

Analyse der RNA Multiloop Strukturen in der Datenbank RNALoops

Studienprojekt
im Bachelorstudiengang Data Science

vorgelegt von

Lukas Hunold

am Institut für Algorithmische Bioinformatik



im Januar 2023

Erstbetreuer: Prof. Stefan Janssen
Zweitbetreuer: Katharina Maibach

Inhaltsverzeichnis

1 Einordnung	3
2 Einleitung	5
3 Analyse der Datenbank	9
3.1 Download der Daten und Parsing	10
3.2 Kontrolle und Filtern der Daten	11
3.2.1 Qualität der Multiloops	11
3.2.2 Veggleich mit RNAStrand	14
3.2.3 Heterogenität der Daten	15
3.3 Clustering der Strukturen nach planaren Winkeln	16
3.3.1 Winkel nach Loop-Typ	16
3.3.2 Detaillierte Übersichten von Winkel-Clustern	17
3.3.3 Einfluss von Parametern auf Winkel	21
4 Zusammenfassung	23

1 Einordnung

RNA-Moleküle spielen bei diverse biologische Prozesse in Zellen eine bedeutsame Rolle, darunter die Proteinbiosynthese, die Regulation der Genexpression und die Katalyse von chemischen Reaktionen [1]. Eine wichtige Eigenschaft von RNA-Molekülen ist ihre Fähigkeit, sich in komplexe dreidimensionale Strukturen zu falten, die ihnen ermöglichen, diese verschiedenen Funktionen auszuführen. Eine bestimmte Art von RNA-Struktur, die in vielen verschiedenen RNA-Molekülen gefunden werden kann, ist der sogenannte Multiloop [2].

RNA (Ribonukleinsäure) ist eine Art von Nukleinsäure, die aus Nukleotid-Monomeren besteht. Jedes Nukleotid in der RNA besteht aus einer Zucker-Molekül (Ribose), einer Phosphatgruppe und einer stickstoffhaltigen Base. Es gibt vier Arten von stickstoffhaltigen Basen in der RNA: Adenin (A), Guanin (G), Cytosin (C) und Uracil (U). Ein Multiloop entsteht, wenn eine Kette von Basen geschlossen wird und damit einen Ring aus Abschnitten mit ungepaarten Basen (Strand) und gepaarten Basen (Helix) bildet [Abb. 3.7]. Voraussetzung für einen Multiloop ist, dass mindestens drei Helices (unterbrochen von Strands) vorkommen. Bei nur zwei Helices spricht man dagegen von einem internal loop, bei nur einer Helix von einem Hairpin.[3]

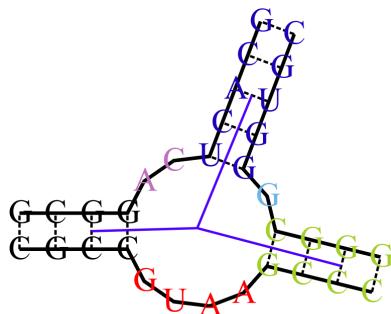


Abbildung 1.1: Vereinfachte Darstellung der Sekundärstruktur eines RNA-Multiloops. Man erkennt die vier verschiedenen Basen und deren Bindungen. Die blauen Vertices deuten die Vektoren an mit Hilfe derer später die Winkel zwischen Helices berechnet werden.

Multiloops sind relevant für die oben genannten Funktionen der RNA. Sie können auch bei der Stabilität und dem Faltungsverhalten von RNA-Molekülen eine Rolle spielen. Die Untersuchung der Struktur und Funktion von RNA-Multiloops kann Einblicke in

1 Einordnung

die Art und Weise geben, wie diese Moleküle ihre verschiedenen Funktionen in Zellen ausführen [4]. Im Allgemeinen gestalten sich solche Analysen allerdings als schwierig, da die Datenlage im Bezug auf RNA Strukturen nicht sehr umfassend ist. Das liegt im Wesentlichen am hohen experimentellen Aufwand, bei der Vermessung dieser Strukturen. Heutige Standardverfahren mit hoher Auflösung sind NMRS (nuclear magnetic resonance spectroscopy) [5], Röntgenkristallographie [6] und Elektronenmikroskopie [7], die aber alle zeitaufwendig und teuer sind.

In den letzten Jahren wurde viel dazu untersucht, wie Methoden des maschinellen Lernens und ähnliche datenanalytische Verfahren diese Experimente ergänzen können [Abb. 3.8]. Allein 2021 wurden mindestens drei solcher Verfahren basierend auf deep learning veröffentlicht [8]–[10]. Ein Vergleich dieser und weiterer Methoden wird beispielsweise in [1] durchgeführt. Der heilige Gral in dem Feld wäre die vollständige Vorhersage der Sekundärstruktur (Bindungen der Basen) und Tertiärstruktur (3D Anordnung) eines RNA Komplexes aus der gegebenen Sequenz der Basen. Eine Vielzahl von Problemen lässt bis heute jedoch offen, wie weit so etwas möglich ist. Das Thema soll hier nicht weiter vertieft werden, einige Aspekte werden in [1] diskutiert.

Name	Authors	Year	Method
CROSS	Delli Ponti <i>et al.</i>	2017	ANN ^a
DMfold	Wang <i>et al.</i>	2019	LSTM ^b
SPOT-RNA	Singh <i>et al.</i>	2019	CNN ^c + BLSTM ^d
E2Efold	Chen <i>et al.</i>	2019	CNN ^c + Transformer ^e
RNA-state-inf	Willmott <i>et al.</i>	2020	BLSTM ^d
RPRes	Wang <i>et al.</i>	2021	BLSTM ^d + ResNet ^f
MXfold2	Sato <i>et al.</i>	2021	BLSTM ^d + ResNet ^c
UFold	Fu <i>et al.</i>	2021	CNN ^c

Abbildung 1.2: Übersicht von Veröffentlichungen über Deep-Learning Verfahren zur Vorhersage der Sekundärstruktur von RNA Komplexen [1] (ANN = Artificial Neural Network, LSTM = Long-Short-Term Memory, CNN = Convolutional Neural Network, BLSTM = Bidirectional Long-Short-Term Memory)

Elementar für alle diese Ansätze ist eine solide Datengrundlage. Für den Bereich der Multiloops wurden dazu 2022 die Datenbank RNALoops erstellt [11], die Strukturen aus der PDB (Protein Database) [12] extrahiert und mittels einer Web REST-API frei zur Verfügung stellt. Sie enthält Informationen über ca. 80.000 Multiloops und wird regelmäßig aktualisiert. Neben den Sequenzen der Loops und ihrer Position in der Gesamtstruktur des PDB Komplexes werden auch Winkelinformationen mitgeliefert. Darüber hinaus gibt es Visualisierungen der Sekundär- und Tertiärstruktur. Die Daten können als PDF und in Bildform heruntergeladen werden.

Die vorliegende Arbeit beschäftigt sich damit die Datenbank besser zu verstehen, auf mögliche Schwächen zu untersuchen und die darin enthaltenen Multiloops zu analysieren. Insbesondere sollen dabei die Winkelinformationen in den Blick genommen werden, da diese speziell sind für die RNALoops Datenbank und nicht direkt in anderen Datenbanken oder der PDB zu finden sind.

2 Einleitung

Die Datenbank RNALoops wurde im Sommer 2022 von Forschern des Institute of Computing Science der Poznan University of Technology unter der Leitung von Prof. Marta Szachniuk und Prof. Maciej Antczak veröffentlicht und beinhaltet ca. 80.000 Multiloop-Strukturen [Abb. 3.8]. Das zugehörige Paper “RNALoops: a database of RNA multiloops” beschreibt, den Extraktionsprozess aus der PDB, den Aufbau der Datenbank und wie User mit der Web API interagieren können [11]. Die Details über zugrundeliegende Algorithmen für das Bestimmen der Sekundärstruktur, das Identifizieren von Multiloops und das Bestimmen der Eigenschaften, werden wesentlich in vorherigen Veröffentlichungen der Gruppe diskutiert [4], [13], [14]. Ausnahme sind die charakteristischen Winkel der Helices eines Multiloops, hier findet sich im RNALoops Paper eine genaue Beschreibung, wie diese bestimmt werden.

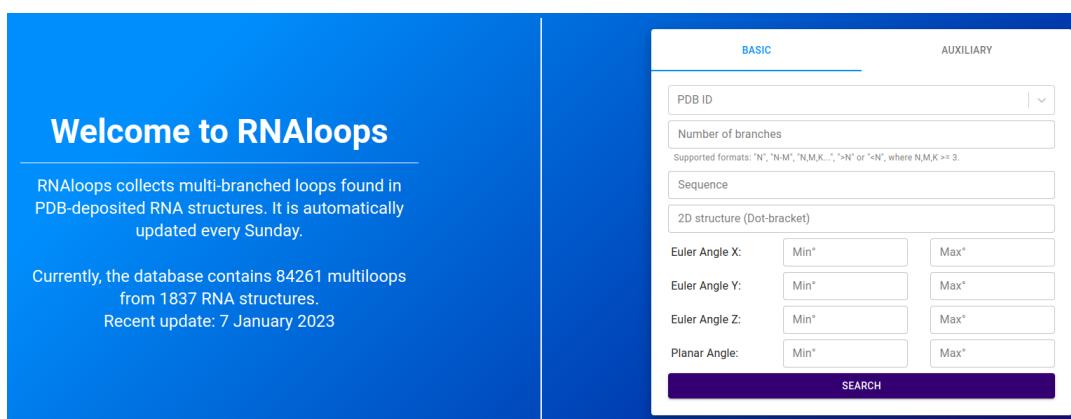


Abbildung 2.1: Homepage der RNALoops Datenbank mit aktuellem Stand und Suchmaske um Multiloops mit bestimmten Kriterien zu finden.

