

Kathrin Hädicke

Statistische Analyse der chemischen Verschiebungen von
RNA-Molekülen

eingereicht als

Bachelorarbeit

an der

HOCHSCHULE MITTWEIDA

UNIVERSITY OF APPLIED SCIENCES

Fakultät Mathematik/Naturwissenschaften/Informatik

Mittweida, August 2010

Erstprüfer:

Zweitprüfer:

Vorgelegte Arbeit wurde verteidigt am:

Bibliografische Beschreibung

Hädicke, Kathrin:

Statistische Analyse der chemischen Verschiebungen von RNA-Molekülen. – 2010. – 95 S.

Mittweida, Hochschule Mittweida – University of Applied Sciences,

Fakultät Mathematik/Naturwissenschaften/Informatik,

Bachelorarbeit, 2010

Referat

Ziel der Bachelorarbeit ist es, die chemischen Verschiebungen statistisch auszuwerten und die Parameter der individuellen Statistik für die mögliche Zuordnung der Signale von RNA Molekülen zu verwenden, die noch strukturell unaufgeklärt sind. Parallel dazu wird versucht diese Informationen zu nutzen, um die jeweilige Sekundärstruktur einzelner RNA Abschnitte zu bestimmen. Anhand einer unbekanntes RNA soll die aufgestellte Statistik ausgewertet und überprüft werden. In späteren Projekten ist geplant auf der Grundlage dieser Statistik eine neuartige Zuordnungsstrategie zu entwickeln und in einen Algorithmus und somit in eine Software umzusetzen.

Motivation

„Die Therapie mit RNA-Molekülen – ob Antisense-RNA, Aptamere, Ribozyme oder siRNA – sollte man keineswegs abschreiben. Im Gegenteil:

Das Potenzial ist gewaltig.“ [1]

Die Strukturaufklärung von Biomolekülen wie Ribonukleinsäuren ist eines der bedeutendsten Ziele - nicht nur in der Erforschung und Entwicklung von Arzneistoffen. Im Rahmen mehrerer klinischer Studien wurde das riesige Leistungsvermögen von RNA-Molekülen entdeckt. Sie „erkennen ihr Zielmolekül eindeutig und docken nicht an „falschen“ Strukturen an“ [1].

Zur Aufklärung der Struktur von RNA-Molekülen dient neben der Röntgenstrukturanalyse vor allem die sogenannte NMR-Spektroskopie (Nuclear Magnetic Resonanz).

Danksagung

Diese Bachelorarbeit entstand an der Hochschule Mittweida. Mein Dank gilt in erster Linie Prof. Dr. rer. nat. Dirk Labudde. Bei ihm möchte ich mich für seinen Themenvorschlag, seine wissenschaftliche Betreuung der Arbeit und seine Ermutigungen bedanken. Mein Dank für Hilfeleistungen zu informatischen Fragestellungen und zur Programmierung gelten meinem Betreuer Dipl.-Inf. (FH) Daniel Stockmann und Stefan Schildbach. Desweiteren gilt ein besonderer Dank den Mitgliedern der Arbeitsgruppe, die stets ein offenes Ohr hatten und für das angenehme Arbeitsklima im Büro sorgten.

Inhaltsverzeichnis

Abbildungsverzeichnis.....	VI
Tabellenverzeichnis.....	VIII
Formelzeichen und Abkürzungen	IX
1 Einleitung	1
2 Grundlagen	3
2.1 RNA.....	3
2.1.1 Bausteine der RNA.....	3
2.1.2 Struktur der RNA	5
2.1.3 Arten der RNA.....	8
2.1.4 Biologische Bedeutung der RNA	10
2.2 NMR-Spektroskopie.....	11
2.2.1 Theoretische Grundlagen	11
2.2.2 Eindimensionale und Zweidimensionale NMR-Experimente	12
2.3.3 TOCSY-Experiment.....	15
2.3.4 NOESY-Experiment.....	16
2.3.5 HSQC-Experiment.....	17
2.3 Chemische Verschiebungen	18
2.3.1 Chemische Verschiebungen innerhalb eines Nukleotids.....	20
2.3.2 Chemische Verschiebungen sequentiell mit aufeinanderfolgenden Nukleotiden...21	
2.3.3. Einflüsse auf chemische Verschiebungen	22
2.4 Zweidimensionale Normalverteilung.....	23
3 Data-mining.....	25
3.1 Datenbanken	25
3.1.1 PDB.....	25
3.1.2 BMRB.....	27
3.2 Datensatz.....	28
3.3 Vorgehensweise bei der Datenanalyse	30
3.4 WCStats	32
3.4.1 Einlesen der Dateien	32
3.4.2 Aufteilung nach Kriterien.....	32
3.4.3 Visualisierung der Datenpunkte.....	33
3.4.4 Erstellen der TOCSY-, NOESY- und HSQC-Distributionsdatei	35
3.4.5 Implementieren der Wahrscheinlichkeitsformel zum Auswerten der Statistik	36
4 Resultate und Diskussion	37

Inhaltsverzeichnis

4.1 Kovarianzellipsen.....	37
4.2 Statistik der ¹ H- und ¹³ C-chemischen Verschiebungen in einem Nukleotid.....	39
4.2.1 2D-Statistik.....	39
4.2.2 Abhängigkeit der chemischen Verschiebung in der Hauptbase	45
4.2.2 Statistik von chemischen Verschiebungen mit allen Variationen.....	49
4.2.3 5'GG-Start und 3'CC-Ende	52
4.2.4 Unterscheidung der Datengruppen nach ihrer Sekundärstruktur	55
4.3 Statistik der ¹ H- chemischen Verschiebungen zwischen zwei Nukleotiden.....	63
4.3.1 Kreuzsignal mit dem Vorgänger.....	63
4.3.2 Kreuzsignal mit dem Nachfolger.....	66
4.3.2 Unterscheidung der Datengruppen nach ihrer Sekundärstruktur	69
4.4 Statistische Auswertung von TOCSY-, NOESY- und HSQC-Experimenten anhand der RNA FZL4	72
5 Ausblick.....	76
5.1 Data-mining	76
5.2 Erweiterung der Statistiken für die Experimente.....	76
5.3 Prüfen der Statistik.....	77
5.4 RASP	77
5.5 Zuordnungsstrategie	79
Literaturverzeichnis	80
Selbstständigkeitserklärung.....	83
Anhang.....	84

Abbildungsverzeichnis

Abbildung 1:Das zentrale Dogma der Molekularbiologie von F. Crick.....	1
Abbildung 2:Aufbau eines Nukleotids am Beispiel eines ATP.....	3
Abbildung 3:Pentosen von RNA und DNA.....	4
Abbildung 4:Natürliche in DNA und RNA vorkommende Nukleinbasen.....	5
Abbildung 5:RNA-Molekülstrukturinformationsebenen anhand der RNA 3PHP (pdb).....	6
Abbildung 6:RNA-Sekundärstrukturmotive.....	7
Abbildung 7: mögliche Watson-Crick-Paarungen.....	8
Abbildung 8:Typische Kleeblattstruktur einer aus 76 Nukleotiden bestehenden phenylalaninübertragenden tRNA der Hefe	9
Abbildung 9:Genexpression bei Eukaryonten.....	10
Abbildung 10:Strukturformel von TMS	12
Abbildung 11:links: Eindimensionales NMR-Spektrum, rechts: Aufbau eines eindimensionalen NMR- Experiments. [10].....	12
Abbildung 12:oben: Zweidimensionales NMR-Spektrum, unten: Aufbau eines zweidimensionalen NMR- Experiments.....	13
Abbildung 13:Entstehung eines 2D-NMR-Spektrums durch zweidimensionale Fourier- Transformation.....	14
Abbildung 14:Das ^1H - ^1H -TOCSY-Spektrum der RNA TASL2 [22].....	15
Abbildung 15: ^1H - ^1H -NOESY-Spektrum erfasst in Deuteriumoxid.....	16
Abbildung 16:Typische intra- und interresiduale Signale in einer RNA-Helix.....	16
Abbildung 17:Schematische Darstellung der RNA 1pjj (pdb).....	19
Abbildung 18:Chemischen Verschiebungen im Cytosin und im Cytidin.....	20
Abbildung 19:Chemischen Verschiebungen im Adenin und im Adenosin.....	20
Abbildung 20:Kreuzsignal zwischen der chemische Verschiebung des H8-Atoms von Adenin mit der chemischen Verschiebung des H1'-Atoms vom Zucker des Vorgängers	21
Abbildung 21:Kreuzsignal zwischen der chemische Verschiebung vom H6-Atom des Cytosin oder Uracil mit der chemischen Verschiebung vom H5-Atom des nachfolgenden Cytosin oder Uracil	21
Abbildung 22:Jährliches Wachstum von RNA-Strukturen in der Proteindatenbank.....	26
Abbildung 23:FASTA file von 1A60.....	26
Abbildung 24:Ausschnitt aus dem STAR File der RNA 1A60.....	27
Abbildung 25:Workflow.....	31
Abbildung 26:Tabstop-getrennte Textdatei.....	32
Abbildung 27:Erstelltes Streudiagramm.....	34

Abbildungsverzeichnis

Abbildung 28:Aufbau der Distributionsdatei für das TOCSY-Experiment	35
Abbildung 29:Aufbau einer Statistik-Datei der RNA FZL4 für das TOCSY-Experiment.	36
Abbildung 30:Streudiagramm	37
Abbildung 31:Verschiedene Streudiagramme mit verschiedenen Clustergrößen und Überschneidungsflächen	38
Abbildung 32:Streudiagramm.	40
Abbildung 33:Streudiagramm	40
Abbildung 34:Streudiagramm	42
Abbildung 35:Streudiagramm	42
Abbildung 36:Streudiagramm	43
Abbildung 37:Streudiagramm	43
Abbildung 38:Streudiagramm	44
Abbildung 39:Streudiagramm	44
Abbildung 40:Streudiagramme	45
Abbildung 41:Streudiagramme	46
Abbildung 42:Streudiagramme	47
Abbildung 43:Streudiagramme	48
Abbildung 44:Streudiagramm	50
Abbildung 45:Streudiagramm	50
Abbildung 46:Streudiagramm	51
Abbildung 47:Streudiagramm.	51
Abbildung 48:Streudiagramm	52
Abbildung 49:Streudiagramme	53
Abbildung 50:Streudiagramme	54
Abbildung 51:Watson-Crick-Basen versus non-Watson-Crick-Basen	56
Abbildung 52:Watson-Crick-Basen versus non-Watson-Crick-Basen	57
Abbildung 53:Streudiagramm	59
Abbildung 54:Streudiagramm	59
Abbildung 55:Streudiagramm	60
Abbildung 56:Streudiagramm	60
Abbildung 57:Streudiagramm	61
Abbildung 58:Streudiagramm	62
Abbildung 59:Streudiagramm	62
Abbildung 60:Streudiagramm	64
Abbildung 61:Streudiagramm	65
Abbildung 62:Streudiagramm.	65
Abbildung 63:Streudiagramm	67

Abbildungsverzeichnis

Abbildung 64:Streudiagramm	67
Abbildung 65:Streudiagramm	68
Abbildung 66:Streudiagramm	68
Abbildung 67:Streudiagramm	70
Abbildung 68:Streudiagramm	70
Abbildung 69:Streudiagramm	71
Abbildung 70:Streudiagramm	71
Abbildung 71:Sekundärstruktur der RNA FZL4.....	72
Abbildung 72:Prozentuale Verteilung der Signale an den ersten fünf Stellen.....	74
Abbildung 73:Sekundärstruktur der RNA TASL1.	76
Abbildung 75:Bindungen innerhalb der Sekundärstruktur eines RNA-stems.	78
Abbildung 74:H5-H6 Diagramm der RNA 1A60.	78

Tabellenverzeichnis

Tabelle 1:Informationen anhand des Basentriplets GAC	29
Tabelle 2:PDB und BMRB ID der für die Statistik verwendeten reinen RNA-Strukturen.....	84

Formelzeichen und Abkürzungen

δ	chemische Verschiebung in ppm
ω	Lamorfrequenz
ρ	Korrelationskoeffizient
μ	Mittelwert
σ	Standardabweichung
A	Adenin
Å	Ångström
AMP	Adenosinmonophosphat
ATP	Adenosintriphosphat
BMRB	Biological Magnetic Resonance Data Bank
C	Cytosin
CMP	Cytidinmonophosphat
DNA	Desoxyribonucleic Acid (Desoxyribonukleinsäure)
G	Guanin
GMP	Guanosinmonophosphat
hnRNA	heterogene nukleare RNA
HSQC	heteronuclear single quantum coherence (Heteronukleare Einzelquantumkohärenz)
I	Kernspin-Quantenzahl
miRNA	microsRNA
mRNA	messenger RNA
NMR	Nuclear Magnetic Resonance (Kernspinresonanz)
NOESY	nuclear overhauser enhancement spectroscopy (Kern-Overhauser-Effekt-Spektroskopie)
nWC	non Watson Crick
PDB	Protein Data Bank
ppm	parts per million
Pu	Purin
Py	Pyrimidin
RNA	Ribonucleic Acid (Ribonukleinsäure)
RNAi	RNA-Interferenz
rRNA	ribosomal RNA
siRNA	small-interfering RNA
snRNA	small nuclear RNA
T	Thymin

Formelzeichen und Abkürzungen

TMS	Tetramethylsilan
tnRNA	tiny-noncodingRNA
TOCSY	total correlated spectroscopy (Vollständige Korrelationsspektroskopie)
tRNA	transfer RNA
U	Uracil
UMP	Uridinmonophosphat)
WC	Watson-Crick
wwPDB	Worldwide Protein Data Bank

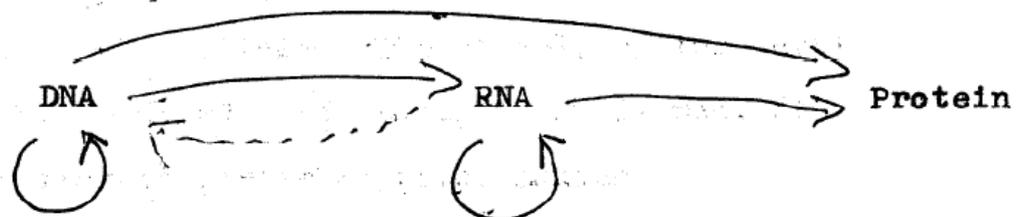
1 Einleitung

Die Geschichte der RNA (Ribonucleic Acid) begann in den 50er Jahren des 20. Jahrhunderts. Francis Crick erläuterte seine Vorstellungen über die Proteinsynthese (Abbildung 1) und postulierte später sein zentrales Dogma der Molekularbiologie. Schon damals erkannte er die außergewöhnliche Funktion der RNA – die Übertragung der genetischen Information von der DNA zum Protein.

