal{P}_1, \mathcal{P}_2))_{j \in \{1,\ldots,N\}}$. Durch die Distanz als Funktion über die Zeit hinweg, kann nun eine Aussage über das gemeinsame räumliche Verhalten von zwei (Protein-)Tracks getroffen werden, so z.B. ob sich die beiden Tracks auf einander zu oder von einander weg bewegen.

Weiter kann neben der Distanz noch eine weitere räumliche Komponente betrachtet werden: die Clusterzugehörigkeit. Wie bereits im vorherigen Kapitel der räumlichen Analyse kann auch hier in jedem Frame eine räumliche Clusteranalyse vorgenommen werden, z.B. mit dem DBSCAN Algorithmus (vgl. Kapitel 4.2.4) oder auch dem EMM (vgl. Kapitel 4.2.3). Durch die frameweise Anwendung der Clustermethode wird in jedem Frame j einem Track von Protein i eine Cluster ID, ID_{ij} mit $j = 1, \ldots, N$ und $i = 1, 2$, zugeteilt. Besteht zwischen zwei (Protein-)Tracks ein Zusammenhang, so sollten diese Tracks möglichst oft demselben Cluster zugeordnet werden. Daher sollte auch diese Komponente in dem Zusammenhangsmaß enthalten sein.

Weiter kann die Anzahl $N^\star$ an Frames, in denen beide (Protein-)Tracks gemeinsam gemessen wurden, genauer betrachtet werden. Sind zwei Proteintracks nur in einem Frame (oder nur sehr wenigen Frames) gemeinsam gemessen worden, so kann vermutet werden, dass eine mögliche räumliche Nähe eher dem Zufall entspricht. (Protein-)Tracks, welche über mehrere Frames hinweg eine räumliche Nähe aufbauen oder besitzen, lassen hingegen eher einen vorhandenen Zusammenhang vermuten. Auch diese Eigenschaft sollte in das Maß einfließen.

Im Folgenden ist nun das Ziel die drei vorher beschriebenen Eigenschaften zu einem Zusammenhangsmaß zusammenzufügen. Dabei sei beachtet, dass bei der Distanz eher ein kleiner Wert einem Zusammenhang zwischen zwei Proteintracks entspricht. Bei der Clusterzugehörigkeit ist hingegen eine hohe Anzahl an Frames positiv, in denen beide Tracks

demselben Cluster zugeordnet wurden. Bei der Anzahl an gemeinsam gemessenen Frames ist ebenfalls eine hohe Anzahl gut, da dies einen möglichst hohen Informationsgehalt repräsentieren würde.
Da i.Allg. ein Korrelationsmaß auf das Intervall $[-1, 1]$ begrenzt ist, kann hier ebenfalls eine obere/untere Grenze für das Zusammenhangsmaß thematisiert werden, um eine gute Interpretation des Maßes zu gewährleisten.

5.2.2. Formulierung eines empirischen Zusammenhangsmaßes für zwei Proteintracks

Um nun die einzelnen Bestandteile der Motivation sinnvoll zu einem Maß zusammenzufügen, müssen sie zunächst in eine jeweils passende Form gebracht werden. Ein Ziel ist dabei, dass stets ein großer Wert einen hohen Zusammenhang zweier (Protein-)Tracks repräsentiert.
Für die Distanz zweier (Protein-)Tracks ist jedoch ein kleiner Wert positiv zu werten. Daher sollte hier eine Umformung mit Hilfe eines Vergleichswertes vorgenommen werden. Dabei wurde sich an Moran's I orientiert, welches ein Maß für räumliche Autokorrelation ist (vgl. u.a. Bivand, Pebesma und Gòmez-Rubio, 2013, Kapitel 9, S. 284 ff, [4]). Im Fall von Moran's I wird das arithmetische Mittel als Vergleichswert herangezogen und die Differenz gebildet. Auf den hier vorliegenden Fall kann die Idee adaptiert werden, jedoch wird -im Gegensatz zu Moran's I- jede Distanz mit dem (globalen) Maximum der Distanzen zweier (Protein-)Tracks $\mathcal{D}_{max}$ verglichen. Dieses ist gegeben durch

$$\mathcal{D}_{max} = \max_{j=1,\dots,N, \mathcal{P}_1 \in \Psi_1, \mathcal{P}_2 \in \Psi_2} dist_j(\mathcal{P}_1, \mathcal{P}_2) I_{[dist_j \neq \texttt{NA}]},$$

wobei Ψ_i die Menge aller Tracks von Objekt/Protein i ist, $i = 1, 2$. Somit ergibt sich zunächst

$$\sum_{j=1}^{N} (\mathcal{D}_{max} - dist_j(\mathcal{P}_1, \mathcal{P}_2) I_{[dist_j \neq \texttt{NA}]}),$$

wobei I_A die Indikatorfunktion ist, für die gilt, dass sie den Wert eins annimmt, sofern das Argument A erfüllt ist und null sonst. Für Track-Paare mit geringen Distanzen und somit einem hohen Zusammenhang konvergiert diese Summe nun gegen $N^\star \mathcal{D}_{max}$, für Track-Paare mit hohen Distanzen konvergiert die Summe hingegen gegen Null, sodass

bereits eine untere Grenze durch Null gegeben ist. Um eine Beschränkung nach oben zu gewährleisten muss die Summe nun noch durch $N^\star \mathcal{D}_{max}$ dividiert werden, sodass sie nun folgender erster Bestandteil des Zusammenhangsmaßes ergibt

$$\sum_{j=1}^{N} \frac{\mathcal{D}_{max} - dist_j(\mathcal{P}_1, \mathcal{P}_2)}{N^\star \mathcal{D}_{max}} I_{[dist_j \neq \texttt{NA}]}. \tag{16}$$

Somit ist durch den Vergleich der frameweisen mit der maximalen Distanz und anschließender Division durch $N^\star \mathcal{D}_{max}$ eine Skalierung auf dem Intervall $[0, 1]$ erfolgt, wobei ein Zusammenhang durch einen Wert nahe eins widergespiegelt wird.

Bei der Clusterzugehörigkeit ist bereits eine hohe Anzahl an Frames mit gleicher Clusterzuordnung zweier (Protein-)Tracks als ein Indiz für einen Zusammenhang zu werten. Jedoch sollte noch eine Skalierung vorgenommen werden:

$$\frac{1}{N^\star} \sum_{j=1}^{N} I_{[ID_{1j} = ID_{2j}]} I_{[dist_j \neq \texttt{NA}]}.$$

Somit wird hier die Anzahl an Frames, in denen beide (Protein-)Tracks demselben Cluster zugeordnet wurden, bestimmt und durch die maximale Anzahl an möglichen Übereinstimmungen skaliert. Dadurch entspricht weiterhin ein hoher Wert einem starken räumlichen Zusammenhang und ferner ist der Ausdruck nach oben durch eins beschränkt.

Abschließend wird nun die „Länge des gemeinsamen Weges von zwei (Protein-)Tracks“ betrachtet, d.h. die Anzahl an Frames, in denen beide (Protein-)Tracks simultan gemessen wurden. Auch hier ist wie zuvor ein hoher Wert, d.h. eine hohe Anzahl, positiv für einen möglichen Zusammenhang zu werten. Denn je länger zwei (Protein-)Tracks zusammen gemessen werden konnten, desto mehr Informationen konnten gesammelt werden. Diese Anzahl sollte jedoch, wie zuvor die Clusterzugehörigkeit, skaliert werden, um eine Beschränkung nach oben zu gewährleisten. An dieser Stelle kann die Skalierung mit Hilfe von N erfolgen. Da nicht alle (Protein-)Tracks durchgängig über alle Frames hinweg gemessen werden, sondern eher eine kürzere Verweildauer haben, wurde an dieser Stelle N_{max} genutzt. Dabei gibt N_{max} das Minimum der beiden Maxima der gemessenen Tracklängen des jeweiligen Objekts/Proteins an, d.h.

$$N_{max} = \min\left(\max\{\text{Tracklängen von Objekt 1}\}, \max\{\text{Tracklängen von Objekt 2}\}\right).$$

Die skalierte Form ergibt sich dann zu

$$\frac{N^\star}{N_{max}}.$$

Nun können die drei Komponenten durch entsprechende Gewichte w_k, $k = 1, 2, 3$, zu einem Zusammenhangsmaß für zwei (Protein-)Tracks zusammengefügt werden:

$$\begin{aligned} \varrho(\mathcal{P}_1, \mathcal{P}_2) &= w_1 \cdot \sum_{j=1}^{N} \frac{\mathcal{D}_{max} - dist_j(\mathcal{P}_1, \mathcal{P}_2)}{N^\star \mathcal{D}_{max}} I_{[dist_j \neq \texttt{NA}]} \\ &+ w_2 \cdot \frac{1}{N^\star} \sum_{j=1}^{N} I_{[ID_{1j}=ID_{2j}]} I_{[dist_j \neq \texttt{NA}]} \\ &+ w_3 \cdot \frac{N^\star}{N_{max}}, \end{aligned} \tag{17}$$

wobei stets $w_1 + w_2 + w_3 = 1$ gilt. Dieses Maß repräsentiert somit alle zuvor argumentierten möglichen Zusammenhänge. Durch die Gewichte ist es weiter möglich, den einzelnen Komponenten einen unterschiedlichen Einfluss in der Berechnung des Zusammenhangs zu geben. Speziell kann durch Wahl des Gewichtes w_3 der Fokus auf (Protein-)Tracks mit einer langen, aber auch entsprechend kurzer gemeinsamer Wegstrecke gelegt werden.
Weiter können folgende Eigenschaften des Zusammenhangsmaßes festgehalten werden:

- Das Zusammenhangsmaß ist auf das Intervall $[0, 1]$ beschränkt, wobei ein Wert nach Null auf keinen Zusammenhang hinweist, ein Wert nahe eins hingegen auf einen starken Zusammenhang.
- Weiter ist dieses Maß durch seine Konstruktion Skaleninvariant, sodass die Anwendung auf Daten unterschiedlicher Skala möglich ist.

Die Wahl der Gewichte wird in der Analyse des simulierten Beispiels noch näher erläutert.

5.3. Analyse der Protein-Trackingdaten

In Folgenden soll nun das Zusammenhangsmaß angewendet und evaluiert werden. Dafür wird zunächst das simulierte Beispiel aus Kapitel 5.1.1 betrachtet. Die Analyse erfolgt dabei für verschiedene Gewichtungen. Anschließend werden die experimentellen Daten aus Kapitel 5.1.2 mit Hilfe des Zusammenhangsmaßes untersucht. Die Analysen wurden mit Hilfe der Programmiersoftware R (2015, [38]) durchgeführt.

5.3.1. Validierung des Zusammenhangsmaßes anhand eines simulierten Beispiels

Zunächst soll das Zusammenhangsmaß auf einem simulierten Trackingbeispiel getestet und evaluiert werden. Dafür wird das Beispiel aus Abbildung 5.1 genutzt. Es ist deutlich zu erkennen, dass in diesem Beispiel viele verschiedene Situationen enthalten sind. So bewegen sich Track 1 des grünen Objekts und Track 1 des roten Objekts (kurz: Trackpaar 1 grün/1 rot) aufeinander zu und bleiben anschließend in einer räumlich nahen Umgebung. Hier sollte somit ein hoher Wert für das Zusammenhangsmaß berechnet werden. Weiter kreuzen sich Track 2 des grünen und Track 2 des roten Objekts einmalig (kurz: Trackpaar 2 grün/2 rot) und haben somit für einen gewissen Zeitraum eine räumliche Nähe. Dadurch wird hier ein Zusammenhang vorliegen, es sollte jedoch ein kleinerer Wert (durch die lediglich kurze räumliche Nähe) berechnet werden. Bei den Trackpaaren 3 grün/3 rot, 4 grün/4 rot und 8 grün/8 rot gibt es für einen gewissen Zeitraum eine räumliche Nähe. Bei diesen Trackpaaren ist jedoch immer ein (Protein-)Track recht kurz. Somit liegt hier ein Zusammenhang vor, welcher jedoch durch die Kürze des gemeinsamen Beobachtungszeitraums weniger auffällig sein kann.

Um nun das Zusammenhangsmaß anwenden zu können, wird zunächst eine frameweise Clusterung durchgeführt. Dafür wird hier der DBSCAN Algorithmus mit $\epsilon = 3$ und $MinPts = 2$ verwendet. Es sei zu beachten, dass durch die Parameterwahl der frameweisen Clusterung hier ebenfalls die räumliche Nähe beurteilt werden kann. Wird ein kleiner Wert für ϵ gewählt, so werden kleine Cluster gebildet und somit eine kleine räumliche Distanz positiv gewertet. Der Parameter ϵ sollte jedoch auch nicht zu klein gewählt werden, da sonst keine Cluster gebildet werden können und folglich alle Proteine dem Background zugeordnet werden.

Wird nun das Zusammenhangsmaß aus (17) mit den Gewichten $w_1 = w_2 = w_3 = 1/3$ genutzt, so ergeben sich folgende Werte in Tabelle 5.2 für das Zusammenhangsmaß. Es ist gut zu erkennen, dass in diesem Fall, wie zuvor diskutiert, Trackpaar 1 grün/1 rot ein hoher Zusammenhang zugeordnet wird. Weiter ist bei Trackpaar 2 grün/2 rot ebenfalls eine hoher Zusammenhang vorhanden. Die Tracks 3 und 4 des grünen Objekts/Proteins weisen die höchsten Werte für den Zusammenhang mit Track 3 bzw. Track 4 des roten Objekts/Proteins auf. Somit ist hier genau das Ergebnis vorzufinden, was vorab vermutet

		grüne Tracks							
		Track 1	Track 2	Track 3	Track 4	Track 5	Track 6	Track 7	Track 8
rote Tracks	Track 1	**0.8926**	0.5426	0.3967	0.6011	0.2537	0.4912	0.3127	**0.8800**
	Track 2	0.4511	**0.8710**	0.3919	0.3220	0.4001	0.3359	0.2799	0.4630
	Track 3	0.4393	0.4041	**0.8200**	0.2332	0.3456	0.2113	0.2810	0.3914
	Track 4	0.7102	0.5397	0.3350	**0.7871**	0.2687	**0.5904**	0.2293	0.8732
	Track 5	0.5015	0.6006	0.5817	0.2761	**0.4757**	0.2634	**0.3303**	0.4758
	Track 6	0.5688	0.4825	0.3353	0.3262	0.2527	0.2825	0.3164	0.5528
	Track 7	0.3183	0.1943	0.1727	0.2446	0.0778	0.1460	0.2108	0.3220
	Track 8	0.7264	0.3369	0.2769	0.6552	0.2270	0.5455	0.2343	0.7866

Tabelle 5.2: Ergebnis der Anwendung des Zusammenhangsmaßes auf simulierte Beispieltracks mit Gewichtung $w_1 = w_2 = w_3 = 1/3$; die dick gedruckten Einträge entsprechen den jeweils höchsten Werten des Zusammenhangsmaßes für die einzelnen **grünen** Tracks.

wurde.

Für Track 8 des grünen Objekts/Proteins liegt, der höchste Zusammenhang mit Track 1 des roten Objekt/Proteins vor, welcher aus der räumlichen Nähe der zwei Tracks plausibel erscheint. Ein weiterer hoher Zusammenhang ist jedoch auch mit Track 4 und Track 8 des roten Objekt/Proteins zu beobachten. Dies ergibt sich, da in den entsprechenden Frames, in denen Track 8 des grünen Objekts/Proteins gemessen wurde, die Tracks 4 und 8 des roten Objekts/Proteins ebenfalls räumlich sehr nahe waren. Der Zusammenhang ist hier jedoch geringer, da die Anzahl gemeinsam gemessener Frames für Track 8 des roten Objekts/Proteins geringer bzw. bei Track 4 des roten Objekts/Proteins die Distanz größer war (im Vergleich zu Track 1 des roten Objekts/Proteins).

Ändert man die Blickrichtung und betrachtet in Tabelle 5.2 z.B. die letzte Zeile und somit die Zusammenhänge für Track 8 des roten Proteins, so wird hier der größte Zusammenhang für das Trackpaar 8 grün/8 rot berechnet. Dies spiegelt die vorab erörterte Vermutung wieder.

Für Track 5 des grünen Objekts/Proteins ist der höchste Zusammenhang mit Track 5 des roten Objekts/Proteins gegeben, wobei dieser deutlich geringer ist, als die zuvor be-

trachteten Maxima des Zusammenhangsmaß. Dies ist plausibel, da Track 5 des roten Objekts/Proteins in den gemessenen Frames (von Track 5 grün) räumlich weiter entfernt verläuft als im Vergleich das Trackpaar 1 grün/1 rot. Auch für Track 6 des grünen Objekts/Proteins sind die berechneten Zusammenhänge plausibel. So ist die Track 6 des grünen Tracks in den Frames, in denen es gemessen wurde, zu Track 4, 8 und 1 des roten Objekts/Proteins am nähsten. Weiter ist Track 6 des grünen Objekts/Proteins ein recht kurzer Track, so dass an dieser Stelle auch die Anzahl gemeinsam gemessener Frames eine Bedeutung spielt, wie man an dem Wert für den Zusammenhang des Trackpaares 6 grün/2 rot sehen kann. Hier ist ebenfalls eine räumliche Nähe vorhanden, die zwei Tracks haben jedoch eine geringere Anzahl gemeinsam gemessener Frames, sodass hier ein geringerer Zusammenhang berechnet wird.

Für Track 7 des grünen Objekts/Proteins sind die Werte des Zusammenhangsmaßes sehr ähnlich und vergleichsweise gering. Diese berechneten Werte resultieren dabei hauptsächlich aus der Anzahl gemeinsam gemessener Frames. Dies war zu erwarten, da Track 7 des grünen Objekts/Proteins lediglich zu zwei Zeitpunkten gemessen wurde und eher am Rand lokalisiert ist.

Somit liefert das Zusammenhangsmaß mit der Gewichtung $w_1 = w_2 = w_3 = 1/3$ plausible Ergebnisse. Eine weitere Überlegung ist nun, die Anzahl an gemeinsam gemessenen Frames weniger stark und dadurch die räumliche Struktur mehr zu gewichten. Dies wurde für dieses Beispiel in Tabelle 7 in Anhang B gemacht. Dort wurden die Gewichte $w_1 = w_2 = 0.4$ und $w_3 = 0.2$ genutzt.

Es ist gut zu sehen, dass sich die Werte des Zusammenhangsmaßes sehr zu denen aus Tabelle 5.2 ähneln und auch die Ordnungen bei den Tracks 2 bis 8 des grünen Objekts/Proteins gleich bleiben. Für Track 1 des grünen Objekts/Proteins ist nun der höchste Zusammenhang mit Track 1 des roten Objekts/Proteins gegeben, der zweithöchste mit Track 8 des roten Objekts/Proteins., sodass hier die Ränge getauscht haben. Dies ist durch das Beispiel weiterhin plausibel und das Ergebnis, welches nach der ersten Überlegung zu erwarten war. Somit liefert auch diese Gewichtung gute Ergebnisse, schwächt aber den Einfluss der Länge „des gemeinsamen Weges“ etwas ab.

Weiter kann nun noch der Einfluss des dritten Parts im Zusammenhangsmaß vollständig gekürzt werden, d. h. es kann $w_3 = 0$ gesetzt werden. Für den Fall, dass die weiteren Ge-

wichte gleichwertig sind, d. h. $w_1 = w_2 = 0.5$, sind die Werte des Zusammenhangsmaßes in Tabelle 5.3 abgetragen. Im Gegensatz zu den vorherigen zwei Gewichtungen, ist nun für

		grüne Tracks							
		Track 1	Track 2	Track 3	Track 4	Track 5	Track 6	Track 7	Track 8
rotes Tracks	Track 1	0.8390	0.3139	0.2451	0.7016	0.1305	0.5868	0.3691	0.8200
	Track 2	0.3266	**0.9565**	0.2878	0.2830	0.3501	0.3538	0.3198	0.3446
	Track 3	0.3090	0.2562	**0.9800**	0.1498	0.2684	0.1669	0.3716	0.2371
	Track 4	0.5653	0.3096	0.1525	**0.9806**	0.1530	**0.7357**	0.2439	0.8098
	Track 5	0.3523	0.5010	0.5226	0.2142	**0.4635**	0.2451	**0.3954**	0.3137
	Track 6	0.6532	0.5238	0.3030	0.3393	0.2290	0.3238	0.3746	0.6292
	Track 7	0.3774	0.1914	0.1591	0.2669	0.0167	0.1690	0.2662	0.3830
	Track 8	**0.8896**	0.3054	0.2153	0.7827	0.1405	0.6683	0.3014	**0.9799**

Tabelle 5.3: Ergebnis der Anwendung des Zusammenhangsmaßes auf simulierte Beispieltracks mit Gewichtung $w_1 = w_2 = 0.5$ und $w_3 = 0$; die dick gedruckten Einträge entsprechen den jeweils höchsten Werten des Zusammenhangsmaßes für die einzelnen **grünen** Tracks.

Track 1 des grünen Objekts/Proteins der Wert des Zusammenhangsmaß mit Track 8 des roten Objekt/Proteins am höchsten. Weiter ist erneut, wie in der vorherigen Gewichtung, nimmt das Zusammenhangsmaß den höchsten Wert für das Trackpaar 8 grün/8 rot an, wobei der Wert nun noch einmal deutlich angestiegen ist. Dies resultiert aus der Tatsache, dass nun nur noch die Distanz und die Clusterzugehörigkeit in das Zusammenhangsmaß einfließen. So ist z. B. bei Track 1 des grünen Objekts/Proteins der Weg aufeinander zu mit höherer Distanz negativ gewertet und die Anzahl an gemeinsam gemessenen Frames kann dies nicht ausgleichen. Hingegen ist bei Trackpaar 1 grün/8 rot der gemeinsame Weg verhältnismäßig kurz, jedoch verlaufen die zwei Tracks in dieser Zeit räumlich nah, sodass ein höherer Zusammenhang berechnet wird. Bei dieser Gewichtung ist somit der Fokus lediglich auf der räumlichen Nähe. Ein Annähern der Tracks würde durch die zusätzlichen Frames, in denen eine relativ hohe Dichte vorliegt, eher negativ gewertet werden.
Somit ist (sofern kein weiteres Expertenwissen vorliegt) eine Gewichtung mit $w_i \neq 0, i = 1, 2, 3$, empfehlenswert. Dabei ist die Wahl von w_3 davon abhängig, wie sehr man die

Anzahl gemeinsam gemessener Frames gewichten will. Hier haben sich sowohl die Gewichtung $w_1 = w_2 = w_3 = 1/3$ als auch $w_1 = w_2 = 0.4$ und $w_3 = 0.2$ als gut erwiesen.
An dieser Stelle sei nochmals darauf hingewiesen, dass N_{max} in (17) nicht die insgesamt gemessene Anzahl an Frames repräsentiert. Dies wird an der folgenden Modifikation des obigen Beispiels deutlich:
Dafür werden im obigen Beispiel die Tracks 1, 4 und 5 des roten Objekts/Proteins wie folgt gekürzt:

- von Track 1 werden die Beobachtungen in den ersten zwei Frames entfernt,
- von Track 4 werden die Beobachtungen in den letzten zwei Frames entfernt und
- von Track 5 werden die ersten zwei gemessenen Beobachtungen (in den Frames 2 und 3) entfernt.

Das daraus resultierende Beispiel ist in Abbildung 15 in Anhang C zu sehen. In diesem Fall ist die Anzahl insgesamt gemessener Frames 10, N_{max} hingegen ist 8. Wenn man nun auf dieses Beispiel das Zusammenhangsmaß mit der Gewichtung $w_1 = w_2 = w_3 = 1/3$ anwendet und zum einen $N_{max} = 8$, zum anderen $N = 10$ statt N_{max} verwendet, so erhält man die Werte des Zusammenhangsmaß für die modifizierten Tracks, welche in Tabelle 5.4 abgetragen sind. Es ist gut zu erkennen, dass alle Zusammenhänge unter Verwendung von N kleinere Werte erhalten und somit (klare) Zusammenhänge ggf. nicht erkannt werden können. Dieser Unterschied wird mit ansteigender Anzahl an Frames, N, stets größer, da die Tracks (meistens) kürzer sind als N. Dieser Unterschied ist jedoch nur dann vorhanden, sofern $w_3 \neq 0$.

