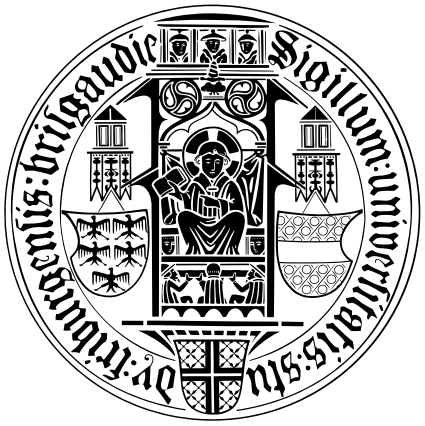
Diplomarbeit

Kinetiken von RNA-RNA Hybridisierungen

Daniel Maticzka

20. April 2009



Lehrstuhl für Bioinformatik
Institut für Informatik
Albert-Ludwigs-Universität Freiburg

Gutachter:	Prof. Dr. Rolf Backofen
	Dr. Sebastian Will
Betreuer:	Martin Mann
	Dr. Anke Busch

Inhaltsverzeichnis

1 Einleitung

1.1 Motivation

Ribonukleinsäure (RNA) ist ein Molekül mit vielfältigen Funktionen innerhalb des Stoffwechsels aller Lebewesen. Als messenger RNA (mRNA) übernimmt es beispielsweise den Transport von Gen-Transkripten zu den Ribosomen [4]. Bei der Proteinsynthese erfolgt die Zuordnung der Basentriplets der mRNA zu Aminosäuren über transfer RNAs (tRNAs) [24].

Abgesehen von seiner Rolle als Informationsbote kann RNA in Form von Ribozymen chemische Reaktionen katalysieren. Diese Eigenschaft wurde schon 1967 von Carl Woese vorgeschlagen [40]. Eines der ersten Ribozyme wurde 1968 von Thomas R. Cech entdeckt [45]. Die Entdeckung eines selbstspleißenden Exons in einer ribosomalen RNA [20] führte zu der Vermutung, dass ein ausschließlich aus RNA bestehendes System von Molekülen zur Selbstreproduktion fähig sein könnte. Von Walter Gilbert wurde 1986 die RNA-Welt Hypothese aufgestellt [11]. Nach dieser Hypothese werden auf RNA basierende selbstreproduzierende Systeme als Basis für die Evolution der Lebewesen angesehen.

Durch die Entdeckungen microRNA (miRNA) im Jahr 1993 [32] sowie der RNA-Interferenz im Jahr 1998 [34] konnte die Regulation der Proteinsynthese als weitere Funktion der RNA nachgewiesen werden. Die Forschungsergebnisse zu diesem Thema wurden 2002 vom Magazin Science als “breakthrough of the year” bezeichnet [3]. Im Jahr 2006 wurde dann für die Entdeckung der RNA-Interferenz der Nobelpreis für Medizin verliehen.

Für die Funktionalität der RNA Moleküle ist die von ihnen angenommene räumliche Struktur von zentraler Bedeutung, die sogenannte Tertiärstruktur. Im Jahr 1973 wurde über Röntgenstrukturanalyse erstmals die räumliche Struktur einer tRNA ermittelt [17]. Lange Zeit waren lediglich die räumlichen Strukturen einiger tRNAs bekannt, durch den Einsatz von Kernspinresonanzspektroskopie konnte die Anzahl bekannter Strukturen jedoch erweitert werden [36]. Die physikalische Bestimmung der Tertiärstruktur bleibt jedoch weiterhin sehr aufwendig.

Die Bestimmung der von einer RNA angenommenen Struktur ist ein wichtiges Aufgabengebiet der Bioinformatik. Ansätze zur Bestimmung der Struktur beschränken sich hauptsächlich auf die Bestimmung von Sekundärstrukturen. Für die Vorhersage dieser Strukturen sind effiziente Algorithmen verfügbar. Eine Sekundärstruktur be-

steht aus der Menge der von einer RNA ausgebildeten Basenpaarbindungen, bildet also noch nicht die vollständige räumliche Struktur ab. Die Bindungen der Sekundärstruktur sind zu einem überwiegenden Teil für die Stabilität der Struktur verantwortlich und bilden damit die Basis für die endgültige räumliche Struktur. Zur Ausbildung der räumlichen Struktur tragen nur einige wenige zusätzliche Bindungen bei. Darum wird davon ausgegangen, dass die Betrachtung von Sekundärstrukturen für die Untersuchung der Funktionalität von RNAs ausreichend exakte Ergebnisse liefert [42].

Die Ausbildung einer Sekundärstruktur wird als Faltung bezeichnet. Unter dem Begriff der Strukturvorhersage wird üblicherweise die Bestimmung energetisch optimaler Strukturen verstanden. Die Güte dieser Vorhersagen hängt von der Qualität der ermittelten Energieparameter ab [35]. Über die Berechnung der kanonischen Zustandssumme lässt sich leicht die Wahrscheinlichkeit von Strukturen bestimmen [25]. Zudem ist die Berechnung suboptimaler Strukturen möglich [43]. Diese Methoden betrachten jedoch nur die Endprodukte des Faltungsvorgangs nach Erreichen des thermodynamischen Equilibriums. Über den eigentlichen Vorgang der Faltung werden keine Aussagen gemacht.

Da viele RNA Moleküle sehr kurzlebig sind, könnte die für das Erreichen des Equilibriums benötigte Zeit die Lebenszeit des Moleküls übersteigen. Viele RNAs können von einer einmal angenommenen Struktur in eine andere wechseln. Im Fall sogenannter Riboswitches können diese Veränderungen durch externe Faktoren ausgelöst werden oder geschehen nach Ablauf einer gewissen Zeitspanne [29]. Einige RNAs bilden zu Beginn der Faltung sogenannte metastabile Zustände aus. Die anschließende Umstrukturierung zu den stabilen Strukturen des thermodynamischen Equilibriums erfolgt nur langsam. Für Viroide konnte nachgewiesen werden, dass die Funktion einer RNA sowohl von den stabilen als auch von den metastabilen Strukturen abhängen kann [31].

Die ausschließliche Betrachtung der stabilen Strukturen des thermodynamischen Equilibriums durch die Methoden der Strukturvorhersage ist zur Untersuchung dieser RNAs nicht geeignet, denn es muss die Wahrscheinlichkeit des Auftretens von Strukturen über den gesamten Zeitraum der Faltung betrachtet werden. Diese Art der Analyse wird unter dem Begriff der Kinetik zusammengefasst. Die Methoden zur Berechnung von Kinetiken bauen im Allgemeinen auf der mathematischen Modellierung des Faltungsvorgangs durch einen stochastischen Prozess auf [8]. Dieses Modell wurde erfolgreich zur Berechnung der Faltungskinetiken von RNAs eingesetzt [41].

Die bereits existierenden Ansätze zur Berechnung von Faltungskinetiken beschränken sich auf die Faltung einzelner RNA Moleküle. Interaktionen mehrerer RNAs, sogenannte Hybridisierungen, bilden die Grundlage der Funktionsweise von Ribozymen und der auf miRNA und siRNA basierenden RNA-Interferenz. Um auch das Faltungsverhalten dieser wichtigen Klasse von Strukturen betrachten zu können, sollen

die bestehenden Ansätze zur Berechnung von Kinetiken im Rahmen dieser Diplomarbeit auf Hybridisierungen von zwei RNA Molekülen erweitert werden.

1.2 Vorhergehende Arbeiten

Ein auf stochastischen Simulationen basierender Algorithmus zur Berechnung von Faltungskinetiken wurde im Jahr 2000 von Flamm et al. in [6] vorgestellt. Die elementaren Schritte zur Modellierung des Faltungsvorgangs werden bei dieser Methode durch Einfügen (Insertion) und Löschen von Basenpaaren (Deletion) sowie das Verschieben einzelner Bindungen (Shift) umgesetzt. Im Gegensatz dazu stehen Ansätze, welche den Übergang zwischen Strukturen durch Hinzufügen oder Löschen ganzer helikaler Strukturen modellieren [27]. Da Insertions, Deletions und Shifts lediglich minimale strukturelle Veränderungen bewirken, ist bei Verwendung dieser Operationen die Berechnung wirklichkeitsnäherer Faltungstrajektorien zu erwarten [7]. Leider ist die Durchführung der stochastischen Simulationen sehr rechenintensiv.

Die hier vorgestellten Methoden basieren auf dem abstrakten Modell der Energielandschaften. Energielandschaften wurde erfolgreich zur Modellierung der Faltung von Proteinen und RNA eingesetzt [5]. In einem Artikel von Flamm et al. wurde eine Partitionierung der Energielandschaft in sogenannte Makrostates vorgeschlagen [9]. Für die Identifikation der durch die Makrostates zusammengefassten Strukturen ist eine vollständige Betrachtung der Energielandschaft nötig, das sogenannte Fluten.

Wolfinger et al. präsentierten 2004 in [41] eine auf der in [6] beschriebenen Energielandschaft aufbauende Methode zur Berechnung von Kinetiken auf Basis der Übergangswahrscheinlichkeiten eines Markov-Prozesses. Die Berechnung dieser Übergangswahrscheinlichkeiten ist jedoch sehr speicherintensiv und eignet sich nur zur Betrachtung weniger Zustände. Durch die Partitionierung der Energielandschaft in Makrostates kann die Anzahl der Zustände auch für größere Energielandschaften soweit reduziert werden, dass eine Berechnung über diese Methode anwendbar ist. Vergleiche zu Kinetiken, die über die vollständige Energielandschaft berechnet wurden, zeigten nur geringe qualitative Unterschiede. Dies ermöglicht die Berechnung von Kinetiken in einem Bruchteil der für die stochastischen Simulationen benötigen Zeit.

Einen guten Überblick über die Methoden zur Berechnung von Kinetiken für RNA bietet der Übersichtsartikel [8] von Christoph Flamm und Ivo Hofacker.

Das in den oben vorgestellten Arbeiten verwendete Maß für die Stabilität einer Sekundärstruktur ist die Gibbsche freie Energie. Diese kann nach der 1981 von Michael Zuker vorgestellten Methode über die experimentell bestimmten Energiebeiträge der enthaltenen Sekundärstrukturelemente berechnet werden [46]. Durch Erweiterung der Sekundärstrukturelemente auf Bindungen zwischen zwei Nukleotidsequenzen ist

eine Berechnung der freien Energie für die Bindungen zwischen zwei RNA Molekülen möglich.

Für erste Ansätze zur Bestimmung der freien Energie von Hybridisierungen wurden lediglich die zwischen den teilnehmenden RNAs ausgebildeten Bindungen betrachtet, Sekundärstrukturen innerhalb der RNAs wurden ignoriert [14, 30]. Die 2006 von Ulrike Mückstein in [28] vorgestellte Methode zur Berechnung der Energie von Hybridisierungen berücksichtigt diese Strukturen in Form eines zusätzlichen Energiebeitrags. Die Energie berechnet sich bei diesem Ansatz aus der Energie der Strukturen der Interaktion zwischen den RNAs sowie der Energie, die nötig ist damit der Bereich dieser Interaktion innerhalb der einzelnen RNA ungebunden vorliegt.

Für Interaktionen zwischen kleinen RNAs, also miRNA oder siRNA, und mRNA im Rahmen der RNA-Interferenz konnte gezeigt werden, dass die Sekundärstruktur der mRNA einen entscheidenden Einfluss auf die Wahrscheinlichkeit einer erfolgreichen Interaktion hat [21]. Unter der Annahme, dass dieses Ergebnis auf beliebige Hybridisierungen verallgemeinert werden kann, folgt die Berechnung der Energien für diese Arbeit der von Ulrike Mückstein vorgestellten Methode.

Für die Berechnung der Energie von Hybridisierungen wird auf das Vienna RNA Package[1], eine Sammlung von Programmen zur Vorhersage von RNA Sekundärstrukturen, zurückgegriffen [14]. Unter anderem sind eine Implementation des Zuker Algorithmus zur Bestimmung der energetisch optimalen Struktur [46] sowie eine Implementation des Algorithmus von McCaskill zur Bestimmung der Wahrscheinlichkeit von Basenpaaren über die Berechnung der kanonischen Zustandsmenge [25] enthalten. Die im Vienna Package verwendeten Energieparameter entstammen aus [38] und [23].

1.3 Übersicht

In Kapitel 2 werden alle für die Beschreibung von Kinetiken nötigen Grundlagen gelegt. Nach der Definition der Energielandschaft und des auf der Energielandschaft aufbauenden Markov-Prozesses wird das für die Berechnungen der Kinetiken einzelner RNAs verwendete Modell beschrieben. In Kapitel 3 erfolgt die Erweiterung der Kinetiken von Faltungen einzelner Moleküle auf Hybridisierungen von zwei RNAs. Nach der Vorstellung der zur Berechnung der Kinetiken notwendigen Algorithmen in Kapitel 4, folgt in Kapitel 5 die Vorstellung der Ergebnisse der für diese Arbeit durchgeführten Experimente. Eine abschließende Zusammenfassung der Ergebnisse erfolgt in Kapitel 6. Dort werden auch offene Probleme angesprochen und mögliche Erweiterungen des Ansatzes diskutiert.

[1] Das Vienna RNA Package ist frei verfügbar unter http://www.tbi.univie.ac.at/RNA/.

2 Grundlagen

In diesem Kapitel wird zunächst auf das Modell der Energielandschaft, welches für die Simulation von Faltungsvorgängen Verwendung findet, eingegangen. Die Energielandschaft dient als Abstraktion des durch sie repräsentierten Vorgangs. Dann erfolgt die Definition von stochastischen Prozessen, insbesondere der Markov-Prozesse. Die Beschreibung eines Markov-Prozesses über die Zustände einer Energielandschaft ermöglicht die Berechnung von Faltungskinetiken. Nach der Vorstellung der in dieser Arbeit verwendeten Übergangsraten erfolgt die Definition einer Energielandschaft zur Berechnung von Kinetiken einzelner RNAs.

2.1 Die Energielandschaft

Eine Energielandschaft ist ein topologischer Raum, dessen Struktur hervorragend zur Modellierung von Faltungsprozessen geeignet ist. Dieses Modell wurde von Wolfinger et al. erfolgreich zur Modellierung der Faltung einzelner RNAs verwendet [41], kann aber auch zur Beschreibung weiterer physikalischer Vorgänge benutzt werden. Durch Argumentation über Elemente der Energielandschaft ist eine von der konkreten Energielandschaft unabhängige Analyse von Faltungsvorgängen möglich. Zum Abschluss dieses Abschnitts werden einige dieser Elemente der Energielandschaft definiert.

2.1.1 Aufbau der Energielandschaft

Eine Energielandschaft besteht formal aus einem Zustandsraum, einer Energiefunktion sowie einer Nachbarschaftsrelation. Diese sollen im Folgenden definiert werden.

Die Menge aller für die Modellierung eines physikalischen Vorgangs zu betrachtenden Zustände wird als Zustandsraum $\mathcal{X}$ bezeichnet.

Definition 1 (Zustandsraum $\mathcal{X}$)**.**
Bezeichne $\mathcal{X}$ den Zustandsraum einer Energielandschaft. Dieser besteht aus einer Menge von Zuständen

$$\mathcal{X} = \{x_1, \ldots, x_n\}\,.$$

Über eine Energiefunktion wird jedem Zustand der Energielandschaft eine Energie zugeordnet. Diese Energie repräsentiert die "Höhe" des Zustands in der Landschaft.

Definition 2 (Energiefunktion E)**.**
Sei $\mathcal{X}$ ein Zustandsraum. Dann gilt:

$$E : \mathcal{X} \to \mathbb{R}$$

Traversierungen einer Energielandschaft erfolgen immer über miteinander benachbarte Zustände. Zwei Zustände einer Landschaft sind miteinander benachbart, falls sie durch regelhafte strukturelle Änderungen, sogenannte Moves, aufeinander abgebildet werden können. Die Gesamtheit der verwendeten Moves wird als Move Set bezeichnet.

Definition 3 (Move, Move Set)**.**
Gegeben ein Zustand $x \in \mathcal{X}$. Bezeichne $Pot(\mathcal{X})$ die Potenzmenge des Zustandsraums. Dann ist ein ***Move*** *m eine Abbildung*

$$m : \mathcal{X} \to Pot(\mathcal{X}),$$

wobei keiner der Zustände auf sich selbst abgebildet werden darf

$$\forall x \in \mathcal{X} : m(x) \neq x.$$

Die Abbildung

$$\mathcal{N} : \mathcal{X} \to Pot(\mathcal{X})$$

wird als ***Move Set*** *der Moves $m_1, ..., m_k$ bezeichnet, falls gilt*

$$\mathcal{N}(x) = m_1(x) \uplus \ldots \uplus m_k(x).$$

Voraussetzung für die Konvergenz der später verwendeten Markov-Prozesse sind Symmetrie und Ergodizität des verwendeten Move-Set.

Definition 4 (Symmetrie eines Move-Set)**.**
Ein Move-Set $\mathcal{N}$ ist symmetrisch, falls gilt

$$\forall x, y \in \mathcal{X} : y \in \mathcal{N}(x) \Leftrightarrow x \in \mathcal{N}(y).$$

Ein Move Set wird als ergodisch bezeichnet, falls durch endlichmalige Anwendung eines darin enthaltenen Moves jeder Zustand in jeden anderen Zustand überführt werden kann.

Definition 5 (Ergodizität eines Move-Set)**.**
Ein Move-Set $\mathcal{N} : \mathcal{X} \to \mathcal{P}ot(\mathcal{X})$ ist genau dann ***ergodisch****, wenn für jedes Paar $x, y \in \mathcal{X}$ ein Pfad*

$$p = p_1 \ldots p_k,\ p_i \in \mathcal{X}$$

existiert mit

$$x = p_1 \text{ und } y = p_k$$

sowie

$$\forall i \in [2, k] \ : \ p_i \in \mathcal{N}(p_{i-1}).$$

Die durch das Move Set definierte Nachbarschaft eines Zustands kann nun zur Definition der symmetrischen Nachbarschaftsrelation herangezogen werden.

Definition 6 (Nachbarschaft, Nachbarschaftsrelation $\sim$)**.**
Gegeben ein symmetrisches Move Set $\mathcal{N}$ und Zustände $x, y \in \mathcal{X}$.
Die Teilmenge $\mathcal{N}(x) \subseteq \mathcal{X}$ wird als ***Nachbarschaft*** *des Zustands x bezeichnet.*
Die symmetrische ***Nachbarschaftsrelation*** *$\sim$ ist definiert durch*

$$x \sim y \Leftrightarrow [(x \in \mathcal{N}(y)) \wedge (y \in \mathcal{N}(x))]$$

Aufbauend auf diesen Elementen folgt nun die formale Definition der Energielandschaft.

Definition 7 (Energielandschaft)**.**
Gegeben einen Strukturraum $\mathcal{X}$, eine Energiefunktion E und eine symmetrische Nachbarschaftsrelation $\sim$. Dann ist die ***Energielandschaft*** *definiert durch das Tripel*

$$(\mathcal{X}, E, \sim).$$

2.1.2 Elemente der Energielandschaft

Wenn die Energiefunktion mehreren Zuständen der Energielandschaft den gleichen Wert zuweist, gilt diese als degeneriert.

Definition 8 (Degeneriertheit)**.**
Wenn gilt

$$\exists x \in X, \exists y \in \mathcal{X} : x \neq y \ \wedge \ E(x) = E(y)$$

wird die Energielandschaft als ***degeneriert*** *bezeichnet.*

Dies ist für die in dieser Arbeit verwendete Energiefunktion der Fall, es wird also ausschließlich mit degenerierten Energielandschaften gearbeitet. Für die folgende Definition der Minima ist jedoch eigentlich eine eindeutige Ordnung der Zustände anhand ihrer Energie nötig. Diese ist aber ohne weiteres nur für nicht degenerierte Energielandschaften möglich. Um dennoch eine eindeutige Ordnung zu definieren, wird als Zusatzkriterium die lexikographische Ordnung bezüglich einer eindeutigen Zeichenkettenrepräsentation der Zustände verwendet.

Definition 9 (Ordnung $<$)**.**
Sei $ZK(x)$ die eindeutige Zeichenkettenrepräsentation des Zustands $x \in X$. Sei $<_{lex}$ eine Ordnungsrelation über diese Zeichenketten.
Dann ist die Ordnungsrelation $<$ für zwei Zustände x und y einer Energielandschaft definiert durch

$$x < y \ \Leftrightarrow \ E(x) < E(y) \ \vee \ (E(x) = E(y) \ \wedge \ ZK(x) <_{lex} ZK(y))$$

und

$$x > y \ \Leftrightarrow \ x \neq y \ \vee \ \neg(x < y).$$

Diese Ordnung ist eine strenge Totalordnung, zwischen zwei Zuständen gilt also immer genau eine der Beziehungen

$$x < y, \ x > y \text{ oder } x = y.$$

Als Minimum wird eine Struktur bezeichnet, die keinen kleineren Nachbarn nach der totalen Ordnung $<$ besitzt.

Definition 10 (lokales Minimum)**.**
Ein Zustand $\hat{x} \in \mathcal{X}$ mit der Eigenschaft

$$\nexists\, x \in \mathcal{N}(\hat{x}) : x < \hat{x}$$

*ist ein **lokales Minimum**.*

Das kleinste lokale Minimum einer Energielandschaft wird als globales Minimum bezeichnet.

Definition 11 (globales Minimum, mfe)**.**
*Ein Zustand $\hat{x} \in \mathcal{X}$ ist das **globale Minimum**, wenn gilt*

$$\nexists x \in \mathcal{X} \;:\; x < \hat{x}.$$

*Die Energie des globalen Minimums wird als **minimum free energy (mfe)** bezeichnet.*

Für die spätere Partitionierung der Zustände der Landschaft in Makrostates ist vor allem der Begriff des Basins (Talkessel) wichtig. Der tiefste Punkt eines Basins ist ein Minimum und jeder Zustand der Energielandschaft gehört genau einem Basin an. Die Zugehörigkeit eines Zustands zu einem Basin wird über die Methode des steilsten Abstiegs, den Gradient Walk, definiert. Bei dieser Wanderung durch die Energielandschaft wird so lange der kleinste Nachbar eines Zustands bezüglich der Ordnung $<$ als Nachfolger ermittelt, bis die Wanderung in einem Minimum endet. Der Zustand von dem diese Wanderung ausgeht, wird dem Basin dieses Minimums zugeordnet. Im Folgenden sollen Walk und Gradient-Walk formal definiert werden.

Unter einem Walk wird die Traversierung von Zuständen einer Energielandschaft verstanden. Zwei aufeinanderfolgende Zustände eines Walks müssen miteinander benachbart sein.

Definition 12 (Walk)**.**
*Ein **Walk** w auf dem Zustandsraum einer Energielandschaft ist eine Folge von Zuständen*

$$w = (x_1, x_2, \ldots, x_n), x_i \in \mathcal{X}$$

für die gilt

$$\forall i \in [2, n] \;:\; x_{i+1} \sim x_i).$$

Verschiedene Typen von Walks unterscheiden sich durch die Art der Auswahl des Folgezustands x_{i+1} bei gegebenem Zustand x_i sowie der Bestimmung des Endzustands, nach dessen Erreichen keine weiteren Zustände mehr besucht werden. Bei

einem Random-Walk auf einer Energielandschaft erfolgt die Auswahl des Nachfolgezustands beispielsweise zufällig. Der in Abschnitt 4.1 definierte Rejectionless Monte-Carlo-Algorithmus gehört zu der Klasse der Random Walks.