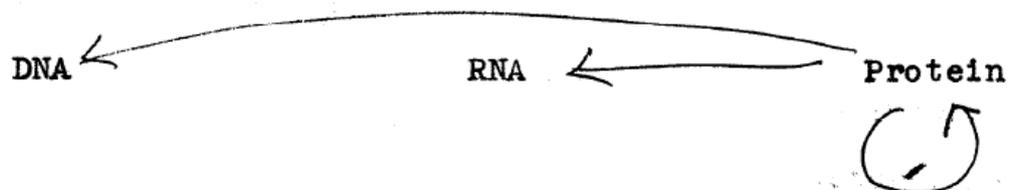
The Doctrine of the Triad.

The Central Dogma: "Once information has got into a protein it can't get out again". Information here means the sequence of the amino acid residues, or other sequences related to it.

That is, we may be able to have



but never



where the arrows show the transfer of information.

Abbildung 1: Das zentrale Dogma der Molekularbiologie von F. Crick
[http://profiles.nlm.nih.gov/SC/B/B/F/T/_/scbbft.pdf]

Heute weiß man, dass die RNA viele unzählige Funktionen aufweist. Sie fungiert nicht nur als Informations- und Transportträger zur Genregulation, sondern besitzt auch als Ribozym die Fähigkeit Reaktionen zu katalysieren. Aus dieser Fülle an Eigenschaften und Funktionen drängt sich folgende Frage auf: Sind RNA die Vorläufer von Organismen?

Eines steht jedoch fest, durch die Fähigkeit der RNA als Speicher, Überträger und Vervielfältiger der genetischen Information zu fungieren, rückt sie immer mehr in den Mittelpunkt des wissenschaftlichen Interesses.

1 Einleitung

Die NMR-Spektroskopie (Nuclear Magnetic Resonanz) ist neben der Röntgenstrukturanalyse die wichtigste spektroskopische Methode zur Strukturaufklärung von Biomolekülen wie Ribonukleinsäuren. Sie wurde 1946 von Felix Bloch und Edward Mills Purcell entwickelt und später durch Richard R. Ernst mit der Einführung der Fourier-Transformation grundlegend verbessert. Die NMR-Spektroskopie basiert auf dem Kernmagnetismus, der durch das magnetische Moment eines Atomkerns bestimmt wird. Nahezu alle Atomkerne besitzen einen Kernspin. Ist dieser von Null verschieden, zeigen sie ein magnetisches Moment und können somit in einem äußeren Magnetfeld diskrete Zustände annehmen. Aus den daraus resultierenden Wechselwirkungen ist die Resonanzfrequenz eines Kerns gegeben. Die Auswertung der NMR-Experimente beinhaltet das Zuordnen der einzelnen Signale des entstehenden Spektrums zu den Atomkernen des Moleküls. Dies ist eine schwierige und zeitintensive Aufgabe, die viel Handarbeit und spezielles Wissen erfordert. [15], [16], [13]

2 Grundlagen

In diesem Kapitel werden die wesentlichsten Grundlagen und Schwerpunkte zu den Themen NMR-Spektroskopie, chemische Verschiebungen, Ribonukleinsäure als auch mathematische Voraussetzungen zur zweidimensionalen Normalverteilung erläutert.

2.1 RNA

Die Ribonukleinsäure, kurz RNS (engl.: RNA) ist ein Makromolekül welches aus einer Kette von mehreren Nukleotiden besteht und infolgedessen als Polynukleotid bezeichnet wird.

Dieses Kapitel dient der Erläuterung biologisch-chemischer Grundlagen von RNA-Molekülen.

2.1.1 Bausteine der RNA

Ribonukleinsäuren sind im Grundaufbau den Desoxyribonukleinsäuren sehr ähnlich. RNA bestehen ebenfalls aus Pyrimidin- und Purinbasen, einem Zuckermolekül und einem Phosphatrest. Diese drei Bausteine bilden das Nukleotid (Abbildung 2), wobei mehrere dieser Nukleotide die RNA-Kette erzeugen. Neben dem Nukleotid gibt es noch ein Nukleosid, welches aus der Nukleinbase und der Pentose gebildet wird. [4], [13]

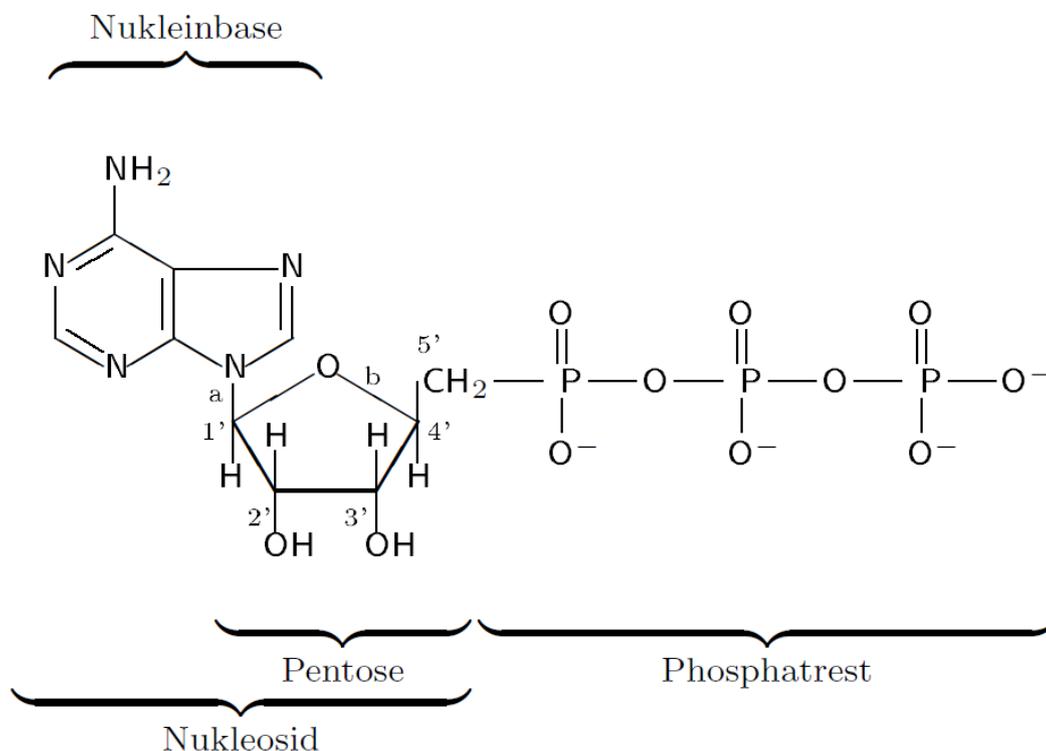


Abbildung 2: Aufbau eines Nukleotids am Beispiel eines ATP. Es besteht aus einer Nukleinbase, einer Pentose und einem Phosphatrest. [13]

Phosphatgruppe

Phosphatgruppen haben die Aufgabe einzelne Nucleoside der RNA durch Esterbindungen miteinander zu kettenförmigen, einzelsträngigen Polynucleotiden zu verknüpfen.

Zur Synthese der RNA müssen Nucleotidtriphosphate vorliegen, welche die Polymerase bei der Kettenbildung zu Diphosphaten spaltet. Dabei wird Energie für die Polymerisation freigesetzt. [13]

Pentose

Der in RNA verwendete Zucker ist eine Ribose, welcher sich durch die 2' OH-Gruppe von der Desoxyribose unterscheidet (Abbildung 3). Sie liegt in Furanoseform bestehend aus vier Kohlenstoff- und einem Sauerstoffatom vor. [13]

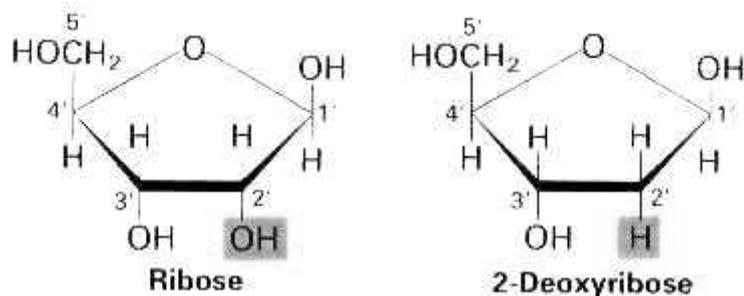


Abbildung 3: Pentosen von RNA und DNA. Sie unterscheiden sich im Vorhandensein von einer OH-Gruppe am C2. [<http://ludwig-sun1.unil.ch/~vjongene/molbio/pics/ribose.jpg>], Stand 30.11.2009

Nucleinbasen

Adenin, Guanin, Uracil und Cytosin sind die Nucleinbasen der Ribonucleinsäure und werden meist mit A für Adenin, G für Guanin, C für Cytosin und U für Uracil abgekürzt. Gemeinsam mit der Pentose bilden sie vier unterschiedliche Nucleoside, das Adenosin, Guanosin, Cytidin und Uridin. Neben diesen Grundformen existierenden weitere zahlreiche, durch posttranskriptionale Modifizierung, entstehende Abwandlungen. Diese sind größtenteils in den Nucleosiden von tRNAs und rRNAs vorzufinden. Sie dienen unter anderem zur Feinregulierung der Struktur sowie zur Einstellung von Aktivität und Spezifität der Moleküle. [13], [24]

Bindet sich ein Phosphat an das Nucleosid entsteht je nach Nucleinbase ein Nucleotid mit dem Namen AMP (Adenosinmonophosphat), GMP (Guanosinmonophosphat), CMP (Cytidinmonophosphat) oder UMP (Uridinmonophosphat). [24]

Chemisch betrachtet kann man die Basen in zwei Gruppen einteilen, den Purinbasen Adenin und Guanin sowie den Pyrimidinbasen Cytosin und Uracil. Beide gehören zur Klasse der

2 Grundlagen

Stickstoffbasen, wobei Purine aus einem Doppelring und Pyrimidine aus einem einfachen Ring aufgebaut sind. [24]

Die einzige der vier Nukleinbasen, die sich von den Basen der DNA unterscheidet ist Uracil welches in der RNA anstelle von Thymin vorkommt (Abbildung 4). [13]