5.3.2. Analyse experimenteller Trackingdaten

Im Folgenden sollen nun experimentelle Trackingdaten im Hinblick auf einen möglichen Zusammenhang untersucht werden. Dafür wird nun für jeden Zeitpunkt nach Stimulation der Datensatz der ersten gemessenen Zelle analysiert.
Eine Darstellung des ersten zu analysierenden Datensatzes, erhoben 0 Minuten nach der Stimulation, war bereits in Kapitel 5.1.2 in Abbildung 5.3 zu sehen. Dort ist gut zu erkennen, dass sich die Tracks auf der x-Achse im Bereich zwischen 0 und 520 nm sowie auf

		rote Tracks					
		Track 1		Track 4		Track 5	
		$N_{max} = 8$	$N = 10$	$N_{max} = 8$	$N = 10$	$N_{max} = 8$	$N = 10$
grüne Tracks	Track 1	0.9737	0.9071	0.6614	0.5947	0.5090	0.4590
	Track 2	0.5685	0.5018	0.5364	0.4697	0.6160	0.5660
	Track 3	0.4550	0.3967	0.3582	0.3082	0.6170	0.5670
	Track 4	0.6344	0.6011	0.8204	0.7871	0.3095	0.2761
	Track 5	0.2954	0.2537	0.3104	0.2687	0.4007	0.3673
	Track 6	0.5162	0.4912	0.6154	0.5904	0.2884	0.2634
	Track 7	0.3294	0.3127	0.2459	0.2293	0.3469	0.3303
	Track 8	0.9085	0.8418	0.8536	0.7869	0.4744	0.4244

Tabelle 5.4: Vergleich des Zusammenhangs der drei modifizierte Tracks des Trackingbeispiels unter Verwendung von N_{max} bzw. N im dritten Part des Zusammenhangsmaßes mit Gewichtung $w_1 = w_2 = w_3 = 1/3$.

der y-Achse zwischen 0 und 275 nm bewegen. Weiter liegen in dem Datensatz 933 Tracks des EGFR-Proteins und 1177 Tracks des PTB-Protein vor. Weiter wurden insgesamt 150 Frames gemessen. Die Längen der (Einzel-)Tracks, d. h. die Anzahl an Frames in denen

Protein	minimale Länge	mediane Länge	mittlere Länge	maximale Länge
EGFR	10	20	27.4802	150
PTB	2	5	11.1589	139

Tabelle 5.5: Übersicht der Tracklängen für den experimentellen Datensatz in Abhängigkeit des Proteins.

die Tracks gemessen wurden, in Abhängigkeit des Proteins, können Tabelle 5.5 entnommen werden. Es ist gut zu erkennen, dass für das PTB-Protein kürzere Tracks vorliegen als für das EGFR-Protein, sowohl im Mittel als auch im Median.
Um bereits einen ersten Eindruck über mögliche zusammenhängende Trackpaare zu erhalten, wird zunächst die paarweise Distanz berechnet. Anschließend kann mit Hilfe eines Cutpoints für die mittlere Distanz eine Auswahl an Trackpaarungen getroffen werden. Für verschiedene Cutoffs ist dieses Vorgehen in Abbildung 5.4 zu sehen, wobei hier beispiel-

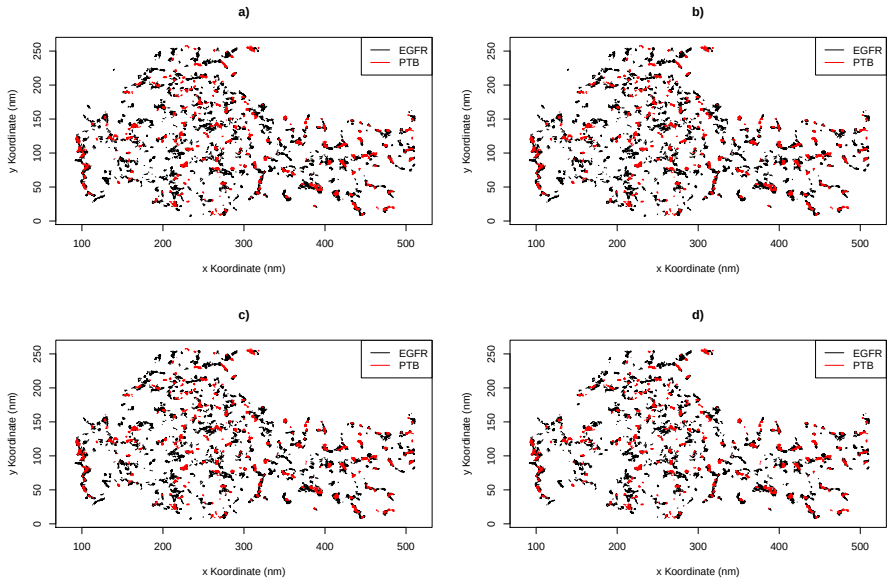

Abbildung 5.4: Auswahl der Tracks in Abhängigkeit eines festen Cutoffs für die mittlere Distanz der Trackpaare: a) Cutoff = 100, b) Cutoff = 50, c) Cutoff = 25, d) Cutoff = 15.

haft 100, 50, 25 und 15 als beliebige Cutoffs genutzt wurden. Es ist gut zu erkennen, dass sich mit kleiner werdendem Cutoff die Dichte der Proteintracks verringert. Eine deutliche Reduzierung ist jedoch nur für einzelne Regionen zu erkennen. Wie viele Trackpaare, in Abhängigkeit vom Cutoff, ausgewählt werden, ist in Abbildung 16 in Anhang C zu sehen. Die fehlende globale Reduzierung kann dabei weiter durch verschiedene Arten der Verläufe der Trackpaare erklärt werden. So können zwei Tracks, die lediglich wenige (z. B. zwei) Frames zusammen gemessen wurden, aber eine räumliche Nähe haben, eine geringe mittlere Distanz aufweisen. Für EGFR-Track 3 sind beispielhaft die Distanzen zu fünf PTB-Tracks bei einer mittleren Distanz kleiner als 5 nm betrachtet worden. Die Distanzkurven sind in Abbildung 17 in Anhang C abgetragen. Es ist gut zu sehen, dass unter diesen fünf Trackpaaren sowohl kurze als auch lange gemeinsam gemessene Trackpaarungen vorliegen. Je mehr gemeinsame Beobachtungspunkte jedoch vorhanden sind, desto besser kann ein möglicher Zusammenhang validiert werden. Dies unterstreicht noch einmal die Relevanz, weshalb neben der Distanz weitere Faktoren in das Zusammenhangsmaß mit aufgenommen wurden.

Das Zusammenhangsmaß nutzt nun neben der Distanz zweier Trackpaare auch die Clus-

ter ID sowie die Länge des „gemeinsamen Weges“. Wenn nun das Zusammenhangsmaß aus Formel (17) auf den experimentellen Datensatz angewendet wird, so erhält man eine Matrix der Dimension 933×1177 mit 1 098 141 Werten für den Zusammenhang als Einträge.

Wie in der Simulation zu sehen war, ist es ratsam alle drei Gewichte ungleich Null zu wählen. Daher wurden für die experimentellen Daten zum einen die Gewichtung mit $w_1 = w_2 = w_3 = 1/3$ sowie die Gewichtung $w_1 = w_2 = 0.4$ und $w_3 = 0.2$ genutzt. Weiter müssen noch die Parameter für den DBSCAN Algorithmus definiert werden, da auch eine Clusterzugehörigkeit mit in das Zusammenhangsmaß einfließt. Dazu wurde zunächst die Parameterwahl aus Masip et al. (2016, [29]) genutzt, d.h. $\epsilon = 40$ und $MinPts = 5$. Wie

	Gewichtung	
Kennzahl	$w_1 = w_2 = w_3 = 1/3$	$w_1 = w_2 = 0.4, w_3 = 0.2$
Minimum	0.1425	0.1696
25%-Quantil	0.5129	0.6067
Median	0.5659	0.6700
Mittelwert	0.5592	0.6610
75%-Quantil	0.6117	0.7241
Maximum	0.9631	0.9643
Varianz	0.0049	0.0066
Standardabweichung	0.0700	0.0814
Anzahl `NA`s	824 523	824 523
N_{max}	139	139

Tabelle 5.6: Übersicht wichtiger Kennzahlen der errechneten Zusammenhänge in Abhängigkeit der Gewichtung mit Parameterwahl $\epsilon = 40$ und $MinPts = 5$ für den DBSCAN Algorithmus.

man in Tabelle 5.6 sehen kann, ist die Streuung der Werte des Zusammenhangsmaßes mit dieser Parameterwahl für den DBSCAN Algorithmus sehr gering. Dies resultiert aus der hohen Parametereinstellung bzgl. ϵ, da so die Cluster sehr groß gefasst werden und somit auch viele Tracks häufig dem selben Cluster zugeordnet werden. Daher wurde noch eine weitere Parameterwahl für den DBSCAN Algorithmus genutzt: $\epsilon = 10$ und $MinPts = 2$.

Dabei wurde simultan auch $MinPts$ verändert, sodass die Parameter keine zu starke DBSCAN Einstellung darstellen. Eine Übersicht der Lagemaße sowie anderen Kennzahlen des Zusammenhangs für eben diese Parameterwahl in Abhängigkeit der zwei Gewichtungen ist in Tabelle 5.7 zu sehen. Für diese Parameterwahl ist nun die Streuung etwas höher sowie der Mittelwert und der Median deutlich kleiner. Durch diese Parameterwahl des DBSCAN Algorithmus unterscheiden sich die Werte für den Zusammenhang somit mehr und hohe Werte werden nur seltener angenommen, wodurch interessante Paarungen besser identifiziert werden können. Daher scheint diese Parameterwahl eine sinnvollere Wahl

	Gewichtung	
Kennzahl	$w_1 = w_2 = w_3 = 1/3$	$w_1 = w_2 = 0.4, w_3 = 0.2$
Minimum	0.0034	0.0027
25%-Quantil	0.1797	0.2069
Median	0.2327	0.2701
Mittelwert	0.2288	0.2645
75%-Quantil	0.2786	0.3242
Maximum	0.8736	0.9160
Varianz	0.0064	0.0088
Standardabweichung	0.0800	0.0937
Anzahl `NA`s	824 523	824 523
N_{max}	139	139

Tabelle 5.7: Übersicht wichtiger Kennzahlen der errechneten Zusammenhänge in Abhängigkeit der Gewichtung mit Parameterwahl $\epsilon = 10$ und $MinPts = 2$ für den DBSCAN Algorithmus.

zu sein.

Weiter ist in Tabelle 5.7 gut zu erkennen, dass das Zusammenhangsmaß bei einer Gewichtung mit $w_1 = w_2 = w_3 = 1/3$ eine etwas kleinere Varianz sowie Standardabweichung besitzt. Weiter liegen die Mittelwerte und auch die Mediane bei beiden Gewichtungen nah beieinander: die Mittelwerte betragen 0.2288 bzw. 0.2645 sowie die Mediane 0.2327 und 0.2701.

Sollen nun diejenigen Trackpaare gesucht werden, welche einen möglichst hohen Zusam-

menhang aufweisen, so könnten die Werte des Zusammenhangsmaß sortiert werden und die Trackpaarungen mit einem möglichst hohen Rang gewählt werden. Weitere Möglichkeiten für die Wahl eines Cutoffs bestehen in einer beliebigen Wahl sowie über die Berechnung von Quantilen.

Hier wird zunächst die Wahl des Cutoffs mittels Quantile betrachtet. Für die DBSCAN Parameter $\epsilon = 10$ und $MinPts = 2$ berechnet sich das 95%- bzw. das 99%-Quantil für eine Gewichtung mit $w_1 = w_2 = w_3 = 1/3$ zu 0.3262 bzw. 0.4430. Im Vergleich dazu beträgt das 95%- bzw. 99%-Quantil der Werte des Zusammenhangsmaßes unter Verwendung der DBSCAN Einstellungen aus Masip et al. für die Gleichgewichtung 0.6592 bzw. 0.6938. Auch hier wird erneut deutlich, dass durch die DBSCAN Parameterwahl nach Masip et al. der berechnete Zusammenhang oft hohe Werte annimmt. Für eine Gewichtung des Zusammenhangsmaßes mit $w_1 = w_2 = 0.4, w_3 = 0.2$ ergeben sich das 95%- bzw. das 99%-Quantil mit $\epsilon = 10$ und $MinPts = 2$ zu 0.3774 bzw. 0.4792 (0.7773 bzw. 0.8010 für eine Parametereinstellung nach Masip et al.). Bei beiden Gewichtungen des Zusammenhangsmaßes ist deutlich zu erkennen, dass für eine Wahl der DBSCAN Parameter $\epsilon = 10$ und $MinPts = 2$ sowohl das 95%-Quantil als auch das 99%-Quantil für beide Gewichtungen kleiner als 0.5 sind. Für die Parametereinstellung aus Masip et al. hingegen sind beide Quantile jeweils größer als 0.5. Dies scheint im ersten Moment ein großer Unterschied, wenn man sich jedoch Tabelle 5.6 in Erinnerung ruft, so sieht man, dass für die Parameterwahl nach Masip et al. die Werte für das Zusammenhangsmaß im Mittel und auch im Median generell größere Werte annimmt im Vergleich zur Wahl $\epsilon = 10$ und $MinPts = 2$. Somit ist es hier nur natürlich, dass der Unterschied auch in den Quantilen deutlich zu sehen ist.

Nutzt man nun das 99%-Quantil für beide Gewichtungen (mit $\epsilon = 10$ und $MinPts = 2$) als Cutoff, so erhält man eine Auswahl an Tracks, welche für die Gleichgewichtung in Abbildung 5.5 bzw. für die Gewichte $w_1 = w_2 = 0.4, w_3 = 0.2$ in Abbildung 19 in Anhang C abgebildet ist. Für die DBSCAN Parameter nach Masip et al. sind die Ergebnisse nach Nutzung der Quantil-Cutoffs für beide Gewichtungen in den Abbildungen 20 bis 23 in Anhang C zu finden. Auch hier kann man durch die Wahl des Cutoffs eine Ausdünnung der Daten erkennen. Durch die Auswahl der Tracks mittels des Cutoffs, können nun interessante Regionen oder aber auch Cluster besser identifiziert werden. Um jedoch eine

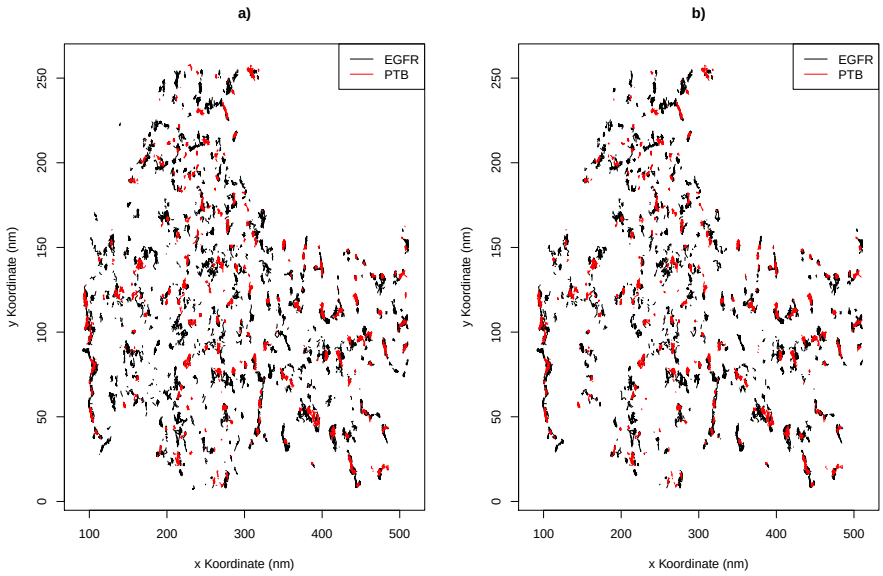

Abbildung 5.5: Übersicht der Trackauswahl bei einer Gewichtung von $w_1 = w_2 = w_3 = 1/3$ ($\epsilon = 10$ und $MinPts = 2$) und dem 99%-Quantil als Cutoff: a) Darstellung des Datensatzes mit allen enthaltenden Tracks; b) Darstellung der Tracks, welche zu einem Trackpaar mit einem errechneten Zusammenhang $\geq$ dem 99%-Quantil gehören.

noch stärkere Auswahl treffen zu können, wird im nächsten Schritt nun noch eine weitere Wahl des Cutoffs betrachtet.

Um eine noch detailliertere Auswahl an Trackpaarungen treffen zu können, sollte der Cutoff nicht zu viele, aber auch nicht zu wenige Trackpaarungen ausschließen. Wie in Tabelle 5.7 zu sehen, ist die Differenz der Maxima und des Mittelwertes verhältnismäßig hoch. Daher lässt sich vermuten, dass viele Trackpaarungen einen kleinen Wert für den Zusammenhang aufweisen und nur wenige einen hohen Zusammenhang. Dies wird in Abbildung 18 in Anhang C nochmals verdeutlicht. Es ist gut zu erkennen, dass die meisten Zusammenhänge Werte ≤ 0.4 annehmen. Um nun einen geeigneten Cutoff wählen zu können, wurde Tabelle 5.8 herangezogen. Aus dieser Tabelle kann man gut erkennen, dass für die Gewichtung mit $w_1 = w_2 = w_3 = 1/3$ ein Cutoff von 0.65 oder 0.7 eine gute Wahl wäre. Für eine Gewichtung mit $w_1 = w_2 = 0.4$ und $w_3 = 0.2$ wären hingegen die Cutoffs 0.75 und 0.8 eine gute Wahl.

Mit diesen Cutoffs wurde nun ebenfalls eine Auswahl der Tracks vorgenommen. Somit

	Gewichtung	
Cutoff	$w_1 = w_2 = w_3 = 1/3$	$w_1 = w_2 = 0.4, w_3 = 0.2$
0.55	2119 (0.7744)	2422 (0.8852)
0.6	1970 (0.7200)	2173 (0.7942)
0.65	1844 (0.6739)	2090 (0.7638)
0.7	166 (0.0607)	1971 (0.7203)
0.75	32 (0.0117)	1872 (0.6842)
0.8	10 (0.0037)	973 (0.3424)
0.85	1 (0.0004)	28 (0.0102)

Tabelle 5.8: Übersicht der Anzahl (Anteil, in %) an Werten des Zusammenhangsmaß größer oder gleich dem Cutoff in Abhängigkeit der Gewichtung.

würden bei Gewichtung $w_1 = w_2 = w_3 = 1/3$ insgesamt noch 1844 bzw. 166 Trackpaare weiter betrachtet, bei einer Gewichtung mit $w_1 = w_2 = 0.4, w_3 = 0.2$ würden 1872 bzw. 973 Trackpaare weiter betrachtet. Dabei sei darauf hingewiesen, dass diese Anzahlen Trackpaarungen beschreiben.

Das Ergebnis der Aussortierung, basierend auf den entsprechenden Cutoffs für die zwei Gewichtungen, sind in den Abbildungen 5.6 sowie 24, 25 und 26 im Anhang B zu finden. Es ist deutlich zu erkennen, dass für beide Cutoffs und auch beide Gewichtungen nun weniger Tracks abgebildet werden und somit die Auswahl über diese Cutoffs weniger Tracks zulassen. Dadurch können (kleinere) Cluster deutlicher und besser identifiziert werden. Durch das Zusammenhangsmaß werden somit räumlich abhängige Tracks gefunden und können mit Hilfe eines Cutoffs aussortiert werden. Weiter können durch dieses Vorgehen auch Cluster bzw. interessante Regionen gefunden werden. Die resultierende Menge an Trackpaarungen ist nach der Auswahl mittels Cutoff um so kleiner, je größer der Cutoff ist. Dies wird auch nochmals in Abbildung 26 in Anhang B deutlich. Hier ist der Cutoff mit 0.80 am größten und die Anzahl an Trackpaaren mit 973 deutlich geringer als die Gesamtanzahl aller möglichen Trackpaarungen. Durch dieses Auswahlverfahren sind jedoch auch nur noch wenige Cluster an Tracks vorhanden, welche dadurch aber auch einen hohen Zusammenhang aufweisen.

Das Zusammenhangsmaß mit der Gleichgewichtung ($w_1 = w_2 = w_3 = 1/3$) sowie der Ge-

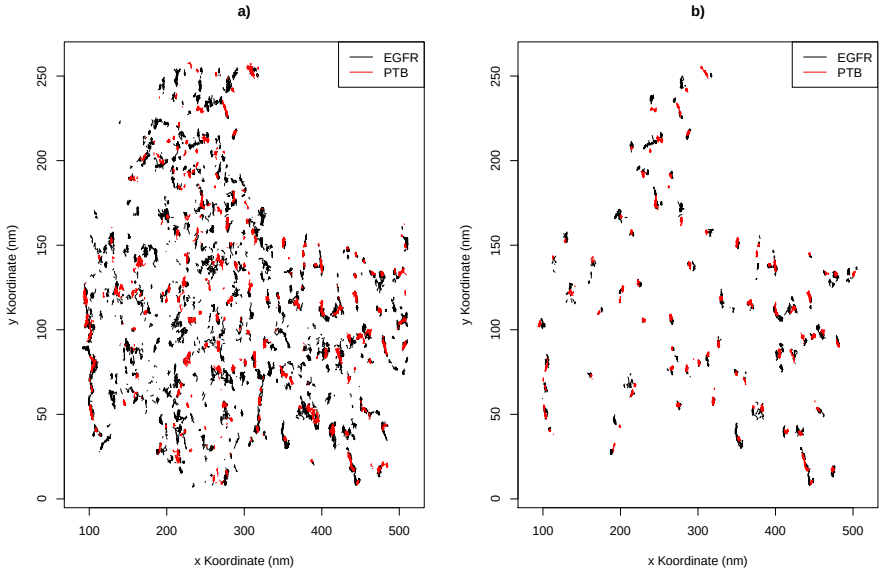

Abbildung 5.6: Übersicht der Trackauswahl bei einer Gewichtung von $w_1 = w_2 = w_3 = 1/3$ und einem Cutoff von 0.70: a) Darstellung des Datensatzes mit allen enthaltenden Tracks; b) Darstellung der Tracks, welche zu einem Trackpaar mit einem Zusammenhang ≥ 0.70 gehören.

wichtung mit $w_1 = w_2 = 0.4$ und $w_3 = 0.2$ kann nun auch auf weitere Daten angewendet werden, welche jeweils an einer Zelle 2, 5 und 10 Minuten nach Stimulation erhoben wurden. Da in Masip et al. (2016, [29]) beschrieben wird, dass die Proteine mit Zunahme der Zeit nach Stimulation zu Clustern mit größerem Durchmesser tendieren, wurden auch für diese 3 Datensätze sowohl die DBSCAN Parameter $\epsilon = 10$ und $MinPts = 2$ sowie $\epsilon = 40$ und $MinPts = 5$ (wie in Masip et al. vorgeschlagen) genutzt. Tabelle 8 in Anhang B enthält einen Überblick über die wichtigsten Kennzahlen der Werte des Zusammenhangsmaß für die obigen Datensätze in Abhängigkeit der jeweiligen Einstellungen. Es ist gut zu erkennen, dass (mit zunehmendem zeitlichen Abstand zur Stimulation) sich die Werte des Minimum, des Maximum und der Varianz der Werte des Zusammenhangsmaß für $\epsilon = 10, MinPts = 2$ und $\epsilon = 40, MinPts = 5$ „annähern“, d.h. sich nur noch geringfügig unterscheiden, wohingegen Mittelwert und Median weiterhin deutlich variieren. Daher ist hier weiterhin die Parameterwahl für den DBSCAN Algorithmus mit $\epsilon = 10, MinPts = 2$ zu bevorzugen.

Auch für diese drei Datensätze ist es nun interessant über den Wert des Zusammenhangs-

maß eine Auswahl der Tracks zu definieren. Dies wird erneut mit Hilfe der 95%- und 99%-Quantile sowie über einen frei gewählten Cutoff erfolgen. Letztere wird in diesem Fall für alle drei Zeitpunkte auf 0.75 festgelegt, aufgrund der generell höher ausfallenden Werte des Zusammenhangsmaß nach Stimulation (vgl. z.B. Maximum für 0 Minuten nach Stimulation und 2 Minuten nach Stimulation). Die entsprechenden Grafiken sind in Abbildung 5.7 sowie Abbildungen 27 bis 43 in Anhang C zu finden. Auch hier ist erneut

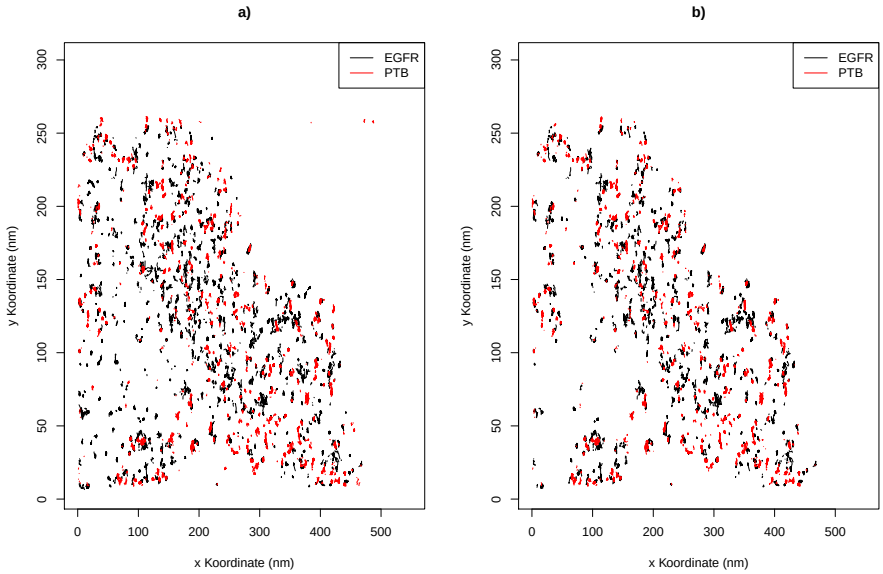

Abbildung 5.7: Übersicht der Trackauswahl 10 Minuten nach Stimulation bei einer Gewichtung von $w_1 = w_2 = 0.4$ und $w_3 = 0.2$ und einem Cutoff von 0.75: a) Darstellung des Datensatzes mit allen enthaltenden Tracks; b) Darstellung der Tracks, welche zu einem Trackpaar mit einem Wert des Zusammenhangsmaß ≥ 0.75 gehören.

zu sehen, dass die Auswahl mit Hilfe eines Cutoffs die Anzahl an Tracks reduziert wurde, wodurch interessante Regionen oder auch Cluster gefunden und identifiziert werden können. Wie bereits für den Datensatz 0 Minuten nach Stimulation verbleiben durch die Wahl des Cutoffs mittels eines Quantils mehr Tracks als durch einen (beliebig gewählten) festen Cutoff. Dabei ist zu beachten, dass je später die Datenerhebung nach Stimulation der Zelle durchgeführt wird, desto höher sollte der Cutoff gewählt werden. Dies wird an Abbildung 5.7 deutlich. Hier wurde der Cutoff größer als das 99%-Quantil gewählt, jedoch ist die Differenz zwischen dem 99%-Quantil und 0.75 gering, sodass die Auswahl der

Trackpaare dennoch recht hoch ist.
Abschließend kann trotzdem festgehalten werden, dass durch einen festen Cutoff die Wahl der Tracks eingegrenzt wird und interessante Regionen bzw. Cluster so leichter identifiziert werden können.

