Albert-Ludwigs-Universität Freiburg
Lehrstuhl für Bioinformatik
Prof. Dr. Rolf Backofen

Signifikanz von RNA-RNA Interaktionen und von RNA-Sequenz-Struktur-Alignments

Bachelorarbeit

Betreuer:
Steffen Heyne
Andreas Richter

von
Benjamin Schulz
5. Januar 2009 - 6. April 2009

Danksagungen

Ich danke **Professor Dr. Rolf Backofen** für die Möglichkeit an seinem Lehrstuhl diese Bachelorarbeit schreiben zu dürfen.

Ich danke **Steffen Heyne** und **Andreas Richter** für die Betreuung meiner Bachelorarbeit.

Ich danke **Dr. Olaf Ronneberger** für die Einführung in Support-Vektor-Maschinen.

Außerdem danke ich den **Mitarbeitern** des Lehrstuhls für Bioinformatik, dass sie meine dauerhafte Belegung des Lehrstuhl-Grids mit Humor genommen haben.

Inhaltsverzeichnis

1 Zusammenfassung

In dieser Arbeit wurden Signifikanzuntersuchungen für die am Lehrstuhl für Bioinformatik der Universität Freiburg entwickelten Programme LocARNA und IntaRNA angestellt. LocARNA bewertet anhand eines Sequenz-Struktur-Alignments die strukturelle, sowie sequenzielle Ähnlichkeit zwischen zwei ncRNA Sequenzen, IntaRNA bewertet mögliche Bindestellen zwischen ncRNA und mRNA, unter Berücksichtigung nicht nur der Sequenz, sondern auch der Sekundärstruktur der Sequenzen.
Es wurde analysiert, wie sich die ausgegebenen Bewertungen dieser zwei Programme in Abhängigkeit von den Eigenschaften Länge, AU gegen GC Anteil und minimaler freier Energie der eingegebenen Sequenzen verändern. Dazu wurden große Mengen an zufälligen Sequenzen erzeugt, die Verteilung bei Sequenzen mit gleichen Eigenschaften untersucht und geprüft, wie sich die Verteilung bei Variation der Länge und des AU zu GC Mengenverhältnisses ändert.
Bei LocARNA wurde mit den Daten eine Support Vektor Maschine trainiert, die nun für Sequenzpaare die zu erwartende Verteilung angeben kann. Mit dieser Verteilung als Nullmodell ist es möglich, die P-Werte, und damit die Signifikanz, der von LocARNA ausgegebenen Bewertungen zu bestimmen.
Bei IntaRNA wurde festgestellt, dass die ncRNA Sequenzen einen Einfluss auf die Ausgabe haben, der sich nicht allein durch Länge, AU-Anteil und freier Energie erklären lässt. Hier sind weitere Untersuchungen nötig, bevor Gesetzmäßigkeiten bestimmt werden können mit denen die Signifikanz bewertet werden kann.

2 Einleitung

2.1 Bioinformatik

Die Bioinformatik hatte einen großen Anteil an der in den letzten Jahren vollendeten Sequenzierung des menschlichen Genoms und ist an der Weiterverarbeitung der aus der Sequenzierung erhaltenen Informationen beteiligt. Die Methoden der Sequenzierung von DNA und RNA sind immer besser und schneller geworden, sodass in relativ kurzer Zeit sehr große Mengen sequenziert werden können. Derzeitige Sequenzierungsmaschinen können pro Tag bis zu eine Milliarde Basen sequenzieren [15]. So ist es nicht praktikabel, für jede gefundene einzelne RNA Laborversuche durchzuführen, um ihre Aufgabe in der Zelle herauszufinden. Es werden Hilfsmittel benötigt, die eingrenzen, welche Funktionen die jeweiligen RNAs erfüllen, und mit welcher anderen RNA, DNA oder mit welchem Protein sie wahrscheinlich interagieren. Nur so kann man die große Menge an Daten, die man aus der Sequenzierung erhält, auf im Labor durchführbare Experimentenanzahlen reduzieren und größeres Verständnis für die Vorgänge in Zellen erlangen.

2.2 RNA

Ribonukleinsäure (Englisch: Ribonucleinacid = RNA) ist ein Grundbestandteil der Zelle. Nach Meinung mancher Wissenschaftler war die Entstehung von RNA im Urtümpel eventuell der erste Schritt auf dem Weg zu lebenden Organismen (Die RNA Welt Hypothese [7]).

RNA besteht, wie auch DNA (=Desoxribonucleinacid), aus einer Kette aus Zucker, Phosphat und den Basen Adenin, Cytosin, Guanin und Uracil bzw. Thymin in DNA.
In heutigen Organismen gibt es viele verschiedene Typen von RNA. Man kann sie in zwei Gruppen unterteilen, die kodierende RNA und die nicht kodierende RNA (ncRNA). Die kodierende RNA oder auch auch MessengerRNA (mRNA) wird als Abbild eines Teils der DNA gebildet (Transkription) und kodiert die nötigen Informationen für die Synthese eines Proteins. An den Ribosomen wird das von der mRNA kodierte Protein schließlich erzeugt (Translation).
Die Klasse der nicht kodierenden RNA umfasst dagegen alle RNA Sequenzen, die kein Protein kodieren. Dazu zählen etwa die „Transfer-RNA“ (tRNA), welche während der Translation den

Basentripeln der mRNA die passende Aminosäure zuordnet, oder die ribosomale RNA (rRNA), die ein wichtiger Teil der Ribosomen ist. Auch sogenannte Ribozyme, katalytisch aktive RNA Sequenzen, zählen zu den nicht kodierenden RNAs. Für diese Arbeit sind die „small interfering RNA“ (siRNA), die „micro RNA“ (miRNA) und die „small regulatory RNA“ (sRNA) wichtig. Vertreter dieser RNA können sich an mRNAs binden und so die Translation unterbinden, erst ermöglichen oder den Abbau der mRNA einleiten. Sie sind also ein wichtiger Bestandteil der Translationsregulation.

2.2.1 RNA Struktur

Man unterscheidet bei RNA zwischen verschiedenen Strukturen [4]:
Die Primärstruktur gibt die Abfolge der Basen der RNA an. Beispiel der Primärstruktur einer tRNA: GGAUUCGUGGCUCAAUGGUAUCGCGUCUGACUCCAGAUCAGAAGGUUGCGUGUUCGAUUCACGUCGGGUUCA. Jeder Buchstabe steht für die Base des Nukleotids an der entsprechenden Stelle (A = Adenin, C = Cytosin, G = Guanin und U = Uracil).

Die Sekundärstruktur gibt die zweidimensionale Anordnung des RNA Strangs an. Wie Proteine auch, liegt die RNA nicht einfach als Faden vor, sondern es bilden sich schwache Verbindung (Wasserstoffbrücken) zwischen den Basen der RNA aus, sodass sie sich zu einer zweidimensionalen Form zusammen faltet. Abbildung 2.1 zeigt die Sekundärstruktur der obigen Beispielsequenz.
Einer Sekundärstruktur lässt sich ein Energiewert (sogenannte „freie Energie“) zuordnen, wel-

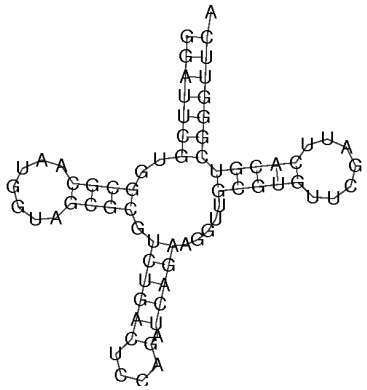

Abbildung 2.1: Sekundärstruktur einer tRNA

cher der Energiedifferenz zwischen der ungefalteten Sequenz und der gefalteten Sequenz entspricht. Die niedrigst mögliche Energie einer Sequenz nennt man „minimale freie Energie“ (*mfe*) dieser Sequenz. Eine niedrige *mfe* ist ein guter Hinweis auf eine hohe Stabilität der Sequenz, da viel Energie aufgewendet werden müsste, um sie wieder zu entfalten.
Außerdem kann bei einer Sequenz auch das Strukturensemble betrachtet werden. Es enthält al-

le Sekundärstrukturen, in die eine Sequenz falten kann. Dem Ensemble lässt sich ebenfalls eine Energie (die „Ensemble Energie“) zuordnen [2]. Diese berechnet sich wie folgt:

$$E_{ensemble}(S) = -RT * ln\left(\sum_{Q \in S} e^{-\frac{E(Q)}{RT}}\right) \tag{2.1}$$

E(Q) ist die freie Energie einer Sekundärstruktur Q, S ist die Menge (das Ensemble) aller möglichen Strukturen, in die sich eine Sequenz falten kann, T ist die Temperatur, R ist die allgemeine Gaskonstante und $E_{ensemble}(S)$ ist die Energie des Ensembles aller Strukturen S.
Das Ensemble und seine Energie werden betrachtet, da sich herausgestellt hat, dass RNA in der Zelle nicht unbedingt in genau der einen energetisch minimalen Struktur vorliegt. Es können auch strukturell deutlich andere Sekundärstrukturen ausgebildet werden, die nur eine gering höhere Energie besitzen als die *mfe*.
Oft wird also nicht nur die ideale Struktur betrachtet, sondern auch Strukturen, die energetisch nicht ganz ideal sind, sich aber deutlich von der Struktur mit minimaler freier Energie unterscheiden.

Außerdem ist eine RNA auch noch in der dritten Dimension gefaltet. Diese Form wird Tertiärstruktur genannt. Bei den meisten Untersuchungen von RNA wird die Tertiärstruktur jedoch nicht betrachtet, sondern nur die Primär- und Sekundärstruktur. Dies hängt damit zusammen, dass die Berechnung der Tertiärstruktur sehr aufwendig ist.

2.3 Programme

Die Programme IntaRNA und LocARNA wurden beide am Lehrstuhl für Bioinformatik der Universität Freiburg entwickelt. Sie sollen helfen, bei neu entdeckten RNA Sequenzen deren Aufgabe in der Zelle einzugrenzen.
LocARNA und IntaRNA betrachten hierbei sowohl Primär- wie Sekundärstruktur der RNA Sequenzen.

2.3.1 IntaRNA

IntaRNA [2] ist ein von Anke Busch, Andreas S. Richter und Professor Rolf Backofen am Lehrstuhl für Bioinformatik der Technischen Fakultät an der Albert-Ludwigs-Universität Freiburg entwickeltes Programm. Es befasst sich mit dem Problem, Interaktionsstellen zwischen siRNA/miRNA/sRNA und mRNA zu finden. siRNA und miRNA treten in Eukaryoten auf, während sRNA in Prokaryoten auftreten [8]. Diese Arbeit konzentriert sich auf sRNA.
sRNAs besitzen die Fähigkeit an bestimmte mRNAs zu binden, um die Translation zu hemmen,

zu ermöglichen oder um die mRNA abzubauen. Je eine sRNA kann nur mit einer geringen Anzahl mRNAs interagieren. Es sind bereits viele verschiedene sRNA Sequenzen bekannt, aber das Wissen, welche sRNAs mit welchen mRNAs interagieren ist noch sehr beschränkt [18, 2]. Daher wurde IntaRNA entwickelt. Es soll helfen, mögliche Bindestellen zwischen den RNA Strängen zu finden.
Die Energie, die als Bewertung der Bindestelle von IntaRNA ausgegeben wird, berechnet sich aus zwei Teilen. Der erste Teil ist die Hybridisierungsenergie der Bindestelle; sie basiert auf dem Energiemodel von RNAhybrid [14]. Als zweites fließt die Energie ein, die nötig ist, um die Bindestelle frei zu legen. Da sowohl mRNA wie sRNA in einer Sekundärstruktur vorliegen können, sind die Bindestellen zwischen den Sequenzen nicht unbedingt direkt erreichbar, sondern können innerhalb der Struktur gebunden sein. Der Energieunterschied zwischen dem Strukturensemble bei dem die Bindestelle frei ist und dem Strukturensemble ohne Einschränkung wird in IntaRNA berechnet und zur Hybridenergie addiert (zur Berechnung der Energie des Ensembles, siehe Gleichung 2.1). Da die Hybridenergie negativ ist, wirkt sich diese Energiedifferenz (ED) abschwächend auf die Güte der Bindestelle aus.

$$E(S_1,S_2) = H(S_1,S_2) + ED(S_1) + ED(S_2) \tag{2.2}$$

Die Energie $E(S_1,S_2)$ der Interaktion zwischen Sequenz S_1 und S_2 besteht aus der Hybridenergie $H(S_1,S_2)$ der Interaktion und den Energiedifferenzen $ED(S_1)$ und $ED(S_2)$ der zwei Sequenzen.
Falls die Energiedifferenzen in der Summe größer als die Hybridenergie sind, wird dieses Sequenzpaar verworfen, daher ist die ausgegebene Energie immer negativ.

IntaRNA unterstützt die Spezifizierung einer Seedregion. Dies ist ein meist zwischen 5 und 8 Basen langer Teil der Bindestelle, bei dem im Normalfall alle Basen gebunden sind. Diese Region wirkt als erster Anker, mit dessen Hilfe die sRNA an die mRNA andockt und von dort wird die Interaktion ausgeweitet bis zur energetisch optimalen Länge [2].