Ein Gradient Walk folgt immer der Richtung des steilsten Abstiegs. So lange ein Zustand kleinere Nachbarn bezüglich der Ordnung $<$ besitzt, wird von diesen der kleinste als Nachfolgezustand ausgewählt. Existieren keine Nachbarn kleinerer Ordnung, endet der Walk. Aufgrund der Definition eines lokalen Minimums als Zustand ohne Nachbarn kleinerer Ordnung ist dieser Endzustand ein lokales Minimum. Der Gradient-Walk endet also immer in einem lokalen Minimum.

Definition 13 (maximal expandierter Gradient Walk)**.**
Sei $\hat{w}$ ein Walk

$$\hat{w} = (x_1, x_2, \ldots, x_n),\ x_i \in \mathcal{X}$$

mit zusätzlicher Eigenschaft

$$\forall i \in [1, n-1]\ :\ x_{i+1} \sim x_i\ \wedge\ \nexists\, y \sim x_i\ :\ y < x_{i+1}.$$

*Dann wird $\hat{w}$ als **Gradient Walk** bezeichnet. Falls zusätzlich gilt*

$$x_n \text{ ist lokales Minimum,}$$

*der Walk also in einem lokalen Minimum endet, gilt er als **maximal expandiert**. Bezeichne desweiteren*

$$\hat{w}(x) = x_n$$

den Endzustand des mit Zustand $x \in \mathcal{X}$ beginnenden, vollständig expandierten Gradient Walks.

Zu jedem Minimum einer Energielandschaft existiert ein zugehöriges Basin. Ein Zustand ist Teil des einem Minimum zugeordneten Basins, falls ein bei dieser Struktur gestarteter und vollständig expandierter Gradient Walk in diesem Minimum endet.

Definition 14 (Basin eines Minimums B)**.**
*Sei $\hat{x} \in \mathcal{X}$ ein Minimum und $\hat{w}$ ein Gradient-Walk auf $\mathcal{X}$. Dann ist das **Basin eines Minimums** definiert durch die Menge*

$$B(\hat{x}) = \{x\ :\ x \in \mathcal{X}\ \wedge\ \hat{w}(x) = \hat{x}\}\,.$$

Sei $M \subseteq \mathcal{X}$ die Menge der Minima. Dann ist die Menge der Basins einer Energielandschaft gegeben durch

$$\mathcal{B}(\mathcal{X}) = \{B(\hat{x})\ :\ \hat{x} \in M\}\,.$$

Sind zwei miteinander benachbarte Zustände Teil verschiedener Basins, gehören sie zu der Kontaktfläche zwischen den jeweiligen Basins.

Definition 15 (Kontaktfläche C)**.**
Seien $\hat{x}, \hat{y} \in \mathcal{X}$ Minima einer Energielandschaft.
Dann ist die Kontaktfläche zwischen den zugehörigen Basins $B(\hat{x})$ und $B(\hat{y})$ definiert durch die Menge der Paare

$$C_{\hat{x}\hat{y}} = \{(x, y) \mid x \in B(\hat{x}), y \in B(\hat{y}) \ : \ x \sim y\}.$$

Oftmals ist eine vollständige Betrachtung aller Zustände einer Energielandschaft nicht durchführbar. In diesem Fall ist es möglich, Teilmengen des Zustandsraums auf eine kleinere Menge von Zuständen abzubilden. Diese neuen Zustände werden als Makrostates bezeichnet. Zur besseren Unterscheidung werden die ursprünglichen Zustände dann auch Mikrostates genannt. Um sicherzustellen, dass auch bei Nutzung von Makrostates alle Zustände in die Berechnung einfließen, müssen die Makrostates eine Partitionierung des Zustandsraums bilden. Für diese Arbeit werden die Basins der Minima als Makrostates verwendet.

Definition 16 (Partitionierung des Zustandsraums)**.**
Sei $M \subseteq \mathcal{X}$ die Menge der Minima der Energielandschaft $(\mathcal{X}, E, \sim)$.
Dann bilden die Basins der Minima eine Partitionierung des Zustandsraums

$$\mathcal{X} = \biguplus_{\hat{x} \in M} B(\hat{x}).$$

Eine anschauliche Darstellung der Basins einer Energielandschaft zeigt Abbildung 2.1.

2.2 Stochastische Prozesse

Zur Beschreibung von zeitlich geordneten Vorgängen, deren Verhalten durch zufällige Ereignisse bestimmt wird, eignet sich das mathematische Modell der stochastischen Prozesse. Im Folgenden soll zunächst die Theorie der stochastischen Prozesse erläutert werden. Dies führt zur Definition der für diese Arbeit verwendeten zeithomogenen Markov-Prozesse. Abschließend folgt eine Analyse der Konvergenz stochastischer Prozesse.

2.2.1 Kategorisierung stochastischer Prozesse

Stochastische Prozesse können nach der Art der repräsentierten Zustände sowie der verwendeten Zeiteinheiten kategorisiert werden. Bei einer abzählbaren Menge von Zuständen spricht man von einem wertdiskreten Prozess. Sind die betrachteten Zeiteinheiten abzählbar, wird der Prozess als zeitdiskret, ansonsten als zeitstetig bezeichnet.

Definition 17 (Stochastischer Prozess)**.**
Gegeben eine Zustandsmenge

$$X = \{x_1, \ldots, x_n\}.$$

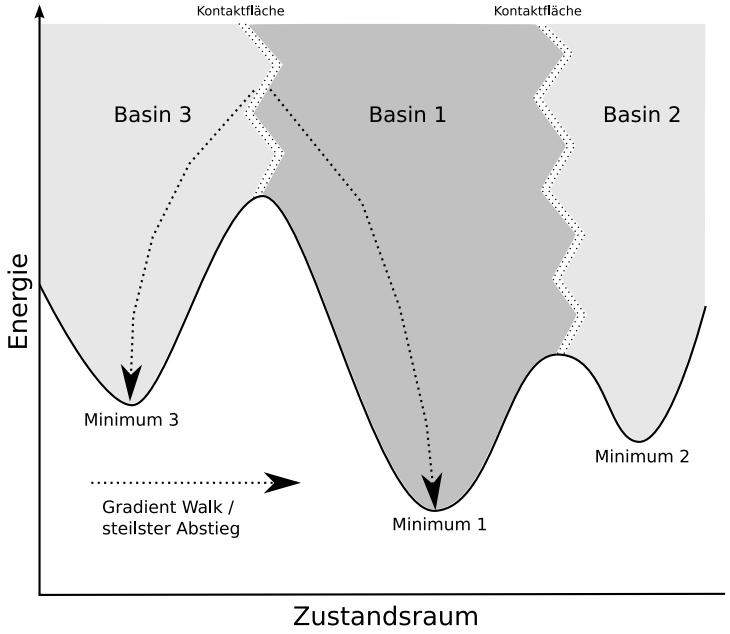


Abbildung 2.1: Idealisierte Darstellung von drei Basins einer Energielandschaft. Die Basins sind aufsteigend nach der Energie ihres zugehörigen Minimums nummeriert. Jeder von einem Zustand des Basins ausgehende und maximal expandierte Gradient-Walk endet in dem Minimum des Basins. Zwischen den Basins 3 und 1 sowie zwischen 1 und 2 sind Kontaktflächen aufgetragen. Diese bestehen aus Paaren miteinander benachbarter Zustände, die verschiedenen Basins zugeordnet sind.

Sei die Wahrscheinlichkeit des Auftretens eines Zustands $x \in X$ zu Zeitpunkt t gegeben durch

$$\mathrm{Pr}_x(t) = \mathrm{Pr}(\text{ “Zustand x zu Zeitpunkt t“}).$$

Gegeben eine Startverteilung

$$\pi^0 = (\pi_1^0, \ldots, \pi_n^0)^T$$

mit

$$\pi_x^0 = \mathrm{Pr}_x(0)$$

sowie eine Matrix von Übergangswahrscheinlichkeiten

$$P(t) = \begin{pmatrix} p_{11}(t) & \cdots & p_{1n}(t) \\ \vdots & \ddots & \vdots \\ p_{n1}(t) & \cdots & p_{nn}(t) \end{pmatrix}$$

mit

$$p_{xy}(t) = \mathrm{Pr}(\text{ ”Übergang von Zustand x nach Zustand y in Zeit t”}).$$

Dann definiert das Tripel

$$\left(X, \pi^0, P(t)\right)$$

einen ***Stochastischen Prozess****.*

2.2.2 Zeithomogener Markov-Prozess

Im Folgenden soll eine Zusatzannahme über die Wahrscheinlichkeit des Übergangs zwischen Zuständen gemacht werden, die Markov-Eigenschaft. Erfüllt ein stochastischer Prozess diese Eigenschaft, so ist die Auftretenswahrscheinlichkeit eines Zustands lediglich von dessen Vorgänger und nicht von der vollständigen Historie des Prozesses abhängig.

Ein stochastischer Prozess mit Markov-Eigenschaft wird im zeitdiskreten Fall als Markov-Kette, im zeitstetigen Fall als Markov-Prozess bezeichnet. Ist zusätzlich der Übergang zwischen Zuständen unabhängig vom Zeitpunkt dieses Übergangs wird der Markov-Prozess als zeithomogen bezeichnet. In dieser Arbeit werden ausschließlich zeithomogene Markov-Prozesse verwendet.

Definition 18 (Zeithomogener Markov-Prozess)**.**
Sei

$$\left(X, \pi^0, P(t)\right)$$

ein zeitkontinuierlicher stochastischer Prozess mit endlicher Zustandsmenge

$$X = \{x_1, \ldots, x_n\}\,.$$

Sei $Z(t)$ eine Zufallsvariable für einen Zustand zum Zeitpunkt t. Sei die ***Markov-Eigenschaft***

$$\forall n \in \mathbb{N},\ 0 < t_1 < \ldots < t_n,\ s > 0,\ i_1, \ldots, i_n, j \in X\ :$$

$$\Pr[Z(t_n) \mid Z(t_{n-1}) = i_{n-1},\ \ldots\ , Z(t_1) = i_1] =$$
$$\Pr[Z(t_n) \mid Z(t_{n-1}) = i_{n-1}]$$

erfüllt. Gelte zusätzlich ***Zeithomogenität*** *für die Übergangswahrscheinlichkeiten*

$$\begin{aligned}\Pr[Z(t+s) = j \mid Z(s) = i] &= \Pr[Z(t) = j \mid Z(0) = i] \\ &= p_{ij}(t).\end{aligned}$$

Dann ist dieser stochastische Prozess ein ***zeithomogener Markov-Prozess****.*

Die Bestimmung der Übergangswahrscheinlichkeiten soll nun von den sogenannten Übergangsraten abhängig sein. Eine solche Übergangsrate gibt die Stärke des Flusses von einem Zustand zu einem anderen an. Demzufolge ist der Fluss und somit die Übergangsrate zwischen nicht benachbarten Zuständen 0. Für die Berechnung

der Übergangsrate zwischen zwei benachbarten Zuständen existiert eine Vielzahl von Berechnungsmöglichkeiten, von denen in dieser Arbeit die Methoden nach Metropolis und nach Kawasaki verwendet und verglichen werden. Der Zusammenhang zwischen Übergangswahrscheinlichkeiten und Übergangsraten wird durch die sogenannte Master-Equation definiert.

Definition 19 (Master Equation für $\Pr_x(t)$)**.**
*Sei $\Pr_x(t)$ die Wahrscheinlichkeit eines Zustands x zum Zeitpunkt t. Bezeichne k_{xy} die Rate für den Übergang von Zustand x zu Zustand y. Dann soll der Markov-Prozess durch die **Master Equation** für die Wahrscheinlichkeit eines Zustands x zum Zeitpunkt t*

$$\frac{\partial \Pr_x(t)}{\partial t} = \sum_{y \neq x} [\Pr_y(t) k_{yx} - \Pr_x(t) k_{xy}] \tag{2.1}$$

bestimmt sein.

Definition 20 (Lösung der Master Equation)**.**
Gegeben eine Matrix von Übergangsraten

$$K = \begin{pmatrix} k_{11} & \dots & k_{1n} \\ \vdots & \ddots & \vdots \\ k_{n1} & \dots & k_{nn} \end{pmatrix}$$

und den Vektor der Verteilung der Zustände zum Zeitpunkt t

$$\overrightarrow{P_t} = (\Pr_1(t), \dots, \Pr_n(t))^T .$$

Dann kann die Master Equation als Vektormultiplikation

$$\frac{\partial}{\partial t} \overrightarrow{P_t} = K \overrightarrow{P_t} \tag{2.2}$$

*ausgedrückt werden. Dies führt zu **Lösung der Master Equation**:*

$$\overrightarrow{P_t} = e^{tK} \overrightarrow{P_0} = e^{tK} \pi^0 \tag{2.3}$$

2.2.3 Konvergenz des Markov-Prozesses

Im Folgenden soll näher auf die Konvergenz von Markov-Prozessen zu einer stationären Verteilung eingegangen werden.

Definition 21 (stationäre Verteilung)**.**
Bezeichne $\Pr_x(t)$ die Wahrscheinlichkeit eines Zustands $x \in X$ zu Zeitpunkt t. Dann sei der Spaltenvektor

$$\pi^* = (\pi_1^*, \dots, \pi_n^*)^T$$

*die **stationäre Verteilung**, falls für alle Zustände $x \in X$ gilt*

$$\lim_{t \to \infty} \Pr_x(t) = \pi_x^*.$$

Für die stationäre Verteilung gilt also

$$\pi^* = \pi^* P(t).$$

Nach der Markov-Theorie besitzt ein Markov-Prozess genau eine stationäre Verteilung, falls sie irreduzibel ist und das Kriterium der Detailed Balance erfüllt. Irreduzibilität bedeutet, dass der Prozess nicht in mehrere voneinander unabhängige Prozesse aufgeteilt werden kann.

Definition 22 (Irreduzibilität)**.**
Ein Markov-Prozess heißt ***irreduzibel****, wenn gilt*

$$\forall x, y \in X : \exists t \leq 0 : p_{xy}(t) > 0,$$

also jeder Zustand von jedem anderen Zustand aus erreicht werden kann.

Diese Definition steht also in unmittelbarem Zusammenhang mit der Ergodizität des verwendeten Move Set. Wäre das verwendete Move Set nicht ergodisch, dann gäbe es Zustände, die von einer oder mehreren Teilmengen des Zustandsraums nicht erreicht werden könnten. Dadurch wäre eine Reduzierung auf einen oder mehrere voneinander unabhängige Zustandsräume möglich.

Definition 23 (Reversibilität, Detailed Balance)**.**
Ein Markov-Prozess ist ***reversibel****, falls für die stationäre Verteilung π^* und alle Zustände $x, y \in X$ gilt*

$$\pi_y^* k_{yx} = \pi_x^* k_{xy}. \tag{2.4}$$

Diese Gleichung wird auch als ***Detailed Balance*** *bezeichnet.*

Das Kriterium der Detailed Balance ist also sowohl von der stationären Verteilung als auch von der verwendeten Funktion zur Berechnung der Übergangsraten abhängig.

Definition 24 (Existenz einer stationären Verteilung)**.**
Ein irreduzibler Markov-Prozess der das Kriterium der Detailed Balance erfüllt, besitzt eine ***eindeutige stationäre Verteilung*** *π^* und es gilt*

$$\lim_{t \to \infty} P_x(t) = \pi_x^*$$

für jede Startverteilung π^0.

Der in dieser Arbeit verwendete stochastische Prozess soll eine stationäre Verteilung erreichen, die der Boltzmann Verteilung entspricht. Darum ist die Erfüllung der Detailed Balance für die Boltzmann Verteilung und die verwendeten Übergangsraten zu zeigen. Dieser Beweis folgt nach der Definition der Übergangsraten im folgenden Abschnitt.

2.2.4 Markov-Prozesse über eine Energielandschaft

In Verbindung mit einer Energielandschaft kann ein Markov-Prozess zur Berechnung von Kinetiken benutzt werden. Für kleine Mengen von Zuständen kann das Verhalten des Prozesses über beliebig lange Zeiträume leicht über die Lösung der Gleichung 2.3 berechnet werden. Bei praktisch relevanten Fällen ist jedoch meist eine Reduktion des Zustandsraums durch Verwendung von Makrostates erforderlich. Alternativ können stochastische Simulationen von Gleichung 2.1 durch Monte Carlo Algorithmen erfolgen.

Markov-Prozess über die vollständige Energielandschaft

Definition 25 (Markov-Prozess über Mikrostates)**.**
Gegeben eine Energielandschaft

$$(\mathcal{X}, E, \sim).$$

Sei π^0 eine Anfangsverteilung der Länge $|\mathcal{X}|$. Sei $P(t)$ eine Matrix von Übergangswahrscheinlichkeiten mit $|\mathcal{X}|$ Zeilen und Spalten. Dann definiert das Tripel

$$\big(\mathcal{X}, \pi^0, P(t)\big)$$

den Markov-Prozess über die Zustände der Energielandschaft.

Zur Abgrenzung des unten definierten Markov-Prozesses über die Makrostates einer Energielandschaft wird dieser auch als **Markov-Prozess über die Mikrostates** einer Energielandschaft bezeichnet.

Markov-Prozess über die Makrostates einer Energielandschaft

Definition 26 (Markov-Prozess über Makrostates)**.**
Gegeben eine Energielandschaft

$$(\mathcal{X}, E, \sim).$$

Sei $\mathcal{B}(\mathcal{X})$ die Menge der Basins der Energielandschaft. Sei π^0 eine Anfangsverteilung der Länge $|\mathcal{B}(\mathcal{X})|$. Sei $P(t)$ eine Matrix von Übergangswahrscheinlichkeiten mit $|\mathcal{B}(\mathcal{X})|$ Zeilen und Spalten. Dann definiert das Tripel

$$\big(\mathcal{B}(X), \pi^0, P(t)\big)$$

*den **Markov-Prozess über die Makrostates** der Energielandschaft. Dieser wird im Folgenden auch als Makrostate-Prozess bezeichnet.*

2.3 Übergangsraten

Für diese Arbeit wurden zwei Übergangsraten implementiert und untersucht. Die Metropolis Übergangsrate, auch Metropolis Kriterium genannt, wurde von Nicholas Metropolis in [26] eingeführt. Die Kawasaki Übergangsrate wurde in [16] definiert und

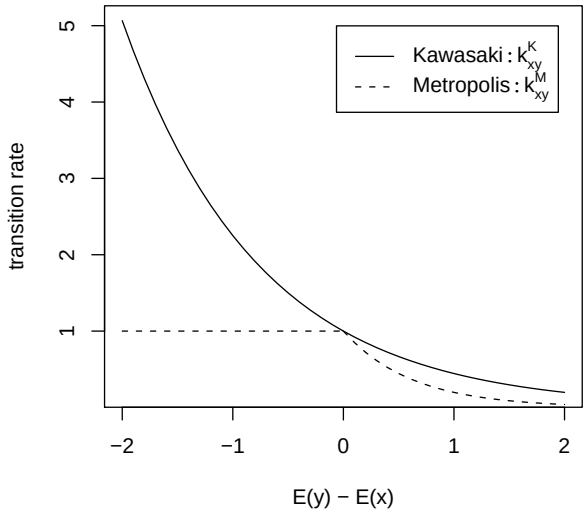


Abbildung 2.2: Vergleich der Übergangsraten. Die Kawasaki Übergangsrate liegt für alle Energiedifferenzen über der Metropolis Übergangsrate. Dies ist besonders ausgeprägt für den Bereich der negativen Energiedifferenzen.

ist ebenfalls nach dem Autor des Artikels benannt. Die Berechnung der Übergangsraten erfolgt in Abhängigkeit der Energie der Zustände. Als Energiefunktion dient üblicherweise die Funktion aus der dem Markov-Prozess zugeordneten Energielandschaft. Da die Definition der Übergangsraten an dieser Stelle unabhängig von einer konkreten Energielandschaft erfolgt, soll daher zunächst die Existenz einer Energiefunktion $E : X \to \mathbb{R}$ als gegeben angenommen werden. Für die auf die Definition der Übergangsraten folgenden Beweise ist zusätzlich die Boltzmann Verteilung nötig.

2.3.1 Boltzmann-Verteilung

Um die Wahrscheinlichkeit eines Zustands bestimmen zu können, muss zunächst die Wahrscheinlichkeitsverteilung der Zustände bekannt sein. Für alle in dieser Arbeit verwendeten Mengen von Zuständen wird die Boltzmann-Verteilung angenommen. Grundlage dieser Wahl ist die Maximum-Entropie-Methode [15]. Diese besagt, dass aus allen möglichen Verteilungen diejenige zu wählen ist, welche die Entropie maximiert. Unter der Annahme, dass die durchschnittliche Energie eines Systems dem Erwartungswert entspricht, ist dies die Boltzmann-Verteilung. Demnach ist die Wahrscheinlichkeit eines Zustands $x \in X$ bei Temperatur T gegeben durch

$$\Pr(x) = \frac{e^{-E(x)/k_B T}}{Z} \tag{2.5}$$

mit der kanonischen Zustandssumme

$$Z = \sum_{y \in X} e^{-E(y)/k_B T}. \tag{2.6}$$

Die Temperatur T wird in Kelvin angegeben, k_B ist die Boltzmann-Konstante. Der Term $e^{-E(x)/k_B T}$ wird als Boltzmann-Gewicht bezeichnet. Dementsprechend besteht die kanonische Zustandssumme aus der Summe der Boltzmann-Gewichte aller Zustände.

Die Boltzmann-Konstante ist eine Proportionalitätskonstante mit der Dimension Energie/Temperatur. Da alle in dieser Arbeit verwendeten Energien auf ein Mol Moleküle bezogen sind, wird von nun an die aus der Boltzmann-Konstante abgeleitete Universelle Gaskonstante R verwendet.

2.3.2 Metropolis

Die Metropolis Übergangsrate bewertet den Übergang zu einem Zustand höherer Energie anders, als den Übergang zu einem Zustand niedrigerer Energie. Die Übergangsrate zu einem Zustand höherer Energie ist gegeben durch das Boltzmann-Gewicht der Energiedifferenz zwischen Ziel- und Ausgangszustand. Alle Übergänge zu Folgezuständen niedrigerer Energie werden pauschal mit einer Übergangsrate von 1 bewertet.

Definition 27 (Metropolis Übergangsrate k_{xy}^M)**.**
Sind zwei Zustände $x, y \in X$ benachbart, gilt für die Übergangsrate nach Metropolis

$$k_{xy}^M = \begin{cases} 1 & \text{, falls } E(y) \leq E(x) \\ e^{\frac{-(E(y)-E(x))}{RT}} & \text{, sonst} \end{cases},$$

ansonsten

$$k_{xy}^M = 0.$$

Nun gilt es zu beweisen, dass bei Verwendung der Metropolis Übergangsrate das Kriterium der Detailed Balance des Markov-Prozesses erfüllt ist, dieser also im thermodynamischen Equilibrium konvergiert.