RNA Komplexe nehmen im thermischen Gleichgewicht eine feste Struktur im dreidimensionalen Raum an (Tertiärstruktur). Während viel Forschung betrieben wird um diese Formierung vollständig zu verstehen, beschränkt sich die vorliegende Arbeit auf ein bestimmtes räumliches Feature von Multiloops. Hierbei handelt es sich um den Winkel, in dem eine Helix relativ zu den Loop selber (dem Teil der ungepaarten Basen) steht. Insbesondere werden die Winkel der Helices zueinander betrachtet, da die räumliche Orientierung der Helices zueinander einen Einfluss auf die Funktion der RNA haben kann. Zur Bestimmung dieser Winkel heißt es in [11]:

2 Einleitung

The mutual positions of all pairs of adjacent helices protruding from the loop are designated for each multiloop. For this purpose, planar and Euler angles are computed between directional vectors of these helices. The beginning and the end of each vector are in the geometric centers of the multiloop and helix, respectively. The first point is the centroid in the set of all non-hydrogen atoms that belong to the first base pairs of outgoing helices. The second point is based on all non hydrogens from the third base pair in the helix or the first pair if the helix has < 3 basepairs. Planar angle is computed according to Equation (1) between two directional vectors, \vec{a} and \vec{a}' , projected onto the plane. Euler angles, a , b , and c , reflect the orientation of a directional vector to the other. They define rotations to be made about the three coordinate axes to superimpose two helices. The helix-representing vectors, \vec{a} and \vec{b} , are projected onto the planes perpendicular to all axes of the coordinate system. An angle between the vectors computes from Equation (1) separately for each dimension.

Dabei beschreibt Gleichung (1) die Berechnung der Winkel zwischen zwei Helices durch Projektion:

$$\phi = \arccos \left[(\vec{a} \cdot \vec{b}) / (|\vec{a}| \cdot |\vec{b}|) \right] \quad (2.1)$$

Die planaren Winkel werden dabei durch Projektion der Helices auf die Ebene durch beide Helices bestimmt, die Eulerwinkel durch Projektion auf alle drei Raumebenen eines zuvor gewählten kartesischen Koordinatensystems.

Dieses Projekt beschränkt sich auf die Untersuchung der planaren Winkel, da sie unabhängig vom gewählten Koordinatensystem sind, und damit zuverlässiger. Sie beschreiben die relativen Winkel zweier benachbarter Helices zueinander, wenn die Vektoren dieser Helices zusammen in eine Ebene projiziert werden. Die Vektoren bestimmen sich dabei wie beschrieben aus der Verbindung zwischen dem geometrischen Mittelpunkt des gesamten Multiloops und dem Mittelpunkt des dritten (bzw. ersten falls kein drittes existiert) Basenpaar der Helix.

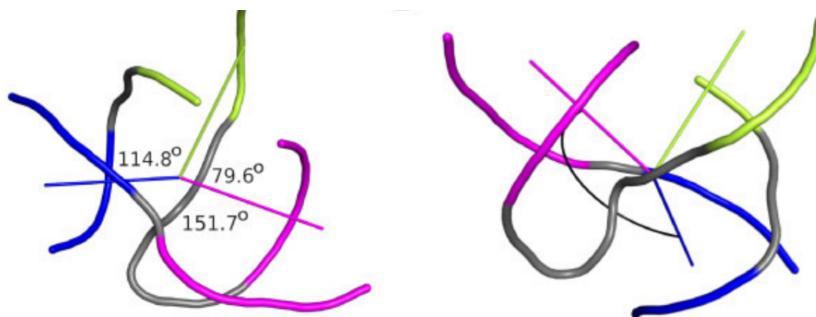


Abbildung 2.2: Beispielgrafik zur Bestimmung der Winkel in einem Multiloop [11]. Man kann erahnen wie schwierig es ist die richtige Wahl der Vektoren zu treffen, um die räumliche Lage der Helices möglichst gut beschreiben zu können. Die Wahl der Autoren von RNALoops wird im Text beschrieben.

Bereits diese Festlegung erlaubt Spielraum (es könnte auch das zweite oder vierte Basenpaar gewählt werden). Im Allgemeinen sind die berechneten Winkel nicht einzigartig, sondern es gäbe viele weitere Definitionen, die man betrachten könnte. Die Idee ist aber, die relative Lage der Helices zueinander möglichst gut zu repräsentieren, ohne dabei die Faltung der RNA zu stark zu berücksichtigen. Im weiteren Verlauf der Helix kann diese sehr stark verdreht sein, und daher werden die Winkel unzuverlässig. Die gewählten Parameter erlauben eine gute Einschätzung der Winkel, in denen die Helices auf dem Loop stehen.

Im Verlaufe dieser Arbeit wird beschrieben, wie Multiloops gemäß dieser Winkel zusammengefasst werden können und welche Parameter die Formation der Winkel beeinflussen könnten. Dazu ist aber zunächst eine gewisse Menge an Vorarbeit notwendig, indem die Daten gewonnen, strukturiert und geprüft werden, um anhand einer soliden Grundlage genauere Untersuchungen durchführen zu können. Die folgenden Abschnitte beschreiben diesen Prozess bestehend aus (1) Download der Daten und Parsing, (2) Prüfen der Intgrität der vorliegenden Strukturen, (3) Vergleich mit anderen Datenbanken, (4) Clustering der Strukturen gemäß planarer Winkel und (5) Analyse des Einflusses von Parametern auf die 3D Struktur.

3 Analyse der Datenbank

Im Folgenden werden die einzelnen Schritte der Analyse der Datenbank beschrieben. Diese wurden alle mit Hilfe von Python inklusive zusätzlicher Packages durchgeführt. Zu nennen sind hier insbesondere:

selenium <https://pypi.org/project/selenium> - zum durchsuchen der Website

textract <https://pypi.org/project/textract> - zum Parsen heruntergeladener PDFs

numpy, pandas, scipy - zur Datenverarbeitung und Auswertung

pyplot, seaborn, IPython - für die Darstellung der Daten und Ergebnisse

BioPython <https://pypi.org/project/biopython> - zum Verarbeiten von PDB Dateien

sqlalchemy <https://pypi.org/project/SQLAlchemy> - zum Vergleich mit RNAStrands

scikit-learn <https://pypi.org/project/scikit-learn> - für das Clustering von Multiloops

Der gesamte Code für die Analyse ist in einem Repository zusammengefasst und minimal kommentiert, sowie dokumentiert. Die Daten sind nur lokal gespeichert, da sie insgesamt sowohl von der Menge der Dateien als auch dem Speicherbedarf sehr umfangreich sind. Zugriff auf alle diese Ressourcen kann bei Interesse jederzeit gewährt werden.

Im Folgenden werden die einzelnen Prozessschritte kurz und übersichtlich zusammengefasst. Für Details sollte der Quellcode konsultiert werden.

3 Analyse der Datenbank

3.1 Download der Daten und Parsing

Die Daten der ca. 80.000 Multiloops können von der RNALoops Website

<https://rnaloops.cs.put.poznan.pl>

abgerufen werden. Jeder Multiloop besitzt eine eindeutige ID und weitere Informationen, die im PDF Format heruntergeladen werden können. Dazu stehen Visualisierungen der Sekundärstruktur des Loops alleine und innerhalb des RNA Komplexes zur Verfügung. Im ersten Schritt des Projektes wurden alle diese Daten gesammelt und aufbereitet. Aus der PDF erhält man insbesondere folgende Feature über ein Multiloop:

- Die vollständige Sequenz des Multiloops als Basenfolge und in Dot-Bracket-Notation
- Der Loop-Type (Anzahl der Helices die vom Multiloop abgehen)
- Die Home-Struktur in der PDB, aus der dieser Multiloop gewonnen wurde
- Sequenzen aller Abschnitte (Strands und Helices) im Multiloop
- Länge aller Abschnitte (Anzahl Basen bzw. Anzahl Basenpaare)
- Start- und Endposition der Abschnitte in der Home-Struktur
- Euler- und Planar-Winkel der Helices (siehe Einleitung)

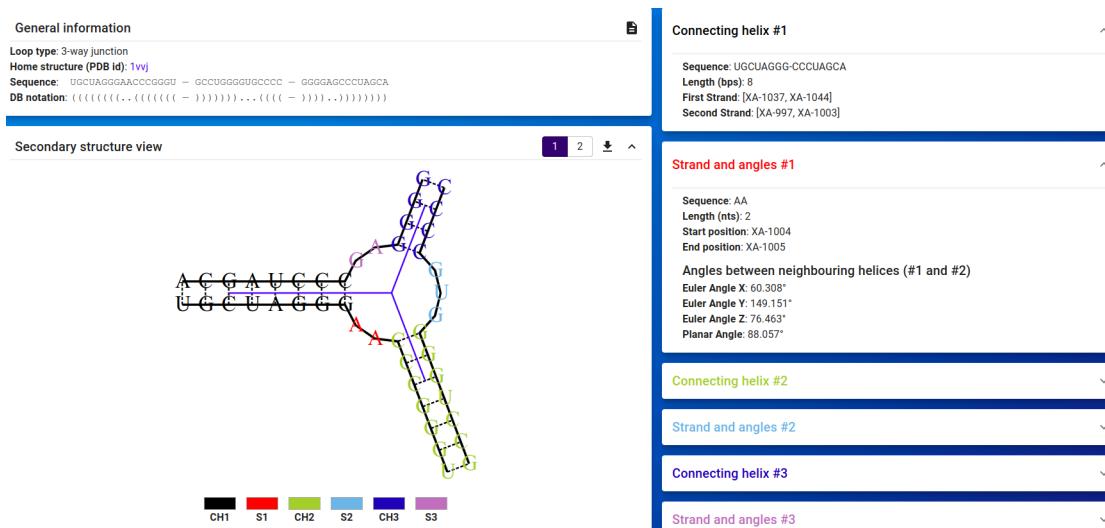


Abbildung 3.1: Beispieleintrag in der RNALoops Datenbank. Mittels PDF-Download können die rechts angegebenen Daten für alle Abschnitte heruntergeladen werden, zusammen mit den Daten oben über Sequenz und Home-Struktur.