2.3.2 LocARNA

LocARNA [22] wurde von Sebastian Will, Kristin Reiche, Ivo L. Hofacker, Peter Stadler und Rolf Backofen von den Universitäten Freiburg, Wien und Leipzig entwickelt. Es führt ein Alignment zwischen zwei ncRNA Sequenzen durch. Im Gegensatz zu normalen Sequenzalignments wird hier jedoch nicht nur die Sequenz sondern auch die Sekundärstruktur, also die Basenpaarung versucht miteinander in Einklang zu bringen.
Bei bestimmten ncRNA Familien (wie auch bei Proteinen) ist die Sekundärstruktur die funktionsbestimmende Eigenschaft. Dies geht so weit, dass eine ähnliche Struktur ein guter Hinweis auf eine ähnliche Funktion ist. Wie etwa bei der Klasse der tRNA Sequenzen zu sehen ist, in der alle RNA Sequenzen eine kleeblattförmige Sekundärstruktur (ähnlich Abbildung 2.1) aufweisen.

Daher versucht LocARNA durch Einbeziehung der Struktur aussagekräftige Ähnlichkeiten zwischen ncRNA Sequenzen zu finden.

LocARNA vergleicht zwei Sequenzen A und B durch die Aufstellung eines Sequenzalignment **A** in Kombination mit einem Stukturalignment **S**. **S** enthält Basenquadrupel der Form (ij,kl), wobei (i,k) gepaarte Basen in Sequenz A sind, (j,l) gepaarte Basen in Sequenz B sind und i mit j, sowie k mit l in **A** aligniert sind (Vergleich Abbildung 2.2).

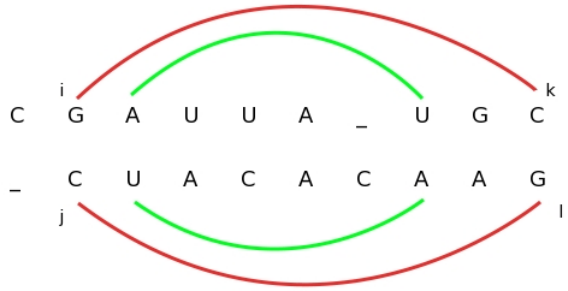

Abbildung 2.2: Ein Sequenz-Struktur-Alignment mit Beispiel Basenquadrupel (ij,kl), wie es bei LocARNA durchgeführt wird

Der Score (die Güte) des Alignments wird entsprechend Formel 2.3 berechnet.

$$\sum_{(ij,kl)\in\mathbf{S}} (\Psi^A_{ik} + \Psi^B_{jl}) + \sum_{(i,j)\in\mathbf{A}_s} (\sigma(A_i, B_j) - N_{gap} \cdot \gamma \tag{2.3}$$

Ψ^A_{ik} hängt direkt von der Paarungswahrscheinlichkeit der Basen i,k in Sequenz A ab. $\mathbf{A}_s$ enthält alle alignierten Basen des Alignments **A**, die in keinem Quadrupel von **S** vertreten sind. $\sigma(i,j)$ ist eine Bewertung der Übereinstimmung der Basen i und j. γ ist die Anzahl an Gaps im Alignment und N_{gap} ein Faktor, mit dem die Gap Anzahl bestraft wird.
Die Formel kann im Detail in [22] nachgelesen werden.
Je größer der Score ist, desto ähnlicher sind sich die Sequenzen.

Mit LocARNA ist es möglich große Mengen von ncRNA Sequenzen in Cluster zu unterteilen. Je ähnlicher zwei Sequenzen sind, desto näher zusammen hängen sie im Cluster. Idealerweise lassen sich aus dem Cluster dann Teilbäume von funktionsgleichen RNA Sequenzen ablesen. In dieser Arbeit wird LocARNA für den Vergleich von jeweils nur zwei Sequenzen benutzt und untersucht, wie man die Signifikanz dieser Paarung feststellen kann.

2.4 Thema

Momentan geben diese Programme nicht-normierte Scores bzw. Energiewerte aus. Ist man an der Signifikanz des Alignments oder der Interaktion interessiert, sind diese nicht ausreichend. Dafür benötigt man noch eine große Vergleichsprobe, die mit Hilfe zufälliger Sequenzen mit gleichen Eigenschaften (meist die permutierten Eingabesequenzen) erzeugt wurde. Durch die Einordnung des eigentlichen Scores bzw. der Energie in Verhältnis zur Vergleichsprobe kann die Signifikanz bestimmt werden. Je nach Größe der Vergleichsprobe kann dieser Schritt sehr rechenintensiv werden.
Das Ziel ist, diesen Schritt zu umgehen, indem der Effekt der einzelnen Eigenschaften der Sequenzen auf Score und Energie analysiert wird und Regeln, die diesen Effekt beschreiben aufgestellt werden. Mit deren Hilfe wäre es möglich, ohne eine Vergleichsprobe zu erzeugen, zu sagen, wie die Vergleichsprobe aussieht. So könnte am Ende aus dem Score/der Energie und den Eingabesequenzen direkt ein P-Wert errechnet werden. Ein P-Wert gibt an, wie wahrscheinlich ein entsprechend hoher Wert für genau diese Merkmalskombination der Sequenzen ist. Dies liefert eine Antwort darauf, ob bei dieser Paarung die Interaktion bzw. die Ähnlichkeit signifikant ist.

Im Lauf der Arbeit wurde versucht, Zusammenhänge zwischen dem Score/der Energie und den feststellbaren Eigenschaften der Eingabesequenzen zu erkennen.
Als Eigenschaften wurden von jeder Sequenz die Länge, das AU zu GC Verhältnis (repräsentiert durch den prozentualen Anteil der A- und U-Basen an der Gesamtsequenz) sowie die „minimale freie Energie“ festgehalten. Bei IntaRNA wurden außerdem mögliche Zusammenhänge zwischen der Energie und der Länge der Interaktionsstelle betrachtet.

2.5 Arbeiten zu verwandten Themen

Viele der die in dieser Arbeit benutzten Methoden basieren auf bereits existierenden Arbeiten, die sich mit verwandten Problemen beschäftigt haben.
In [14] von M.Rehmsmeier *et al.* wurden miRNA/mRNA-Duplexe untersucht. Die Verteilung der Duplexenergie bei gleichen Eigenschaften der Testsequenzen konnte dort als Extremwertverteilung identifiziert werden. Es wurde auch versucht, den Effekt der Sequenzlänge auf die Duplexenergie zu entfernen.
Ähnliches wurde in dieser Arbeit getan, allerdings nicht bei miRNA/mRNA-Duplexen, sondern bei Interaktionen zwischen sRNA und mRNA. Die Verteilung der Interaktionsenergie gleicht ebenfalls einer Extremwertverteilung. Zusätzlich wurde nicht nur in Bezug auf die Länge der Sequenzen versucht zu normalisieren, sondern auch in Bezug auf das AU zu GC Verhältnis.

Die Idee zur Nutzung einer Support-Vektor-Maschine(SVM) basiert auf „fast and reliable pre-

diction of noncoding RNAs“ von Stefan Washietl *et al.* [19]. Dort wurde eine SVM benutzt um funktionale RNA zu finden. Der darin diskutierte Ansatz ist, dass funktionale RNA im Allgemeinen sehr stabil ist, womit man aus der minimalen freien Energie schließen kann, ob eine RNA Sequenz funktional ist oder nicht. Dafür wurden zufällig geshuffelte Sequenzen verschiedener Länge und verschiedenem Verhältnis zwischen AU und GC erzeugt. Je Eigenschaftspaar aus Länge und Verhältnis wurden zehntausend Sequenzen genommen, deren minimale freie Energie berechnet und die definierenden Eigenschaften (hier Mittelwert und Standardabweichung) der entstehenden Verteilungen extrahiert. Eine Support Vektor Maschine wurde benutzt um die Zusammenhänge zwischen den entstandenen Verteilungen und den Sequenzeigenschaften zu lernen. Eine Untersuchung einer RNA Sequenz mit Hilfe dieser SVM ist dann in der Lage zu beurteilen, ob die freie Energie der Sequenz außerordentlich niedrig ist, was darauf hindeuten würde, dass die Sequenz funktional ist.

2.6 Überblick

Im ersten Kapitel wurde auf den Hintergrund der Bioinformatik eingegangen und erläutert was bereits zur Thematik dieser Arbeit publiziert wurde. Das folgende Kapitel gibt einen kurzen Überblick über die benutzten Methoden.
Im dritten Kapitel werden die Ergebnisse vorgestellt, die im vierten Kapitel diskutiert werden. Daran schließt sich der Ausblick an, sowie ein Überblick über die benutzten Programme, die verwendete Hardware und das Abbildungs-, Tabellen- und Literaturverzeichnis.

3 Methoden

In diesem Kapitel werden die Methoden vorgestellt, die in dieser Arbeit Anwendung fanden. Außerdem werden für diese Arbeit wichtige Begriffe erläutert.

3.1 Verteilungen

In dieser Arbeit wurde mit Stichproben gearbeitet, die große Ähnlichkeit mit Extremwertverteilungen oder Normalverteilungen aufweisen.

3.1.1 Extremwertverteilung nach Gumbel

Eine Extremwertverteilung ist eine stetige Wahrscheinlichkeitsverteilung. Sie ist ein Mittel zur Beschreibung von extremen Ereignissen, wie etwa der Wahrscheinlichkeit von starken Überschwemmungen bei Flüssen. Eine Gumbel-Extremwertverteilung [12] lässt sich durch zwei Parameter μ und β beschreiben (siehe Formel 3.1 und 3.2). Wobei μ den Ort (Location) des Maximums in der Dichtefunktion angibt, während β die Breite (Scale) der Verteilung festlegt. In Abbildung 3.1 sind Dichteverteilungen bei unterschiedlichem μ und β abgetragen.

$$\text{Verteilungsfunktion:} f(x) = e^{-e^{\frac{\mu - x}{\beta}}} \tag{3.1}$$

$$\text{Dichtefunktion:} f(x) = \frac{e^{\frac{x-\mu}{\beta}} \cdot e^{e^{\frac{x-\mu}{\beta}}}}{\beta} \tag{3.2}$$

Log-Log Test auf Extremwertverteilung

Der Log-log Test [14] macht sich zunutze, dass die Gumbel-Extremwertverteilung durch eine doppelte e-Funktion dargestellt werden kann (Formel 3.1). Wenn man also den Logarithmus des Logarithmus der Funktion nimmt, erhält man eine Gerade. Falls die Verteilung keine Gumbel Extremwertverteilung ist, kommt dagegen keine Gerade heraus.
Dieser Zusammenhang wurde auch benutzt, um die Parameter der Verteilung zu finden. Es

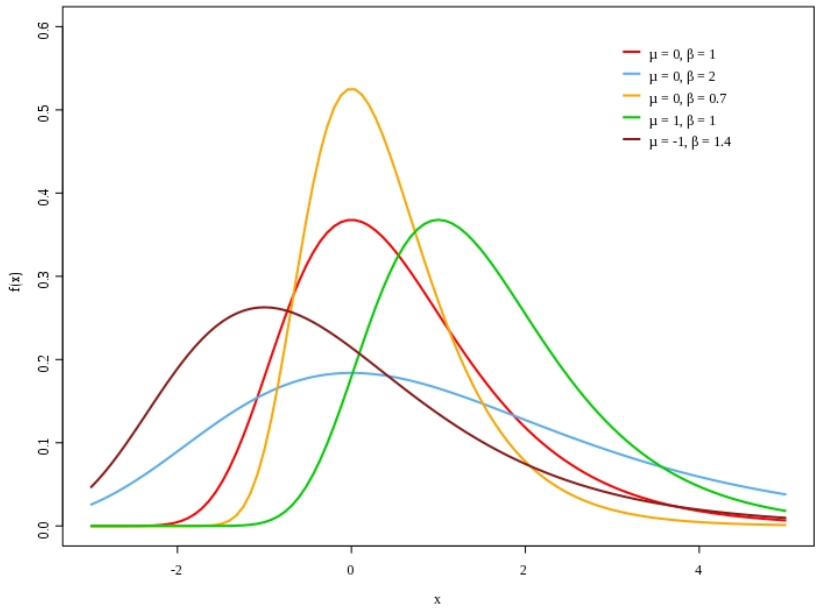

Abbildung 3.1: Dichtefunktion der Gumbel-Extremwertverteilungen bei verschiedenem μ und β [20]

wurde eine Regression der „Log-log“-Geraden gemacht und daraus errechnet, welche Parameter die Gumbel-Extremwertverteilung besitzt.

3.1.2 Normalverteilung

Eine Normalverteilung [11] oder auch Gauß'sche Glockenkurve genannt, ist eine oft beobachtete stetige Verteilung. Sie hat die Schiefe 0 und ist damit symmetrisch. Die Standardnormalverteilung ist eine besondere Art der Normalverteilung. Sie hat die Eigenschaft, dass die Fläche unter der Kurve 1 ergibt und dass ihr Mittelwert 0 ist. Eine Normalverteilung ist durch die zwei Parameter, Mittelwert μ und Standardabweichung σ, vollständig beschreibbar (siehe Formel 3.3). Falls man von einer Datenmenge weiß, dass sie normalverteilt ist, kann man die Kurve aus den Daten einfach dadurch gewinnen, dass man den Mittelwert findet, um den die Daten streuen und die Standardabweichung berechnet. Beides ist in der Regel schnell berechenbar.

$$\text{Dichtefunktion: } f(x) = \frac{1}{\sigma \cdot \sqrt{2\pi}} \cdot e^{-\frac{(x-\mu)^2}{2 \cdot \sigma^2}} \tag{3.3}$$

Kolmogorow-Smirnow-Test

Der Kolmogorow-Smirnow-Test [21] wurde von Andrej Niklajewitsch Kolmogorow und Wladimir Iwanowitsch Smirnow entwickelt und ist ein statistischer Test zur Prüfung, ob zwei Stichproben durch die selbe Wahrscheinlichkeitsverteilung entstanden sein können. Entweder werden zwei Stichproben angegeben, oder eine Stichprobe wird gegen eine angenommene Verteilungsfunktion getestet. Bewertet wird der Punkt mit der größten Abweichung zwischen der Stichprobe und der Verteilung. Ist diese Abweichung zu groß, wird die Hypothese verworfen (Mögliche Grenzwerte siehe z.b. [10]).
Der KS-Test ist nicht auf die Normalverteilungen beschränkt, sondern kann genutzt werden um gegen beliebige Verteilungen zu testen. In dieser Arbeit wurde er nur für den Test auf Normalverteilung benutzt.