Beweis 1 (Metropolis Übergangsrate garantiert Detailed Balance)**.**
Gegeben:
Zustandsraum X,
Kanonische Zustandssumme Z über alle Zustände aus X,
stationäre Verteilung π^*, Boltzmann verteilt. D.h. $\pi_x^* = e^{-E(x)/RT}/Z$

Zu zeigen:
$\forall x, y \in \mathcal{X} \ : \ x \sim y \ : \ \pi_x^* k_{xy}^M = \pi_y^* k_{yx}^M$

Beweis:
$E(x) < E(y)$:

$$\begin{aligned}\pi_x^* k_{xy}^M &= e^{\frac{-E(x)}{RT}} Z^{-1} e^{\frac{-(E(y)-E(x))}{RT}} \\ &= e^{\frac{-E(y)}{RT}} Z^{-1} \\ &= e^{\frac{-E(y)}{RT}} Z^{-1} \cdot 1 \\ &= \pi_y^* k_{yx}^M\end{aligned}$$

$E(x) > E(y)$:

$$\textit{analog zu } E(x) < E(y)$$

□

2.3.3 Kawasaki

Definition 28 (Kawasaki Übergangsrate k_{xy}^K).
Sind zwei Zustände $x, y \in X$ benachbart, gilt für die Übergangsrate nach Kawasaki

$$k_{xy}^K = e^{\frac{-(E(y)-E(x))}{2RT}},$$

ansonsten

$$k_{xy}^K = 0.$$

Analog zu Metropolis muss auch für Kawasaki die Detailed Balance erfüllt sein.

Beweis 2 (Kawasaki Übergangsrate garantiert Detailed Balance).
Gegeben:
Zustandsraum X,
Kanonische Zustandssumme Z über alle Zustände aus X,
stationäre Verteilung π^*, Boltzmann verteilt. D.h. $\pi_x^* = e^{-E(x)/RT}/Z$

Zu zeigen:
$\forall x, y \in \mathcal{X} \ : \ x \sim y \ : \ \pi_x^* k_{xy}^K = \pi_y^* k_{yx}^K$

Beweis:

$$\begin{aligned}\pi_x^* k_{xy}^K &= e^{\frac{-E(x)}{RT}} Z^{-1} e^{\frac{-(E(y)-E(x))}{2RT}} \\ &= e^{\frac{-E(x)-E(y)}{2RT}} Z^{-1} \\ &= e^{\frac{-E(y)}{RT}} Z^{-1} e^{\frac{-(E(x)-E(y))}{2RT}} \\ &= \pi_y^* k_{yx}^K\end{aligned}$$

□

2.3.4 Erweiterung auf Makrostates

Die oben beschriebenen Übergangsraten können nur auf die Übergänge eines Markov-Prozesses über Mikrostates angewandt werden. Die auf Übergänge zwischen Makrostates erweiterte Berechnung von Übergangsraten wird in diesem Abschnitt beschrieben. Grundlage der Übergangsraten zwischen Makrostates bleiben jedoch weiterhin die Übergangsraten zwischen einzelnen Zuständen. Auch für Übergänge zwischen Makrostates kann also zwischen Metropolis und Kawasaki gewählt werden.

Jeder Übergang zwischen zwei Basins erfolgt über die Zustände ihrer Kontaktflächen. Dementsprechend werden die Übergangsraten zwischen Makrostates durch die Übergangsraten zwischen den Zuständen der jeweiligen Kontaktflächen berechnet. Die Übergangsraten werden zusätzlich durch die Wahrscheinlichkeit, den Ausgangszustand innerhalb seines Basins anzutreffen, bewertet.

Definition 29 (Übergangsraten zwischen Makrostates)**.**
Gegeben eine Energielandschaft

$$(\mathcal{X}, E, \sim)$$

und den Markov-Prozess

$$\big(\mathcal{B}(\mathcal{X}), \pi^0, P(t)\big)$$

über die Makrostates dieser Energielandschaft. Seien $\alpha, \beta \in \mathcal{B}(\mathcal{X})$ Makrostates der Energielandschaft. Sei k_{xy} die Übergangsrate zwischen zwei Zuständen $x, y \in \mathcal{X}$. Bezeichne Z_α die kanonische Zustandssumme der Zustände in Basin α. Unter der Annahme, die Zustände innerhalb eines Makrostates seien Boltzmann verteilt, ist die Wahrscheinlichkeit, einen Zustand x aus einem Makrostate α zu wählen, gegeben durch

$$\Pr(x|\alpha) = e^{-E(x)/RT}/Z_\alpha.$$

Dann ist die ***Übergangsrate von Makrostate*** α ***nach Makrostate*** β *gegeben durch*

$$\begin{aligned}
\hat{k}_{\alpha\beta} &= \sum_{x\in\alpha}\sum_{y\in\beta} k_{xy}\Pr(x|\alpha) && (2.7)\\
&= \sum_{x\in\alpha}\sum_{y\in\beta} k_{xy}e^{-E(x)/RT}/Z_\alpha && (2.8)\\
&= \frac{1}{Z_\alpha}\sum_{x\in\alpha}\sum_{y\in\beta} k_{xy}e^{-E(x)/RT} && (2.9)
\end{aligned}$$

Auch die Übergangsraten zwischen Makrostates müssen das für die Konvergenz wichtige Kriterium der Detailed Balance erfüllen, darum folgt nun der Beweis über diese Eigenschaft.

Beweis 3 (Makrostate Übergangsraten erfüllen Detailed Balance).
Gegeben:
Zustandsraum $\mathcal{X}$
Makrostates des Zustandsraums $\mathcal{B}(\mathcal{X})$,
Kanonische Zustandssumme Z *über den Zustandsraum* $\mathcal{X}$,
Kanonische Zustandssummen Z_α, Z_β *über Makrostates* $\alpha, \beta \in \mathcal{B}(\mathcal{X})$

Zu zeigen:
$\pi^*_\alpha \hat{k}_{\alpha\beta} = \pi^*_\beta \hat{k}_{\beta\alpha}$ *mit* $\pi^*_\alpha = \frac{Z_\alpha}{Z}$ *und* $\pi^*_\beta = \frac{Z_\beta}{Z}$

Beweis:

$$\begin{aligned}
\pi^*_\alpha \hat{k}_{\alpha\beta} &= \frac{Z_\alpha}{Z Z_\alpha} \sum_{x\in\alpha} \sum_{y\in\beta} k_{xy} e^{-E(x)/RT} && (2.10)\\
&= \sum_{x\in\alpha} \sum_{y\in\beta} \pi^*_x k_{xy} && (2.11)\\
&= \sum_{x\in\alpha} \sum_{y\in\beta} \pi^*_y k_{yx} && (2.12)\\
&= \frac{Z_\beta}{Z Z_\beta} \sum_{x\in\alpha} \sum_{y\in\beta} k_{yx} e^{-E(y)/RT} && (2.13)\\
&= \pi^*_\beta \hat{k}_{\beta\alpha} && (2.14)
\end{aligned}$$

Die Äquivalenz der Ausdrücke 2.11 und 2.12 wurde für die Übergangsraten nach Metropolis und Kawasaki in den Beweisen 1 und 2 gezeigt.

□

Die Identifikation der Makrostates und die Berechnung der Übergangsraten zwischen den Makrostates erfolgen durch das Fluten der Energielandschaft. Dieser Algorithmus wird in Abschnitt 4.2 beschrieben.

2.3.5 Vergleich

Die erste Verwendung der Kawasaki Übergangsrate zur Berechnung von Kinetiken für RNAs geht auf die Dissertation von Christoph Flamm [7] zurück. Dort wird argumentiert, dass die Wahl einer bestimmten Übergangsrate unter zwei Bedingungen lediglich kleine Auswirkungen auf die generelle Dynamik eines Systems hat. Die gewählte Übergangsrate muss das Kriterium der Detailed Balance erfüllen. Außerdem sollte das Move Set keine zu großen Strukturänderungen bewirken. Da kleine strukturelle Änderungen nur kleine Änderungen der Energie bewirken, kann dann ein schneller Ausgleich zwischen den Zuständen erfolgen.

Abbildung 2.3 zeigt die Metropolis und Kawasaki Übergangsraten in Abhängigkeit von der Energiedifferenz der benachbarten Zustände. Vor allem für Kawasaki wird dabei deutlich, dass die Übergangsraten für hohe positive und negative Energiedifferenzen weit auseinander liegen. Zwischen kleineren positiven und negativen Energiedifferenzen besteht jedoch nur ein relativ geringer Unterschied.

Der hauptsächliche Unterschied zwischen den beiden Übergangsraten liegt in der Festlegung der Metropolis Raten für negative Energieunterschiede. Dadurch werden bei Metropolis alle Übergänge zu Zuständen niedriger Energie gleich bewertet und somit eine künstliche obere Grenze für die Übergangsraten eingeführt. Bei Kawasaki existiert diese Einschränkung nicht. Zudem werden Übergänge zu energetisch vorteilhafteren Zuständen progressiv höher bewertet als Übergänge zu weniger vorteilhaften.

Aufgrund höherer Raten zu energetisch besseren Zuständen für die Übergangsraten nach Kawasaki, scheinen stochastische Prozesse mit dieser Übergangsrate etwas schneller im Equilibrium zu konvergieren als mit der Metropolis Übergangsrate. Dies führt bei Verwendung von stochastischen Simulationen zu einer schnelleren Berechnung der Kinetiken. In Abschnitt 5.2.2 wird eine Gegenüberstellung der beiden Methoden bei der Berechnung von Kinetiken von Hybridisierungen unternommen.

2.4 Ribonukleinsäure

Im diesem Abschnitt folgen auf eine allgemeine Beschreibung der RNA die Definitionen der Sekundärstruktur sowie deren Energie. Mit diesen Grundlagen kann dann in Abschnitt 2.5 die Energielandschaft für Kinetiken einzelner RNAs beschrieben werden.

Desoxyribonukleinsäure (DNA, desoxyribonucleic acid) und Ribonukleinsäure (RNA, ribonucleic acid) sind sogenannte Polynukleotide, die aus den Grundbausteinen Nukleobase, Zucker und Phosphorsäurediester aufgebaut sind. Ein Molekül bestehend aus einer der Nukleobasen, einem Zucker und einem Phosphorsäurediester wird als Nukleotid bezeichnet. Diese Nukleotide sind an den Zuckermolekülen über Phosphorsäurediester miteinander verknüpft und bilden eine Kette. Die Abfolge der verschiedenen Nukleotide in dieser Kette wird als Primärstruktur oder Nukleotidsequenz bezeichnet.

Ribose und Desoxyribose bestehen jeweils aus einem Ring von fünf Kohlenstoffatomen, die mit 1' bis 5' nummeriert sind. Die Bindung an den Phosphorsäurediester erfolgt über das 5' Atom. Über eine OH-Gruppe am 3' Atom kann die Bindung zum Phosphorsäurediester eines weiteren Nukleotids gebildet werden. Dies hat zur Folge, dass jede Nukleotidsequenz von RNA oder DNA zwei verschiedene Enden hat, das 5' und das 3' Ende. Die Notation erfolgt üblicherweise ausgehend vom 5' Ende zum

3' Ende der Nukleotidsequenz.

Im molekularen Aufbau unterscheiden sich DNA und RNA durch die enthaltenen Nukleobasen und den Zucker. Die Nukleobasen der DNA sind Adenin, Guanin, Cytosin und Thymin; der Zucker ist eine Desoxyribose. Bei der RNA ist statt Thymin die Nukleobase Uracil enthalten, der Zucker ist eine Ribose.

Jede der Nukleobasen kann sich mit einer weiteren Nukleobase über Wasserstoffbrückenbindungen zu einem sogenannten Basenpaar verbinden. Ist diese Bindung für zwei Basen möglich, werden diese als Komplementärbasen bezeichnet. Bei den Nukleobasen der RNA ist eine solche Bindung zwischen Adenin und Uracil, Cytosin und Guanin (Watson-Crick Paare) sowie Guanin und Uracil (Wobble Paar) möglich. DNA liegt üblicherweise als doppelsträngige Helix zueinander komplementärer Einzelstränge vor, RNA ist in der Regel einsträngig.

Die Beschreibung der Struktur von RNA und DNA erfolgt anhand einer Klassifikation in Primär-, Sekundär-, Tertiär- und Quartärstruktur. Diese sind nach der Komplexität der durch sie beschriebenen Strukturinformationen geordnet, wobei die Strukturinformationen einer Ebene von denen der untergeordneten Ebenen abhängig sind. Die Primärstruktur beschreibt die Abfolge der Nukleotide in der Polynukleotidkette. Die Sekundärstruktur enthält die von den Nukleotiden gebildeten Basenpaare. Die Tertiärstruktur bezeichnet die räumliche Anordnung der durch die Basenpaare miteinander verbundenen Nukleotidkette. Unter der Quartärstruktur versteht man einen Komplex aus mehreren Nukleotidketten.

Die Funktionalität einer RNA wird durch ihre räumliche Struktur bestimmt. Ansätze zur Bestimmung dieser Struktur gelten jedoch als noch nicht ausgereift. Üblicherweise erfolgt darum lediglich die Betrachtung von Sekundärstrukturen [7]. Diese Einschränkung hat eine Reduktion der Komplexität des Problems zur Folge, die es einer mathematischen Analyse zugänglich macht und die praktische Durchführbarkeit der nötigen Berechnungen ermöglicht. Aus diesen Gründen werden auch in dieser Arbeit lediglich Sekundärstrukturen betrachtet und nicht die vollständige räumliche Konformation.

In einer Arbeit von Christoph Flamm wird argumentiert, warum trotz einer Beschränkung auf Sekundärstrukturen ausreichend gute Ergebnisse zu erwarten sind [6]. Obwohl durch die Sekundärstruktur nicht die vollständige Struktur der RNA abgebildet wird, bildet sie gewissermaßen das Gerüst für die Ausbildung der komplexeren räumlichen Struktur. Es konnte gezeigt werden, dass die Bindungen innerhalb der Sekundärstruktur zu einem Großteil für die Stabilität der gesamten Struktur verantwortlich sind. Es besteht also ein starker Zusammenhang zwischen Sekundärstruktur und Tertiärstruktur von RNA. Bei der Bildung der endgültigen räumlichen Struktur werden nur einige wenige zusätzliche Bindungen ausgebildet. Diese Bindungen werden bei steigender Temperatur als erste wieder aufgelöst.

2.4.1 Primärstruktur

Die Begriffe Nukleotidsequenz oder Primärstruktur bezeichnen die Abfolge der Nukleotide einer RNA. Falls nicht anders angegeben, beginnt die Notation am 5'-Ende und endet am 3'-Ende der Sequenz. Die Bezeichnung der Nukleotide erfolgt durch den jeweiligen Anfangsbuchstaben.

Definition 30 (Primärstruktur S)**.**
Sei Σ das Alphabet der Nukleotide Adenin, Cytosin, Guanin und Uracil mit

$$\Sigma = \{A, C, G, U\}\,.$$

*Dann ist die **Primärstruktur** S der Länge $n \in \mathbb{N}$ gegeben durch*

$$S \in \Sigma^n.$$

2.4.2 Sekundärstruktur

Für die Elemente der Sekundärstrukturen sollen ausschließlich Bindungen zwischen Adenin und Uracil, zwischen Cytosin und Guanin sowie zwischen Guanin und Uracil erlaubt werden. Diese werden als gültige Basenpaare bezeichnet.

Definition 31 (gültiges Basenpaar)**.**
*Sei S eine Primärstruktur. Dann bezeichnet (i, j) ein **gültiges Basenpaar**, falls gilt*

$$\{S_i, S_j\} \in \{\{A, U\}, \{C, G\}, \{G, U\}\}\,.$$

Die Sekundärstruktur einer RNA besteht aus der Menge der von ihr ausgebildeten gültigen Basenpaare.

Definition 32 (Sekundärstruktur P, Menge aller Sekundärstrukturen $\mathcal{P}$)**.**
*Sei S eine Primärstruktur. Dann sei eine **Sekundärstruktur** über S eine Menge von Paaren*

$$P = \{(i, j) \mid i < j \ \wedge\ (i, j) \text{ ist gültiges Basenpaar } \}$$

unter den Bedingungen, dass jedes Nukleotid Teil höchstens eines Basenpaars ist

$$\forall (i, j), (i', j') \in P : i = i' \Rightarrow j = j'$$

und die non-crossing Eigenschaft

$$\forall (i, j), (i', j') \in P : i < i' \Rightarrow j > j'$$

*erfüllt ist. Eine Sekundärstruktur, die keine Basenpaare enthält, wird als offene Kette bezeichnet. Unter dem **externen Basenpaar** einer Sekundärstruktur versteht man das Basenpaar welches von keinem weiteren Basenpaar überspannt wird. Für das externe Basenpaar $(i, j) \in P$ gilt also*

$$\nexists (k.l) \in P : k < i \ \wedge\ j < l.$$

*Bezeichne zusätzlich $\mathcal{P}$ die **Menge aller Sekundärstrukturen** über eine Primärstruktur.*

Die Beschränkung auf Sekundärstrukturen mit sich nicht überkreuzenden Basenpaaren ermöglicht eine Repräsentation durch die sogenannte Dot-Bracket Notation. Die Position eines ungebundenen Basenpaars wird durch einen Punkt, öffnende und schließende Bindungen durch entsprechende Klammern gekennzeichnet. Dementsprechend wäre beispielweise

```
5' ACUGGCACAGG 3'
   .(((...))).
```

die Dot-Bracket Notation der Sekundärstruktur $P = \{(2, 10), (3, 9), (4, 8)\}$. Im Folgenden soll jeder Sekundärstruktur eine Energie zugeordnet werden.

2.4.3 Energie einer Sekundärstruktur

Für die Energie einer Sekundärstruktur wird das Modell der Gibbschen freien Energie verwendet.

Definition 33 (Gibbsche freie Energie)**.**
Gegeben die Enthalpie U, die Entropie S und absolute Temperatur T bezeichnet

$$G = U - TS \tag{2.15}$$

*die **Gibbsche freie Energie** eines Systems das bei dieser Temperatur im thermischen Gleichgewicht mit seiner Umgebung steht. Als Maßeinheit der Gibbschen freien Energie wird im Folgenden kcal/mol verwendet.*

Dieses System kann eine Menge von Gasmolekülen im thermodynamischen Gleichgewicht sein, lässt sich aber auch zur Beschreibung von Molekülen in einer Lösung verwenden. Für diese Arbeit wird die freie Energie einer Lösung von RNA Molekülen betrachtet. Die Enthalpie oder innere Energie des Systems umfasst im Fall von RNA die Energiebeiträge der Basenpaarbindungen, die Entropie ist das Maß für Unordnung. Somit beschreibt die Gibbsche freie Energie, im Folgenden kurz freie Energie, eine Differenz zwischen Ordnung und Unordnung.

Da weder Entropie noch Enthalpie direkt gemessen werden können, wird in Experimenten die Differenz

$$\Delta G = \Delta U - T\Delta S \tag{2.16}$$

bestimmt. Durch die Bestimmung dieser Differenzen für verschiedene Basensequenzen lassen sich Energiebeiträge für einzelne Sekundärstrukturelemente herleiten [10]. Durch Summation über die Energiebeiträge aller in einer Struktur enthaltener Strukturelemente kann letztlich die freie Energie dieser Konformation in Abhängigkeit von der Temperatur berechnet werden.

Die Berechnung der freien Energie soll nach dem Zuker Modell erfolgen [46]. Bei dieser Methode wird die freie Energie einer Sekundärstruktur aus der Summe der enthaltenen Sekundärstrukturelemente berechnet. Der Zuker Algorithmus unterscheidet die

Elemente Stacking, innere Schleife, Bulge und k-Multiloop. Nach Definition 32 darf eine Sekundärstruktur keine sich überkreuzenden Basenpaare enthalten, Pseudoknoten werden also nicht betrachtet.

Definition 34 (Stacking)**.**
Sei P eine Sekundärstruktur. Zwei Basenpaare $(i,j),(i',j') \in P$ mit $i < i' < j' < j$, deren öffnende und schließende Bindungen in der Nukleotidsequenz direkt aufeinanderfolgen, also gilt

$$i' = i + 1 \ \wedge j = j' + 1,$$

bilden ein ***Stacking****.*

Definition 35 (Hairpin-Loop)**.**
Sei P eine Sekundärstruktur. Ein Basenpaar $(i,j) \in P$ bildet einen ***Hairpin-Loop****, falls die Sekundärstruktur kein weiteres Basenpaar zwischen i und j enthält:*

$$\forall (k,l) \in P \backslash (i,j) \ : \ k < i \ \wedge \ j < l.$$

Definition 36 (innere Schleife, Bulge)**.**
Sei P eine Sekundärstruktur. Zwei Basenpaare $(i,j),(i',j') \in P$ mit $i < i' < j' < j$ und $(i' - i + j - j') > 2$ bilden eine ***innere Schleife****, falls der Bereich zwischen diesen Basenpaaren kein weiteres Basenpaar enthält:*

$$\forall (k,l) \in P : \left([k < i \wedge l > j] \ \vee \ \left[k > i' \wedge l < j'\right]\right)$$

Falls die öffnenden oder die schließenden Bindungen der Basenpaare in der Basensequenz direkt aufeinanderfolgen, also gilt

$$(i' = i + 1) \vee (j = j' + 1),$$

wird eine innere Schleife auch ***Bulge*** *genannt.*

Definition 37 (k-Multiloop)**.**
Sei P eine Sekundärstruktur. Die Basenpaare

$$(i_1, j_1), \ldots, (i_k, j_k) \in P \text{ und } (i_0, j_{k+1}) \in P$$

bilden einen ***k-Multiloop****, falls gilt*

$$\forall \ 0 \leq l \leq k \ : \ j_l < i_{l+1}$$

und

$$\forall \ 0 \leq l, \ l' \leq k \ : \ \nexists (i',j') \in P \ : \ i' \in [j_l, i_{l+1}] \ \vee \ j' \in [j_{l'}, i_{l'+1}].$$

Nach der Definition der Sekundärstrukturelemente kann nun die Definition der freien Energie einer Sekundärstruktur als Summe der Energiebeiträge ihrer Sekundärstrukturelemente definiert werden.

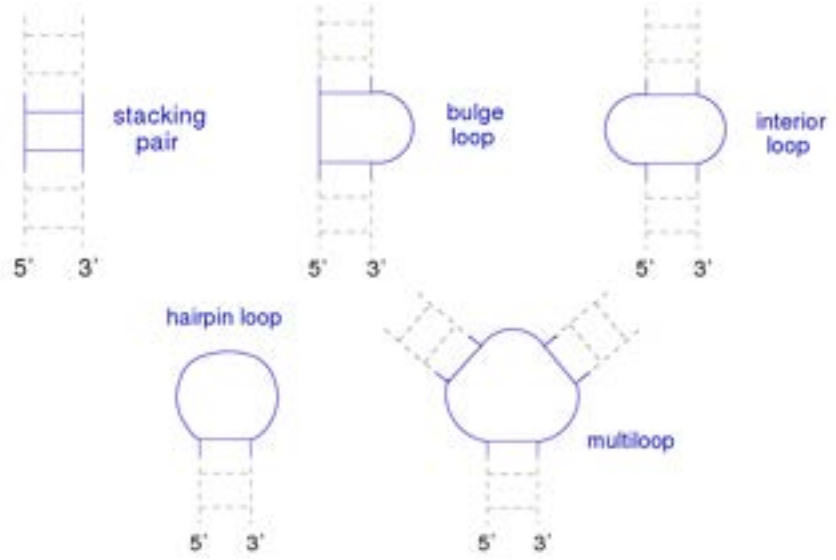


Abbildung 2.3: RNA Sekundärstrukturelemente, entnommen aus [1].

Definition 38 (Energie einer Sekundärstruktur)**.**
Sei P eine Sekundärstruktur über die Sequenz S. Sei ein Sekundärstrukturelement eindeutig bestimmt über das externe Basenpaar. Bezeichne $E^P_{i,j}$ den Energiebeitrag des Sekundärstrukturelements mit externem Basenpaar (i,j). Dann ist die Energie einer Sekundärstruktur gegeben durch

$$E(P) = \sum_{(i,j)\in P} E^P_{i,j}.$$

Dementsprechend wird der offenen Kette eine Energie von 0 zugewiesen.