Alle diese Informationen werden in einer Tabelle zusammengefasst, sodass typische Anwendungen zur Datenverarbeitung (Excel, Python/Pandas) unkompliziert damit arbeiten können. Zu jedem Multiloop kann zusätzlich die Sekundärstruktur als Grafik angezeigt werden. Damit sind alle wesentlichen Informationen der Datenbank aggregiert und vorbereitet. In einem nächsten Schritt werden noch einige Bereinigungen durchgeführt, so wurden beispielsweise etwa 3000 Strukturen gefunden, die in allen oben genannten Features übereinstimmen und damit als identisch angesehen werden. Diese wurden aus den Daten entfernt. Weiter gibt es ca. 2000 Strukturen, die zwar in der Datenbank sind, über die aber keine weiteren Informationen außer der ID und der Sequenz vorliegen, auch diese wurden entfernt, da hier keine Untersuchungen möglich sind.

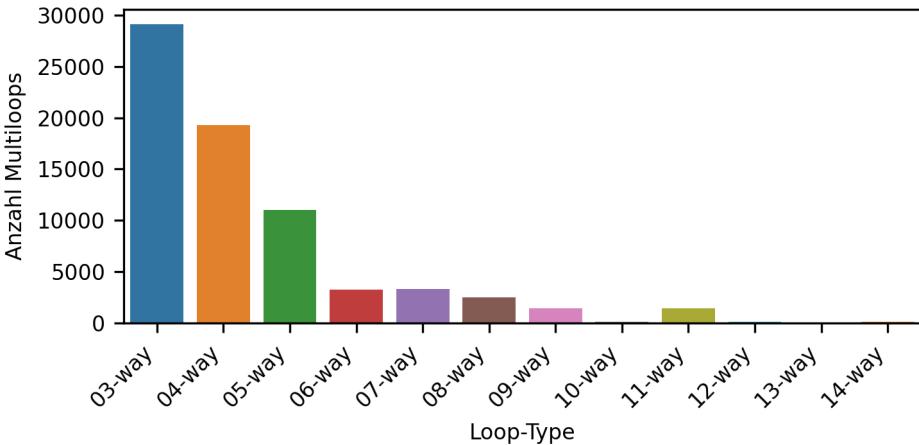


Abbildung 3.2: Zusammensetzung der RNALoops Datenbank nach Loop-Type. Dieser repräsentiert die Anzahl der Helices im jeweiligen Multiloop. Man sieht, dass die meisten Strukturen 3, 4 oder 5 Helices besitzen und sehr weniger mehr als 8.

Insgesamt erhält man nach diesem Prozess etwa 75.000 Multiloops mit 145 Features. Diese große Zahl an Features kommt daher zustande, dass es Multiloops mit 14 Strands und Helices gibt, die für jeden dieser Abschnitte Länge, Position und Winkel definieren. Üblicherweise besitzen die Multiloops aber wesentlich weniger Helices (oft nur 3 oder 4) und haben dementsprechend weniger Feature [Abb. 3.8]. Die meisten Analysen werden im Folgenden getrennt nach Loop-Typ durchgeführt, da die unterschiedliche Anzahl an Features einen Vergleich schwierig macht. Häufig werden auch nur Loop-Types bis 8 betrachtet, da darüber zu wenige Strukturen vorhanden sind um sinnvolle Vergleiche durchführen zu können.

Während die beschriebenen Schritte es ermöglichen die vollständigen Daten der Datenbank zu sammeln, zielt der nächste Abschnitt darauf ab diese Daten zu kontrollieren und gegebenenfalls Strukturen zu entfernen, wenn diese als problematisch eingestuft werden.

3.2 Kontrolle und Filtern der Daten

3.2.1 Qualität der Multiloops

Das Extrahieren von Multiloops aus der PDB ist nicht trivial. Zwar betragen die Strukturen selber häufig aus einer überschaubaren Anzahl an Basen (< 100), die Komplexe in die sie eingebettet sind können aber extrem groß sein (z.B. in Ribosomen). Da ein Multiloop nicht ein Abschnitt entlang der Sequenz ist, sondern an seinen Helices unterbrochen ist, können die Strands an sehr unterschiedlichen Positionen innerhalb der Home-Struktur liegen [Abb. 3.7]. Genauer besteht die Home-Struktur aus verschiedenen Ketten (chains), die durch diverse Bindungen verknüpft sein können. Die Ketten selber sind durchgängige Sequenzen von Basen (verbunden durch das Phosphat Rückgrat). Ein Multiloop kann sich über mehrere Ketten erstrecken, wenn die beiden Strände einer Helix in verschiedenen Ketten liegen, das kommt allerdings sehr selten vor. Ein Strand dagegen muss der Sequenz einer Kette folgen und kann nicht unterbrochen sein.

Diese Situation macht das Auffinden eines Multiloops in einer Home-Struktur kompliziert. Dazu kommt die Variabilität der Bindungen innerhalb der RNA. Es gibt Basenpaarungen verschiedener Stärken, die nicht alle gleich behandelt werden können. Insbesondere werden kanonische und nicht-kanonische Paarungen unterschieden. Erstere sind die “klassischen”

3 Analyse der Datenbank

C-G und A-U Bindungen, die am häufigsten und am stärksten sind. Daneben gibt es aber auch alle möglichen anderen Kombinationen von Paaren, die dann aber weniger Wasserstoffbrückenbindungen besitzen und daher schwächer sind.

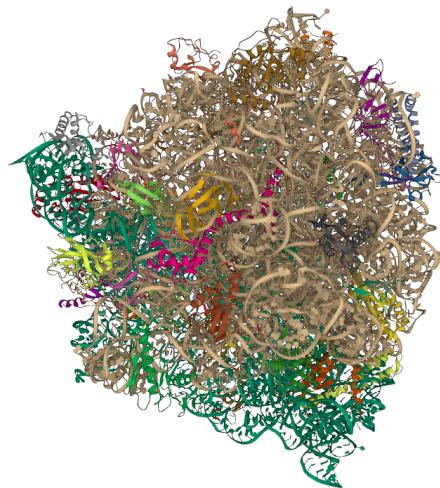


Abbildung 3.3: Beispiel für die dreidimensionale Struktur eines Ribosoms aus der PDB (ID 1VVJ). Irgendwo in diesem Komplex könnte ein (oder mehrere) Multiloops zu finden sein. Spezielle Algorithmen können diese aus der Gesamtsequenz extrahieren.

Für die Multiloops der RNALoops Datenbank ist die Festlegung so, dass Basen als ungepaart betrachtet werden, wenn sie nicht-kanonisch gepaart sind. Das heißt insbesondere solche Basen können (neben den echt-ungepaarten Basen) Teile der Strands bilden. Noch komplizierter ist die Situation bei den Helices. Hier gibt es praktisch keine einheitliche Konvention wann ein Basenpaar als Teil der Helix betrachtet wird. Kanonische Bindungen werden sicher dazu gezählt, bei allen weiteren ist es Ansichtssache, ob die Bindung noch zur Helix gezählt wird, oder ob die Helix dort endet, weil eine “ungepaarte” Base erreicht wurde.

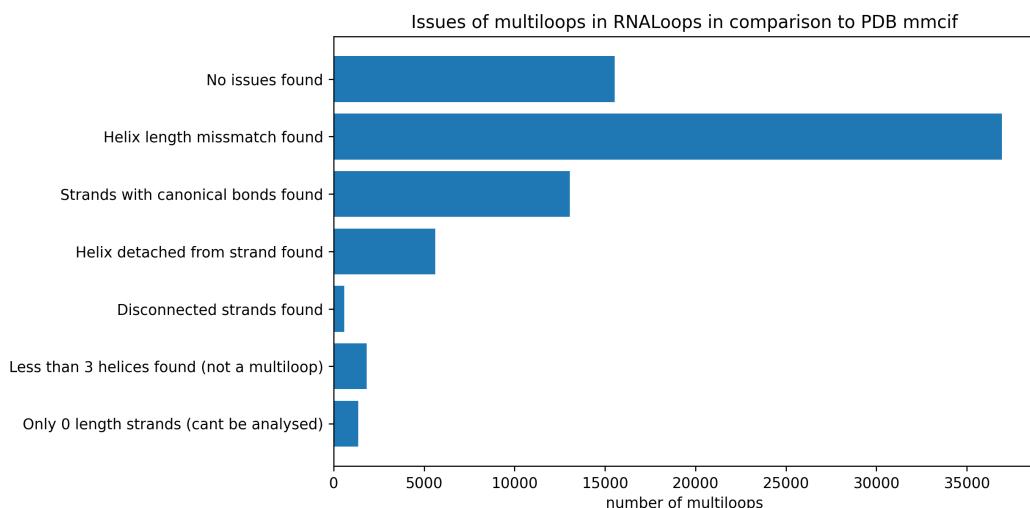


Abbildung 3.4: Multiloops sortiert nach Qualitätsstufe. Details siehe Text.