Quantil-Quantil Plot

Ein QQ-Plot [16] ist eine graphische Methode, die Übereinstimmung zwischen einer Stichprobe und einer Normalverteilung zu bewerten. Dabei wird die Stichprobe aufsteigend sortiert und auf der y-Achse abgetragen, während auf der x-Achse eine Standardnormalverteilung abgetragen wird.
Falls die Stichprobe normalverteilt ist, bildet sich eine exakte Diagonale in der Graphik (Vergleich Abbildung 4.2).

3.2 P-Wert, Signifikanz und Nullmodell

P-Wert und Nullmodell sind Begriffe, die bei statistischen Tests verwendet werden. Ein statistischer Test prüft die Gültigkeit einer Hypothese anhand von Beobachtungen. Meist wird gegen die Annahme getestet, dass eine Beobachtung durch Zufall (die Nullhypothese) statt durch die angenommene Hypothese entstanden ist. Ein einfacher Test stellt daher ein sogenanntes Nullmodell auf, das die durch Zufall entstandene Wahrscheinlichkeitsverteilung widerspiegelt. Basierend auf diesem Modell wird analysiert, wo die Beobachtung relativ zur Verteilung liegt. Dazu wird der P-Wert (**P**robability-Value) aufgestellt, der die Wahrscheinlichkeit liefert, dass die Beobachtung durch Zufall entstanden ist:

$$PW(b) = P(Y > b|H_0) \tag{3.4}$$

$PW(b)$ ist der P-Wert der Beobachtung b, der sich aus der Wahrscheinlichkeit bildet, dass ein beliebiges $Y \in H_0$ größer ist als b, wenn das Nullmodell H_0 gilt. Dies ist der „linke“ P-Wert, für die Annahme, dass nicht-zufällige Beobachtungen größer als zufällige Beobachtungen sind. Der

rechte P-Wert dagegen nimmt an, dass nicht-zufällige Beobachtungen kleiner sind, und spiegelt dementsprechend die Wahrscheinlichkeit wider, dass ein beliebiges Y kleiner als b ist.

Der P-Wert für einen Vergleich oder eine Interaktion gibt an, wie wahrscheinlich es ist, dass ein so hoher oder höherer Score (eine so niedrige oder niedrigere Energie) mit den Eigenschaften der Ausgangssequenzen durch Zufall entsteht. Um das adäquat angeben zu können wird eine oder beide Sequenzen zufällig permutiert und nochmal der Score (die Energie) berechnet. Nach einigen Wiederholungen (etwa 1000-10000) ist es möglich, die Häufigkeit des eigentlichen Scores (und besserer) in der erzeugten Verteilung anzugeben, womit sich eine Wahrscheinlichkeit schätzen lässt. Auf diese Weise lässt sich ein empirischer P-Wert berechnen:

$$PW_{emp}(b) = \frac{\#S(n) \geq S(b)}{\#S(n)} \tag{3.5}$$

Der empirische P-Wert $PW_{emp}(b)$ einer Beobachtung b ist die Anzahl Scores aus permutierten Sequenzen $S(n)$, die größer als der Score $S(b)$ von b sind, geteilt durch die Anzahl aller Scores aus permutierten Sequenzen $S(n)$.

Ein Wert wird als signifikant bezeichnet, wenn es unwahrscheinlich ist, dass er durch Zufall entstanden ist. Auf den P-Wert bezogen heißt das, dass eine Interaktion oder ein Vergleich signifikant ist, wenn der P-Wert sehr niedrig ist, zum Beispiel kleiner als 0.01.
Falls der Score eines Vergleichs signifikant ist, heißt das, es ist wahrscheinlich kein Zufall. Im Fall eines Vergleichs kann man annehmen, dass diese Sequenzen wirklich sehr ähnlich sind, so ähnlich, dass sie nahe verwandt sind, oder (fast) die gleiche Aufgabe erfüllen (zum Beispiel in verschiedenen Spezies).
Falls die Anzahl an Scores aus zufälligen Sequenzen groß genug ist, lässt sich ein Modell (das Nullmodell) der Wahrscheinlichkeitsverteilung aus der Stichprobe approximieren. Mit diesem Modell lässt sich der eigentliche, nicht empirische, P-Wert berechnen.

In dieser Arbeit wird versucht, ein Nullmodell für ein Sequenzpaar, das verglichen wird, oder bei dem eine Interaktion gesucht wird, zu finden, um den P-Wert der Interaktion oder des Vergleichs zu erhalten; allerdings ohne zuerst sehr viele zufällige Sequenzen erzeugen zu müssen. Dafür wird nach Regeln gesucht, die in Abhängigkeit von den Eigenschaften der Sequenzen einen Schluss auf das Nullmodell zulassen.

3.3 Zufallssequenzen

Für die Analysen wurden pseudo-zufällige Sequenzen erzeugt, bei denen die Länge und das Verhältnis zwischen A, U, G und C festgelegt wurde. Dies wird dadurch erreicht, dass eine künstliche Sequenz erzeugt wird, die genau den Vorgaben entspricht. Dann werden so viele Permutatio-

nen davon erzeugt, wie zufällige Sequenzen benötigt werden. Dazu wird das Programm Shuffle benutzt (siehe Kapitel 7.1).

3.4 Regression

Eine Regression ist ein Verfahren um eine Beziehung zwischen einer abhängigen Variable (genannt Zielvariable) und einer unabhängigen Variable zu finden. Es wird zum Beispiel benutzt, um zu einer Punktmenge, von der bekannt ist, dass sie etwa auf einer Linie verteilt ist, eine Gerade zu finden, die den Punkten möglichst nahe liegt. Die verschiedenen Methoden der Regression unterscheiden sich darin, wie sie mit Ausreißern umgehen. Hier muss unterschieden werden, ob man möchte, dass alle Punkte möglichst gut beschrieben werden, was bei einem starken Ausreißer eine deutliche Verschiebung der Regressionsgeraden bewirkt. Oder ob man zum Beispiel bereits aus der Erzeugung der Daten weiß, dass es zu seltenen aber extremen Messfehlern kommt. In diesem Fall braucht man eine Gerade, welche die meisten Punkte gut beschreibt, auf die extreme Ausreißer aber nur einen sehr geringen Effekt haben.

3.4.1 Methode der kleinsten Quadrate

In dieser Arbeit wurde bei allen Regressionen die Methode der kleinsten Quadrate (least squares) benutzt. Sie ist eine weit verbreitete Regressionsmethode, die sich dadurch definiert, dass Fehler mit dem Quadrat der Abweichung bestraft werden. Es wird also von nur geringen Messfehlern ohne extreme Ausreißer ausgegangen. Es wird das Minimum folgender Formel gesucht:

$$\min_{\vec{x}} \left(\sum_{i=1}^{n} (y_m - y_i)^2 \right) \tag{3.6}$$

Wobei die y_i die abhängige Variable, x die unabhängige Variable, und y_m der per Regression geschätzte Wert von y_i ist.

3.5 Support-Vektor-Maschinen

Eine Support-Vektor-Maschine (SVM) ist ein Verfahren des maschinellen Lernens[3, 5]. Sie kann zur Klassifikation oder zur Regression benutzt werden. Eine SVM wird zuerst mit einem Datensatz trainiert, bei dem die Klassenzugehörigkeiten beziehungsweise die Zielvariable bereits bekannt sind. Dabei werden die Regeln, nach denen sich die Datenmenge aufbaut, gelernt. Später kann die SVM mit diesen Regeln bei neuen Daten die unbekannte Klassenzugehörigkeit oder die unbekannte Zielvariable bestimmen.

Die Trainingsdaten werden von der SVM in einen höherdimensionalen Raum abgebildet, um dann nach einer Hyperebene zu suchen. Diese Ebene soll bei einer Klassifikation die Punkte möglichst gut in die vorgegebenen Klassen unterteilen. Bei einer Regression wird eine Hyperebene gesucht, die möglichst gut durch die Trainingspunkte verläuft und also vorhersagen kann, welche Zielvariable zu welchen Eingabevariablen passt.
Eine als Klassifikator benutzte SVM hat die Einschränkung, dass sie nur mit binären Klassenzugehörigkeiten umgehen kann, also nur zwei Klassen trennen kann. Falls mehr als zwei Klassen vorliegen, wird meist für jedes mögliche Klassenpaar eine SVM aufgerufen und aus den entstandenen Hyperebenen eine alle Klassen beschreibende Ebene zusammengestellt.
Verwendung finden Klassifikator-SVMs unter anderem in der Bildverarbeitung. Dort werden sie zum Beispiel benutzt, um mit Hilfe einer Datenbank mit Bildern von bekannten Objekten die SVM zu trainieren, so dass auch auf fremden Bildern diese Objekte erkannt werden.

4 Versuche und Ergebnisse

In diesem Kapitel werden die Versuche, und deren Ergebnisse dargestellt.

4.1 LocARNA

Es wurde LocARNA Version 1.3.5 benutzt und alle Einstellungen auf den Standardwerten belassen.

Als erstes sollte herausgefunden werden, welche Form die Scoreverteilung einnimmt, wenn man Länge und Basenverhältnis konstant hält. Hierfür wurden viele gleichlange Sequenzen mit gleicher Basenfrequenz je paarweise verglichen. Nur wenn die Scores eine eindeutige Verteilung einnehmen, lässt sich in einem späteren Schritt mit ihnen ein Nullmodell aufstellen und damit ein P-Wert als Maß der Signifikanz errechnen.
Für eine möglichst stetige Verteilung wurden in diesem Schritt je zehntausend Scores pro ausgesuchter Merkmalskombination erzeugt. Dies entspricht zwanzigtausend RNA Sequenzen, da keine Sequenz doppelt benutzt wurde. Damit ist sichergestellt, dass die einzelnen Scores unabhängig voneinander sind.
Um den Einfluss der Merkmalsparameter auf die Verteilung zu erkennen, wurde für folgende Kombinationen aus Länge und AU-Anteil je eine Verteilung von zehntausend Scores erzeugt:

- Länge 40nt(Nukleotide), 80nt, 120nt, 160nt, 200nt, 240nt, 280nt, 320nt, 360nt und 400nt bei AU-Anteil von 50%.
- Länge 160nt bei Variation des AU-Anteils zwischen 25% und 75% in Schritten von 5%.

Es wurden nur gleichlange Sequenzen miteinander verglichen. Der AU-Anteil wurde dagegen unabhängig voneinander in beiden Sequenzen variiert. Insgesamt liegen damit 10 Verteilungen zur Untersuchung der Länge und 121 Verteilungen zur Untersuchung des AU zu GC Verhältnisses vor.

4.1.1 Verteilungsuntersuchung

Bei allen Analysen hat die Verteilung der Scores große Ähnlichkeit mit einer Normalverteilung (siehe Abbildung 4.1). Ein Test mit Quantil-Quantil Plots zeigt eine gute Gerade (siehe Abbildun-

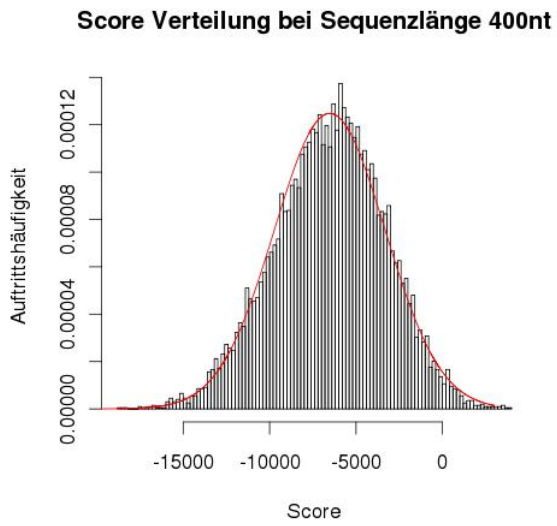

Abbildung 4.1: Histogramm von 10000 Scores gegen eine Normalverteilung (rot)

gen 4.2). Im weiteren wird daher von einer Normalverteilung ausgegangen, so dass sie durch einen Erwartungswert (Mittelwert der Verteilung) und eine Standardabweichung vollständig beschrieben werden kann.
Es wurde auch der Kolmogorow-Smirnow-Test auf den Daten durchgeführt (Tabelle 4.1). Zu beachten ist hier, dass der KS-Test, abgesehen von der zu testenden Stichprobe, auch den Mittelwert und die Standardabweichung der Vergleichsverteilung als Eingabe benötigt. Der KS Test schreibt vor, dass die Parameter nicht aus der Stichprobe gezogen werden, sondern unabhängig davon sind.
Da aber keine unabhängige Vergleichsprobe vorhanden ist, wurden die Parameter trotzdem aus der Stichprobe ermittelt. Dies hat zur Folge, dass die P-Werte kritisch betrachtet werden müssen. Die P-Werte sind recht unterschiedlich, der größte Teil liegt aber im ein- und zweistelligen Prozentbereich, somit wird weiter von einer Normalverteilung ausgegangen.