Die Berechnung der freien Energie erfolgt für diese Arbeit durch die vom Vienna RNA Package bereitgestellte Funktion `energy_of_struct`.

2.5 Kinetiken von RNA-Faltungen

In diesem Abschnitt soll zunächst die in [7] für Kinetiken einzelner RNAs verwendete Energielandschaft definiert werden. Danach wird der Markov-Prozess über die vollständige Energielandschaft sowie der Markov-Prozess über Makrostates definiert.

2.5.1 Energielandschaft

Für die Definition der Energielandschaft wird ein Zustandsraum, eine Energiefunktion sowie eine Nachbarschaftsrelation benötigt. Der Zustandsraum soll aus der Menge der erlaubten Sekundärstrukturen $\mathcal{P}$ bestehen, als Energiefunktion dient die in

Abschnitt 2.4.3 definierte Energiefunktion $E : \mathcal{P} \to \mathbb{R}$. Nun muss lediglich die Nachbarschaftsrelation über die Angabe der zu verwendeten Moves definiert werden.

Hier soll der Übergang zu einem benachbarten Zustand durch das Einfügen oder Löschen eines Basenpaars sowie das Verschieben einer der Bindungen eines Basenpaares geschehen. Die mit diesen Operationen assoziierten Moves werden als Insertion, Deletion und Shift bezeichnet. Ein Shift Move verschiebt eine der Bindungen eines Basenpaares nach links oder rechts bis zur nächsten Position, an der wieder ein gültiges Basenpaar gebildet werden kann.

Das Move Set bestehend aus Insertion und Deletion ist das elementarste ergodische und symmetrische Move Set. Durch die zusätzliche Verwendung von Shift Moves sollen Vorgänge wie etwa die Wanderung von Bulges entlang von Stackings vereinfacht werden [7].

Definition 39 (Insertion, Deletion, Shift)**.**
Gegeben ein Zustand $x \in \mathcal{P}$ und $i, j, k \in \mathbb{N}$.
Ein Move m_{ins} ist eine Insertion, falls gilt

$$m_{ins}(x) = x \cup (i, j) \text{ und } m_{ins}(x) \in \mathcal{P}.$$

Ein Move m_{del} ist eine Deletion, falls gilt

$$m_{del}(x) = x \backslash (i, j) \text{ und } m_{del}(x) \in \mathcal{P}.$$

Ein Move m_{shift} ist ein Shift-Move, falls gilt

$$\begin{aligned} m_{shift}(x) \quad = \quad & \big([x \cup (i,j) \cap (i, j \pm k)] \wedge \big[\nexists (i, j \pm k') : k' < k : x \cup (i, j \pm k') \in \mathcal{P}\big]\big) \vee \\ & \big([x \cup (i,j) \cap (i \pm k, j)] \wedge \big[\nexists (i, j \pm k') : k' < k : x \cup (i \pm k', j) \in \mathcal{P}\big]\big) \end{aligned}$$

und $m_{shift}(x) \in \mathcal{P}$.

Die Zusammenfassung dieser Moves zu einem Move Set ermöglicht die Definition einer symmetrischen Nachbarschaftsrelation.

Definition 40 (Move Set $\mathcal{N}_{single}$, Nachbarschaftsrelation $\sim_{single}$)**.**
Gegeben ein Zustand $x \in \mathcal{P}$.
Sei $\mathcal{N}_{single} : \mathcal{P} \to Pot(\mathcal{P})$ das Move Set für Faltungskinetiken einzelner RNAs, falls gilt

$$\mathcal{N}_{single}(x) = m_{ins}(x) \uplus m_{del}(x) \uplus m_{shift}(x).$$

Dann ist die symmetrische Nachbarschaftsrelation definiert durch

$$x \sim_{single} y \Leftrightarrow [(x \in \mathcal{N}_{single}(y)) \wedge (y \in \mathcal{N}_{single}(x))]\,.$$

Damit ist die Energielandschaft für Faltungskinetiken einzelner RNAs vollständig definiert durch das Tripel

$$(\mathcal{P}, E, \sim_{single}) \tag{2.17}$$

3 Energielandschaften von Hybridisierungen

In diesem Kapitel wird die von mir durchgeführte Erweiterung der im vorigen Kapitel vorgestellten Energielandschaft für Faltungen einzelner RNAs auf Hybridisierungen beschrieben.

Nach der Definition der Struktur von Hybridisierungen werden die Zustände der Energielandschaft definiert. Diese repräsentieren nun nicht mehr einzelne Strukturen, sondern Ensembles von Hybridisierungen. Nach Definition der Energiefunktion sowie der Nachbarschaftsrelation wird eine Einschränkung für die Zustände der Energielandschaft vorgestellt, die Limitierung der Größe des Interaktionsbereichs der Hybridisierungen.

Die dadurch bewirkte Verkleinerung des Zustandsraums ermöglicht erst die Anwendbarkeit des Makrostate-Prozesses zur Berechnung der Kinetiken. Zudem ermöglicht die Parametrisierung des maximalen Interaktionsbereichs eine Anpassung an verschiedene Längen der für die Hybridisierung verwendeten Sequenzen.

3.1 Hybridisierungen

Eine Hybridisierung besteht aus zwei RNA-Strängen S^1 und S^2 zwischen denen Basenpaarbindungen ausgebildet werden. Für die Anzahl der Nukleotide der Stränge S^1 und S^2 gilt $|S^1| = n$ und $|S^2| = m$. Dann ist eine Hybridisierung H wie folgt definiert.

Definition 41 (Interaktion I, Hybridisierung H)**.**
Gegeben zwei Nukleotidsequenzen S^1 und S^2 sowie zwei Sekundärstrukturen P^1 und P^2 über diese Sequenzen. Gegeben eine ***Interaktion*** *I bestehend aus der Menge der zwischen den Sequenzen ausgebildeten Basenpaare*

$$I = \left\{(i, i') \mid S^1_i \text{ und } S^2_{i'} \text{ bilden ein gültiges Basenpaar}\right\}.$$

Sei jedes Nukleotid höchstens Teil eines Basenpaares dieser Interaktion und gelte die non-crossing Eigenschaft

$$\forall\ (i,i'),(j,j') \in I\ :\ i = j \Rightarrow i' = j'\ \wedge\ i < j \Rightarrow i' < j'.$$

Aufgrund dieser Eigenschaft kann eine eindeutige Ordnung über die Basenpaare der Interaktion eingeführt werden und es gilt

$$\forall\ (i,i'),(j,j') \in I\ :\ i < j \Rightarrow (i,i') < (j,j').$$

Wenn zwischen den Basenpaaren der Interaktion keine Bindungen der Sekundärstrukturen beginnen oder enden und weiterhin jedes Nukleotid der Sequenzen an höchstens einem Basenpaar beteiligt ist, also gilt

$$\forall\ (i,i') \in I \quad : \quad \left[\nexists (k,l) \in P^1 \ : \ i \leq k \leq j \ \vee \ i \leq l \leq j\right] \ \wedge \\ \left[\nexists (k',l') \in P^2 \ : \ i' \leq k' \leq j' \ \vee \ i' \leq l' \leq j'\right],$$

ist eine ***Hybridisierung*** *H durch das Tripel*

$$(I, P^1, P^2) \tag{3.1}$$

definiert.

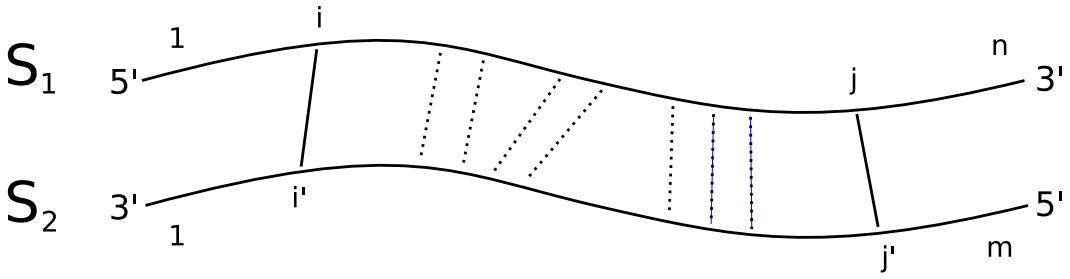


Abbildung 3.1: Stilisierte Darstellung einer Hybridisierung zwischen den RNA Sequenzen S^1 und S^2. S^1 wurde in Richtung $5' \rightarrow 3'$ aufgetragen, S^2 in umgekehrter Reihenfolge. Sequenz S^1 besteht aus n, Sequenz S^2 aus m Nukleotiden. Das kleinste Basenpaar der Interaktion wurde in dieser Abbildung mit (i, i'), das größte Basenpaar mit (j, j') gekennzeichnet.

Im Gegensatz zur Definition der Energielandschaft über Sekundärstrukturen in Abschnitt 2.17, wo jedes Element des Zustandsraums für eine konkrete Sekundärstruktur stand, sollen die Zustände der Energielandschaft für Hybridisierungen Mengen von Hybridisierungen repräsentieren. Solche Mengen von Strukturen werden als Ensembles bezeichnet.

Definition 42 (Zustände der Energielandschaft)**.**
Gegeben die Mengen $\mathcal{P}^1$ und $\mathcal{P}^2$ aller Sekundärstrukturen über die Sequenzen S^1 und S^2. Dann ist ein Zustand der Energielandschaft gegeben durch das Ensemble von Hybridisierungen mit fester Interaktion I

$$H(I) = \left\{(I, P^1, P^2) : P^1 \in \mathcal{P}^1 \wedge P^2 \in \mathcal{P}^2\right\}.$$

Gelte im Folgenden die abkürzende Schreibweise $H = H(I)$ für die Zustände der Energielandschaft. Da diese durch die Angabe einer Interaktion vollständig definiert sind, sollen H und I synonym verwendet werden. Bezeichne $\mathcal{H}$ die Menge aller Zustände für zwei gegebene Nukleotidsequenzen.

Die Definition eines Zustands der Energielandschaft als Ensemble von Hybridisierungen hat eine deutliche Reduktion der Anzahl der Zustände im Vergleich zur Repräsentation jeder einzelnen Hybridisierung zur Folge. Da zur Beschreibung eines Zustands lediglich die Angabe einer Interaktion notwendig ist, müssen die Sekundärstrukturen P^1 und P^2 nicht explizit repräsentiert werden. Zusätzlich verringert sich die Komplexität des benötigten Move Set.

3.2 Die Energiefunktion

Die Energie für das durch eine Interaktion definierte Ensemble von Strukturen setzt sich aus zwei Komponenten zusammen: der Energie für die Bindungen der Interaktion sowie zwei Energiebeiträgen für die in dem Ensemble enthaltenen Sekundärstrukturen. Zunächst soll die Energie der Interaktion definiert werden.

3.2.1 Energie der Interaktion

Der Energiebeitrag für die Bindungen der Interaktion basiert, wie die in Abschnitt 2.4.3 definierte Berechnung der freien Energie von Sekundärstrukturen, auf der Zuker Methode. Die freie Energie der Interaktion setzt sich also ebenfalls aus der Summe der Energiebeiträge der enthaltenen Sekundärstrukturelemente zusammen. Dafür müssen diese zunächst für zwischen zwei Sequenzen ausgebildete Basenpaare definiert werden. Für die Energiebeiträge der Strukturelemente der Interaktion werden die gleichen Energiebeiträge benutzt wie für die Berechnung von Sekundärstrukturen. Eine Interaktion kann die Strukturelemente Stacking, innere Schleife und Bulge enthalten.

Definition 43 (Stacking)**.**
Gegeben eine Interaktion I. Zwei Basenpaare $(i,i'),(j,j') \in I$ mit $i < j$ bilden ein Stacking, falls gilt

$$j = i + 1 \ \wedge \ j' = i' + 1.$$

Definition 44 (innere Schleife, Bulge)**.**
*Gegeben eine Interaktion I. Zwei Basenpaare $(i,i),(j',j') \in I$ mit $i < i'$ und $j < j'$ sowie $(i' - i + j' - j) > 2$ bilden eine **innere Schleife**, falls der Bereich zwischen diesen Basenpaaren kein weiteres Basenpaar enthält:*

$$\forall (k,k') \in I : \big([k < i \wedge k' < i'] \ \vee \ [k > i' \wedge k' > j']\big)$$

Falls die öffnenden oder die schließenden Bindungen der Basenpaare in der Basensequenz direkt aufeinanderfolgen, also gilt

$$(i' = i + 1) \vee (j' = j + 1),$$

*wird eine innere Schleife auch **Bulge** genannt.*

Unter diesen Voraussetzungen kann nun die Definition der Energie einer Interaktion erfolgen.

Definition 45 (Energie einer Interaktion $E_i(I)$)**.**
*Sei I eine Interaktion zwischen den Sequenzen S^1 und S^2. Sei ein Sekundärstrukturelement dieser Interaktion eindeutig bestimmt über das kleinste enthaltene Basenpaar. Bezeichne desweiteren $E^P_{i,i'}$ den Energiebeitrag des Sekundärstrukturelements (i, i'). Das größte Basenpaar einer Interaktion ist nicht Teil eines Sekundärstrukturelements, darum wird für dessen Energiebeitrag 0 angenommen. Dann ist die **Energie einer Interaktion** gegeben durch*

$$E_i(I) = \sum_{(i,i')\in I} E^P_{i,i'}.$$

3.2.2 Energie für die Zugänglichkeit des Interaktionsbereichs

Die Energie für die Zugänglichkeit des Interaktionsbereichs modelliert den Einfluss der in einem Zustand enthaltenen Sekundärstrukturen. In Definition 41 wurde für eine Hybridisierung gefordert, dass zwischen den Basenpaaren der Interaktion keine Bindungen der Sekundärstrukturen beginnen oder enden dürfen. Die Bereiche der Nukleotidsequenzen einer Hybridisierung zwischen dem kleinsten und größten Basenpaar der Interaktion werden nun als Interaktionsbereiche bezeichnet.

Definition 46 (Interaktionsbereich)**.**
*Gegeben eine Hybridisierung über die Nukleotidsequenzen S^1 und S^2 mit der Interaktion I. Seien $(i, i'), (j, j') \in I$ das kleinste und größte Basenpaar der Interaktion. Dann sind die **Interaktionsbereiche** der Sequenzen S^1 und S^2 gegeben durch $S^1_i \dots S^1_j$ und $S^2_{i'} \dots S^2_{j'}$.*

Die Energie für die Zugänglichkeit des Interaktionsbereichs gibt nun den Energiebeitrag an, der im Schnitt aufgewendet werden muss, damit die Nukleotide der Interaktionsbereiche innerhalb der Sekundärstrukturen ungepaart vorliegen und somit für die Interaktion zur Verfügung stehen. Für die Definition dieser Energie wird die Energie eines Ensembles von Sekundärstrukturen benötigt. Diese berechnet sich aus den Energien der Elemente des Ensembles.

Definition 47 (Ensemble, Ensembleenergie)**.**
*Sei $\tilde{\mathcal{P}}$ eine Menge von Sekundärstrukturen. Diese Menge wird als **Ensemble** bezeichnet. Sei zusätzlich*

$$Z_{\tilde{\mathcal{P}}} = \sum_{P\in\tilde{\mathcal{P}}} e^{-\frac{E(P)}{RT}}$$

die kanonische Zustandssumme über das Ensemble $\tilde{\mathcal{P}}$.
*Dann ist die **Energie des Ensembles** gegeben durch*

$$E^{ens}(\tilde{\mathcal{P}}) = -RT \ln(Z_{\tilde{\mathcal{P}}}).$$

Außerdem soll über Ensembles von Sekundärstrukturen mit ungepaarten Bereichen argumentiert werden.

Definition 48 (Sekundärstrukturen mit ungepaartem Bereich $\mathcal{P}_{i,j}^{unpaired}$).
Gegeben eine Basensequenz S und die Menge $\mathcal{P}$ der Sekundärstrukturen über diese Sequenz. Dann bezeichne

$$\mathcal{P}_{i,j}^{unpaired} = \{P \mid P \in \mathcal{P} \wedge \nexists (k,l) \in P \ : \ (i \leq k \leq j) \vee (i \leq l \leq j)\}$$

die ***Menge aller Sekundärstrukturen*** *von S in denen der* ***Bereich von*** *i* ***bis*** *j* ***ungepaart*** *vorliegt. Die Wahrscheinlichkeit, dass dieser Bereich ungepaart vorliegt ist gegeben durch*

$$Pr_{i,j}^{unpaired} = \frac{Z_{\mathcal{P}_{i,j}^{unpaired}}}{Z_{\mathcal{P}}}.$$

Damit kann der Energiebeitrag für die Zugänglichkeit eines Bereichs definiert werden. Dieser wird aus der Differenz zwischen der Energie des Ensembles aller Strukturen, in denen dieser Bereich ungepaart vorliegt, und der Energie des Ensembles aller gültigen Strukturen gebildet.

Definition 49 (Energiebeitrag für Zugänglichkeit eines Bereichs E_a).
Der ***Energiebeitrag*** *$E_a(i,j)$* ***für die Zugänglichkeit des Bereichs*** *$S_i \dots S_j$ der Basensequenz S für die Hybridisierung ist gegeben durch*

$$E_a(i,j) = E^{ens}(\mathcal{P}_{i,j}^{unpaired}) - E^{ens}(\mathcal{P}).$$

Die Sekundärstrukturen des Zustands, dessen Interaktion keine Basenpaare enthält, unterliegen keiner Einschränkung. Dementsprechend ist der Energiebeitrag für die Zugänglichkeit des Interaktionsbereichs in diesem Fall 0.

Diese Wahl des Energiebeitrags für die Zugänglichkeit eines Bereichs kann durch folgenden Zusammenhang zwischen der Wahrscheinlichkeit eines ungepaarten Bereichs und des für die Zugänglichkeit dieses Bereichs definierten Energiebeitrags gerechtfertigt werden. Es gilt

$$Pr_{i,j}^{unpaired} = e^{\frac{-E_a(i,j)}{RT}}. \tag{3.2}$$

Dieser Zusammenhang soll nun bewiesen werden.

Beweis 4 (Zusammenhang von Energie und Wahrscheinlichkeit).
Zu zeigen:

$$Pr_{i,j}^{unpaired} = e^{\frac{-E_a(i,j)}{RT}}$$

Beweis:

$$\begin{aligned}
Pr_{i,j}^{unpaired} &= \frac{Z_{i,j}^{unpaired}}{Z} \\
&= e^{\left(\ln(Z_{i,j}^{unpaired}) - \ln(Z_{i,j})\right)} \\
&= e^{\frac{\left(RT\ln(Z_{i,j}^{unpaired}) - RT\ln(Z_{i,j})\right)}{RT}} \\
&= e^{\frac{-\left(E^{ens}(\mathcal{P}_{i,j}^{unpaired}) - E^{ens}(\mathcal{P})\right)}{RT}} \\
&= e^{\frac{-E_a(i,j)}{RT}}
\end{aligned}$$

□

3.2.3 Gesamtenergie

Mit der Energie der Interaktion sowie den Energiebeiträgen für die Zugänglichkeit der Interaktionsbereiche kann nun die Energie eines Zustands definiert werden.

Definition 50 (Energie eines Zustands)**.**
*Gegeben einen Zustand H über die Nukleotidsequenzen S^1 und S^2. Sei $(i, i') \in I$ das kleinste und $(j, j') \in I$ das größte Basenpaar der Interaktion I des Zustands. Damit sind die Interaktionsbereiche für S^1 und S^2 gegeben durch $S^1_i, \ldots, S^1_j$ sowie $S^2_{i'}, \ldots, S^2_{j'}$. Bezeichne E^1_a die Energie für die Zugänglichkeit eines Bereichs von S^1 und E^2_a die Energie für die Zugänglichkeit eines Bereichs von S^2. Dann ist die **Energie des Zustands** H gegeben durch*

$$E_h(H) = E_i(I) + E^1_a(i, j) + E^2_a(i', j'). \tag{3.3}$$

3.3 Einschränkung des Interaktionsbereichs

Nach der Definition einer Hybridisierung als Tripel aus einer Interaktion und zwei Sekundärstrukturen folgte die Definition eines Zustands als Ensemble aller Hybridisierungen mit gemeinsamer Interaktion. Die damit einhergehende Verkleinerung des Zustandsraums ist aber alleine noch nicht ausreichend, um die Anwendbarkeit der Algorithmen zur Berechnung der Kinetiken zu ermöglichen. Über die im Folgenden beschriebene Beschränkung der Interaktionsbereiche auf eine maximale Länge soll die Größe des Zustandsraums weiter reduziert werden.

Definition 51 (Hybridisierung mit eingeschränktem Interaktionsbereich)**.**
Sei $\mathcal{H}$ die Menge der Hybridisierungen. Bezeichne desweiteren (i, i') das kleinste Basenpaar und (j, j') das größte Basenpaar einer Interaktion. Dann ist die Menge der Hybridisierungen mit durch $w \in \mathbb{N}^+$ eingeschränktem Interaktionsbereich gegeben durch

$$\mathcal{H}_w = \left\{ I \in \mathcal{H} : |j - i + 1| \leq w \ \wedge \ |j' - i' + 1| \leq w \right\}.$$

Im Folgenden wird w als Hybridisierungsfenster bezeichnet.

Sei die Menge der Hybridisierungen mit eingeschränktem Interaktionsbereich $\mathcal{H}_w$ der Zustandsraum der Energielandschaft. Das Move Set für diese Energielandschaft besteht analog zu dem in 2.17 beschriebenen Move Set aus den Moves Insertion, Deletion und Shift. Dazu kann Definition 39 der Moves nach Ersetzen der Menge der Sekundärstrukturen $\mathcal{P}$ durch die Menge der eingeschränkten Hybridisierungen $\mathcal{H}_w$ direkt übernommen werden. Die auf diese Weise definierte symmetrische Nachbarschaftsrelation sei $\sim_h$.

Damit ist die Energielandschaft für RNA-RNA Hybridisierungen mit Hybridisierungsfenster w gegeben durch das Tripel

$$(\mathcal{H}_w, E, \sim_h). \tag{3.4}$$

Die Markov-Prozesse über Mikro- und Makrostates der Energielandschaft ergeben sich aus Definitionen 25 und 26. Damit kann nun die Berechnung von Kinetiken für Hybridisierungen durchgeführt werden.

4 Algorithmen

Für die Berechnung von Kinetiken sollen zwei Methoden verwendet werden, die auf den in Abschnitt 2.2 definierten Markov-Prozessen über Energielandschaften basieren: stochastische Simulation und die Berechnung der Übergangswahrscheinlichkeiten des Makrostate-Prozesses.

Die Berechnung von Kinetiken über stochastische Simulationen erfolgt in zwei Schritten. Zunächst wird eine Reihe von stochastischen Simulationen durchgeführt, jede dieser Simulationen liefert eine Faltungstrajektorie. Die Zustände dieser Trajektorien sind entsprechend der Master Equation nach Definition 19 verteilt. Zusätzlich sind Start- und Endzeitpunkt jedes Zustands bekannt. Für jeden Zustand kann nun die Anzahl der Trajektorien ermittelt werden, in denen dieser zu einem bestimmten Zeitpunkt auftritt. Diese Häufigkeit bildet mit steigender Zahl berechneter Trajektorien eine immer bessere Abschätzung der Auftretenswahrscheinlichkeiten der Zustände. Die Durchführung der stochastischen Simulation erfolgt durch den in Abschnitt 4.1 beschriebenen Gillespie Algorithmus.