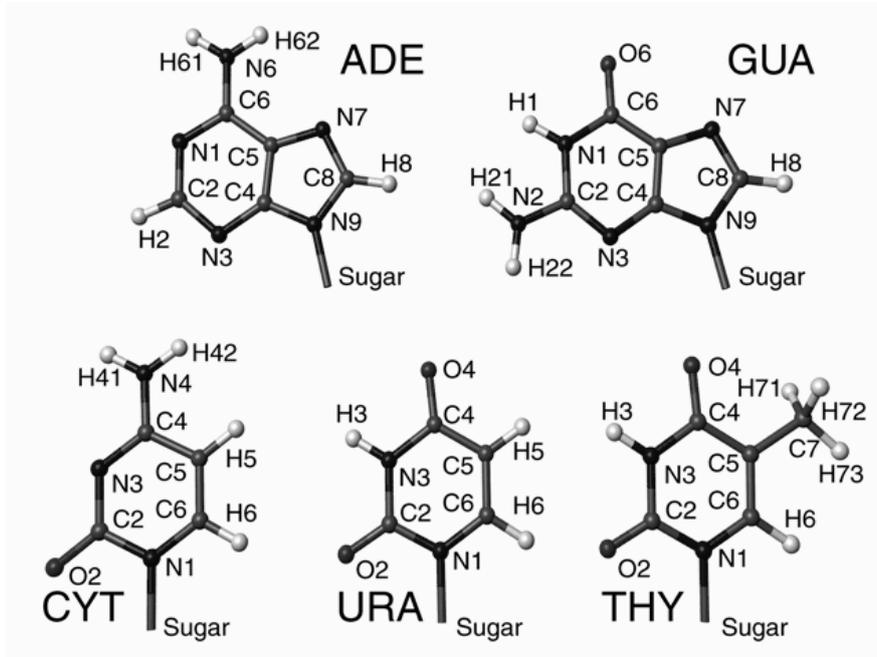


Abbildung 4: Natürliche in DNA und RNA vorkommende Nukleinbasen.
[<http://www.bmrb.wisc.edu/referenc/nomenclature/>],
Stand 21.06.2010

2.1.2 Struktur der RNA

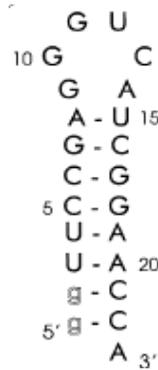
Die RNA Polynukleotide treten vorwiegend als Einzelstrang auf. Durch ausgedehnte intramolekulare Basenpaarungen ist es möglich, dass sie doppelhelikale Regionen ausbilden. Diese benötigen bestimmte Strukturinformationen, welche in drei Ebenen (siehe Abbildung 5) gespeichert sind. [6], [5]

Primärstruktur

Als Primärstruktur bezeichnet man die Abfolge der Nukleotide, d.h. die Nukleotidsequenz, welche die unterste Ebene der Strukturinformationsebenen darstellt. 5' und 3' bilden Anfang und Ende der Sequenz. In Abbildung 5 (a) ist die Primärstruktur der RNA 3PHP (pdb) abgebildet.

2 Grundlagen

GGUUCCGAGGGUCAUCGGAACCA



(a) Primärstruktur

(b) Sekundärstruktur

(c) Tertiärstruktur

Abbildung 5: RNA-Molekülstrukturinformationsebenen anhand der RNA 3PHP (pdb).
[Mario Schubert, Proteindatenbank]

Sekundärstruktur

Obwohl die RNA üblicherweise einzelsträngig ist, kommt es vor, dass sie sich faltet und doppelsträngige Regionen ausbildet (siehe Abbildung 5 (b)). Sie ist also in der Lage mit sich selbst Basenpaarungen einzugehen. Dabei verknüpfen sich die einzelnen Basen über Wasserstoffbrückenbindungen miteinander, wodurch es zur Ausbildung verschiedener Motive kommt. In Abbildung 6 sind ausgewählte Sekundärstrukturmotive aufgeführt. Das Bekannteste ist der sogenannte Hairpin loop. [13], [23]

Watson-Crick-Paare (WC-Paare) sind Basenpaare, bei denen bestimmte Atome an den Wasserstoffbrückenbindungen beteiligt sein müssen. Exemplarisch ist ein AU-Paar nur dann ein WC-Paar, wenn es genau zwei Wasserstoffbrückenbindungen bildet, wobei eines der beiden H-Atome am N6 des Adenin mit dem O2 des Uracil sowie das H3 des Uracil mit dem N1 des Adenin eine dieser Brücken ausbildet. Abbildung 7 zeigt die Watson-Crick-Paare der RNA-Nukleinbasen. [13]

In der RNA existieren wesentlich mehr Fehlpaarungen als in der DNA, was durch die vielfältigen Funktionen der RNA begründet ist. Dabei entstehen sogenannte Wobble, reverse Paare (Abbildung 7) oder auch Hoogsteenpaare. [13]

2 Grundlagen

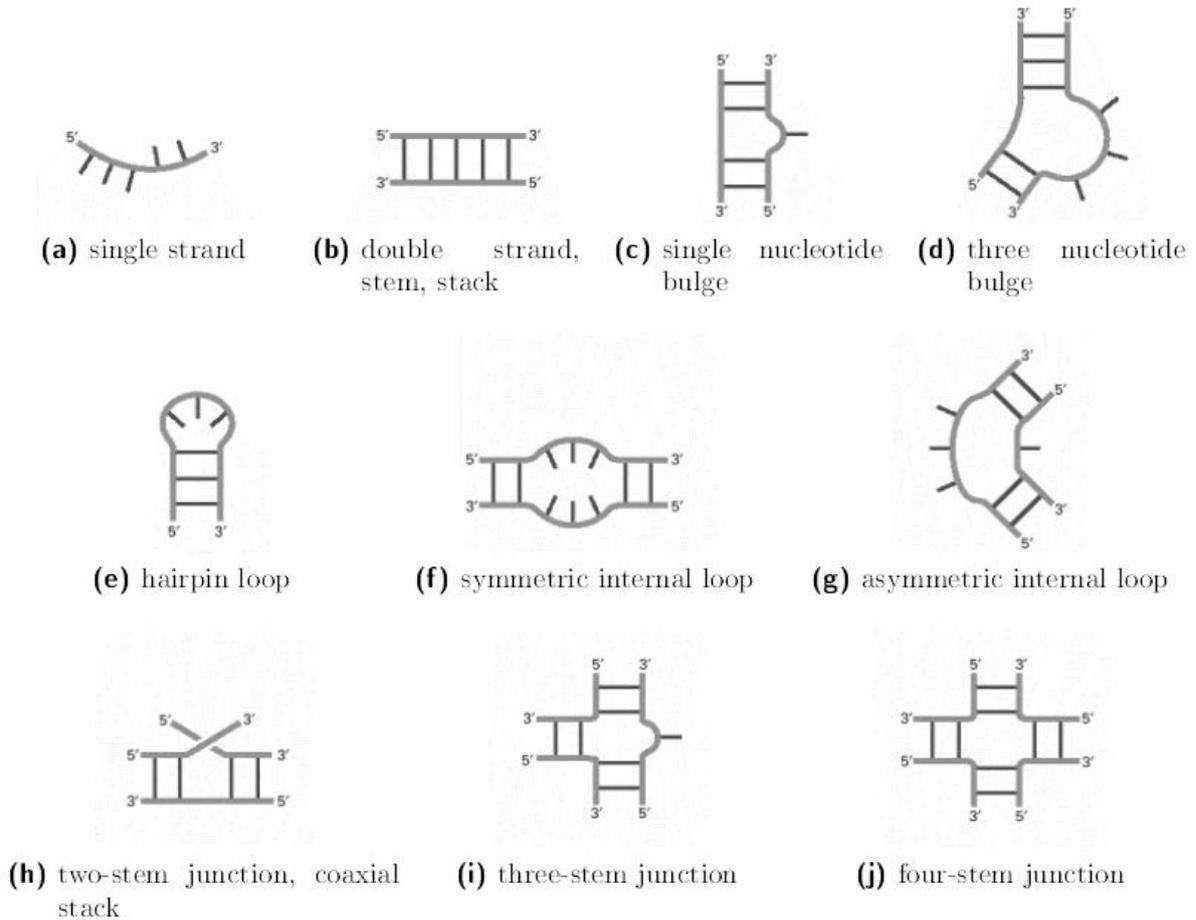


Abbildung 6: RNA-Sekundärstrukturmodelle. Die einzelnen Strukturen, außer (b), entstehen wenn sich nicht komplementäre Basen gegenüberstehen und somit keine Wasserstoffbrückenbindungen ausbilden. [13]

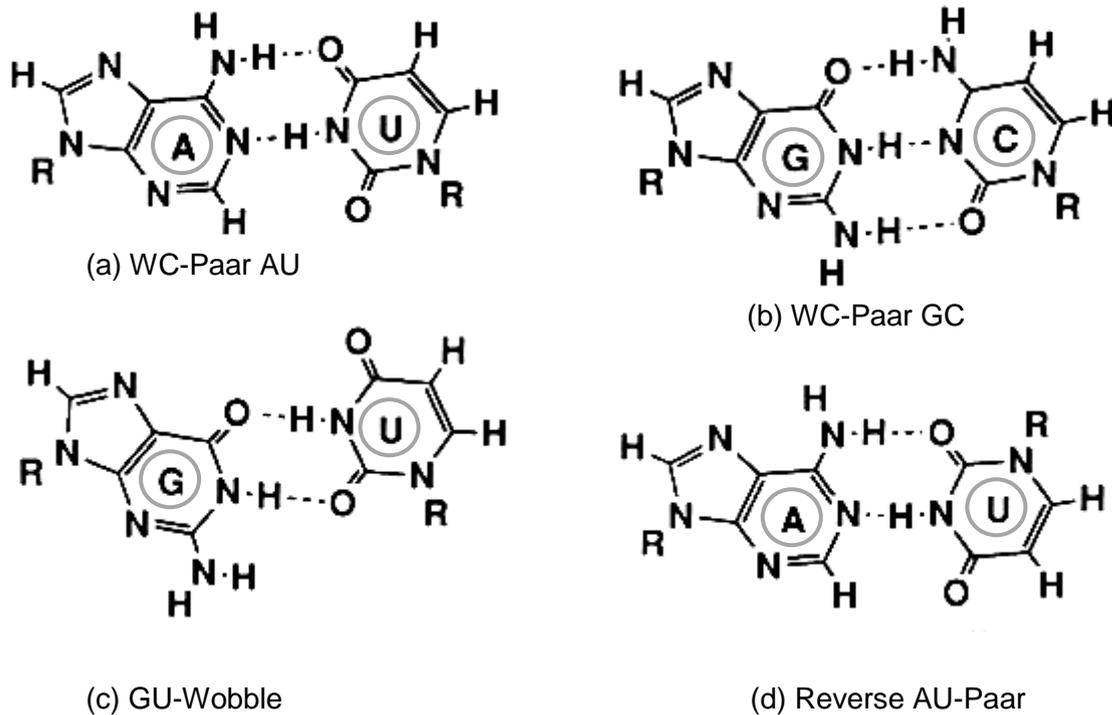


Abbildung 7: Abbildungen (a) und (b) zeigen die beiden möglichen Watson-Crick-Paarungen. Nicht alle AU- bzw. GC-Paare sind Watson-Crick-Paare. (c) und (d) zeigen jeweils eine mögliche in RNAs vorkommende Fehlpaarung. [13]

Tertiärstruktur

Der dreidimensionale Aufbau der Moleküle wird als Tertiärstruktur (siehe Abbildung 5 (c)) bezeichnet, die sich mit Hilfe von NMR- oder Röntgenstrukturanalysen bestimmen lässt. Der räumliche Abstand zwischen einzelnen Atomen sowie der Bindungswinkel sind von Interesse.

2.1.3 Arten der RNA

Jeder Organismus bildet verschiedene Typen von Ribonukleinsäuren aus, die unterschiedliche Aufgaben erfüllen und entsprechend ihrer Funktion benannt sind. In den 50er und 60er Jahren des 20. Jahrhunderts erforschte und verstand man die meisten Arten. Die bekanntesten sind mRNA, tRNA und rRNA. [5], [3], [4]

mRNA (Boten-RNA, engl.: messenger RNA) wird aus der hnRNA durch Spleißen gebildet um die Information der DNA zur Herstellung der Aminosäuresequenz vom Zellkern zu den Ribosomen im Zytoplasma zu übertragen. [5], [3], [26]

2 Grundlagen

tRNA (transfer RNA) ist ein einzelsträngiges zwischen 70 und 93 Nukleotide langes Molekül. Abbildung 8 veranschaulicht die ausgeprägte als Kleeblatt- oder L-Struktur bekannte Sekundärstruktur, die durch intramolekulare Zusammenlagerung des tRNA-Einzelstranges entsteht. tRNAs überbringen die Aminosäure zum Ort der Proteinbiosynthese, den Ribosomen. Für jede der kanonischen Aminosäuren gibt es mindestens eine spezifische tRNA. [5], [3], [4], [26]

