Bachelor Thesis

p-Wert Statistiken von IntaRNA Vorhersagen

Fabio Gutmann

Gutachter: Prof. Dr. Rolf Backofen
Betreuer: Dr. Martin Raden

Albert-Ludwigs-Universität Freiburg

Technische Fakultät

Institut für Informatik

Lehrstuhl für Bioinformatik

13. August 2019

Bearbeitungszeit

13. 05. 2019 – 13. 08. 2019

Gutachter

Prof. Dr. Rolf Backofen

Betreuer

Dr. Martin Raden

Erklärung

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.

Ort, Datum	Unterschrift

Zusammenfassung

IntaRNA identifiziert energieminimale RNA-RNA Interaktionen und gibt somit Auskunft über mögliche Interaktionsstellen. Durch diese lassen sich z.B. Vorhersagen für mRNA targets für gegebene ncRNAs treffen. IntaRNA gibt allerdings nur vereinfachte Informationen über die Aussagekraft dieser Vorhersagen. Um die Signifikanz dieser Vorhersagen abschätzen zu können, wird in dieser Arbeit eine Methode zur Bestimmung des p-Wertes für IntaRNA Energien ausgearbeitet. Um diesen zu approximieren, werden zufällige Sequenzen aus den Ursprünglichen unter Erhaltung der Mono- und Dinukleotidfrequenz generiert und deren minimalen Interaktionsenergien von IntaRNA bestimmt. Die permutierten Sequenzen besitzen somit ähnliche Eigenschaften wie die Ausgangssequenzen. Aus diesen lässt sich der p-Wert durch das Angleichen einer Wahrscheinlichkeitsdichtefunktion und geschickter Integration bestimmen. Im Folgenden wird der Einfluss verschiedener Parameter auf den p-Wert einer Interaktion untersucht. Ebenfalls wurde versucht echte und unechte targets durch den p-Wert zu unterscheiden. Allerdings legen die Ergebnisse dieser Arbeit nahe, eine solche Unterscheidung anhand des p-Werts nicht möglich ist. Es besteht lediglich ein starker Zusammenhang zwischen p-Wert und MFE einer Interaktion.

Inhaltsverzeichnis

1. Einführung

RNA-RNA Interaktionen sind für regulierende Prozesse in allen Organismen von entscheidender Bedeutung [1]. Sie sind für grundlegende Prozesse in Zellen wie z.B. der Translation, also der Synthese von Proteinen, verantwortlich [2]. Die Vorhersage dieser Interaktionen stellt ein elementares Problem der aktuellen molekularen biologischen und biomedizinischen Forschung dar [3]. Kleine nichtcodierende RNAs (sRNAs) üben in Eu- und Prokaryoten die posttranskriptionale Regulation von Boten-RNA (mRNA) vor der Translation in Proteine aus [4]. Bioinformatische Analysen zur *in-silico* Vorhersage solcher Interaktionen nehmen somit immer weiter an Bedeutung zu. IntaRNA ist ein Programm zur schnellen und akkuraten Vorhersage von RNA Interaktionen. Hierfür wird nicht nur die Interaktionsenergie, sondern auch die *accessibility* der interagierenden Sequenzen berücksichtigt [1].

1.1. Biologischer Hintergrund

Ribonukleinsäure (RNA) ist ein Polynukleotid, also eine Kette aus kovalent verbundenen Nukleotiden. Jedes Nukleotid enthält hierbei entweder eine der beiden Purin-Basen Adenin und Guanin oder eine der beiden Pyrimidin-Basen Cytosin und Uracil, einen Zucker- und einen Phosphatteil. RNA enthält im Gegensatz zur DNA die Base Uracil anstatt Thymin. Die Primärstruktur von RNA-Ketten ist somit die Abfolge dieser Nukleotide. Die Sequenz lässt sich somit beschreiben als $S = s_1 s_2 \dots s_n$ mit $s_i \in \{A, C, G, U\}$, wobei n die Länge der Nukleotidkette darstellt.

Man bezeichnet die Watson-Crick Basenpaare A-U und G-C als komplementär, da sich zwischen ihnen Wasserstoffbrücken bilden, welche die Sekundärstruktur und somit auch die Funktion von RNA Molekülen bestimmen [5]. Außerdem kann in seltenen Fällen das Wobble-Basenpaar G-U entstehen [6]. Verbundene Basen werden als Basenpaar bezeichnet und der Prozess, bei dem sich ein Strang an einen anderen bindet nennt man Hybridisierung. Das lange Polynukleotid kann sich somit falten und einen Doppelstrang bilden, auch intramolekulare Sekundärstruktur genannt, welche die dreidimensionale Struktur definiert [7].

Auch die intermolekulare Sekundärstruktur entsteht durch komplementäre Basen, wobei diese sich hierbei in zwei verschiedenen Molekülen befinden. Eine sekundäre Struktur für S wird definiert als Menge P von geordneten Basenpaaren mit:

$$P \subseteq \{\, (i,j) \mid 1 \leq i < j \leq n,\ s_i \text{ und } s_j \text{ komplementär} \,\} \tag{1}$$

Außerdem müssen beliebige verschiedene Basenpaare stets disjunkt sein, dass heißt wenn $(i,j) \neq (i',j')$, dann $\{i,j\} \cap \{i',j'\} = \emptyset$. Eine Nukleobase kann also nur mit einer anderen Nukleobase Wasserstoffbrücken bilden.

Abbildung 1.: Beispiel einer verschachtelten Sekundärstruktur

Eine Sekundärstruktur heißt verschachtelt, wenn $\forall\ (i,j),(i',j') \in P$ mit $(i,j) \neq (i',j')$ gilt: $i < j < i' < j'$, das heißt (i,j) ist vor (i',j'), oder $i < i' < j' < j$, das heißt (i,j) schließt (i',j') ein. In Abbildung 1 ist ein Beispiel einer verschachtelten Sekundärstruktur zu sehen.

Bei zweidimensionaler Betrachtung ist leicht zu erkennen, dass sich verschachtelte Basenpaare nicht schneiden, weshalb sie auch als nicht-kreuzend bezeichnet werden. Dies ist im dreidimensionalen Raum natürlich nicht der Fall. Der Rechenaufwand zur

Vorhersage kreuzender Strukturen ist deutlich erhöht [8], weshalb diese von IntaRNA und in dieser Arbeit nicht berücksichtigt werden [9].

Oft erfüllen RNA Moleküle durch ihre dreidimensionale Struktur (Tertiärstruktur) eine regulierende Funktion auf andere RNA Moleküle aus. Für diese RNA-RNA Interaktionen wird eine stabile Struktur der Moleküle vorausgesetzt [10]. Die Stabilität wird leicht durch komplementäre Stränge beeinflusst, welche untereinander Basenpaare bilden. Deutlich stärker jedoch beeinflussen benachbarte Basen durch Stapelung (stacking), auch Basenstapelkraft genannt, die Stabilität [11].

Sei ein Basenpaar gegeben als $(i, j) \in P$ und eine Base an Position m mit $i < m < j$. Dann wird m als zugänglich bezeichnet, falls es kein Basenpaar $(i', j') \in P$ gibt, sodass $i < i' \leq m \leq j' < j$. Die Menge der Basen, die von dem Basenpaar (i, j) zugänglich sind, wird Schleife (loop) genannt. Die Größe der Schleife ist gegeben durch die Anzahl eingeschlossener ungepaarter Basen. Analog wird ein Basenpaar (i', j') als zugänglich bezeichnet, wenn es direkt von (i, j) eingeschlossen ist. Weiter lassen sich diese Strukturelemente, wie in Abbildung 2 zu sehen, nach folgenden Regeln kategorisieren:

1. hairpin loop, *falls es kein zugängliches Basenpaar gibt*,
2. multi-loop, *falls es mehr als ein zugängliches Basenpaar gibt*,
3. stacked pair, *falls genau ein Basenpaar* (i', j') *zugänglich ist mit* $i - i' = 1$ *und* $j - j' = 1$,
4. bulge loop, *falls ein Basenpaar* (i', j') *zugänglich ist, sodass entweder* $i - i' > 1$ *oder* $j - j' > 1$,
5. internal loop, *falls ein Basenpaar* (i', j') *zugänglich ist, sodass sowohl* $i - i' > 1$ *als auch* $j - j' > 1$.

Über die freie Energie von RNA Strukturen lassen sich Rückschlüsse auf die Stabilität der Struktur ziehen. Ein positiver Wert beschreibt dabei den Energiewert, der beim Zerfall der Struktur (in Form von Wärme) frei wird. Ein negativer Wert hingegen ist jene Energie, die zum Auftrennen der Basenpaarung notwendig wäre. Nach dem

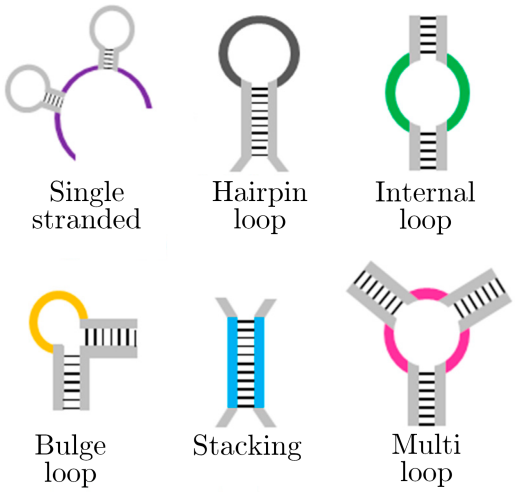

Abbildung 2.: Übersicht über die Sekundärstrukturelemente der RNA. Grafik adaptiert von Richard Sullivan [12, Scheme 2].

zweiten Grundsatz der Thermodynamik nimmt ein System bei konstanter Temperatur und konstantem Volumen immer den Zustand an, welcher die geringste freie Energie besitzt. Die Minimale Freie Energie (MFE) Struktur ist somit die Stabilste.

Zur Berechnung dieser Energie wird das Nearest Neighbor Modell verwendet [1, 13]. Dieses nutzt zur Berechnung der Energie eines Basenpaares die Energie des Paares selbst und die angrenzenden Nachbarn im selben Strang. Bei diesem wird die Struktur in kleinere Elemente zerlegt, welche zur Abschätzung der Gesamtenergie genutzt werden. Die Energie der gesamten Struktur wird als Differenz zur offenen Kette (also ohne Basenpaarungen) angegeben. Die Energie einer verschachtelten Sekundärstruktur P lässt sich somit als Summe der Einzelenergien wie folgt abschätzen:

$$E(S) = \sum_{(i,j)\in P} \begin{cases} e^H(i,j) & : \text{bei einem hairpin loop} \\ e^{SBI}(i,j,k,l) & : \text{bei einem stack, bulge oder internal loop} \\ e^M(i,j,x,x') & : \text{bei einem multi-loop,} \end{cases}$$

wobei (k,l) in e^{SBI} das eingeschlossene Basenpaar darstellt und x beim multi-loop

die Anzahl ungepaarter Basen und x' die Anzahl eingeschlossener Helices ist.

Es ist wichtig zu beachten, dass die intramolekulare Sekundärstruktur vor der Intermolekularen gebildet wird. Es kann also von Nöten sein, bereits gebildete Basenpaarungen wieder zu trennen, um neue Paarungen zwischen Molekülen zu bilden.

Gegeben seien zwei Sequenzen S und S' mit den intermolekularen Basenpaaren $(i, i'), (j, j') \in [1 \dots n] \times [1 \dots n']$, wobei $i \leq j$ und $i' \leq j'$ und (i, i') komplementär zu (j, j') sind. Die Energien zum Freilegen der Paarungen zwischen $i \dots j$ und $i' \dots j'$ (mit lila markiert in Abbildung 3) sind durch die Strafterme $\Delta_{i \dots j}$ und $\Delta_{i' \dots j'}$ gegeben, welche aus den Wahrscheinlichkeiten $Pr^{SS}(i \dots j)$ und $Pr^{SS}(i' \dots j')$ abgeleitet werden. Diese geben an, mit welcher Wahrscheinlichkeit die Basen an der jeweiligen Interaktionsstelle ungepaart (single stranded) sind. Die Formel hierzu beinhaltet die Gaskonstante R und die Temperatur T und berechnet sich wie folgt: $\Delta_{i \dots j} = -RT \cdot \ln\left(Pr^{SS}(i \dots j)\right)$ [7]. Im neu gebildeten Duplex, der durch die intermolekularen Basenpaare gebildet wird, ist die Energie $D^{i,j}_{i',j'}$ gespeichert. Wenn nun S und S' intermolekulare Basenpaare bilden (in Abbildung 3 mit blau markiert), dann ist die Interaktionsenergie gegeben als

$$I^{i,j}_{i',j'} = D^{i,j}_{i',j'} + \Delta_{i \dots j} + \Delta_{i' \dots j'}.$$

Nur Interaktionsenergien kleiner null sind von Interesse. Eine Interaktionsenergie von exakt null entspricht keiner Interaktion. Eine Struktur mit einer Energie größer null würde sofort zur offenen Kette zerfallen, da diese eine niedrigere Energie besitzt. Die Interaktion mit dem kleinsten Energiewert ist die Optimale.

In vielen RNA-RNA Interaktionen lassen sich seed Regionen beobachten [14]. Diese Regionen sind Bereiche, in denen die Stränge beider Moleküle nahezu perfekt komplementär sind und somit stackings bilden können, die deutlich zur Stabilität beitragen. IntaRNA lässt sich durch Parameter so einstellen, dass eine bestimmte Anzahl stackings in der seed Region nötig ist, damit die Interaktion berücksichtigt

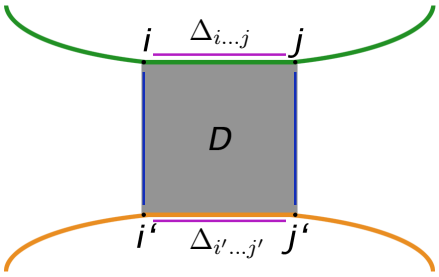

Abbildung 3.: Übersicht des von IntaRNA genutzten Energiemodells. Zu sehen ist die Duplexenergie $D^{i,j}_{i',j'}$, sowie die beiden Strafterme $\Delta_{i...j}$ und $\Delta_{i'...j'}$, die die Energie zum Freilegen der Interaktionsstelle darstellen.

wird. Gleichermaßen lässt sich ein Limit für die maximale Anzahl ungepaarter Basen in der seed Region einstellen.