5.3.3. Diskussion der Ergebnisse

In diesem Kapitel wurde das Zusammenhangsmaß aus Kapitel 5.2.2 sowohl auf ein simuliertes Beispiel, als auch auf experimentelle Daten angewendet. Das Maß besteht dabei aus 3 Summanden, welche durch entsprechende Gewichte versehen werden. Dabei fließt zum einen die räumliche Struktur durch eine Funktion der Distanz, zum anderen die Clusterzugehörigkeit sowie der „gemeinsame Weg“ ein, welcher über die Anzahl an Frames gemessen wird, in welchen beide Einzeltracks simultan gemessen wurden.
Mit Hilfe des simulierten Beispiels konnte das Zusammenhangsmaß zunächst evaluiert und validiert werden. Es konnte gezeigt werden, dass das Maß die Beziehungen der Tracks gut repräsentiert und Trackpaare mit einem starken Zusammenhang durch einen hohen Wert des Maßes repräsentiert werden. Weiter konnten durch die Wahl verschiedener Gewichtungen gezeigt werden, dass es (ohne a priori Wissen) von Vorteil ist, alle drei Gewichte ungleich Null zu wählen. Durch diese Wahl werden sowohl die räumliche Struktur, als auch die Anzahl gemeinsam gemessener Frames, d.h. die Länge des gemeinsamen Weges, beachtet. Weiter wird durch eine derartige Gewichtung einem hohen Wert für einen Zusammenhang bei Trackpaaren vorgebeugt, welche nur in wenigen Frames zusammen gemessen wurden, in dieser kurzen Zeit aber dennoch räumlich nahe verlaufen. Ein Beispiel für solch ein Trackpaar wäre ein langer Track und ein sehr kurzer Track, welcher nur sehr kurz neben dem anderen Track aufleuchtet, z.B. für zwei Frames. Ein weiteres Beispiel wäre ein Trackpaar, bei dem sich die Trackpaare lediglich am Ende des einen bzw. am Anfang des anderen Tracks überschneiden, d.h. einer der Track endet, wobei in räumlicher Näher der andere Track startet. Somit ist es ratsam hier $w_3 \neq 0$ zu wählen.
Mit derartigen Gewichten wurde anschließend ein experimenteller Datensatz analysiert. Auch hier hat das Zusammenhangsmaß sinnvolle Ergebnisse geliefert mit den gewählten Gewichtungen. Durch die Wahl eines geeigneten Cutoffs für die Werte des Zusammen-

hangsmaßes konnte anschließend eine Auswahl an interessanten Trackpaaren erfolgen. Durch diesen Schritt konnte neben einer deutlichen Reduzierung der Anzahl an Tracks auch eine mögliche Identifizierung von Clustern bzw. ROIs ermöglicht werden.
Somit konnte gezeigt werden, dass das Zusammenhangsmaß sinnvolle Ergebnisse, sowohl für ein simuliertes Beispiel als auch für experimentelle Daten, liefert. Dabei müssen jedoch zwei Parametereinstellungen beachtet werden:

1. die Wahl der Gewichte w_1, w_2, w_3 und
2. die Wahl der Clustermethode mit zugehörigen Parametern.

Durch diese Einstellungen können sowohl der räumliche, als auch der zeitliche Zusammenhang unterschiedlich gewichtet werden. Weiter kann über die Clustermethode gesteuert werden, wann zwei Proteine demselben Cluster zugeordnet werden und somit räumlich nah sind, sodass ein Zusammenhang zu vermuten ist.
Durch die Wahl der Gewichte w_1, w_2 und w_3 wird weiter reguliert, wie viel Gewicht dem räumlichen bzw. dem zeitlichen Zusammenhang zugeteilt werden soll. Dadurch kann der Fokus des Zusammenhangsmaßes an die jeweilige Fragestellung angepasst werden. Es sei jedoch erneut darauf hingewiesen, dass $w_3 \neq 0$ gewählt werden sollte.

6. Zusammenfassung und Ausblick

In dieser Arbeit wurden Proteindaten zunächst zeitlich, anschließend räumlich und abschließend räumlich-zeitlich analysiert. In der zeitlichen Analyse war das Ziel eine Segmentierung an Proteinzeitreihen vorzunehmen. In der räumlichen Analyse sollten anschließend Clustereigenschaften der Proteine in der Zellmembran untersucht werden. Abschließend war das Ziel der räumlich-zeitlichen Analyse Trackpaarungen zu finden, welche einen hohen Zusammenhang aufweisen. Die Proteindaten wurden dabei am Max-Planck Institut für molekulare Physiologie Dortmund in der Arbeitsgruppe um Dr. Peter J. Verveer mit Hilfe von Fluoreszenzmikroskopie erhoben.

In der zeitlichen Analyse wurde zur Segmentierung ein Bayessches hierarchisches Modell zunächst für ChIP-Seq-Daten entwickelt, welches mittels MCMC-Methoden implementiert wurde. Das Modell dient der Segmentierung, wobei es nicht - wie die meisten Segmentierungsmethoden - log-Ratios benötigt, sondern statt dessen, die beobachteten Count-Daten. Weiter benötigt dieses Modell die Anzahl an Segmenten. Bei einer geeigneten Wahl dieser Anzahl lieferte das Modell für ChIP-Seq-Daten gute Ergebnisse und eine sinnvolle Segmentierung.

Anschließend wurde dieses Modell an Proteinzeitreihen angewendet, da auch hier die Segmentierung Analyseziel war. Hier musste zum einen ebenfalls die Anzahl der Segmente angegeben werden und zum anderen auch die Background Counts, da die Proteinzeitreihen lediglich aus der Lichtintensität eines Spottes zu den Zeitpunkten besteht. Letzteres Problem konnte schnell und einfach gelöst werden, indem als Background Counts der Backgroundwert der Proteinzeitreihen zuzüglich einem Fehler genutzt wurde. Mit diesen Informationen konnte das Bayessche hierarchische Modell auf die Proteindaten angewendet werden und lieferte auch hier gute Ergebnisse und sinnvolle Segmentierungen.

Da jedoch stets die Anzahl der Segmente bekannt sein musste, um das Bayessche hierarchische Modell anzuwenden, wurde abschließend in der zeitlichen Analyse ein Reversible Jump Schritt eingeführt. Dadurch musste die Segmentanzahl nun nicht mehr als fest angesehen werden. Die Häufigkeit eines Reversible Jump Schritts, konnte dabei mit Hilfe entsprechender Gewichte gesteuert werden. Weiter war die Berechnung der resultierenden Segmentierung, basierend auf geeigneten Cutoffs, flexibler, da durch unterschiedliche Cu-

toffs unterschiedlich feine Segmentierungen resultierten. Mit Hilfe des Reversible Jump Schritts lieferte das Modell weiterhin, sowohl für ChIP-Seq-Daten als auch für Proteindaten, gute Ergebnisse und war nun deutlich flexibler.
Für die räumliche Analyse der Proteindaten wurden verschiedene Clustermethoden verglichen, um die Anzahl an Proteinen in Clustern schätzen zu können. Die Methoden wurden dabei zunächst auf einer Single Colour Simulation, d.h. einer Simulation mit nur einem Protein, angewendet. In dieser Simulationsstudie wurde gezeigt, dass die Methoden bei geeigneter Parameterwahl gute Ergebnisse lieferten und man mit Hilfe eines Ablaufschemas Vor- und Nachteile der einzelnen Methoden besser nutzen bzw. ausgleichen konnte. Dieses Schema ist in Abbildung 4.9 abgebildet und wurde anschließend sowohl auf experimentelle Single Colour als auch auf simulierte Dual Colour Daten angewendet. Es konnte gezeigt werden, dass durch die vorgestellte Vorgehensweise nicht nur Zeit eingespart werden kann, sondern auch die Parameterwahl für einige Methoden vereinfacht wurde, da nun Vorwissen aus Methoden ohne nötige Parametereinstellung gewonnen werden konnte.
Auch bei der Anwendung der Methoden und des Schemas auf eine Dual Colour Simulationsstudie konnte gezeigt werden, dass die Vorgehensweise gute Ergebnisse liefert.
Abschließend wurde in der räumlich-zeitlichen Analyse (von Dual Colour Daten) ein Zusammenhangsmaß für zwei Proteintracks unterschiedlicher Proteine entwickelt. Dafür wurde zunächst differenziert, welche Eigenschaften zweier Proteintracks einen Zusammenhang repräsentieren. Basierend auf diesen Überlegungen wurde das Zusammenhangsmaß zusammengesetzt, wobei es aus drei Summanden mit zugehörigem Gewicht besteht. Diese beinhalten die Distanz, die Clusterzugehörigkeit pro Frame bzw. Zeitpunkt und die Anzahl an Frames, in welchen beide Proteintracks zusammen gemessen wurden.
In einem ersten simulierten Beispiel wurde das Zusammenhangsmaß zunächst evaluiert, wobei u.a. die Wahl der Gewichte thematisiert wurde. Es zeigte sich, dass eine Gewichtung mit $w_3 = 0$, sodass die Anzahl an Frames, in denen beide Proteintracks gemessen wurden, keinen Einfluss hat, nicht zu bevorzugen ist. Zwei weitere Gewichtungen lieferten hingegen die gewünschten Ergebnisse und repräsentierten die vorhandene Abhängigkeitsstruktur in diesem Beispiel.
Anschließend wurde das Zusammenhangsmaß mit den Gewichten, welche sich zuvor als sinnvoll erwiesen hatten, auf experimentelle Daten angewendet. Auch hier lieferte das

Zusammenhangsmaß gute Ergebnisse. Weiter konnte durch die Wahl eines Cutoffs für die Werte des Zusammenhangsmaß eine Auswahl an Trackpaaren erfolgen. Durch diese Auswahl wird nicht nur die Datenmenge reduziert, es können auch interessante Regionen bzw. Cluster identifiziert werden.

Dabei sei stets darauf hingewiesen, dass in allen drei Analysen Parameter oder Startwerte gewählt werden müssen:
In der zeitlichen Analyse muss neben einer Startsegmentierung und (Start-)Segmentanzahl auch die Iterationenanzahl, der Burn-In sowie die Gewichtung des Reversible Jump spezifiziert werden. Durch die Einstellung eben dieser Parameter können sich die Berechnungen, je nach Einstellung, unterscheiden. Die Konvergenz der Markov Kette kann jedoch mit Hilfe entsprechender Prüfmethoden gewährleistet werden. Die resultierende Segmentierung hingegen, kann basierend auf der Wahl des Cutoffs weiterhin flexibel sein.
In der räumlichen Analyse können die Parametereinstellungen durch das vorgeschlagene Schema vereinfacht werden (vgl. Abbildung 4.9 in Kapitel 4.4.1). Dabei sollte mit einer Wahl von interessanten Regionen, den ROIs, begonnen werden. Diese können auf Basis von Expertenwissen oder aber mit Hilfe von ASH identifiziert werden. Bei der Wahl eben dieser Regionen muss zum einen die Größe der ROIs als auch die Anzahl beachtet werden, wodurch die Rechenzeit und auch die Güte der Schätzungen abhängig sein kann. Da auf den ROIs Vorwissen, z.B. mit Hilfe der Gammics Methode, berechnet wird, sollten die ROIs die Daten bzw. verschiedene Situationen gut repräsentieren. Ist dies nicht der Fall und die ROIs sind z.B. zu klein gewählt worden, so kann das Ergebnis, welches als Vorwissen für weitere Methoden genutzt wird, verfälscht sein.
Bei der räumlich-zeitlichen Analyse ist die Wahl der Gewichte der einzelnen Summanden im Zusammenhangsmaß sowie die Wahl der Clustermethode entscheidend. Diese Spezifikationen können einzelnen Eigenschaften mehr oder weniger Gewicht zuteilen, aber auch das Ergebnis verfälschen. Wird bei der Clustermethode z.B. der zugehörige Parameter so eingestellt, dass stets sehr große Cluster gebildet werden, so kann aus dem entsprechenden Summanden des Zusammenhangsmaß nur wenig zusätzliche Information gewonnen werden.

Anschließend an die hier durchgeführten Analysen, sind noch weitere Untersuchungen denkbar. So kann im Rahmen der zeitlichen Analyse, neben einer effizienteren Implementierung des Modells, der Reversible Jump weiter verfeinert werden. Hier kann z.B. die Nutzung einer anderen Verteilung innerhalb des Merge- oder des Split-Schrittes in Betracht gezogen werden. Durch eine entsprechende Verteilung, kann im Gegensatz zur Gleichverteilung der Split- bzw. der Merge-Schritt auf bestimmte Regionen fokussiert werden. So kann der Split-Schritt bei bestimmten Datensätzen in der Mitte der Beobachtungen sinnvoller sein, als an den Rändern. Der Merge-Schritt hingegen kann an den Rändern sinnvollere Ergebnisse liefern als in der Mitte der Beobachtungen.
In der räumlichen Analyse kann die Untersuchung dahingehend erweitert werden, dass nun auch andere Clustereigenschaften betrachtet werden können. So kann besonders im Dual Colour Fall von Interesse sein, wo genau die Cluster der zwei Proteine im Raum liegen. Dadurch würden Rückschlüsse über einen möglichen Zusammenhang der Cluster der zwei Proteine möglich gemacht werden. Dies ist, gerade aus biologischer Sicht, von Interesse.
Im Falle einer räumlich-zeitlichen Analyse von Proteintracks ist eine weitere Untersuchung und mögliche Erweiterung des Zusammenhangsmaßes denkbar. Eine Erweiterung könnte dabei ein statistischer Signifikanztest sein (z.B. ein Permutationstest), aber auch eine weitere Untersuchung der möglichen Gewichtungen bzw. Parameter. Weiter ist die Anwendung des Zusammenhangsmaßes auch auf anderen Daten möglich. So ist die Durchführung von Tracking Methoden aus dem Bereich des Animal Tracking denkbar. Eine erste Adaption dieser Methoden wurde bereits in einer Bachelor Arbeit durchgeführt, wobei sich jedoch zunächst auf zwei exemplarische Tracks beschränkt wurde. Dies sollte erweitert werden.
Weiter ist eine Modellierung der Bewegung der Proteine denkbar, wobei sowohl die Abhängigkeit einzelner Tracks, eine abrupte Änderung des Weges, wie auch ein Vereinigen bzw. das Trennen von Tracks beachtet werden sollte. Dazu könnte sich an Arbeiten zum Animal Tracking, aber auch über Zellbewegungen orientiert werden.

Literatur

[1] Arnau, Vicente, Sergio Mars und Ignacio Marín: *Iterative Cluster Analysis of Protein Interaction Data.* Bioinformatics, 21(3):364–378, 2005.

[2] Baddeley, Adrian und Rolf Turner: *spatstat: An R Package for Analyzing Spatial Point Patterns.* Journal of Statistical Software, 12(1):1–42, 2005.

[3] Barry, Daniel und J. A. Hartigan: *A Bayesian Analysis for Change Point Problems.* Journal of the American Statistical Association, 88(421):309–319, 1993.

[4] Bivand, Roger S., Edzer J. Pebesma und Virgilio Gómez-Rubio: *Applied Spatial Data Analysis with R.* Use R! Springer New York, 2. Auflage, 2013.

[5] Campbell, Neil A. und Jane B. Reece: *Biologie.* 6. Auflage. Pearson Verlag, 2006.

[6] Dunham, M. H., Yu Meng und Jie Huang: *Extensible Markov model.* In: *Data Mining, 2004. ICDM '04. Fourth IEEE International Conference on*, Seiten 371–374, Nov 2004.

[7] Eilers, Paul H. C. und Renée X. de Menezes: *Quantile smoothing of array CGH data.* Bioinformatics, 21(7):1146–1153, 2005.

[8] Eisen, Michael B., Paul T. Spellman, Patrick O. Brown und David Botstein: *Cluster analysis and display of genome-wide expression patterns.* Proceedings of the National Academy of Sciences, 95(25):14863–14868, 1998.

[9] Ester, Martin, Hans-Peter Kriegel, Jörg Sander und Xiaowei Xu: *A density-based algorithm for discovering clusters in large spatial databases with noise.* Seiten 226–231. AAAI Press, 1996.

[10] Fryzlewicz, P. und S. Subba Rao: *Multiple-change-point detection for auto-regressive conditional heteroscedastic processes.* Journal of the Royal Statistical Society: Series B (Statistical Methodology), 76(5):903–924, 2014.

[11] Gelfand, A.E., P. Diggle, P. Guttorp und M. Fuentes: *Handbook of Spatial Statistics*. Chapman & Hall/CRC Handbooks of Modern Statistical Methods. CRC Press, 2010.

[12] Geweke, John: *Evaluating the Accuracy of Sampling-Based Approaches to the Calculation of Posterior Moments*. In: *In Bayesian Statistics*, Seiten 169–193. University Press, 1992.

[13] Green, Peter J.: *Reversible jump Markov chain Monte Carlo computation and Bayesian model determination*. Biometrika, 82(4):711–732, 1995.

[14] Gruenbaum, Yosef, Ayelet Margalit, Robert D. Goldman, Dale K. Shumaker und Katherine L. Wilson: *The nuclear lamina comes of age*. Nature Reviews Molecular Cell Biology, (1):21–31, 2005.

[15] Guelen, Lars, Ludo Pagie, Emilie Brasset, Wouter Meuleman, Marius B. Faza, Wendy Talhout, Bert H. Eussen, Annelies de Klein, Lodewyk Wessels und Wouter de Laat: *Domain organization of human chromosomes revealed by mapping of nuclear lamina interactions*. Nature, (7197):948–951, 2008.

[16] Gurarie, Eliezer: *bcpa: Behavioral change point analysis of animal movement*, 2014. R package version 1.1.

[17] Hartigan, John A.: *Clustering Algorithms*. John Wiley & Sons, Inc., New York, NY, USA, 1975.

[18] Heidelberger, Philip und Peter D. Welch: *Simulation Run Length Control in the Presence of an Initial Transient*. Operations Research, 31(6):1109–1144, 1983.

[19] Herrmann, Sabrina, Holger Schwender, Katja Ickstadt und Peter Müller: *A Bayesian changepoint analysis of ChIP-Seq data of Lamin B*. Biochimica et Biophysica Acta (BBA) - Proteins and Proteomics, 1844(1, Part A):138 – 144, 2014.

[20] Hupé, Philippe, Nicolas Stransky, Jean-Paul Thiery, François Radvanyi und Emmanuel Barillot: *Analysis of array CGH data: from signal ratio to gain and loss of DNA regions*. Bioinformatics, 20(18):3413–3422, 2004.

[21] Ibach, Jenny, Yvonne Radon, Márton Gelléri, Michael H. Sonntag, Luc Brunsveld, Philippe I. H. Bastiaens und Peter J. Verveer: *Single Particle Tracking Reveals that EGFR Signaling Activity Is Amplified in Clathrin-Coated Pits.* PLoS ONE, 10(11):1–22, 2015.

[22] Jaqaman, Khuloud, Dinah Loerke, Marcel Mettlen, Hirotaka Kuwata, Sergio Grinstein, Sandra L. Schmid und Gaudenz Danuser: *Robust single-particle tracking in live-cell time-lapse sequences.* Nature Methods, 5:695 – 702, 2008.

[23] Johnson, Stephen C.: *Hierarchical clustering schemes.* Psychometrika, 32(3):241–254, 1967.

[24] Kanungo, T., D. M. Mount, N. S. Netanyahu, C. D. Piatko, R. Silverman und A. Y. Wu: *An efficient k-means clustering algorithm: analysis and implementation.* IEEE Transactions on Pattern Analysis and Machine Intelligence, 24(7):881–892, Jul 2002.

[25] Kaufman, Leonard und Peter J. Rousseeuw: *Finding roups in data: An introduction to cluster analysis*, Kapitel Partitioning Around Medoids (Program PAM), Seiten 68–125. John Wiley & Sons, Inc., 2008.

[26] Kranstauber, Bart, Roland Kays, Scott D. LaPoint, Martin Wikelski und Kamran Safi: *A dynamic Brownian bridge movement model to estimate utilization distributions for heterogeneous animal movement.* Journal of Animal Ecology, 81(4):738–746, 2012.

[27] Kranstauber, Bart und Marco Smolla: *move: Visualizing and Analyzing Animal Track Data*, 2016. R package version 2.0.0.

[28] Lai, Weil R., Mark D. Johnson, Raju Kucherlapati und Peter J. Park: *Comparative analysis of algorithms for identifying amplifications and deletions in array CGH data.* Bioinformatics, 21(19):3763–3770, 2005.

[29] Masip, Martin E., Jan Huebinger, Jens Christmann, Ola Sabet, Frank Wehner, Antonios Konitsiotis, Günther R. Fuhr und Philippe I. H. Bas-

TIAENS: *Reversible cryo-arrest for imaging molecules in living cells at high spatial resolution.* Nature Methods, 13:665–672, 2016.

[30] MATÉRN, BERTIL: *Spatial Variation.* Springer, Berlin, 2. Auflage Auflage, 1986.

[31] MATHEW, LISHA A., PAUL R. STAAB, LAURA E. ROSE und DIRK METZLER: *Why to account for finite sites in population genetic studies and how to do this with Jaatha 2.0.* Ecology and Evolution, 3(11):3647–3662, 2013.

[32] MATLAB: *version 7.10.0 (R2010a).* The MathWorks Inc., Natick, Massachusetts, 2010.

[33] MATTESON, DAVID S. und NICHOLAS A. JAMES: *A Nonparametric Approach for Multiple Change Point Analysis of Multivariate Data.* Journal of the American Statistical Association, 109(505):334–345, 2014.

[34] MESSINA, TROY C., HIYUN KIM, JASON T. GIURLEO, und DAVID S. TALAGA: *Hidden Markov Model Analysis of Multichromophore Photobleaching.* The Journal of Physical Chemistry B, 110(33):16366–16376, 2006. PMID: 16913765.

[35] MORAN, P. A. P.: *Notes on continuous stochastic phenomena.* Biometrika, 37(1-2):17–23, 1950.

[36] OLSHEN, A.B. und E.S. VENKATRAMAN: *Circular binary segmentation for the analysis of array-based DNA copy number data.* Biostatistics, 5(4):557–572, 2004.

[37] PARKER, PETER J.: *ChIP-Seq: Advantages and Challenges of a Maturing Technology.* Nature reviews. Genetics, 10(10):669–680, 2009.

[38] R CORE TEAM: *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria, 2015.

[39] REENTS, REINHARD, MELANIE WAGNER, JÜRGEN KUHLMANN und HERBERT WALDMANN: *Synthese und Anwendung fluoreszenzmarkierter Ras-Proteine in der Bildgebung lebender Zellen.* Angewandte Chemie, 116(20):2765–2768, 2004.

[40] RIPLEY, B. D.: *Modelling Spatial Patterns.* Journal of the Royal Statistical Society. Series B (Methodological), 39(2):172–212, 1977.

[41] Robert, Christian P. und George Casella: *Monte Carlo statistical methods.* Springer Texts in Statistics. Springer-Verlag, New York, Second Auflage, 2004.

[42] Robert, Christian P. und Judith Rousseau: *A Mixture Approach to Bayesian Goodness of Fit.* Technical Report 02009, Cahiers du CEREMADE, Université Paris, Dauphine, 2002.

[43] Roberts, G. O., A. Gelman und W. R. Gilks: *Weak convergence and optimal scaling of random walk Metropolis algorithms.* Annals of Applied Probability, 7:110–120, 1997.

[44] Schäfer, Martin, Yvonne Radon, Thomas Klein, Sabrina Herrmann, Holger Schwender, Peter J. Verveer und Katja Ickstadt: *A Bayesian mixture model to quantify parameters of spatial clustering.* Computational Statistics & Data Analysis, 92:163 – 176, 2015.

[45] Schäffler, Arne und Nicole Menche: *Biologie, Anatomie, Physiologie : kompaktes Lehrbuch für die Pflegeberufe.* Nummer 4. überarbeitete Auflage. Urban & Fischer, München (u.a.), 2000.

[46] Scott, David W. und Stephan R. Sain: *Multidimensional Density Estimation.* In: C.R. Rao, E.J. Wegman und J.L. Solka (Herausgeber): *Data Mining and Data Visualization*, Band 24 der Reihe *Handbook of Statistics*, Seiten 229 – 261. Elsevier, 2005.

[47] Seshan, Venkatraman E. und Adam Olshen: *DNAcopy: DNA copy number data analysis.* R package version 1.38.0.

[48] Siegmund, David: *Boundary Crossing Probabilities and Statistical Applications.* Ann. Statist., 14(2):361–404, 06 1986.

[49] Sunaga, D. Y., J. C. Nievola und M. P. Ramos: *Statistical and Biological Validation Methods in Cluster Analysis of Gene Expression.* In: *Machine Learning and Applications, 2007. ICMLA 2007. Sixth International Conference on*, Seiten 494–499, Dec 2007.

[50] Syrjala, Stephen E.: *A Statistical Test for a Difference between the Spatial Distributions of Two Populations.* Ecology, 77(1):75–80, 1996.

[51] Wang, Pei, Young Kim, Jonathan Pollack, Balasubramanian Narasimhan und Robert Tibshirani: *A method for calling gains and losses in array CGH data.* Biostatistics, 6(1):45–58, 2005.