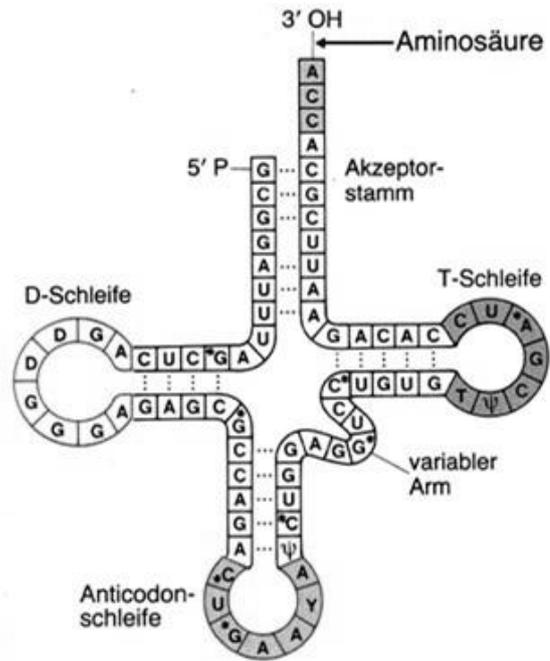


Abbildung 8: Typische Kleeblattstruktur einer aus 76 Nukleotiden bestehenden phenylalaninübertragenden tRNA der Hefe [27]

rRNA (ribosomale RNA) ist Bestandteil der Ribosomen. Sie trägt zur Struktur und zur katalytischen Funktion der Ribosomen bei. Die rRNA besitzt drei Untergruppen: die 28S rRNA, 18S rRNA und die 5,8S rRNA. [5], [3], [24]

hnRNA (heterogene nukleare RNA, engl.: heterogen nuclear RNA) ist ein primäres Transkriptionsprodukt, welches nach der Ablesung eines Gens auf der DNA entsteht und aus Exons und Introns aufgebaut ist. [26]

snRNA (kleine nukleare RNA, engl.: small nuclear RNA) ist als Bestandteil des sogenannten Spleißosoms am Spleißen der hnRNA beteiligt. [26]

tnRNA (kleine RNA, engl.: smallRNA oder tiny-noncodingRNA) sind etwa 28 Nukleotide lang und dienen zur Regulation von zellulären Prozessen. Die bekanntesten Molekülarten sind die miRNA (engl.: microRNA) und siRNA (engl.: small interfering RNA). Oftmals werden RNAi, die vor allem für die Forschung, insbesondere im medizinischen Bereich von Interesse sind, als siRNA bezeichnet. [1], [5]

2.1.4 Biologische Bedeutung der RNA

In Lebewesen ist die genetische Information in Form von DNA gespeichert. Durch Genexpression oder Proteinbiosynthese (Abbildung 9) werden aus dieser Information die Grundbausteine der Zelle, die Proteine, hergestellt. Bei der Proteinbiosynthese wird die DNA als erstes in RNA umgeschrieben. Dieser Vorgang wird als Transkription bezeichnet. Das entstehende Transkript nennt man mRNA. Sie dient bei der Translation als Vorlage bei der Synthese von Proteinen. Neben der mRNA ist außerdem die tRNA sowie Ribosomen am Prozess beteiligt. Ribosomen kommen im Zellplasma aller Lebewesen vor und bestehen aus rRNA und Proteinen, welche einen Komplex bilden.

Die RNA besitzt aber weitaus mehr Funktionalitäten, welche vor allem in den letzten Jahren entdeckt wurden. Dazu gehört unter anderem die Apoptose (programmierter Zelltod), die Zelldifferenzierung, die RNA-Interferenz oder der Abbau von DNA und RNA der Viren, die in die Zelle eingedrungen sind. [13]

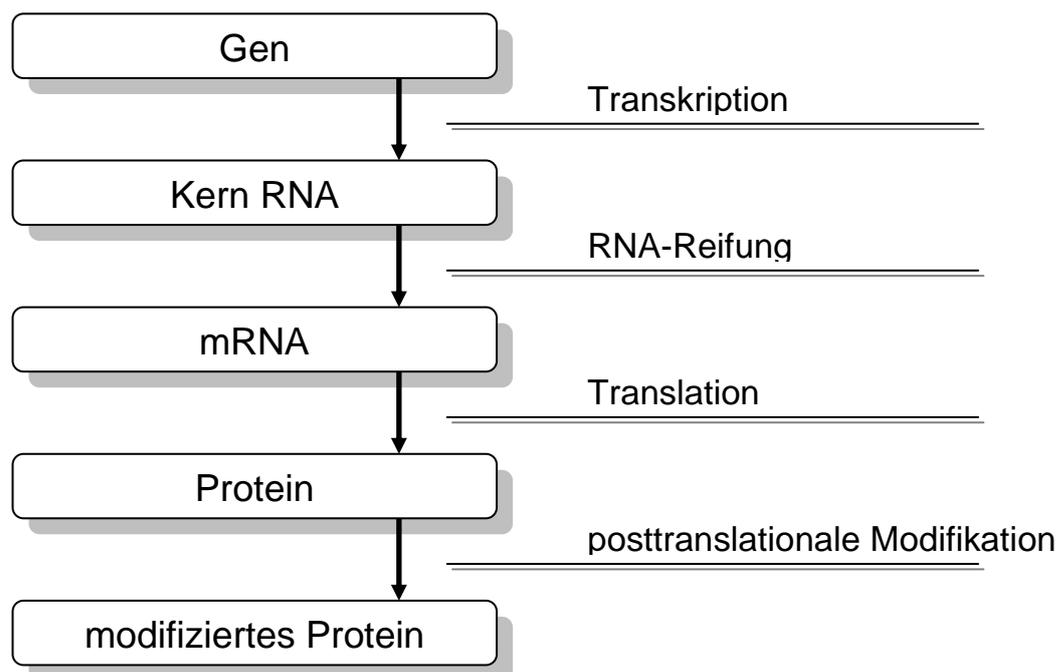


Abbildung 9: Genexpression bei Eukaryonten.

Vor der Translation im Zellplasma erfolgt die sogenannte RNA-Reifung oder Prozessierung der Kern-RNA im Zellkern. Dazu gehören unter anderem das RNA-splicing und das RNA-editing. Dieser Vorgang wird auch posttranskriptionale Modifizierung genannt.

2.2 NMR-Spektroskopie

Die kernmagnetische Resonanz-Spektroskopie (kurz Kernspinresonanz- oder NMR-Spektroskopie) ist eine der wichtigsten Methoden zur Strukturaufklärung von Molekülen wie Proteinen oder Ribonukleinsäuren. Die intramolekulare Dynamik, die molekularen Erkennungsprozesse, Reaktionskinetiken oder auch Proteinfaltungen können dabei untersucht werden. Anwendung findet die NMR-Spektroskopie vor allem in der organischen Chemie und neuerdings auch in der Biochemie und Strukturbiologie. [13], [2], [7]

Das Prinzip der NMR-Spektroskopie soll in diesem Kapitel kurz erläutert werden.

2.2.1 Theoretische Grundlagen

Der Atomkern besteht aus Protonen und Neutronen, die genauso wie Elektronen einen Spin aufweisen. Ist die Zahl der Protonen und Neutronen gerade, so kompensieren sich die Spins und der Atomkern hat einen Gesamtspin von Null. Ist sie jedoch ungerade, findet diese Kompensierung nicht statt und der Spin hat einen definierbaren Wert. Dieser wird mit I (Kernspin-Quantenzahl) bezeichnet, beschreibt den Gesamtdrehimpuls und kann sowohl ganze als auch halbzahlige Werte annehmen. [8], [2], [9], [13], [10]

Vor allem Kerne mit $I=1/2$ sind besonders gut für die kernmagnetische Resonanz-Spektroskopie geeignet. Hierbei sind zwei Ausrichtungsmöglichkeiten zu betrachten, die man antiparallel ($-1/2$) und parallel ($+1/2$) nennt, wobei letztere energieärmer ist. Mit Hilfe von elektromagnetischer Strahlung, die genau die passende Frequenz besitzt, kann ein Atomkern aus der parallelen (\uparrow) in die antiparallele (\downarrow) Ausrichtung angeregt werden. Dadurch wird der vorher parallel ausgerichtete Kern auf ein höheres Energieniveau gebracht. [8], [2], [9], [13], [10]

Bringt man Kerne mit $I=1/2$ in ein starkes permanentes Magnetfeld, so wird eine Bewegung der Elektronen in den Molekülen ausgelöst. Dies hat zur Folge, dass ein elektrischer Strom und dadurch wiederum ein zweites Magnetfeld induziert wird. Der Atomkern wird dementsprechend durch die ihm umgebene Elektronenhülle magnetisch abgeschirmt. Diese Abschirmung wird als chemische Verschiebung bezeichnet. Gibt man nun eine bekannte Standard- bzw. Referenzsubstanz, z.B. Tetramethylsilan (TMS = $\text{Si}(\text{CH}_3)_4$, siehe Abbildung 10) hinzu, kann man die Frequenzunterschiede der Kerne gegen TMS bestimmen. TMS, dessen chemische Verschiebung als 0 ppm definiert wird, ist eine inerte Verbindung und wird aus praktischen Gründen als Referenz verwendet, da es im Gegensatz zu anderen Substanzen nur ein Signal liefert, welches außerdem sehr intensiv ist. Der Einsatz von TMS hat sich vor allem bei ^1H -, ^{13}C - und ^{29}Si -NMR-Spektroskopie durchgesetzt. Für andere Kernarten gibt es weitere, gelistete Referenzsubstanzen. [8], [2], [9], [13], [10]

2 Grundlagen

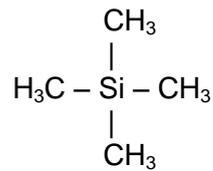


Abbildung 10: Strukturformel von TMS

2.2.2 Eindimensionale und Zweidimensionale NMR-Experimente

Eindimensionale NMR-Experimente bestehen aus den beiden Phasen Präparation und Detektion (Abbildung 11). Die Präparationsphase besteht im einfachsten Fall aus einem einzelnen kurzen und starken 90° -Puls, der dazu führt, dass die Magnetisierung aus dem ursprünglichen Zustand auf der z-Achse in die y-Achse ausgelenkt wird. Danach rotieren die Kerne um die z-Achse und induzieren eine Spannung. Diese nimmt während der Detektionsphase immer weiter ab. Demzufolge werden die Daten als FID (Freier Induktionszerfall, engl. Free Induction Decay) bezeichnet. Im Anschluss an die beiden Phasen werden die Daten durch Fourier Transformation in Frequenzdaten übersetzt. [10]

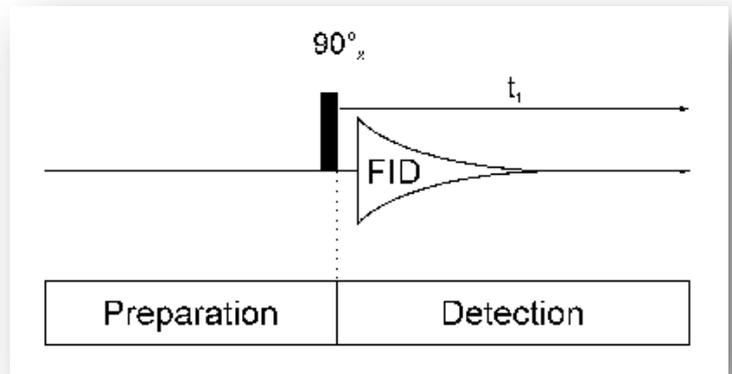
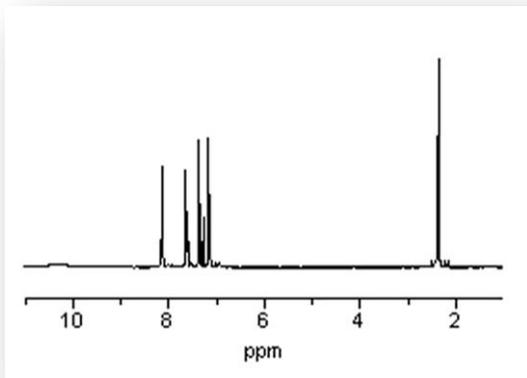


Abbildung 11: links: Eindimensionales NMR-Spektrum

[http://www.chemgapedia.de/vsengine/vlu/vsc/de/ch/3/anc/nmr_spek/zweidimensionale_nmr.vlu/Page/vsc/de/ch/3/anc/nmr_spek/m_50/nmr_10_1/zweideinl_m50te0803.vscml.html, Stand 28.06.2010]

rechts: Aufbau eines eindimensionalen NMR- Experiments. [10]

Zweidimensionale NMR-Experimente sind für komplexe Moleküle geeignet und deren Spektren besitzen zwei Frequenzachsen, Abszisse und Ordinate. Die dritte Dimension bilden die Intensitäten der Spins. Bei einem 2D-Spektrum lagern sich mehrere verschiedene Signale des Kernspins übereinander wodurch die Interpretation eines 1D-Spektrums nahezu unmöglich ist. Durch die Einführung von weiteren spektralen Dimensionen können die Überlagerungen aufgelöst werden. Im Gegensatz zu eindimensionalen NMR-Experimenten

2 Grundlagen

verfügen die 2D-Experimente über vier Phasen: die Präparationsphase, eine indirekte Evolutionszeit, eine Mischzeit und die Detektionsphase (siehe Abbildung 12). [8], [10]