4.1.2 Variation der Länge

Aus den erzeugten Verteilungen bei verschiedenen Längen lassen sich bereits Schlüsse auf den Zusammenhang zwischen Länge und dem zu erwartenden Score machen.
Es ist zu erkennen, dass bei den bisher erzeugten Daten, die ein ausgeglichenes AU zu GC Verhältnis haben, mit steigender Länge der Score sinkt und die Streuung zunimmt (siehe Abbildung 4.3).

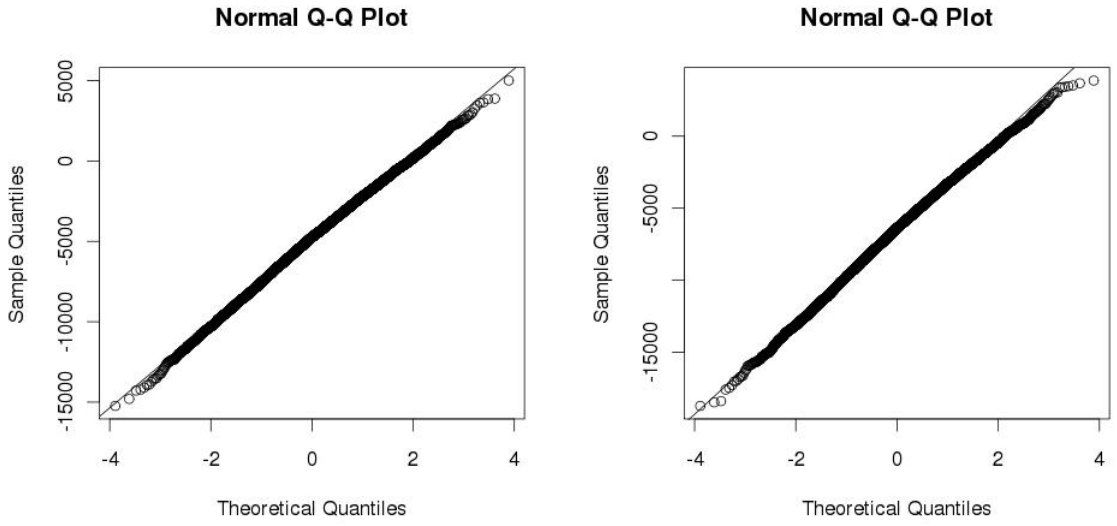

Abbildung 4.2: die Quantil-Quantil Plots für Länge 280 und 400

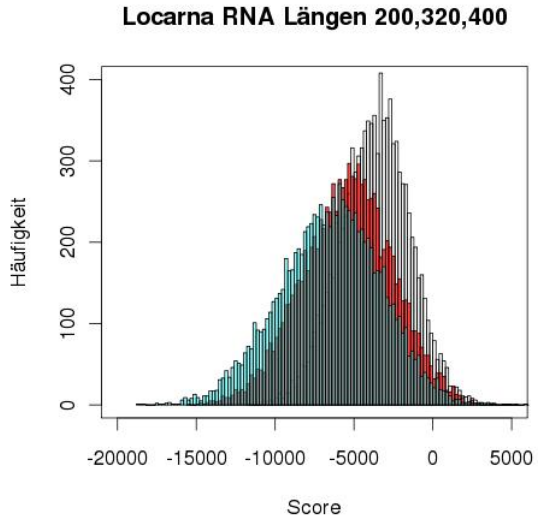

Abbildung 4.3: Verteilungen bei verschiedenen Längen: Weiß: 200nt, Rot: 320nt, Blau: 400nt

Länge	% AU Sequenz 1	% AU Sequenz 2	P-Wert KS-Test
40	50	50	0.072349857592287
80	50	50	0.445699025180607
120	50	50	0.242174053595498
160	50	50	0.0182206265241274
200	50	50	0.0151472722199515
240	50	50	0.0676771536417728
280	50	50	0.00252088454118615
320	50	50	0.0627066443475985
360	50	50	0.00525053040279577
400	50	50	0.00511934474215903
160	35	35	0.148345523625484
160	40	35	0.433152728405506
160	45	35	0.208966714835374
160	50	35	0.756407843628611
160	55	35	0.0582532316913031
160	60	35	0.0558045964732868
160	65	35	0.0924318059629353
160	70	35	0.0218098157353803
160	75	35	0.0077649167197531
160	45	40	0.27985898882795
160	50	40	0.39422339993509
160	55	40	0.117868243231518
160	65	40	0.0268069189090214
160	70	40	0.0427729419744578
160	75	40	0.00178271952233500
160	45	45	0.0994947168981077
160	60	45	0.0709847295185381
160	65	45	0.0654659897470311
160	75	45	0.0140986332375779
160	50	50	0.214818085077511
160	55	50	0.046915522266683
160	60	50	0.0981656779501656
160	65	50	0.249841000649402
160	60	55	0.187726873726165
160	65	55	0.0585121320244996
160	70	55	0.112324235237953
160	65	60	0.0158644561525767
160	70	60	0.158394588714055
160	75	60	0.0411487304404953
160	70	65	0.211363084485297
160	75	65	0.0616676681631476

Tabelle 4.1: Ausschnitt der P-Werte des KS Tests bei verschiedenen Scoreverteilungen

4.1.3 Einfluss der minimalen freien Energie

Bei der Untersuchung von möglichen Zusammenhängen zwischen dem Score und der *mfe* muss darauf geachtet werden, dass die minimale freie Energie selbst wiederum von Faktoren, wie der Länge oder dem AU zu GC Verhältnis beeinflusst wird. Daher müssen für diese Analysen Scores benutzt werden, deren zugrunde liegende Sequenzen sich in Bezug auf Länge und AU-Anteil gleichen. Nur so kann ein der *mfe* zugehöriger Zusammenhang erkannt werden.
Eine Untersuchung der Verteilungen daraufhin, ob die *mfe* einen Einfluss auf den Score hat, zeigte keinerlei Abhängigkeiten (siehe Abbildung 4.4). Die minimale freie Energie ist also keine Eigenschaft, die einen Einfluss auf den Score hat.

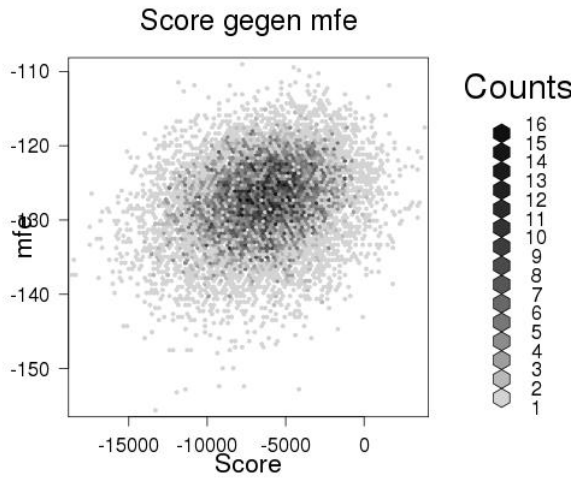

Abbildung 4.4: Dichteplot: Minimale freie Energie gegen Score. Sequenzlänge 400nt

4.1.4 Variation der Länge in größerem Umfang

Zur exakten Analyse des Zusammenhangs von Länge zum berechneten Score wurden weitere Verteilungen erzeugt. Jede Verteilung besteht aus zehntausend Scores, die mit Hilfe von zwanzigtausend Sequenzen gleicher Länge und gleichem AU zu GC Verhältnis erzeugt wurden. Von jeder Verteilung wird der Erwartungswert und die Standardabweichung betrachtet.
Zur Untersuchung des Längenzusammenhangs wurden Verteilungen für Länge 600, 800 sowie weitere im Bereich zwischen 20 und 400 erzeugt. Bei allen Verteilungen war das AU zu GC Verhältnis ausgeglichen. Mit diesen Daten, die nun eine gute Abdeckung des Bereichs zwischen 20 und 800 Basen Länge darstellen, lässt sich erkennen, dass der Erwartungswert $E(\ell)$ der Scores mit steigender Länge ℓ linear abnimmt (Abbildung 4.5). Eine lineare least-square Regression

wurde durchgeführt und folgende Formel konnte abgeschätzt werden:

$$E(\ell) = -14.3 \cdot \ell - 709$$

Die Standardabweichung σ scheint mit der Wurzel der Länge zu wachsen (Abbildung 4.5). Mit Hilfe einer least-square Regression an die Wurzelfunktion ($y = b \cdot \sqrt{x} + d$) konnte der folgende Zusammenhang zwischen Score und Länge ermittelt werden:

$$\sigma = 169 \cdot \sqrt{\ell} - 162$$

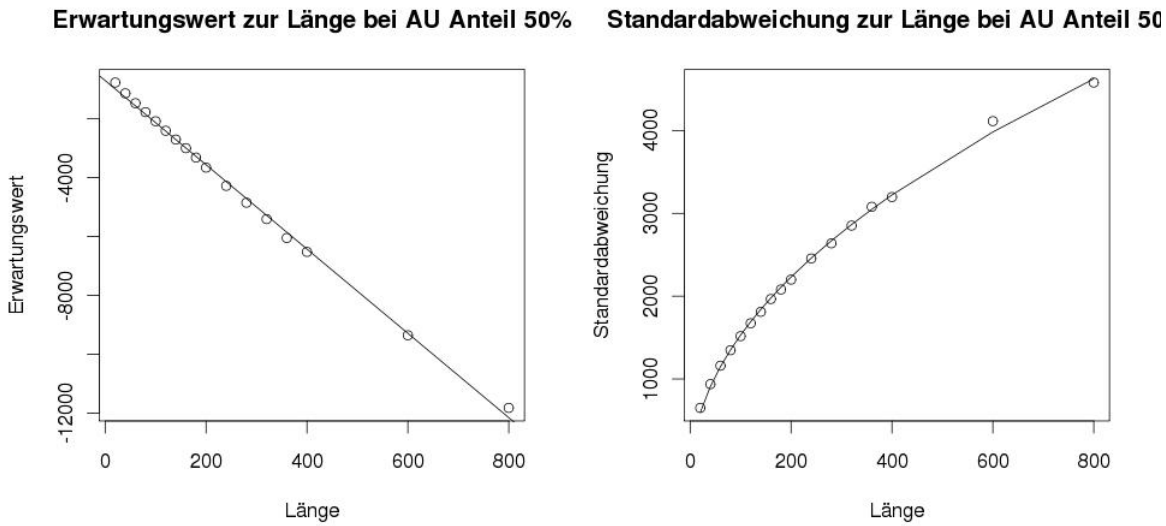

Abbildung 4.5: Linearer und wurzelförmiger Zusammenhang zur Länge

4.1.5 Variation des AU-Anteils

Die Analyse auf den bisherigen Daten zum Zusammenhang zwischen Erwartungswert bzw. Standardabweichung und der Variation des AU zu GC Verhältnisses ist weniger eindeutig. Es lässt sich allerdings ein stetiger Zusammenhang erkennen (Abbildung 4.6). Ebenfalls zu erkennen ist, dass die entstandenen Ebenen symmetrisch sind. Somit ist die Verteilung identisch, unabhängig davon, ob Sequenz1 AU-Anteil x hat und Sequenz2 Anteil y, oder umgekehrt Sequenz1 y und Sequenz2 x. Die Art, wie LocARNA vorgeht, ließ dies bereits im Vorfeld vermuten, hier sei es trotzdem noch einmal ausdrücklich erwähnt.

Zur Verifikation der bisherigen Erkenntnisse und Untersuchung des Zusammenhangs der Verteilung mit dem AU-Anteil wurden weitere Verteilungen erzeugt, siehe Tabelle 4.2.

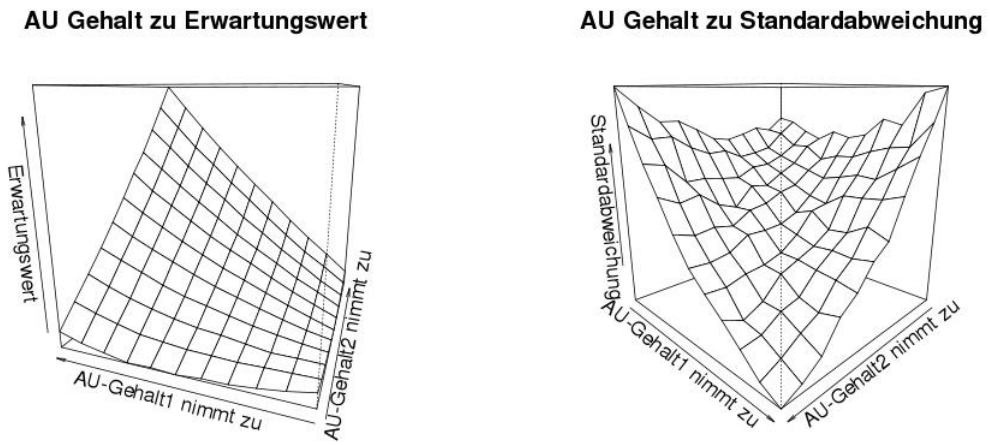

Abbildung 4.6: Verhalten des Erwartungswerts und der Standardabweichung bei Änderung des AU-Anteils in den Sequenzen

Länge in nt	AU-Anteil Seq1	AU-Anteil Seq2	Anzahl Verteilungen
20 bis 600	25%	25%	10
20 bis 600	75%	75%	10
80	25% bis 75%	25% bis 75%	121
100	25% bis 75%	25% bis 75%	121
200	25% bis 75%	25% bis 75%	83

Tabelle 4.2: Die Erzeugung der Verteilungen für Variation des AU-Anteils bei 200nt Länge dauerte länger als erwartet (> eine Woche), daher wurde nur etwas mehr als die Hälfte des 11 mal 11 Verteilungen großen Felds berechnet. Die fehlende Hälfte ist allerdings, wie bereits festgestellt, mit der berechneten Hälfte deckungsgleich, womit hier nur ein geringer Genauigkeitsverlust vorliegt.