A. Weiterführende Rechnungen

A.1. Umformung der Parameter einer Beta-Verteilung

Für eine Beta-verteilte Zufallsvariable $X \sim \text{Beta}(\alpha, \beta)$ sind folgende Lage- und Streuungsmaße bekannt:

$$\begin{aligned} \text{E}(X) &= \mu = \frac{\alpha}{\alpha+\beta}, \\ \text{Var}(X) &= \sigma^2 = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}, \\ \text{VarK}(X) &= \rho = \frac{\sigma}{\mu} = \sqrt{\frac{\beta}{(\alpha+\beta+1)\alpha}}. \end{aligned}$$

Die Varianz sowie der Variationskoeffizient können nun mit Hilfe von μ und einer Hilfsvariable $\theta = \alpha + \beta$ (für eine einfachere Notation) umformuliert werden. Die Varianz ergibt sich dann zu:

$$\begin{aligned} \text{Var}(X) = \sigma^2 &= \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)} \\ &= \frac{\mu\theta \cdot \theta(1-\mu)}{\theta^2(\theta+1)} \\ &= \frac{\cancel{\theta^2}(\mu-\mu^2)}{\cancel{\theta^2}(\theta+1)} \\ &= \frac{\mu-\mu^2}{\theta+1}. \end{aligned}$$

Der quadrierte Variationskoeffizient errechnet sich daraus dann wie folgt:

$$\begin{aligned} \rho^2 = \frac{\sigma^2}{\mu^2} &= \frac{\mu-\mu^2}{\theta+1} \cdot \frac{1}{\mu^2} \\ &= \frac{\mu-\mu^2}{\mu^2(\theta+1)}. \end{aligned} \tag{18}$$

In der Reparametrisierung nach Robert und Rousseau (2002, [42]) wird als Parameter zum einen ein Lagemaß $\epsilon \in]0,1[$ und zum anderen ein Streuungsmaß $\delta > 0$ verwendet. Die Reparametrisierung sieht dann wie folgt aus:

$$\text{Be}(\epsilon\delta, \delta(1-\epsilon)).$$

Für das Lagemaß wird hier der Erwartungswert gewählt, d.h. $\epsilon = \mu$. Für das Streuungsmaß könnte der Variationskoeffizient genutzt werden. Da aber in Formel (18) noch θ und somit

α und β enthalten sind, wird diese Formel nach θ umgeformt und dieser Ausdruck gewählt. Somit ergibt sich zusammenfassend

$$\epsilon = \mu \quad \text{und} \quad \delta = \frac{\mu - \mu^2}{\rho^2 \mu^2} - 1.$$

Wenn dies nun in die obige Reparametrisierung eingesetzt wird, ergibt sich für die Shapeparameter der Beta-Verteilung α und β

$$\begin{aligned} \alpha &= \epsilon\delta = \mu(\frac{\mu - \mu^2}{\rho^2\mu^2} - 1) \\ &= \frac{1-\mu}{\rho^2} - \mu \\ &\quad \text{und} \\ \beta &= \delta(1-\mu) = (\frac{\mu - \mu^2}{\rho^2\mu^2} - 1)(1-\mu) \\ &= (\frac{\cancel{\mu}(1-\mu)}{\cancel{\mu}(\rho^2\mu)} - 1) - (1-\mu) \\ &= \frac{(1-\mu)^2}{\mu\rho^2} - (1-\mu). \end{aligned} \tag{19}$$

Da im Fall des hierarchischen Modells noch beachtet werden muss, ob das Segment eine erhöhte Lamin B-Konzentration besitzt und daher auch 2 Beta-Verteilungen mit unterschiedlichen Erwartungswert genutzt werden, müssen die Formeln aus (20) dahingehend unterschieden werden. Somit erhält man abschließend die Formeln

$$\alpha_l = \frac{1-\mu_l}{\rho^2} - \mu_l \quad \text{und} \quad \beta_l = \frac{(1-\mu_l)^2}{\mu_l\rho^2} - (1-\mu_l), \quad \text{mit } l \in \{0,1\}.$$

A.2. Integration der a posteriori Verteilung

Die a posteriori Verteilung ist proportional zum Produkt aus der Likelihood und der a priori Verteilung. Für das hierarchische Modell aus Kapitel 3.2.2 ergibt sie sich (durch die

Umrechnung der Shape-Parameter der Beta-Verteilung) zu

$$\begin{aligned} p(p, w_1, \mu_0, \mu_1, \rho, \pi, t|c, N) \propto & \prod_{k=1}^{K} \prod_{i=t_{k-1}}^{t_k-1} \binom{n_i}{c_i} p_k^{c_i} (1-p_k)^{n_i-c_i} \frac{1}{K-2} \frac{1}{t_{k+1}-t_{k-1}} \cdot \\ & \cdot \prod_{k=1}^{K} \left[\left(\frac{\Gamma(\alpha_0+\beta_0)}{\Gamma(\alpha_0)\Gamma(\beta_0)} p_k^{\alpha_0-1} (1-p_k)^{\beta_0-1} \right)^{1-w_k} \cdot \right. \\ & \cdot \left(\frac{\Gamma(\alpha_1+\beta_1)}{\Gamma(\alpha_1)\Gamma(\beta_1)} p_k^{\alpha_1-1} (1-p_k)^{\beta_1-1} \right)^{w_k} \cdot \\ & \left. \cdot \pi^{w_k} (1-\pi)^{1-w_k} \frac{1}{0.75} \frac{1}{0.9-0.25} \frac{1}{\Gamma(2)} \rho \exp(\rho) \frac{\Gamma(2)}{\Gamma(1)\Gamma(1)} \right]. \end{aligned}$$

An dieser Stelle kann man nun nach $w_k = 0$ bzw. $w_k = 1$ unterscheiden und die a posteriori Verteilung aufteilen. Daraus folgt, dass

$$p(p, w_1, \mu_0, \mu_1, \rho, \pi, t|c, N) \propto \prod_{k=1}^{K} \prod_{i=t_{k-1}}^{t_k-1} \binom{n_i}{c_i} p_k^{c_i} (1-p_k)^{n_i-c_i} \frac{1}{K-2} \frac{1}{t_{k+1}-t_{k-1}} \cdot \prod_{k=1}^{K} A_1,$$

wobei

$$\begin{aligned} A_1 = & \left[\left(\frac{\Gamma(\alpha_0+\beta_0)}{\Gamma(\alpha_0)\Gamma(\beta_0)} p_k^{\alpha_0-1} (1-p_k)^{\beta_0-1} \right)^{1-w_k} \left(\frac{\Gamma(\alpha_1+\beta_1)}{\Gamma(\alpha_1)\Gamma(\beta_1)} p_k^{\alpha_1-1} (1-p_k)^{\beta_1-1} \right)^{w_k} \pi^{w_k} \cdot \right. \\ & \left. (1-\pi)^{1-w_k} A_2 \right] \end{aligned}$$

und

$$A_2 = \frac{1}{0.75} \frac{1}{0.9-0.25} \frac{1}{\Gamma(2)} \rho \exp(\rho) \frac{\Gamma(2)}{\Gamma(1)\Gamma(1)} = \frac{1}{0.4875} \cdot \rho \cdot \exp(\rho) = 2.051282 \cdot \rho \cdot \exp(\rho).$$

Im Fall von A_1 kann man nun eine Fallunterscheidung bzgl. w_k machen und erhält folgende Schreibweise

$$A_1 = \begin{cases} \frac{\Gamma(\alpha_0+\beta_0)}{\Gamma(\alpha_0)\Gamma(\beta_0)} p_k^{\alpha_0-1} (1-p_k)^{\beta_0-1} \cdot (1-\pi) \cdot C & , w_k = 0 \\ \frac{\Gamma(\alpha_1+\beta_1)}{\Gamma(\alpha_1)\Gamma(\beta_1)} p_k^{\alpha_1-1} (1-p_k)^{\beta_1-1} \cdot \pi \cdot C & , w_k = 1 \end{cases}.$$

An dieser Stelle wird nun das Produkt über k weiter zusammengefasst und man bekommt folgendes Ergebnis:

$$
\begin{aligned}
p(p, w_1, \mu_0, \mu_1, \rho, \pi, t|c, N) \propto & \prod_{k=1}^{K} \prod_{i=t_{k-1}}^{t_k-1} \binom{n_i}{c_i} p_k^{c_i} (1-p_k)^{n_i-c_i} \frac{1}{K-2} \frac{1}{t_{k+1}-t_{k-1}} \cdot \\
& \cdot \prod_{k=1}^{K} \left[\left(\frac{\Gamma(\alpha_0+\beta_0)}{\Gamma(\alpha_0)\Gamma(\beta_0)} p_k^{\alpha_0-1} (1-p_k)^{\beta_0-1} \right)^{1-w_k} \cdot \right. \\
& \cdot \left(\frac{\Gamma(\alpha_1+\beta_1)}{\Gamma(\alpha_1)\Gamma(\beta_1)} p_k^{\alpha_1-1} (1-p_k)^{\beta_1-1} \right)^{w_k} \cdot \\
& \left. \cdot \pi^{w_k} (1-\pi)^{1-w_k} \underbrace{\frac{1}{0.75} \frac{1}{0.9-0.25} \frac{1}{\Gamma(2)} \rho \exp(\rho) \frac{\Gamma(2)}{\Gamma(1)\Gamma(1)}}_{\frac{1}{0.4875}\rho\exp(\rho)} \right] \\
= & \prod_{k=1}^{K} \prod_{i=t_{k-1}}^{t_k-1} \binom{n_i}{c_i} p_k^{c_i} (1-p_k)^{n_i-c_i} \frac{1}{K-2} \frac{1}{t_{k+1}-t_{k-1}} \cdot \\
& \cdot \left[\left(\frac{\Gamma(\alpha_0+\beta_0)}{\Gamma(\alpha_0)\Gamma(\beta_0)} p_k^{\alpha_0-1} (1-p_k)^{\beta_0-1} \right)^{1-w_k} \cdot \right. \\
& \left. \cdot \left(\frac{\Gamma(\alpha_1+\beta_1)}{\Gamma(\alpha_1)\Gamma(\beta_1)} p_k^{\alpha_1-1} (1-p_k)^{\beta_1-1} \right)^{w_k} \frac{1}{0.4875} \rho \exp(\rho) \right] \\
= & \prod_{k=1}^{K} \prod_{i=t_{k-1}}^{t_k-1} \begin{cases} \binom{n_i}{c_i} p_k^{c_i} (1-p_k)^{n_i-c_i} \frac{1}{K-3} \cdot \\ \frac{1}{t_{k+1}-t_{k-1}} \frac{\Gamma(\alpha_0+\beta_0)}{\Gamma(\alpha_0)\Gamma(\beta_0)} p_k^{\alpha_0-1} \cdot \\ (1-p_k)^{\beta_0-1} \cdot (1-\pi) \cdot A_2 & , w_k = 0 \\ \binom{n_i}{c_i} p_k^{c_i} (1-p_k)^{n_i-c_i} \frac{1}{K-3} \cdot \\ \frac{1}{t_{k+1}-t_{k-1}} \frac{\Gamma(\alpha_1+\beta_1)}{\Gamma(\alpha_1)\Gamma(\beta_1)} p_k^{\alpha_1-1} \cdot \\ (1-p_k)^{\beta_1-1} \cdot \pi \cdot A_2 & , w_k = 1 \end{cases} .
\end{aligned}
$$

Wenn man nun bezüglich p_k integriert ergibt sich

$$\begin{aligned}
\int p(p, w_1, \mu_0, \mu_1, \rho, \pi, t|c, N)dp_k \propto \int \prod_{k=1}^{K} \prod_{i=t_{k-1}}^{t_k-1} \binom{n_i}{c_i} p_k^{c_i}(1-p_k)^{n_i-c_i} \frac{1}{K-2} \frac{1}{t_{k+1}-t_{k-1}} \cdot \\
\cdot \prod_{k=1}^{K} \left[\left(\frac{\Gamma(\alpha_0+\beta_0)}{\Gamma(\alpha_0)\Gamma(\beta_0)} p_k^{\alpha_0-1}(1-p_k)^{\beta_0-1} \right)^{1-w_k} \cdot \right. \\
\cdot \left(\frac{\Gamma(\alpha_1+\beta_1)}{\Gamma(\alpha_1)\Gamma(\beta_1)} p_k^{\alpha_1-1}(1-p_k)^{\beta_1-1} \right)^{w_k} \cdot \\
\cdot \pi^{w_k}(1-\pi)^{1-w_k} \frac{1}{0.75} \frac{1}{0.9-0.25} \frac{1}{\Gamma(2)} \cdot \\
\left. \rho \exp(\rho) \frac{\Gamma(2)}{\Gamma(1)\Gamma(1)} \right] dp_k \\
= \prod_k \begin{cases}
\prod_i \binom{n_i}{c_i} \frac{1}{K-2} \frac{1}{t_{k+1}-t_{k-1}} \frac{\Gamma(\alpha_0+\beta_0)}{\Gamma(\alpha_0)\Gamma(\beta_0)} (1-\pi) \cdot \\
\frac{1}{0.4875} \rho \exp \rho \int p_k^{c_i}(1-p_k)^{n_i-c_i} p_k^{\alpha_0-1} \cdot \\
(1-p_k)^{\beta_0-1} dp_k & , w_k = 0 \\
\prod_i \binom{n_i}{c_i} \frac{1}{K-2} \frac{1}{t_{k+1}-t_{k-1}} \frac{\Gamma(\alpha_1+\beta_1)}{\Gamma(\alpha_1)\Gamma(\beta_1)} \pi \cdot \\
\frac{1}{0.4875} \rho \exp \rho \int p_k^{c_i}(1-p_k)^{n_i-c_i} p_k^{\alpha_1-1} \cdot \\
(1-p_k)^{\beta_1-1} dp_k & , w_k = 1
\end{cases} \\
= \prod \begin{cases}
\prod \binom{n_i}{c_i} \frac{1}{K-2} \frac{1}{t_{l+1}-t_{l-1}} \frac{\Gamma(\alpha_0+\beta_0)}{\Gamma(\alpha_0)\Gamma(\beta_0)} (1-\pi) \cdot \\
\frac{1}{0.4875} \rho \exp \rho \int p_k^{c_i+\alpha_0-1} \cdot \\
(1-p_k)^{n_i-c_i+\beta_0-1} dp_k & , w_k = 0 \\
\prod \binom{n_i}{c_i} \frac{1}{K-2} \frac{1}{t_{l+1}-t_{l-1}} \frac{\Gamma(\alpha_1+\beta_1)}{\Gamma(\alpha_1)\Gamma(\beta_1)} \pi \cdot \\
\frac{1}{0.4875} \rho \exp \rho \int p_k^{c_i+\alpha_1-1} \cdot \\
(1-p_k)^{n_i-c_i+\beta_1-1} dp_k & , w_k = 1
\end{cases} .
\end{aligned}$$

Das Integral in obiger Formel $\int p_k^{c_i+\alpha_0-1}(1-p_k)^{n_i-c_i+\beta_0-1}dp_k$ bzw. $\int p_k^{c_i+\alpha_1-1}(1-p_k)^{n_i-c_i+\beta_1-1}$ dp_k ist dabei die Dichte einer Beta-Verteilung mit Parametern $(c_i+\alpha_0, n_i-c_i+\beta_0)$ bzw. $(c_i+\alpha_1, n_i-c_i+\beta_1)$.

A.3. Restriktionen der Parameter der Beta-Verteilung

Durch die Wahl der Reparametrisierung der Beta-Verteilung nach Robert und Rousseau (2002, [42]) und der anschließenden Rückrechnung ergibt sich eine Restriktion für ρ. In der allgemein gebräuchlichen Beta-Verteilung $\text{Be}(\alpha, \beta)$ ist stets $\alpha, \beta > 0$ gefordert. Wir fordern

weiter, wie in Kapitel 3.2.3 bereits beschrieben, dass $\beta > 1$ gelten soll. Dies resultiert aus dem Wunsch, dass für extreme Werte die Wahrscheinlichkeit p_k, $k = 1, \ldots, K$, nicht stets 1 sein soll. Aus der Forderung, dass α stets positiv und ungleich 0 sein soll, ergibt sich nun

$$\begin{aligned} \alpha > 0 &\Leftrightarrow \frac{1-\mu}{\rho^2} - \mu > 0 \\ &\Leftrightarrow \frac{1-\mu}{\rho^2} > \mu \\ &\Leftrightarrow 1 - \mu > \mu\rho^2 \\ &\Leftrightarrow \frac{1-\mu}{\mu} > \rho^2 \\ &\Leftrightarrow \rho < \sqrt{\frac{1-\mu}{\mu}}. \end{aligned} \tag{20}$$

Aus der Bedingung $\beta > 1$ ergibt sich wiederum

$$\begin{aligned} \beta > 1 &\Leftrightarrow \frac{(1-\mu)^2}{\mu\rho^2} - (1-\mu) > 1 \\ &\Leftrightarrow \frac{(1-\mu)^2}{\mu\rho^2} > 1 + (1-\mu) = 2 - \mu \\ &\Leftrightarrow (1-\mu)^2 > (2-\mu)\mu\rho^2 \\ &\Leftrightarrow \frac{(1-\mu)^2}{(1-\mu)\mu} > \rho^2 \\ &\Leftrightarrow \rho^2 < \frac{(1-\mu)^2}{2\mu - \mu^2} \\ &\Leftrightarrow \rho < \sqrt{\frac{(1-\mu)^2}{2\mu - \mu^2}}. \end{aligned} \tag{21}$$

Da beide Bedingungen in Restriktionen für ρ resultieren, kann man nun schauen, welche Restriktion strenger ist. In diesem Fall ist es die zweite Restriktion (21), welche aus der Bedingung $\beta > 1$ resultiert ist. Daher reicht es in diesem Fall auch diese Restriktion abzufragen, da wenn (21) erfüllt ist, so ist auch (20) erfüllt.

An dieser Stelle sei noch einmal darauf hingewiesen, dass diese Herleitung im allgemeinen Fall dargelegt ist. In der Anwendung im hierarchischen Modell muss für μ jeweils μ_l, $l \in 0, 1$, eingesetzt werden.

A.4. Konvergenzdiagnostik

Bei MCMC-Verfahren ist die Beurteilung, ob Konvergenz vorliegt oder nicht, oft subjektiv. Es existieren jedoch verschiedene Konvergenzdiagnostiken, welche man zur objektiven Bewertung heranziehen kann, z.B. die Geweke Diagnostik oder auch die Heidelberger und Welch Diagnostik (Geweke, 1992, [12] bzw. Heidelberger und Welch, 1983, [18]).
Bei der Geweke Diagnostik ist die Idee, dass bei Konvergenz ein Test bzgl. der Lage für zwei Teile einer Markov Kette nicht ablehnen wird. Das bedeutet, falls eine Markov Kette $X_1, \ldots, X_N$ ab einer bestimmten Iteration $i^\star$ konvergiert ist, so hat sie keine großen Schwankungen und unterschiedet sich in der Lage nicht mehr signifikant. Würde man hier einen frühen (z.B. n_1 Iterationen nach $i^\star$) und einen späten Teil (z.B. die letzten n_2 Iterationen) mit Hilfe eines Tests auf Lageverschiebung vergleichen, so kann darüber eine Aussage getroffen werden, ob die Kette ihre stationäre Verteilung angenommen hat. Die Teststatistik vergleicht dabei die Mittelwerte der zwei Teilfolgen und standardisiert sie mit Hilfe von entsprechenden Spektraldichteschätzern. Dadurch kann eine Testentscheidung über die Standardnormalverteilung erfolgen.
Bei der Heidelberger und Welch Diagnostik ist hingegen die Idee, dass bei einer Konvergenz die Markov Kette in einem stationären Zustand und somit der beobachtete Prozess schwach stationär ist. Weiter nutzen Heidelberger und Welche eine Cramer von Mises Statistik, wessen Werte bei einem schwach stationären Prozess gegen eine Brownsche Brücke konvergieren. Dies testet die Heidelberger und Welch Diagnostik und kann so eine Aussage bzgl. der Konvergenz treffen.

B. Tabellen

Start-Segmentzahl	Iterationen	q-Quantil q=0.5	q=0.55	q=0.6	q=0.65	q=0.7	q=0.75	q=0.8
$K = 10$	5000	51	44	40	35	32	25	21
$K = 10$	10000	49	45	41	35	31	25	21
$K = 10$	25000	49	44	40	35	30	25	21
$K = 10$	50000	49	44	40	36	30	25	21
$K = 10$	100000	49	44	40	35	30	25	21
$K = 10$	150000	49	44	40	35	30	25	21
$K = 20$	5000	49	44	40	35	30	25	21
$K = 20$	10000	49	44	40	35	30	25	21
$K = 20$	25000	49	44	40	35	30	25	21
$K = 20$	50000	49	44	40	35	30	25	21
$K = 20$	100000	49	44	40	35	30	25	21
$K = 20$	150000	49	44	40	35	30	25	21

Tabelle 1: Geschätzte Anzahl an Segmenten des hierarchischen Modells mit Reversible Jump und Gewichtung (0.8, 0.1, 0.1) in Abhängigkeit von der Iterationenanzahl sowie des genutzten Quantils.

Simulationsparameter	$\hat{p}_{prior}$	$\hat{p}_{ash}$	$\hat{p}_{EMM25}$	$\hat{p}_{EMM55}$	$\hat{p}_{db50}$	$\hat{p}_{db75}$	$\hat{p}_{db100}$	$\hat{p}_{gammics}$
p=0.4, μ=4, r=15	0.38	0.38	0.35	0.71	0.31	0.61	0.84	0.42
p=0.4, μ=4, r=15	0.36	0.28	0.37	0.74	0.36	0.65	0.87	0.43
p=0.4, μ=4, r=15	0.37	0.37	0.37	0.74	0.32	0.62	0.86	0.40
p=0.4, μ=4, r=15	0.36	0.36	0.39	0.75	0.39	0.68	0.87	0.42
p=0.4, μ=4, r=30	0.38	0.38	0.39	0.73	0.33	0.62	0.83	0.46
p=0.4, μ=4, r=30	0.37	0.29	0.34	0.72	0.30	0.63	0.86	0.47
p=0.4, μ=4, r=30	0.37	0.32	0.36	0.72	0.33	0.63	0.84	0.42
p=0.4, μ=4, r=30	0.36	0.31	0.38	0.74	0.33	0.67	0.88	0.53
p=0.4, μ=8, r=15	0.37	0.20	0.43	0.73	0.49	0.61	0.79	0.41
p=0.4, μ=8, r=15	0.37	0.25	0.41	0.71	0.43	0.57	0.78	0.37
p=0.4, μ=8, r=15	0.36	0.22	0.44	0.72	0.46	0.63	0.83	0.40
p=0.4, μ=8, r=15	0.37	0.25	0.44	0.74	0.49	0.65	0.82	0.42
p=0.4, μ=8, r=30	0.35	0.19	0.43	0.73	0.47	0.65	0.83	0.44
p=0.4, μ=8, r=30	0.36	0.21	0.43	0.72	0.48	0.63	0.80	0.45
p=0.4, μ=8, r=30	0.37	0.20	0.42	0.72	0.46	0.63	0.82	0.40
p=0.4, μ=8, r=30	0.38	0.23	0.39	0.70	0.45	0.61	0.81	0.41
p=0.8, μ=4, r=15	0.75	0.25	0.52	0.83	0.52	0.79	0.90	0.76
p=0.8, μ=4, r=15	0.73	0.22	0.53	0.86	0.52	0.78	0.90	0.79
p=0.8, μ=4, r=15	0.73	0.37	0.55	0.86	0.55	0.81	0.90	0.81
p=0.8, μ=4, r=15	0.71	0.24	0.54	0.84	0.52	0.77	0.88	0.79
p=0.8, μ=4, r=30	0.74	0.29	0.48	0.83	0.46	0.77	0.90	0.76
p=0.8, μ=4, r=30	0.70	0.35	0.46	0.84	0.51	0.80	0.93	0.81
p=0.8, μ=4, r=30	0.75	0.37	0.48	0.83	0.49	0.76	0.89	0.80
p=0.8, μ=4, r=30	0.69	0.23	0.49	0.84	0.53	0.77	0.91	0.82
p=0.8, μ=8, r=15	0.70	0.28	0.66	0.86	0.82	0.87	0.91	0.80
p=0.8, μ=8, r=15	0.72	0.24	0.66	0.86	0.81	0.88	0.91	0.79
p=0.8, μ=8, r=15	0.76	0.38	0.67	0.86	0.82	0.88	0.92	0.81
p=0.8, μ=8, r=15	0.68	0.27	0.68	0.88	0.85	0.89	0.93	0.83
p=0.8, μ=8, r=30	0.76	0.38	0.59	0.83	0.76	0.85	0.91	0.79
p=0.8, μ=8, r=30	0.71	0.36	0.62	0.84	0.77	0.87	0.91	0.79
p=0.8, μ=8, r=30	0.74	0.37	0.64	0.84	0.79	0.88	0.92	0.79
p=0.8, μ=8, r=30	0.67	0.27	0.65	0.86	0.79	0.89	0.94	0.82

Tabelle 2: Geschätzte Proportionen an Proteinen in Clustern in den jeweiligen Simulationen und die entsprechenden Methoden.

Tabelle 3: Geschätzte Anzahl an Clustern in Abhängigkeit des Simulationssettings sowie des simulierten Bildes (links: simulierte Anzahl an Proteinen, rechts: mit ASH geschätzte Anzahl an Clustern); dabei entspricht grün $\widehat{=}$ der Simulation des grünen Proteins im Dual Colour Fall bzgl. der einzelnen Single Colour Simulation, rot1 $\widehat{=}$ der Simulation des roten Proteins in Setting 1, rot2 $\widehat{=}$ der Simulation des roten Proteins in Setting 2 und rot 3 $\widehat{=}$ der Simulation des roten Proteins in Setting 3.