IntaRNA gibt lediglich die Differenz der freien Energie im Vergleich zur offenen Kette einer Struktur als score aus. Um Aussagen über die Signifikanz dieses scores zu treffen, wird der p-Wert benutzt. Dieser wird aus den scores einer großen Stichprobenzahl zufälliger Sequenzen, die aus den ursprünglichen Sequenzen unter Bewahrung bestimmter Eigenschaften entstanden sind, approximiert.

1.2. p-Wert Statistik

Der p-Wert gibt anhand von Stichproben Aussage darüber, wie extrem ein Ergebnis ist [15]. Man nutzt dazu die Nullhypothese, welche dem Gegenteil der zu testenden Hypothese entspricht. Es stellt somit die Wahrscheinlichkeit dar, bei Wiederholung eines Experiments Werte zu erhalten, die gleich oder extremer als der Wert der Hypothese sind. Beim linksseitigen Test entspricht der p-Wert der Wahrscheinlichkeit einen Wert kleiner oder gleich dem zu testenden Wert zu erhalten. Der p-Wert des rechtsseitigen Test hingegen entspricht der Wahrscheinlichkeit, einen Wert größer oder gleich dem zu testenden Wert zu erhalten. Bei linksseitigem Test ist dies die Wahrscheinlichkeit

einen zufälligen Wert X kleiner m unter Annahme der Nullhypothese zu erhalten

$$P(X \leq m|H_0).$$

Es wird grundsätzlich zwischen empirischem und nicht-empirischem p-Wert unterschieden.

Sei Y eine Menge Stichproben und x eine zu untersuchende Beobachtung. Der empirische p-Wert eines linksseitigen Tests ergibt sich dann aus der Anzahl Stichproben, die kleiner oder gleich als die zu untersuchende Beobachtung sind, dividiert durch die Anzahl der Stichproben.

$$p_{empirisch}(x) = \frac{\#(Y \leq x)}{\#Y}$$

Der empirische p-Wert konvergiert für eine größer werdende Stichprobenzahl gegen den echten p-Wert.

Mit Hilfe einer großen Anzahl Stichproben lässt sich die eigentliche Wahrscheinlichkeitsdichtefunktion $f(b)$ approximieren [16]. Diese ist stets nichtnegativ, also $f(b) \geq 0 \ \forall \ b \in \mathbb{R}$ und ist immer normiert, also $\int_{-\infty}^{\infty} f(b)dx = 1$. Sie kann dazu benutzt werden, den eigentlichen, nicht-empirischen p-Wert zu ermitteln. Der p-Wert entspricht der Wahrscheinlichkeit, bei einer zufälligen Stichprobe X einen Wert kleiner oder gleich einer Beobachtung x unter Annahme der Nullhypothese zu erhalten.

$$P(X \leq x|H_0) = \int_{-\infty}^{x} f(b)db$$

Die Alternativhypothese zu einem guten score einer RNA-RNA Interaktion besagt, dass dieser evolutionär bedingt ist. Im Gegensatz dazu ist die Nullhypothese die Annahme, dass er rein zufällig entstanden ist und kein evolutionärer Zusammenhang besteht. Um den p-Wert einer von IntaRNA getroffenen Vorhersage zu ermitteln, werden die Sequenzen unter Erhaltung bestimmter Eigenschaften randomisiert. Der p-Wert gibt somit die Wahrscheinlichkeit an, dass eine stabilere Struktur als die

Beobachtete aus RNA-Ketten mit ähnlichen Eigenschaften wie die ursprüngliche Sequenz durch Zufall entsteht.

1.3. Randomisieren von RNA Ketten

Zur Angabe, wie ähnlich sich zwei Sequenzen sind, wird oft die *Evolutionäre Distanz* benutzt. Diese gibt an, wie viel mindestens an einer Sequenz geändert werden muss, um sie in eine Andere zu überführen [17]. Hierbei können Nukleotide gelöscht, eingefügt oder durch ein anderes ersetzt werden. Da gestapelte Basenpaare eine große Rolle bei der Stabilität von RNA Ketten spielen, ist es von Nöten nicht nur die Mono-, sondern auch die Dinukleotidfrequenz beim Randomisieren beizubehalten [18]. Um zufällige Sequenzen unter Erhaltung der Eigenschaften der ursprünglichen Sequenz zu generieren, wird in dieser Arbeit der Algorithmus von Altschul und Erickson verwendet. Es wurde hierzu eine Implementierung von P. Clote [19] für Python 2 [1] genutzt und für Python 3 modifiziert.

Abbildung 4.: Dinukleotide in einer RNA Kette

1.3.1. Algorithmus von Altschul und Erickson

Der Algorithmus von Stephen Altschul und Bruce Erickson baut auf Eulerkreisen auf [17]. Er behält die Häufigkeit von Mono- und Dinukleotiden bei und erzeugt alle permutierten Sequenzen mit gleicher Wahrscheinlichkeit. Die ausgegebenen Sequenzen sind von selber Länge als die originale Sequenz. Der Algorithmus kann theoretisch

[1] `http://clavius.bc.edu/~clotelab/RNAdinucleotideShuffle/ShuffleCodeParts/altschulEriksonDinuclShuffle.txt`

so erweitert werden, dass auch die Trinukleotidhäufigkeit beibehalten wird. Eine zufällige Permutation zu finden, die die Dinukleotidhäufigkeit beibehält, ist äquivalent damit, einen zufälligen Eulerkreis in einem gerichteten Multigraphen zu finden [20]. Ein Eulerweg ist eine Abfolge von Kanten eines Graphen, sodass jede Kante genau einmal besucht wird, Start- und Endknoten spielen hierbei keine Rolle.

Gegeben sei eine zu permutierende Sequenz $S = s_1 s_2 s_3 \dots s_n$. Gegeben sei ein leerer Graph G, genannt Dinukleotid-Graph. Für jedes Mononukleotid $s_i \in S$ wird nun einmalig ein Knoten in G erstellt und für jedes Vorkommen des Dinukleotids $s_i s_j \in S$ wird im Anschluss eine Kante von s_i nach s_j in G eingefügt. Für jeden Knoten $s \in G$ sei eine geordnete Liste aller ausgehenden Kanten, genannt Kantenliste gegeben. Die Menge aller Kantenlisten für G wird Kantenordnung $K(S)$ von G genannt. In Abbildung 5 ist der Dinukleotid-Graph und in Tabelle 1 die Kantenordnung zur Sequenz $S = GAUGGCAU$ zu sehen. Der entstandene Dinukleotid-Graph ist ein gerichteter Multigraph, der Schleifen enthalten kann. S definiert hierbei eindeutig die Kantenordnung $K(S)$, und umgekehrt definiert $K(S)$ eindeutig die Sequenz S. $K(S)$ definiert ebenso einen Eulerweg in G: Folge beginnend bei s_1 der ersten Kante aus der Kantenliste von s_1 zu s_2 und streiche diese Kante. Folge nun der ersten Kante aus der Kantenliste von s_2 und so weiter.

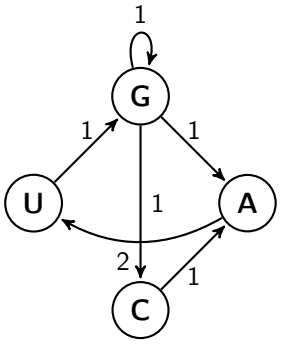

Abbildung 5.: Der Graph zu S

Kantenlisten

G	$\overset{1}{GA}$	$\overset{4}{GG}$	$\overset{5}{GC}$
A	$\overset{2}{AU}$	$\overset{7}{AU}$	
U	$\overset{3}{UG}$		
C	$\overset{6}{CA}$		

Tabelle 1.: Die Kantenordnung zu S

Eine permutierte Sequenz S' hat denselben Dinukleotid-Graphen G wie S, da die Häufigkeit der Dinukleotide nicht verändert wird. Die Kantenordnungen $K(S)$ und $K(S')$

unterscheiden sich nur durch die Reihenfolge der Dinukleotide in den Kantenlisten. S' hat ebenfalls die selben Start- und Endnukleotide s_1 und s_n wie S.

Um eine neue Sequenz S' aus S zu generieren, wird zuerst die Reihenfolge der Dinukleotide einzeln in jeder der Kantenlisten der Kantenordnung $K(S)$ zufällig verändert. Hierdurch entsteht ein neuer Pfad durch den Graphen, der an s_1 beginnt und bei einem Knoten endet, dessen Kantenliste leer ist. Dieser letzte Knoten s_n muss logischerweise bei S' derselbe wie bei S sei. Außerdem muss der Pfad ein Eulerweg sein. Um dies effizient zu überprüfen, kann das Eulersche Kantenordnungs-Theorem genutzt werden. Hierzu wird der Letzte-Kanten-Graph Z definiert, welcher ein Subgraph aus G darstellt, der nur die letzte Kante aus jeder der Kantenlisten (außer s_n) enthält. Das Theorem besagt nun, dass die Kantenordnung $K(S)$ eulersch ist, falls jeder der Knoten in Z eine Kante zum finalen Knoten s_n besitzt. Der Beweis für dieses Theorem kann in der Publikation von Altschul und Erickson gefunden werden [17].

Der Algorithmus um, eine Sequenz S in eine permutierte Sequenz S' zu überführen, lässt sich somit in fünf Schritte unterteilen:

1. Aus S wird ein Dinukleotid-Graph G mit Kantenordnung $K(S)$ erstellt
2. Für jeden Knoten $s \in G$ (außer s_n) eine Kante aus der Kantenliste von s zufällig auswählen und an das Ende der Kantenliste setzten
3. Den Letzte-Kanten-Graph bilden und überprüfen, ob neue Kantenordnung eulersch ist (also ob jeder Knoten in Z eine Kante zu s_n besitzt). Falls die neue Kantenordnung nicht eulersch ist, gehe zurück zu Schritt zwei.
4. Für jede der Kantenlisten in $E(S)$ die verbleibenden Kanten zufällig anordnen (außer der Letzten). Somit ist die neue Kantenordnung $K(S')$ gegeben.
5. Die permutierte Sequenz S' ergibt sich nun daraus, dass man an s_1 beginnend jeder Kante aus der zugehörigen Kantenliste folgt. Hierbei wird die jeweils erste Kante $s_i s_j$ aus der Kantenliste entfernt, s_i zur Sequenz S hinzugefügt und zur Kantenliste s_j gesprungen. Dies wird solange wiederholt, bis alle Kantenlisten leer sind.

1.3.2. Sonstige Methoden zur Permutation von Sequenzen

Es gibt einige weitere Methoden, mit dem zufällige Sequenzen unter Erhaltung der Eigenschaften permutiert werden können. Eine davon ist die Markov Methode, die auf Markov Ketten aufbaut. Sie erhält allerdings die ursprünglichen Eigenschaften einer Sequenz beim Randomisieren nicht immer, sondern nur im Durchschnitt, weshalb sie für diese Arbeit nicht geeignet ist. Eine andere Methode ist die Permutations-Methode. Die von ihr erzeugten Sequenzen erhalten die Eigenschaften der ursprünglichen Sequenz, wie der Algorithmus von Altschul und Erickson, immer. Jedoch erhält sie nur die Mono- und nicht die Dinukleotidhäufigkeit der Ausgangssequenz, weshalb auch sie keine Anwendung in dieser Arbeit findet.

1.3.3. Die shuffle-Modi

Für ein Sequenzpaar aus query und target gibt es drei mögliche shuffle-Modi. Entweder kann nur das query permutiert werden, nur das target oder jedoch beide. Wenn nur das target geshuffeld wird, also das query nicht verändert wird, ähnelt das Hintergrundmodell dem genomweiten Hintergrundmodell von CopraRNA. Der gewählte shuffle-Modus sollte keinen zu großen Einfluss auf den p-Wert haben. Dies wird in Abschnitt 2.2 genauer untersucht.

1.4. Dichteverteilungen

1.4.1. Die Normalverteilung

Die Normal- oder Gauß-Verteilung ist eine stetige, symmetrische Wahrscheinlichkeitsverteilung. Sie besitzt die Dichtefunktion $f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp(-\frac{(x-\mu)^2}{2\sigma^2})$. Die Kurve ist normiert durch $\int_{-\infty}^{\infty} f(x)dx = 1$. Der Lageparameter $\mu \in \mathbb{R}$ definiert gleichzeitig den Erwartungswert und das Maximum der Verteilung. Die Varianz und die Breite

der Kurve sind durch den Skalenparameter $\sigma^2 > 0$ gegeben, wobei σ die Standardabweichung der Verteilung ist. Die Anwendung dieser symmetrischen Kurve macht nur Sinn, falls eine gleichmäßige Streuung der Messwerte in beide x-Richtungen zu erwarten ist.