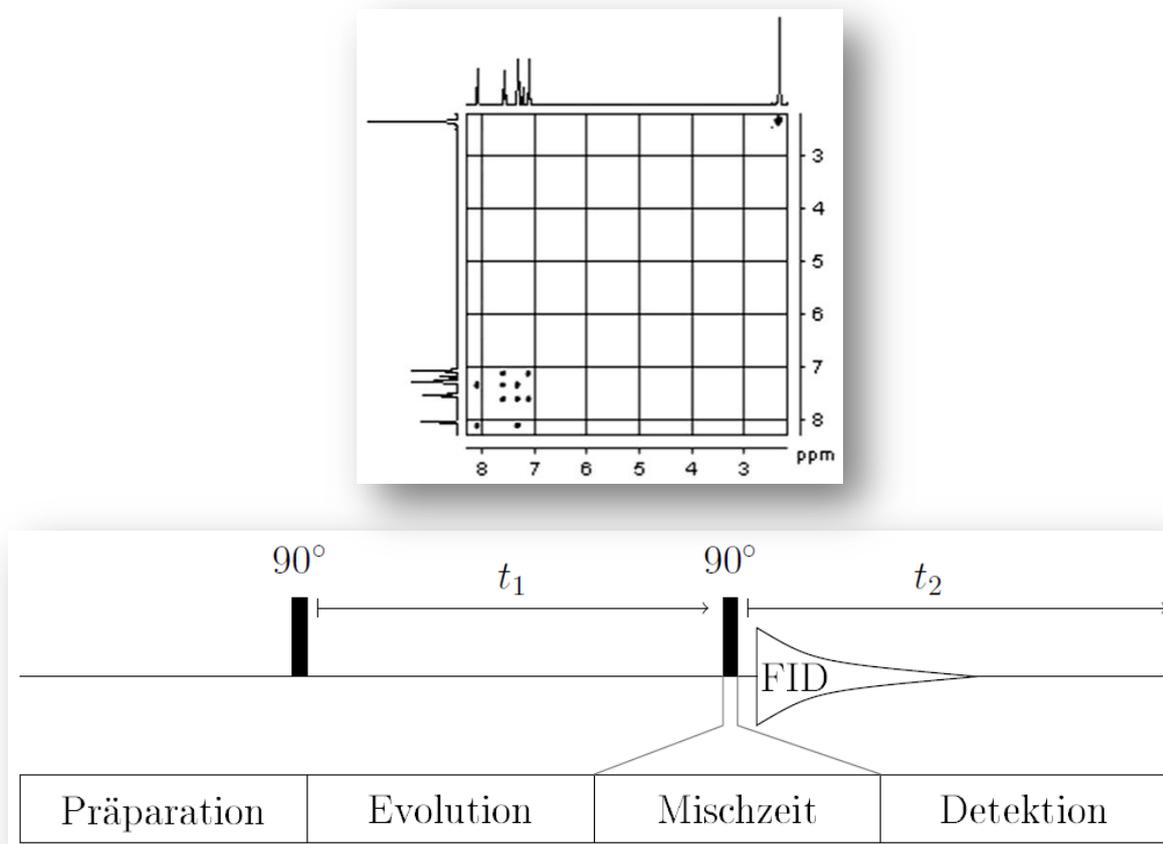


Abbildung 12: oben: Zweidimensionales NMR-Spektrum

[http://www.chemgapedia.de/vsengine/vlu/vsc/de/ch/3/anc/nmr_spek/zweidimensionale_nmr.vlu/Page/vsc/de/ch/3/anc/nmr_spek/m_50/nmr_10_1/zweideinl_m50te0803.vscml.html, Stand 28.06.2010]

unten: Aufbau eines zweidimensionalen NMR- Experiments. Der FID ist das Signal indem in der Empfängerspule eine Spannung induziert wird. [10]

Nach der Präparationsphase präzedieren die Spins frei innerhalb einer festgesetzten Zeit t_1 , in der unter anderem die chemische Verschiebung des ersten Kerns „markiert“ wird. In der Mischzeit wird die Magnetisierung von einem Kern auf einen anderen entweder durch skalare Kopplung oder durch dipolare Wechselwirkungen übertragen. Zum Abschluss des Experiments erfolgt die sogenannte Datenakquisition. Zu dieser Zeit t_2 wird die Magnetisierung mit der chemischen Verschiebung des zweiten Kerns markiert. Somit erhält man, wie in Abbildung 13 ersichtlich, nach der Fourier-Transformation ein triviales, d.h. ein gewöhnliches eindimensionales Spektrum. Führt man nun mehrere NMR-Einzelexperimente durch, wobei man nur jeweils t_1 variiert, so ist das gemessene Signal bei verschiedenen Evolutionszeiten unterschiedlich stark ausgeprägt. Die bei jedem Einzelexperiment

2 Grundlagen

aufgenommenen Signale in der Detektionszeit t_2 werden als Abfolge von Momentaufnahmen abgebildet. Das endgültige, als Höhenliniendiagramm dargestellte 2D-Spektrum entsteht aus einer weiteren Fourier-Transformation, wobei die Momentaufnahmen in Frequenzen umgerechnet werden. Dieses Spektrum enthält verschiedene Peaks, die jeder für sich in jeder der beiden Frequenzrichtungen ω_1 und ω_2 einer Kernsorte zuzuordnen sind. Bei einem homonuklearen Spektrum liegen die Peaks auf einer Diagonalen ($\omega_1 = \omega_2$) und haben dieselbe Kernsorte. Als Kreuzsignale werden die Peaks bezeichnet, die nicht auf der Diagonalen liegen, sondern symmetrisch zu dieser liegen. Sie enthalten die eigentlich wichtigen Informationen, da sie Auskunft über die Wechselwirkung zweier Kerne geben. [10], [8], [13]

Zur Ermittlung der kompletten 3D-Struktur von RNA-Molekülen reichen drei zweidimensionale NMR-Experimente aus. Diese werden im Folgenden näher betrachtet.

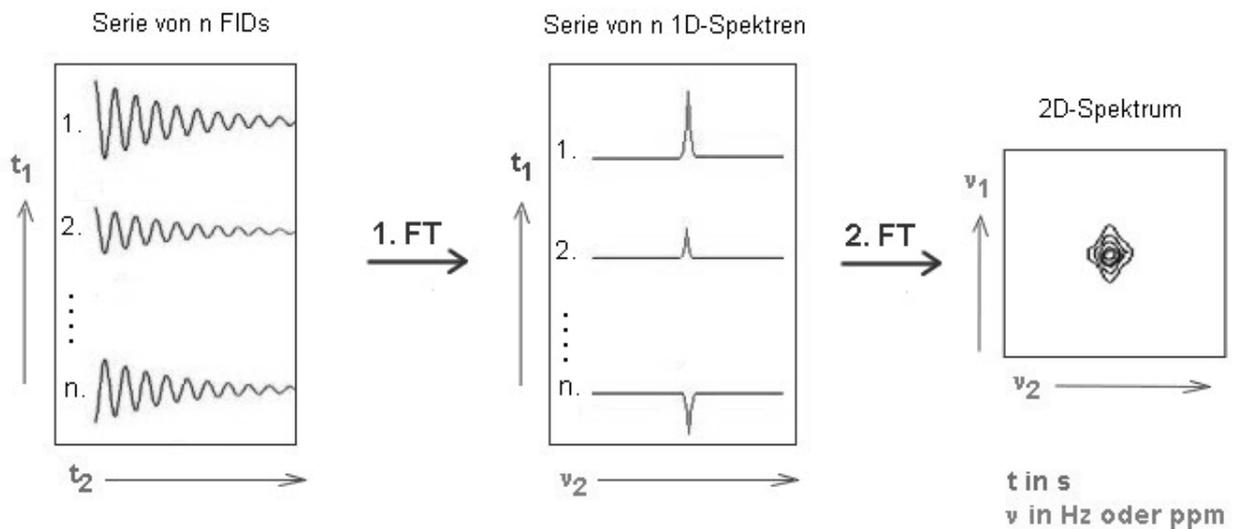


Abbildung 13: Entstehung eines 2D-NMR-Spektrums durch zweidimensionale Fourier-Transformation.

[http://www.chemgapedia.de/vsengine/vlu/vsc/de/ch/3/anc/nmr_spek/zweidimensional_e_nmr.vlu/Page/vsc/de/ch/3/anc/nmr_spek/m_50/nmr_10_1/zweideinl_m50te0803.vscml.html, Stand 28.06.2010]

2.3.3 TOCSY-Experiment

Eine Möglichkeit der zweidimensionalen NMR-Spektroskopie ist das sogenannte TOCSY-Experiment (**T**otal **C**orrelation **S**pectroscopy, vollständige Korrelationsspektroskopie). In diesem Experiment werden alle Protonen eines Spinsystems, hier Nukleotide, miteinander korreliert. Dabei wird die Magnetisierung durch einen mehrstufigen Transfer über skalare Kopplung über das gesamte System eines Nukleotids verteilt. Für jedes Nukleotid erhält man ein charakteristisches Signalmuster, anhand dessen man die Nukleotide besser identifizieren kann als mit der eindimensionalen NMR-Spektroskopie. Im Wesentlichen wird das TOCSY-Experiment dafür verwendet um zwischen intraresidualen und interresidualen Kontakten zu unterscheiden. [13], [8], [10]

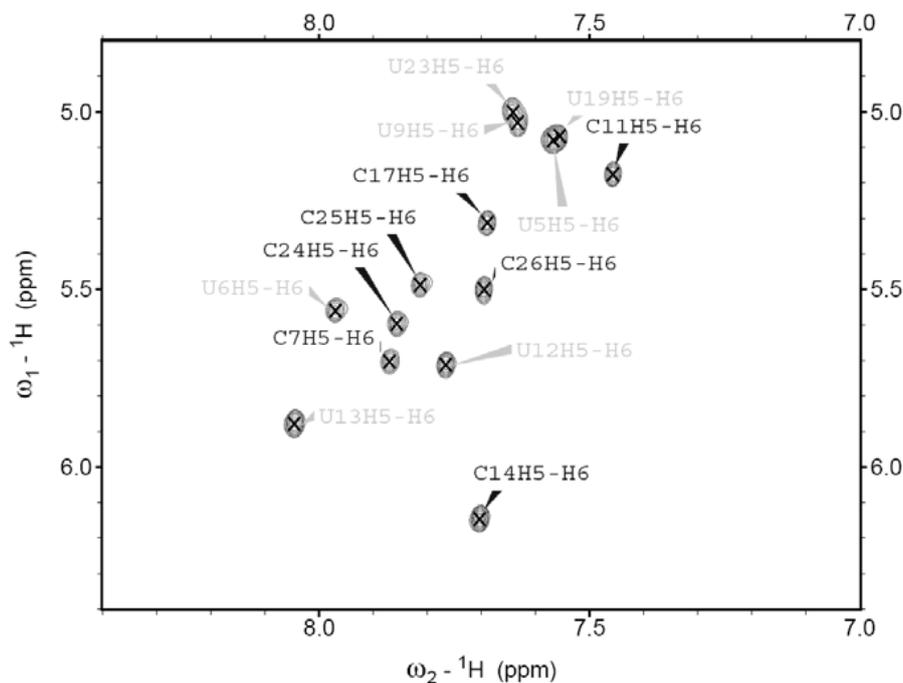


Abbildung 14: Das ^1H - ^1H -TOCSY-Spektrum der RNA TASL2 [22]

Das ^1H - ^1H -TOCSY-Spektrum einer RNA (siehe Abbildung 14) zeigt für jedes Uracil und jedes Cytosin genau einen H5-H6 Peak. Diese sind meist sehr gut im Spektrum verteilt. Genau aus diesem Grund wird es häufig als sogenanntes Fingerprint verwendet. Die H5-H6-Kreuzsignale treten ebenfalls im NOESY-Spektrum auf, falls der Abstand zwischen dem H5- und H6-Protonen in Uracil und Cytosin sehr klein ist, etwa 2,44-2,48 Å. [22]

2.3.4 NOESY-Experiment

Zur Bestimmung der Struktur und den dazu benötigten intramolekularen Abständen ist das NOESY-Experiment (**N**uclear **O**verhauser and **E**nhancement (**E**ffect) **S**pectroscopy) entscheidend. Basierend auf den dipolaren Wechselwirkungen der Kerne bedient man sich des sogenannten Kern-Overhauser-Effekts (Nuclear Overhauser Effect, kurz: NOE). Die dipolaren Wechselwirkungen werden durch den Raum transferiert, wobei deren Stärke proportional zu $1/r^6$ ist (r ist hierbei der Abstand zwischen den beteiligten Kernen). Die Korrelation zweier Kerne ist demzufolge von ihrem räumlichen Abstand abhängig wobei der Magnetisierungsübertrag dieser nur beobachtet wird, wenn ihr Abstand weniger als 5 Å beträgt. [15], [10], [8], [16]