Um einen Überblick dieser Aspekte für die vorliegende Datenbank zu erhalten wurden die gegebenen Daten mit den “Originaldaten” aus der PDB verglichen. Hierzu wurden die mmcif Dateien [15] der Datenbank verwendet, die Informationen über die Bindungen aller

Basen in einer Home-Struktur bereitstellen. Hierzu sei erwähnt, dass diese Sekundärstruktur mittels eines Algorithmus aus den gemessenen Atompositionen gewonnen wird [2]. Eine häufig verwendete Implementierung stellt das Programm RNAView dar [16]. Auch dieser Algorithmus macht Annahmen, und daher sind die Daten der mmcif Dateien selber nicht ohne jeden Zweifel. Hier werden sie aber dennoch für einen Vergleich mit der RNALoops Datenbank verwendet. Dazu werden die Bindungen jener Basen extrahiert, die in den jeweiligen Multiloops der RNALoops Datenbank vorkommen. Es wird dann betrachtet, ob diese Bindungen kompatibel sind mit den Strukturen die in RNALoops präsentiert werden. Das Ergebnis ist in Abb. [Abb. 3.8] zu sehen.

Dabei werden die Kategorien von oben nach unten “*problematischer*” und ein Multiloop der in mehrere Kategorien fällt wird immer der problematischsten zugeordnet. Im Detail werden folgende Strukturen unterschieden:

1. “*No issues found*”: Diese Multiloops sind kompatibel mit den Daten der PDB. Etwa 20% der Strukturen fallen in diese Gruppe.
2. “*Helix length mismatch found*”: Hier wurde die Länge mindestens einer Helix zu einer anderen Länge bestimmt, als sie in RNALoops aufgeführt ist. Diese Abweichung ist in der Regel nicht problematisch, da der Loop an sich nicht betroffen ist und die Winkel ohnehin nur anhand der ersten Paare bestimmt werden. Die Tatsache, dass etwa die Hälfte der Strukturen in diese Kategorie fällt zeigt, dass wie oben diskutiert im Allgemeinen unklar ist, welche Basenpaare für die Helix berücksichtigt werden. Für den Vergleich wurden verschiedene Varianten getestet, aber es war nicht möglich die Längen aus RNALoops zu reproduzieren.
3. “*Strands with canonical bounds found*”: In diesen Strukturen gab es Basen in Strands, die kanonische Bindungen hatten, und damit eigentlich nicht in den Strand gehören. Es ist unklar, ob RNALoops einen Algorithmus verwendet, der hier tatsächlich keine kanonischen Basen vorhersagt oder ob es andere Gründe gibt, warum die Basen trotzdem für Strands verwendet wurden.
4. “*Helix detached from strand found*”: Hier wurden Helices gefunden, die nicht mit den umgebenen Strands verbunden waren. Das ist ein schwerwiegender Defekt, da hier die Winkel voraussichtlich nicht korrekt bestimmt werden können. Möglicherweise hängt diese Kategorie mit Pseudoknoten zusammen (s.u.). Auch hier ist nicht klar, warum diese Fehler auftreten.
5. “*Disconnected strands found*”: Hier wurden Strands gefunden, die in sich nicht vollständig verbunden sind (sich z.B. über zwei nicht verbundene Ketten erstrecken). In diesem Fall sollte die Struktur nicht als Multiloop identifiziert werden.
6. “*Less than 3 helices found (not a multiloop)*”: Hier wurden basierend auf den Bindungen der mmcif Datei keine drei Helices gefunden, sodass es sich eigentlich nicht um einen Multiloop handelt.
7. “*Only 0 length strands (cant be analysed)*”: Es gibt Multiloops, die nur aus Helices bestehen, da alle Strands die Länge 0 haben. In diesem Fall kann die Qualitätsanalyse nicht zuverlässig durchgeführt werden.

Um einschätzen zu können, ob es sich bei diesen Problemen tatsächlich um Fehler in der Extraktion der Multiloops aus der PDB handelt, oder es andere Ursachen gibt (Fehler in mmcif, Fehler in Qualitätskontrolle, ...) müsste weiter untersucht werden. Mangels Zeit wird an dieser Stelle aber stattdessen eine Filterung der Daten nach drei Leveln vorgeschlagen:

L1: Alle Daten aus RNALoops nach Entfernen von Duplikaten

L2: Zusätzliches Entfernen von Multiloops der Kategorien 5 und 6.

L3: Zusätzliches Entfernen von Multiloops der Kategorie 4.

In allen Anwendungen, in denen das Entfernen eines kleinen Teiles der Daten (<5%) kein Problem ist, wird L2 als sinnvoller Kompromiss eingeschätzt. Damit kann die Datenqualität erhöht werden ohne zu viel Verlust in Kauf nehmen zu müssen. Wie beschrieben wären weitere Untersuchungen notwendig, um die Datenqualität vollständig einschätzen zu können.

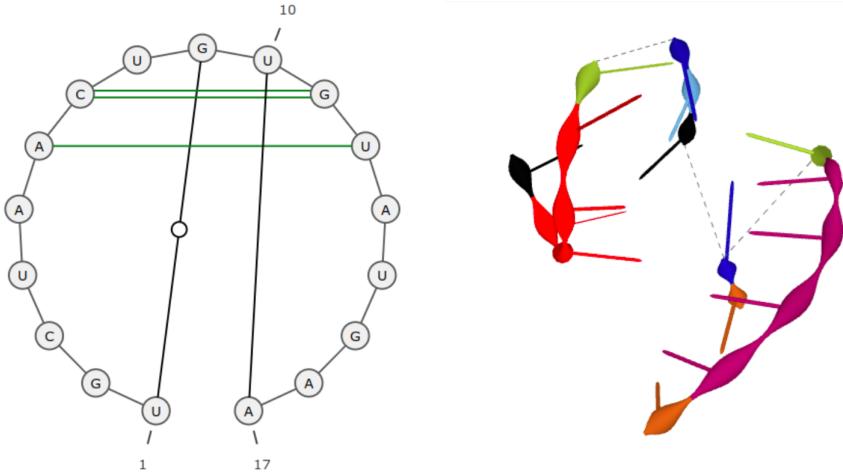


Abbildung 3.5: Sekundärstruktur (links) und dreidimensionale Konfiguration (rechts) eines Pseudoknotens. Man erkennt, dass es schwierig ist in dieser Struktur sinnvolle Winkel zwischen den Helices zu definieren, insbesondere da die Helices einzelne Teile des Multiloops miteinander verbinden (und damit unter Umständen doppelt gezählt werden).

An dieser Stelle sei noch erwähnt, dass die Analyse der Multiloops auch dadurch erschwert wird, dass Pseudoknoten in die Datenbank mit aufgenommen werden. Dabei handelt es sich um Multiloops, die nicht unmittelbar verbunden sind, sondern entfernte Bindungen eingehen [Abb. 3.9]. Die Analyse dieser Strukturen ist komplex und es ist nicht sicher, wie sinnvoll die beschriebenen Winkel bei solch auseinandergezogenen Multiloops bestimmt werden können.

3.2.2 Vergleich mit RNASTrand

RNASTrand ist eine etablierte Datenbank für RNA Strukturen [17]. Eine Möglichkeit die Qualität der Daten in RNALoops einzuschätzen könnte ein Vergleich mit dieser Datenquelle sein. Hierzu wurde zunächst die Schnittmenge der beiden Quellen im Bezug auf die verwendeten Home-Strukturen der PDB bestimmt [Abb. 3.10]. Leider ist die Schnittmenge sehr klein, sodass ein sinnvoller Vergleich nicht möglich ist.

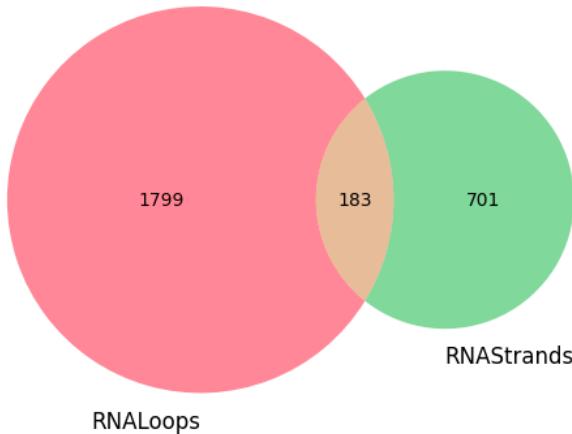


Abbildung 3.6: Schnittmenge von RNALoops und RNAstrand im Bezug auf Home-Strukturen in der PDB.

Darüber hinaus hat RNAstrand keine detaillierten Informationen über die Multiloops, sodass es sich nicht lohnt hier weitere Untersuchungen durchzuführen.