1.4.2. Die Allgemeine Extremwertverteilung

Die Allgemeine Extremwertverteilung ist eine Gruppe von Wahrscheinlichkeitsverteilungen, darunter die Gumbel-, Fréchet und Weibull Verteilungen. Die Wahrscheinlichkeitsdichtefunktion ist gegeben als $f(x) = \frac{1}{\sigma} t(x)^{\xi+1} e^{-t(x)}$ mit

$t(x) = \begin{cases} (1 + \xi(\frac{x-\mu}{\sigma})), \text{ falls } \xi \neq 0 \\ e^{-(x-\mu)/\sigma}, \text{ falls } \xi = 0 \end{cases}$. Der Lageparameter μ gibt, wie auch bei der Normalverteilung, den Ort der Kurve auf der x-Achse an. Unter σ versteht man erneut den Skalenparameter, der die Streuung bzw. Variabilität der Kurve definiert. Da die Kurve allerdings nicht symmetrisch ist, ist σ nicht die Standardabweichung und μ nicht der Erwartungswert. Im Gegensatz zur Normalverteilung besitzt die Allgemeine Extremwertverteilung einen zusätzlichen Parameter, den Formparameter $\xi \in \mathbb{R}$. Dieser definiert die Gestalt der Kurve, wie in Abbildung 6 zu sehen. Mit größer werdendem ξ verlagert sich die Schräge der Kurve nach rechts und mit kleiner werdendem ξ nach links. Hierbei nimmt die Steigung auf der jeweils anderen Seite der Kurve zu. Ebenfalls deutlich zu sehen ist, dass die Kurve mit größer werdendem σ breiter wird und sich auf der x-Achse mit μ verschieben lässt. Durch den zusätzlichen Parameter kann der tail, also die besonders extremen Werte, deutlich besser im Vergleich zur Normalverteilung approximiert werden.

1.4.3. Die Gumbel-Verteilung

Die Gumbel-Verteilung ist ein Spezialfall der Allgemeinen Extremwertverteilung, bei dem der Formparameter $\xi = 0$ ist. Die Dichteverteilung ergibt sich somit aus

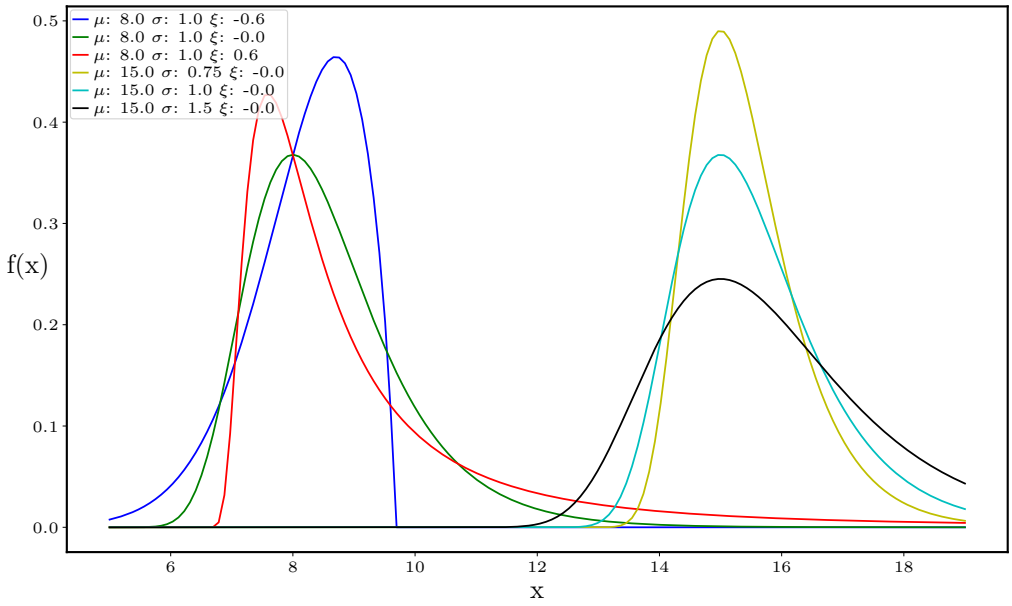

Abbildung 6.: Verschiedene Dichtefunktionen der Allgemeine Extremwertverteilung mit unterschiedlichen Parametern

der Allgemeinen Extremwertverteilung zu $f(x) = \frac{1}{\sigma} e^{-(\frac{x-\mu}{\sigma} + e^{-\frac{x-\mu}{\sigma}})}$. Sie ist ebenfalls asymmetrisch, jedoch durch den fehlenden zusätzlichen Parameter schlechter für das Anpassen des tails zu gebrauchen. Sie wird oft zur Vorhersage der Wahrscheinlichkeit von Naturkatastrophen verwendet.

2. Versuche und Ergebnisse

Um die Einflüsse verschiedenster Parameter auf den p-Wert eines Sequenzpaares zu untersuchen, wurden mehrere zielgerichtete Experimente durchgeführt. Jede in diesen Experimenten verwendete Sequenz ist in Anhang A aufgeführt.

2.1. Verteilung von IntaRNA scores

Um den nicht-empirischen p-Wert zu ermitteln, muss eine Wahrscheinlichkeitsdichtefunktion an die Verteilung der IntaRNA scores gefittet werden. In einem ersten Versuch wurde somit ein Histogramm der Verteilung der scores aus permutierten Sequenzen für verschiedene Beispiele erstellt. Im Anschluss wurde versucht, verschiedene Wahrscheinlichkeitsdichtefunktionen an das Histogramm anzupassen.

Da IntaRNA keine scores größer null ausgibt, ist in Abbildung 7 nicht die gesamte Verteilung zu sehen. Dies führt dazu, dass der Mittelwert deutlich weiter im negativen Bereich liegt als dies der Fall wäre, wenn diese scores nicht fehlen würden. Es ist deutlich zu erkennen, dass die Symmetrie der Normalverteilung dazu führt, dass der verschobene Mittelwert nicht gut ausgleichen werden kann. Die flexibleren Verteilungen der Gumbel und Allgemeinen Extremwertverteilung gleichen dieses Problem wesentlich besser aus. Für die Bestimmung des p-Werts ist in den meisten Fällen (besonders für kleine p-Werte) jedoch das fitting des tails von größerer Bedeutung, da von $-\infty$ integriert wird. Die Verschiebung des Mittelwerts führt in Abbildung 7 dazu, dass

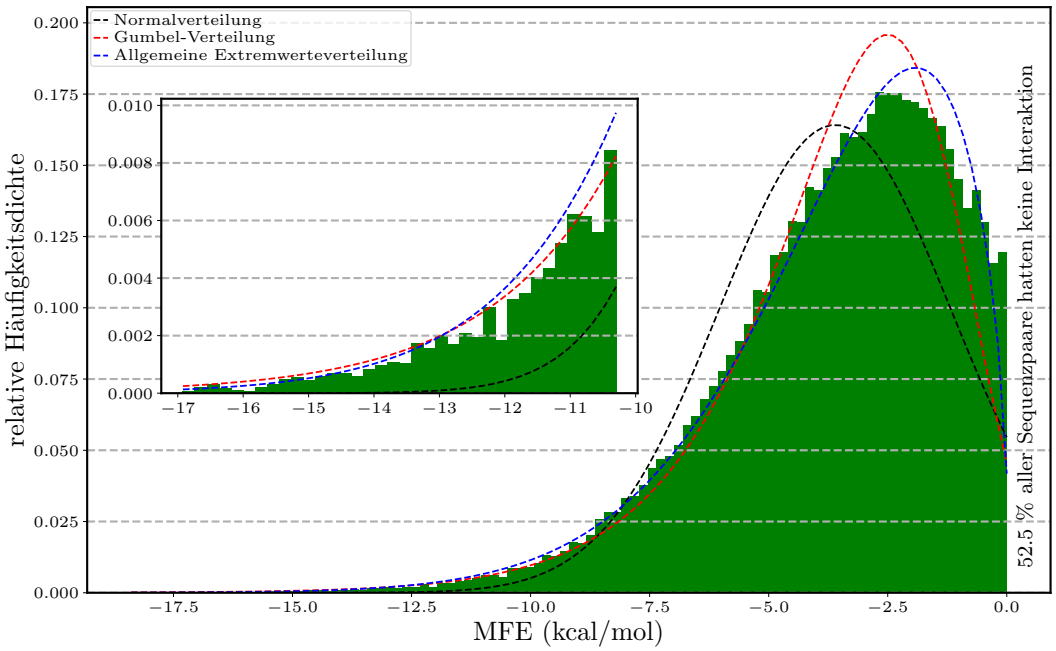

Abbildung 7.: Verteilung von 100000 IntaRNA scores aus den Sequenzen rq1 und rt1 mit gefitteten Dichteverteilungen. Es wurden sowohl query als auch target permutiert. Ebenfalls zu sehen ist eine vergrößerte Darstellung des tails.

die Normalverteilung für kleine scores weit unter dem tatsächlichen Verlauf liegt. Da dieses Problem häufig auftritt, wurde die Normalverteilung nicht weiter berücksichtigt. Da die Gumbel Verteilung nur ein Spezialfall der Allgemeinen Extremwertverteilung ohne den zusätzlichen Formparameter ist und dies in den meisten Fällen zu einer ungenaueren Approximation führt, wurde für die weitere Berechnung des p-Werts die Allgemeine Extremwertverteilung genutzt.

2.2. Einfluss der Anzahl permutierter Sequenzen und des shuffle-Modus

Da der p-Wert aus zufällig generierten Sequenzen berechnet wird, ist eine kleine Menge Stichproben oft nicht ausreichend, um eine akkurate Verteilung an die Messwerte anzupassen. Um den Einfluss der Anzahl Sequenzen, die zur Approximation des p-Werts genutzt werden, zu untersuchen, wurde eine große Anzahl IntaRNA scores generiert. Dies wurde für drei verschiedene query/target Paare mit den jeweils drei möglichen shuffle-Modi wiederholt. Aus all diesen Messwerten wurde die tatsächlich verwendete Anzahl permutierter Sequenzen schrittweise vergrößert. Hierbei wurden immer die jeweils ersten scores verwendet, um den jeweiligen p-Wert zu berechnet. Für eine größer werdende Zahl scores sollte der p-Wert einen Grenzwert besitzen, also gegen einen festen Wert konvergieren.

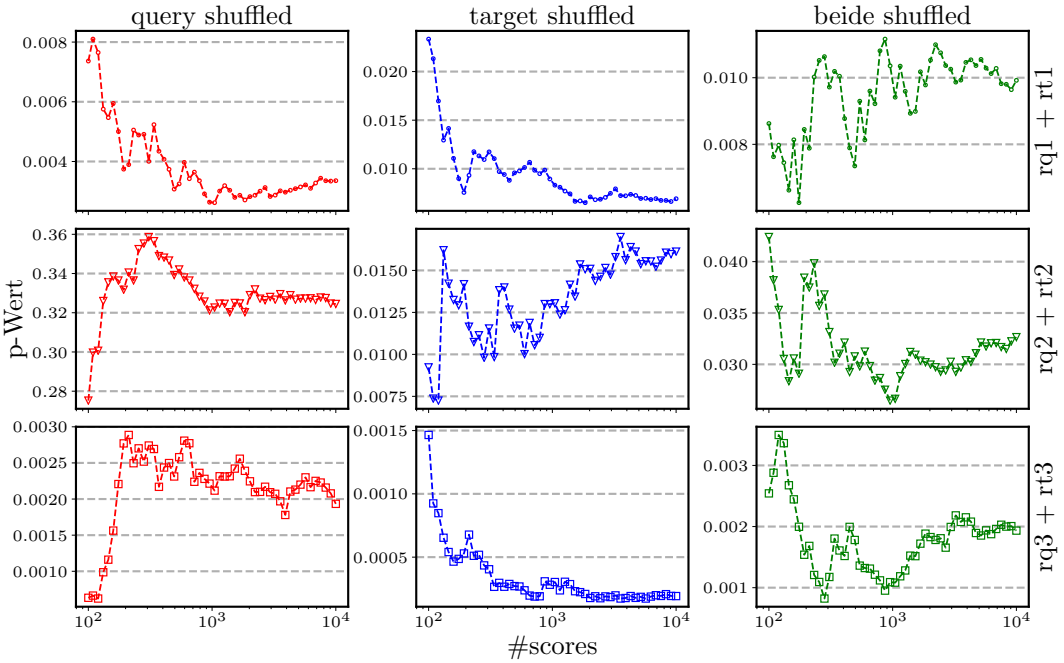

Abbildung 8.: p-Werte für steigende Anzahl genutzter scores anhand von drei Sequenzpaaren für alle drei shuffle-Modi

In Abbildung 8 ist für jedes der drei Sequenzpaare eine Konvergenz des p-Werts gegen

einen festen Wert zu erkennen. Ebenfalls ist eine deutliche Abnahme der Varianz der p-Werte mit steigender Anzahl scores zu beobachten.