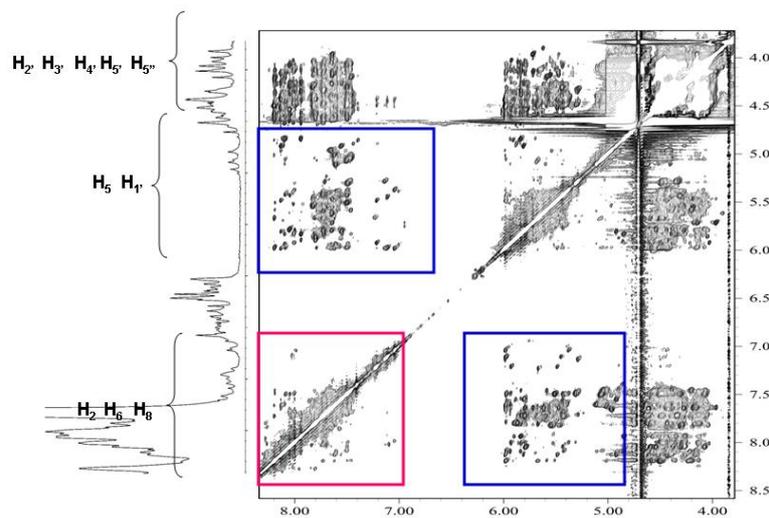


Abbildung 15: ^1H - ^1H -NOESY-Spektrum erfasst in Deuteriumoxid. Von links unten nach rechts oben ist die „Wasserlinie“ ersichtlich, an der sich die Kreuzsignale spiegeln (blaue Kästen). [22]

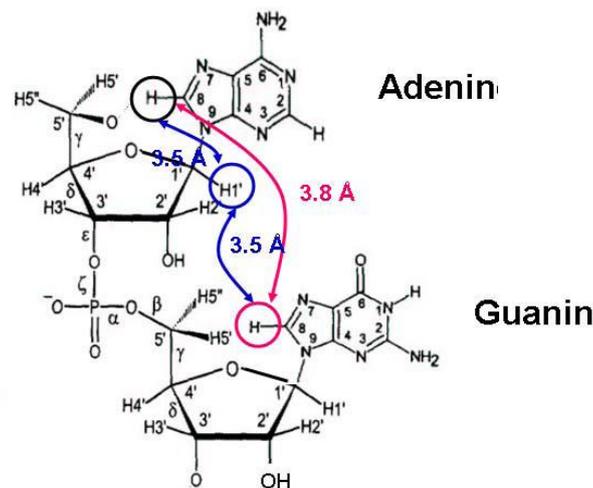


Abbildung 16: Typische intra- und interresiduale Signale in einer RNA-Helix.
Das H8 von Guanin bildet mit dem H1' des Zuckers von Adenosin bzw. das H1' des Zuckers von Adenosin bildet mit dem H8 (blau) von Adenin ein Signal und das H8 von Guanin bildet mit dem H8 von Adenin (pink) ebenfalls ein Signal. [22]

Welche Signale man in einem ^1H - ^1H -NOESY-Spektrum (siehe Abbildung 15) sieht, hängt von der Tertiärstruktur der RNA ab. Man erkennt zwischen allen Protonen Kreuzsignale, was zur Folge hat, dass auch Protonen miteinander korreliert worden, welche zwar in der Primärstruktur weit voneinander entfernt sind, deren Abstand in der räumlichen Struktur aber weniger als 5 Å beträgt. Daher treten also nicht nur intraresiduale Kreuzsignale auf, sondern es kommt auch zu interresidualen Kontakten zwischen Protonen (siehe Abbildung 16). Folglich stellt diese Korrelation der Protonen aufgrund ihrer räumlichen Beziehung die wichtigste Strukturinformation in der NMR-Spektroskopie dar. [15], [10], [8], [16], [22]

2.3.5 HSQC-Experiment

Bei dem HSQC-Experiment (**H**eteronuclear **S**ingle **Q**uantum **C**oherence) handelt es sich um ein *inverses protonendetektierendes* Experiment. Es dient zur Detektion von Kernen mit geringer Empfindlichkeit. Bei dieser Methode werden bei den unempfindlichen Kernen Magnetisierungen erzeugt, die auf die empfindlichen Kerne, hier ^1H , übertragen werden. Letztlich werden die „leicht zu messenden“ ^1H -Resonanzen detektiert. [10], [8]

Außerdem ist es in der Lage überlappende Resonanzen in der Protonendimension durch Entzerrung des Spektrums deutlich getrennt darzustellen. Aus diesem Grund basieren viele mehrdimensionale Experimente im Wesentlichen auf dem HSQC-Experiment. [10], [8]

Im Vergleich zu homonuklearen Spektren hat das HSQC keine Diagonale, da während den Zeiten t_1 und t_2 völlig unterschiedliche Kerne gemessen werden. Der große Nachteil des HSQC-Experiments ist, dass es schon auf kleine Abweichungen von den exakten Impulswinkeln sehr empfindlich reagiert. Desweiteren erfordert das Experiment fünf genau eingestellte 90° und 180° -Impulse. [10], [8]

In einem ^1H - ^{13}C -HSQC-Spektrum werden die Korrelationen zwischen direkt gebundenen Kohlenstoff- und Wasserstoffatomen sichtbar. Folgende Kreuzsignale sind im Spektrum ersichtlich: H2-C2 und H8-C8 in Adenin und Guanin, H5-C5 und H6-C6 in Uracil und Cytosin sowie H1'-C1' in allen Nukleinbasen. Analoge Experimente sind die ^{15}N -HSQC. [10], [8]

2.3 Chemische Verschiebungen

Die kernumgebenden Elektronen erzeugen im Molekül ein schwaches Magnetfeld, welches eine Veränderung des Kernspinniveaus bewirkt. Die Resonanzfrequenz des betrachtenden Kerns wird beeinflusst und folglich der Kern abgeschirmt. Diese Abschirmung wird als chemische Verschiebung δ bezeichnet und ist von der direkten chemischen Umgebung des betrachteten Kerns abhängig. Je höher die Elektronendichte, desto größer ist die Abschirmung des Kerns und somit ist die chemische Verschiebung kleiner. Sie ist eine relative Größe und wird anhand Gleichung (1) berechnet. Zur Vereinfachung wird der Zahlenwert mit 10^6 multipliziert und die δ in ppm (parts per million) angegeben, da die Einheit unabhängig von der Magnetfeldstärke ist. [21], [20], [10]

$$(1) \quad \delta = \frac{\nu(\text{Probe}) - \nu(\text{TMS})}{\nu(\text{TMS})} 10^6 \text{ ppm}$$

$$\nu(\text{TMS}) = \text{Resonanzfrequenz von TMS}$$

[9], [10]

Aufgrund der Geräteunabhängigkeit der chemischen Verschiebung konnten für jede Kernart, da voneinander verschiedene Kerne unterschiedliche chemische Umgebungen haben, Verschiebungstabellen erstellt werden, in denen δ für die meisten Strukturgruppen abgelesen werden kann. [21], [20], [10]

Nahezu alle bekannten chemischen Verschiebungen sind positiv, da die Referenzsubstanz TMS die Wasserstoff- und Kohlenstoff-Atome sehr stark abschirmt und es fast keine weiteren Gruppen mit derart hoher Abschirmung gibt. Im Spektrum werden die chemischen Verschiebungen auf der Abszisse von rechts nach links aufgetragen. Chemisch nicht äquivalente Kerne ergeben im Spektrum getrennte Resonanzsignale. [21], [20], [10]

2 Grundlagen

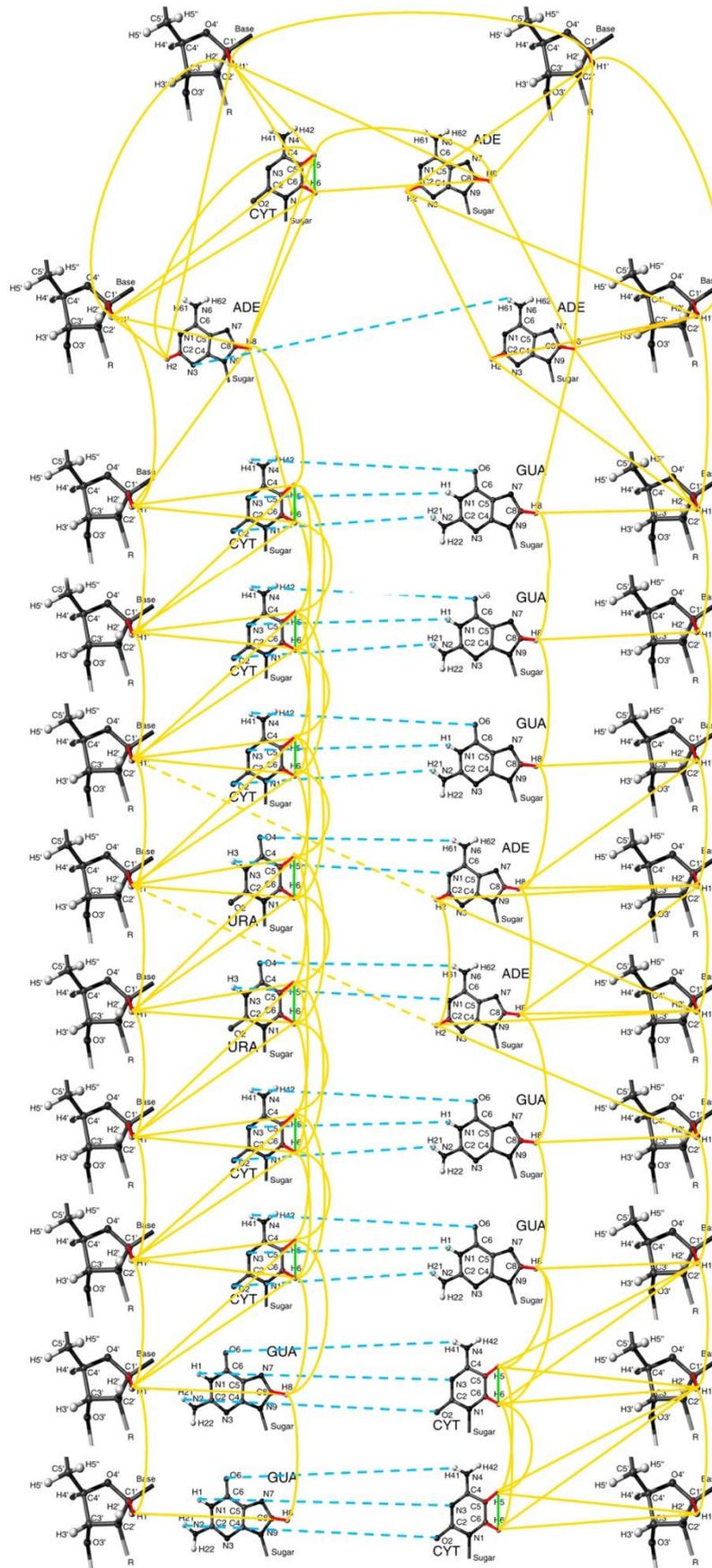


Abbildung 17: Schematische Darstellung der RNA 1pjj (pdb).

Gelbe Linien beschreiben die im NOESY, rote die im HSQC und grüne die im TOCSY erzeugten Kreuzsignale der chemischen Verschiebungen. Die blauen Linien stellen die Wasserstoffbrückenbindungen zwischen den Nucleinbasen der RNA dar.

[modifiziert nach: <http://www.bmrw.wisc.edu/referenc/nomenclature>], Stand 21.06.2010

2.3.1 Chemische Verschiebungen innerhalb eines Nukleotids

Chemische Verschiebungen innerhalb eines Nukleotids gehören zur Gruppe der intraresidualen Signale in einem NMR-Spektrum. Abbildung 18 stellt die chemischen Verschiebungen im Cytosin und im Cytidin dar. Deutlich zu erkennen sind dabei links die Verschiebungen von H5-H6 im TOCSY und NOESY, C5-H5 und C6-H6 im HSQC und rechts C1'-H1', H1'-H6 sowie H1'-H5.