3.2.3 Heterogenität der Daten

Heterogenität der Daten ist wichtig im Bezug auf zwei Merkmale:

1. Organismus: Multiloops werden aus Home-Strukturen der PDB entnommen, die wiederum einem Organismus zugeordnet werden können. Es ist wichtig, dass die Daten von unterschiedlichen Organismen stammen, um evolutionäre Effekte auf die Formation der Strukturen einschätzen zu können.
2. Chain-label: Als chain-label wird im Folgenden das Label bezeichnet, das der Kette eines Multiloops zugeordnet wird, wenn dieser in der PDB eingetragen wird. Eine Home-Struktur besteht aus mehreren solcher Ketten, die wiederum in verschiedenen Strukturen auftauchen können. Bekannte Ketten sind die ribosomalen RNA-Stränge 16s, 23s, etc. oder diverse Ausprägungen der transfer RNA wie e-side oder a-side tRNA. Heterogenität im Bezug auf chain-label ist wichtig, da manche Ketten in diversen Home-Strukturen auftreten können, dadurch häufig in RNALoops erscheinen, tatsächlich aber praktisch identische Multiloops darstellen.

Leider ist in beiden Aspekten keine gute Heterogenität der Daten gegeben.

Auf wenige Organismen entfällt bereits ein großer Teil der Multiloops [Tab. 3.1]. Gleiches gilt für die Chain-label [Tab. 3.2].

Diese Verteilungen sind üblich in der Analyse von RNA Strukturen, da bestimmte Organismen und RNA-Typen sehr gut untersucht sind, andere dagegen kaum. Gründe dafür sind das Interesse an bestimmten Strukturen (z.B. von Homo Sapiens), die Einfachheit der Untersuchung (z.B. escherichia coli) und allgemein die Bedeutung der Struktur für die Biologie (z.B. ribosomale RNA). Durch diese Homogenität der Daten werden allgemein Analysen von RNA Strukturen erschwert.

Eine Möglichkeit die Datenlage etwas besser zu gestalten ist das Aggregieren von identischen Sequenzen. In RNALoops kommen viele Sequenzen mehrfach in Form von unterschiedlichen Multiloops vor. Das kann daran liegen, dass zufällig zwei Strukturen Multiloops mit der gleichen Sequenz ausbilden, wahrscheinlicher ist aber, dass es sich um ein und dieselbe Struktur handelt, die nur in mehreren Ausführungen in der PDB auftaucht. Um das zu berücksichtigen werden für alle folgenden Untersuchungen Multiloops der gleichen Sequenz aggregiert. Die Winkel werden dabei gemittelt (Median). So erhält man eine übersichtlichere und heterogenere Datenbasis [Tab. 3.3].

3 Analyse der Datenbank

Organismus	Multiloops	Home-Strukturen
thermus thermophilus hb8	27568	330
escherichia coli	16699	439
saccharomyces cerevisiae	9469	186
homo sapiens	2699	110
oryctolagus cuniculus	2673	48
haloarcula marismortui	2623	62
deinococcus radiodurans	935	45
Andere	7995	1335

Tabelle 3.1: Organismen mit den meisten Multiloops in RNALoops. Der viel studierte *thermus thermophilus* dominiert klar, gefolgt vom ebenfalls sehr häufig untersuchten *escherichia coli*. Auch RNA des Menschen und von Kaninchen findet sich häufig in der Datenbank. Insgesamt ist die Datenlage im Bezug auf den Organismus sehr stark auf wenige Vertreter fokussiert. Allerdings sieht man in der rechten Spalte, dass die anderen Organismen aus deutlich mehr verschiedenen Home-Strukturen stammen.

Chain-label	Multiloops	Home-Strukturen
23s ribosomal rna	33691	558
16s ribosomal rna	15514	571
25s ribosomal rna	5735	94
18s ribosomal rna	4863	177
28s ribosomal rna	2887	68
5s ribosomal rna	811	121
5.8s ribosomal rna	680	168
Andere	8291	1537

Tabelle 3.2: Chains mit den meisten Multiloops in RNALoops. Ribosomale RNA dominiert hier vollständig, insbesondere 23s und 16s kommen extrem oft vor. Wieder stammen die anderen Vertreter aus vielen verschiedenen Home-Strukturen, die aufgelisteten dagegen erneut aus verhältnismäßig wenigen.

3.3 Clustering der Strukturen nach planaren Winkeln

3.3.1 Winkel nach Loop-Typ

Im vorherigen Abschnitt wurde bereits diskutiert, dass die Datenmenge durch Aggregation verringert werden kann. Konkret reduziert sich die Zahl der Multiloops dadurch auf 7856. Diese wiederum können nach Loop-Type unterschieden werden, da Strukturen mit unterschiedlichen Loop-Typen nicht sinnvoll im Bezug auf die Winkel verglichen werden können. Weiter werden nur planare Winkel betrachtet, da diese wie diskutiert zuverlässiger sind.

Insgesamt sind die Winkel sehr homogen über den gesamten möglichen Wertebereich von 0° bis 180° verteilt [Abb. 3.7]. Allerdings zeigen sich gerade bei höheren Loop-Typen Cluster. Das Ziel dieses Abschnittes ist es solche Cluster zu identifizieren und abzuschätzen welche gemeinsamen Charakteristika die Multiloops mit ähnlichen Winkeln besitzen.

3.3 Clustering der Strukturen nach planaren Winkeln

Kategorie:	rRNA	tRNA	Rest
Multiloops:	65573	1989	2737
Multiloops mit verschiedener Sequenz ODER verschiedenem Host (z.B. 23s vs 16s):	7104	866	1437
Multiloops mit verschiedener Sequenz:	6396	600	1065

Tabelle 3.3: Einfluss von Aggregieren auf Verteilung der Multiloops. Die beiden häufigsten Arten von Strukturen sind rRNA und tRNA. Da diese sehr dominant sind wird die Analyse erschwert. Aggregieren gleicher Sequenzen reduziert die Zahl der rRNA Multiloops um ca. Faktor 10 und die anderen um ca. Faktor 3. So wird die Dominanz von rRNA reduziert und vermutlich doppelt aufgenommene Multiloops entfernt.

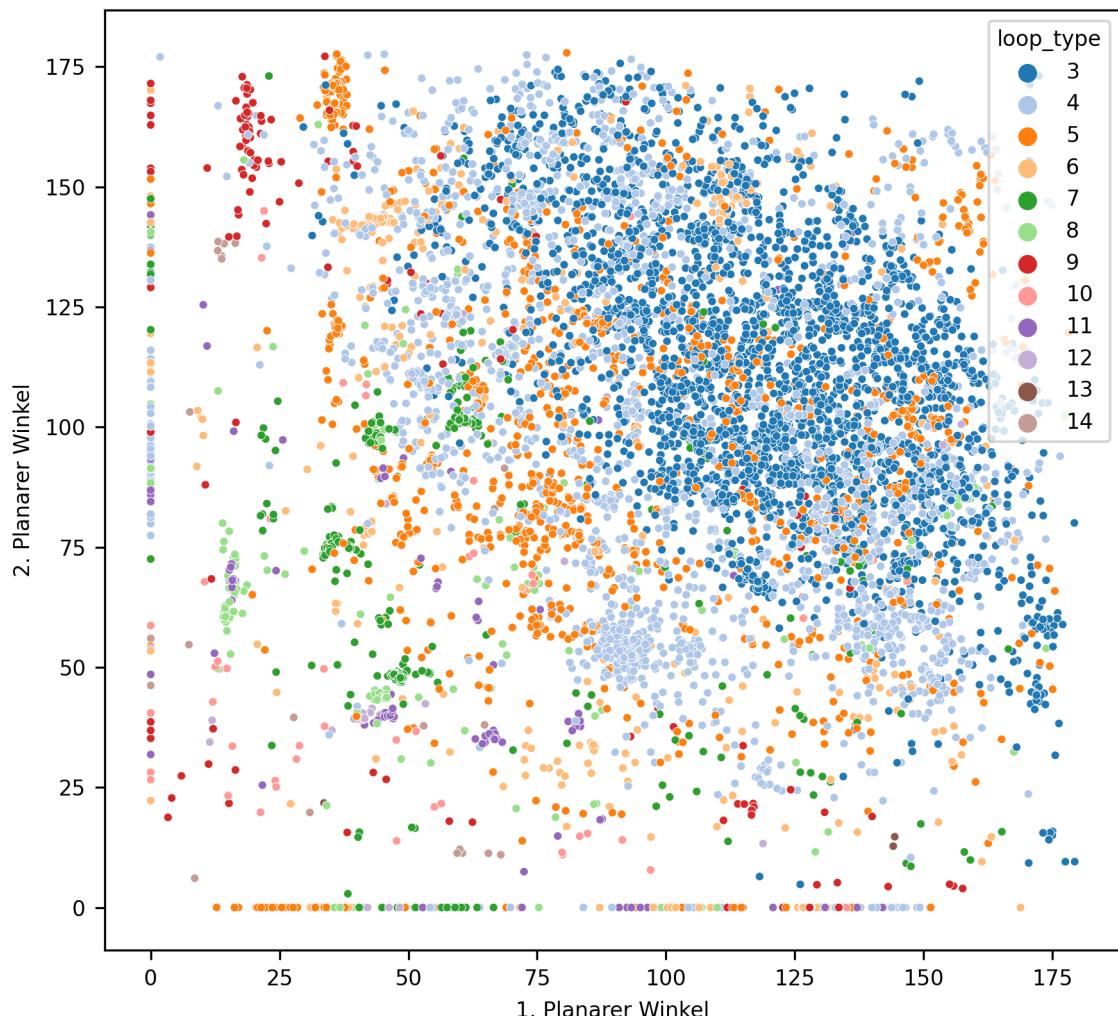


Abbildung 3.7: Verteilung der ersten beiden planaren Winkel nach Loop-Type. Für ein vollständiges Bild müssten auch noch die weiteren Winkel betrachtet werden, aber man erhält auch so bereits einen Eindruck über die Verteilung.