Simulationssetting	grün	ASH	rot 1	ASH	rot 2	ASH	rot 3	ASH
p=0.4, μ=4, r=15	202	150	207	150	202	130	202	120
p=0.4, μ=4, r=15	194	120	222	130	194	120	194	130
p=0.4, μ=4, r=15	195	150	208	150	195	180	195	150
p=0.4, μ=4, r=15	196	150	183	100	196	180	196	150
p=0.4, μ=4, r=30	205	150	210	150	205	120	205	150
p=0.4, μ=4, r=30	194	120	205	150	194	150	194	120
p=0.4, μ=4, r=30	205	130	211	100	205	150	205	130
p=0.4, μ=4, r=30	210	130	204	150	210	150	210	200
p=0.4, μ=8, r=15	92	80	102	100	92	100	92	120
p=0.4, μ=8, r=15	114	100	121	100	114	130	114	120
p=0.4, μ=8, r=15	102	90	106	120	102	120	102	120
p=0.4, μ=8, r=15	102	100	96	90	102	100	102	100
p=0.4, μ=8, r=30	89	80	98	90	89	80	89	90
p=0.4, μ=8, r=30	101	90	102	100	101	100	101	90
p=0.4, μ=8, r=30	96	80	89	90	96	100	96	100
p=0.4, μ=8, r=30	103	90	108	90	103	100	103	120
p=0.8, μ=4, r=15	444	100	405	120	444	120	444	130
p=0.8, μ=4, r=15	424	90	468	120	424	100	424	120
p=0.8, μ=4, r=15	402	150	454	170	402	150	402	150
p=0.8, μ=4, r=15	432	100	416	120	432	150	432	120
p=0.8, μ=4, r=30	441	120	426	180	441	150	441	150
p=0.8, μ=4, r=30	428	150	418	150	428	180	428	180
p=0.8, μ=4, r=30	396	150	428	200	396	150	396	180

p=0.8, μ=4, r=30	427	100	411	100	427	150	427	150
p=0.8, μ=8, r=15	211	120	183	130	211	150	211	130
p=0.8, μ=8, r=15	217	100	200	100	217	100	217	130
p=0.8, μ=8, r=15	203	150	188	170	203	150	203	150
p=0.8, μ=8, r=15	211	120	233	130	211	150	211	130
p=0.8, μ=8, r=30	180	150	196	150	180	180	180	170
p=0.8, μ=8, r=30	192	150	214	120	192	120	192	120
p=0.8, μ=8, r=30	197	150	198	120	197	150	197	130
p=0.8, μ=8, r=30	210	120	210	150	210	150	210	130

	grün			rot 1			rot 2			rot 3		
Simulation	$\bar{\hat{r}}$	$\bar{\hat{\mu}}$	$\bar{\hat{p}}$	$\bar{\hat{r}}$	$\bar{\hat{\mu}}$	$\bar{\hat{p}}$	$\bar{\hat{r}}$	$\bar{\hat{\mu}}$	$\bar{\hat{p}}$	$\bar{\hat{r}}$	$\bar{\hat{\mu}}$	$\bar{\hat{p}}$
p=0.4,μ=4,r=15	9.16	3.74	0.43	9.08	6.82	0.43	8.48	3.57	0.78	9.39	7.08	0.82
p=0.4,μ=4,r=30	8.98	3.73	0.46	8.61	6.89	0.42	8.31	3.64	0.81	8.76	6.63	0.81
p=0.4,μ=8,r=15	9.80	3.84	0.43	9.60	6.85	0.42	8.45	3.68	0.81	8.94	6.67	0.81
p=0.4,μ=8,r=30	8.43	3.74	0.44	8.40	6.59	0.44	8.24	3.62	0.81	8.17	6.66	0.84
p=0.8,μ=4,r=15	14.71	3.68	0.50	12.80	6.13	0.45	12.84	3.60	0.80	13.27	6.36	0.81
p=0.8,μ=4,r=30	15.20	3.81	0.49	11.61	5.93	0.45	13.07	3.50	0.79	11.95	6.01	0.81
p=0.8,μ=8,r=15	15.01	3.81	0.48	12.87	5.95	0.42	13.34	3.72	0.80	12.31	6.00	0.81
p=0.8,μ=8,r=30	15.34	3.69	0.54	12.78	5.80	0.45	13.44	3.75	0.81	11.92	5.87	0.82

Tabelle 4: Durchschnittliche Schätzung mit Hilfe der Gammics Methode für alle Parametereinstellungen der Simulationen, wobei $\bar{\hat{r}}$ die Schätzung des durchschnittlichen Radius, $\bar{\hat{\mu}}$ die Schätzung für die durchschnittliche Größe und $\bar{\hat{p}}$ die Schätzung der durchschnittlichen Proportion an Proteinen in Clustern repräsentiert; weiter steht hier „grün“ für die Simulation der grünen Proteine, „rot 1“ für die Simulation der roten Proteine in Setting 1, „rot 2“ für die Simulation der roten Proteine in Setting 2 und „rot 3“ für die Simulation der roten Proteine in Setting 3.

Simulationsparameter	$\hat{p}_{prior}$	$\hat{p}_{EMM25}$	$\hat{p}_{EMM55}$	$\hat{p}_{db50}$	$\hat{p}_{db75}$	$\hat{p}_{db100}$
p=0.4, μ=4, r=15	0.37	0.39	0.73	0.33	0.63	0.85
p=0.4, μ=4, r=30	0.36	0.35	0.73	0.33	0.64	0.87
p=0.4, μ=8, r=15	0.37	0.43	0.72	0.46	0.61	0.80
p=0.4, μ=8, r=30	0.37	0.40	0.73	0.45	0.62	0.82
p=0.8, μ=4, r=15	0.71	0.55	0.85	0.55	0.79	0.91
p=0.8, μ=4, r=30	0.69	0.48	0.84	0.49	0.79	0.92
p=0.8, μ=8, r=15	0.71	0.66	0.85	0.81	0.87	0.91
p=0.8, μ=8, r=30	0.70	0.63	0.86	0.80	0.89	0.92

Tabelle 5: Durchschnittlich geschätzte Proportion der Proteine in Clustern für die roten Proteine aus dem Dual Colour Datensatz für Szenario 1 (unabhängige Proteine).

Simulationsparameter	$\hat{p}_{prior}$	$\hat{p}_{EMM25}$	$\hat{p}_{EMM55}$	$\hat{p}_{db50}$	$\hat{p}_{db75}$	$\hat{p}_{db100}$
p=0.4, μ=4, r=15	0.37	0.36	0.67	0.33	0.55	0.74
p=0.4, μ=4, r=30	0.37	0.36	0.67	0.33	0.56	0.74
p=0.4, μ=8, r=15	0.37	0.41	0.66	0.44	0.55	0.71
p=0.4, μ=8, r=30	0.37	0.43	0.66	0.47	0.56	0.70
p=0.8, μ=4, r=15	0.73	0.55	0.83	0.54	0.75	0.86
p=0.8, μ=4, r=30	0.72	0.56	0.84	0.54	0.75	0.86
p=0.8, μ=8, r=15	0.71	0.68	0.85	0.82	0.86	0.89
p=0.8, μ=8, r=30	0.72	0.67	0.84	0.81	0.85	0.89

Tabelle 6: Durchschnittlich geschätzte Proportion der Proteine in Clustern für die roten Proteine aus dem Dual Colour Datensatz für Szenario 2 (korrelierte Clusterzentren).

		grünes Protein							
		Track 1	Track 2	Track 3	Track 4	Track 5	Track 6	Track 7	Track 8
rotes Protein	Track 1	**0.8712**	0.4511	0.3361	0.6413	0.2044	0.5295	0.3353	0.8560
	Track 2	0.4013	**0.9052**	0.3502	0.3064	0.3801	0.3431	0.2959	0.4157
	Track 3	0.3872	0.3450	**0.8840**	0.1999	0.3147	0.1935	0.3173	0.3297
	Track 4	0.6522	0.4477	0.2620	**0.8645**	0.2224	**0.6485**	0.2351	0.8479
	Track 5	0.4418	0.5608	0.5580	0.2514	**0.4708**	0.2561	**0.3563**	0.4109
	Track 6	0.6026	0.4990	0.3224	0.3315	0.2432	0.2990	0.3397	0.5834
	Track 7	0.3419	0.1932	0.1673	0.2536	0.0533	0.1552	0.2330	0.3464
	Track 8	0.7917	0.3243	0.2522	0.7062	0.1924	0.5946	0.2611	**0.8639**

Tabelle 7: Ergebnis der Anwendung des Zusammenhangsmaßes auf simulierte Beispieltracks mit Gewichtung $w_1 = w_2 = 0.4$ und $w_3 = 0.2$; die dick gedruckten Einträge entsprechen den jeweils höchsten Werten des Zusammenhangsmaß für die einzelnen **grünen** Tracks.

(w_1, w_2, w_3)	$(\epsilon, MinPts)$	Zeit	Min.	q_{25}	Med.	Mittel.	q_{75}	Max.	Var.
$(1/3, 1/3, 1/3)$	$(10, 2)$	2	0.0075	0.1929	0.2394	0.2417	0.2829	0.9698	0.0061
$(0.4, 0.4, 0.2)$	$(10, 2)$	2	0.0050	0.2218	0.2762	0.2770	0.3262	0.9811	0.0078
$(1/3, 1/3, 1/3)$	$(40, 5)$	2	0.0075	0.5252	0.5724	0.5684	0.6161	0.9698	0.0053
$(0.4, 0.4, 0.2)$	$(40, 5)$	2	0.0050	0.6207	0.6758	0.6690	0.7261	0.9811	0.0067
$(1/3, 1/3, 1/3)$	$(10, 2)$	5	0.0047	0.1970	0.2464	0.2439	0.2880	0.9974	0.0058
$(0.4, 0.4, 0.2)$	$(10, 2)$	5	0.0029	0.2275	0.2854	0.2802	0.3339	0.9982	0.0073
$(1/3, 1/3, 1/3)$	$(40, 5)$	5	0.0088	0.5296	0.5794	0.5727	0.6213	0.9974	0.0055
$(0.4, 0.4, 0.2)$	$(40, 5)$	5	0.0092	0.6266	0.6851	0.6748	0.7338	0.9982	0.0069
$(1/3, 1/3, 1/3)$	$(10, 2)$	10	0.0022	0.2052	0.2505	0.2499	0.2914	0.9998	0.0058
$(0.4, 0.4, 0.2)$	$(10, 2)$	10	0.0013	0.2374	0.2903	0.2869	0.3380	0.9997	0.0072
$(1/3, 1/3, 1/3)$	$(40, 5)$	10	0.0022	0.5383	0.5838	0.5788	0.6247	0.9998	0.0054
$(0.4, 0.4, 0.2)$	$(40, 5)$	10	0.0013	0.6370	0.6903	0.6816	0.7379	0.9997	0.0066

Tabelle 8: Übersicht wichtiger Kennzahlen der errechneten Zusammenhänge in Abhängigkeit der Gewichtung und der Parameterwahl für den DBSCAN Algorithmus; dabei ist mit Zeit der Zeitpunkt der Datenerhebung nach Stimulation, mit Min. das Minimum, mit q_{25} das 25%-Quantil, mit Med. der Median, mit Mittel. der Mittelwert, mit q_{75} das 75%-Quantil, mit Max. das Maximum und mit Var. die Varianz gemeint.

C. Abbildungen

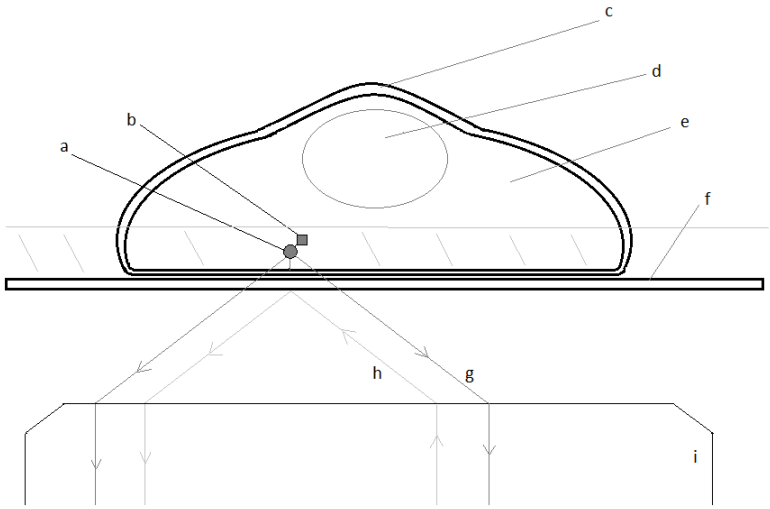

Abbildung 1: Schematische Darstellung des experimentellen Aufbaus einer TIRF-Fluoreszenzmikroskopie; dabei ist a: das Protein (z.B. ein Ras-Protein), b: das Fluoreszenzprotein, c: die Plasmamembran, d: der Zellkern, e: das Zytoplasma, f: der Objektträger, g: die Emission, h: das Licht zur Stimulation und i: das Objektiv bzw. die Linse des Objektivs.

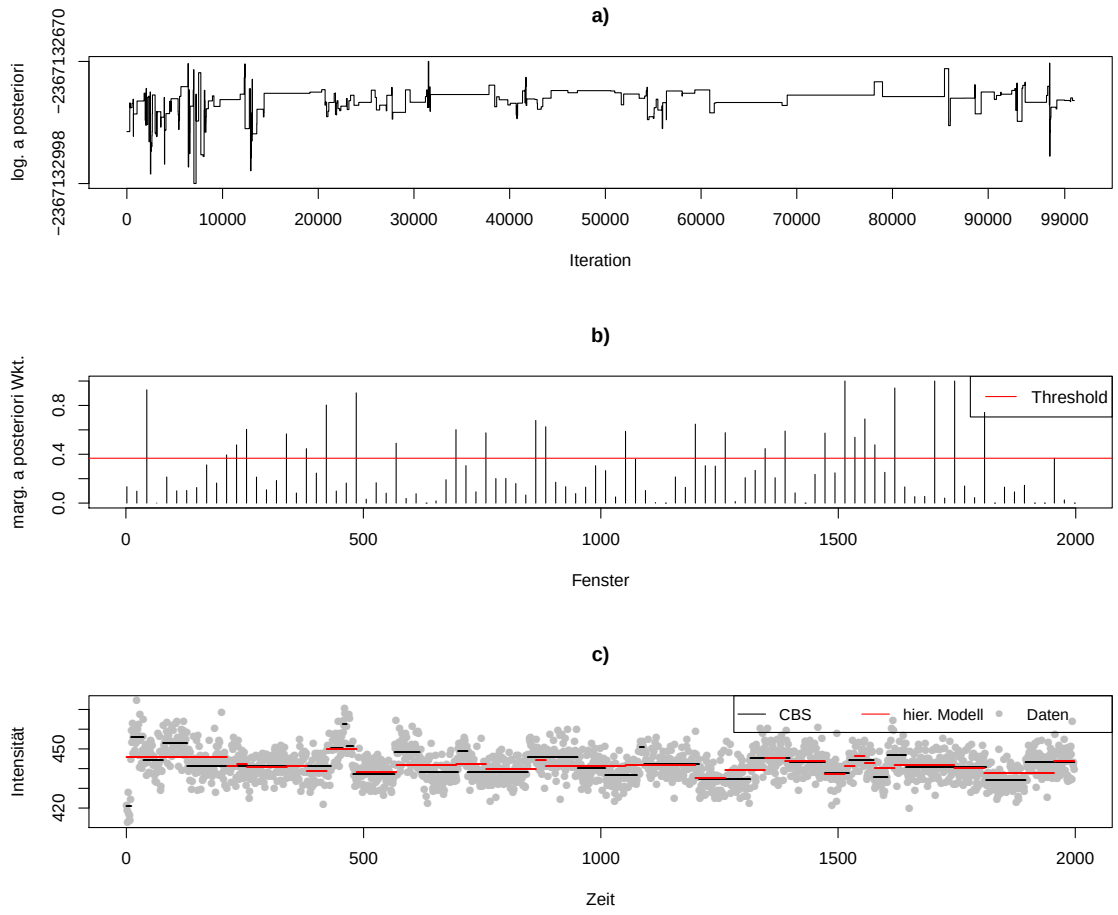

Abbildung 2: Ergebnis der 99 000 MCMC-Iterationen (100 000 Iterationen insgesamt, jedoch davon 1 000 Iterationen Burn-In) des Random Walk Metropolis Hastings Algorithmus mit einer zufälligen Startsequenz für eine exemplarische Proteinzeitreihe aus Experiment 2: a) logarithmierte a posteriori, b) marginale posteriori Wahrscheinlichkeit für mindestens einen CP auf einem Gitter mit ca. 100 Fenstern und c) die resultierende Segmentierung, basierend auf den CPs mit höchster marginaler a posteriori Wahrscheinlichkeit aus b) im Vergleich zur Segmentierung des CBS Algorithmus.

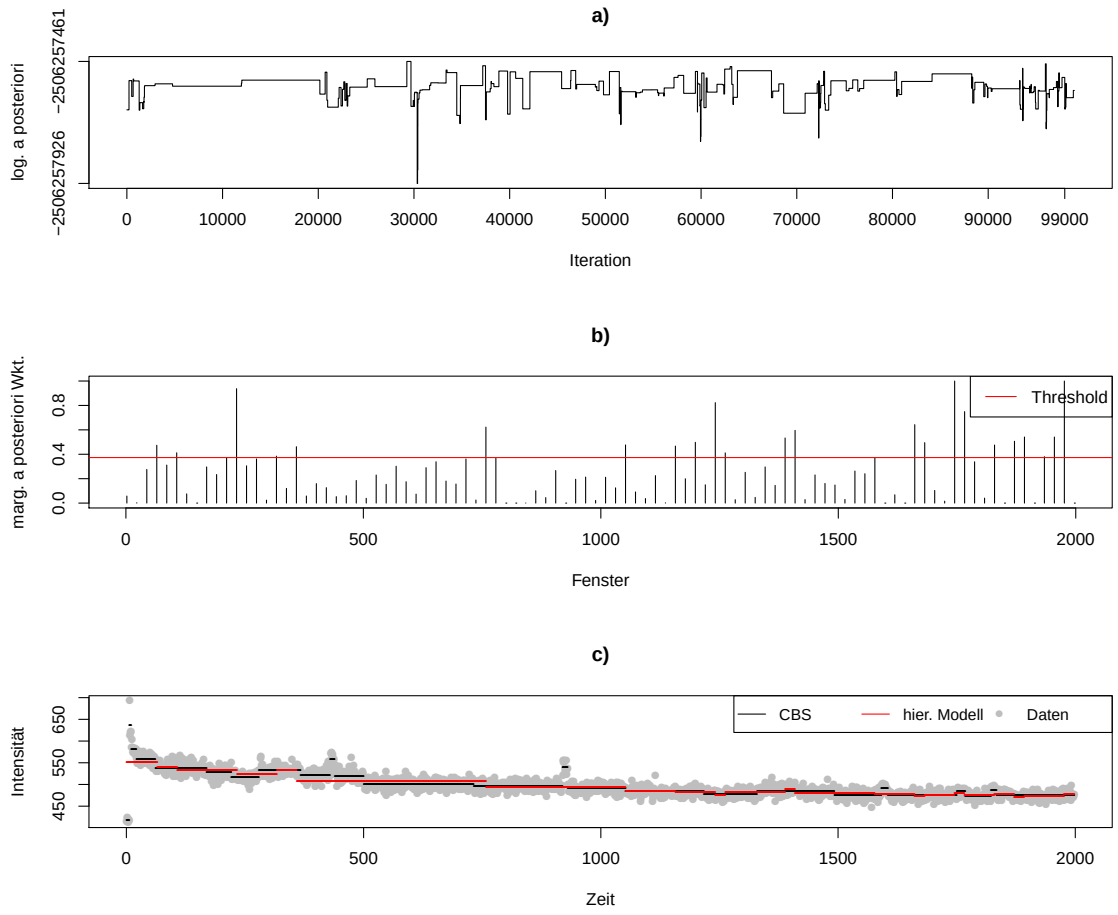

Abbildung 3: Ergebnis der 99 000 MCMC-Iterationen (100 000 Iterationen insgesamt, jedoch davon 1 000 Iterationen Burn-In) des Random Walk Metropolis Hastings Algorithmus mit einer zufälligen Startsequenz für eine exemplarische Proteinzeitreihe aus Experiment 3: a) logarithmierte a posteriori, b) marginale a posteriori Wahrscheinlichkeit für mindestens einen CP auf einem Gitter mit ca. 100 Fenstern und c) die resultierende Segmentierung, basierend auf den CPs mit höchster marginaler a posteriori Wahrscheinlichkeit aus b) im Vergleich zur Segmentierung des CBS Algorithmus.

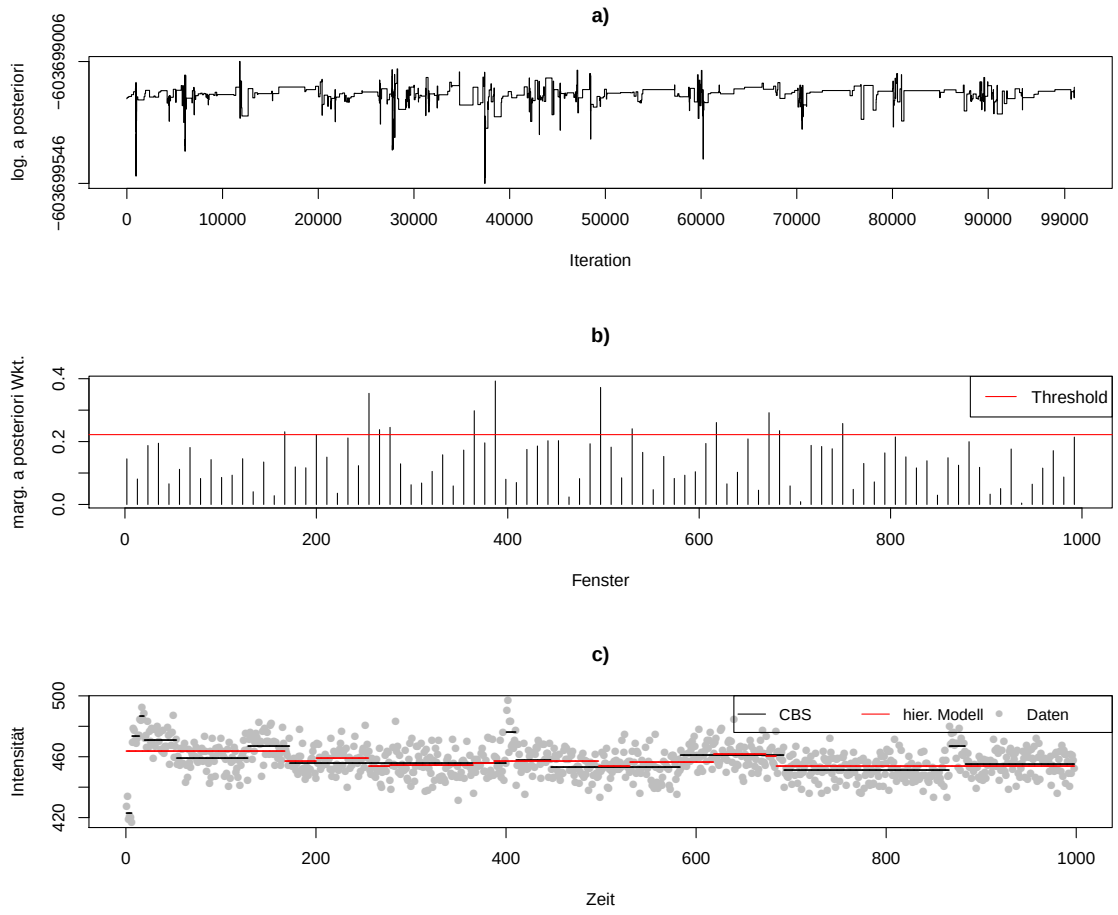

Abbildung 4: Ergebnis der 99 000 MCMC-Iterationen (100 000 Iterationen insgesamt, jedoch davon 1 000 Iterationen Burn-In) des Random Walk Metropolis Hastings Algorithmus mit einer zufälligen Startsequenz für eine exemplarische Proteinzeitreihe aus Experiment 4: a) logarithmierte a posteriori, b) marginale a posteriori Wahrscheinlichkeit für mindestens einen CP auf einem Gitter mit ca. 100 Fenstern und c) die resultierende Segmentierung, basierend auf den CPs mit höchster marginaler a posteriori Wahrscheinlichkeit aus b) im Vergleich zur Segmentierung des CBS Algorithmus.

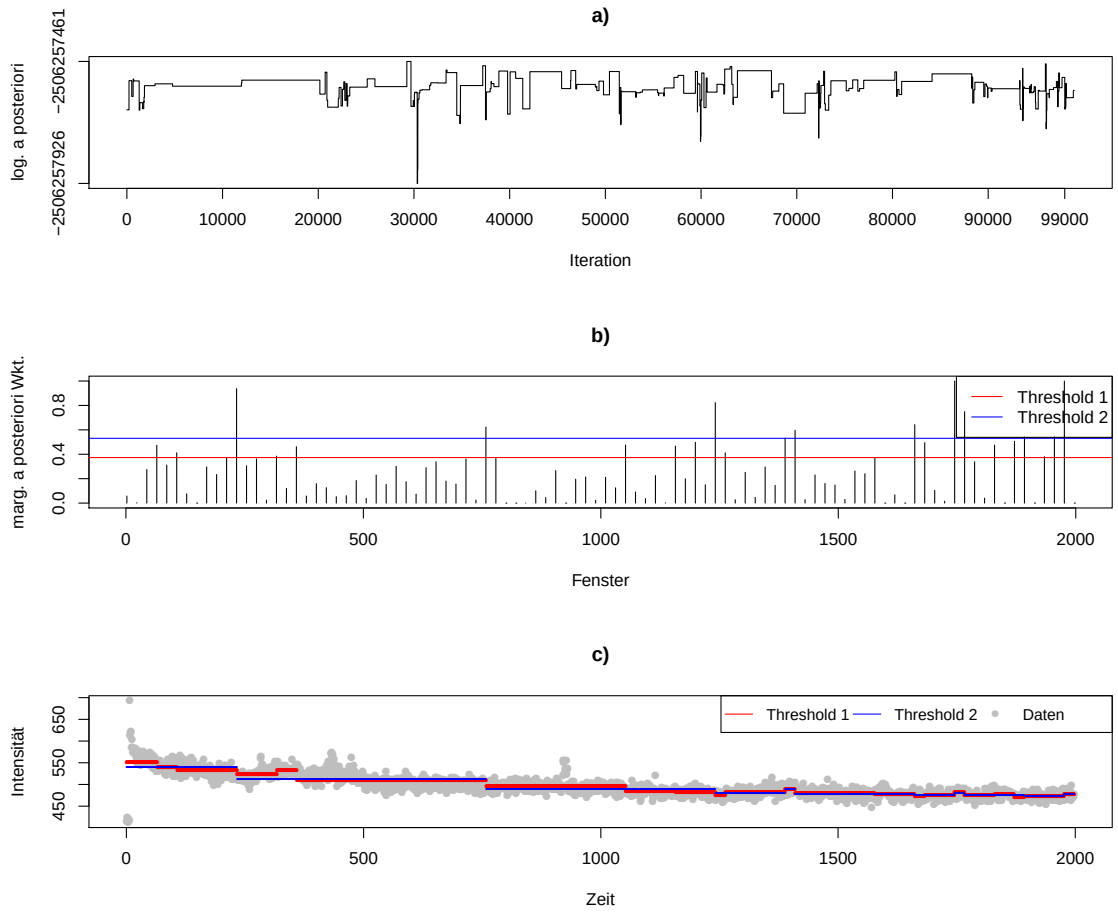

Abbildung 5: Ergebnis der 99 000 MCMC-Iterationen (100 000 Iterationen insgesamt, jedoch davon 1 000 Iterationen Burn-In) des Random Walk Metropolis Hastings Algorithmus mit einer zufälligen Startsequenz und beiden Thresholds für eine exemplarische Proteinzeitreihe aus Experiment 3: a) logarithmierte a posteriori, b) marginale a posteriori Wahrscheinlichkeit für mindestens einen CP auf einem Gitter mit ca. 100 Fenstern und c) die resultierende Segmentierung, basierend auf den CPs mit höchster marginaler a posteriori Wahrscheinlichkeit aus b) in Abhängigkeit vom Threshold.