Dieses nun deutlich größere Datenfeld ermöglicht es, weitergehende Erkenntnisse zu erlangen. So ist durch die ersten zwei weiteren Scoreanalysen klar, dass die Verschiebung des Scores ins negative bei steigender Länge nur bei einem AU-Anteil von genau 50% den obigen Formeln entspricht. Wenn aber der Anteil ein anderer ist, ändert sich damit auch der Zusammenhang zwischen Länge und Erwartungswert (siehe Abbildung 4.7). So sinkt bei einem AU-Anteil von 75% nicht etwa der Erwartungswert mit der Länge, sondern er steigt. Auch die Standardabweichung verhält sich etwas anders, je nach Verhältnis der Basen, wobei hier der Unterschied nicht so groß ist wie beim Erwartungswert.

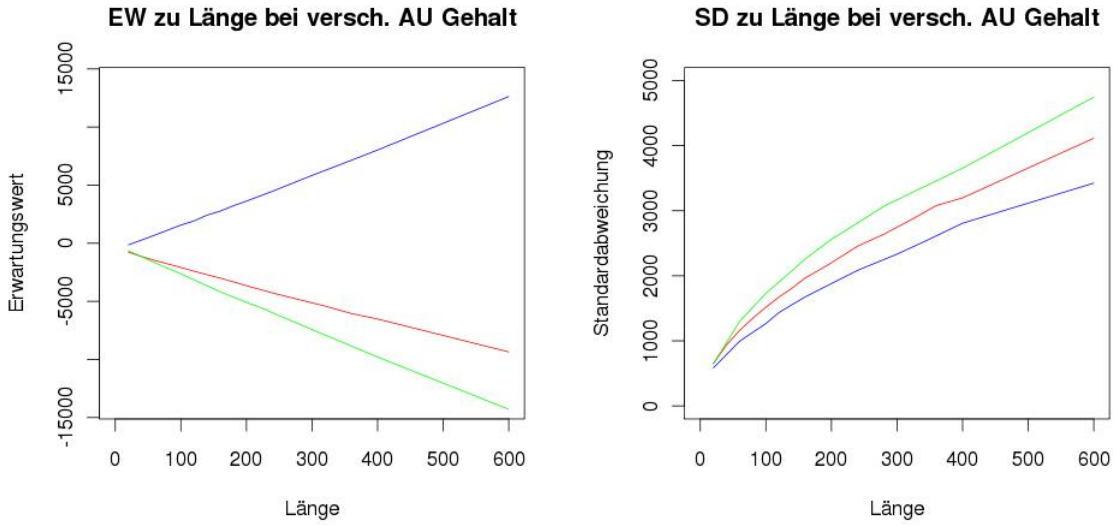

Abbildung 4.7: Erwartungswert und Standardabweichung bei verschiedenen Längen und AU-Anteilen, Grün: 25% AU-Anteil, Rot: 50% AU-Anteil, Blau: 75% AU-Anteil

4.1.6 SVM

Da keine der die Verteilung beeinflussenden Eigenschaften direkt herausrechenbar ist, wird nun versucht Support-Vektor-Maschinen mit den Daten zu trainieren, damit sie die Zusammenhänge erkennen und Vorhersagen zur Verteilung bei anderen Eigenschaftskombinationen machen können. Insgesamt wurden zwei SVMs trainiert, eine für den Erwartungswert und eine für die Standardabweichung. Die Parameter und das Vorgehen waren bei beiden SVMs identisch.
Als SVM Implementierung wird LibSVM benutzt [3].
Es wurden folgende Parameter beim Training der SVM benutzt:

- Typ: ν-SVR
- Kernel: radial basis function (rbf)

Feineinstellungen an den Parametern ν, γ und ε der SVM haben nur sehr geringe Veränderungen gezeigt, daher wurden sie nach einigen Tests auf den Standardwerten belassen ($\nu = 0.1$, $\gamma = 0.3333$, $\varepsilon = 0.001$). Eine Ausnahme sind die Kosten für Fehlregression, da eine leichte Erhöhung hier die Qualität der Regression verbessern konnte. Sie wurden auf 3 gesetzt, Standard ist 1.

Die Eingabedaten wurden auf einen Bereich zwischen -1 und 1 normalisiert:

Parameter	-1	0	1
AU-Anteil	25%	50%	75%
Sequenzlänge	-	0nt	600nt
Standardabweichung	-	0	4000
Erwartungswert	-10000	0	10000

Auf eine exakte Anpassung aller Parameter auf den Bereich zwischen -1 und 1 wurde verzichtet, um saubere Umrechnungen zu gewährleisten.

Abweichungen der Regression sind bei sehr langen Sequenzen (länger als 300nt) zu beobachten. Da bisher nur sehr wenige Verteilungen mit so großer Länge erzeugt wurden, scheinen diese von der SVM als tolerierbare Fehler angesehen zu werden, womit die Hyperebene nicht sehr gut durch diese Punkte verläuft.
Ein größerer Datenpool würde hier Abhilfe schaffen. Bei der momentanen Geschwindigkeit von LocARNA würde die Berechnung einer ausreichend großen Menge an Punkten allerdings die zeitlichen Grenzen der Arbeit sprengen (siehe Laufzeiten in Kapitel 7.2). Um trotzdem ausreichend Daten zur Verfügung stellen zu können, wird auf das bereits festgestellte Verhältnis zwischen Länge und Score aufgebaut: So ist durch die vorherigen Ergebnisse mit LocARNA klar, dass der Erwartungswert linear und die Standardabweichung in Form einer Wurzelfunktion mit der Länge variieren, wenn der AU zu GC Anteil konstant gehalten wird. Mit ausreichend Punkten gleichen AU-Anteils kann die Geradenfunktion des Erwartungswerts und die Kurve der Standardabweichung für diesen speziellen Fall per Regression ermittelt werden und so ist es möglich bis zu einem gewissen Grad die Parameter der Verteilungen vorherzusagen, die jenseits des berechneten Bereichs liegen. Eine Durchführung dieses Vorgehens für alle 55 betrachteten AU-Anteile (je Sequenzen von 25% bis 75% in in 5% Schritten, 11*10/2 = 55 mögliche Kombinationen) ermöglicht es eine große Menge an Punkten zu extrapolieren auch in Bereichen, in denen bisher die Daten gefehlt haben.
Allerdings wird eine Regression über die schnell zu berechnenden kurzen Sequenzen mit steigender Entfernung von den zugrunde liegenden Verteilungen immer weniger exakt, daher ist dies kein Weg um für beliebige Längen Werte zu finden. Da nur Daten bis zur Länge von 200 Basen für sämtliche AU-Anteil Kombinationen berechnet wurden, ist nicht davon auszugehen, dass jenseits von 400, maximal 600, Basen noch eine ausreichend hohe Trefferquote gewahrt ist.
Bis zu dieser Länge allerdings sollte die auf diesen Daten trainierte SVM mit nur sehr geringem

Fehler den Erwartungswert und die Standardabweichung vorhersagen können.

Es wurde ein Perlskript für die Ausgabe des P-Werts geschrieben. Als Eingabe werden zwei Sequenzen und der Score des Alignments erwartet. Das Skript ermittelt die Sequenzeigenschaften und ruft die Erwartungswert- und Standardabweichungs-SVM damit auf. Auf Basis des erhaltenen Erwartungswert und der erhaltenen Standardabweichung wird der P-Wert des Scores berechnet und ausgegeben.

Test der SVM

Um die Genauigkeit der von der SVM ausgegebenen P-Werte beurteilen zu können, wurden zufällig ausgewählte paarweise Alignments aus der Bralibase 2.1 Bibliothek [23] genommen. Diese Alignments sind Vergleiche zwischen je zwei Sequenzen, die tatsächlich in Organismen auftreten und von denen man bereits die Familienzugehörigkeit kennt. Da die SVM nur für Sequenzenpaare trainiert wurde, die in etwa gleich lang sind, wurden nur entsprechend annähernd gleichlange Sequenzpaare ausgewählt (Tabelle 4.4).

Je Test wurde der LocARNA Score des Sequenzpaars berechnet. Mit dem Sequenzpaar und dem Score wurde die SVM aufgerufen, um den P-Wert zu erhalten. Als Gegenprobe wurden zehntausend Scores mit durch zufällige Permutation (Mononukleotidshuffling) aus den Sequenzen gewonnenen neuen Sequenzen berechnet. Wie auch schon bei den Daten zum Trainieren der SVM, wurde jede permutierte Sequenz nur für die Berechnung von genau einem Score verwendet. Auf die Verteilung der entstandenen Scores wurde eine Normalverteilung angepasst und berechnet, wie groß der prozentuale Anteil ist, der noch größer als der Ausgangsscore ist. Dieser Wert ist der Vergleichs P-Wert. Die Ergebnisse sind in Tabelle 4.3 zu sehen.

Identifier	P-Wert(Nullmodell)	P-Wert(SVM)
U2.apsi-50.sci-89.no-1	$< 1.0 \cdot 10^{-16}$	$< 1.0 \cdot 10^{-16}$
5 8S rRNA.apsi-61.sci-90.no-1	$2.220 \cdot 10^{-15}$	$6.106 \cdot 10^{-15}$
Cobalamin.apsi.46.sci.66.no1	0.2307	0.24045
SRP euk arch.apsi-37.sci-94.no-1	$2.273 \cdot 10^{-12}$	$1.554 \cdot 10^{-15}$
K chan RES.apsi-69.sci-78.no-1	$9.240 \cdot 10^{-09}$	$1.522 \cdot 10^{-11}$
K chan RES.apsi-57.sci-68.no-1	$3.082 \cdot 10^{-05}$	$3.884 \cdot 10^{-07}$

Tabelle 4.3: P-Wert Berechnung mit Nullmodell gegen Berechnung mit trainierter SVM

Identifier	Länge	Anzahl A/U/G/C
U2.apsi-50.sci-89.no-1 Sequenz1	192	41/57/44/45
U2.apsi-50.sci-89.no-1 Sequenz2	191	38/58/49/46
5 8S rRNA.apsi-61.sci-90.no-1 Sequenz1	153	43/31/39/40
5 8S rRNA.apsi-61.sci-90.no-1 Sequenz2	153	41/37/33/42
Cobalamin.apsi.46.sci.66.no1 Sequenz1	204	58/54/41/51
Cobalamin.apsi.46.sci.66.no1 Sequenz2	201	44/25/60/72
SRP euk arch.apsi-37.sci-94.no-1 Sequenz1	302	71/85/55/91
SRP euk arch.apsi-37.sci-94.no-1 Sequenz2	298	66/66/66/100
K chan RES.apsi-69.sci-78.no-1 Sequenz1	114	14/28/32/40
K chan RES.apsi-69.sci-78.no-1 Sequenz2	114	20/32/28/34
K chan RES.apsi-57.sci-68.no-1 Sequenz1	114	20/40/21/33
K chan RES.apsi-57.sci-68.no-1 Sequenz2	114	23/36/24/31

Tabelle 4.4: Details zu den SVM-Test Sequenzen

4.2 IntaRNA

Es wurde IntaRNA Version 1.1.1 benutzt und die Parameter wie folgt gewählt:

- Fenstergröße: 150 nt
- maximale Distanz zwischen zwei gepaarten Basen: 100 nt
- exakte Anzahl an gepaarten Basen in der Seedregion: 7 nt
- exakte Anzahl ungepaarter Basen in der Seedregion: 0 nt
- maximale Länge der Hybridstelle: 80 nt
- RNAplfold[1] wurde benutzt um die ED Werte zu berechnen
- RNAup[13] wurde benutzt für die Berechnung der ED Werte der bindenden RNA

Alle anderen Parameter wurden auf den Standardwerten belassen.

IntaRNA sucht zu einer ncRNA die passende mRNA, mit der sie interagiert und so deren Translation regelt. Dafür wird IntaRNA meist mit einer ncRNA und einer Datenbank mit mRNA Sequenzen aufgerufen. Daher werden hier anders als bei LocARNA für einen Vergleich nicht Ergebnisse aus je 2 noch nicht benutzten Sequenzen verwendet, sondern der Anwendungsfall wird imitiert und Interaktionen einer ncRNA mit zehntausend mRNA Sequenzen durchgeführt. Um zu verhindern, dass der Einfluss dieser einzelnen ncRNA einen zu großen Einfluss hat, wurden die Vergleiche zehn bis sechzig mal mit unterschiedlichen ncRNA Sequenzen wiederholt.
Das untersuchte Feld ist in Tabelle 4.5 zu sehen. Die ncRNA Länge wurde in Schritten von 40nt variiert, die mRNA Länge in Schritten von 60nt und der AU-Anteil in Schritten von 5%.
Im Gegensatz zu LocARNA wurden hier die Sequenzen wieder verwendet. Zum Beispiel im

#Versuche	ncRNA Länge	mRNA Länge	AU-Anteil ncRNA	AU-Anteil mRNA
60	20-300	400	50%	50%
10	160	100,200-500,760,1000	50%	50%
60	160	400	25%-75%	50%
10	160	400	50%	25%-75%

Tabelle 4.5: Bei IntaRNA durchgeführte Versuche. Je Versuch wurden Interaktionen zwischen einer ncRNA und zehntausend mRNA Sequenzen gesucht.

ersten Fall wurden die selben 10000 mRNA Sequenzen gegen die verschieden langen ncRNA Sequenzen getestet. So kann sichergestellt werden, dass nur die Eigenschaft das Ergebnis beeinflusst, die man variiert.