3.3.2 Detaillierte Übersichten von Winkel-Clustern

Für das Clustering werden verschiedene Verfahren der scikit-learn Bibliothek verwendet. Gute Ergebnisse zeigt ein hierarchisches CLustering mit ward-Abstand als Metrik. Dadurch können lokale Häufungen von Winkeln identifiziert werden. Da die Größe der Cluster sehr

3 Analyse der Datenbank

stark variiert, wird zudem ein Threshold gesetzt um Cluster zuschneiden zu können. So kann das hierarchische Clustern mit geringerer Zahl von Clustern durchgeführt werden und nachträglich die Anzahl der Punkte pro Cluster reduziert werden um eine höhere Dichte von Punkten zu erhalten. Als Maß für den Threshold wird der mittlere Abstand entlang aller relevanten Winkelachsen genommen und je nach Wahl z.B. bei 3° , 5° oder 8° abgeschnitten.

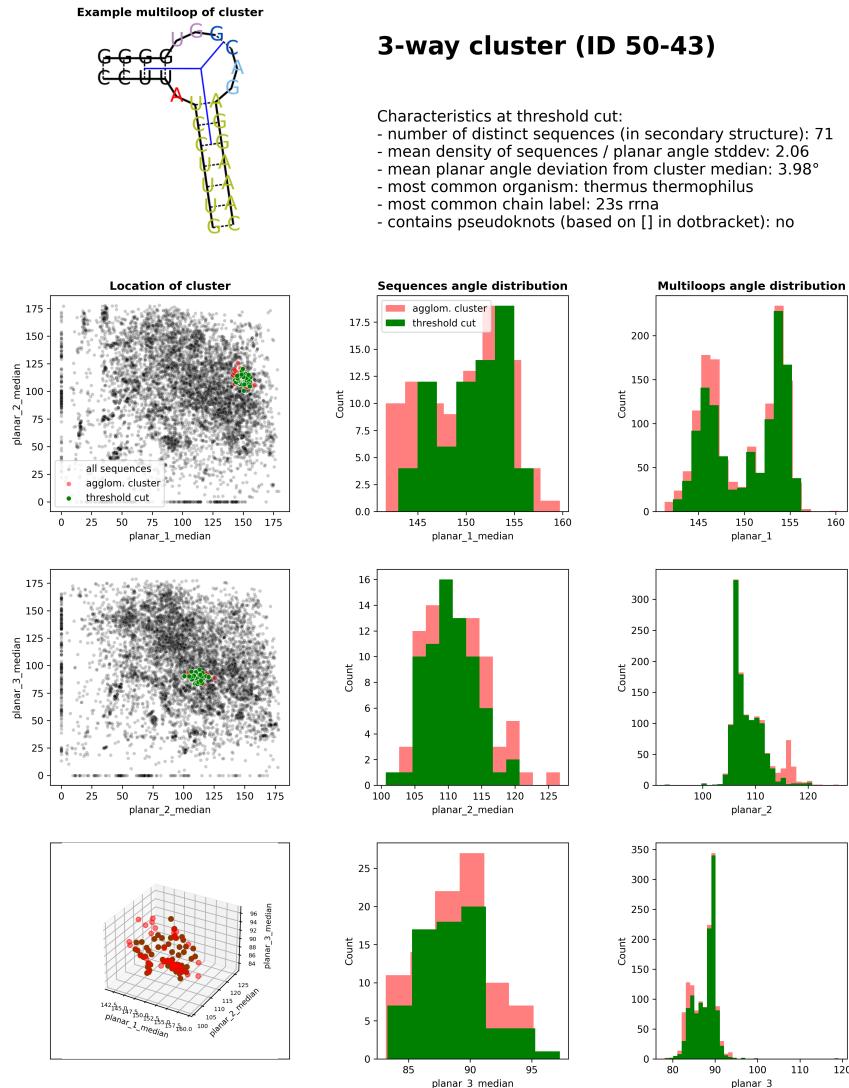


Abbildung 3.8: Erster Teil der Gesamtübersicht eines Clusters (Verteilung der Winkel).

Um ein möglichst umfassendes Bild zu erhalten werden alle Charakteristika eines Clusters in einer Übersicht zusammengefasst [Abb. 3.8], [Abb. 3.9], [Abb. 3.10]. So können eine Reihe von Fragen beantwortet werden, z.B.:

1. Stammen die Multiloops mit ähnlichen Winkeln alle von einem oder von sehr wenigen Organismen?
2. Stammen die Multiloops mit ähnlichen Winkeln alle von einem oder von sehr wenigen Chain-Labels?
3. Ist die Verteilung der Winkel innerhalb eines Clusters normalverteilt oder gibt es Besonderheiten?

3.3 Clustering der Strukturen nach planaren Winkeln

4. Wie nah beieinander liegen die Winkel innerhalb eines Clusters und können das Messfehler sein?
5. Haben Multiloops eines Clusters gleiche oder ähnliche Sequenzen, was sind Unterschiede?
6. Spielt die konkrete Verteilung der Basen (C, G, A, U) eine Rolle für die Winkel?

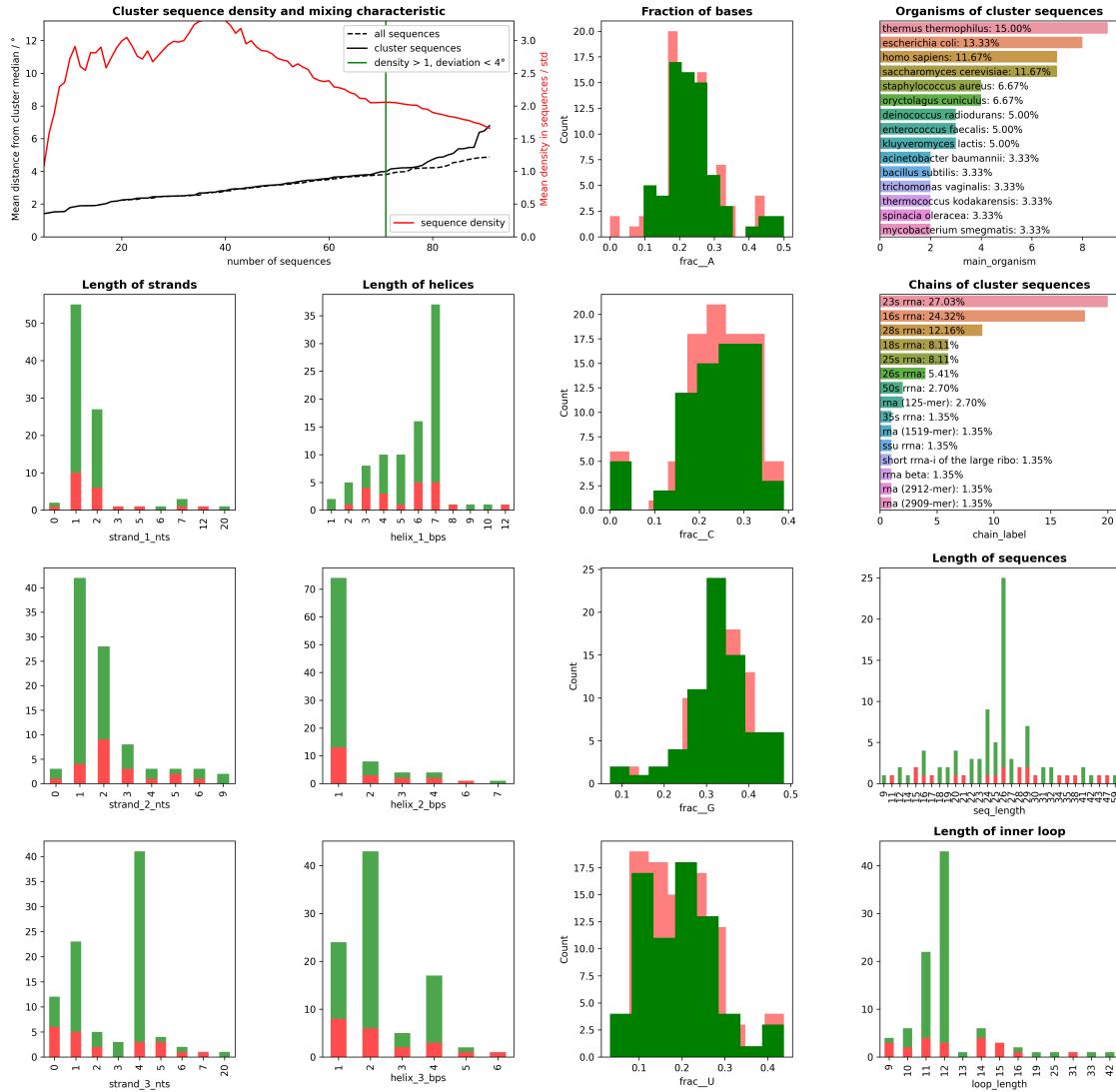


Abbildung 3.9: Zweiter Teil der Gesamtübersicht eines Clusters (Charakteristika).