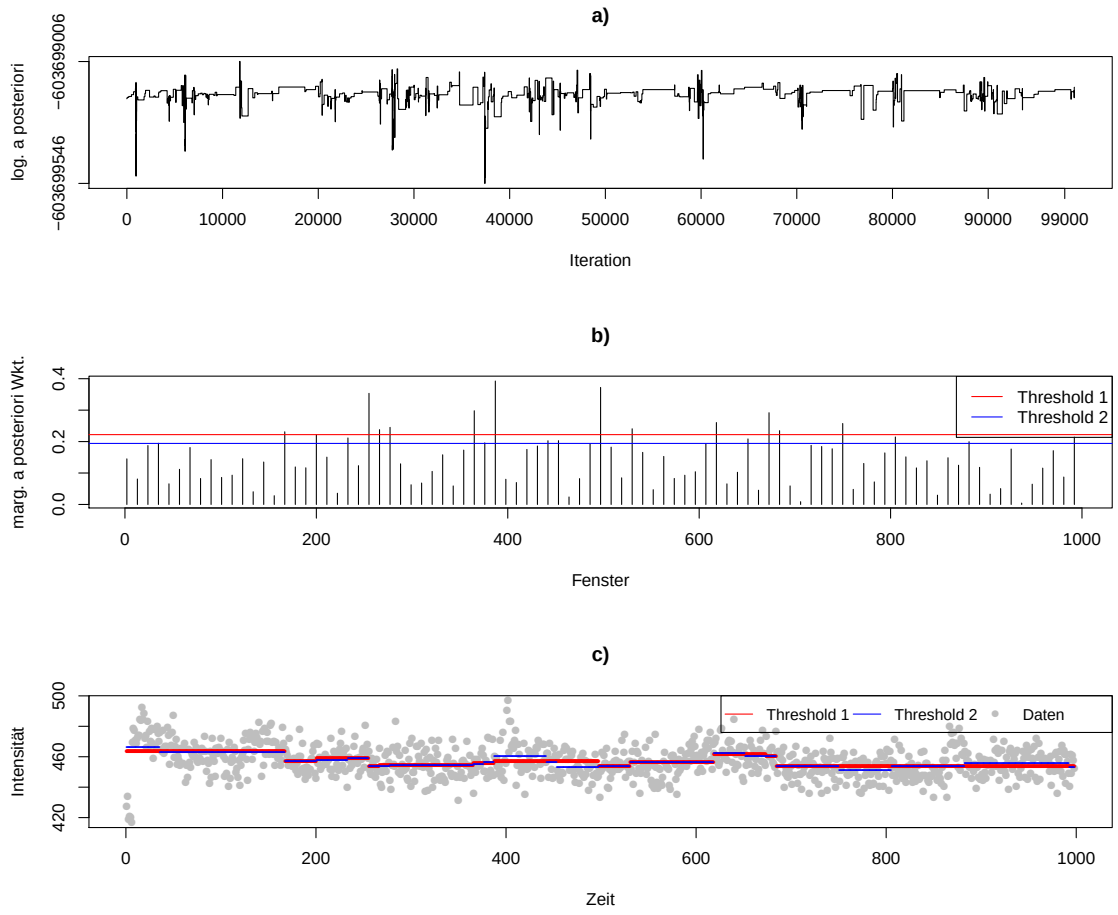

Abbildung 6: Ergebnis der 99 000 MCMC-Iterationen (100 000 Iterationen insgesamt, jedoch davon 1 000 Iterationen Burn-In) des Random Walk Metropolis Hastings Algorithmus mit einer zufälligen Startsequenz und beiden Thresholds für eine exemplarische Proteinzeitreihe aus Experiment 4: a) logarithmierte a posteriori, b) marginale a posteriori Wahrscheinlichkeit für mindestens einen CP auf einem Gitter mit ca. 100 Fenstern und c) die resultierende Segmentierung, basierend auf den CPs mit höchster marginaler a posteriori Wahrscheinlichkeit aus b) in Abhängigkeit vom Threshold.

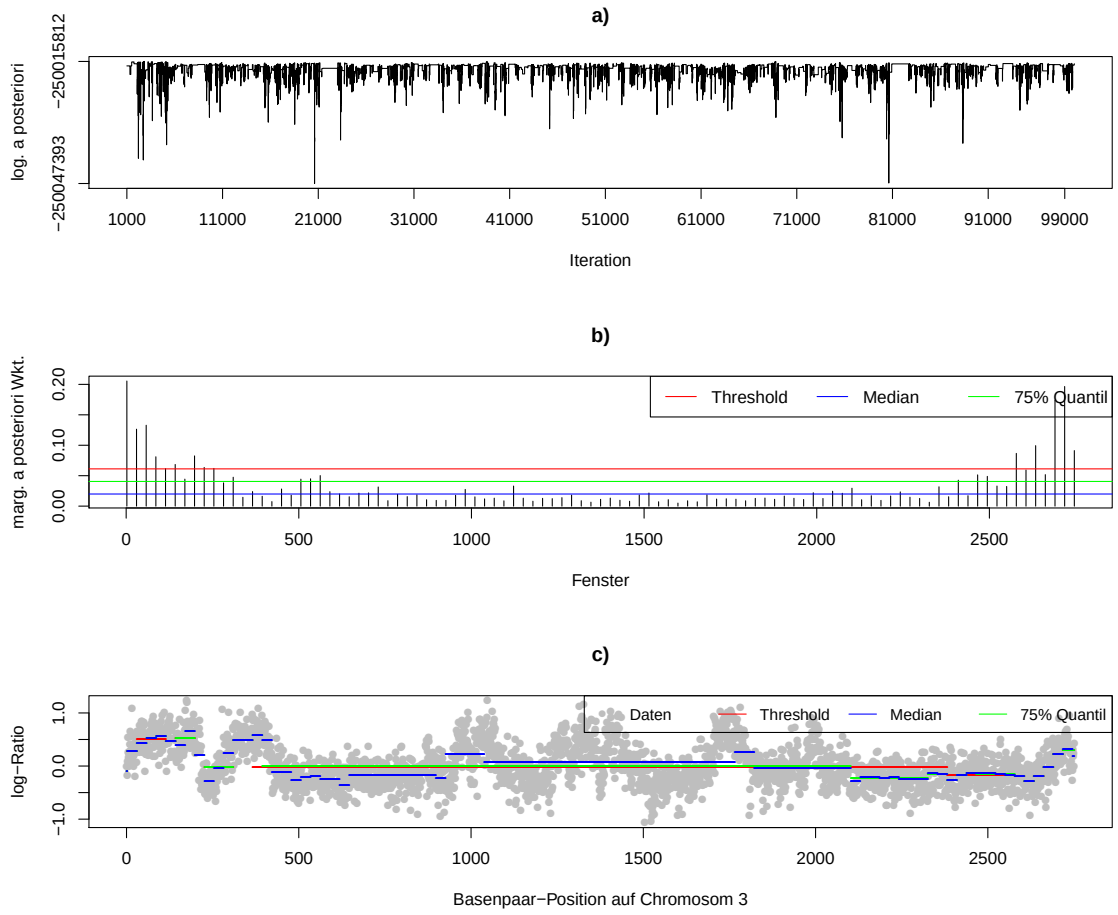

Abbildung 7: Ergebnis der 99 000 MCMC-Iterationen (100 000 Iterationen insgesamt, jedoch davon 1 000 Iterationen Burn-In) des Random Walk Metropolis Hastings Algorithmus mit einer zufälligen Startsequenz und dem Reversible Jump für die ChIP-Seq-Daten: a) logarithmierte a posteriori, b) marginale a posteriori Wahrscheinlichkeit für mindestens einen CP auf einem Gitter mit ca. 100 Fenstern und c) die resultierende Segmentierung, basierend auf den CPs mit höchster marginaler a posteriori Wahrscheinlichkeit aus b) im Vergleich zur Segmentierung des CBS Algorithmus.

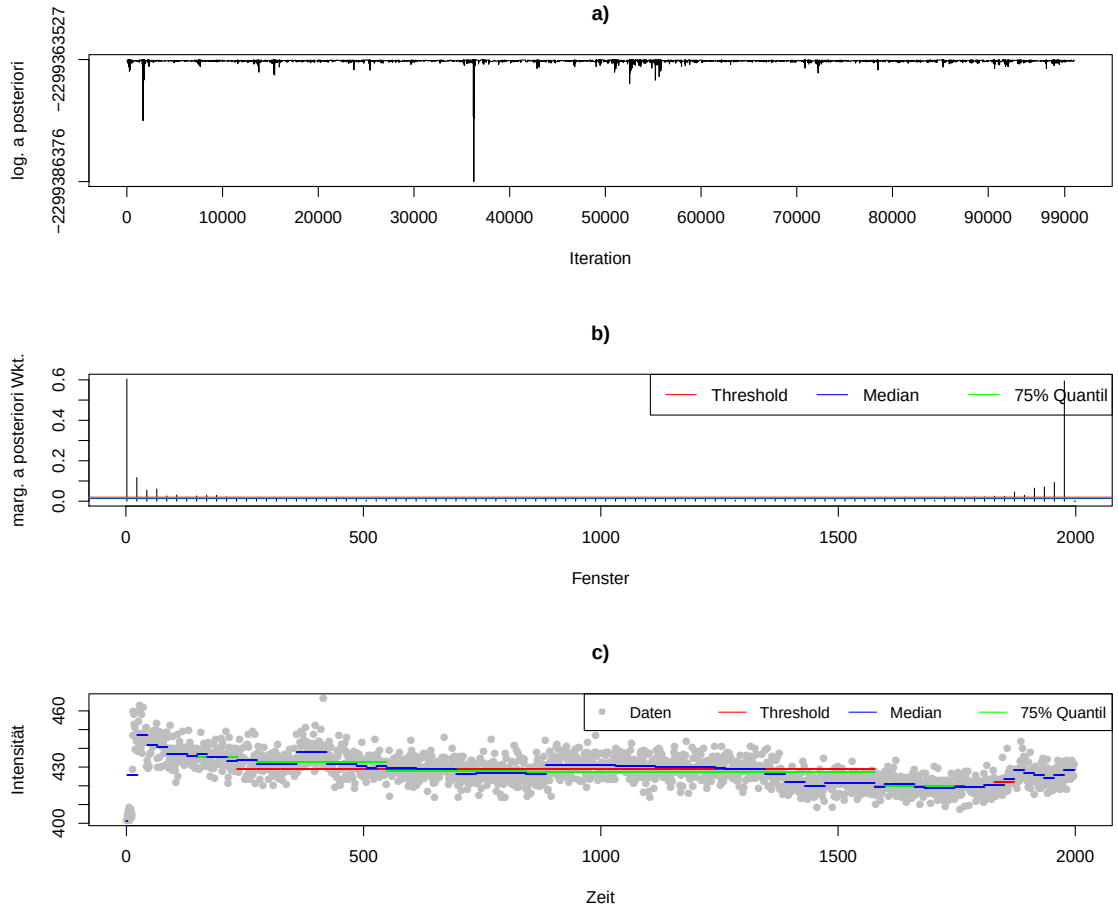

Abbildung 8: Ergebnis der 99 000 MCMC-Iterationen (100 000 Iterationen insgesamt, jedoch davon 1 000 Iterationen Burn-In) des Random Walk Metropolis Hastings Algorithmus mit einer zufälligen Startsequenz und dem Reversible Jump für die exemplarische Proteinzeitreihe aus Experiment 1: a) logarithmierte a posteriori, b) marginale a posteriori Wahrscheinlichkeit für mindestens einen CP auf einem Gitter mit ca. 100 Fenstern und c) die resultierende Segmentierung, basierend auf den CPs mit höchster marginaler a posteriori Wahrscheinlichkeit aus b) im Vergleich zur Segmentierung des CBS Algorithmus.

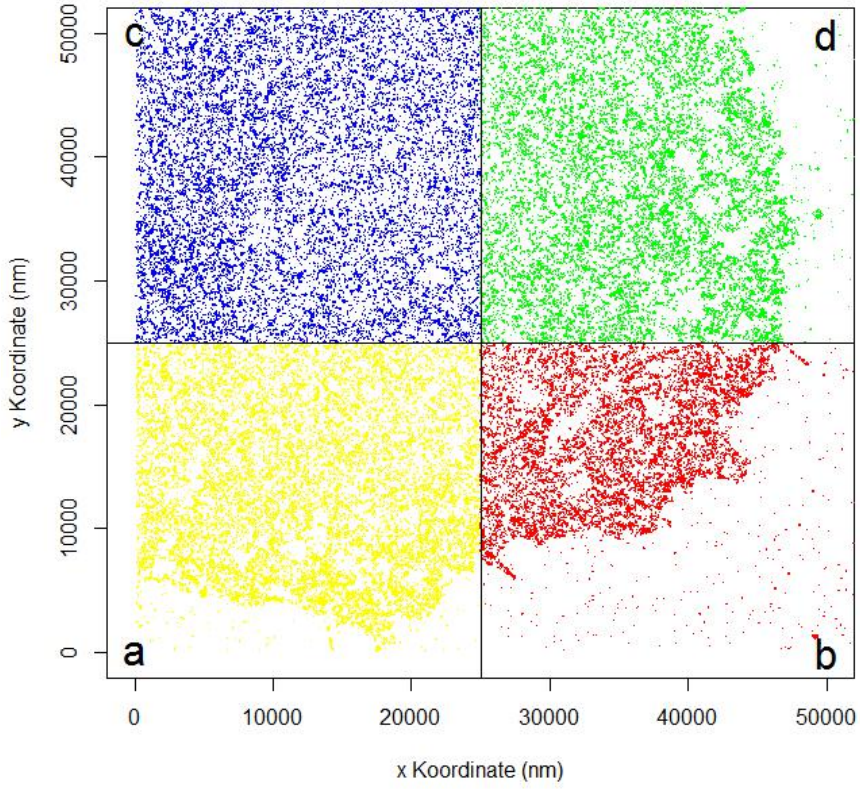

Abbildung 9: Zellaufteilung der räumlichen experimentellen Daten für eine Analyse mittels des EMM.

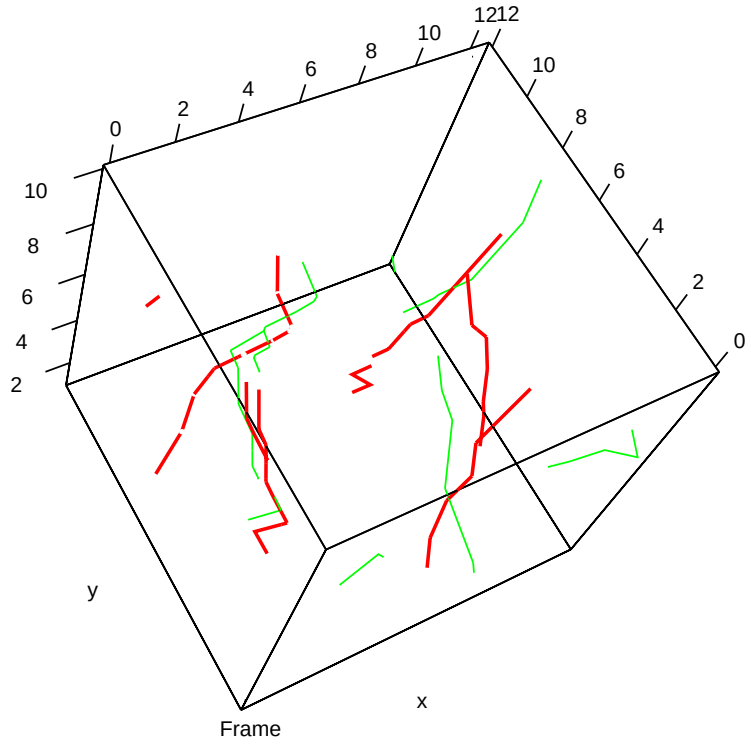

Abbildung 10: Räumliche Darstellung der 16 simulierten Tracks (die farbliche Kennzeichnung entspricht dabei der Zugehörigkeit zu Protein 1 bzw. 2).

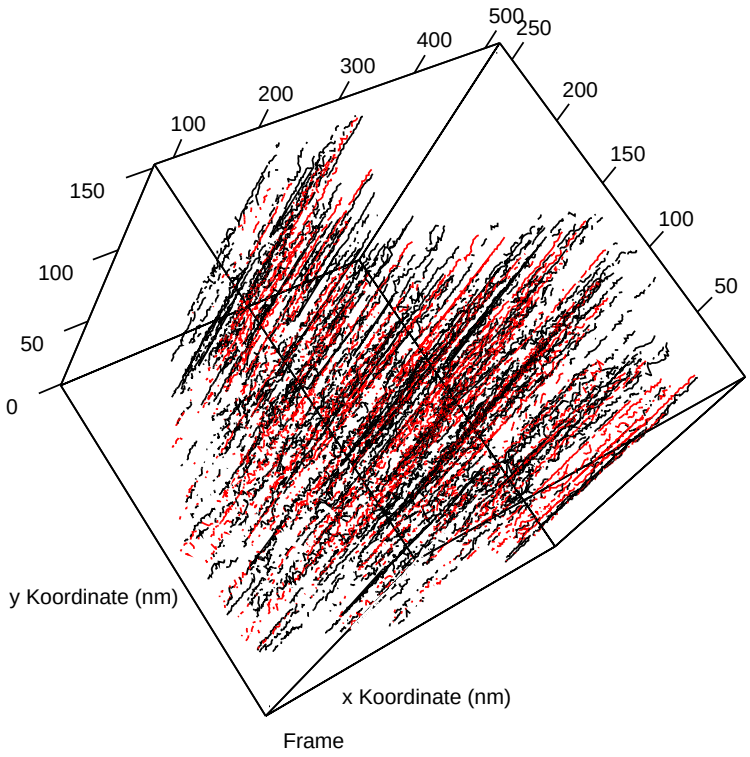

Abbildung 11: Räumliche Darstellung der experimentellen Daten (die farbliche Kennzeichnung entspricht dabei der Zugehörigkeit zu den zwei Proteinen: schwarze Tracks entstammen einem EGFR Protein, rote Tracks dem PTB Protein).

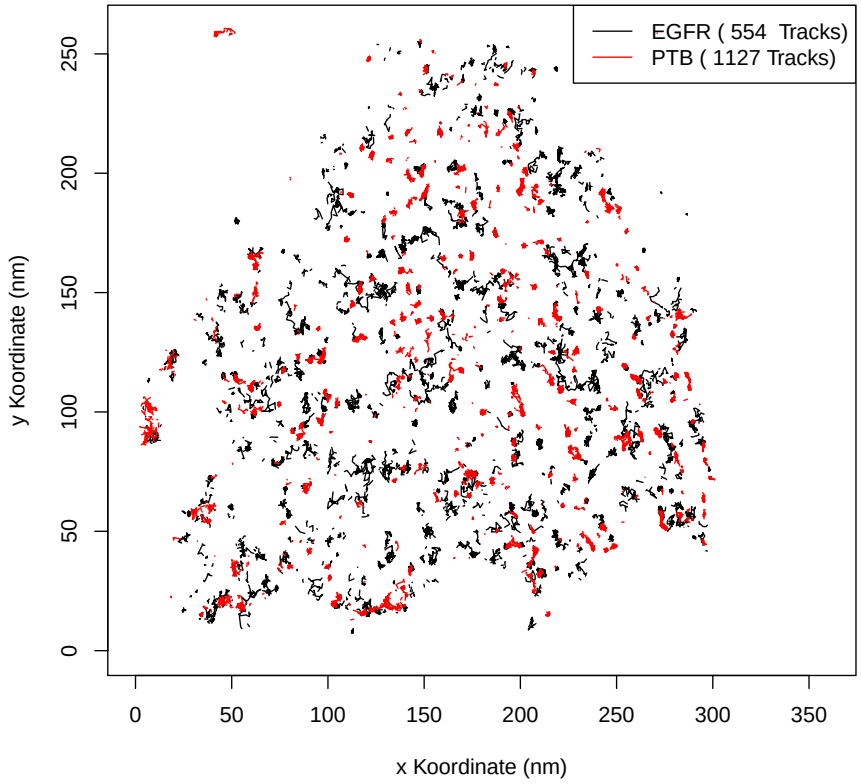

Abbildung 12: Darstellung aller EGFR- und PTB-Tracks aus dem Experiment 2 Minuten nach der Stimulation für Zelle 1.

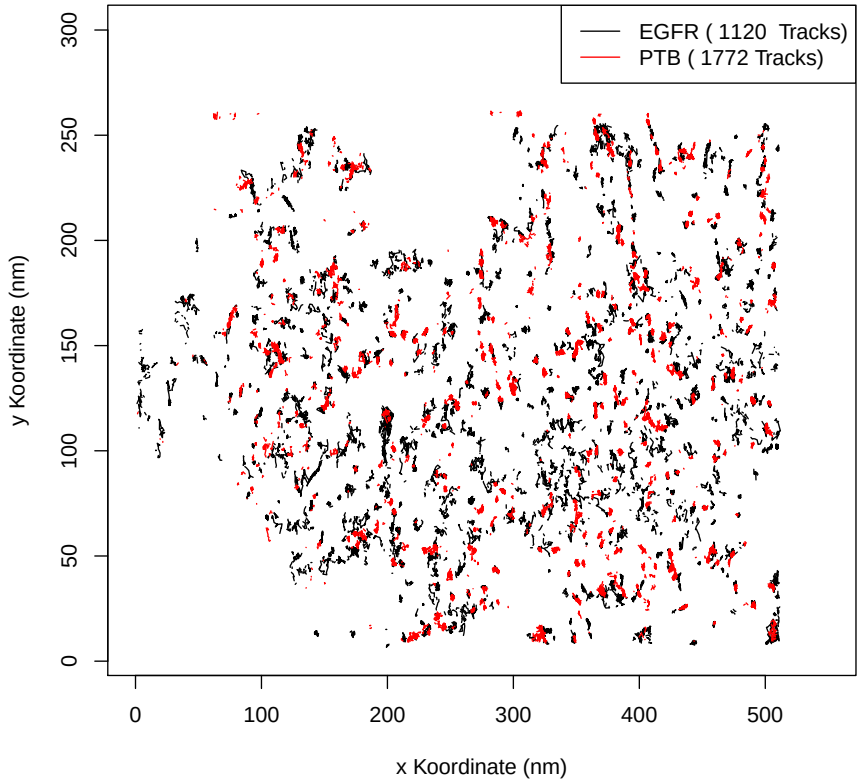

Abbildung 13: Darstellung aller EGFR- und PTB-Tracks aus dem Experiment 5 Minuten nach der Stimulation für Zelle 1.

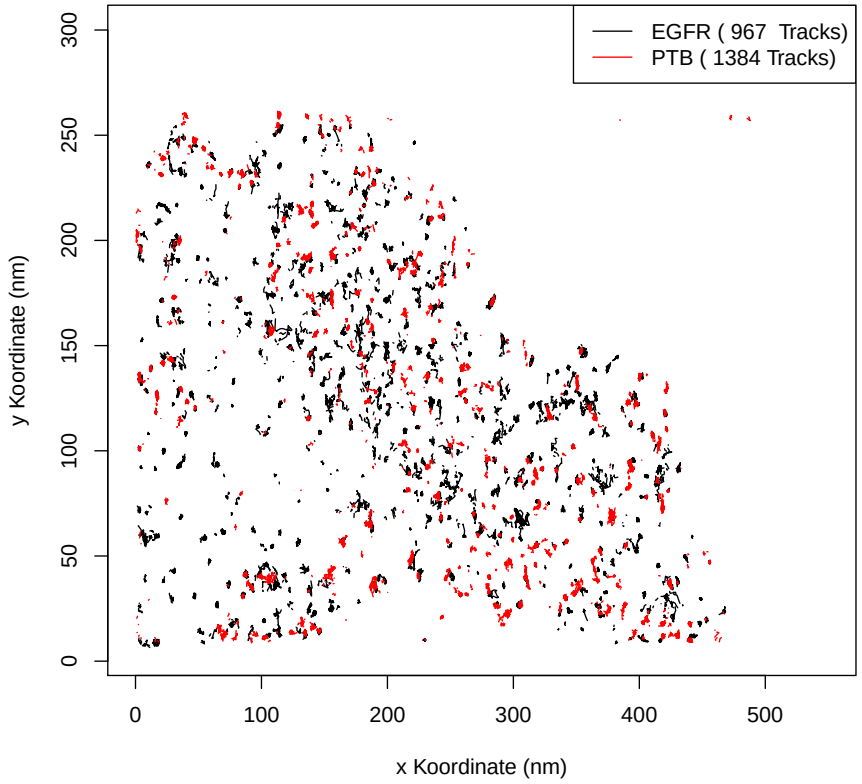

Abbildung 14: Darstellung aller EGFR- und PTB-Tracks aus dem Experiment 10 Minuten nach der Stimulation für Zelle 1.