Die Berechnung von Kinetiken über die Übergangswahrscheinlichkeiten des Makrostate-Prozesses erfolgt in drei Schritten. Zunächst müssen die Zustände der Makrostates ermittelt und die Übergangsraten zwischen diesen Zuständen berechnet werden. Aus der resultierenden Matrix der Übergangsraten kann nun die Matrix der Übergangswahrscheinlichkeiten berechnet werden. Über die Matrix der Übergangswahrscheinlichkeiten kann dann problemlos die Wahrscheinlichkeitsverteilung der Zustände für beliebige Zeitpunkte ermittelt werden. Die Berechnung der Ratenmatrix erfolgt durch den in Abschnitt 4.2 vorgestellten Landscape Flooding Algorithmus. Die Umwandlung in die Matrix der Übergangswahrscheinlichkeiten und die Berechnung der eigentlichen Kinetik erfolgen durch das in Abschnitt 4.3 vorgestellte Programm `treekin`.

Für diese Diplomarbeit wurden von mir die Energielandschaft für Hybridisierungen, der Gillespie Algorithmus sowie der Landscape Flooding Algorithmus implementiert. Die Implementation erfolgte über die Energy Landscape Library (ELL) [22], einer Bibliothek generischer Algorithmen zur Untersuchung von Kinetiken.[1] Die ELL stellt eine abstrakte Repräsentation der Energielandschaft zur Verfügung. Durch Ableitung von dieser abstrakten Klasse können beliebige Energielandschaften definiert werden. Alle in der ELL integrierten Algorithmen operieren auf dieser abstrakten Repräsen-

[1] Die ELL ist unter http://www.bioinf.uni-freiburg.de/SW/ELL verfügbar.

tation und sind damit für beliebige Energielandschaften einsetzbar.

Die ELL stellt einige Algorithmen zur Traversierung von Energielandschaften bereit, darunter den für die Definition der Basins verwendeten Gradient Walk. Außerdem sind, unter anderem, Energielandschaften für Gitterproteine sowie die in Abschnitt 2.5 vorgestellte Energielandschaft für RNA enthalten. Die für diese Arbeit durchgeführten Kinetiken von RNA Faltungen wurden durch Nutzung dieser Energielandschaft durchgeführt.

4.1 Gillespie

Eine Möglichkeit, die eigentlichen Faltungskinetiken zu berechnen, besteht in der stochastischen Simulation des Markov-Prozesses durch eine Rejectionless Monte Carlo Methode. Der hier vorgestelle Algorithmus basiert auf der von Daniel Gillespie in [12, 13] vorgestellten Methode für zeitstetige Simulationen. Der Gillespie Algorithmus realisiert einen Random Walk auf der Energielandschaft, dessen Zustände entsprechend der Master Equation verteilt sind. Eine Simulation des Faltungsvorgangs durch den Gillespie Algorithmus resultiert in einer einzelnen Faltungstrajektorie, also einer zeitlich geordneten Folge von Zuständen. Durch Berechnung einer großen Anzahl von Faltungstrajektorien kann für jeden Zustand die Auftretenswahrscheinlichkeit zu einem beliebigen Zeitpunkt der Simulation ermittelt werden. Die Notation des Algorithmus in Pseudocode findet befindet sich auf Seite 44.

Die Simulation beginnt mit einem Startzustand zum Zeitpunkt $t = 0$. Ausgehend von diesem Zustand erfolgt die Bestimmung eines Nachfolgezustands sowie die Berechnung der für diesen Übergang benötigten Zeit dt. Der gewählte Nachfolgezustand wird nun zum aktuellen Zustand und die für diesen Übergang ermittelte Zeitspanne dt wird der aktuellen Zeit t aufgeschlagen. Diese Wahl neuer Nachfolgezustände erfolgt so lange, bis die während der Simulation vergangene Zeit t einen zuvor gewählten Zeitraum überschreitet.

Demnach müssen in jedem Simulationsschritt zwei Fragen beantwortet werden. Zunächst muss ein Nachfolger für den aktuellen Zustand ermittelt werden. Außerdem soll die für diesen Übergang benötigte Zeit bestimmt werden.

Der Übergang von einem Zustand x zu einem Zustand y soll mit Wahrscheinlichkeit

$$\Pr(x \to y) = \frac{k_{xy}}{\sum_{z \in N(x)} k_{xz}} \tag{4.1}$$

erfolgen.

Diese Wahl eines Nachfolgezustands mit Wahrscheinlichkeit $\Pr(x \rightarrow y)$ wird durch den Algorithmus auf folgende Weise umgesetzt: Zunächst werden die Übergangsraten zu allen Nachbarn von x aufsummiert. Nun erfolgt die Wahl einer Zufallszahl aus dem Intervall $\left[0, \sum_{z \in N(x)} k_{xz}\right]$ als Schwellenwert. Die Übergangsraten zu den Nachbarzuständen werden ein weiteres mal so lange aufsummiert, bis diese neue Summe die gewählte Schwelle überschreitet. Der beim Überschreiten des Schwellenwerts aktuelle Nachbar wird als Nachfolgezustand gewählt.

Die für den Übergang zu diesem Zustand benötigte Zeit dt wird über die Wahrscheinlichkeit bestimmt, dass im Zeitraum zwischen t und $t + dt$ kein Übergang zu einem der Nachbarn des Ausgangszustands erfolgt und ist damit abhängig von den Übergangsraten aller Nachbarn. Diese Zeit dt soll nach der Exponentialverteilung mit Mittelwert

$$\mu = \frac{1}{\sum_{z \in N(x)} k_{xz}}$$

gewählt werden. Durch Verwendung der Inversen der Verteilungsfunktion lassen sich unter Angabe einer im Intervall [0,1] gleichverteilt gezogenen Zufallszahl, nach der Exponentialverteilung verteilte Zufallszahlen generieren. Die Bestimmung der Übergangszeit kann also durch Auswertung der Inversen der Exponentialverteilung mit Mittelwert μ erfolgen. Unter Angabe einer Zufallszahl $rnd \in [0, 1]$ erfolgt die Berechnung der Übergangszeit dt durch Auswertung von

$$dt = \ln\left(\frac{1}{rnd}\right)\frac{1}{\mu}.$$

4.1.1 Überprüfung der Implementation

Um die korrekte Implementation des Gillespie Algorithmus zu zeigen, wurde die Berechnung einer von Michael Wolfinger et al. in [41] beschriebenen Faltungskinetik für die sogenannte xbix-Sequenz wiederholt. Dabei handelt es sich um eine künstliche Basensequenz mit nichttrivialem Faltungsverhalten. Die xbix-Sequenz besteht aus lediglich 20 Nukleotiden und kann 3886 Sekundärstrukturen ausbilden. Diese Größe ermöglicht noch die direkte Berechnung der Faltungskinetik durch Integration der Ratenmatrix über den vollständigen Markov-Prozess der Energielandschaft.

Durch die bereits in der ELL vorhandene Beschreibung einer Energielandschaft über die Sekundärstrukturen einzelner RNAs unter Verwendung der Moves Insertion, Deletion und Shift, konnte eine stochastische Simulation der Faltungskinetik von xbix durchgeführt werden. Für diese Kinetik wurden 10.000 Simulationsläufe bis zu einer Zeit von 10^6 unter Verwendung der Metropolis Übergangrate durchgeführt.

Abbildung 4.1 zeigt diese Kinetik neben einer aus der Arbeit von Michael Wolfinger entnommenen Kinetik. Diese wurde ebenfalls unter Verwendung der Metropolis

Algorithm 1: Gillespie

Eingabe : *maxTime* - Zeitbeschränkung des Walks, *origin* - Startzustand
Ausgabe: *trace* - Liste der Tupel (Zustand,Endzeit)

```
time ← 0
while (time < maxTime) do
    // Bestimmung der Summe der Übergangsraten
    flux ← 0
    forall (neighbor ∈ N(origin)) do
        flux ← flux + k(E(origin), E(neighbor))
    end
    totalFlux ← flux
    // Wahl des Nachfolgezustands
    threshold ← rnd([0, flux])
    flux ← 0
    forall (neighbor ∈ N(origin)) do
        flux ← flux + k(E(origin), E(neighbor))
        if (flux ≥ threshold) then
            origin ← neighbor
            // Berechnung der Übergangszeit
            randomNumber ← rnd([0, 1])
            deltaT ← log(1/randomNumber)/totalFlux
            time ← time + deltaT
            trace.add((origin, time))
            end forall
        end
    end
end
return trace
```

Übergangsrate erstellt, die Berechnung der Kinetik erfolgte jedoch über numerische Integration der Ratenmatrix des vollständigen Markov-Prozesses statt durch stochastische Simulation. Die resultierenden Kinetiken sind quantitativ und qualitativ nahezu identisch. Die über stochastische Simulationen erstellte Kinetik zeigt einige kleine Unregelmäßigkeiten in den Wahrscheinlichkeitsverläufen, dies sollte aber durch eine Erhöhung der Anzahl der Simulationsläufe behoben werden können.

Damit ist die korrekte Funktionsweise der von mir durchgeführten Implementation des Gillespie Algorithmus für ein Beispiel nachgewiesen.

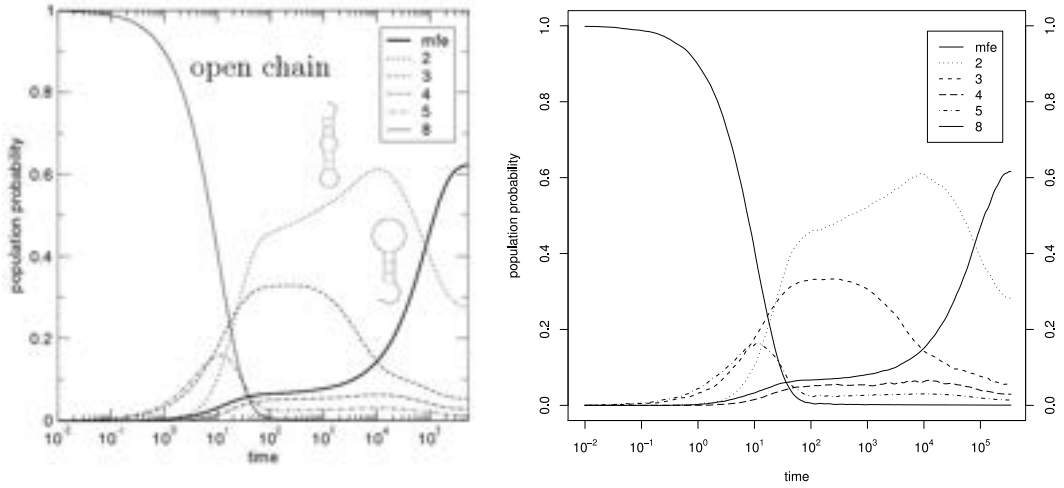


Abbildung 4.1: Faltungskinetiken von xbix. Die Berechnung dieser Kinetiken erfolgte unter Verwendung der Metropolis Übergangsraten. (links) Berechnung der Kinetik durch numerische Integration der Ratenmatrix des Markov-Prozesses über die vollständige Energielandschaft. Entnommen aus [41]. (rechts) Berechnung der Kinetik durch Auswertung von 10.000 stochastischen Simulationen durch den Gillespie Algorithmus.

4.2 Landscape Flooding

Für die Berechnung der Matrix der Übergangsraten K für einen Markov-Prozess über Makrostates nach Definition 29 ist die Kenntnis der Basins der Energielandschaft sowie deren Kontaktflächen nötig. Eine naive Methode zur Berechnung der Basins ist die Durchführung eines Gradient Walks für jeden Zustand der Energielandschaft. Eine effizientere Möglichkeit ist das Fluten der Energielandschaft, das Landscape Flooding. Diese Methode zur Bestimmung der Makrostates einer Energielandschaft wurde von Flamm et al. in [9] vorgeschlagen. Die Notation des Algorithmus in Pseudocode findet befindet sich auf Seite 47.

Unter Landscape Flooding wird allgemein die sequentielle Analyse aller Strukturen einer Energielandschaft verstanden, wobei diese nach aufsteigender Energie sortiert abgearbeitet werden. Da die verwendeten Energielandschaften degeneriert sind, also mehrere Zustände die gleiche Energie haben können, wird hier die totale Ordnung $<$ zur Ordnung der Zustände genutzt. Dies garantiert die eindeutige Reihenfolge der Abarbeitung der Zustände. Die Verwaltung der noch abzuarbeitenden Zustände erfolgt durch eine Priority Queue. Das in die Priority Queue eingeordnete Element niedrigster Ordnung wird als top-Element bezeichnet.

Vor Beginn des eigentlichen Flutens werden alle lokalen Minima der Energielandschaft ermittelt und in Verbindung mit dem Index des zugehörigen Basins in die Priority Queue eingefügt. Beim Fluten wird nun so lange das jeweilige top-Element aus der Priority Queue entnommen und verarbeitet, bis die Queue keine weiteren Elemente mehr enthält. Diese Verarbeitung besteht aus der Ermittlung der Nachbarn des top-Elements. Im Fall eines größeren Nachbarn nach der Ordnung $<$ wird dieser, falls nicht bereits in einem der vorherigen Schritte geschehen, in die Priority Queue eingefügt.

Zusätzlich sollen jedoch die Basins der Energielandschaft und deren Kontaktflächen ermittelt werden. Dazu ist eine zusätzliche Fallunterscheidung während der Analyse eines Nachbarn höherer Ordnung nötig. Nachbarn von niedrigerer Ordnung als der aktuelle Zustand müssen nicht mehr betrachtet werden, da sie aufgrund der Bearbeitungsreihenfolge schon bearbeitet, und die Übergänge von und zu diesen Nachbarn dementsprechen bereits berechnet wurden. Bei Nachbarn höherer Ordnung werden zwei Fälle unterschieden: der Nachbar kann bereits Teil der Priority Queue sein, oder nicht. Bezeichne *top* das aktuelle top-Element und *neighbor* den gerade betrachteten Nachbarn.

1. *Der Nachbar des top-Elements ist noch nicht in der Priority Queue enthalten.* Daraus folgt, dass das top-Element der niedrigste Nachbar dieses Zustands sein muss, denn andernfalls wäre dieser Nachbar bereits bei der Abarbeitung eines der vorherigen top-Elemente niedrigerer Ordnung in die Priority Queue eingefügt worden. Damit gilt für die Endpunkte der Gradient Walks beginnend bei den beiden Zuständen $\hat{w}(top) = \hat{w}(neighbor)$, die Zustände sind also Teil des gleichen Basins. Da die Basinzugehörigkeit der in der Priority Queue verwalteten Zustände bekannt ist, kann diese für *neighbor* übernommen werden und der Nachbar wird in Verbindung mit dem Index seines Basins in die Priority Queue eingefügt.

2. *Der Nachbar des top-Elements ist bereits in der Priority Queue enthalten.* Dann ist die Basinzugehörigkeit beider Zustände bekannt. Falls die Zustände Teil verschiedener Basins sind, bilden sie also einen Teil der Kontaktfläche dieser Basins.

Algorithm 2: Landscape Flooding

Eingabe : *minima* - Liste der Minima
Ausgabe: K - Matrix der Übergangsraten zwischen den Makrostates

```
PriorityQueue queue
Z[1,...,|minima|] ← 0                    /* Kanonische Zustandssummen der Basins */
C(i,j)_{1≤i≤|minima|, 1≤j≤|minima|} ← 0   /* Übergangsraten zwischen Basins */

// Vorbereitung der Datenstruktur
forall i ∈ {1,...,|minima|} do
    queue ← minima[i]                                  /* registriere Minima */
    queue(m[i]) ← i                        /* speichere das zugeordneten Basin */
end

// Flutung der Landschaft
while (queue not empty) do
    top ← queue.top
    currentEnergy ← E(top)
    currentBasin ← queue(top)

    // aktualisiere Summe der Boltzmann-Gewichte für dieses Basin
    Z[currentBasin] ← Z[currentBasin] + exp(−currentEnergy/RT)

    forall (neighbor ∈ N(top)) do
        if (top < neighbor) then
            if (neighbor ∉ queue) then      /* erstes Auftreten des Zustands */
                queue ← neighbor
                queue(neighbor) ← currentBasin
            else                    /* Nachbar ist Kandidat für Kontaktfläche */
                neighborBasin ← queue(neighbor)
                if currentBasin ≠ neighborBasin then
                    // aktualisiere Energien der Kontaktflächen
                    C[currentBasin, neighborBasin] ←
                        C[currentBasin, neighborBasin]+
                        k(currentEnergy, E(neighbor)) exp(−currentEnergy/RT)
                    C[neighborBasin, currentBasin] ←
                        C[neighborBasin, currentBasin]+
                        k(E(neighbor), currentEnergy)) exp(−E(neighbor)/RT)
                end
            end
        end
    end
end

// Berechnung der Übergangsraten
K(i,j)_{1≤i≤|minima|, 1≤j≤|minima|} ← 0
forall (i ∈ {1,...,|minima|}) do
    forall (j ∈ {1,...,|minima|}) do
        K(i,j) ← C[i,j]/Z[i]
    end
end

return K
```

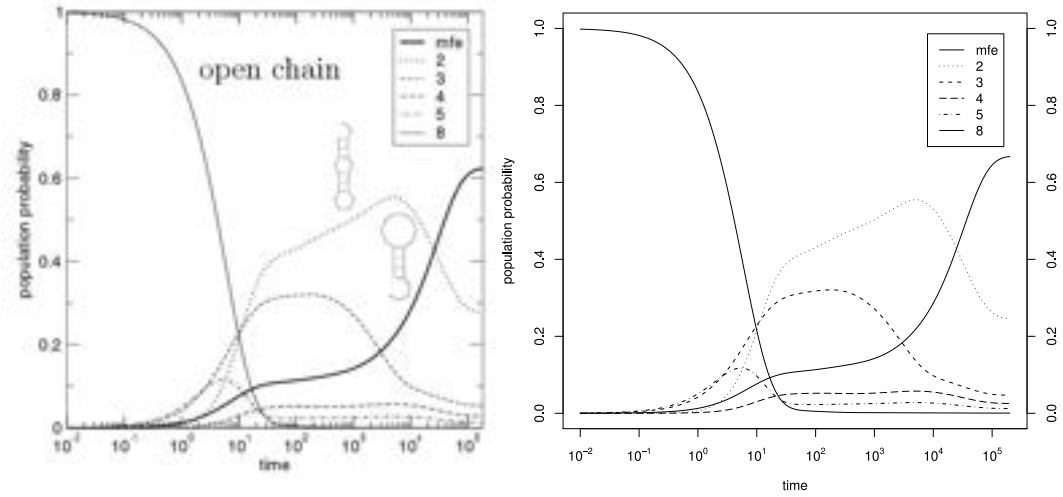


Abbildung 4.2: Faltungskinetiken von xbix. Für die Berechnung dieser Kinetiken wurde der Makrostate-Prozess unter Verwendung der Metropolis Übergangsraten benutzt. (links) Entnommen aus [41]. (rechts) Berechnung durch den für diese Arbeit implementierten Flutungsalgorithmus und `treekin`.

4.2.1 Überprüfung der Implementation

Zur Überprüfung der korrekten Implementation des Landscape Flooding Algorithmus wurde die Matrix der Übergangsraten für den Makrostate-Prozess der xbix-Sequenz berechnet. Die Berechnung der Übergangsraten erfolgte nach Metropolis. Für die anschließende Berechnung der Matrix der Übergangswahrscheinlichkeiten und der daraus resultierenden Kinetik wurde das Tool `treekin` benutzt. Abbildung 4.2 zeigt diese Kinetik neben einer aus [41] entnommenen Kinetik. Diese wurde ebenfalls unter Verwendung der Metropolis Übergangsraten durch Integration der Matrix der Übergangsraten des Makrostateprozesses berechnet. Die beiden Kinetiken sind identisch. Damit wurde die korrekte Funktionsweise der Implementation des Landscape Flooding Algorithmus für ein Beispiel nachgewiesen.

4.3 Berechnung der Makrostate-Kinetik

Nachdem die Matrix der Übergangsraten zwischen den Makrostates der Energielandschaft mittels des Landscape Flooding Algorithmus berechnet wurde, muss diese in die Matrix der Übergangswahrscheinlichkeiten überführt werden. Dies und die anschließende Berechnung der eigentlichen Kinetik werden durch das Programm `treekin` durchgeführt. Dieses wurde von Michael Wolfinger im Rahmen seiner Diplomarbeit “The energy landscape of RNA folding” [42] entwickelt. Die Diplomarbeit von Hannes Kochniß [19] enthält einen guten Überblick über die von `treekin` durch-

geführten Berechnungen, an denen sich die folgende Beschreibung orientiert.

Der Zusammenhang zwischen Übergangswahrscheinlichkeiten und Übergangsraten wird durch die Master Equation aus Gleichung 2.2 beschrieben. Die Lösung dieser Differentialgleichung ist

$$\overrightarrow{P_t} = e^{tK}\overrightarrow{P_0}.$$

Für die Berechnung Wahrscheinlichkeitsverteilung der Zustände $\overrightarrow{P_t}$ muss also das Matrixexponential e^{tK} gelöst werden.

Zur Berechnung dieses Matrixexponentials durch `treekin` werden Diagonalmatrizen verwendet. Als solche werden quadratische Matrizen bezeichnet, deren Elemente nur auf der Hauptdiagonale einen von Null verschiedenen Wert annehmen. Sei Λ eine Diagonalmatrix. Dann ist $U = e^{t\Lambda}$ ebenfalls eine Diagonalmatrix mit den Werten

$$u_{ii} = e^{t\lambda_{ii}}, \tag{4.2}$$

kann also leicht berechnet werden.

Um dies ausnutzen zu können, muss K zunächst in eine Diagonalmatrix umgewandelt werden. Nach dieser sogenannten Diagonalisierung gilt $K = N\Lambda M$ mit den Restmatrizen N und M sowie der diagonalen Matrix Λ, welche die Eigenwerte von K enthält. Nach dieser Diagonalisierung lässt sich das Matrixexponential

$$e^{tK} = Ne^{t\Lambda}M$$

durch Anwendung von Gleichung 4.2 lösen. Die für die Berechnung der Wahrscheinlichkeitsverteilungen $\overrightarrow{P_t}$ nötige Diagonalisierung der Ratenmatrix kann für n Makrostates in Zeit $O(n^3)$ unter Verwendung von $O(n^2)$ Speicher durchgeführt werden.

5 Ergebnisse

In diesem Kapitel werden zunächst die Energielandschaften von Hybridisierungen untersucht. Es erfolgt eine Abschätzung der Größe des Zustandsraums in Abhängigkeit der verwendeten Sequenzen. Außerdem wird die Auswirkung auf die Berechenbarkeit des Makrostate-Prozesses erläutert. Nach einem Vergleich zwischen stochastischer Simulation sowie der Berechnung der Übergangswahrscheinlichkeiten des Makrostate-Prozesses werden über Kawasaki und Metropolis Übergangsraten berechnete Kinetiken gegenübergestellt. Abschließend erfolgen ein Vergleich zur Strukturvorhersage sowie die Berechnung von Kinetiken einer validierten Hybridisierung.

Verwendete Sequenzpaare

Bei der RNA-Interferenz handelt es sich um einen Mechanismus zur Genregulation, bei dem zielerkennende RNAs Hybridisierungen mit mRNAs ausbilden. Bei diesen zielerkennenden RNAs kann es sich um siRNA oder miRNA handeln. Während die miRNA und siRNA Moleküle lediglich aus etwa 20 Nukleotiden bestehen, enthalten mRNAs deutlich mehr Nukleotide. Der Bereich der Interaktion zwischen den beiden Molekülen, die Ziel- oder Targetregion, ist jedoch relativ kurz. Die mRNA einer solchen Interaktion wird auch als Target-Sequenz bezeichnet.