4.2.1 Verteilungsuntersuchung

Die erhaltenen Energien (je einzelnen Versuch betrachtet) scheinen extremwertverteilt zu sein (mit umgekehrtem Vorzeichen).
Mit Hilfe des Log-log Tests konnte verifiziert werden, dass die Energien tatsächlich in etwa einer Extremwertverteilung entsprechen (siehe Abbildung 4.8). Da der Log-log Test eine gute

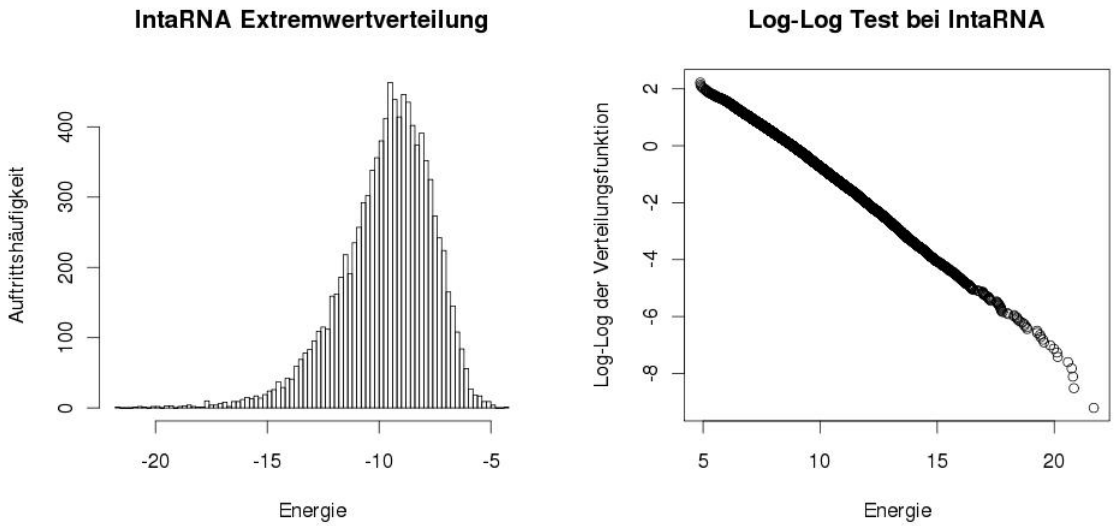

Abbildung 4.8: Stichprobe einer Energieverteilung und zugehöriger Log-Log Test. ncRNA Länge: 300nt, mRNA Länge 400nt

Näherung zeigt, wird im weiteren angenommen, dass die Energie tatsächlich einer Gumbel-Extremwertverteilung entspricht. Als Gegenprobe wurde mit Hilfe des GNU R Pakets „EVD“ (ExtremValueDistribution) geprüft, ob die allgemeine Extremwertverteilung die Verteilung noch besser beschreiben kann. Im Gegensatz zur Gumbelverteilung besitzt die allgemeine Extremwertverteilung drei Parameter, über die sie sich definiert, μ (Location), σ (Scale) und ξ (Shape) (siehe auch Formel 4.1). Falls $\xi \mapsto 0$, ist die allgemeine Extremwertverteilung mit der Gumbelverteilung identisch.
EVD hat nur sehr geringe Änderungen gegenüber der Gumbelverteilung gezeigt, daher wird weiter eine Gumbelverteilung als die korrekte Beschreibung der Energieverteilung angesehen.

$$\text{allg. Extremwertverteilung: } e^{-(1+\xi * \frac{x-\mu}{\sigma})^{\frac{-1}{\xi}}} \tag{4.1}$$

Aufgetretene Abweichungen

Es lässt sich feststellen, dass bei sehr kurzen ncRNA Längen (bis $\sim$60 Nukleotide) die Energiewerte von der angenommenen Extremwertverteilung abweichen und ein abgeschnittenes Ende

haben (Abbildung 4.9). Die Abweichung von der Extremwertverteilung lässt sich damit begrün-

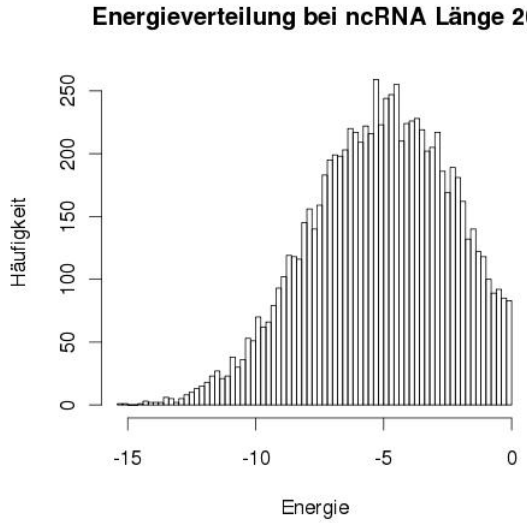

Abbildung 4.9: Verteilung der Energien bei ncRNA Länge 20 nt

den, dass bei so kurzen Längen die Interaktionsstelle nicht mehr als lokal angenommen werden kann, sondern global die gesamte ncRNA umfasst. Der Wechsel von einer lokalen zu einer globalen Interaktion scheint mit einer Änderung der Art der Verteilung einher zu gehen.
Das abgeschnittene Ende ergibt sich dadurch, dass bei kurzen Sequenzen oft keine Interaktionsstelle gefunden werden kann, die eine so hohe Energie aufweist, dass sie die Entfaltungsenergie übersteigt. In einem solchen Fall wird die Interaktion von IntaRNA verworfen.

4.2.2 Einfluss der minimalen freien Energie und der Interaktionslänge

Die Untersuchung des Einflusses der Länge auf die zwei Parameter einer Gumbel-Extremwertverteilung, Location und Scale, zeigt abgesehen vom eigentlichen Verlauf in Abhängigkeit von der Länge deutlich, dass es mindestens eine weitere Eigenschaft geben muss, die den Ort und die Breite der Verteilung beeinflusst. In Abbildung 4.10 repräsentiert jeder Punkt die Verteilung eines Versuchs, aufgespalten nach Location (linke Graphik) und Scale (rechte Graphik). In beiden Graphiken streuen die Versuche gleicher Länge deutlich. Das Basenverhältnis scheidet aus, da bei den erzeugten Daten für die Untersuchung der Länge das AU zu GC Verhältnis konstant gehalten wurde. Wie in den Dichteplots 4.11 und 4.12 zu sehen ist, kommt weder die *mfe* noch die Länge der Interaktion als Kandidat für die gesuchte Eigenschaft in Frage. Beide lassen keinen Zusammenhang zur Energie erkennen.

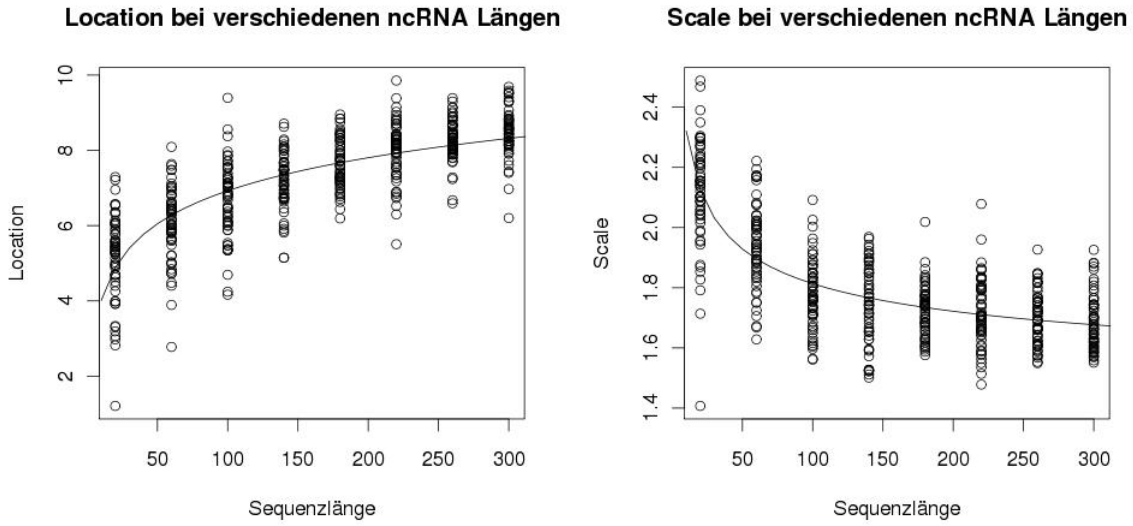

Abbildung 4.10: Location und Scale bei verschiedenen ncRNA Sequenzen und ncRNA Längen

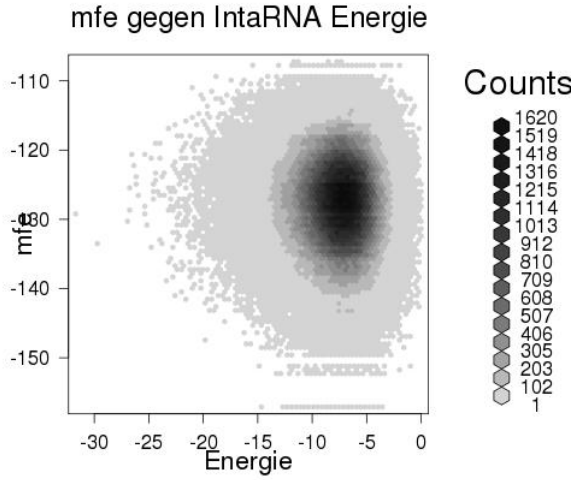

Abbildung 4.11: *mfe* der ncRNA gegen Energie der Interaktion(IntaRNA)

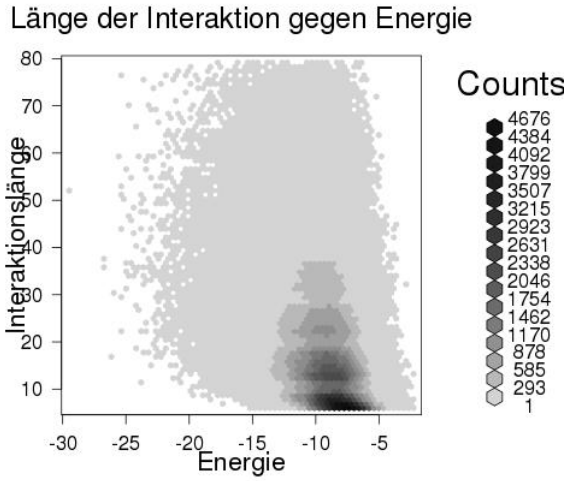

Abbildung 4.12: Länge der Interaktion im Verhältnis zur Energie der Interaktion(IntaRNA)

4.2.3 Location/Scale in Abhängigkeit von der Sequenzlänge

Es lässt sich erkennen, dass Location bei steigender ncRNA Länge steigt, allerdings nicht linear, sondern je größer die Länge wird, desto geringer ist die Steigerung. Eine Logarithmusfunktion beschreibt den Zusammenhang sehr gut. Eine Regression über die Formel $y = a \cdot log(x) + c$ wurde durchgeführt, und die Werte $a = 1.266$ und $c = 1.096$ erhalten.
Scale dagegen nimmt mit steigender Länge ab, nähert sich also immer weiter der x-Achse an. Per Regression konnte die Formel $y = a \cdot x^b + c$ an Scale angepasst werden, mit $a = 2.0374$, $b = -0.3269$ und $c = 1.3613$ (siehe Abbildung 4.10).

Sehr ähnlich verhalten sich Location und Scale auch, wenn man die Länge der mRNA variiert. Location steigt ebenfalls in etwa entsprechend einer Logarithmusfunktion, hier konnte per Regression die Werte $a = 2.406$ und $c = -6.956$ gefunden werden. Scale nähert sich genau wie bei Veränderung der ncRNA Länge immer mehr der x-Achse an. Es konnte die Formel $\text{Scale} = 1.8666 \cdot x^{-0.2211} + 1.27165$ als gute Abschätzung ermittelt werden (Abbildung 4.13).