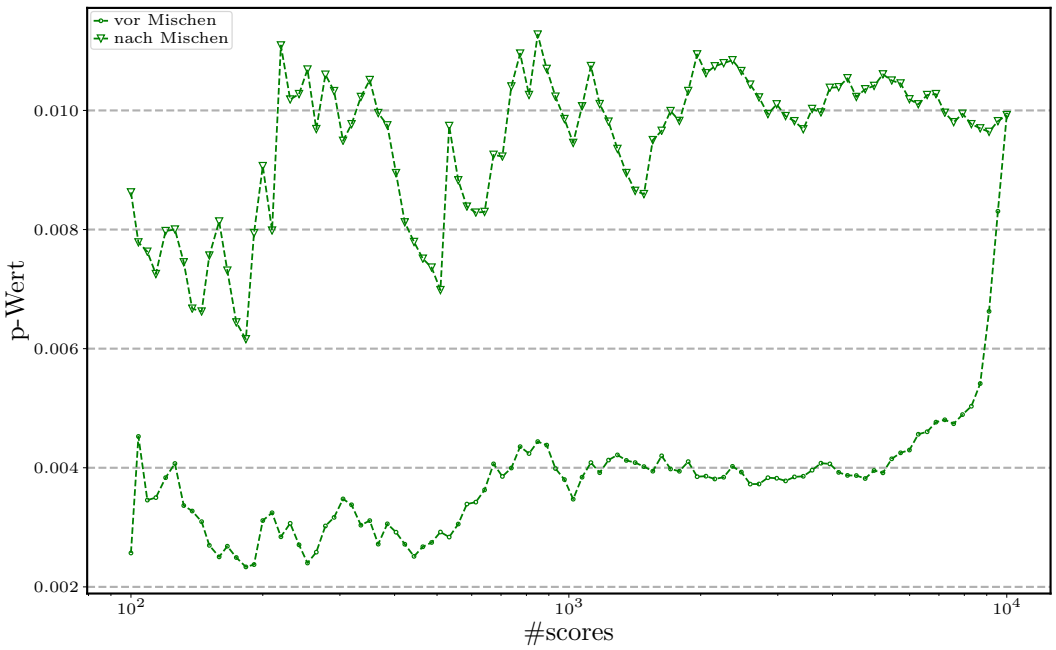

Abbildung 9.: Anomalie im Verlauf des p-Werts für steigende Anzahl IntaRNA scores der Sequenzen rq1 und rt1 vor und nach Mischen

Im Rahmen der Untersuchung fanden sich zum Teil Anomalien des Verlaufs, wie in Abbildung 9 zu sehen. Im Beispiel wurden sowohl query als auch target geshuffled. Ab etwa 8000 verwendeter scores steigt der p-Wert abrupt an. Um eine Erklärung für diese Beobachtung zu finden, wurden die Einzelwerte der Stichprobe untersucht. Hierbei fanden sich besonders viele kleine Werte am Ende der Messreihe. Daraufhin wurden die scores neu gemischt, dass heißt deren Anordnung neu verteilt. Die scores selbst blieben hierbei identisch, weshalb in beiden Verläufen der selbe Endpunkt zu finden ist. Durch das Durchmischen verschwand die Anomalie vollständig, wie ebenfalls in Abbildung 8 zu sehen.

Der Einfluss des shuffle-Modus wurde ähnlich wie die generelle Verteilung von IntaRNA scores untersucht, indem anhand verschiedener Sequenzpaare eine Verteilung für jeden der drei shuffle-Modi erstellt wurde. Dies wurde für die Sequenzpaare rq1 + rq1 bis

rq6 + rt6 wiederholt.

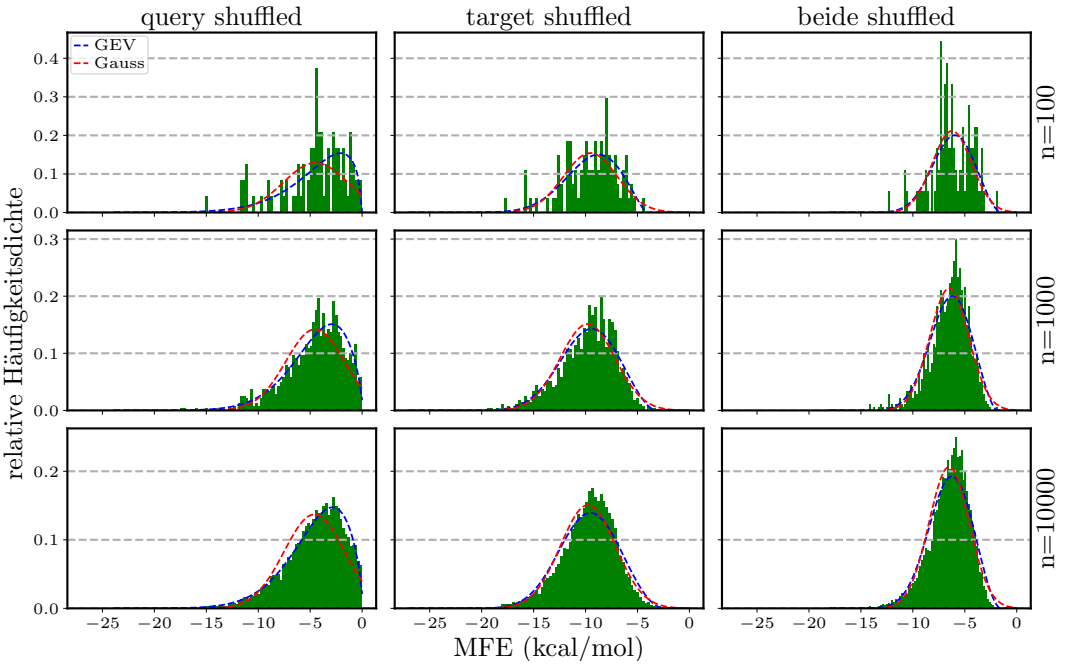

Abbildung 10.: Verteilungen von IntaRNA scores der Sequenzen rq2 und rt2 für die drei shuffle-Modi mit steigender Anzahl permutierter Sequenzen. Auf den Verteilungen wurden jeweils eine Allgemeine Extremwertverteilung und Normalverteilung gefittet.

In Abbildung 10 ist die Verteilung von rq2 + rt2 für die drei shuffle-Modi mit jeweils steigender Anzahl verwendeter Permutationen zu sehen. Es ist deutlich zu erkennen, dass die Verteilung definierter wird, je mehr permutierte Sequenzen verwendet werden. Allerdings sind keine klaren Trends aufgrund des shuffle-Modus erkenntlich. Diese befinden sich zumeist in der selben Größenordnung.

2.3. Korrelation zwischen p-Wert und MFE

Um den Zusammenhang zwischen p-Wert und MFE genauer zu untersuchen, wurden für verschiedene sRNA die besten targets, also die targets mit der niedrigsten MFE, aus der Studie von Patrick R. Wright [21] ausgewählt. Diese können zusammen mit den echten targets im Anhang A eingesehen werden. Alle targets haben eine Länge von 300 nt, die sRNAs sind zwischen 84 nt und 122 nt lang. Für jedes dieser targets wurde der p-Wert anhand von 2500 zufällig generierten, permutierten Sequenzen für jeden der drei shuffle-Modi ausgerechnet. Die p-Werte wurden dann für die jeweiligen MFE-Werte aufgetragen und Spearman's Rangkorrelationskoeffizient ermittelt.

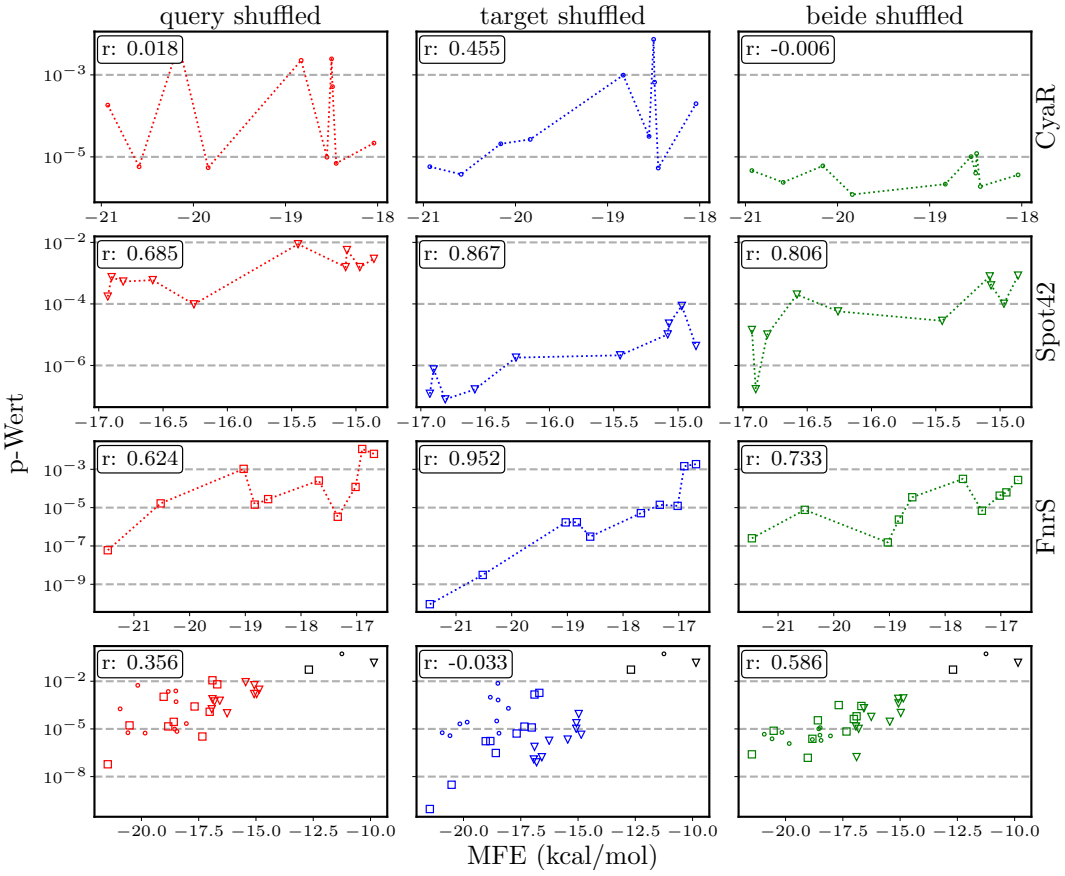

Abbildung 11.: p-Werte von drei echten sRNA queries mit den jeweils am besten interagierenden targets und einem echten target (schwarz) mit ihren MFE Werten

Die Abbildung 11 zeigt die p-Werte von drei sRNAs mit ihren zugehörigen targets mit niedrigsten MFE scores. Die sRNAs Spot42 mit 109 nt Länge und FnrS mit 122 nt Länge zeigen eine starke Korrelation zwischen p-Wert und MFE. Der p-Wert der kürzeren sRNA CyaR mit 84 nt Länge korreliert im Gegensatz dazu nur schwach mit dem MFE-Wert der targets, außer wenn nur das target geshufflet wird. Dennoch ist anhand des kumulativen Plots eine deutliche sequenzübergreifende Korrelation zwischen p-Wert und MFE zu erkennen. Ebenfalls ersichtlich ist, dass die p-Werte der targets zu jeder sRNA gruppiert beieinander liegen. In diesen sind in schwarz die p-Werte der echten targets jeder sRNA markiert. Diese liegen allerdings um mehrere Größenordnungen höher als die der besten targets.

2.4. Einfluss der Längendifferenz zwischen query und target

Um die Auswirkungen der Längendifferenz zwischen query und target zu analysieren, wurden drei Sequenzpaare mit Länge 300 nt zufällig generiert. Das target wurde hierbei jeweils auf der Gesamtlänge festgesetzt, während das query von anfangs 10 nt in 10 nt Schritten bis zur vollen Länge vergrößert wurde. Bei jeder Iteration wurden für die drei shuffle-Modi 2500 IntaRNA scores zufällig generiert. Der p-Wert wurde durch das Anpassen einer Allgemeinen Extremwertverteilung an diese Energiewerte approximiert.

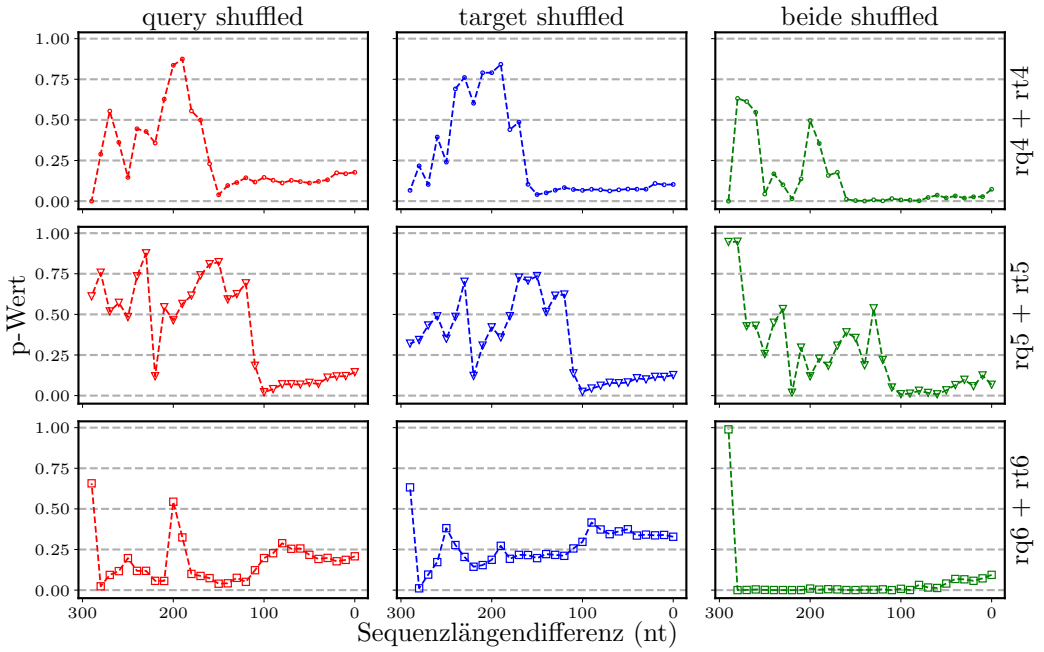

Abbildung 12.: p-Werte für verschiedene Längenunterschiede zwischen query und target

In Abbildung 12 ist der Verlauf der p-Werte für den kleiner werdenden Sequenzlängenunterschied zwischen query und target zu sehen. Für kurze queries ist eine sehr starke Schwankung des p-Werts bei jedem der drei Sequenzpaare zu erkennen. Diese Schwankung ist jedoch bei jedem Beispiel ab einer fixen Längendifferenz stark

abnehmend. Sehr auffällig ist, dass dies unabhängig vom shuffle-Modus immer an etwa derselben Stelle der Fall ist. Für die Sequenzen rq4 + rq4 beginnt dies ab einer Differenz von circa 220 nt, also einer Länge des querys von 120 nt. Das Paar rq5 + rt5 hat dieses Phänomen etwas später ab einer Differenz von circa 110 nt, also einer query-Länge von 190 nt. Das letzte Beispiel rq6 + rt6 hingegen ist bereits ab 290 nt Differenz recht konstant, also ab einer Länge des querys von 10 nt. Um diese Beobachtung zu erklären, wurde mit IntaRNA die beste Interaktionsstelle des jeweiligen Sequenzpaars überprüft. Hierbei hat das Sequenzpaar rq4 + rt4 diese für das query zwischen den Nukleotiden an Position 126 bis 134 und das target zwischen 51 bis 59. Beim Beispiel rq5 + rt5 liegt die Interaktionsstelle für das query zwischen 158 und 195 nt und für das target zwischen 165 und 205 nt. Das letzte Sequenzpaar interagiert besonders gut zwischen den Nukleotiden 5 bis 39 für das query und 149 bis 172 für das target. Da das target jedoch nicht in seiner Länge variiert wurde, spielen die Interaktionsstellen hierfür eine untergeordnete Rollte. Es ist allerdings in allen drei Beispielen zu sehen, dass der p-Wert etwa an der besten Interaktionsstelle des querys deutlich abnimmt. Auch die Schwankung des p-Werts nimmt massiv ab.