Diese Übersichten wurden für alle Cluster angefertigt, sortiert nach Loop-Typ. Außerdem wurde sowohl der Threshold als auch die Anzahl der initialen Cluster variiert. Insgesamt erhält man so eine große Menge an Informationen über die Multiloops der RNALoops Datenbank, die in Zukunft noch weiter analysiert werden können. Der Fokus in diesem Projekt lag auf der Sortierung und Kontrolle der Daten, sowie der Aufbereitung für einen möglichst umfassenden Überblick.

Die vollständigen Ergebnisse liegen vor und können bei Interesse zur Verfügung gestellt werden. Einige allgemeine Ergebnisse sind die Folgenden:

- Weiterhin liegt eine starke Homogenität der Daten vor. Mit Abstand die meisten Cluster werden von ribosomaler RNA dominiert. Insbesondere 16s und 23s ist sehr stark vertreten. Es wurden auch Analysen durchgeführt ohne rRNA, allerdings ist die

3 Analyse der Datenbank

Datenlage dann so überschaubar, dass kaum eindeutige Cluster identifiziert werden können.

- Bei den Organismen gibt es auch einige, die sehr häufig vorkommen. Das ist allerdings nicht so problematisch wie bei den chain-labels und führt in der Regel nicht zu Problemen bei der Auswertung.
- Die Verteilung der Winkel innerhalb eines Clusters ist häufig annähernd normalverteilt, die Abweichungen liegen im Bereich von wenigen Grad. Messungenauigkeiten liegen in der gleichen Größenordnung, daher ist zu vermuten, dass die Unterschiede in diesen Multiloops nicht anhand von einfachen Merkmalen der Multiloops aufgelöst werden können.
- Sequenzen eines Clusters haben häufig sehr ähnliche Sequenzen, die sich nur in wenigen Basen unterscheiden. Vermutlich handelt es sich hier um gleiche chain-labels mit leichten Änderungen in der Sequenz, die z.B. durch evolutionäre Einflüsse verursacht wurden. Mit den bisherigen Methoden können solche Sequenzen nicht aggregiert werden, da nur exakt Gleiche zusammengefasst werden. Es kommt aber auch vor, dass sehr unterschiedliche Sequenzen ähnliche Winkel haben, was weiter zu untersuchen wäre.
- Die konkrete Verteilung der Basen scheint kaum eine Rolle für die Winkel zu spielen. Ob an einer Position C, G, A oder U steht ist nicht entscheidend, sondern eher die Struktur des Multiloops. Diese Hypothese kann allerdings anhand der Daten aufgrund ihrer Heterogenität auch nicht abschließend geklärt werden.

Parts sequences in Cluster (max 50 shown):

```

GGUUCCC-GGGAGCU|U|C-G|A|GU-GC|AUAA
GUUCCC-GGGAGC|U|C-G|A|G-C|AUAA
GGUUUCC-GGGAGCU|U|C-G|A|GU-GC|AGAA
GGUUGGC-GCCAGCU|U|A|I|G-C|AUAA
GGCGAGC-GCUUAGCU|U|C-G|A|AC-GU|AAAA
GAGC-GUC|U|C-G|A|G-C|AAAA
CGGCUCCGGG-CCCGGGGCCG|AGGGAGCGAGACCCGUCGCC|GCG-CGC|ACGGGG|CG...
GUUCCC-GGGAGC|U|C-G|A|GU-GC|AUAA
GUACAG-CUGUGC|U|A-U|AU|GU|AUAA
GGUUGGC-GCCAGCU|U|A-U|GU-GC|AUAA
GGUUCCC-GGGAGC|U|C-G|A|GU-GC|ACAA
GGUCAUC-GAUGACU|U|C-G|A|GU-GC|AAAA
GGCUAUC-GAUAGCU|U|C-G|A|GU-GC|AAAA
GUGIAGC-GCUACAU|U|C-G|A|AC-GU|AAAA
GGUUCCC-GGGAGC|U|C-G|A|GU-GC|AGAA
GGUUCCC-GGGAGC|U|C-g|A|GU-GC|AGAA
GCAUGUG-CACGUGC|U|A-C-G|G|A|U-A|GGA
GGUACACC-GGUUGGC|U|C-G|A|GC-GU|AGAA
GACAAGC-GCUUUGU|U|C-G|A|CU-GG|AAAA
GGCGAGC-GCUCGCU|U|C-G|A|GC-GC|AAAA
CGAGC-GCUCGCU|U|C-G|A|GC-GC|AAAA
GGUGCGC-GCGCGCU|U|C-G|A|GU-GC|AAAA
UCCUJUUG-CAAAGGA|G|A|C-G|GU|GU|G|A
GGUUCCC-GGGACCU|U|C-G|A|G-C|AUAA
UCCC-GGG|U|C-G|A|GU-GC|AGAA
UCCUUU-AAAGGA|G|A|C-G|GU|CCUU-GGGG|A
CACC-GGU|U|C-G|A|GC-GU|AGAA
GUCACC-GGUUGGC|U|C-G|A|GC-GU|AGAA
GGGGGGC-GCCCCCU|U|C-G|A|AC-GU|AAAA
UUU-GGA|G|U-A|I|AAAAU
CGGGU-GCCUG|GGGUGCC|CC-GG|GGAGCC|CUAG-CUAG|GGAAC
GGCAC-GGUUGC|U|A|U|G-C|AC
GGCAAGC-GCUUUGC|U|C-G|A|CC-GG|AAAA
GGCAGGC-GCCUGCU|U|C-G|A|GU-GC|AAAA
GGCAGC-GCUCG|U|C-G|A|GC-GC|AAAA
UU-AA|G|A-U|AU|A|U
GGGGC-GCCCC|U|C-G|A|AC-GU|AAAA
UCCU-AGGA|G|A|C-G|GU|CCUU-GGGG|A
GGGGGGC-GUCCCCU|U|C-G|A|AC-GU|AAAA
CCAAUA-AAAUGG|G|A|C-G|AG|UAC-GGU|G
UCCUJUUG-CAAAGGA|G|A|C-G|GU|CCUU-GGGG|A
C-G|CCAAGUC|C-G|AGGCCAGC|C-G|GACGGUGUGAGGCCGUAGC
GGGGGAG-GUCCCCU|U|C-G|A|AC-GU|AAAA
UCUCA-UGAGA|G|A|C-G|GU|CCUU-GGGG|G
GGUCCCC-GGGGGC|U|C-G|A|GC-GU|AGAA
AUCUUAG-CUAAGGU|G|A|C-G|GU|CCUU-GGGG|G
CCUUUAGGU-GCCAUAAGG|CGCUACC-GGUAGCG|UAAA|AG-CU|
UCCUU-AAGGA|G|A|C-G|GU|CCUU-GGGG|A
GU-AU|U|G|AA-U|AU|A|U|A
CCGUUAG-CUAACGG|G|A|C-G|GA|CCCC-GGGG|G

```

Abbildung 3.10: Dritter Teil der Gesamtübersicht eines Clusters (Sequenzen im Cluster).

3.3.3 Einfluss von Parametern auf Winkel

Abschließend wurde noch eine Analyse der Feature-Importance durchgeführt, um einschätzen zu können wie sehr einzelne Feature die Winkel beeinflussen [Abb. 3.11]. Hierbei gibt es aber eine Reihe von Problemen, insbesondere erneut die starke Dominanz weniger chain-labels und Organismen, sodass diese Untersuchung ebenfalls nicht vollständig verlässlich ist.

	way-3	way-4	way-5	way-6	way-7	way-8
loop_type -	27.34	41.47	41.94	45.03	38.52	39.51
strand_1_nts -	26.39	38.10	38.77	40.60	33.08	30.25
strand_2_nts -	22.33	30.61	27.97	29.59	18.92	11.73
strand_3_nts -	11.47	18.63	15.82	14.16	8.50	6.86
strand_4_nts -	11.47	8.34	7.15	6.16	5.18	4.67
strand_5_nts -	11.47	8.34	3.50	3.83	4.30	3.29
strand_6_nts -	11.47	8.34	3.50	2.09	3.43	2.52
strand_7_nts -	11.47	8.34	3.50	2.09	3.41	2.27
strand_8_nts -	11.47	8.34	3.50	2.09	3.41	2.19
parts_seq -	1.07	2.08	0.94	0.87	1.36	1.11
chain_label -	1.04	0.81	0.90	0.84	1.33	1.08
main_organism -	0.98	0.79	0.86	0.79	1.18	1.03

Abbildung 3.11: Einfluss unterschiedlicher Feature auf die Streuung der Winkel. Jede Spalte beschreibt einen Loop-Typ (3 bis 8). In der ersten Zeile werden Multiloops nur nach Loop-Typ aufgeteilt. Die mittlere Streuung (Standardabweichung) der Winkel aller Multiloops in diesen Gruppen ist in den Feldern der ersten Zeile eingetragen (in °). Teilt man anschließend die Multiloops weiter auf gemäß der Länge des ersten Strands verringert sich die mittlere Streuung in allen Gruppen, da diese jetzt nur noch Multiloops mit der gleichen ersten Strandlänge enthalten. Der Unterschied ist allerdings sehr gering, also hat dieses Feature eher einen geringen Einfluss auf die Winkel. Gleichermaßen gilt für den zweiten Strand. Erst wenn auch der dritte festgelegt wird können die Winkel genauer bestimmt werden (mehr als doppelt so gut), da jetzt die Struktur des Multiloops schon sehr gut definiert wird. Allerdings ist zu beachten, dass die geringere Streuung nicht zwangsläufig durch eine Korrelation der Winkel mit den Features kommen muss, sondern auch durch die sinkende Zahl an Multiloops pro Gruppe zustande kommen kann. Es sind daher weitere Untersuchungen notwendig, um den Einfluss der Feature zu studieren. In den unteren Zeilen sind die Multiloops nach Sequenz aufgeteilt, hier hat man nur noch Streuungen im Bereich von einem Grad, die auf Messfehler zurückgeführt werden könnten.