4.2.4 Location/Scale in Abhängigkeit vom AU-Anteil der ncRNA Sequenzen

Das Verhalten von Location und Scale bei Veränderung des AU-Anteils in der ncRNA ist uneindeutig. Es scheint eine Tendenz zu geben, dass Location zwischen 50% und 75% AU-Anteil sinkt, je größer der Anteil wird. Allerdings sind viele Ausreißer vorhanden, die eine exaktere Bestimmung dieses Verhaltens schwierig machen. Ein AU-Anteil unter 50% scheint dagegen keinen weiteren Einfluss auf Location zu haben. Alle Daten zwischen 25% AU-Anteil und 50% AU-Anteil

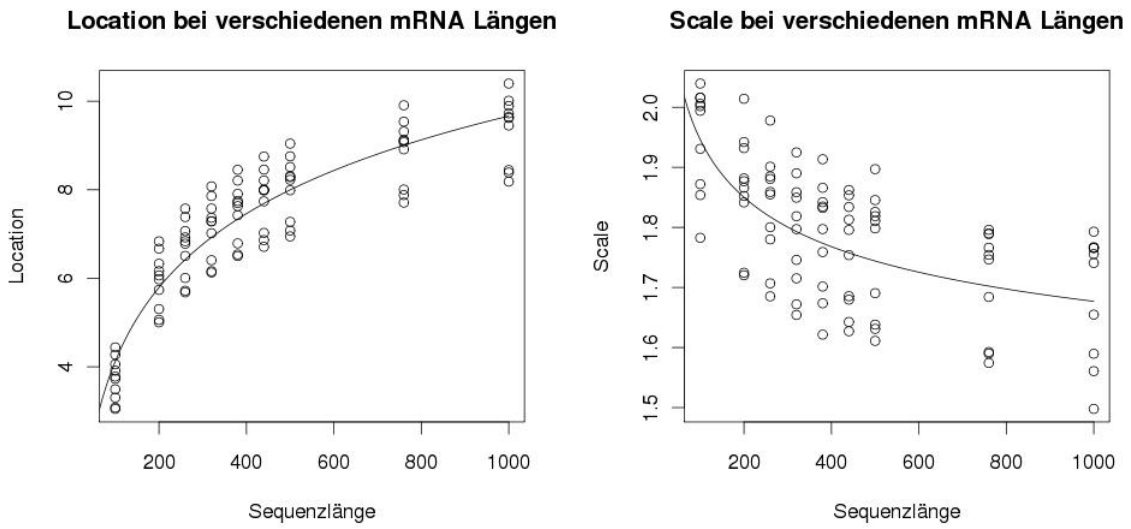

Abbildung 4.13: Location und Scale bei verschiedenen ncRNA Sequenzen und mRNA Längen

haben etwa den selben Wert.
Bei Scale sieht es ähnlich aus. Scale sinkt etwas, mit steigendem AU-Anteil oberhalb von 50%. Im Bereich zwischen 25% und 50% AU-Anteil ist Scale dagegen unabhängig vom AU-Gehalt, soweit sich dies bei der Streuung sagen lässt (Abbildung 4.14).

4.2.5 Location/Scale in Abhängigkeit vom AU-Anteil der mRNA Sequenzen

Die Veränderung des AU-Anteils der mRNA lässt wieder ein etwas besser erkennbares Bild zu. Hier sinkt Location eindeutig, falls der AU-Anteil die 45% übersteigt. Die Veränderung scheint linear zu sein.
Unterhalb von 45% erhöht sich allerdings die Ungenauigkeit. Ein Teil der Location Punkte scheint zu sinken je näher der Gehalt dem getesteten Minimum von 25% AU kommt, während andere die in etwa lineare Veränderung, die zwischen 45% und 75% erkennbar ist, auch unterhalb von 45% fortführen, also mit sinkendem AU-Anteil weiter steigen.
Scale verhält sich nahezu exakt genauso. Oberhalb von 50% AU-Anteil lässt sich ein relativ sauberer Zusammenhang erkennen. Je näher der AU-Anteil der Testgrenze von 75% kommt, desto weiter sinkt Scale. Unterhalb von 50% dagegen erhöht sich die Breite der Scale Verteilung. Teilweise sinkt in diesem Bereich der Scale wieder, teilweise steigt er aber auch weiter (Abbildung 4.15).

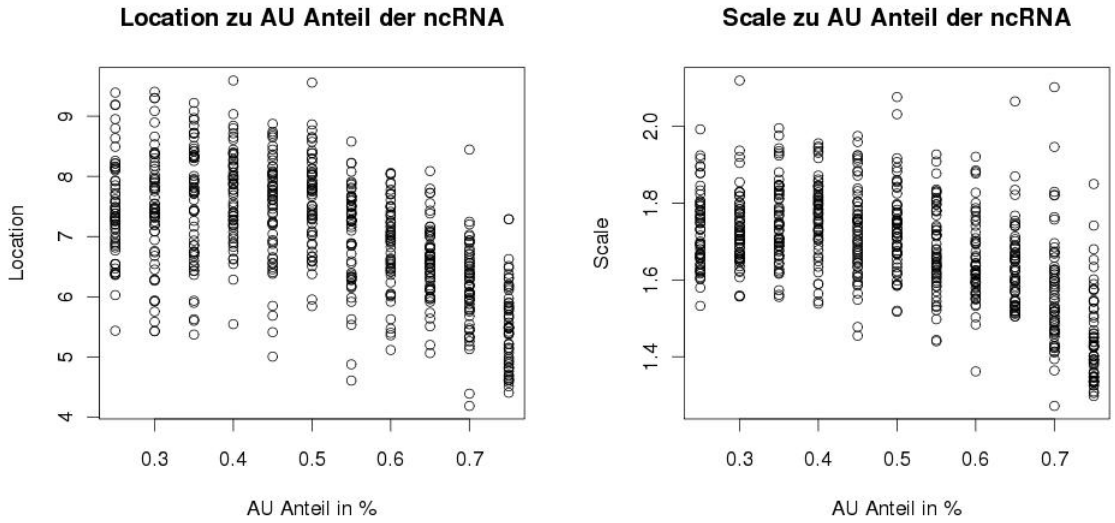

Abbildung 4.14: Location und Scale bei verschiedenen ncRNA Sequenzen und ncRNA AU-Anteilen

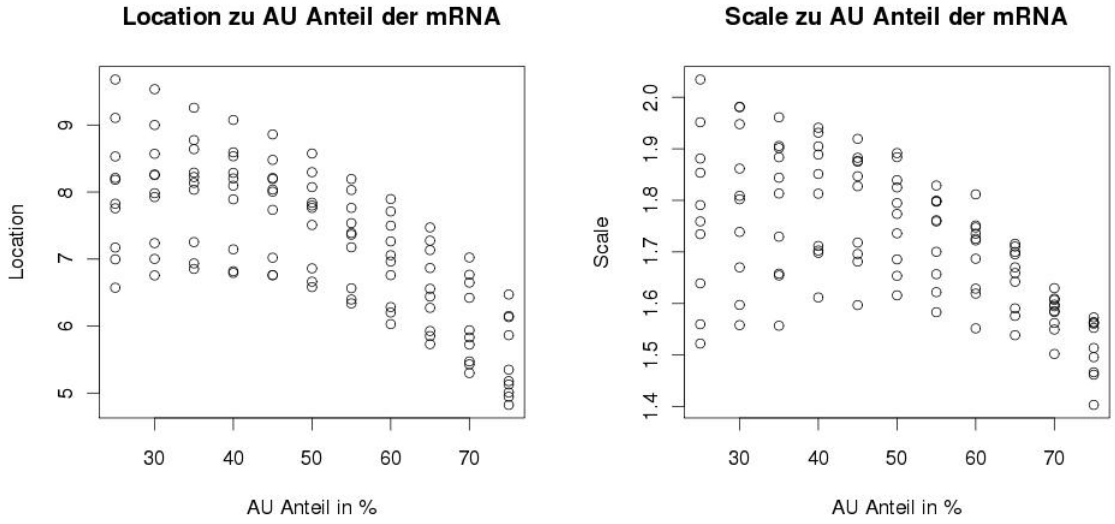

Abbildung 4.15: Location und Scale bei verschiedenen ncRNA Sequenzen und mRNA AU-Anteilen

5 Diskussion

In dieser Arbeit wurden Signifikanzuntersuchungen bei LocARNA und IntaRNA durchgeführt. Es konnte gezeigt werden, dass die Scoreverteilung bei LocARNA einer Normalverteilung gleicht und die Verteilung der Energie bei IntaRNA einer Gumbel-Extremwertverteilung folgt.

5.1 LocARNA

Die Untersuchungen zeigen, dass bei LocARNA der zu erwartende Score linear von der Länge der Eingabesequenzen abhängt. Die Streuung um den Erwartungswert wird mit steigender Länge ebenfalls größer, allerdings nicht linear, sondern mit der Wurzel der Länge.
Die Scores hängen auch sehr stark vom AU zu GC Verhältnis ab. Je nachdem wie dieses Verhältnis liegt, steigt der Score linear mit der Länge der Sequenzen oder er sinkt linear mit der Länge. Je größer der Anteil an Adenin und Uracil ist, desto höher ist der Score.
Dies könnte damit zusammenhängen, dass A und U nur schwächere Bindungen eingehen, sich also mit wenig Aufwand eine unpassende Sekundärtstruktur in eine andere verschieben lässt, die eventuell besser mit der Vergleichssequenz übereinstimmt.
Um dies genauer sagen zu können, müssten aber noch weitere Analysen gemacht werden, wie etwa eine Untersuchung, wie sich der strukturspezifische Teilscore zum sequenzspezifischen Teilscore verhält. Dominiert eventuell ein Score unter bestimmten Umständen den anderen deutlich? Diese Informationen werden allerdings von LocARNA derzeit nicht ausgegeben.

Der Test der SVM zeigt teilweise Abweichungen zwischen den P-Werten, die mit einem Nullmodell errechnet wurden und den von der SVM ausgegebenen P-Werten. Ein Grund ist sicher, dass die SVM mit Sequenzen trainiert wurde, die genauso viel Adenin, wie Uracil bzw. genauso viel Guanin wie Cytosin enthalten, in der Natur so exakte Gleichverhältnisse jedoch nicht häufig auftreten, sodass mit Sequenzen getestet wurde, deren Mononukleotidverhältnis nicht ausgeglichen ist (Tabelle 4.4). Allerdings muss auch bedacht werden, dass ein Unterschied zwischen 10^{-9} und 10^{-11} auf dem Papier zwar groß aussieht, aber auch schon ein P-Wert von 10^{-9} eine so niedrige Auftrittswahrscheinlichkeit hat, dass man im Mittel mehr als eine Milliarde Scores berechnen muss, um einmal durch Zufall darauf zu treffen. Die meisten Verteilungen, mit denen in dieser Arbeit die P-Wert Analyse durchgeführt wurde, sowie die als Gegenprobe beim SVM Test erzeugten Verteilungen, bestehen aus maximal zehntausend Scores. Die P-Werte liegen also

zum Großteil in Bereichen, die außerhalb der berechneten Stichproben liegen. Daher lässt sich eine gewisse Ungenauigkeit nicht verhindern.
Eine relativ einfache, wenn auch zeitaufwendige Möglichkeit, diesen Fehler zu minimieren, wäre weitere Scores zu berechnen. Je größer die Stichprobe an Scores pro Verteilung ist, desto besser sollten auch die extremen Bereiche abgedeckt sein, und so würde auch der Fehleinschätzung der Scores entgegengewirkt.

Der sechste Test in Tabelle 4.3, für den das Nullmodell einen P-Wert von 10^{-5} ausgibt, während die SVM einen P-Wert von 10^{-7} angibt, ist dagegen bedenklich. Der Nullmodell P-Wert ist noch so hoch, dass er nur knapp außerhalb des Testfelds liegt. Eigentlich müsste hier eine hohe Qualität der SVM Vorhersage vorliegen, was aber nicht der Fall ist. Hier sind weitere Untersuchungen nötig, um den Grund für diese Abweichung festzustellen.

5.2 IntaRNA

Die große Streuung im für IntaRNA erstellten Nullmodell werfen die Frage auf, ob das Modell ungünstig gewählt wurde. Wäre es sinnvoll gewesen, die Versuche mit einem anderen Nullmodell durchzuführen, als das hier in der Arbeit benutzte? Zum Beispiel, ein ähnliches Modell, wie es bei LocARNA benutzt wurde: Für jede Sequenzkombination werden neue Sequenzen benutzt, statt die selbe ncRNA mit 10000 mRNAs zu untersuchen. Damit würde der bei den durchgeführten Versuchen deutlich sichtbare Unterschied zwischen den Interaktionen der einzelnen ncRNAs vermutlich in der Masse verschwinden und nicht mehr in den Graphiken auftauchen. Allerdings ist fraglos ein Grund da, der zu diesen Unterschieden führt.
Insofern kann man sagen, ein anderes Nullmodell wäre vielleicht problemloser handhabbar gewesen, aber nur weil Teile der Problemstellung ignoriert würden, womit die Korrektheit der so erzeugten Ergebnisse zweifelhaft wäre.

Festzuhalten bleibt, dass einige Erkenntnisse gewonnen werden konnten, auch wenn der letzte Schritt, zu einer beliebigen Sequenzkombination den P-Wert der Energie angeben zu können, fehlt.
So konnte eindeutig die Verteilung der Energie bei festgehaltener Merkmalskombination (und ncRNA) als Gumbel-Extremwertverteilung identifiziert werden, sowie das Verhalten bei Änderung der Länge bestimmt werden und auch das Verhalten bei verschiedenen AU zu GC Verhältnissen konnte eingegrenzt werden.
Auch über den Grund für die Unterschiede zwischen den Verteilungen der einzelnen ncRNA Sequenzen konnten Erkenntnisse gewonnen werden. So wird die Streuung zwischen den ncRNA Verteilungen größer, falls der AU-Anteil der mRNA unter 40% fällt, während Veränderungen in der Länge keinen Einfluss haben.