2.5. Unterscheidung von echten und unechten targets

Ziel dieser Untersuchung war eine Überprüfung, ob eine Unterscheidung von echten und unechten targets, also regulierenden und nicht-regulierenden targets, anhand des p-Werts möglich ist. In der Studie von Patrick R. Wright [21] wurde für eine Menge von Sequenzpaaren experimentell verifiziert, ob regulatorische Effekte gemessen werden können. In einigen Fällen konnte keine Regulation bestätigt werden. Diese targets sind nicht erwiesenermaßen unreguliert, dies ist allerdings zu erwarten. Im Folgenden werden diese targets als unwahrscheinliche targets betitelt. Für jedes dieser Sequenzpaare wurde die IntaRNA Interaktionsenergie und der p-Wert ermittelt.

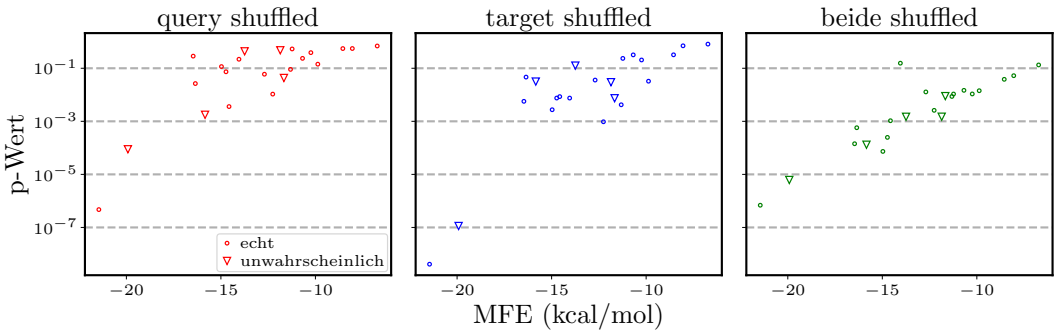

Abbildung 13.: p-Werte und Energiewerte ausgewählter echter und unechter targets

In Abbildung 13 sind die p-Werte und Energiewerte dieser Sequenzpaare zu sehen. Für keinen der drei shuffle-Modi ist eine Tendenz der echten oder unwahrscheinlichen targets zu erkennen. Wie auch schon in Abschnitt 2.3 besteht allerdings eine eindeutige Korrelation des p-Werts mit der Interaktionsenergie.

Im Anschluss wurde untersucht, ob eine Korrelation zwischen dem aus permutierten Sequenzen generierten und dem genomweiten p-Wert besteht. Hierfür wurden erneut die besten zehn targets der drei sRNAs Spot42, FnrS und CyaR aus [21] ausgewählt. Für jedes dieser Sequenzpaare wurde der p-Wert auf zweierlei Art bestimmt und miteinander verglichen. Zum einen wurde er aus 2500 zufällig permutierten Sequenzen wie zu Beginn dieser Arbeit beschrieben approximiert. Zum anderen wurde er durch

den IntaRNA Webserver genomweit berechnet. Hierzu werden als Hintergrundmodell lediglich alle möglichen Sequenzen aus dem angegebenen Genom verwendet und nicht alle durch Permutation möglichen Sequenzen. Somit ist das Hintergrundmodell des genomweiten p-Werts nur eine kleine Teilmenge des Hintergrundmodells des zufällig generierten p-Werts.

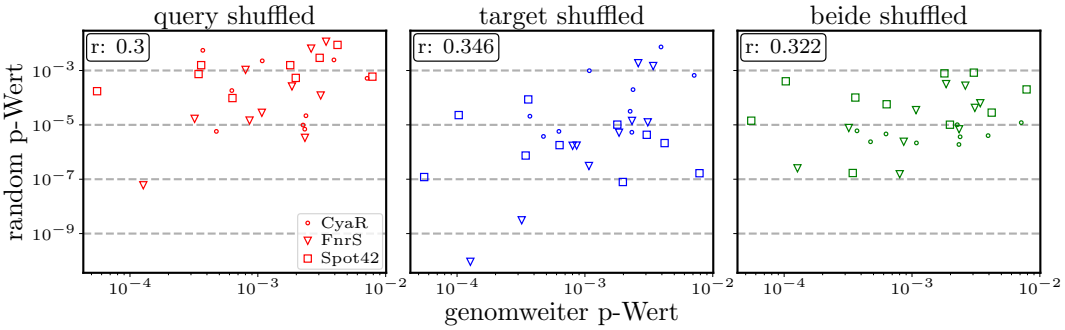

Abbildung 14.: Korrelation zwischen dem genomweiten p-Wert mit dem random p-Wert für die top-targets der jeweiligen sRNA

In Abbildung 14 sind für jedes Sequenzpaar die beiden p-Werte sowie Spearman's Rangkorrelationskoeffizient aufgetragen. Anhand des Rangkoeffizients sowie der starken Streuung der p-Werte ist ersichtlich, dass keine Korrelation zwischen den beiden p-Werten besteht. Allerdings ist der zufällig generierte p-Wert außer in wenigen Ausnahmen immer einige Größenordnungen kleiner als der genomweite p-Wert.

In einem letzten Versuch wurde überprüft, ob anhand des genomweiten p-Werts eine Unterscheidung zwischen echten und unechten targets möglich ist. Erneut wurden dieselben Sequenzpaare wie auch schon beim zufällig generierten p-Wert ausgewählt und dessen p-Wert bestimmt. Diese wurden zusammen mit ihrer Interaktionsenergie in einer Grafik aufgetragen.

In Abbildung 15 ist zu erkennen, dass auch in diesem Fall keinerlei Unterscheidung möglich ist. Allerdings besitzt der genomweite p-Wert, wie auch der zufällig Generierte, einen eindeutigen Zusammenhang mit der MFE der Strukturen.

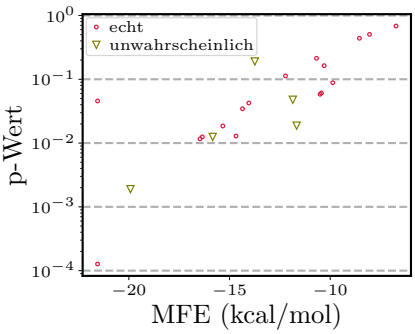

Abbildung 15.: Genomweite p-Werte und Energiewerte ausgewählter echter und unwahrscheinlicher targets

3. Diskussion

3.1. Verteilung von IntaRNA scores

Es konnte klar gezeigt werden, dass IntaRNA scores nicht normalverteilt, sondern extremwertverteilt sind. Besonders bei Verteilungen, die Energiewerte größer null beinhalten würden, welche von IntaRNA jedoch nicht berücksichtigt werden, ist die Normalverteilung durch den verschobenen Mittelwert sehr ungenau. Mit einer Allgemeinen Extremwertverteilung hingegen lassen sich selbst diese Verteilungen ausreichend genau approximieren. Allerdings ist das fitting des tails selbst mit dieser nicht trivial, sodass die eigentliche Verteilung oft leicht unter- oder überschätzt wird. Diese Ungenauigkeiten lassen sich nicht generell beheben, da jedes Sequenzpaar eine unterschiedliche Verteilung besitzt. Im Mittel wird die echte Verteilung durch die Allgemeine Extremwertverteilung allerdings am besten genähert. Der p-Wert sollte somit als grober Richtwert betrachtet werden und nicht als absolute Größe.

3.2. Einfluss der Anzahl permutierter Sequenzen und des shuffle-Modus

Wie erwartet konvergiert der p-Wert für eine größer werdende Anzahl permutierter Sequenzen gegen einen festen Wert. Aus langen Sequenzen sind mehr verschiedene Permutationen möglich, als aus kurzen Sequenzen. Somit sollte für lange Sequenzen

auch eine größere Anzahl Permutationen verwendet werden, damit die meisten davon zur Approximation des p-Werts genutzt werden können. Jedes der drei Sequenzpaare ist von unterschiedlicher Länge und zeigt deshalb ab einer unterschiedlichen Anzahl scores eine Konvergenz. In allen Beispielen ist der p-Wert ab etwa 1000 permutierten Sequenzen recht stabil, weshalb dies als Richtwert für Sequenzen mit ähnlicher Länge angesehen werden sollte, um ein aussagekräftiges Ergebnis zu garantieren. Da die Laufzeit zur Berechnung des p-Werts nur Komplexität $O(n)$ besitzt, wobei n der Anzahl permutierter Sequenzen entspricht, sollte immer erwägt werden, eine größere Menge zu verwenden, falls eine präzisere Berechnung des p-Werts notwendig ist.

Wie in Abbildung 9 zu erkennen, kann für eine zu kleine Stichprobenzahl der p-Wert stark schwanken. Da die scores völlig zufällig generiert sind, ist es somit auch möglich, dass besonders extreme Werte sehr dicht aufeinander entstehen. Auch dieses Problem lässt sich nur durch eine größere Stichprobenzahl beheben.

Durch den fehlenden Trend in den Verteilungen, die zur Untersuchung des shuffle-Modus erzeugt wurden, lassen sich keine generellen Aussagen über den Einfluss des shuffle-Modus treffen. Aus der Untersuchung des Einflusses des shuffle-Modus lässt sich jedoch eine Hypothese aufstellen, die die Beobachtungen erklären könnten. Wenn eine Sequenz mit vielen erreichbaren Interaktionsstellen nicht geshuffled (also festgehalten) wird, sind deutlich niedrigere Interaktionsenergien möglich, als wenn eine Sequenz mit keinen bis kaum erreichbaren Interaktionsstellen festgehalten wird. Die direkte Korrelation des p-Werts mit der MFE der Sequenzen, welche in Abschnitt 3.3 diskutiert wird, führt zu einem größeren/kleineren möglichen p-Wert.

3.3. Korrelation zwischen p-Wert und MFE

Wie erwartet besteht eine Korrelation des p-Werts mit der MFE der targets. Durch die kleine Stichprobenzahl ist allerdings keine generelle Korrelation bewiesen, diese liegt allerdings nahe. Die echten targets, deren p-Wert im kumulativen Plot aufgetragen

sind, folgen diesem Trend ebenfalls. Das Gruppieren der p-Werte jeder sRNA ist durch die ähnlichen Eigenschaften, etwa dem selben GC Gehalt erklärbar. In diesem Beispiel ist es allein anhand des p-Werts nicht möglich, echte von unechten targets zu unterscheiden. Die echten targets liegen durch ihren deutlich größeren p-Wert nicht in der Nähe der besten targets. Um eine Unterscheidung anhand des p-Werts zu ermöglichen, hätten diese einen deutlich kleineren p-Wert haben müssen. Es scheint vielmehr keinen Zusammenhang zwischen der bewiesenermaßen regulierenden Funktion der targets und des auf dieser Weise berechneten p-Werts zu geben.

3.4. Einfluss der Längendifferenz zwischen query und target

Aus dem Verlauf des p-Werts in Abbildung 12 lässt sich eine Hypothese aufstellen, die die abrupte Abnahme der Schwankung des p-Werts erklären könnte. Da das target in seiner Länge konstant bleibt und lediglich das query variiert wird, muss die beste Interaktionsstelle des gesamten querys mit dem gesamten target betrachtet werden. Sobald diese Interaktionsstelle Teil des querys wird, wird die Interaktionsenergie und somit auch der p-Wert kleiner und konstanter. Durch die nachfolgende Untersuchung der Beispiele lässt sich diese Hypothese bestätigen. Die Interaktionsstellen passen recht genau zu den Längen, bei denen die Schwankung abnimmt. Oft ist eine minimale Sequenzlänge für eine Interaktion mit guter Energie nötig, weshalb sich die starke Schwankung im Bereich der großen Sequenzlängenunterschiede erklären lässt. Durch das weitere Verlängern nach Hinzukommen der Interaktionsstelle steigen die Strafterme für die Berechnung der MFE durch IntaRNA an, weshalb auch der p-Werts ab diesem Punkt leicht steigt. Durch das Verlängern des querys werden zum Teil auch neue Interaktionsstellen an sowohl query wie target freigelegt. Um generelle Aussagen über den Einfluss der Längendifferenz treffen zu können, müsste die Anzahl Stichproben jedoch deutlich erhöht werden. Auch müssten zufällige queries der erforderlichen

Länge generiert werden und nicht nur ein Teil desselben querys. Dies war in dieser Arbeit aus Zeitgründen allerdings nicht möglich, da jeder p-Wert in Abbildung 12 anhand von 2500 zufällig generierten IntaRNA scores approximiert wurde. Somit hatte jedes der Sequenzpaare eine Laufzeit von circa 15 Stunden. In einer Folgearbeit wäre eine Untersuchung dieses Sachverhalts mit größerer Rechenleistung und somit größerer möglicher Stichprobenzahl denkbar.