Für die folgenden Untersuchungen wurden ausschließlich Sequenzpaare aus mRNA und siRNA sowie mRNA und miRNA betrachtet. Die bei der RNA-Interferenz auftretenden Interaktionen haben eine geringe Komplexität, wodurch sie gut für anfängliche Betrachtungen geeignet sind. Für die Vorhersage dieser Interaktionen sind viele Programme verfügbar. Dadurch lassen sich die Ergebnisse der Kinetiken leicht mit denen der thermodynamischen Strukturvorhersage vergleichen. Zudem kann die Länge der mRNA unter Beibehaltung der vorhergesagten oder experimentell bestimmten Zielregion variiert werden, um beispielsweise den Aufwand für die Berechnung der Kinetiken zu senken.

Konvention zur Darstellung der Kinetiken

Abbildung 5.1 zeigt zwei Darstellungen einer durch Integration der Ratenmatrix zwischen Makrostates erstellten Kinetik. Die Energielandschaft dieser Kinetik enthält 1185 Basins, deren Wahrscheinlichkeitsverläufe links abgebildet sind. Hier ist erkennbar, dass nur eine geringe Zahl der Makrostates jemals eine signifikante Auftretenswahrscheinlichkeit erreicht.

Die Makrostates mit geringer Höchstwahrscheinlichkeit scheinen in drei Kategorien zu fallen. Eine große Anzahl von Makrostates hat während des Faltungsvorgangs eine relativ konstante Wahrscheinlichkeit nahe 0%. Einige Makrostates erreichen ihr Maximum etwa in dem Bereich, in dem das Basin der offenen Struktur aufhört, der wahrscheinlichste Makrostate zu sein. Die Wahrscheinlichkeit dieser Makrostates sinkt im zweiten Abschnitt der Kinetik jedoch ebenfalls auf sehr niedrige Niveaus ab. Die dritte Gruppe zeigt in der zweiten Hälfte des Faltungsvorgangs einen stetigen Zuwachs, die maximale Wahrscheinlichkeit dieser Makrostates liegt aber generell unter 5%.

Die rechte Seite der Abbildung zeigt eine Auswahl der Wahrscheinlichkeitsverläufe, für die während der Faltung eine maximale Wahrscheinlichkeit von mindestens 5% erreicht wird. Die Makrostates sind mit der Nummer des assoziierten lokalen Minimums bezeichnet, die Nummerierung der lokalen Minima erfolgt sortiert nach aufsteigender Energie. Ausgenommen von dieser Regel sind die Makrostates der Hybridisierung, deren Interaktion keine Basenpaare enthält, und der mfe-Hybridisierung. Diese werden mit open und mfe bezeichnet. Die Konventionen dieser Darstellung gelten, falls nicht anders angegeben, für alle weiteren Kinetiken in dieser Arbeit.

Alle in diesem Kapitel verwendeten Sequenzen sind in Anhang A aufgeführt.

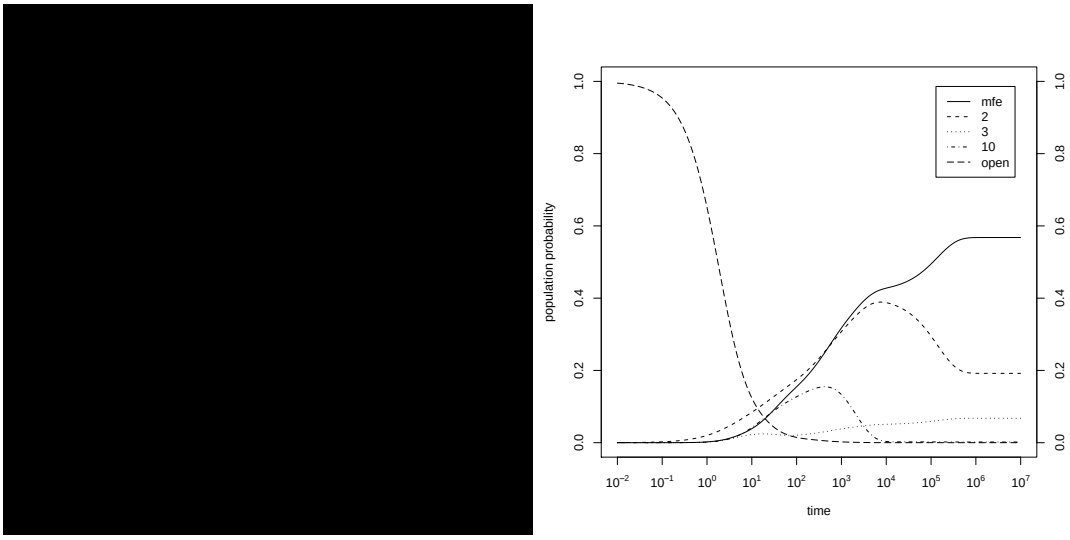


Abbildung 5.1: Kinetiken von hsa-miR-1 und KCNE1, berechnet über den Makrostate-Prozess. (links) Vollständiger Plot über alle 1158 Basins. (rechts) Auswahl der Basins deren Wahrscheinlichkeit ein Maximum von mindestens 5% erreicht.

5.1 Energielandschaft von Hybridisierungen

Die Energielandschaft der Hybridisierungen bildet die Basis für die Berechnung der Kinetiken. Darum werden in diesem Abschnitt die Eigenschaften der Energielandschaft sowie deren Einfluss auf die Berechnung der Kinetiken betrachtet.

5.1.1 Überprüfung der Implementation des Move Set

Voraussetzung für die Konvergenz der Markov-Prozesse im thermodynamischen Equilibrium ist die Ergodizität des Move Set. Um diese für die Implementation des verwendeten Move Set an einem Beispiel aufzuzeigen, wurden zwei Algorithmen zur Aufzählung des vollständigen Zustandsraums entworfen.

Der erste Algorithmus setzt auf der Implementation der Energielandschaft auf. Dieser Algorithmus benutzt zwei Datenstrukturen für die Verwaltung der Zustände - eine Menge noch nicht bearbeiteter Zustände *todo* und eine Menge noch zu bearbeitender Zustände *seen*. Zu Beginn des Algorithmus wird *todo* mit der offenen Struktur initialisiert. Der Algorithmus endet, sobald *todo* keine Zustände mehr enthält.

So lange die Menge *todo* noch Elemente enthält, wird eines dieser Elemente entnommen, verarbeitet und in die Menge *seen* eingefügt. Dabei werden alle Nachbarn des entnommenen Elements generiert. Falls einer dieser Nachbarn noch nicht bearbeitet wurde und auch nicht zur Bearbeitung vorgemerkt wurde, also weder in *todo* noch in *seen* enthalten ist, wird er in die Menge der noch zu bearbeitenden Zustände *todo* eingefügt. Der Algorithmus generiert also die Nachbarn aller von der offenen Hybridisierung aus erreichbaren Strukturen.

Es ist leicht einzusehen, dass nach Beendigung des Algorithmus alle Zustände, die von der offenen Struktur aus erreicht werden können in *seen* enthalten sind. Nach Definition 5 ist ein Move Set ergodisch, wenn zwischen jedem Paar von Zuständen der Energielandschaft ein Pfad miteinander benachbarter Zustände existiert. Durch die Aufzählung dieses Algorithmus wurde nun gezeigt, dass ein solcher Pfad zwischen der offenen Struktur und jeder anderen Struktur in *seen* existiert. Zusammen mit der Symmetrie des Move Set folgt daraus, dass ein solcher Pfad zwischen allen in *seen* enthaltenen Strukturen existiert. Falls *seen* dem vollständigen Strukturraum entspricht, wurde demnach Ergodizität für das betrachtete Beispiel nachgewiesen.

Um die Vollständigkeit des durch den ersten Algorithmus berechneten Zustandsraums zu zeigen, wurde ein zweiter, von der Implementation der Energielandschaft unabhängiger, Algorithmus entworfen. Dieser generiert alle nach Definition 51 gültigen Hybridisierungsstrukturen auf eine vom Move Set unabhängige Weise. Durch den Vergleich dieser Strukturmenge mit der Menge *seen* des ersten Algorithmus kann also die Ergodizität der Implementation des Move Set für beliebige Beispiele nachgewiesen werden.

Dieser Nachweis wurde für eine mRNA der Länge 182 sowie eine siRNA der Länge 19 aus [33] für ein Hybridisierungsfenster von 14 Nukleotiden erbracht. Von beiden Algorithmen wurden äquivalente Zustandsräume mit jeweils 18.936.277 Hybridisierungsstrukturen generiert. Damit ist die Funktionalität der Implementation des Move Set für dieses Beispiel nachgewiesen.

5.1.2 Größe des Zustandsraums

Die Größe des Zustandsraums der Energielandschaft hat direkten Einfluss auf Laufzeit und Speicherbedarf der verwendeten Algorithmen. Insbesondere die Laufzeiten für das Finden der Minima sowie das Fluten der Landschaft sind von dieser direkt abhängig. Darum ist es wichtig, die Anzahl möglicher Hybridisierungen für zwei gegebene Sequenzen zumindest abschätzen zu können. Dabei ist zu erwarten, dass die Anzahl der Hybridisierungen sowohl von der Länge der verwendeten Sequenzen als auch von der Größe des verwendeten Hybridisierungsfensters abhängt.

Bei einer einzelnen RNA steigt die Anzahl der möglichen Sekundärstrukturen exponentiell zur Länge der Nukleotidsequenz. Im Vergleich zu Sekundärstrukturen einzelner RNA ist der Strukturraum bei den in dieser Arbeit modellierten Hybridisierungen eingeschränkt. Zum einen werden keine Sekundärstrukturelemente innerhalb der einzelnen RNAs erlaubt, zum anderen wurde die Menge der erlaubten Strukturen durch Einführung des Hybridisierungsfensters zusätzlich eingeschränkt. Wie sich die Größe des Strukturraums in Abhängigkeit der Eingabesequenzen und des Hybridisierungsfensters verhält, soll im Folgenden geklärt werden.

Theoretische Überlegungen

Für einzelne RNAs ist bekannt, dass sich die Zahl der möglichen Sekundärstrukturen exponentiell zur Länge der Nukleotidsequenz verhält. Dies konnte von Michael S. Waterman durch eine Untersuchung der kombinatorischen Eigenschaften von Sekundärstrukturen nachgewiesen werden [39]. Demnach ist die Anzahl der möglichen Sekundärstrukturen für eine Nukleotidsequenz der Länge n durch $O(4^n)$ begrenzt.

Was ist nun für die Anzahl der möglichen Hybridisierungen zu erwarten? Für vergleichbare Sequenzen sollten weniger Hybridisierungen als Sekundärstrukturen ausbildet werden können. Angenommen eine Sequenz S der Länge $n + m$ wird nach Position n in zwei Sequenzen S_1 und S_2 der Längen n und m mit $S = S_1S_2$ aufgeteilt. Die Zahl der von S gebildeten Sekundärstrukturen muss nun immer größer sein, als die Zahl der von S_1 und S_2 gebildeten Hybridisierungen. Die Anzahl erlaubter Basenpaare ist für die Hybridisierungen im Vergleich zu den Sekundärstrukturen eingeschränkt, da lediglich Basenpaarungen erlaubt sind, deren öffnende Bindungen auf der Sequenz S_1 und deren schließende Bindung auf der Sequenz S_2 liegen. Durch

dieses Verbot von Strukturen innerhalb der Sequenzen S_1 und S_2 reduziert sich sich die Menge möglicher Sekundärstrukturelemente in Hybridisierungen um Hairpin Loops und k-Multiloops auf Stackings, Bulges und innere Schleifen. Zusätzlich wird die Zahl der Hybridisierungen durch die Beschränkung der Länge der an der Hybridisierungsstruktur beteiligten Sequenzabschnitte eingeschränkt.

Abhängigkeit von Target Sequenz und Fenstergröße

Für eine erste Untersuchung der Größe des Strukturraums für Hybridisierungen wurde ein in [33] verwendetes Paar von Sequenzen verwendet. Die verwendete mRNA, ein Abschnitt einer für Luciferase kodierenden mRNA, hat eine Länge von 182 Nukleotiden. Die zugehörige siRNA besteht aus 19 Nukleotiden.

Zunächst wurde die Länge der mRNA variiert und die Anzahl der möglichen Hybridisierungen aufgetragen. Die Hybridisierungen unterlagen keiner Beschränkung der Fenstergröße. Das Ergebnis dieses Versuches zeigt der linke Graph von Abbildung 5.2. Hier ist ein exponentieller Zuwachs der Anzahl möglicher Hybridisierungen in Abhängigkeit der Länge der mRNA zu erkennen.

Zusätzlich wurde unter Verwendung der vollständigen Sequenz der mRNA die zulässige Fenstergröße variiert und wiederum die Anzahl möglicher Hybridisierungen ermittelt. Das Ergebnis dieser Untersuchung zeigt der rechte Graph aus Abbildung 5.2. Die Anzahl der Hybridisierungen steigt exponentiell zur Größe des Fensters. Auffällig ist ein "Knick" bei einer Fenstergröße von 18 Nukleotiden. Ab dieser Länge steigt die Anzahl der Hybridisierungen weniger stark an. Dies ließ einen Zusammenhang zur Länge der siRNA von 19 Nukleotiden vermuten. Diese Vermutung konnte in einem weiteren Experiment bestätigt werden.

Abhängigkeit von der Länge der miRNA

Als Datenbasis für diese Evaluierung wurde die miRNA hsa-miR-1 ausgewählt. Diese besteht aus 22 Nukleotiden. In der miRecords Datenbank[1] werden für diese miRNA 107 experimentell bestätigte Zielsequenzen aufgeführt, darunter die mRNA KCNE1. Für diese mRNA werden drei vorhergesagte Zielregionen angegeben.[2] Von dieser Sequenz wurden drei Abschnitte der Länge 179 dergestalt ausgewählt, dass sich der Beginn der vorhergesagten Zielregion jeweils in der Mitte befand. Durch die Wahl dieser Versuchsparameter war eine gezielte Variation der Zielsequenz bei Beibehaltung aller anderen Parameter, insbesondere der miRNA Sequenz gewährleistet.

[1] http://mirecords.umn.edu/miRecords/index.php

[2] Die Zielregionen wurden durch miRanda, MirTarget2, NBmiRTar, PITA, RNA22 und RNAhybrid vorhergesagt.

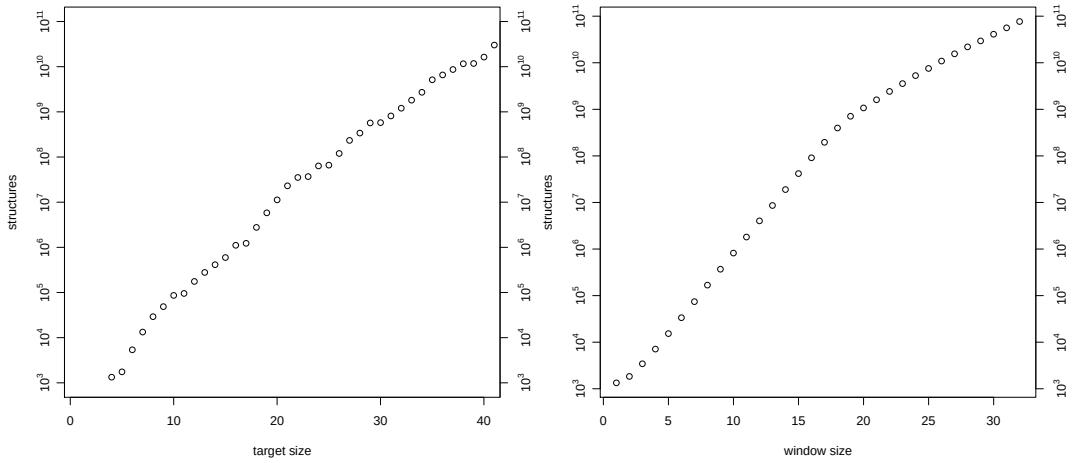


Abbildung 5.2: Größe des Strukturraums in Abhängigkeit von Target Sequenz und Fenstergröße. (links) Anzahl gültiger Hybridisierungen für eine siRNA der Länge 19 und eine mRNA variabler Länge. Die Strukturen unterliegen keinen Einschränkungen bezüglich eines Hybridisierungsfensters. (rechts) Anzahl gültiger Hybridisierungen für eine siRNA der Länge 19 und eine mRNA der Länge 182 unter Variation des Hybridisierungsfensters.

Abbildung 5.3 a) zeigt die Entwicklung der Anzahl der Strukturen für die drei Sequenzpaare bei steigender Fenstergröße. Für jeden der mRNA Abschnitte gilt auch hier eine exponentielle Zunahme der Anzahl gültiger Hybridisierungen in Abhängigkeit der Fenstergröße. Alle betrachteten Sequenzpaare liefern sehr ähnliche Verläufe für die Anzahl von Strukturen. Die Größe des Strukturraums scheint also nur sehr gering von den konkret verwendeten Sequenzen abzuhängen. Auch diese Sequenzpaare zeigen ein Abflachen der Zuwachsrate ab einer Fenstergröße, die grob mit der Länge der miRNA übereinstimmt.

Um diesen Zusammenhang bestätigen zu können, wurde der Versuch mit einem der Targets und drei gekürzten Abschnitten der mRNA wiederholt. Die Sequenz der miRNA wurde auf 17, 12 und 7 Nukleotide beschnitten. Die Entwicklung des Strukturraums in Abhängigkeit von der Fenstergröße für diese drei Längen zeigt Abbildung 5.3 b). Damit konnte der vermutete Einfluss der Länge der miRNA auf die Entwicklung der Größe des Strukturraums nachgewiesen werden.

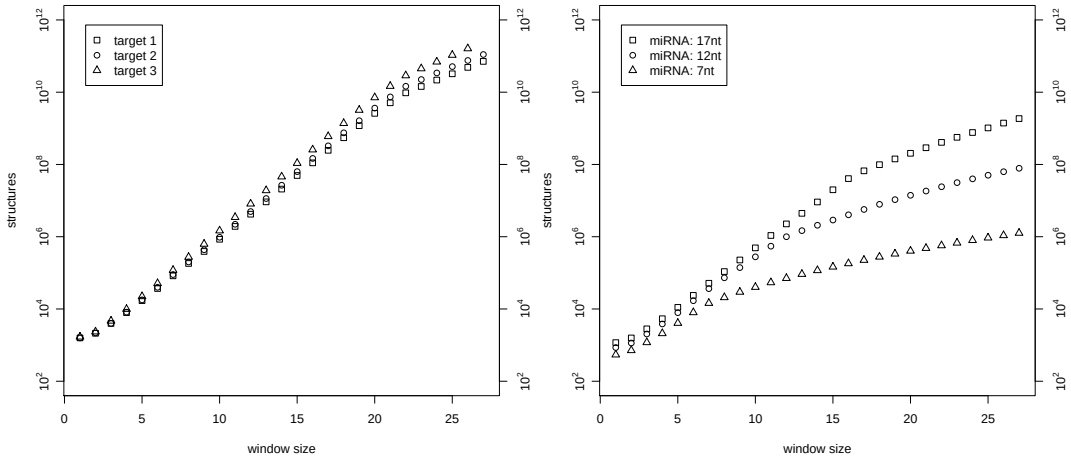


Abbildung 5.3: (links) Hybridisierung der miRNA hsa-miR-1 mit drei ausgewählten Abschnitten der mRNA KCNE1. miRNA: 22 Nukleotide, mRNA: 179 Nukleotide. Die Anzahl der Zustände wächst für alle drei mRNA-Abschnitte exponentiell zur Fenstergröße. Ab einer Fenstergröße von etwa 20 flacht die Kurve leicht ab. (rechts) Hybridisierung von drei gekürzten Abschnitten der miRNA hsa-miR-1 mit einem Abschnitt der mRNA KCNE1. Übersteigt die Fenstergröße die Länge der miRNA, nimmt die Anzahl der Strukturen weniger stark zu.

5.1.3 Speicherplatzbedarf der Makrostate-Kinetik

Für die Berechnung von Kinetiken über die Makrostate-Kinetik müssen zunächst die Basins der Energielandschaft sowie die Übergangsraten zwischen den Basins bestimmt werden. Diese Berechnung erfolgt durch das Fluten der Energielandschaft. Der Zeit- und Speicherplatzbedarf für das Fluten ist direkt abhängig von der Größe des Strukturraums. Dagegen hängen Zeit- und Speicherplatzbedarf für die anschließende Berechnung der Kinetik über das Programm `treekin` lediglich von der Anzahl der Basins ab. Darum wurde zunächst der Zusammenhang zwischen der Größe des Strukturraums und der Anzahl der darin enthaltenen Minima untersucht. Für diese Untersuchung wurden wieder die beiden aus [33] entnommenen Sequenzen der Längen 19 und 182 verwendet.

Abbildung 5.4 a) zeigt für diese Sequenzen die Anzahl von Strukturen und Minima in Abhängigkeit von der Fenstergröße. Es wird deutlich, dass zumindest in dem hier betrachteten Bereich die Anzahl der Minima exponentiell mit der gewählten Fenstergröße steigt. Abbildung 5.4 b) zeigt den Speicherplatzbedarf für das Fluten sowie die Diagonalisierung der Ratenmatrizen für dieses Beispiel. Der Speicherplatzbedarf für das Fluten der Energielandschaft übersteigt den für die Berechnung der Übergangswahrscheinlichkeiten benötigten Speicherplatz für alle betrachteten Fenstergrößen.

Bei einer Fenstergröße von 17 werden für das Fluten der Landschaft knapp 51 Gigabyte Hauptspeicher benötigt. Der für diese Experimente zur Verfügung stehende Rechner verfügt über 64 Gigabyte Hauptspeicher, damit ist dies das größte Fenster, für das eine Berechnung der Makrostate-Kinetik für dieses Sequenzpaar möglich ist. Durch die Verwendung kürzerer Sequenzen kann die maximal mögliche Fenstergröße weiter erhöht werden.

Bei der Verwendung von Paaren aus kleinen RNA und mRNA, wobei die mRNA auf den Bereich der vorhergesagten Zielregion beschnitten wurde, ist dieser Effekt jedoch gering, da die Zielregion viele zu den kleinen RNA komplementäre Nukleotide enthält.Für zwei Sequenzen der Länge 22 wurden bei einer Fenstergröße von 19 noch etwa 40 Gigabyte Hauptspeicher für das Fluten der Energielandschaft benötigt.