4 Zusammenfassung

In der vorliegenden Arbeit wurde die Datenbank RNALoops analysiert. Dazu wurden zunächst alle Daten heruntergeladen und so strukturiert, dass Programme zur Datenanalyse damit arbeiten können. Dann wurde die Datenqualität untersucht, insbesondere im Bezug auf Fehler beim Extrahieren der Multiloops aus der PDB und die Homogenität der Daten (Organsimus/Chains). Anschließend wurden die Multiloops gemäß ihrer Winkel geclustert und umfangreiche Informationen über die einzelnen Cluster präsentiert. Abschließend konnte der Einfluss verschiedener Feature auf die Winkel abgeschätzt werden.

In fast allen genannten Schritten ist die vorliegende Arbeit noch ausbaufähig. Beim Filtern der Daten sollte geklärt werden, warum Fehler in der Datenbank vorliegen und ob diese kritisch für die Analyse sind. Hierzu könnten die Autoren der Datenbank konsultiert werden oder noch weitere Datenquellen als Vergleich betrachtet werden. Im Bezug auf die Homogenität der Daten ist leider nicht viel möglich, da die Datenlage im Bezug auf RNA insgesamt nicht gut ist. Dennoch sollte die Entwicklung genau beachtet werden und jede mögliche zusätzliche Datenquelle für Multiloops berücksichtigt werden.

Bei der Analyse der Cluster ist mit den vorliegenden Daten vermutlich nicht mehr zu erreichen. Sie bietet einen umfassenden Überblick der Daten und erlaubt detaillierte Einblicke in die lokalen Verteilung der Winkel. Mehr Arbeit ist aber notwendig um den Einfluss der Feature besser zu bestimmen. Hierzu könnten diverse weitere Algorithmen eingesetzt werden und insgesamt die Verteilungen und Korrelationen dieser Feature genauer studiert werden. Auch wäre es möglich aus den Daten der PDB weitere Informationen über die Multiloops zu gewinnen und so anhand weiterer Feature die Winkelverteilungen besser zu verstehen. Es bleibt abschließend zu klären wie groß der Effekt von Messfehlern, sowie der Einfluss der gesamten Home-Struktur um den Multiloop herum ist.

Abschließend lässt sich sagen, dass die vorliegende Arbeit ihr Ziel der Analyse der RNALoops Datenbank erreicht hat. Die Daten liegen jetzt vollständig in einfach zu verarbeitender Form vor und sind bereinigt. Informationen über einzelne Multiloops und Cluster wurden umfassend dargestellt und können in vollständiger Form zur Verfügung gestellt werden. Der vollständige Quellcode für alle Analysen liegt ebenfalls für mögliche weitere Untersuchungen vor.

Bibliografie

- [1] M. Szikszai u. a., “Deep learning models for RNA secondary structure prediction (probably) do not generalize across families,” *Bioinformatics*, Jg. 38, Nr. 16, Y. Ponty, Hrsg., S. 3892–3899, Aug. 2022, ISSN: 1367-4803, 1460-2059. DOI: [10.1093/bioinformatics/btac415](https://doi.org/10.1093/bioinformatics/btac415).
- [2] S. Wuchty u. a., “Complete suboptimal folding of RNA and the stability of secondary structures,” *Biopolymers*, Jg. 49, Nr. 2, S. 145–165, Feb. 1999, ISSN: 0006-3525, 1097-0282. DOI: [10.1002/\(SICI\)1097-0282\(199902\)49:2<145::AID-BIP4>3.0.CO;2-G](https://doi.org/10.1002/(SICI)1097-0282(199902)49:2<145::AID-BIP4>3.0.CO;2-G).
- [3] N. B. Leontis u. a., “The building blocks and motifs of RNA architecture,” *Current Opinion in Structural Biology*, Jg. 16, Nr. 3, S. 279–287, Juni 2006, ISSN: 0959440X. DOI: [10.1016/j.sbi.2006.05.009](https://doi.org/10.1016/j.sbi.2006.05.009).
- [4] M. Popenda u. a., “Automated 3D structure composition for large RNAs,” *Nucleic Acids Research*, Jg. 40, Nr. 14, e112–e112, Aug. 2012, ISSN: 1362-4962, 0305-1048. DOI: [10.1093/nar/gks339](https://doi.org/10.1093/nar/gks339).
- [5] B. Fürtig u. a., “NMR Spectroscopy of RNA,” *ChemBioChem*, Jg. 4, Nr. 10, S. 936–962, Okt. 2003, ISSN: 14394227. DOI: [10.1002/cbic.200300700](https://doi.org/10.1002/cbic.200300700).
- [6] J. Lipfert und S. Doniach, “Small-Angle X-Ray Scattering from RNA, Proteins, and Protein Complexes,” *Annual Review of Biophysics and Biomolecular Structure*, Jg. 36, Nr. 1, S. 307–327, Juni 2007, ISSN: 1056-8700, 1545-4266. DOI: [10.1146/annurev.biophys.36.040306.132655](https://doi.org/10.1146/annurev.biophys.36.040306.132655).
- [7] L. T. Chow u. a., “A map of cytoplasmic RNA transcripts from lytic adenovirus type 2, determined by electron microscopy of RNA:DNA hybrids,” *Cell*, Jg. 11, Nr. 4, S. 819–836, Aug. 1977, ISSN: 00928674. DOI: [10.1016/0092-8674\(77\)90294-X](https://doi.org/10.1016/0092-8674(77)90294-X).
- [8] L. Fu u. a., “UFold: Fast and accurate RNA secondary structure prediction with deep learning,” *Nucleic Acids Research*, Jg. 50, Nr. 3, e14–e14, Feb. 2022, ISSN: 0305-1048, 1362-4962. DOI: [10.1093/nar/gkab1074](https://doi.org/10.1093/nar/gkab1074).
- [9] K. Sato u. a., “RNA secondary structure prediction using deep learning with thermodynamic integration,” *Nature Communications*, Jg. 12, Nr. 1, S. 941, Feb. 2021, ISSN: 2041-1723. DOI: [10.1038/s41467-021-21194-4](https://doi.org/10.1038/s41467-021-21194-4).
- [10] L. Wang u. a., “A novel end-to-end method to predict RNA secondary structure profile based on bidirectional LSTM and residual neural network,” *BMC Bioinformatics*, Jg. 22, Nr. 1, S. 169, Dez. 2021, ISSN: 1471-2105. DOI: [10.1186/s12859-021-04102-x](https://doi.org/10.1186/s12859-021-04102-x).
- [11] J. Wiedemann u. a., “RNALoops: A database of RNA multiloops,” *Bioinformatics*, Jg. 38, Nr. 17, P. Robinson, Hrsg., S. 4200–4205, Sep. 2022, ISSN: 1367-4803, 1460-2059. DOI: [10.1093/bioinformatics/btac484](https://doi.org/10.1093/bioinformatics/btac484).
- [12] J. L. Sussman u. a., “Protein Data Bank (PDB): Database of Three-Dimensional Structural Information of Biological Macromolecules,” *Acta Crystallographica Section D Biological Crystallography*, Jg. 54, Nr. 6, S. 1078–1084, Nov. 1998, ISSN: 0907-4449. DOI: [10.1107/S0907444998009378](https://doi.org/10.1107/S0907444998009378).

Bibliografie

- [13] T. Zok u. a., “RNAPdbee 2.0: Multifunctional tool for RNA structure annotation,” *Nucleic Acids Research*, Jg. 46, Nr. W1, W30–W35, Juli 2018, ISSN: 0305-1048, 1362-4962. DOI: [10.1093/nar/gky314](https://doi.org/10.1093/nar/gky314).
- [14] M. Antczak u. a., “RNAPdbee—a webserver to derive secondary structures from pdb files of knotted and unknotted RNAs,” *Nucleic Acids Research*, Jg. 42, Nr. W1, W368–W372, Juli 2014, ISSN: 1362-4962, 0305-1048. DOI: [10.1093/nar/gku330](https://doi.org/10.1093/nar/gku330).
- [15] P. E. Bourne u. a., “The Macromolecular Crystallographic Information File (mmCIF),”
- [16] H. Yang, “Tools for the automatic identification and classification of RNA base pairs,” *Nucleic Acids Research*, Jg. 31, Nr. 13, S. 3450–3460, Juli 2003, ISSN: 1362-4962. DOI: [10.1093/nar/gkg529](https://doi.org/10.1093/nar/gkg529).
- [17] M. Andronescu u. a., “RNA STRAND: The RNA Secondary Structure and Statistical Analysis Database,” *BMC Bioinformatics*, Jg. 9, Nr. 1, S. 340, Dez. 2008, ISSN: 1471-2105. DOI: [10.1186/1471-2105-9-340](https://doi.org/10.1186/1471-2105-9-340).