3.5. Unterscheidung von echten und unechten targets

Die Korrelation zwischen MFE und p-Wert konnte hier erneut, wie bereits in Abschnitt 2.3, bestätigt werden. Allerdings scheint es keinen Zusammenhang zwischen dem p-Wert eines targets und dessen regulierender Funktion zu geben. Es konnte somit in dieser Untersuchung nicht nachgewiesen werden, dass eine Unterscheidung von echten und unechten targets anhand des p-Werts möglich ist. Der genomweite p-Wert nutzt zur Berechnung nur etwa 4000 target Sequenzen aus demselben Genom. Diese sind sich untereinander sehr viel ähnlicher, besitzen also deutlich mehr gleiche Eigenschaften untereinander, als die Sequenzen, die zur Berechnung des aus zufällig permutierten Sequenzen errechneten p-Werts genutzt werden. Aus diesem Grund sind die genomweiten p-Werte in fast allen Fällen um einige Größenordnungen größer als die zufälligen p-Werte. Auch die fehlende Möglichkeit echte von unechten targets anhand des genomweiten p-Werts zu unterscheiden, lässt generell an der Praktikabilität des p-Werts zweifeln.

4. Fazit und Ausblick

Die Ergebnisse dieser Arbeit legen nahe, dass der aus permutierten Sequenzen errechnete p-Wert nicht für die Unterscheidung von echten und unechten targets verwendet werden kann. Es wurde lediglich eine direkte Korrelation des p-Werts mit der MFE der Sequenzpaare aufgezeigt. Auch anhand des genomweiten p-Werts ist keine Unterscheidung von echten und unechten targets möglich. Dies führt dazu, dass es in absehbarer Zukunft kaum Anwendung den p-Wert geben wird. In einer anschließenden Untersuchung wäre jedoch eine Abwandlung der hier beschriebenen Berechnung denkbar, bei dem weitere Eigenschaften der Sequenzen erhalten bleiben würden. Dies könnte zum Beispiel die Erhaltung der Trinukleotidhäufigkeiten anstatt der hier konstanten Mono- und Dinukleotidhäufigkeiten sein. Auch könnten beim Permutieren gewisse Bereiche der Sequenzen nicht verändert werden, um eine noch größere Ähnlichkeit der permutieren Sequenzen zur ursprünglichen Sequenz zu erhalten. Durch diese Abwandlungen könnte eventuell eine Unterscheidung von regulierenden und nicht regulierenden targets möglich sein.

A. Anhang

Im Folgenden ist eine Tabelle mit allen verwendeten RNA-Sequenzen aufgelistet. Die sRNAs CyaR, Spot42 und FnrS, sowie die Sequenzen b* stammen aus dem Genom NC_000913 des Organismus Escherichia coli.

Tabelle 2.: Die jeweils zehn targets mit den niedrigsten Energien zu den verwendeten sRNAs

sRNA	**Beste targets**
CyaR	b0723 b0837 b1182 b1201 b2274 b2481 b3206 b3730 b3987 b4621
Spot42	b1207 b0158 b2193 b4455 b2398 b4647 b3537 b2371 b2832 b4589
FnrS	b0986 b3239 b2306 b1237 b0237 b2245 b3901 b0932 b0843 b2070

Tabelle 3.: Die jeweils echten targets zu den verwendeten sRNAs

sRNA	**Echtes target**
CyaR	b0723
Spot42	b0728
FnrS	b1531

Tabelle 4.: Alle verwendeten RNA Sequenzen mit Name, Länge und Rohsequenz

Name	Länge	Sequenz
CyaR	84 nt	GCUGAAAAACAUAACCCAUAAAAUGCUAGCUGUACCAGGAACCACCUCCUUAGCCUGUGUAAUCUCCCUUACACGGGCUUAUUU
Spot42	109 nt	GUAGGGUACAGAGGUAAGAUGUUCUAUCUUUCAGACCUUUUACUUCACGUAAUCGGAUUUGGCUGAAUAUUUUAGCCGCCCCAGUCAGUAAUGACUGGGGCGUUUUUUA
FnrS	122 nt	GCAGGUGAAUGCAACGUCAAGCGAUGGGCGUUGCGCUCCAUAUUGUCUUACUUCCUUUUUUGAAUUACUGCAUAGCACAAUUGAUUCGUACGACGCCGACUUUGAUGAGUCGGCUUUUUUUU
b0723	300 nt	GGUUUCUUCGCCUCUGCGUUCACCAAAGUGUUCACCCUGCUGGCGCUGUUUUCUAUCUUGAUCCAUGCCUGGAUCGGCAUGUGGCAGGUGUUGACCGACUACGUUAAACCGCUGGCUUUGCGCCUGAUGCUGCAACUGGUGAUUGUCGUUGCACUGGUGG
		UUUACGUGAUUUAUGGAUUCGUUGUGGUGUGGGGUGUGUGAUGAAAUUGCCAGUCAGAGAAUUUGAUGCAGUUGUGAUUGGUGCCGGUGGCGCAGGUAUGCGCGCGGCGCUGCAAAUUUCCCAGAGCGGCCAGACCUGUG
b0728	300 nt	CGCUGACGCGUCUUAUCAGGCCUACGGGACCACCAAUGUAGGUCGGAUAAGGCGCAAGCGCCGCAUCCGACAAGCGAUGCCUGAUGUGACGUUUAACGUGUCCUUAUCAGGCCUACGGGUGACCGACAAUGCCCGGAAGCGAUACGAAAUAUUCGGUCUAC
		GGUUUAAAAGAUAACGAUUACUGAAGGAUGGACAGAACACAUGAACUUACAUGAAUAUCAGGCAAAACAACUUUUUGCCCGCUAUGGCUUACCAGCACCGGUGGGUUAUGCCUGUACUACUCCGCGCGAAGCAGAAGAAG
b1531	300 nt	AAGGUUGCCGAACCCGAAUGACAAGCGCGGCGUACUGGUAAAACUUACCACCGGCGGCGCGGCAAUAUGUGAACAAUGCCAUCAAUUAGUUGGCCAGGACCUGCACCAAGAAUUAACAAAAAACCUGACGGCGGACGAAGUGGCAACACUUGAGUAUUUG
		CUUAAGAAAGUCCUGCCGUAAACAAAAAAGAGGUAUGACGAUGUCCAGACGCAAUACUGACGCUAUUACCAUUCAUAGCAUUUUGGACUGGAUCGAGGACAACCUGGAAUCGCCACUGUCACUGGAGAAAGUGUCAGAGC
b0837	300 nt	GUCAACGCCGAUAAGCUUGAAAAUUGUCUGAUGAGCUUAACGCUGAGUAUCCAGAGCCUGCAGAAGCACGCCAUGCUUGAGAAGGCCUGAAAACUAAGGGGGAGAAAGCGUCUCCCCCUUCAUGUUUAAGUUUUUGUAAAAAUGAAUUUGUUAUCUCCUC
		CACUGACUACGCUUUAAGCCAGAGUCAAUCCGGAGGCGUUAUGCAUCGACAAUCCUUUUUCCUUGUGCCCCUUAUUUGUCUUUCUUCCGCUCUCUGGGCGGCUCCUGCAACGGUAAAUGUCGAAGUACUGCAAGACAAAC
b1182	300 nt	AAAUAUCACCCGGCUAAAAAAUAGAAUAGAAGCAUUCGCCAUAAUGACAUUAAACAUUGUUUGAUAUUUAUCAUAUUAAUAGAAAUAAAGACAUUGACGCAUCCCGCCCGGCUAACUAUGAAUUAGAUGAAGUAAAAUUUAUUAAUAGUUGUAAAACAGG
		AGUUUCAUUACAAUUUAUAUAUUUAAAGAGGCGAAUGAUUAUGACUGAAAUCGUUGCAGAUAAAACGGUAGAAGUAGUUAAAAACGCAAUCGAAACCGCAGAUGGAGCAUUAGAUCUUUAUAAUAAAUAUCUCGAUCAGG
b1201	300 nt	CAUGGAUGAUGUUCAACGACACGGCGACCACUGACUGCCGAUGAAUCCAUUGUGCAUCAGGCAAGGGAAAAGAAAAUUAGCCCAAAUUAUGUUUCAUAGUGAAACAUAUGCUUUAAUGAAUGUUCCAUAUUGAAACUUUUACGUGUAUUAAUACUUAAAA
		UUGCGAGCCGGAACACCUUUUGUCAUAAGGGAUGCGGGAUAUGAGUGGCGCUUUUAACAACGAUGGUCGGGGCAUAUCUCCCUUAAUUGCAACCUCCUGGGAGCGAUGCAAUAAGCUGAUGAAACGGGAGACAUGGAACG
b2274	300 nt	CCUUUAAAUAUUUAUGAUAUUAAAUAUAACAACGUCUGGAUAACGACCCGGCCUGACUGAAUAUCUCACGUUUUCCUGCUUAACAUUCAUGUUAUGAUGCUCUUGAAACUAAUAAUACGUAUGGAAUUGUCACUAUUUAUCUGCUAAUGAAUUCAAACAA
		UGUCAUUCGUUUGAUGAAUAUAUUAUUUCAGGGAAUAAUUAUGACACCUUUAGAACGCAUUACCCAACUUGUUAAUAUCAAUGGUGAUGUUAAUAAUCCUGAUACUCCUCGUCCAUUGUUAUCUCUUGAGGAUUUUUUUA
b2481	300 nt	CUCGUGAUCAAGAUCACAUUCUCGCUUUCCCCUGCGACACGGGUGUCGAAUCCAUUUUUUGCUGAACGUUAAUGACCAUCAUUUUUGUACCGUUCAGAAUCCAGUUAAUACAUAACUUAUUGAAUAUAUUGAGUUAAUCAGAAUGGCAUCCUUUAUGCAA
		UAUGAAAUGCAAUGUUUCAUAUCAUUUUCAAGGAGCCGACAUGAACCGCUUUGUGGUGGCCGAACCACUGUGGUGUACAGGAUGUAAUACCUGUCUCGCUGCCUGUUCGGACGUGCAUAAAACGCAAGGUUUACAGCAAC
b3206	300 nt	UUUAUCUACCAGACGCGAAGCUAUCUUGAGCUAUGGUUACCUAUGCUGGAAACCAACAACCGUAGCUACUUGACGGUCGCCAUUGGUUGUACCGGCGGGAAGCACCGUUCGGUGUAUAUUGCAGAGCAACUGGCAGACUACUUCCGCUCGCGCGGUAAAA
		ACGUCCAGUCACGCCAUCGUACGCUGGAAAAACGUAAACCAUGACCGUCAAGCAAACUGUUGAAAUCACAAACAAGCUGGGCAUGCAUGCCCGGCCUGCAAUGAAGCUGUUUGAAUUAAUGCAGGGUUUUGACGCUGAAG
b3730	300 nt	UCCACCUGACGCUUAAAUUAAGGUACUGCCUUAAUUUUCUGCAGACAAAAGGCGUGACGAUGGUCGAAAAUGGCGCUUUCGUCAGCGGGGAUAAUCCGUUAUUGAACAAUUUAUCCUCUGUCCAUUUCACGAUGAAAAAAAUGUAGUUUUUUCAAGGUGA
		AGCGGUUUAAAUUCGUUCUCAAAUUACAGUCAGGACGCGUAUGUUGAAUAAUGCUAUGAGCGUAGUGAUCCUUGCCGCAGGCAAAGGCACGCGCAUGUAUUCCGAUCUUCCGAAAGUGCUGCAUACCCUUGCCGGGAAAG
b3987	300 nt	CUUUCAAUGCUUGUUUCUAUCGACGACUUAAUAUACUGCGACAGGACGUCCGUUCUGUGUAAAUCGCAAUGAAAUGGUUUAAGCGUGAUAGCAACAGGCAUUGCGGAAAGUGUUCCAUUUUCCGGUCAACAAAAUAGUGUUGCACAAACUGUCCGCUCAA
		UGGACAGAUGGGUCGACUUGUCAGCGAGCUGAGGAACCCUAUGGUUUACUCCUAUACCGAGAAAAAACGUAUUCGUAAGGAUUUUGGUAAACGUCCACAAGUUCUGGAUGUACCUUAUCUCCUUUCUAUCCAGCUUGACU
b4621	300 nt	UACAUUAAUUUUACGUUUAUUAUGAUGAAUCCUCGGCAAAUGAGCGAAAUUUGAUAGAACAUUUAUCUGCUGUACGAUUUAAUGAACUUAAUGAUACAGCAUUACAAAAACAAUCGAAGUUUAUAAAGAUGAUUUCUGAUUGACCAGCCCCCUACCCUAC
		AAUGGUGAUUAAUUUUAUUACUCGCGAUGGAGAUGAUGAUAUGAAUAUAUCAUAUGUGAAUAGCAACAAAACGACAUCUUUACCUGUUGAGCUAGAUGCACUAAAUAAUAAAGAUAUCUCUUAUGCCAAAGAUUUCUUUU
b1207	300 nt	UUGUGGCGAAAGGCGCUAAUCUUUCCCCAUUGCACAGAGCCAUGCUUUAAGCCGGGCAAGCUGAGUUUCGGUGACAACGUCACCUUGUUCCAGACGUUGCAUCGCGCUCUUUAAUACACCGCCUGGAAAGGAUCAUGCCUGGCCCGCACAGUUUUCGGCA
		GAUUCUUUCCACCAAUGGACGCAUGCCUGAGGUUCUUCUCGUGCCUGAUAUGAAGCUUUUUGCUGGUAACGCCACCCCGGAACUAGCACAACGUAUUGCCAACCGCCUGUACACUUCACUCGGCGACGCCGCUGUAGGUC
b0158	300 nt	CUUCCCACAGGCCAUUGCUGUAGAGAUGGAAGCGACGGCAAUCGCCCAUGUCUGCCACAAUUUCAACGUCCCGUUUGUUGUCGUACGCGCCAUCUCCGACGUGGCCGAUCAACAGUCUCAUCUUAGCUUCGAUGAGUUCCUGGCUGUUGCCGCUAAACAG
		UCCAGCCUGAUGGUUGAGUCACUGGUGCAGAAACUUGCACAUGGCUAAGUCACUGUUCAGGGCGCUGGUCGCCCUGUCUUUUCUUGCGCCACUGUGGCUCAACGCCGCGCCGCGCGUCAUCACGCUUUCUCCCGCCAACA
b2193	300 nt	UUGAUUUUUAGUUGUUUUUCUUGAUGAGAAGCUGAUGCAAAAUUCCGUCUUUAUAAUGAAAAUGAUGCCAAAGCGAACGACAAGGUUGUAGUUUUCACUACAUGUCCAUACAUAAAAUGGGGUAACAUUCACGCGCCUGGUAGCGUUACCAACGCUACG
		CUCAAACAUAAUGAUUCUAAUAAAACCUCAGGAGACUACUAUGCCUGAAGCAACACCUUUUCAGGUGAUGAUUGUGGAUGAUCAUCCACUUAUGCGACGCGGUGUUCGUCAGUUACUGGAGCUUGAUCCUGGCUCUGAAG