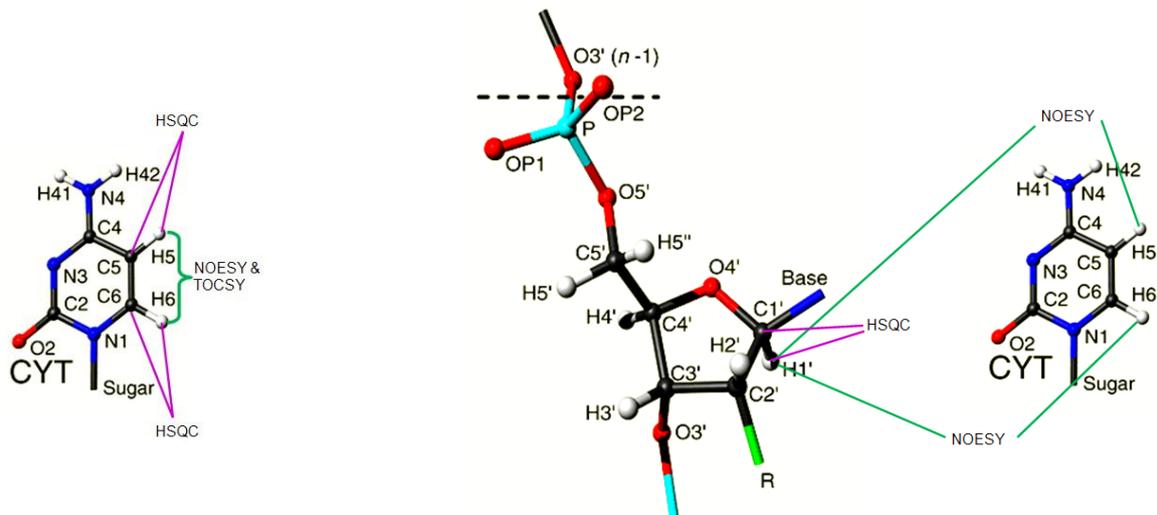


Abbildung 18: Chemischen Verschiebungen im Cytosin und im Cytidin.
[modifiziert nach <http://www.bmrbl.wisc.edu/referenc/nomenclature>], Stand 21.06.2010

Abbildung 19 veranschaulicht die chemischen Verschiebungen im Adenin und im Adenosin. Im Gegensatz zum Cytosin sieht man in den Spektren des Adenin Signale von C2-H2 und C8-H8 innerhalb der Nukleinbase als auch C1'-H1', H1'-H8 sowie H1'-H2 mit dessen Zucker. Diese Signale werden nunmehr immer als Signale in der Hauptbase bezeichnet.

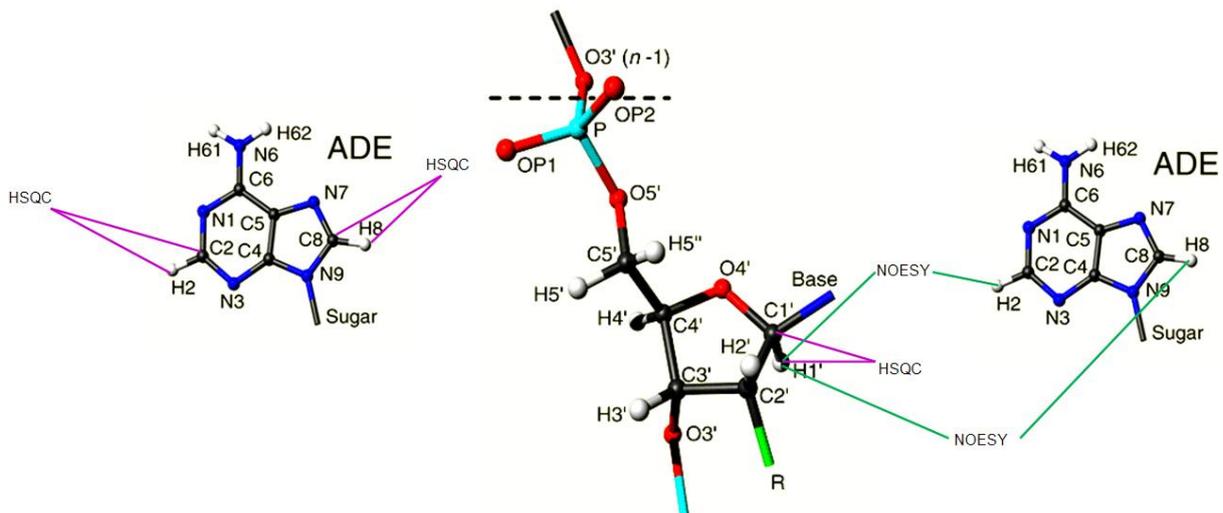


Abbildung 19: Chemischen Verschiebungen im Adenin und im Adenosin.
[modifiziert nach <http://www.bmrbl.wisc.edu/referenc/nomenclature>], Stand 21.06.2010

2.3.2 Chemische Verschiebungen sequentiell mit aufeinanderfolgenden Nukleotiden

Kreuzsignale von chemischen Verschiebungen die von aufeinanderfolgenden Nukleotiden entstehen, sind sequentielle Verschiebungen und werden als interresiduale Signale bezeichnet. Abbildung 20 veranschaulicht chemische Verschiebungen zwischen einem H8-Proton einer Purinbase und einem H1'-Proton des Zuckers des Vorgängers. Diese Signale sieht man in einem NOESY-Spektrum.

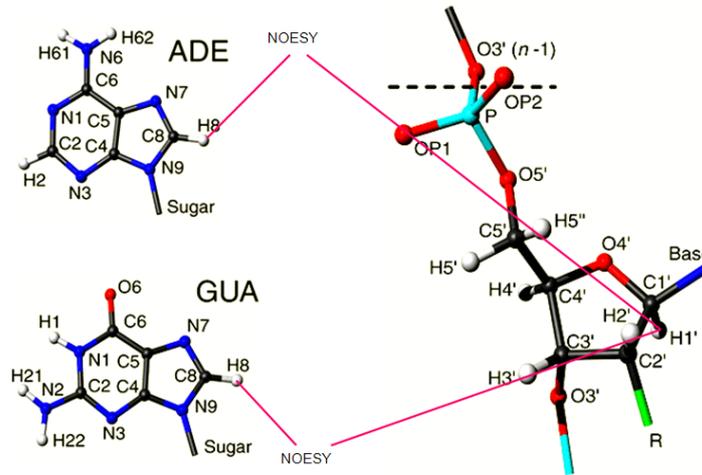


Abbildung 20: Kreuzsignal zwischen der chemischen Verschiebung des H8-Atoms von Adenin mit der chemischen Verschiebung des H1'-Atoms vom Zucker des Vorgängers
[modifiziert nach <http://www.bmrb.wisc.edu/referenc/nomenclature>], Stand 21.06.2010

Abbildung 21 hingegen stellt die Kreuzsignale von einem H6-Proton mit einem H5-Proton des Nachfolgers dar. Diese Signale erkennt man im NOESY-Spektrum. Dabei sind nur Pyrimidinbasen beteiligt, da es im Guanin und Adenin keine H5- und H6-Protonen gibt.

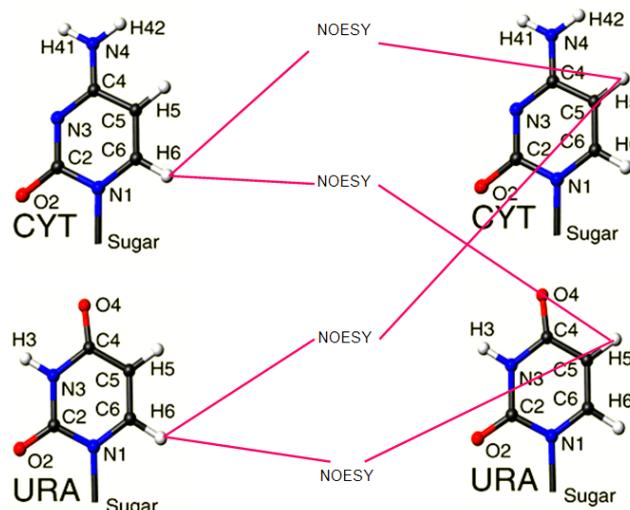


Abbildung 21: Kreuzsignal zwischen der chemischen Verschiebung vom H6-Atom des Cytosin oder Uracil mit der chemischen Verschiebung vom H5-Atom des nachfolgenden Cytosin oder Uracil
[modifiziert nach <http://www.bmrb.wisc.edu/referenc/nomenclature>], Stand 21.06.2010

2.3.3. Einflüsse auf chemische Verschiebungen

Chemische Verschiebungen können durch bestimmte Faktoren beeinflusst werden. Dazu gehören unter anderem intermolekulare Wechselwirkungen wie Wasserstoffbrückenbindungen, Anisotropieeffekte und elektrische Feldeffekte. Weitere Störfaktoren von chemischen Verschiebungen sind Substituenten, aromatische Ringstromeffekte und Lösungsmittel. [20]

Ringstromeffekte in Aromaten

Für die niedrigere Abschirmung der Protonen in aromatischen Systemen, als in aliphatischen Doppelbindungssystemen, ist der Kreis- oder Ringstromeffekt verantwortlich. Er spielt insbesondere für die ^1H -NMR-Spektroskopie eine große Rolle.

Dabei induziert das äußere Magnetfeld in der Ringebene einen Strom der π -Elektronen, der sogenannte Ringstrom, welcher wiederum ein Magnetfeld erzeugt. Dieses ist innerhalb des Ringes dem äußeren Feld entgegen gerichtet. Am Protonenort ist es dagegen in Richtung des äußeren Magnetfeldes gerichtet und verstärkt dieses. Demzufolge ist die Abschirmung der Ringprotonen gering und die chemische Verschiebung hoch. [20]

Wasserstoffbrückenbindungen

In der ^1H -NMR-Spektroskopie ist es möglich, dass Wasserstoffbrückenbindungen zu einer Tieffeldverschiebung führen können, d.h. die Resonanzsignale erscheinen bei chemischen Verschiebungen von zum Teil über 10 ppm. [20]

Elektrische Feldeffekte

Dabei kann es in Verbindungen mit starken polaren Gruppen (z.B. Nitro- und Carbonyl-Gruppen) zum Aufbau eines zusätzlichen elektrischen Feldes und somit zu Veränderungen der Elektronendichte um den betreffenden Kern kommen. [20]

2.4 Zweidimensionale Normalverteilung

Eine Methode zur Datenseparierung von stetigen Messwerten ist die zweidimensionale Normalverteilung. Für verschiedene Datengruppen werden dabei die Wahrscheinlichkeitsdichtefunktionen berechnet, wodurch für jedes unbekannte, in eine Gruppe zu ordnende Wertepaar eine Wahrscheinlichkeit für die Zugehörigkeit zu einer bestimmten Gruppe berechnet wird.

Für diese Berechnung sind folgenden Größen zu definieren:

Der Mittelwert μ berechnet sich aus der Summe aller Messwerte geteilt durch die Gesamtanzahl der Messwerte, kurz:

(2)

$$\mu = \frac{1}{n} \sum_{i=1}^n x_i$$

Die Stichprobenkovarianz $Cov(x_j, x_k)$ der beiden Variablen j und k errechnet sich aus der Gleichung (3):

(3)

$$Cov(x_j, x_k) = \frac{1}{n} \sum_{i=1}^n (x_{ji} - \mu_j)(x_{ki} - \mu_k)$$

Anhand Gleichung (4) lässt sich die Varianz berechnen.

(4)

$$Var(x_j) = \frac{1}{n-1} \sum_{i=1}^n (x_{ji} - \mu_j)^2$$

Der Korrelationskoeffizienten ρ_{jk} berechnet sich laut folgender Gleichung:

(5)

$$\rho(x_j, x_k) = \frac{Cov(x_j, x_k)}{\sqrt{Var(x_k)} \sqrt{Var(x_j)}}$$

2 Grundlagen

Die Standardabweichung σ berechnet sich wie folgt:

(6)

$$\sigma = \sqrt{\text{Var}(x_j)} = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_{ji} - \mu_j)^2}$$

Die Wahrscheinlichkeit für die Zugehörigkeit eines Wertepaares zu einer bestimmten Gruppe bestimmt sich aus dem Integral der Dichtefunktion. Die Gleichung für die zweidimensionale Dichtefunktion lautet:

(7)

$$f(x_1, x_2) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \exp \left\{ -\frac{1}{2(1-\rho^2)} \left[\left(\frac{x_1 - \mu_1}{\sigma_1} \right)^2 - 2\rho \left(\frac{x_1 - \mu_1}{\sigma_1} \right) \left(\frac{x_2 - \mu_2}{\sigma_2} \right) + \left(\frac{x_2 - \mu_2}{\sigma_2} \right)^2 \right] \right\}$$

Nun betrachtet man das Gebiet, welche durch die Ellipsengleichung (8) beschrieben wird:

(8)

$$1 - \rho^2 = \left(\frac{x_1 - \mu_1}{\sigma_1} \right)^2 - 2\rho \left(\frac{x_1 - \mu_1}{\sigma_1} \right) \left(\frac{x_2 - \mu_2}{\sigma_2} \right) + \left(\frac{x_2 - \mu_2}{\sigma_2} \right)^2$$

Die entstandene Ellipse begrenzt das Volumen unter der Dichtefunktion. Die Wahrscheinlichkeit, dass ein Wert zur Verteilung gehört entspricht dem Volumen unter der Glocke innerhalb der Ellipse:

(9)

$$P(X_1, X_2) = \frac{1}{\sqrt{e^{1-\rho^2}}}$$

Die Wahrscheinlichkeit für ein Wertepaar (X_1, X_2) lautet somit:

(10)

$$P(X_1, X_2) = \exp \left\{ -\frac{1}{2} \left[\left(\frac{x_1 - \mu_1}{\sigma_1} \right)^2 - 2\rho \left(\frac{x_1 - \mu_1}{\sigma_1} \right) \left(\frac{x_2 - \mu_2}{\sigma_2} \right) + \left(\frac{x_2 - \mu_2}{\sigma_2} \right)^2 \right] \right\}$$

Das Herleiten und Aufstellen der Wahrscheinlichkeitsformel wurde von Juliana Hildebrand durchgeführt. [13]

3 Data-mining

Dieses Kapitel dient zur Erläuterung der Beschaffung und Unterteilung der Daten, welche als Übungsdatensatz benötigt wurden.