6 Ausblick

Ziel dieser Arbeit war es, einen Weg zu finden, aus den Scores von LocARNA oder den Energiewerten von IntaRNA direkt einen P-Wert zu berechnen. Für LocARNA war dieses Vorhaben erfolgreich. In gewissen Schranken ist es möglich, zu einem berechneten Score direkt einen P-Wert auszugeben. Bei IntaRNA müssen dagegen noch weitere Untersuchungen folgen, bevor es möglich wird, einen P-Wert ohne große Vergleichsprobe zu berechnen.

Ein viel versprechender Ansatzpunkt bei IntaRNA könnte sein, auch das Basenverhältnis der Interaktionsstelle zu betrachtet, statt nur das globale Verhältnis. Die Energie von IntaRNA besteht aus den ED Werten, die vom globalen Verhältnis abhängen, und der Hybridenergie, dieses hängt vom lokalen Basenverhältnis in der Interaktionsstelle ab. Eine Betrachtung beider Verhältnisse würde also der zweiteiligen Energieberechnung entsprechen.
Es könnte auch untersucht werden, wo sich echte ncRNA Sequenzen in der Streuung einordnen. Liegen sie genauso breit verteilt vor, oder existiert vielleicht ein Mechanismus, der in Organismen die ncRNA Sequenzen vorzieht, die an der oberen oder unteren Grenze der Streuung Interaktionen erzeugen. Falls die meisten echten ncRNA Sequenzen nicht so weit streuen, wie in dieser Arbeit die zufälligen ncRNA Sequenzen, könnte man sich damit, ohne den genauen Grund für die Streuung zu kennen, auf die wichtigen, da tatsächlich auftretenden, Verteilungen konzentrieren. Nicht vergessen werden darf aber, dass noch nicht alle ncRNA Sequenzen in allen Lebewesen sequenziert sind. Wenn also die bisher gefundenen Sequenzen sich an einer bestimmten Stelle in der Streuung einordnen, ist dies zwar ein Indiz, aber kein gesicherter Beweis, dass sich alle ncRNA Sequenzen dort anordnen.

Im Weiteren bieten sowohl IntaRNA wie LocARNA einiges an Einstellmöglichkeiten, wie etwa die Länge der Interaktion und die Größe des betrachteten Fensters. Eine Änderung dieser Parameter führt vermutlich zu anderen Scores/Energiewerten, die eventuell wiederum anders auf Länge und AU zu GC Verhältnis reagieren.

Außerdem ist auch die Menge an Parametern der RNA Sequenzen mit Länge und AU zu GC Verhältnis noch nicht ausgeschöpft. In einem weiteren Schritt kann das Verhalten untersucht werden, wenn man die Häufigkeit von Adenin, Uracil, Guanin und Cytosin separat variiert. Diese Verhältnisse waren bei den durchgeführten Untersuchungen festgelegt auf gleich viel Adenin wie Uracil, bzw. Guanin wie Cytosin. Wie an den Testsequenzen für die SVM (Tabelle 4.4)zu sehen ist, kann aber ein solches Gleichverhältnis nicht als gegeben betrachtet werden.

7 Anhang

7.1 Verwendete Software

Shuffle gehört zum Paket SQUID, das ein Teil von HMMER (Version 2.3.2) [6] ist. Es wurde benutzt um pseudo-zufällige Sequenzen zu erzeugen. Wobei eine Beispielsequenz übergeben wurde, welche die gewünschten Parameter (Länge, Anteil A,U,G und C) besitzt. Die zufälligen Sequenzen wurden durch Mononukleotidshuffling aus der Beispielsequenz erzeugt (http://hmmer.janelia.org/).

ViennaPackage 1.7.2 ist eine Bibliothek von Programmen [9], die in der „Theoretische Biochemie“ Gruppe der Universität Wien zusammengestellt wurde. Sie enthält Bibliotheken die für LocARNA benötigt wurden, insbesondere RNAfold, mit dem die Dotplots erstellt wurden, die LocARNA als Eingabe benötigt und mit dessen Hilfe die minimale freie Energie berechnet wurde (http://www.tbi.univie.ac.at/RNA/).

GNU R ist eine Statistik-Software, zu der es eine Reihe von Zusatzpaketen gibt. Die meisten Berechnungen und Grafiken in dieser Arbeit wurden mit R erstellt. Es wurde Version 2.8.1 verwendet (http://www.r-project.org/).

- **EVD** = Extrem Value Distribution. EVD [17] ist ein Zusatzpaket für GNU R, welches benutzt wurde um Extremwertverteilungen bei IntaRNA zu untersuchen. Es wurde Version 2.2-4 verwendet (http://cran.rakanu.com/web/packages/evd/index.html).

- **Hexbin** ist ein Zusatzpaket für GNU R, das die Erstellung von Dichteplots ermöglicht (http://cran.rakanu.com/web/packages/hexbin/index.html).

Libsvm ist eine weit verbreitete SVM Bibliothek [3], die Klassifikation und Regression beherrscht. Hier wurde sie zur Regression benutzt.

Perl ist eine plattformunabhängige Programmiersprache, mit der sich schnell kleine Skripte erstellen lassen. Der sehr einfache Umgang mit regulären Ausdrücken ist eine der Stärken dieser Sprache (http://www.perl.org/).

7.2 Hardware und Laufzeiten

Alle Ergebnisse wurden auf dem Grid des Lehrstuhls für Bioinformatik der Albert-Ludwigs-Universität Freiburg berechnet. Es besteht aus insgesamt elf Rechnern der Typen Intel Xeon 5160, AMD Opteron 275, AMD Opteron 875 und AMD Opteron 2356 mit je 4-8 Kernen, insgesamt 68 Kerne.
Es wurden 52442 Rechenstunden (2185 Rechentage) benötigt, was eine Komplettauslastung des Grids für 32 Tage bedeutet.

7.2.1 Laufzeit LocARNA

Die Laufzeit von LocARNA hängt von mehreren Faktoren ab. Als wichtige Einflussgrößen konnte die Länge der Eingabesequenzen sowie das AU zu GC Verhältnis festgestellt werden. Mit steigender Länge der Sequenzen steigt die Laufzeit sehr schnell an. Außerdem scheint die Rechenzeit zu wachsen, je größer der Anteil an Adenin und Uracil in den Sequenzen ist.
Insgesamt wurde 37860 Stunden (1577 Tage) für LocARNA gerechnet. Auszug aus den Laufzeiten:

Berechnungen	Sequenzlänge	AU-Anteil	Dauer
1’210’000	100nt	verschieden	490 Stunden
1’210’000	160nt	verschieden	4300 Stunden
830’000	200nt	verschieden	8173 Stunden
10’000	100nt	25%	2.7 Stunden
10’000	100nt	75%	6.9 Stunden

7.2.2 Laufzeit IntaRNA

Die Laufzeit von IntaRNA scheint primär von der Länge der Sequenzen abzuhängen, das Basenverhältnis spielt nur eine untergeordnete Rolle. Wobei auch die Länge einen deutlich schwächeren Einfluss auf die Laufzeit hat, als bei LocARNA.
Insgesamt wurde 14582 Stunden (607 Tage) für IntaRNA gerechnet. Auszug aus den Laufzeiten:

Berechnungen	ncRNA Länge	#ncRNAs	mRNA Länge	#mRNAs	Dauer
100’000	160nt	10	100nt	10000	31 Stunden
100’000	160nt	10	760nt	10000	424 Stunden
100’000	160nt	10	1000nt	10000	590 Stunden
4’000’000	20 bis 300	400	400nt	10000	3890 Stunden

7.3 Erzeugte Daten

Es wurden 4'139'000 Scores mit LocARNA berechnet und 13'400'000 Energiewerte mit IntaRNA berechnet. die Aufteilung der Daten nach Länge und AU-Anteil ist in den Tabellen 7.1, 7.2 und 7.3 zu sehen.

Sequenzlänge	% AU-Anteil	Anzahl Scores
20,60,100,140	50	je 20000 => 80000
40,80,120,160,180,200,240,280,320,360,400	50	je 10000 => 110000
600	50	4000
800	50	1000
20,60,100	25	je 20000 => 60000
160,200,220,280	25	je 10000 => 40000
400,600	25	je 5000 => 10000
20,60,100,120,140,160,180,200	75	je 10000 => 80000
240,300	75	je 5000 => 10000
340	75	3000
400,600	75	je 2500 => 5000

Tabelle 7.1: Übersicht aller erzeugten Scores, wenn beide Sequenzen den selben AU-Anteil besitzten

Sequenzlänge	AU-Anteil Seq1 in %	AU-Anteil Seq2 in %	Anzahl Scores
80,100,160	25-75	25-75	je 10000, insgesamt 3630000
200	25-50	25-75	je 10000, insgesamt 660000
200	55	25,30	je 10000, insgesamt 20000
200	55-75	>=AU1 bis 75	je 10000, insgesamt 150000

Tabelle 7.2: Übersicht aller erzeugten Scores, bei unterschiedlichem AU-Anteil

Länge ncRNA	Länge mRNA	Anzahl Energiewerte
160	100,200,260,320,380,440,500,760,1000	je 100000 => 900000

Länge ncRNA	Länge mRNA	Anzahl Energiewerte
20,60,100,140,180,220,260,300	400	je 600000 => 4800000

AU ncRNA	AU mRNA	Anzahl Energiewerte
25-75	50	je 100000 => 1100000
50	25-75	je 600000 => 6600000

Tabelle 7.3: Übersicht aller erzeugten Energiewerte. In den zwei oberen Tabellen sind alle AU-Anteile bei 50%, in der unteren Tabelle sind die ncRNA 160nt lang und die mRNA 400nt lang.

Abbildungsverzeichnis

Tabellenverzeichnis

Literaturverzeichnis

[1] S. H. Bernhart, I. L. Hofacker, and P. F. Stadler. Local RNA base pairing probabilities in large sequences. *Bioinformatics*, 22(5):614–5, 2006.

[2] A. Busch, A. S. Richter, and R. Backofen. IntaRNA: efficient prediction of bacterial sRNA targets incorporating target site accessibility and seed regions. *Bioinformatics*, 24(24):2849–56, 2008.

[3] C.-C. Chang and C.-J. Lin. *LIBSVM: a library for support vector machines*, 2001. Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

[4] P. Clote and R. Backofen. *Computational Molecular Biology: An Introduction*. Mathematical and Computational Biology. Jon Wiley & Sons, Chichester, Aug. 2000. series editor S. Levin. 290 pages.

[5] C. Cortes and V. Vapnik. Support vector networks. *Machine Learning*, 20:273–297, 1995.

[6] R. Durbin, S. Eddy, A. Krogh, and G. Mitchison. *Biological sequence analysis - Probabilistic models of proteins and nucleic acids*. Cambridge University Press, Cambridge, UK, 1998.

[7] W. Gilbert. The RNA world. *Nature*, 319:618, 1986.

[8] S. Gottesman. Micros for microbes: non-coding regulatory RNAs in bacteria. In *Trends in Genetics*, volume 21, pages 399–404. Elsevier Ltd., july 2005.

[9] I. Hofacker and P. Stadler. Vienna RNA package. Paper as Print Copy, 1998.

[10] E. E. R. Institute. Distance learning center - appendix - kolmogorov smirnov test. `http://www.eridlc.com/onlinetextbook/index.cfm?fuseaction=textbook.appendix&FileName=Table7`, 26.März 2009.

[11] U. Krengel. *Einführung in die Wahrscheinlichkeitstheorie und Statistik*. vieweg, Wiesbaden, 8. edition, 2005.

[12] A. Y. Mitrophanov and M. Borodovsky. Statistical significance in biological sequence analysis. *Briefings in Bioinformatics*, 7(1):2–24, 2005.

[13] U. Muckstein, H. Tafer, J. Hackermuller, S. H. Bernhart, P. F. Stadler, and I. L. Hofacker. Thermodynamics of RNA-RNA binding. *Bioinformatics*, 22(10):1177–82, 2006.

[14] M. Rehmsmeier, P. Steffen, M. Höchstmann, and R. Giegerich. Fast and effective prediction of microRNA/target duplexes. In *RNA Journal*, pages 1507–1517. Cold Spring Harbor Laboratory Press., 2004.

[15] Roche. Products and solutions. http://www.454.com/products-solutions/system-features.asp.

[16] R. Schlittgen. *Einführung in die Statistik: Analyse und Modellierung von Daten*. Oldenbourg, München, Wien, 5. edition, 1995.

[17] A. Stephenson. *A User's Guide to the evd Package*. National University of Singapore, 2006.

[18] B. Tjaden, S. S. Godwin, J. A. Opdyke, M. Gullier, D. X. Fu, S. Gottesman, and G. Storz. Target prediction for small, noncoding RNAs in bacteria. *Nucleic Acid Research*, 34(9):2791–2802, 2006.

[19] S. Washietl, I. L. Hofacker, and P. F. Stadler. Fast and reliable prediction of noncoding RNA. In *PNAS*, volume 102, pages 2454–2459. National Academy of Sciences, 2005.

[20] Wikipedia. Gumbel-Verteilung. http://de.wikipedia.org/wiki/Gumbel-Verteilung, 17. Okt. 2008.

[21] Wikipedia. Kolmogorow-Smirnow-Test. http://de.wikipedia.org/wiki/Kolmogorow-Smirnow-Test, 22. Mär. 2009.

[22] S. Will, K. Reiche, I. L. Hofacker, P. F. Stadler, and R. Backofen. Inferring non-coding RNA families and classes by means of genome-scale structure-based clustering. *PLOS Computational Biology*, 3(4):e65, 2007.

[23] A. Wilm, I. Mainz, and G. Steger. An enhanced RNA alignment benchmark for sequence alignment programs. *Algorithms Mol Biol*, 1:19, 2006.

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