b4455	300 nt	UCCUGACCUGGGUUCGGGGGACACCAAAACGUGCCGAGAUGAUCCUGUAACCAUCAUCAGUUGUGAAGUAGUGAUUCACGACUUCAAGGCGCUUUUCAAAAGGGUAUUUUGGCUUUGACAUAUUAGGGGCUAUUCCAUUUCAUCGUCCAACAAAAUGGGU GCAGUACAUACUCGUUGGAAAUCAACACAGGAGGCUGGGAAUGCCGCAGAAAUAUAGAUUACUUUCUUUAAUAGUGAUUUGUUUCACGCUUUUAUUUUUCACCUGGAUGAUAAGAGAUUCACUGUGUGAAUUGCAUAUUA
b2398	300 nt	AUAUGAAUUCCGCCCGAAUGUGUCAACGGCAAAUUGCAACGUGUAGUUUCAAUCGCUGAAAAAUCAGGCAAAUGAACAAUUUUGGAAAACCGCUCGCAUUCAGUAGUUAUUCAUGUCACGGUUUCCUGUAAAGUGGUGUUAUAAAAUGAACUACUAAUAG ACCCACAUACAUUCAGGGAAUUGUUAUGUUCAAGGAGAGGAUGACACCAGAUGAACUUGCCAGACUGACCGGCUAUAGCCGCCAGACCAUUAAUAAAUGGGUACGCAAGGAAGGCUGGACAACGUCACCAAAACCUGGCG
b4647	300 nt	UUCUCCAUGUUUUUCAUAGAGGGCAACCCAUGUCCUGACCUGGGUUCGGGGGACACCAAAACGUGCCGAGAUGAUCCUGUAACCAUCAUCAGUUGUGAAGUAGUGAUUCACGACUUCAAGGCGCUUUUCAAAAGGGUAUUUUGGCUUUGACAUAUUAGGG GCUAUUCCAUUUCAUCGUCCAACAAAAUGGGUGCAGUACAUACUCGUUGGAAAUCAACACAGGAGGCUGGGAAUGCCGCAGAAAUAUAGAUUACUUUCUUUAAUAGUGAUUUGUUUCACGCUUUUAUUUUUCACCUGGAU
b3537	300 nt	AACCGUAUGGUCUGGUUUGAAGAUGAUCAAAUCAGUGCCGAGCUGGUGCAGAUGCGCUUGCUUGCCCCAGAACAAUGGGGCAUGCCGCUGCCUUUAACGCAAAGUUCUAAACCGGUCAUCAAUGCCGAGCACGAUGGUCGCCACUGGCGACGAAUACCAG AACCCAUGCGACUGUUAGAUGAUGCUGUGGAGCGCUCAUCAUGAUGACCAUCAGCGAUAUCAUUGAAAUUAUUGUCGUUUGCGCACUGAUAUUUUUCCCCGCUGGGCUAUCUGGCGCGGCACUCUUUGCGACGCAUUCGCG
b2371	300 nt	CGGAAUCAUUAUUGCCAGCCGGUUUAAUGUCUACACCCGUACUGGUACAGCGUCAUUGGCUGUGAGCGUACUGGGCUUUGUCGUCACGGCUCCCUUGUGGAUUUAUGUCAGUCGACUGGUUUCAUAAUUAAUCAUGUUAUUGAUUUUACAUAAAUGCCGA AGUGUCUUUUCACCUUCGGCUUUUUUCAAGGCAUAAGCAAAUGACAAAUAAUGAAAGCAAAGGGCCGUUUGAAGGCUUAUUAGUUAUCGAUAUGACACAUGUCCUUAAUGGACCUUUCGGAACUCAACUUCUUUGUAAUA
b2832	300 nt	GCAGCGAAUGCGAAAGCGCUUACCGAAGCCAUUGGUGCCCGGGGCGAACGGAUUCUGACGCUGCCACGCGGCUUUUAUUUGAAGAAGAAUUUCACCAGUGCACUACUGGCCCGUCAUUUUCUGAUCCAGUAGCCAUCGCUUUGACCUGCCGCUUUCCGGG CAUAUAAUUACCGCUUCAUUUUUUUGGCAGGGCUUUUUAGAUGUUAUUUGCAUGGAUAACCGAUCCUAACGCCUGGCUUGCGCUCGGUACGCUGACGCUGCUGGAGAUCGUUCUUGGGAUCGACAAUAUUAUUUUCCUUU
b4589	300 nt	CCGUAACCUGUCGGAUCGCCGGAAAGGACCCACAAAAUGAUAAUAAUUAUCAUCUACAUGUCACAACGUGCAUCUACGCCAUCAAACCACGUCAAAUAAUCAAUUAUGACGCAGGUAUCGUAUUAAUUGAUCUGCAUCAACUUAACGUAAAAACAACUUC AGACAAUACAAAUCAGCGACACUGAAUACGGGGCAACCUCAUGUCAACUAAGAACAGAACCCGCAGAACAACAACCCGCAACAUCCGCUUUCCUAACCAAAUGAUUGAACAAAUUAACAUCGCUCUUGAGCAAAAAGGGU
b0986	300 nt	UGUUGGUGUGGCGACGUUGGCAGUUGUUGCAGCAAAUGACAGUGGCGAUGGCGGUUCUCAUAACACCUCAACCACCACGUCCACUACACGCUAAUACUGCUUUAACGCGUCGAAAUUAAAGUUAAAUGAUACCCGGUUGCGAUAUUCGUACCGGGCUUUA GAGAUUUGCAUUUUUUUAAUUAACCUUUCAAGGAUUAAAGGUGCGCCCUCUUAUUUUAUCGAUUUUCGCACUAUUUCUUGCGGGAUGUACGCACAGCCAGCAAAGUAUGGUCGAUACAUUUCGCGCCAGCCUUUUCGAUA
b3239	300 nt	CUACCAGGAAAAGCUACGCAUCUGGCAAGCGCCACCGCAGGUGACGGAACCGGUAAAUCGGCUGGCGGGGAUGCUCCAUAAGUAUCAACAUGCGUUGACCGAUAGUUAAGUCAAAACCGACGCCAAAAGCGUCGGUUUUUUCAUGGCUAUACUUAGCGAU GCACGGCAGAACUCGCCGCGAAACGUGACGGUGGCAACAGAUGAAUAUUUAUACCUUUGAUUUUGAUGAGAUUGAGAGUCAGGAGGAUUUUUAUCGUGACUUUAGCCAAACCUUUGGUCUGGCGAAAGAUAAGGUACGCG
b2306	300 nt	GUUGGCAUUUACUGCCACGGUGCCGGAUCUGCUGAAAAUAGCCCGCGAUAUUAACGCCGCCACGUAUCAACCUUUUACCGCCUUCGGCAUUGCCGCGGUGCUCUAUUUAAUCAUCUCUUAUGUCCUGAUCAGCCUCUUUCGCAGAGCGGAAAAACGCUGG UUGCAGCAUGUGAAACCUUCUUCAACGCACUGAGAAUACGAUGUCCGAGAAUAAAUUAAACGUUAUCGAUUUGCACAAACGCUACGGCGAACAUGAAGUGCUGAAAGGGGUAUCACUGCAAGCGAAUGCCGGAGAUGUAA
b1237	300 nt	CUCACGUGCUGCGAAAUCAUCGGUGUAAAUAGGGCUAUAUGCCGCGUCUUUUCUGGCUAAUUUUAUGAAAAGAUAUUUAUUGGCGGCACAAAAUAAAGAACAAUUUUGAAUUCCUUACAUUCCUGGCUAUUGCACAACUGAAUUUAAGGCUCUAUUAUUA CCUCAACAAACCACCCCAAUAUAAGUUUGAGAUUACUACAAUGAGCGAAGCACUUAAAAUUCUGAACAACAUCCGUACUCUUCGUGCGCAGGCAAGAGAAUGUACACUUGAAACGCUGGAAGAAAUGCUGGAAAAAUUAG
b0237	300 nt	CGCGAGAUUAUAGAGAUCUGGCGCACUAAAAACCAGUAUUUCACAUGAGUCCGCGUCUUUUUACGCACUGCCUCUCCCUGACGCGGGAUAAAGUGGUAUUCUCAAACAUAUCUCGCAAGCCUGUCUUGUGUUGACAACAUUUUCUGCUAACCCUGUGACC UGCAAUACUGUUUUGCGGGUGAUCGACAAGGAGACUUAACGUGUCUGAACUGUCUCAAUUAUCUCCACAGCCGCUGUGGGAUAUUUUUGCCAAAAUCUGUUCUAUUCCUCACCCGUCCUAUCAUGAAGAGCAACUCGCUG
b2245	300 nt	AACCUGUUUGGUGCAGUGGGCGGGUUUAUUGCCCCGAUCCUGCGCGUGAAAGCAGAAACGUUAUUUGCCAGCGAUGCGGCGGGAUUACUGACGCUGGCAGCGGUGGCGGUCAUCGGUUCGCUGAUUAUUUUCACUCUGCGUGUAAAUCGCACUGUUGCGC AGACCGACGUGGCACAUCAUUAAAUAGGUUAAGGAACACGAUGAACGCAUUAUUAAGCAAUCCCCUUUAAAGAACGUUUACGCAAGGGCGAAGUGCAAAUUGGUCUGUGGUUAAGCUCAACGACUGCCUAUAUGGCAGAAA
b3901	300 nt	ACGGCGUCUUCGGCAGCGGACCGACGCUGGAUGAAACCUUCGGUUUAAUCGACACCGCAGAAAAAUCAGCACAAGUAUUAGUGAAGGUUUAUUCGAUGGGCGGCAUGAAACAGACCAUCAGCCGUGAAGAGUUGAUAGCGCUCGGCAAGCGUUUCGGCGU UACGCCACUCGCCAGUGGCGCUGGCGCUGUAAGGAGCAAACAUGAUCCGCAAAGCCUUUGUCAUGCAGGUAAACCCCGACGCCCACGAAGAGUAUCAGCGUCGGCAUAAUCCCAUCUGGCCAGAACUGGAAGCAGUGCUGA
b0932	300 nt	GUAUACCUUGUUUCGCAAUUUAUUGAACCCCGAUCACACCAUAUGCCACCUUUCUGGUCGAUGGCAUUACGCGGUGUAUGUUAUAAAAAUGUAGCAAUAAAGGCGUUUGUACCUGAAAAGAUGAAGAUUCUGCAUAGCGCGAUUUACGCAACAGGAAUAG ACUGAACACCAGACUCUAUAAAAGAUGCUAAAGGUUAUUUAUGACUCAACAGCCACAAGCCAAAUACCGUCACGAUUAUCGUGCGCCGGAUUACCAGAUUACUGAUAUUGACUUGACCUUUGACCUCGACGCGCAAAAGA
b0843	300 nt	UGCCGGUAGCGCGGUCGUCGCAGCGGCAAAAUACCGCGCAGGCUCCAAUAACCGUGAAGGCGUACUGGAUGUGAUCGAUAAAGUUCUUAAACACGAAGCGCCAUUUGACCAAUAAAACCGGCAGCACGCCCGGCAACACAAUGCCUGAAUGAAGGCUGCA UAAUUCGCGGAAAUUAACCCGCUGCAUUAUUUCAGGCGAAAUGAUUAUGAAAAAUUGUCUACUGUUGGGCGCGCUUUUAAUGGGCUUUACUGGUGUGGCGAUGGCGCAAAGUGUCACCGUCGAUGUGCCGAGUGGAUACA
b2070	300 nt	UUUUUUCUCCAGAAUCCCUUUGUAGUCAAAGUUAGCAAAAACCUGUGGAAGUUAAUCAUUAUCAGGCCGAAAUUGAAUUGUAUCCUGAUAAAAUUGCAUUAUCAUUUUUAUUCUUGAGUGGGUUUUCCCCAUAUAGAAAUAAUUUUAAUGGUUAGUUGUU AUUAUUCAUUGAAAUAUAAUAUAACUACCCGCGAAAUAUUAUGAAAACAAAUAUAAAAGUAUUUACAUCGACAGGUGAAUUGACCACUCUUGGCCGUGAACUGGGCAAAGGUGGUGAAGGCGCGGUUUAUGAUAUCGAGG
rq1	60 nt	AGGAUGGGGGAAACCCCAUACUCCUCACACACCAAAUCGCCCGAUUUAUCGGGCUUUUUU
rt1	175 nt	UUUAAAUUAAAAAAUCAUAGAAAAAGUAUCGUUUGAUACUUGUGAUUAUACUCAGUUAUACAGUAUCUUAAGGUGUUAUUAAUAGUGGUGAGGAGAAUUUAUGAAGCUUUUCAAAAGCUUGCUUGUGGCACCUGCAACUCUUGGUCUUUUAGCACCAAUG ACCGCUACUGCUAAU