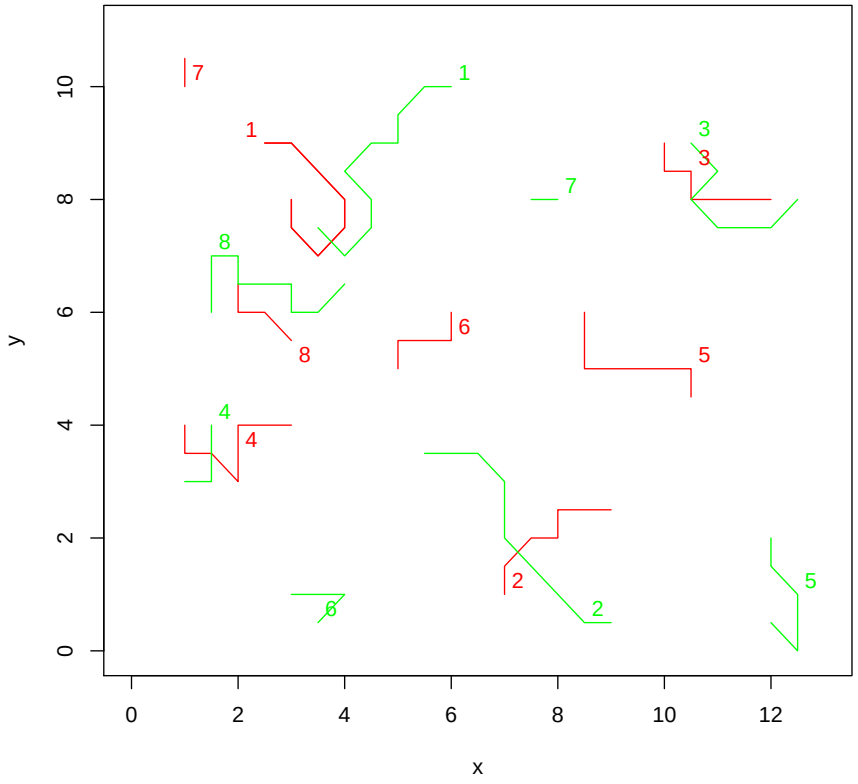

Abbildung 15: Grafische Darstellung der Modifikation der 16 simulierten Tracks (die farbliche Kennzeichnung entspricht dabei der Zugehörigkeit zu Protein 1 (rot) bzw. 2 (grün) sowie die Zahl an den entsprechenden Proteintracks einer möglichen Zuordnung dient).

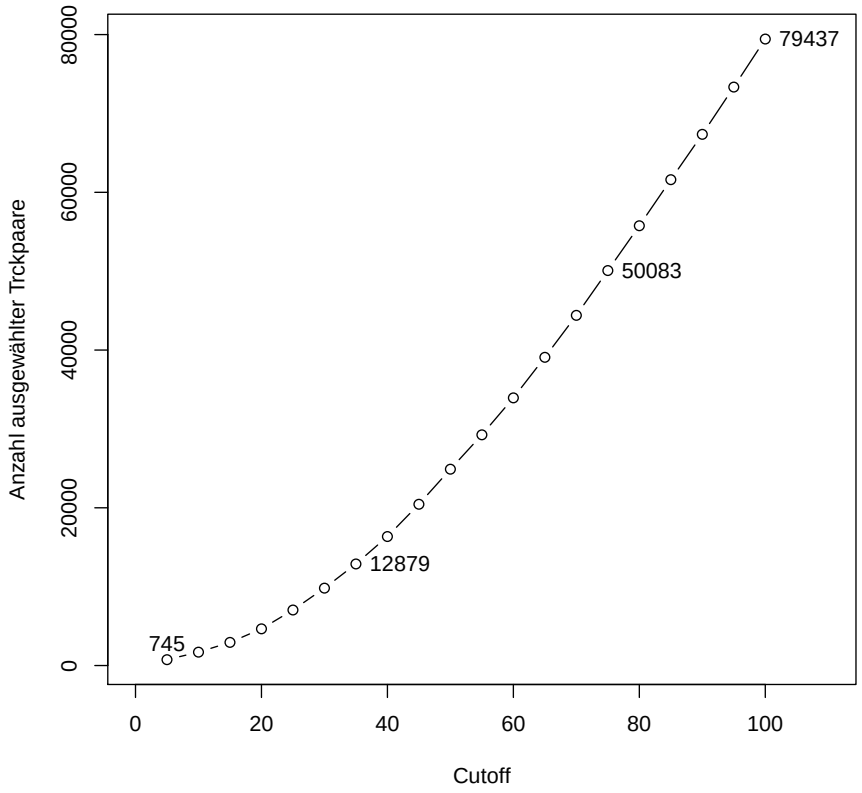

Abbildung 16: Anzahl ausgewählter Trackpaare in Abhängigkeit des Cutoffs für eine Zelle aus dem Experiment 0 Minuten nach Stimulation.

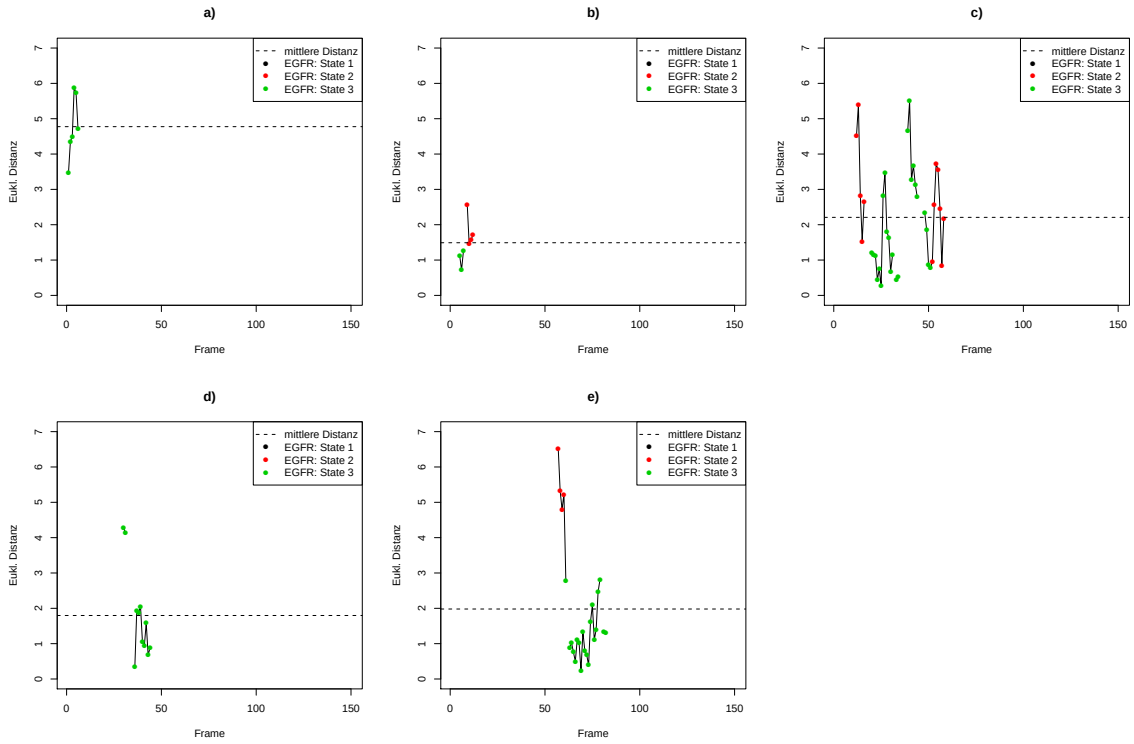

Abbildung 17: Distanzen in Abhängigkeit des Frames von Trackpaaren mit einer mittleren Distanz kleiner 5 nm; hier EGFR-Tracks 3 mit PTB-Track 13 (a)), EGFR-Track 3 und PTB-Track 112 (b)), EGFR-Track 3 und PTB-Track 158 (c)), EGFR-Track 3 und PTB-Track 284 (d)) und EGFR-Track und PTB-Track 459 (e)).

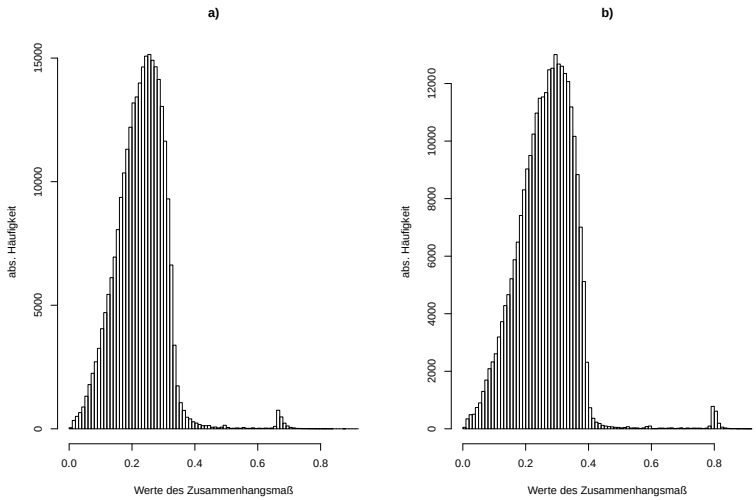

Abbildung 18: Histogram der berechneten Zusammenhänge der experimentellen Daten unter Verwendung der Gewichtung $w_1 = w_2 = w_3 = 1/3$ (a)) und $w_1 = w_2 = 0.4, w_3 = 0.2$ (b)).

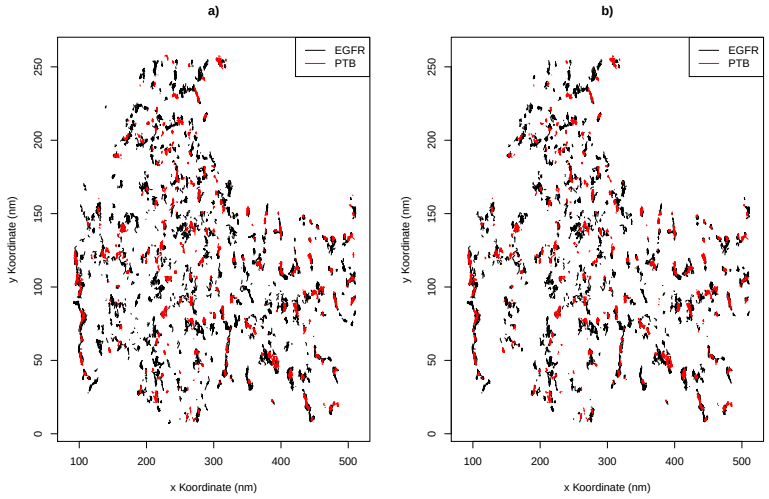

Abbildung 19: Übersicht der Trackauswahl 0 Minuten nach Stimulation bei einer Gewichtung von $w_1 = w_2 = 0.4, w_3 = 0.2$ ($\epsilon = 10$ und $MinPts = 2$) und dem 99%-Quantil als Cutoff: a) Darstellung des Datensatzes mit allen enthaltenden Tracks; b) Darstellung der Tracks, welche zu einem Trackpaar mit einem Zusammenhang $\geq$ dem 99%-Quantil gehören.

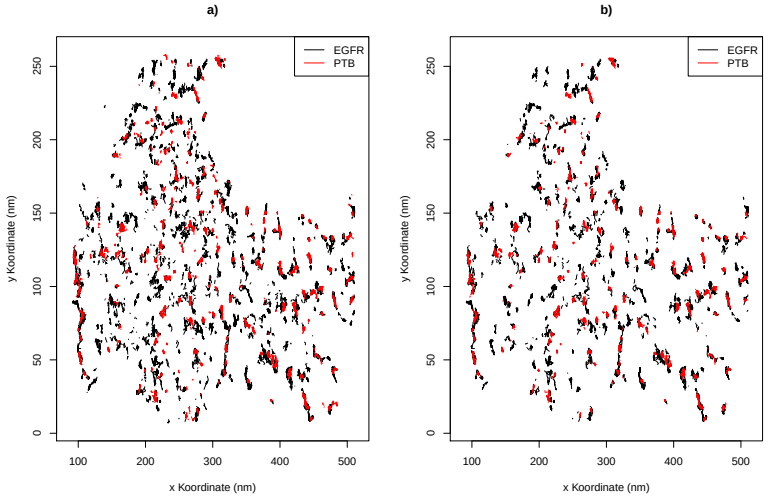

Abbildung 20: Übersicht der Trackauswahl 0 Minuten nach Stimulation bei einer Gewichtung von $w_1 = w_2 = 0.4, w_3 = 0.2$ ($\epsilon = 40$ und $MinPts = 5$) und dem 95%-Quantil als Cutoff: a) Darstellung des Datensatzes mit allen enthaltenden Tracks; b) Darstellung der Tracks, welche zu einem Trackpaar mit einem Zusammenhang $\geq$ dem 95%-Quantil gehören.

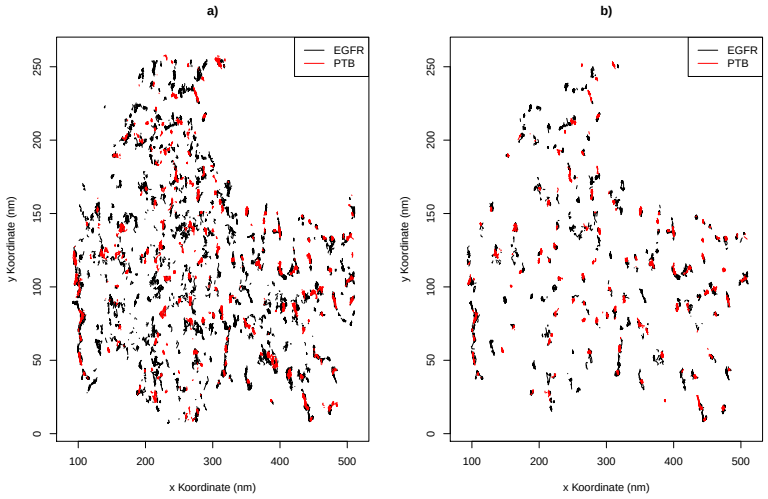

Abbildung 21: Übersicht der Trackauswahl 0 Minuten nach Stimulation bei einer Gewichtung von $w_1 = w_2 = w_3 = 1/3$ ($\epsilon = 40$ und $MinPts = 5$) und dem 99%-Quantil als Cutoff: a) Darstellung des Datensatzes mit allen enthaltenden Tracks; b) Darstellung der Tracks, welche zu einem Trackpaar mit einem Zusammenhang $\geq$ dem 99%-Quantil gehören.

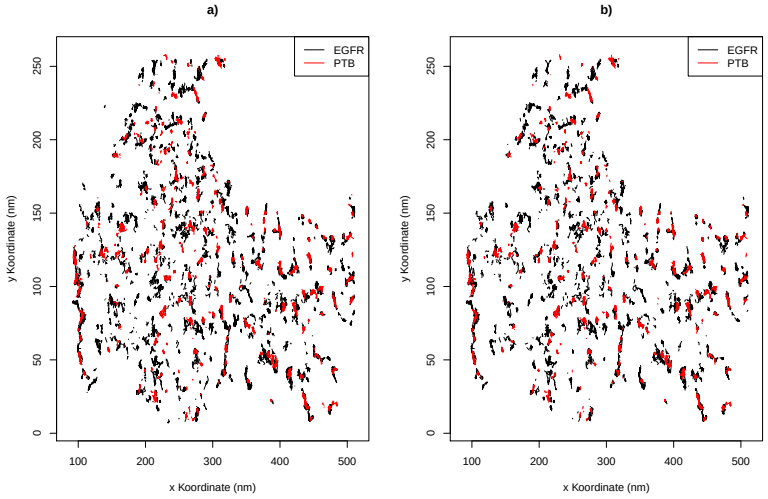

Abbildung 22: Übersicht der Trackauswahl 0 Minuten nach Stimulation bei einer Gewichtung von $w_1 = w_2 = 0.4, w_3 = 0.2$ ($\epsilon = 40$ und $MinPts = 5$) und dem 95%-Quantil als Cutoff: a) Darstellung des Datensatzes mit allen enthaltenden Tracks; b) Darstellung der Tracks, welche zu einem Trackpaar mit einem Zusammenhang $\geq$ dem 95%-Quantil gehören.

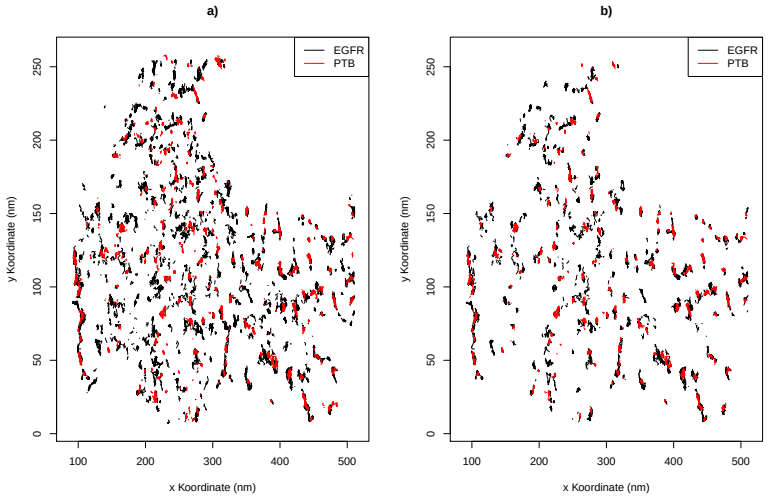

Abbildung 23: Übersicht der Trackauswahl 0 Minuten nach Stimulation bei einer Gewichtung von $w_1 = w_2 = 0.4, w_3 = 0.2$ ($\epsilon = 40$ und $MinPts = 5$) und dem 99%-Quantil als Cutoff: a) Darstellung des Datensatzes mit allen enthaltenden Tracks; b) Darstellung der Tracks, welche zu einem Trackpaar mit einem Zusammenhang $\geq$ dem 99%-Quantil gehören.

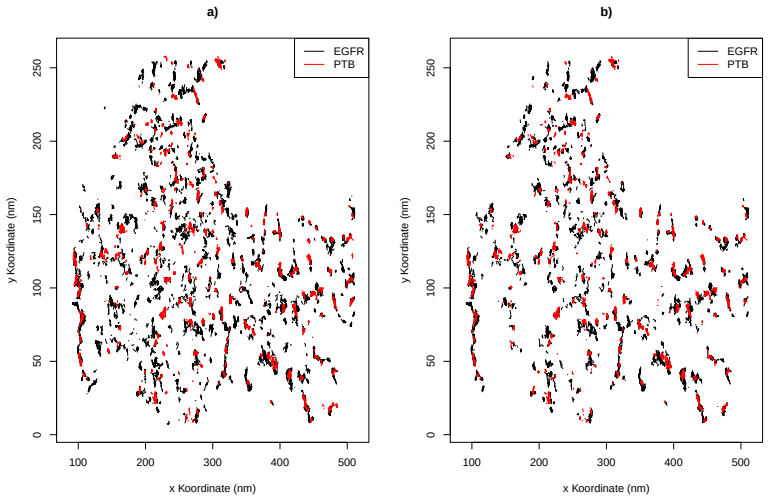

Abbildung 24: Übersicht der Trackauswahl 0 Minuten nach Stimulation bei einer Gewichtung von $w_1 = w_2 = w_3 = 1/3$ und einem Cutoff von 0.65: a) Darstellung des Datensatzes mit allen enthaltenden Tracks; b) Darstellung der Tracks, welche zu einem Trackpaar mit einem Zusammenhang ≥ 0.65 gehören.

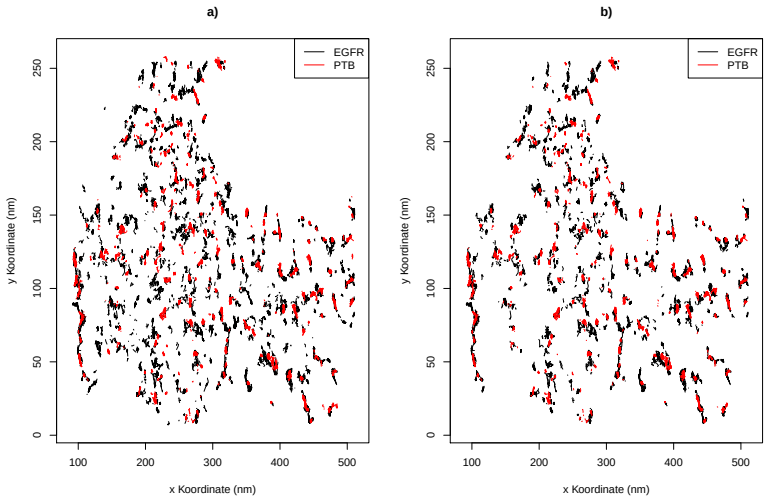

Abbildung 25: Übersicht der Trackauswahl 0 Minuten nach Stimulation bei einer Gewichtung von $w_1 = w_2 = 0.4$ und $w_3 = 0.2$ und einem Cutoff von 0.75: a) Darstellung des Datensatzes mit allen enthaltenden Tracks; b) Darstellung der Tracks, welche zu einem Trackpaar mit einem Zusammenhang ≥ 0.75 gehören.

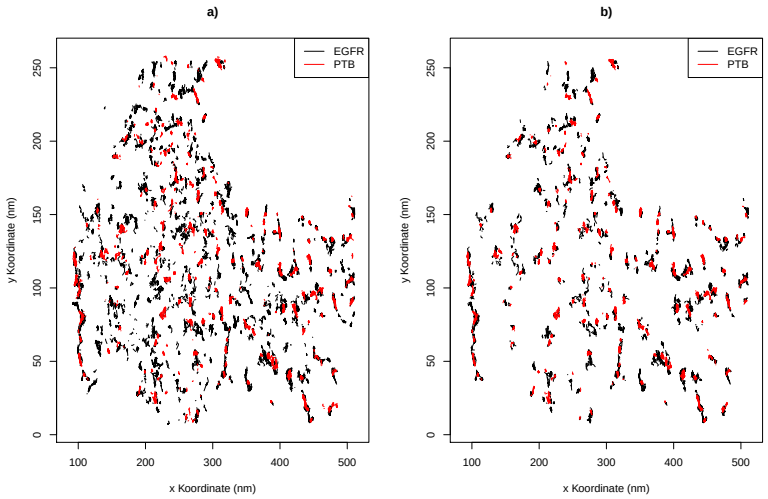

Abbildung 26: Übersicht der Trackauswahl 0 Minuten nach Stimulation bei einer Gewichtung von $w_1 = w_2 = 0.4$ und $w_3 = 0.2$ und einem Cutoff von 0.80: a) Darstellung des Datensatzes mit allen enthaltenden Tracks; b) Darstellung der Tracks, welche zu einem Trackpaar mit einem Zusammenhang ≥ 0.80 gehören.

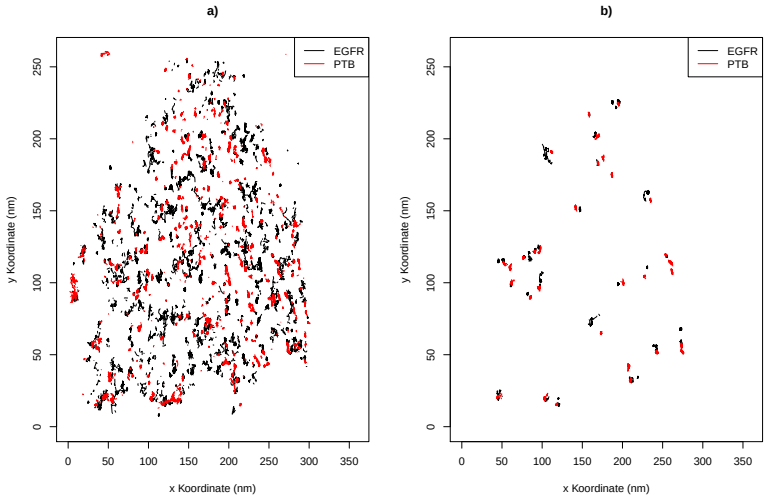

Abbildung 27: Übersicht der Trackauswahl 2 Minuten nach Stimulation bei einer Gewichtung von $w_1 = w_2 = w_3 = 1/3$ und einem Cutoff von 0.75: a) Darstellung des Datensatzes mit allen enthaltenden Tracks; b) Darstellung der Tracks, welche zu einem Trackpaar mit einem Zusammenhang ≥ 0.75 gehören.

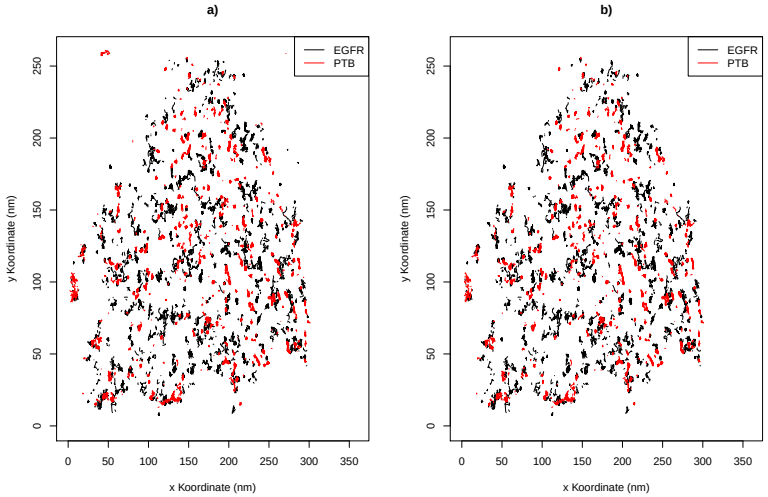

Abbildung 28: Übersicht der Trackauswahl 2 Minuten nach Stimulation bei einer Gewichtung von $w_1 = w_2 = w_3 = 1/3$ und dem 95%-Quantil als Cutoff: a) Darstellung des Datensatzes mit allen enthaltenden Tracks; b) Darstellung der Tracks, welche zu einem Trackpaar mit einem Zusammenhang $\geq$ dem 95%-Quantil gehören.