5.2 Vergleich der verwendeten Algorithmen

5.2.1 Stochastische Simulation und Markov-Prozess

Nun soll ein Vergleich zwischen einer durch stochastische Simulation sowie einer durch Berechnung der Übergangswahrscheinlichkeiten des Makrostate-Prozesses entstandenen Kinetik erfolgen. Für die Berechnung dieser Kinetiken wurde die mRNA hsa-miR-1 der Länge 22 sowie ein Abschnitt der mRNA KCNE1 der Länge 66 verwendet. Dieser Abschnitt wurde auch zur Berechnung aller folgenden Kinetiken von KCNE1 benutzt. Der Abschnitt enthält im mittleren Bereich die Nukleotide der

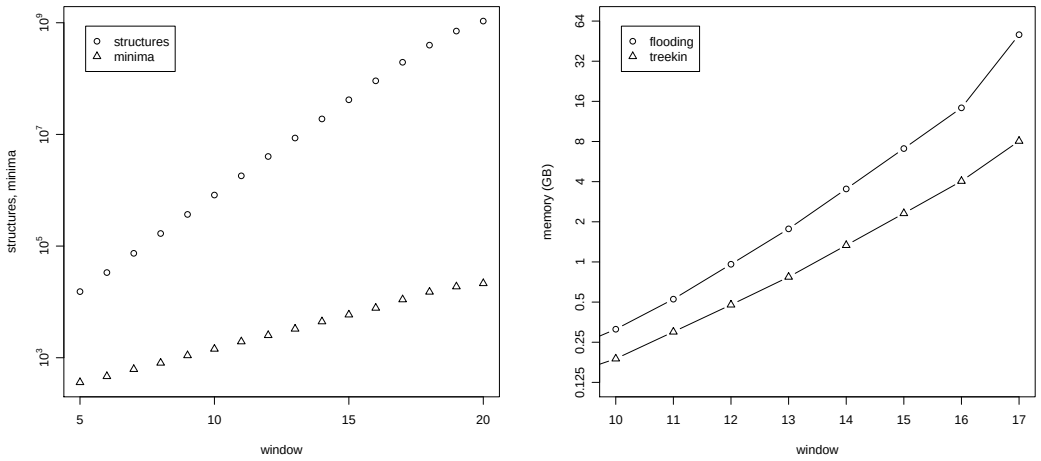


Abbildung 5.4: (links) Anzahl der Strukturen und Minima in Abhängigkeit von der Fenstergröße. (rechts) Speicherbedarf für das Fluten der Energielandschaft sowie die Berechnung der Übergangswahrscheinlichkeiten durch `treekin` in Abhängigkeit von der Fenstergröße.

vorhergesagten Hybridisierung. Die Berechnung der Kinetiken erfolgte unter Verwendung der Kawasaki Übergangsraten sowie einer Fenstergröße von 10. Abbildung 5.5 zeigt diese beiden Kinetiken.

Die Wahrscheinlichkeitsverläufe der stochastischen Simulation enthalten viele kleine Sprünge der Wahrscheinlichkeit, gegen Ende der Simulation kommt es noch einmal zu ausgeprägteren Unregelmäßigkeiten. Durch eine Erhöhung der Anzahl der Simulationsläufe sollte dies jedoch abgemildert werden können. Die Wahrscheinlichkeitsverläufe der zweiten Kinetik sind dagegen absolut glatt. Ansonsten ähneln sich die beiden Kinetiken sehr stark und sind von der allgemeinen Aussagekraft gleichwertig.

Für die Durchführung der stochastischen Simulationen wurden 14 Stunden benötigt, die Berechnung der zweiten Kinetik sowie die dafür nötige Flutung der Energielandschaft erfolgten jedoch in weniger als drei Minuten. Diese Art der Berechnung bietet also einen absolut überlegenen Geschwindigkeitsvorteil.

5.2.2 Metropolis und Kawasaki Übergangsraten

Mit Metropolis und Kawasaki stehen zwei verschiedene Methoden zur Berechnung von Übergangsraten zur Verfügung. In Abschnitt 2.3.5 wurden die Unterschiede zwischen den beiden Methoden analysiert. Um zu überprüfen, ob diese Feststellungen auch hinsichtlich Kinetiken von Hybridisierungen gültig sind, wurde die Kinetik für

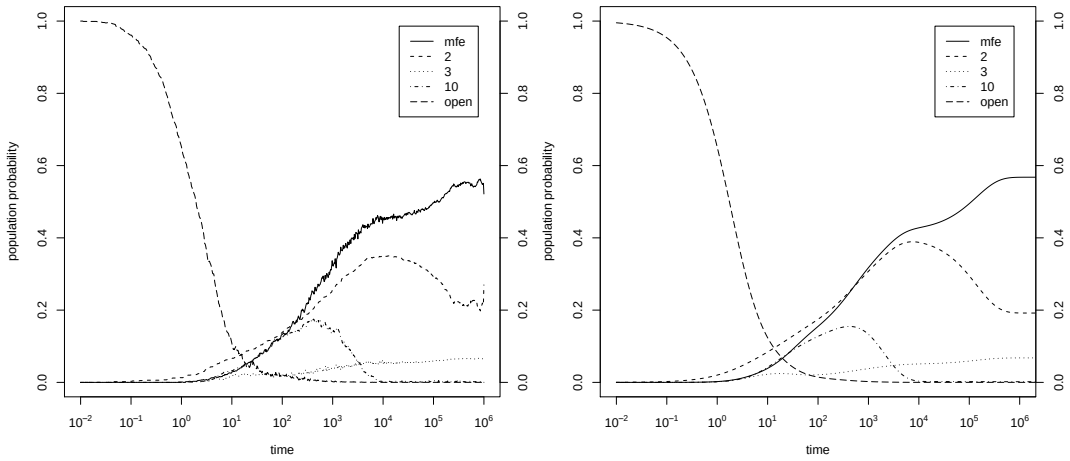


Abbildung 5.5: Vergleich zwischen stochastischer Simulation über die vollständige Energielandschaft und dem Markov-Prozess über Makrostates. (links) Kinetik berechnet durch Auswertungen von 1.000 Simulationen. (rechts) Kinetik des Makrostate-Prozesses.

die Sequenzen hsa-miR-1 und KCNE1 unter Nutzung beider Methoden berechnet. Abbildung 5.6 zeigt diese Kinetiken.

Die untersuchten Kinetiken weisen beträchtliche Unterschiede bezüglich der Höhe der erreichten Wahrscheinlichkeiten auf. Der Wahrscheinlichkeitsverlauf des Basins des ersten lokalen Minimums hat beispielsweise in der Kawasaki Kinetik ein um etwa zehn Prozent höheres Maximum als in der Metropolis Kinetik. Insgesamt sind die Veränderungen der Wahrscheinlichkeiten bei der Metropolis Kinetik schwächer ausgeprägt. Abgesehen davon sind die Wahrscheinlichkeitsverläufe beider Kinetiken jedoch sehr ähnlich.

Die über Kawasaki berechnete Kinetik erreicht das thermodynamische Equilibrium zum Zeitpunkt 10^6, die Metropolis Kinetik erreicht das Equilibrium bei etwa $10^{6,5}$. Die Kawasaki Kinetik konvergiert also in diesem Beispiel wie erwartet schneller als die Metropolis Kinetik. Die Nutzung der Kawasaki Raten hätte für dieses Beispiel einen beträchtlichen Geschwindigkeitsvorteil für die Durchführung einer stochastischen Simulationen zur Folge. Da für die Energielandschaft der Hybridisierungen ein minimales und lokales Move Set gewählt wurde, ist die Verwendung der Kawasaki Übergangsraten problemlos möglich. Aufgrund der schnelleren Konvergenz sollten die Übergangsraten nach Kawasaki auch bei der Durchführung von stochastischen Simulationen von Hybridisierungen bevorzugt werden.

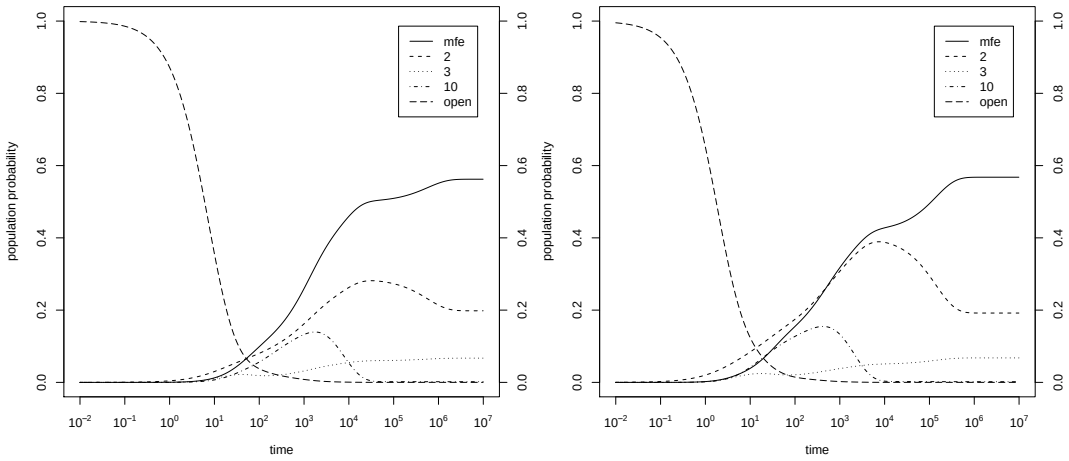


Abbildung 5.6: Kinetiken von hsa-miR-1 und KCNE1, berechnet über den Makrostate-Prozess. (links) Berechnung der Übergangsraten durch Metropolis. (rechts) Berechnung der Übergangsraten durch Kawasaki. Trotz der quantitativen Unterschiede sind sich die Kinetiken qualitativ sehr ähnlich.

5.3 Vergleich zur Strukturvorhersage

Die Untersuchung in Abschnitt 5.1.3 ergab, dass die Berechnung von Kinetiken über Makrostate-Prozesse nur bis zu einer maximalen Fenstergröße von 17 bis 19 durchgeführt werden kann. Um zu ermitteln, ob diese Größe zur Berechnung einer vorhergesagten Interaktion ausreicht, wurde eine Strukturvorhersage über `IntaRNA` durchgeführt. `IntaRNA` ist ein Programm zur Vorhersage von Interaktionen zwischen kleinen RNAs und mRNAs und wurde bereits erfolgreich zur Bestimmung von Targets kleiner RNAs eingesetzt [2]. Die von `IntaRNA` verwendete Energiefunktion entspricht weitestgehend der in dieser Arbeit verwendeten Energie von Hybridisierungen.

Die Strukturvorhersage wurde für die miRNA hsa-miR-1 und den oben verwendeten Abschnitt der mRNA KCNE1 durchgeführt. Die von `IntaRNA` ermittelte Interaktion enthält zwölf Basenpaare, für die Repräsentation dieser Interaktion durch die für die Kinetiken verwendete Energielandschaft ist eine minimale Fenstergröße von 13 erforderlich. Für diese Fenstergröße ist die Berechnung einer Makrostate-Kinetik möglich.

Abbildung 5.7 zeigt die Makrostate-Kinetik für dieses Sequenzpaar. Die Berechnung erfolgte mit Kawasaki Übergangsraten bei einer Fenstergröße von 16. Die Wahrscheinlichkeit des Basins der mfe-Struktur beträgt bei Erreichen des thermodynamischen Equilibriums nahezu 1. Tabelle 5.2 zeigt die von IntaRNA berechnete Interaktion sowie die Strukturen der lokalen Minima der im Plot der Kinetik dargestellten

Makrostates.

Die von `IntaRNA` berechnete Interaktion unterscheidet sich von der mfe-Hybridisierung lediglich durch ein Basenpaar. Die beiden Hybridisierungen können durch einen Shift ineinander überführt werden. Da beiden Hybridisierungen durch die in dieser Arbeit beschriebene Energiefunktion die gleiche Energie zugewiesen wird, ist die von `IntaRNA` berechnete Hybridisierung im Basin der mfe-Hybridisierung der Energielandschaft enthalten. Dass diese Hybridisierung nicht als lokales Minimum erkannt wird, kann also der willkürlichen Definition der Ordnungsrelation $<$ über die lexikographische Ordnung der Zustände zugeschrieben werden.

Damit wurde nachgewiesen, dass trotz der Beschränkung der Fenstergröße durch Berechnung von Makrostate-Kinetiken zur Strukturvorhersage äquivalente Aussagen getroffen werden können.

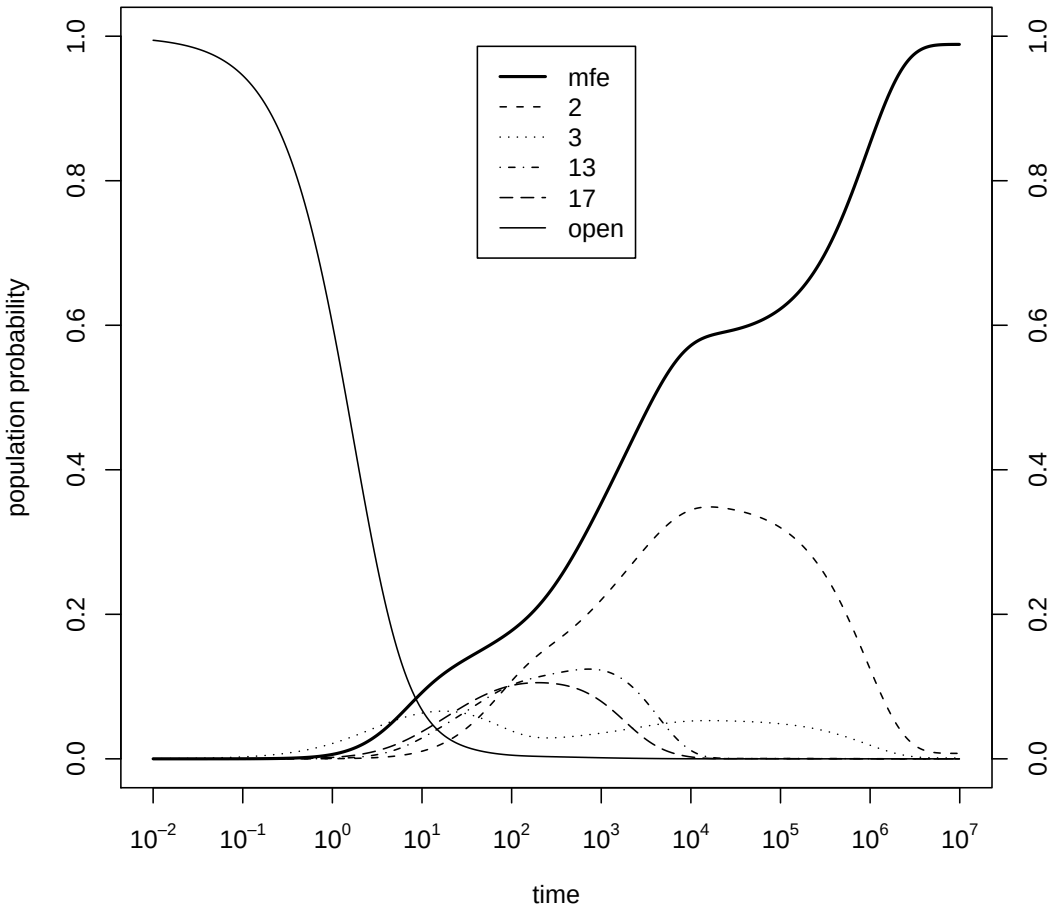


Abbildung 5.7: Makrostate-Kinetik von hsa-miR-1 und einem Abschnitt der mRNA KCNE1 der Länge 66. Die Wahrscheinlichkeit des Basins der mfe-Struktur nach Erreichen des thermodynamischen Equilibriums beträgt fast 1. Die mfe-Struktur entspricht der über IntaRNA gefundenen Hybridisierung. Für die Strukturen und Energien der zugehörigen Minima siehe Tabelle 5.2 auf Seite 64.

Sequence	UGGACACAUCCUGCCUGGCAACCUGAUUUUCUAAUCACAUUCCUCUCAUACUCUUUAUUGUGAUGG&UGGAAUGUAAAGAAGUAUGUAU	Energy
IntaRNA	.....(((((((((((.....&.....))))))).))))))...	-8.7
mfe-0	.....(((((((((.....&.....)))).)))))...	-5.1
mfe-1	.....(((((((.....&.))))))).....	-4.8
mfe-2	.....(((((((.....&.....)))))))..)).....	-3.8

Tabelle 5.1: Ausgewählte Hybridisierungen von hsa-miR-1 mit KCNE1 bei einer Fenstergröße von 10. Alle angegebenen Energiewerte wurden über die in dieser Arbeit beschriebene Methode berechnet. Die mit der validierten Hybridisierung übereinstimmenden Basenpaare wurden markiert. Für die Kinetik dieser Hybridisierungen siehe Abbildung 5.5 auf Seite 60.

Sequence	UGGACACAUCCUGCCUGGCAACCUGAUUUUCUAAUCACAUUCCUCUCAUACUCUUUAUUGUGAUGG&UGGAAUGUAAAGAAGUAUGUAU	Energy
IntaRNA	.....(((((((((((.....&.....))))))).))))))...	-8.7
mfe-0	.....(((((((((((.....&.....))))))).))))))...	-8.7
mfe-1	.....(((((....(((((((.....&.)))))))..))))).....	-5.8
mfe-2	.....(((((((.....&.))))))).....	-4.9
mfe-12	.....((((.((.....(((.....&.....))).....)).)))))..	-2.4
mfe-16	.....((((((.....&.))).)))....	-1.7

Tabelle 5.2: Ausgewählte Hybridisierungen von hsa-miR-1 mit KCNE1 bei einer Fenstergröße von 16. Alle angegebenen Energiewerte wurden über die in dieser Arbeit beschriebene Methode berechnet. Die mit der validierten Hybridisierung übereinstimmenden Basenpaare wurden markiert. Für die Kinetik dieser Hybridisierungen siehe Abbildung 5.7 auf Seite 63.

5.3.1 Auswirkung eines zu klein gewählten Hybridisierungsfensters

Da die maximale Fenstergröße, für die noch eine Berechnung der Makrostate-Kinetik möglich ist, zur Darstellung einer gegebenen Interaktion nicht ausreichend sein kann, soll nun der Einfluss eines zu klein gewählten Fensters untersucht werden. Für die Abschnitt 5.2.2 beschriebenen Kinetiken für hsa-miR-1 und KCNE1 wird eine Fenstergröße von 10 verwendet. Zur Darstellung der durch IntaRNA berechneten Interaktion ist jedoch ein Fenster von mindestens 13 erforderlich.

Tabelle 5.5 zeigt die von IntaRNA berechnete Interaktion sowie die Strukturen der lokalen Minima der in der rechten Kinetik von Abbildung 5.6 dargestellten Makrostates. Die mfe-Interaktion enthält neun Basenpaare. Von diesen Basenpaaren entsprechen fünf den Basenpaaren der über IntaRNA ermittelten Interaktion. Somit scheint eine zu gering gewählte Fenstergröße zumindest nicht zu vollkommen anderen Ergebnissen zu führen, als eine Kinetik mit ausreichend großem Fenster.

5.4 Kinetik einer experimentell nachgewiesenen Hybridisierung

Die regulatorische Wirkung der für die vorhergehenden Kinetiken verwendeten miRNA hsa-miR-1 auf die mRNA KCNE1 wurde experimentell nachgewiesen [44]. Für die Zielregionen auf KCNE1 sind jedoch lediglich die durch Strukturvorhersage ermittelten Interaktionen bekannt. Darum ist nicht klar, ob der für die Berechnung der Kinetiken verwendete Abschnitt der mRNA wirklich für eine erfolgreiche Regulation notwendig ist. In diesem Abschnitt soll die Berechnung von Kinetiken für eine experimentell nachgewiesenen Zielregion durchgeführt werden.

Eine solche Zielregion konnte für die miRNA let-7 und die mRNA lin-41 durch Monica C. Vella et. al nachgewiesen werden [37]. Der für eine Regulation durch let-7 nötige Abschnitt von lin-41 besteht aus zwei konservierten und zu let-7 komplementären Bereichen, LCS1 und LCS2. Diese sind durch einen Abschnitt aus 27 Nukleotiden voneinander getrennt. Der Komplex aus LCS1, dem trennenden Abschnitt und LCS2 wurde als pMV9 bezeichnet. Es konnte gezeigt werden, dass beide zu let-7 komplementäre Bereiche für eine erfolgreiche Regulation notwendig sind. Die Veränderung einiger Basen in dem trennenden Abschnitt unterband die regulatorische Wirkung von let-7. Dieser Komplex wurde als pMV19 bezeichnet. Durch eine Punktmutation des 5. Basenpaars in let-7 wurde die regulatorische Wirkung auf lin-41 abgeschaltet. Eine kompensierende Mutation auf let-41 führte zur Wiederherstellung der Regulation. Damit wurde pMV9 als der minimale Abschnitt von let-41 identifiziert, für den eine erfolgreiche Regulation durch let-7 möglich ist.

Abbildung 5.8 zeigt zwei Makrostate-Kinetiken von let-7 mit LCS1 und LCS2. In beiden Fällen wird das thermodynamische Equilibrium sehr rasch erreicht, die mfe-

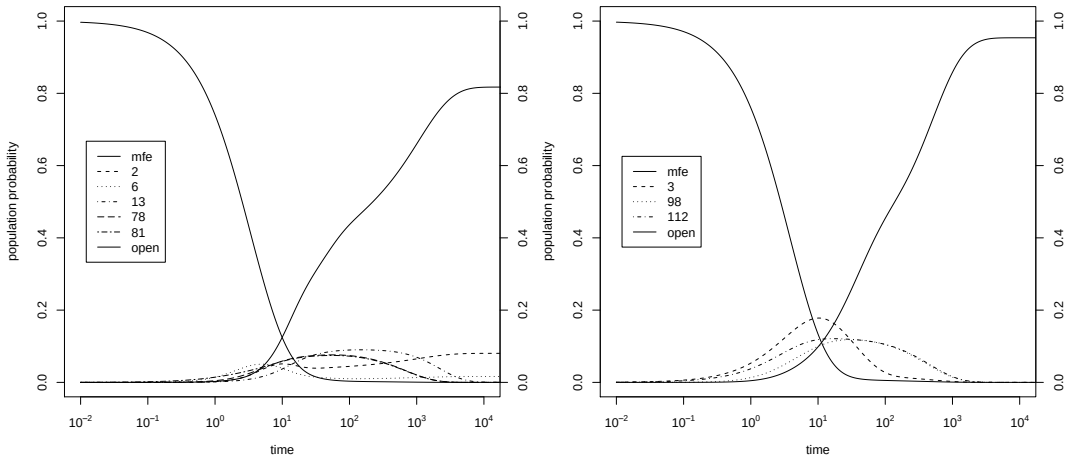


Abbildung 5.8: Makrostate-Kinetiken von let-7 mit LCS1 und LCS2 über Kawasaki Übergangsraten. (links) let-7 und LCS1 mit Fenstergröße 15. (rechts) let-7 und LCS2 mit Fenstergröße 17. Es wurden nur Makrostates abgebildet, die eine maximale Wahrscheinlichkeit von mindestens 5,5% erreichen.

Hybridisierungen sind die einzigen relevanten Strukturen.

Um zu untersuchen, ob die Sequenzen pMV9 und pMV19 Unterschiede im Faltungsverhalten zeigen, wurden für beide Sequenzen Kinetiken erstellt. Abbildung 5.9 zeigt zwei Makrostate-Kinetiken von let-7 mit pMV9 und pMV19. Leider konnte kein Unterschied zwischen dem Faltungsverhalten festgestellt werden, durch den der Ausfall der regulatorischen Wirkung von let-7 auf pMV19 erklärt werden könnte.

Dieses Ergebnis ist konform mit der Vermutung der Autoren der Studie, dass für die erfolgreiche Regulation von lin-41 ein zusätzlicher Faktor nötig sein könnte, der an den Abschnitt der trennende Sequenz von pMV9 bindet.