rq2	84 nt	ACACCGUCGCUUAAAGUGACGGCAUAAUAAUAAAAAAAUGAAAUUCCUCUUUGACGGGCCAAUAGCGAUAUUGGCCAUUUUUUU
rt2	1289 nt	ACGUUAAUGCAAUCAAAAAUGGCUUUAUUCGCUUAAUGCCGUUAACCCUUGCGGGGGCCAUGUUUGUAUUAAUUAACAACGUUUUUCUAAGCUUUGGGGAGGGGUCGUUUUUUUAUUCCUUAGGUAUUCGCCUCGACGCCUCAACCAUUGAAACACUUAA UGGUCUGAAAGGUAUUGGCGGCAACGUAUAUAACGGAACAUUAGGAAUAAUGUCUUUAAUGGCACCGUUCUUUAUUGGCAUGGCGCUGGCAGAAGAGCGUAAAGUCGAUGCGCUGGCGGCUGGGUUGUUAUCCGUUGCAGCAUUUAUGACCGUCACCCCA UAUAGUGUCGGUGAGGCCUAUGCGGUUGGUGCAAACUGGUUAGGUGGGGCGAAUAUCAUCUCCGGGAUUAUUAUUGGCCUGGUGGUGGCAGAAAUGUUUACCUUUAUUGUCCGCCGCAAUUGGGUCAUUAAACUGCCCGACAGCGUACCUGCUUCAGUAU CGCGUUCCUUCUCGGCAUUAAUUCCCGGCUUUAUUAUUCUUUCCGUGAUGGGGAUUAUUGCCUGGGCGUUGAAUACCUGGGGCACCAACUUCCAUCAGAUCAUUAUGGAUACCAUCUCAACCCCACUGGCAUCGUUGGGUAGCGUGGUGGGCUGGGCCUA UGUGAUCUUUGUUCCACUGCUCUGGUUCUUCGGUAUUCAUGGCGCGCUGGCGCUGACCGCACUGGACAACGGCAUUAUGACGCCGUGGGCACUGGAAAAUAUCGCGACCUAUCAGCAAUAUGGUUCCGUCGAAGCGGCGCUGGCAGCCGGUAAGACCUUC CAUAUCUGGGCCAAGCCGAUGCUGGACUCCUUUAUUUUCCUUGGGGGCAGUGGUGCGACUUUAGGCCUGAUCCUGGCUAUCUUUAUCGCCUCUCGCCGUGCUGAUUAUCGUCAGGUGGCAAAACUGGCGCUGCCGUCCGGCAUCUUCCAGAUUAACGAAC CGAUUCUGUUUGGUCUGCCAAUUAUCAUGAACCCGGUGAUGUUUAUCCCGUUUGUACUGGUACAACCGAUUCUGGCGGCAAUCACCCUCGCAGCGUACUACAUGGGCAUUAUUCCUCCGGUGACCAAUAUUGCACCGUGGACCAUGCCAACCGGUCUGGG AGCCUUCUUUAACACCAACGGUAGCGUCGCCGCAUUGCUGGUCGCACUCUUCAACCUUGGAUCGCAACGUUAAUUCUAUCUGCCCUUUGUUGUGGUGGCUAACAAAGCACAAAAUGCGAUUGAUAAAGAAGAGAGCGAAGAAGAUAUCGCUAACGCCCUG AAAUUUUAA
rq3	46 nt	ACGGAUGUGGAGUUCGCUACACCUACUUUGGAGGUGCUUCGACUUU
rt3	175 nt	UUUACUAAUCUGGGGUGAGGUAUAUGUUAUUUCUGGGAACCAGCUUUGGCGUGCAAGCUUUGUUGUUUUAAUGCUGUAGUUCGCUUCAAAAUUUUUGAGUCUUAAAGUGGCCUGACAGGGGCUGAAGCACAACGAAGCAAUUAAAUUUAUUACUUGAAGC GGCUCUAAAUAUUCC
rq4	300 nt	AAAAACGUUUAAUAAAAAAUAUUUAAACAUAUAAUCGCGUGAGGCAUCAUUUGGGGCCCUAAAAUUUUAUAAACCCUUUUAGCCGAUUUUUUUCGCGGUUACCUUUGUUUUUACGCUUUUGCAACGGCCCCGUGGUACCCGCGGAAAAGGCCCUCAUCU CUUUCAGGUUCAUAAAACCAUAAUAUGCAUUCCAUUAAUAAAUUACUAAAUAGGAAUAUAGAUGGUUUAAGUCGCGGCACGAGGAAAAAAAUUGAAUAAUAAUAAAUAAGCCCGAGCAUUCGCGGCAAACGGAACGUUGC
rt4	300 nt	ACGUUAAUGCAAUCAAAAAUGGCUUUAUUCGCUUAAUGCCGUUAACCCUUGCGGGGGCCAUGUUUGUAUUAAUUAACAACGUUUUUCUAAGCUUUGGGGAGGGGUCGUUUUUUUAUUCCUUAGGUAUUCGCCUCGACGCCUCAACCAUUGAAACACUUAA UGGUCUGAAAGGUAUUGGCGGCAACGUAUAUAACGGAACAUUAGGAAUAAUGUCUUUAAUGGCACCGUUCUUUAUUGGCAUGGCGCUGGCAGAAGAGCGUAAAGUCGAUGCGCUGGCGGCUGGGUUGUUAUCCGUUGCAG
rq5	300 nt	UCUAGCGUCCCGCUCAGAAUAAUGGGGAUUACUCUUAUGCUAUCUCUUCCCGGUGGGAGCCGCACUCAGGAGCAUACCGAUGUGUGUAAUAGUAGCAUUACUGUACCAAGCUCCAAACCUAGGUUCCAGAAAGCACUUCGUGGUCCUGAAGUAGAUUUGC CUGACGGAGUCAAAUUGUCGCCGUACUCGUGUUGGAGGCUAACAAGGUCUCGGAUCGUCCACUUUAGUAGAUACAUCAACUUUGGUUCAUUAACAUAGUGCUUGUGUCGGACCCCCGGGCAGCGCCAAACAUUGGCUGGA
rt5	300 nt	CAAUUGUUACUUGUUACCAGAAUCGUGUGCGCGCUACCAGGGCAUACCGGACUAAUCAUUUAACUUUGGGACGAUAACUGAUAGGUGGUAGCCAGCGUUGCAAUGUCCACCUUUUAACCGACGCACCGCUAGGACUAAACACGGGUCCGUUCUUCAAUUU UUAACCACACGACUCUUGAGGUGACAGAUGCUCCGCUAACAGGCGCCCGCGUGGACCUGACUUCUUAGCCCCGCUAAGAUUUUUAUGACGGAUGGCCUGAUCACGGCCCAGCCUGAGCCAACGCCCGCUCGUCUCCGGUU
rq6	300 nt	GUUAAGGCUAUGGUGGAGCGUCGCGGGAGUAUCUGCGUCUGAUUGUGAAAAGUUUUGCAGUGUCUAGACGGUGGAUUGCACGAUACUAUGACUAAAGCGGCCCAACCGAUGUCCGUGUUGCUGCUUCUGAAAUAAAAUACUUGGCACGAUCGCAUCUUCU AUAAGGAGGCUUUGGGUAUUACGAUGUAAGGCGCACGGGGAUGGCACGUGCGUUUGGAAAGCCAACGACUCCCUCAGGGCCCCACCGUGUACGAAUCCACGUGGUGUUCCCGUCUAACUGUUGAUCCAUAUGUGCGGUAU
rt6	300 nt	AAGCGGUUAAACCUUUUACGCCGUCCCACACAAAGAUUGCACAUUACCCUCUAAGAUUUAUGGAUGAACCUCAUCGCUUCGAGCAGUACUAAUCCCACGCCCCCGAUUAGACUUUUUGUGUUCGGCAACUACUCGAGGCUUUGUUCGUCUCCCAUUGGUU CUUCUAUAGCCUACGCGCACCCCGAAGCGCGAUGAGAUCAUGGAGGUGGCAUUGGUACGACUGUAUCGGACUUGCUUCGUCCGGUCCGGCACCCCGGCAUUCAUAAGUAGGGGUCACGCUACGGUCUGGCGCUUGCCCUC

Danksagung

An dieser Stelle möchte ich mich noch bei all denen bedanken, die mich während meiner Bachelorarbeit unterstützt und motiviert haben. Besonderen Dank gebührt meinem Betreuer Martin Raden, der mir stundenlang jegliche Fragen beantwortet hat und mir zielgerichtete Ideen gab. Außerdem bedanke ich mich bei Herr Prof. Dr. Rolf Backofen für die Möglichkeit an seinem Lehrstuhl meine Bachelorarbeit zu schreiben. Abschließend möchte ich mich noch bei meinen Eltern und Freunden für ihre Unterstützung bedanken, ohne die dieses Studium nicht möglich wäre.

Literaturverzeichnis

[1] M. Mann, P. R. Wright, and R. Backofen, "IntaRNA 2.0: enhanced and customizable prediction of RNA-RNA interactions," *Nucleic Acids Research*, vol. 45, no. W1, pp. W435–W439, 2017.

[2] J. Gong, Y. Ju, D. Shao, and Q. C. Zhang, "Advances and challenges towards the study of rna-rna interactions in a transcriptome-wide scale," *Quantitative Biology*, vol. 6, no. 3, pp. 239–252, 2018.

[3] M. Raden, S. M. Ali, O. S. Alkhnbashi, A. Busch, F. Costa, J. A. Davis, F. Eggenhofer, R. Gelhausen, J. Georg, S. Heyne, M. Hiller, K. Kundu, R. Kleinkauf, S. C. Lott, M. M. Mohamed, A. Mattheis, M. Miladi, A. S. Richter, S. Will, J. Wolff, P. R. Wright, and R. Backofen, "Freiburg RNA tools: a central online resource for RNA-focused research and teaching," *Nucleic Acids Research*, vol. 46, no. W1, pp. W25–W29, 2018.

[4] G. Storz, J. Vogel, and K. M. Wassarman, "Regulation by small RNAs in bacteria: expanding frontiers," *Molecular Cell*, vol. 43, no. 6, pp. 880–891, 2011.

[5] C. Biele and H. R. Horton, *Biochemie*, vol. 4. Pearson Studium, 2008.

[6] G. Varani and W. H. McClain, "The G x U wobble base pair. A fundamental building block of RNA structure crucial to RNA function in diverse biological systems," *EMBO Reports*, vol. 1, no. 1, pp. 18 – 23, 2000.

[7] M. Raden, M. M. Mohamed, S. M. Ali, and R. Backofen, "Interactive implementations of thermodynamics-based RNA structure and RNA–RNA interaction prediction approaches for example-driven teaching," *PLOS Computational Biology*, vol. 14, no. 8, pp. 1–19, 2018.

[8] S. Smit, K. Rother, J. Heringa, and R. Knight, "From knotted to nested RNA structures: A variety of computational methods for pseudoknot removal," *RNA*, vol. 14, no. 3, 2008.

[9] R. Gelhausen, "Constrained RNA-RNA interaction prediction," Master's thesis, Albert-Ludwigs University of Freiburg, 2018.

[10] J. M. Engreitz, K. Sirokman, P. McDonel, A. Shishkin, C. Surka, P. Russell, S. R. Grossman, A. Y. Chow, M. Guttman, and E. S. Lander, "RNA-RNA Interactions Enable Specific Targeting of Noncoding RNAs to Nascent Pre-mRNAs and Chromatin Sites," *Cell*, vol. 159, no. 1, pp. 188–199, 2014.

[11] P. Yakovchuk, E. Protozanova, and M. D. Frank-Kamenetskii, "Base-stacking and base-pairing contributions into thermal stability of the DNA double helix," *Nucleic Acids Research*, vol. 34, no. 2, p. 564–574, 2006.

[12] R. Sullivan, M. C. Adams, R. R. Naik, and V. T. Milam, "Analyzing Secondary Structure Patterns in DNA Aptamers Identified via CompELS," *Molecules*, vol. 24, no. 8, 2019.

[13] P. N. Borer, B. Dengler, I. Tinoco, and O. C. Uhlenbeck, "Stability of ribonucleic acid double-stranded helices," *Journal of Molecular Biology*, vol. 86, no. 4, pp. 843–853, 1974.

[14] A. Busch, A. S. Richter, and R. Backofen, "IntaRNA: efficient prediction of bacterial sRNA targets incorporating target site accessibility and seed regions," *Bioinformatics*, vol. 24, no. 24, p. 2849–2856, 2008.

[15] J. Besag and P. Clifford, "Sequential Monte Carlo p-values," *Biometrika*, vol. 78, no. 2, pp. 301–304, 1991.

[16] K. D. Schmidt, *Maß und Wahrscheinlichkeit*, vol. 2. Springer-Verlag Berlin Heidelberg, 2011.

[17] S. F. Altschul and B. W. Erickson, "Significance of nucleotide sequence alignments: a method for random sequence permutation that preserves dinucleotide and codon usage," *Molecular Biology and Evolution*, vol. 2, no. 6, pp. 526–538, 1985.

[18] M. Jiang, J. Anderson, J. Gillespie, and M. Mayne, "uShuffle: A useful tool for shuffling biological sequences while preserving the k-let counts," *BMC Bioinformatics*, vol. 9, no. 1, p. 192, 2008.

[19] P. Clote, F. Ferré, E. Kranakis, and D. Krizanc, "Structural RNA has lower folding energy than random RNA of the same dinucleotide frequency," *RNA*, vol. 11, p. 578–591, 2005.

[20] W. M. Fitch, "Random sequences," *Journal of Molecular Biology*, vol. 163, no. 2, pp. 171 – 176, 1983.

[21] P. R. Wright, A. S. Richter, K. Papenfort, M. Mann, J. Vogel, W. R. Hess, R. Backofen, and J. Georg, "Comparative genomics boosts target prediction for bacterial small rnas," *Proceedings of the National Academy of Sciences*, vol. 110, no. 37, pp. E3487–E3496, 2013.