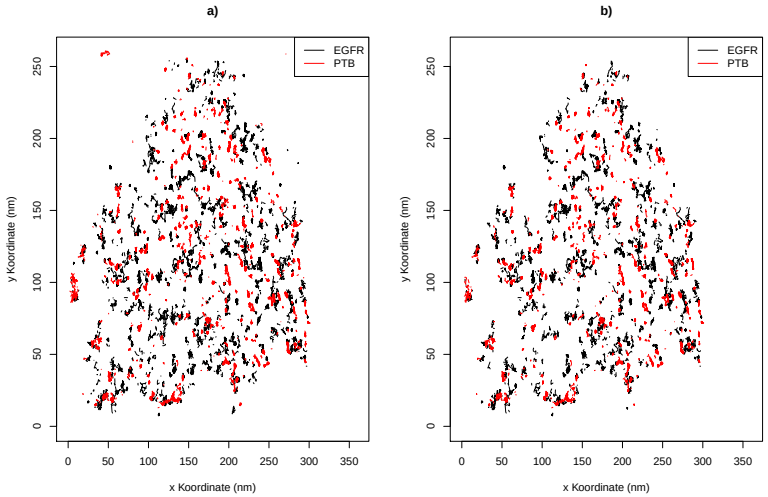

Abbildung 29: Übersicht der Trackauswahl 2 Minuten nach Stimulation bei einer Gewichtung von $w_1 = w_2 = w_3 = 1/3$ und dem 99%-Quantil als Cutoff: a) Darstellung des Datensatzes mit allen enthaltenden Tracks; b) Darstellung der Tracks, welche zu einem Trackpaar mit einem Zusammenhang $\geq$ dem 99%-Quantil gehören.

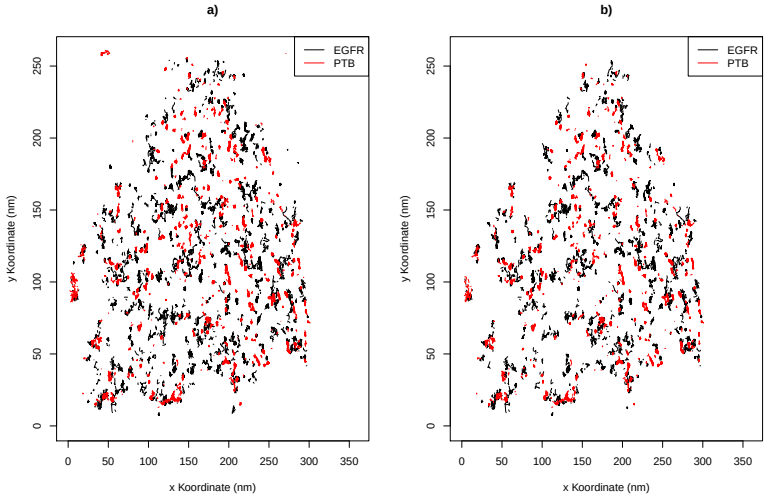

Abbildung 30: Übersicht der Trackauswahl 2 Minuten nach Stimulation bei einer Gewichtung von $w_1 = w_2 = 0.4$ und $w_3 = 0.2$ und einem Cutoff von 0.75: a) Darstellung des Datensatzes mit allen enthaltenden Tracks; b) Darstellung der Tracks, welche zu einem Trackpaar mit einem Zusammenhang ≥ 0.75 gehören.

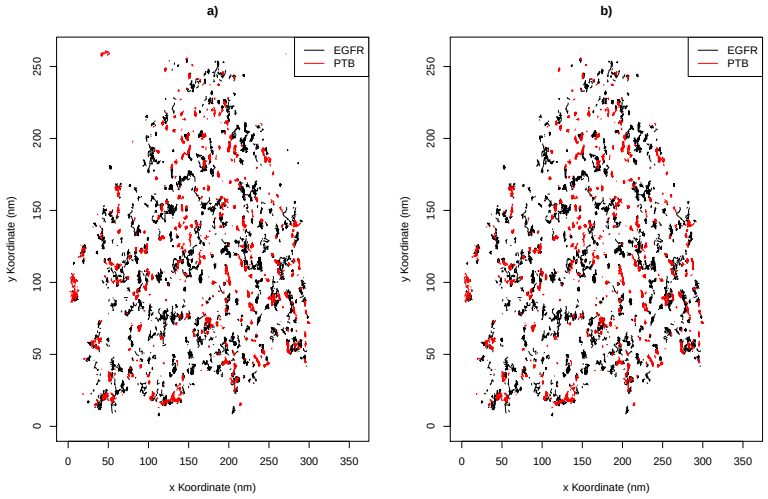

Abbildung 31: Übersicht der Trackauswahl 2 Minuten nach Stimulation bei einer Gewichtung von $w_1 = w_2 = 0.4$ und $w_3 = 0.2$ und dem 95%-Quantil als Cutoff: a) Darstellung des Datensatzes mit allen enthaltenden Tracks; b) Darstellung der Tracks, welche zu einem Trackpaar mit einem Zusammenhang $\geq$ dem 95%-Quantil gehören.

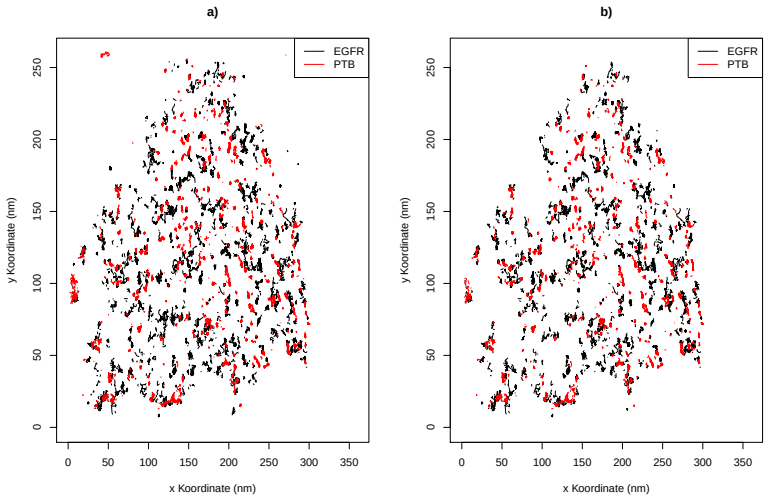

Abbildung 32: Übersicht der Trackauswahl 2 Minuten nach Stimulation bei einer Gewichtung von $w_1 = w_2 = 0.4$ und $w_3 = 0.2$ und dem 99%-Quantil als Cutoff: a) Darstellung des Datensatzes mit allen enthaltenden Tracks; b) Darstellung der Tracks, welche zu einem Trackpaar mit einem Zusammenhang $\geq$ dem 99%-Quantil gehören.

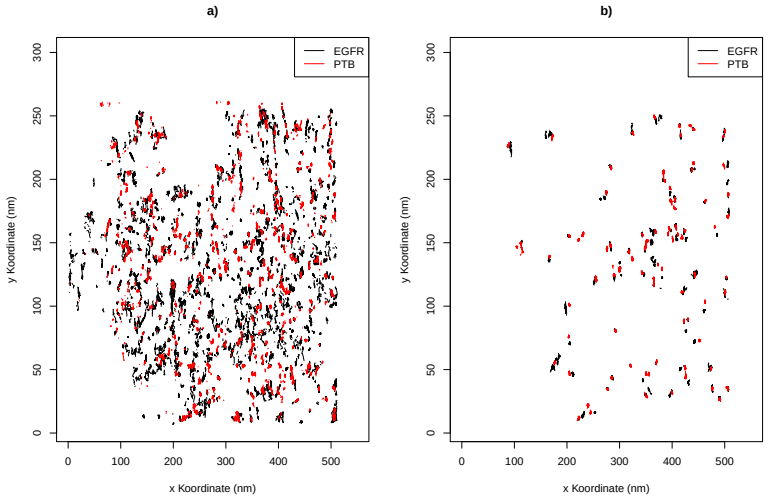

Abbildung 33: Übersicht der Trackauswahl 5 Minuten nach Stimulation bei einer Gewichtung von $w_1 = w_2 = w_3 = 1/3$ und einem Cutoff von 0.75: a) Darstellung des Datensatzes mit allen enthaltenden Tracks; b) Darstellung der Tracks, welche zu einem Trackpaar mit einem Zusammenhang ≥ 0.75 gehören.

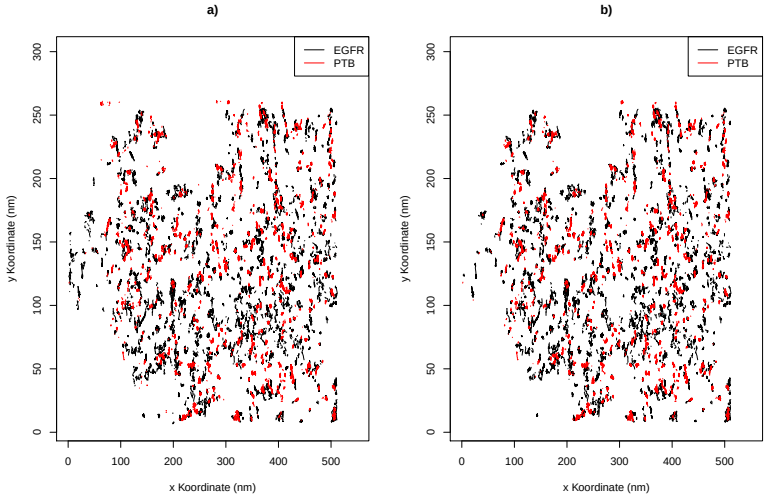

Abbildung 34: Übersicht der Trackauswahl 5 Minuten nach Stimulation bei einer Gewichtung von $w_1 = w_2 = w_3 = 1/3$ und dem 95%-Quantil als Cutoff: a) Darstellung des Datensatzes mit allen enthaltenden Tracks; b) Darstellung der Tracks, welche zu einem Trackpaar mit einem Zusammenhang $\geq$ dem 95%-Quantil gehören.

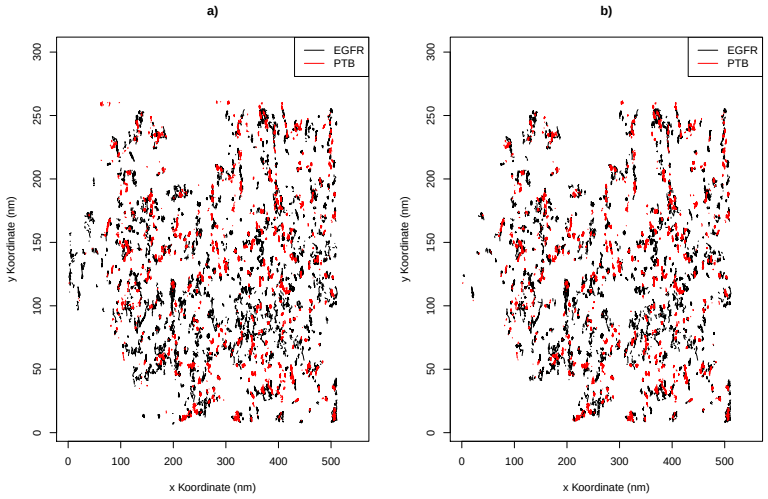

Abbildung 35: Übersicht der Trackauswahl 5 Minuten nach Stimulation bei einer Gewichtung von $w_1 = w_2 = w_3 = 1/3$ und dem 99%-Quantil als Cutoff: a) Darstellung des Datensatzes mit allen enthaltenden Tracks; b) Darstellung der Tracks, welche zu einem Trackpaar mit einem Zusammenhang $\geq$ dem 99%-Quantil gehören.

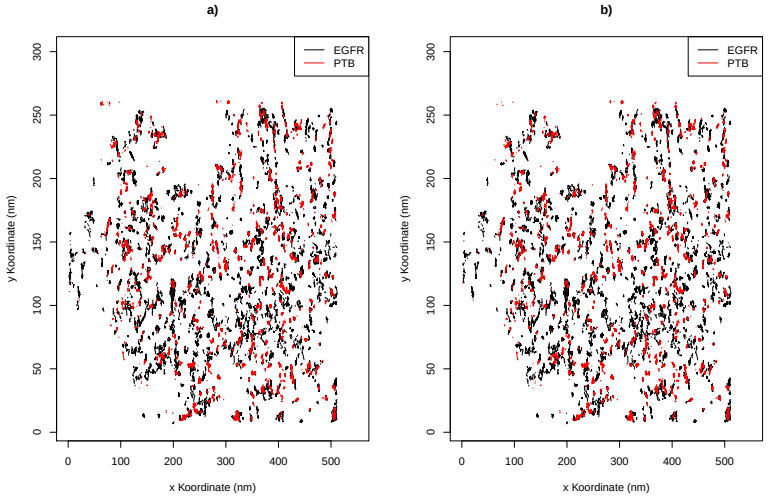

Abbildung 36: Übersicht der Trackauswahl 5 Minuten nach Stimulation bei einer Gewichtung von $w_1 = w_2 = 0.4$ und $w_3 = 0.2$ und einem Cutoff von 0.75: a) Darstellung des Datensatzes mit allen enthaltenden Tracks; b) Darstellung der Tracks, welche zu einem Trackpaar mit einem Zusammenhang ≥ 0.75 gehören.

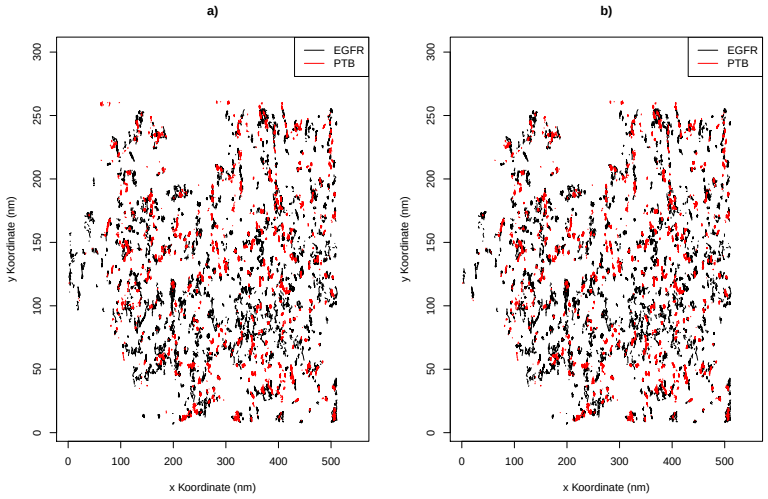

Abbildung 37: Übersicht der Trackauswahl 5 Minuten nach Stimulation bei einer Gewichtung von $w_1 = w_2 = 0.4$ und $w_3 = 0.2$ und dem 95%-Quantil als Cutoff: a) Darstellung des Datensatzes mit allen enthaltenden Tracks; b) Darstellung der Tracks, welche zu einem Trackpaar mit einem Zusammenhang $\geq$ dem 95%-Quantil gehören.

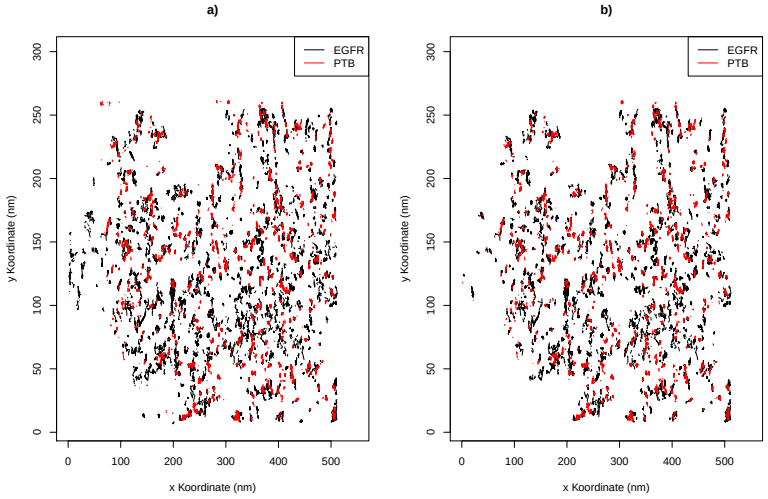

Abbildung 38: Übersicht der Trackauswahl 5 Minuten nach Stimulation bei einer Gewichtung von $w_1 = w_2 = 0.4$ und $w_3 = 0.2$ und dem 99%-Quantil als Cutoff: a) Darstellung des Datensatzes mit allen enthaltenden Tracks; b) Darstellung der Tracks, welche zu einem Trackpaar mit einem Zusammenhang $\geq$ dem 99%-Quantil gehören.

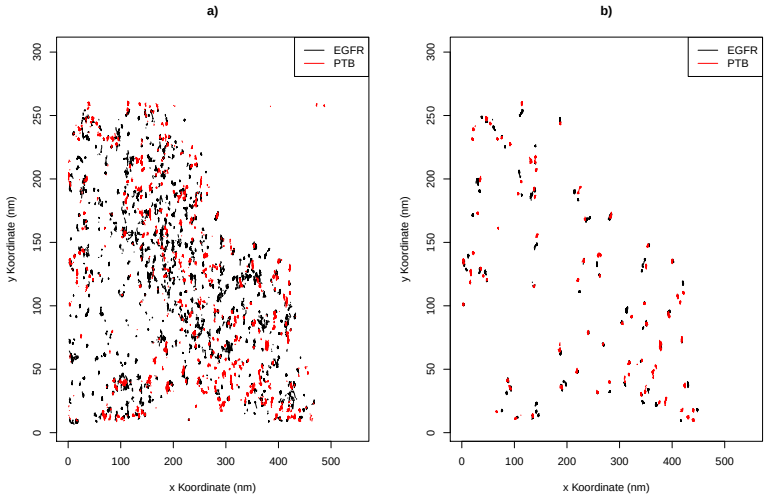

Abbildung 39: Übersicht der Trackauswahl 10 Minuten nach Stimulation bei einer Gewichtung von $w_1 = w_2 = w_3 = 1/3$ und einem Cutoff von 0.75: a) Darstellung des Datensatzes mit allen enthaltenden Tracks; b) Darstellung der Tracks, welche zu einem Trackpaar mit einem Zusammenhang ≥ 0.75 gehören.

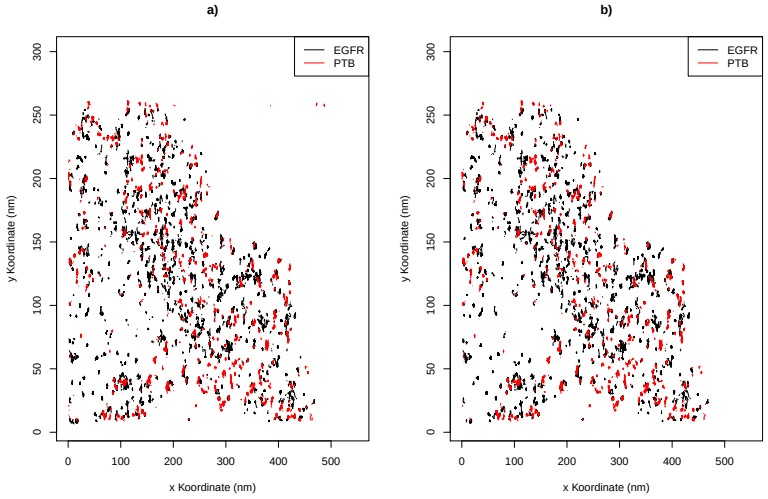

Abbildung 40: Übersicht der Trackauswahl 10 Minuten nach Stimulation bei einer Gewichtung von $w_1 = w_2 = w_3 = 1/3$ und dem 95%-Quantil als Cutoff: a) Darstellung des Datensatzes mit allen enthaltenden Tracks; b) Darstellung der Tracks, welche zu einem Trackpaar mit einem Zusammenhang $\geq$ dem 95%-Quantil gehören.

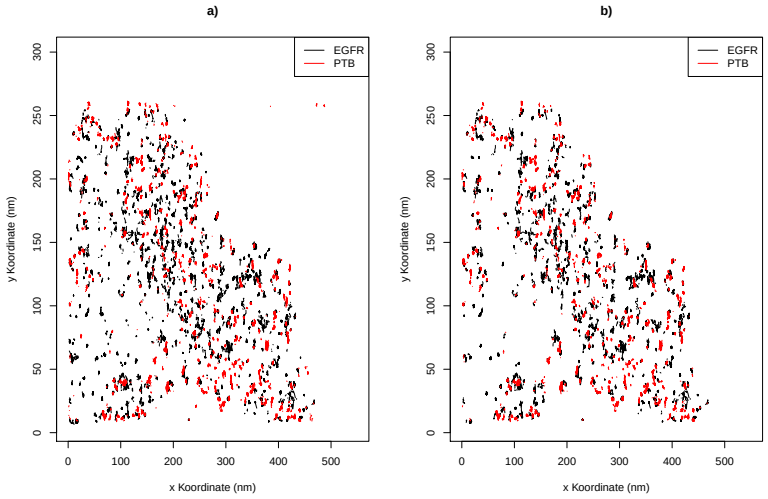

Abbildung 41: Übersicht der Trackauswahl 10 Minuten nach Stimulation bei einer Gewichtung von $w_1 = w_2 = w_3 = 1/3$ und dem 99%-Quantil als Cutoff: a) Darstellung des Datensatzes mit allen enthaltenden Tracks; b) Darstellung der Tracks, welche zu einem Trackpaar mit einem Zusammenhang $\geq$ dem 99%-Quantil gehören.

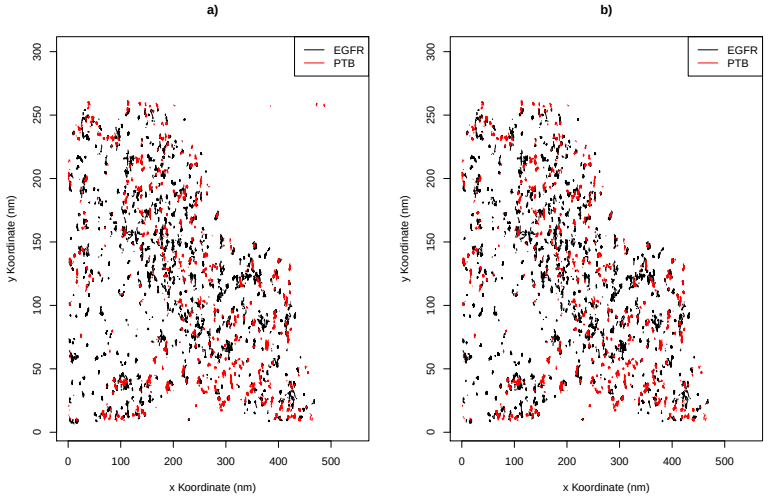

Abbildung 42: Übersicht der Trackauswahl 10 Minuten nach Stimulation bei einer Gewichtung von $w_1 = w_2 = 0.4$ und $w_3 = 0.2$ und dem 95%-Quantil als Cutoff: a) Darstellung des Datensatzes mit allen enthaltenden Tracks; b) Darstellung der Tracks, welche zu einem Trackpaar mit einem Zusammenhang $\geq$ dem 95%-Quantil gehören.

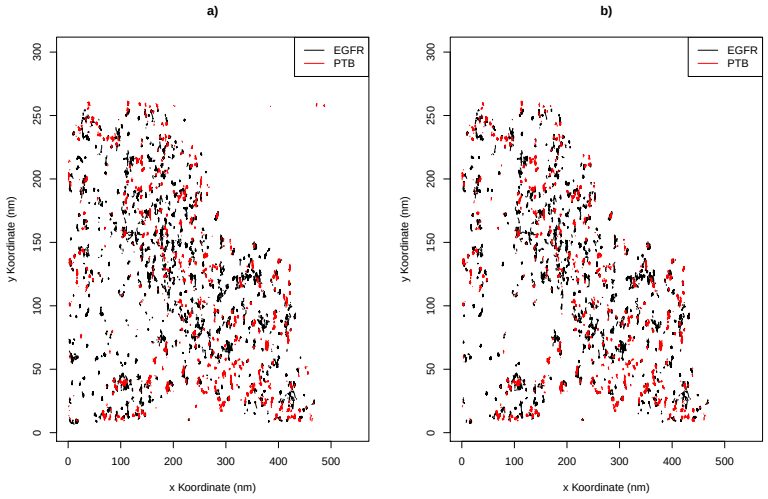

Abbildung 43: Übersicht der Trackauswahl 10 Minuten nach Stimulation bei einer Gewichtung von $w_1 = w_2 = 0.4$ und $w_3 = 0.2$ und dem 99%-Quantil als Cutoff: a) Darstellung des Datensatzes mit allen enthaltenden Tracks; b) Darstellung der Tracks, welche zu einem Trackpaar mit einem Zusammenhang $\geq$ dem 99%-Quantil gehören.

D. Abkürzungsverzeichnis

ASH	Average Shifted Histogram
BAC	Bacterial Artificial Chromosome
Be(μ, ρ)	Beta-Verteilung nach Robert und Rousseau mit Parametern μ und ρ
Bernoulli(π)	Bernoulli-Verteilung mit Erfolgswahrscheinlichkeit π
Beta(α, β)	Beta-Verteilung mit α und β
bp(s)	Basenpaar(e)
CBS	Circular Binary Segmentation
ChIP-Seq	vom englischen „Chromatin-Immunoprecipitation Sequencing“
CP	Changepoint
DBSCAN	Density-Based Spatial Clustering of Applications with Noise
DNA	Desoxyribonukleinsäure
EMM	Extensible Markov Model
Exp(λ)	Exponential-Verteilung mit Parameter λ
Gamma(a, b)	Gamma-Verteilung mit Parameter a und b
$\Gamma(x)$	Gamma-Funktion
MCMC	Markov Chain Monte Carlo
MSE	Mittlerer quadratischer Fehler
$\mathcal{N}(\mu, \sigma^2)$	Normalverteilung mit Parametern μ und σ^2
PALM	Photo Activated Localisation Microscopy
ROI	Region of Interest
TIRF	Total Internal Reflection Fluorescence
Unif(a, b)	Gleichverteilung auf dem Intervall (a, b)

CD