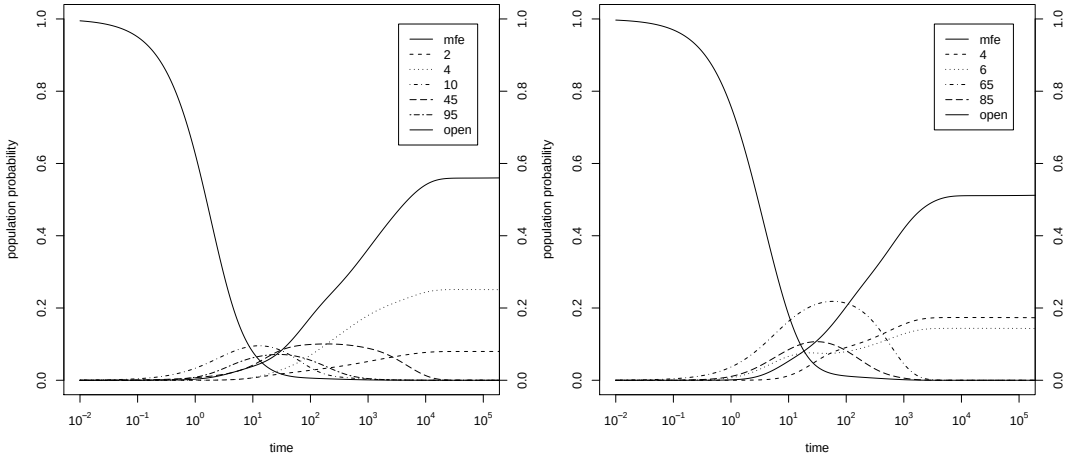


Abbildung 5.9: Makrostate-Kinetiken von let-7 mit pMV9 und pMV19 über Kawasaki Übergangsraten. (links) let-7 und pMV9 mit Fenstergröße 15. Es wurden nur Makrostates abgebildet, die eine maximale Wahrscheinlichkeit von mindestens 7% erreichen. (rechts) let-7 und pMV19 mit Fenstergröße 15. Es wurden nur Makrostates abgebildet, die eine maximale Wahrscheinlichkeit von mindestens 9% erreichen.

6 Zusammenfassung und Ausblick

Unter dem Begriff der Faltungskinetik wird die Analyse des Faltungsvorgangs von RNA zusammengefasst. Das Ergebnis einer solchen Analyse ist die Angabe der Wahrscheinlichkeitsverteilung der während der Faltung auftretenden Strukturen über den Zeitraum der Faltung. Für die Berechnungen von Kinetiken der Faltung von RNA Molekülen existieren bereits einige Ansätze. Diese sind jedoch auf die Beschreibung der Faltung einzelner RNA Moleküle zu Sekundärstrukturen beschränkt. Die Funktionalität vieler RNA Moleküle wird jedoch durch Strukturen bestehend aus mehreren RNA Strängen vermittelt, den Hybridisierungen. Ziel dieser Diplomarbeit war die Erweiterung eines vorhandenen Ansatzes zur Berechnung von RNA Faltungskinetiken auf diese Klasse von Interaktionen.

Zunächst wurde ein existierender Ansatz zur Beschreibung von Faltungsvorgängen vorgestellt, der auf der Modellierung des Faltungsvorgangs als stochastischer Prozess aufbaut. Für die Umsetzung dieses Modells werden drei Komponenten benötigt: die Menge von Strukturen die von einer gegebenen Basensequenz ausgebildet werden können, die Beschreibung der möglichen Übergänge zwischen diesen Strukturen sowie eine Methode zur Bestimmung der Wahrscheinlichkeit eines solchen Übergangs.

Daraufhin wurde gezeigt, wie diese Elemente nach Zusammenfassung zu einer Energielandschaft zur Berechnung von Faltungskinetiken genutzt werden können. Für diese Berechnung wurden zwei Ansätze vorgestellt, die stochastische Simulation des Faltungsvorgangs sowie die Berechnung der Wahrscheinlichkeitsmatrix eines auf der Energielandschaft aufsetzenden Markov-Prozesses.

Die Berechnung von stochastischen Simulationen ist sehr zeitaufwendig, kann aber für beliebig große Zustandsräume angewendet werden. Die Berechnung der Wahrscheinlichkeitsmatrix ist ebenfalls sehr aufwendig und nur für kleine Mengen von Zuständen durchführbar. Mit der Einführung von Makrostates über die Basins einer Energielandschaft wurde eine Methode vorgestellt, die Anzahl der zu betrachtenden Zustände zu verringern und gleichzeitig eine hohe Qualität der Kinetiken sicherzustellen. Durch diese Methode lassen sich auch Kinetiken von größeren Energielandschaften berechnen. Der Zeitaufwand für die Berechnung von Kinetiken nach dieser Methode ist viel geringer als die Berechnung der stochastischen Simulationen.

Die vorgestellten Algorithmen operieren auf der Abstraktion der Energielandschaft und sind damit weitestgehend unabhängig von der eigentlichen Repräsentation des zu betrachtenden Vorgangs. Aufgrund dieser Eigenschaft konnte die Erweiterung des

Ansatzes auf Kinetiken von Hybridisierungen durch die Definition einer Energielandschaft für diese Interaktionen umgesetzt werden.

Die Energielandschaft für Hybridisierungen sowie die zur Berechnung der Kinetiken notwendigen Algorithmen wurden in die Energy Landscape Library integriert. Damit ist nun erstmals die Bestimmung von Kinetiken für Hybridisierungen von RNA Molekülen möglich. Aufgrund der Modularität der ELL sowie der implementierten Algorithmen ist zudem eine Bestimmung von Kinetiken für alle in dieser Bibliothek enthaltenen Energielandschaften möglich. Die Identifikation der Makrostates sowie die Berechnung der für den Markov-Prozess notwendigen Übergangsratenmatrix erfolgt durch den Landscape Flooding Algorithmus innerhalb der ELL. Für die anschließende Berechnung der Übergangswahrscheinlichkeiten wird das Programm `treekin` verwendet.

Für die Berechnung der Kinetiken konnten zwei Einschränkungen identifiziert werden. Die erste Einschränkung besteht in der Art der repräsentierten Hybridisierungen. Nach der in dieser Arbeit verwendeten Definition einer Hybridisierung werden ausschließlich Strukturen mit zusammenhängendem Interaktionsbereich betrachtet, eine Ausbildung von Sekundärstrukturen innerhalb der einzelnen RNAs in diesem Bereich wird nicht erlaubt. Diese Einschränkung könnte durch eine Erweiterung auf mehrere Interaktionsbereiche, zwischen denen die Ausbildung von Sekundärstrukturen innerhalb der einzelnen RNAs erlaubt ist, aufgehoben werden. Um den energetischen Einfluss möglicher Sekundärstrukturen in diesen Zwischenbereichen zu modellieren, müsste zusätzlich die Berechnung der Energie für die Zugänglichkeit dieser Abschnitte erweitert werden.

Die zweite Einschränkung besteht in der Limitierung der Fenstergröße, für welche die Berechnung des Makrostate-Prozesses noch möglich ist. Diese Einschränkung ist auf zwei Faktoren zurückzuführen: Den Speicherplatzbedarf für das Fluten der Energielandschaft sowie für die Berechnung der Übergangswahrscheinlichkeiten durch `treekin`.

Für die Reduktion des Speicherbedarfs bei der Identifikation der Makrostates und der damit verbundenen Berechnung der Übergangsraten bieten sich sogenannte Sampling Ansätze an. Bei diesen wird auf eine vollständige Betrachtung des Zustandsraums verzichtet, stattdessen werden die Übergangsraten zwischen den Makrostates durch gezielte Stichproben abgeschätzt. Ein hybrider Ansatz, der die teilweise Flutung der Energielandschaft mit einem anschließenden Sampling kombiniert, wurde in der Diplomarbeit von Hannes Kochniß [19] beschrieben.

Der Speicherplatzbedarf für die Berechnung der Übergangswahrscheinlichkeiten über `treekin` hängt ausschließlich von der Anzahl der betrachteten Zustände ab. Durch eine weitere Verringerung der Anzahl von Makrostates wäre folglich die Betrachtung größerer Energielandschaften möglich. Der Faltungsprozess von RNA scheint

durch sogenannte Funnels bestimmt zu sein. Diese entsprechen in etwa den in dieser Arbeit verwendeten Basins. Eine Definition dieser Funnels über lokale Minima und Sattelpunkte einer Energielandschaft erfolgte in [18]. Eine Verkleinerung des Zustandsraums könnte nun durch die Zusammenlegung derjenigen Funnels durchgeführt werden, die lediglich durch niedrige Energiebarrieren voneinander getrennt sind. Die Auswirkung auf die Genauigkeit einer auf diese Weise berechneten Kinetik ist jedoch unklar.

Mit den in dieser Arbeit vorgestellten Methoden ist erstmals eine Berechnung von Kinetiken für RNA-RNA Hybridisierungen möglich. Damit wurde die Grundlage für weitere Untersuchungen zum Faltungsverhalten von Hybridisierungen geschaffen.

A Verwendete Sequenzen

xbix

```
5’ CUGCGGCUUUGGCUCUAGCC 3’
```

Sequenzen für den Nachweis der Ergodizität des Move Set

Die folgenden Sequenzabschnitte entstammen aus einer Untersuchung der Auswirkung der Struktur von mRNA auf post-transkriptionelles Gen-Silencing [33].

mRNA:

```
5’ GGAACAAUUGCUUUUACAGAUGCACAUAUCGAGGUGAACAUCACGUACGCGGAAUACUUCGAAA
   UGUCCGUUCGGUUGGCAGAAGCUAUGAAACGAUAUGGGCUGAAUACAAAUCACAGAAUCGUCGU
   AUGCAGUGAAAACUCUCUUCAAUUCUUUAUGCCGGUCUAUAGUGUCACCUAAAU 3’
```

siRNA:

```
5’ AUUUGUAUUCAGCCCAUAU 3’
```

hsa-miR-1

Für die menschliche miRNA hsa-miR-1 sind in der miRecords Datenbank 107 experimentell bestätigte Targets erfasst.

```
5’ UGGAAUGUAAAGAAGUAUGUAU 3’
```

KCNE1

Die menschliche mRNA KCNE1 kodiert für einen spannungsgesteuerten Kaliumkanal und ist ein experimentell bestätigtes Target für die miRNA hsa-miR-1 [44]. Es wurden drei Zielregionen für Interaktionen mit hsa-miR-1 vorhergesagt.

Für die Aufzählung der Strukturen verwendete Sequenzabschnitte

target 1:

5' AUUUAGCAGAAUCCCUGAGGACAUGGCCUCUGAGAAUAGCAGCUGCAUUUCCCAGACUCCCUUG
CAGCUAGCAAGGUUGUGUGACUAAGCCCUGGCCAGUAGGCAUGGAAGUGAAGACUGUAAUGUCC
AAGUAAUCCUUGGAAAGAAAAGAACGUGCCCUUAACUAACUUUGUCCUGCUUC 3'

target 2:

5' CCUGAGUUACCACAGUCCUUGAGAUGAGUGGUUCUUUGGGUUACAAAGUCCUCUGAAAGUCUAG
UGAGAGCUGUGAUCUUUGCCCCACCCGAAUAAUGCAUAUGGACACCACACCUUGCCUGCCGUGU
CCAGGAUUCAUGACCAGUAGCAGCCCAGCUAUGCCUGCCACGUCUCAUGGCCC 3'

target 3:

5' CAUUUUAAAGGGGGAGAGGGAAAAGUAACCGGGAGACAAAUUGAGCCACAUAUUUUCAGACACU
UGUUACCAUAUUUUAAAAUCUGGCUUCACAUACACAGAGUCUUUGCUAUGCACCAUGUACUGUU
CUAAGCUUCUUAAAAAUAGAAUCUCAAUUAUUAUUUUGCAGGCAAUACUCUAU 3'

Für die Berechnung der Kinetiken verwendeter Sequenzabschnitt

5' UGGACACAUCCUGCCUGGCAACCUGAUUUUCUAAUCACAUUCCUCUCAUACUCUUUAUUGUGAU
GG 3'

let-7

5' GUGAGGUAGUAGGUUGUAUAGUACUA 3'

LCS1

5' GUUAUACAACCGUUCUACACUCAUAUGA 3'

LCS2

5' GUUAAUACAACCAUUCUGCCUCCGGA 3'

pMV9

5' GUUAUACAACCGUUCUACACUCAUAUGAACGCGAUGUAAAUAUCGCAAUCCCUUUGUUAAUACA
ACCAUUCUGCCUCCGGA 3'

pMV19

5' GUUAUACAACCGUUCUACACUCAUAUGAAAGUGAUGUAAAUAUAGGAAUGUAUUUGUUAAUACA
ACCAUUCUGCCUCCGGA 3'

Literaturverzeichnis

[1] Anke Busch. *RNA secondary structure design under simple and complex constraints*. Dissertation, Albert-Ludwigs-Universität Freiburg, 2008.

[2] Anke Busch, Andreas S. Richter, and Rolf Backofen. IntaRNA: Efficient prediction of bacterial sRNA targets incorporating target site accessibility and seed regions. *Bioinformatics*, pages 2849–2856, 2008.

[3] Jennifer Couzin. Breakthrough of the year. Small RNAs make big splash. *Science*, 298(5602):2296–7, 2002.

[4] F. H. Crick. The origin of the genetic code. *J. Mol. Biol.*, 38:367–379, Dec 1968.

[5] A. R. Dinner, A. Sali, L. J. Smith, C. M. Dobson, and M. Karplus. Understanding protein folding via free-energy surfaces from theory and experiment. 25(7):331–9, 2000.

[6] C. Flamm, W. Fontana, I. L. Hofacker, and P. Schuster. RNA folding at elementary step resolution. *RNA*, 6(3):325–38, 2000.

[7] Christoph Flamm. *Kinetic Folding of RNA*. Doctor rerum naturalium, University of Vienna, 1998.

[8] Christoph Flamm and Ivo Hofacker. Beyond energy minimization: approaches to the kinetic folding of RNA. *Chemical Monthly*, 139:447–457, 2008.

[9] Christoph Flamm, Ivo L. Hofacker, Peter F. Stadler, and Michael T. Wolfinger. Barrier trees of degenerate landscapes. *Z.Phys.Chem*, 216:155–173, 2002.

[10] S. M. Freier, R. Kierzek, J. A. Jaeger, N. Sugimoto, M. H. Caruthers, T. Neilson, and D. H. Turner. Improved free-energy parameters for predictions of RNA duplex stability. *Proc. Natl. Acad. Sci. U.S.A.*, 83:9373–9377, Dec 1986.

[11] Walter Gilbert. The rna world. *Nature*, 319:618, 1986.

[12] D. T. Gillespie. A General Method for Numerically Simulating the Stochastic Time Evolution of Coupled Chemical Reactions. *Journal of Computational Physics*, 22:403–434, December 1976.

[13] Daniel T. Gillespie. Exact stochastic simulation of coupled chemical reactions. *The Journal of Physical Chemistry*, 81(25):2340–2361, 1977.

[14] Ivo L. Hofacker, Walter Fontana, Peter F. Stadler, Sebastian Bonhoeffer, Manfred Tacker, and Peter Schuster. Fast folding and comparison of RNA secondary structures. *Monatshefte Chemie*, 125:167–188, 1994.

[15] E. T. Jaynes. Information theory and statistical mechanics. *The Physical Review*, 106(4):620–630, 1957.

[16] Kyozi Kawasaki. Diffusion constants near the critical point for time-dependent ising models. I. *Phys. Rev.*, 145(1):224–230, May 1966.

[17] S. H. Kim, G. Quigley, F. L. Suddath, A. McPherson, D. Sneden, J. J. Kim, J. Weinzierl, P. Blattmann, and A. Rich. The three-dimensional structure of yeast phenylalanine transfer RNA: shape of the molecule at 5.5-A resolution. *Proc. Natl. Acad. Sci. U.S.A.*, 69:3746–3750, Dec 1972.

[18] K. Klemm, C. Flamm, and P. F. Stadler. Funnels in energy landscapes. *European Physical Journal B*, 63:387–391, June 2008.

[19] Hannes Kochniß. Ein Hybridkinetik Ansatz für RNA Faltungswahrscheinlichkeiten. Diplomarbeit, Friedrich-Schiller-Universität Jena, August 2008.

[20] K. Kruger, P. J. Grabowski, A. J. Zaug, J. Sands, D. E. Gottschling, and T. R. Cech. Self-splicing RNA: autoexcision and autocyclization of the ribosomal RNA intervening sequence of Tetrahymena. *Cell*, 31(1):147–57, 1982.

[21] Kathy Q. Luo and Donald C. Chang. The gene-silencing efficiency of siRNA is strongly dependent on the local structure of mRNA at the targeted region. *Biochem Biophys Res Commun*, 318(1):303–10, 2004.

[22] Martin Mann, Sebastian Will, and Rolf Backofen. The energy landscape library - a platform for generic algorithms. In *BIRD'07 - 1st international Conference on Bioinformatics Research and Development*, volume 217, pages 83–86. Oesterreichische Computer Gesellschaft, 2007.

[23] D. H. Mathews, M. E. Burkard, S. M. Freier, J. R. Wyatt, and D. H. Turner. Predicting oligonucleotide affinity to nucleic acid targets. *RNA*, 5(11):1458–69, 1999.

[24] J. Heinrich Matthaei, Oliver W. Jones, Robert G. Martin, and Marshall W. Nirenberg. Characteristics and composition of RNA coding units. *Proceedings of the National Academy of Sciences of the United States of America*, 48(4):666–677, 1962.

[25] J. S. McCaskill. The equilibrium partition function and base pair binding probabilities for RNA secondary structure. *Biopolymers*, 29(6-7):1105–19, 1990.

[26] Nicholas Metropolis, Arianna W. Rosenbluth, Marshall N. Rosenbluth, Augusta H. Teller, and Edward Teller. Equation of state calculations by fast computing machines. *The Journal of Chemical Physics*, 21(6):1087 – 1092, 1953.

[27] A. Mironov and A. Kister. RNA secondary structure formation during transcription. *J. Biomol. Struct. Dyn.*, 4:1–9, Aug 1986.

[28] Ulrike Mückstein, Hakim Tafer, Jorg Hackermuller, Stephan H. Bernhart, Peter F. Stadler, and Ivo L. Hofacker. Thermodynamics of RNA-RNA binding. *Bioinformatics*, 22(10):1177–82, 2006.

[29] Jord H. A. Nagel and Cornelis W. A. Pleij. Self-induced structural switches in rna. *Biochimie*, 84(9):913 – 923, 2002.

[30] Marc Rehmsmeier, Peter Steffen, Matthias Hochsmann, and Robert Giegerich. Fast and effective prediction of microRNA/target duplexes. *RNA*, 10(10):1507–17, 2004.

[31] Dirk Repsilber, Sabine Wiese, Marc Rachen, Astrid W. Schröder, Detlev Riesner, and Gerhard Steger. Formation of metastable rna structures by sequential folding during transcription: Time-resolved structural analysis of potato spindle tuber viroid ([minus])-stranded rna by temperature-gradient gel electrophoresis. *RNA*, 5(04):574–584, 1999.

[32] Rhonda L. Feinbau Rosalind C. Lee and Victor Ambros†. The c. elegans heterochronic gene lin-4 encodes small rnas with antisense complementarity to lin-14. *Cell*, 75(5):843–845, 1993.

[33] Stephen I. Rudnick, Jyothishmathi Swaminathan, Marina Sumaroka, Stephen Liebhaber, and Alan M. Gewirtz. Effects of local mRNA structure on posttranscriptional gene silencing. 105(37):13787–92, 2008.

[34] Steven A. Kostas Samuel E. Driver Craig C. Mello Andrew Fire SiQun Xu, Mary K. Montgomery. Potent and specific genetic interference by double-stranded rna in caenorhabditis elegans. *Nature*, 391:806–811, 1998.

[35] D. H. Turner, N. Sugimoto, and S. M. Freier. RNA structure prediction. *Annu Rev Biophys Biophys Chem*, 17:167–192, 1988.

[36] O. C. Uhlenbeck, A. Pardi, and J. Feigon. RNA structure comes of age. *Cell*, 90:833–840, Sep 1997.

[37] M. C. Vella, E. Y. Choi, S. Y. Lin, K. Reinert, and F. J. Slack. The C. elegans microRNA let-7 binds to imperfect let-7 complementary sites from the lin-41 3'UTR. *Genes Dev.*, 18:132–137, Jan 2004.

[38] A. E. Walter, D. H. Turner, J. Kim, M. H. Lyttle, P. Muller, D. H. Mathews, and M. Zuker. Coaxial stacking of helixes enhances binding of oligoribonucleotides and improves predictions of RNA folding. 91(20):9218–22, 1994.

[39] Michael S. Waterman. Secondary structure of single-stranded nucleic acids. *Studies in Foundations and Combinatorics, Advances in Mathematics Supplementary Studies*, 1:167–212, 1978.

[40] Carl Woese. *The Genetic Code*. New York: Harper and Row, 1967.

[41] Michael T. Wolfinger, W. Andreas Svrcek-Seiler, Christoph Flamm, Ivo L. Hofacker, and Peter F. Stadler. Efficient computation of RNA folding dynamics. *Journal of Physics A: Mathematical and General*, 37(17):4731–4741, 2004.

[42] Michael Thomas Wolfinger. The energy landscape of RNA folding. Diplomarbeit, Universität Wien, März 2001.

[43] S. Wuchty, W. Fontana, I. L. Hofacker, and P. Schuster. Complete suboptimal folding of RNA and the stability of secondary structures. *Biopolymers*, 49(2):145–65, 1999.

[44] Huixian Lin Baoxin Li Yanjie Lu Baofeng Yang Zhiguo Wang Xiaobin Luo, Jiening Xiao. Transcriptional activation by stimulating protein 1 and post-transcriptional repression by muscle-specific micrornas of i_{Ks}-encoding genes and potential implications in regional heterogeneity of their expressions. *Journal of Cellular Physiology*, 212(2):358–367, 2007.

[45] AJ Zaug and TR Cech. The intervening sequence RNA of Tetrahymena is an enzyme. *Science*, 231(4737):470–475, 1986.

[46] M. Zuker and P. Stiegler. Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. *Nucleic Acids Res.*, 9:133–148, Jan 1981.

Abbildungsverzeichnis

Tabellenverzeichnis

Algorithmenverzeichnis

Danksagung

Danken möchte ich zunächst Professor Rolf Backofen für die Vergabe dieses interessanten Diplomarbeitsthemas sowie allen Mitarbeitern des Lehrstuhls für Bioinformatik für das angenehme Arbeitsklima und die wertvollen Anmerkungen und Ratschläge zu dieser Arbeit.

Meinen Betreuern Martin Mann und Dr. Anke Busch möchte ich für die fachkundige wissenschaftliche Beratung danken. Einen besonderen Dank möchte ich Martin für seine unentbehrlichen Anmerkungen zu dieser Arbeit aussprechen.

Ohne die Führung von Dr. Sebastian Will durch die Gebirge der Energielandschaften im Rahmen des Tutorats seiner Vorlesung wäre dieser Weg viel beschwerlicher gewesen. Den Hinweis auf die validierte Interaktionsregion verdanke ich Andreas Richter. Marianne Dratwinski unterstützte mich mit ihren Hinweisen zur literarischen Qualität dieser Arbeit.

Ohne die fantastische Unterstützung durch meine Familie wäre mein Studium in Freiburg nicht möglich gewesen. Mein größter Dank gilt jedoch Kim-Jennifer, die mich seit vielen Jahren durch alle Höhen und Tiefen des Lebens begleitet.

Selbstständigkeitserklärung

Hiermit erkläre ich, dass ich diese Abschlussarbeit selbständig verfasst habe, keine anderen als die angegebenen Quellen/Hilfsmittel verwendet habe und alle Stellen, die wörtlich oder sinngemäß aus veröffentlichten Schriften entnommen wurden, als solche kenntlich gemacht habe. Darüber hinaus erkläre ich, dass diese Abschlussarbeit nicht, auch nicht auszugsweise, bereits für eine andere Prüfung angefertigt wurde.

Freiburg im Breisgau, 20. April 2009