3.1 Datenbanken

Die für diese Arbeit notwendigen Daten konnten anhand zweier Datenbanken ermittelt werden, die im Folgenden beschrieben werden.

3.1.1 PDB

Die Proteindatenbank ist eine der bekanntesten öffentlichen Datenbanken und wird vom Research Collaboratory for Structural Bioinformatics verwaltet. Sie beinhaltet die durch NMR-Spektroskopie oder Röntgenstrukturanalyse ermittelten 3D-Strukturdaten von komplexen biologischen Molekülen wie Proteinen und Nukleinsäuren. [11], [13], [6]

1971 wurde die PDB gegründet und enthielt 7 Strukturen. Momentan verfügt sie über 60906 Einträge für Proteine, 1265 für DNA und 793 für RNA. Abbildung 22 zeigt wie sich die RNA-Einträge in den vergangenen Jahren verändert haben.

Im Jahr 2003, mit der Gründung der wwPDB (engl.: world wide Protein Data Bank) durch die RCSB PDB (USA), PDBe (Europe) und PDBj (Japan) stieg die Proteindatenbank zu einer internationalen Organisation auf. [11]

Die Informationen sind in verschiedenen Textdateien gespeichert, welche wöchentlich aktualisiert werden. Es gibt unter anderem das PDB File, das XML File und das FASTA File, auch FASTA-Format genannt. Dieses besteht aus einer Annotationszeile, die mit > beginnt und der sogenannten Sequenzzeile, welche die Abfolge der Basen in 5'-3'-Richtung darstellt (Abbildung 23). Jede Struktur, die in der PDB veröffentlicht wird, erhält eine vierstellige alphanumerische Kennung, die PDB ID. [11]

Des Weiteren bietet die Proteindatenbank eine Vielzahl von Handwerkzeugen und Hilfsmitteln an. Die Nutzer können diese verwenden, um eine einfache oder erweiterte Suche anhand von Annotationen bezüglich Sequenz, Struktur und Funktion durchzuführen. Dies sind meist sogenannte Open-Source-Programme, wie VMD, MDL Chime, Swiss-PDB Viewer, StarBiochem Sirius und VisProt3DS. [11]

3 Data-mining

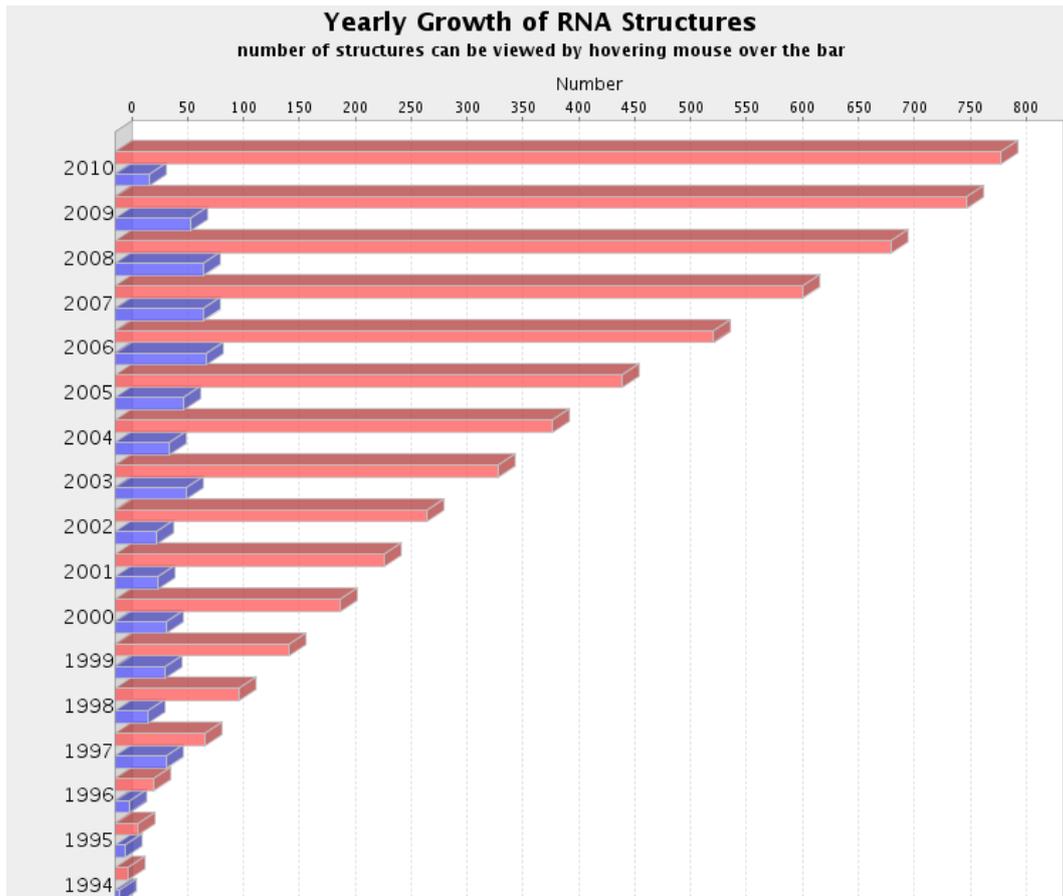


Abbildung 22: Jährliches Wachstum von RNA-Strukturen in der Proteindatenbank. Blau: jährliches Wachstum, Rot: Gesamtanzahl an RNA-Strukturen [http://www.pdb.org/pdb/statistics/contentGrowthChart.do?content=molType-rna&seqid=100], Stand 15.06.2010

bmr_4120-1A60.fasta.txt - Editor

Datei Bearbeiten Format Ansicht ?

```
>1A60 : A | PDBID | CHAIN | SEQUENCE  
GGGAGCUCAACUCUCCCCCCC UUUCCGAGGGUCAUC  
GGAACCA
```

Abbildung 23: FASTA File von 1A60.

3.1.2 BMRB

Die Biological Magnetic Resonance Data Bank ist eine öffentlich-zugängliche Datenbank und umfasst neben NMR Ergebnissen von Peptiden, Proteinen und Nukleinsäuren auch Quellenangaben und verschiedene Informationen zu den NMR Experimenten sowie Computersoftware für biomolekulare NMR. Sie wird von der Universität in Wisconsin-Madison verwaltet und beinhaltet momentan 6083 Einträge zu Proteinen, 166 zu DNA und 124 RNA-Einträge. Im Jahre 2006 trat sie der wwPDB bei.

Die chemischen Verschiebungen der Biomoleküle, welche für diese Arbeit von großem Interesse sind, werden in Textdateien gespeichert. In Abbildung 24 ist ein Ausschnitt aus dem STAR File abgebildet in dem diese enthalten sind. [12], [13]

```

loop_
  _Atom_shift_assign_ID
  _Residue_seq_code
  _Residue_label
  _Atom_name
  _Atom_type
  _Chem_shift_value
  _Chem_shift_value_error
  _Chem_shift_ambiguity_code

1      1      G      C8      C      142.1      .      .
2      1      G      H8      H       8.03      .      .
3      1      G      "C1'"   C       92.7      .      .
4      1      G      "H1'"   H       5.65      .      .
5      1      G      "C2'"   C       74.9      .      .
6      1      G      "H2'"   H       4.50      .      .
7      1      G      "C3'"   C       72.0      .      .
8      1      G      "H3'"   H       4.12      .      .
9      1      G      "C4'"   C       82.5      .      .
10     1      G      "H4'"   H       4.23      .      .

```

Abbildung 24:Ausschnitt aus dem STAR File der RNA 1A60.

[<http://www.bmrwisc.edu/cgi-bin/explore.cgi?format=raw&bmrblid=4120>], Stand 30.11.2009

3.2 Datensatz

Erfasst wurden die chemischen Verschiebungen von RNA-Strukturen, welche sowohl in der PDB, als auch in der BMRB verzeichnet sind. Dabei konnten nur RNA-Strukturen verwendet werden, die keine Verunreinigungen aufwiesen und von denen die Wasserstoffbrückenbindungen berechnet wurden, woraufhin die in den Datenbanken erfassten RNA-Nukleinbasen aufgrund objektiver Kriterien als Watson-Crick- oder non-Watson-Crick-Basen eingestuft wurden. Schließlich blieben von 124 RNA-Einträgen aus der BMRB und PDB 42 reine RNA-Strukturen übrig. Ihre PDB- und BMRB-Identifizierungen sind im Anhang, Tabelle 2 aufgelistet.

Folgende Informationen, die zur Datengruppierung notwendig waren wurden erfasst:

- PDB ID
- BMRB ID
- Art der Hauptbase (A/G/C/U)
- Art der Basenpaarung der Hauptbase (Watson-Crick- oder non-Watson-Crick-Base)
- Art der Vorgängerbase (A/G/C/U/Pu/Py)
- Art der Basenpaarung der Vorgängerbase
- Art der Nachfolgerbase (A/G/C/U/Pu/Py)
- Art der Basenpaarung der Nachfolgerbase
- Nummer der Hauptbase in ihrer Sequenz
- Chemische Verschiebung des C2, H2, C8, H8, C1', H1', C5, H5, C6, H6

Auf der Grundlage von Basentriplets wurden 77 Excel-Dateien erzeugt, in denen, wie in Tabelle 1 ersichtlich, die oben beschriebenen Informationen aufgelistet sind. Bei dem in Tabelle 1 genannten Beispiel ist zu erwähnen, dass es sich bei der Hauptbase um Adenin handelt und es folglich keine C5, H5, C6, H6 chemischen Verschiebungen gibt.

Basentriplets wurden gewählt, um die sequenzielle Abhängigkeit der chemischen Verschiebungen zu bestimmen. Dabei wurde sowohl auf Abhängigkeit des Vorgängers oder des Nachfolgers Bezug genommen als auch auf die Wasserstoffbrückenbindungen der Nukleinbasen.

Die Watson-Crick-Bindung zwischen Guanin und Cytosin ist stabiler als die zwischen Uracil und Adenin, da sie über drei Wasserstoffbrückenbindungen verfügen und somit energetisch günstiger sind. Nukleinbasen die keine Wasserstoffbrückenbindungen eingehen befinden

3 Data-mining

sich meist in einer Schleife (loop) oder sind aufgrund von benachbarten Fehlpaarungen vom gegenüberliegenden Strang weggedreht und somit nicht in der Lage Bindungen auszubilden. Häufig kommt es vor, dass Basen mit mehreren benachbarten Basen Bindungen eingehen, sodass ein genauer Bindungspartner nicht ermittelt werden kann. Besonders Adenin und Guanin, welche die meisten Wasserstoffbrückenbindungen ausbilden können, sind davon betroffen. Allein Adenin besitzt 5 Wasserstoffbrückenbindungsstellen und Guanin sogar 6. [13] Aufgrund der geringen Anzahl an Daten ist es jedoch nicht sinnvoll für die statistische Auswertung auf die verschiedensten Fehlpaarungen einzugehen.

Tabelle 1: Informationen anhand des Basentriplets GAC

WC steht für Watson-Crick-Bindung und nWC für non-Watson-Crick-Bindung. Bei denen mit einem „-“ gekennzeichneten Flächen waren unzureichende Informationen bezüglich der Wasserstoffbrückenbindungen gegeben.

PDB ID	BMRB ID	Vor-gänger	Haupt-base	Nach-folger	Nr.	Vor-gänger WC	Haupt-base WC	Nach-folger WC	C1'	C2	C5	C6	C8	H1'	H2	H5	H6	H8
1bn0	4175	G	A	C	3	WC	WC	WC						6,02	7,8			7,88
1esy	4780	G	A	C	5	WC	nWC	WC	92,969	154,264			139,298	5,917	7,696			7,849
1l1w	5321	G	A	C	5	nWC	nWC	nWC						5,95	8,1			8,02
1l1w	5321	G	A	C	23	WC	nWC	nWC						5,98	8,23			8,02
1ldz	4226	G	A	C	4	WC	WC	WC	93,1	154,3			139,5	5,96	7,76			7,9
1mfj	5559	G	A	C	2	-	-	-	90,952	155,208				5,985	8,014			8,259
1mfj	5559	G	A	C	17	-	-	-	93,041	154,212				5,985	7,838			7,8
1mnx	5919	G	A	C	6	nWC	nWC	nWC	93,4	154,1			139,8	5,92	7,95			7,86
1n8x	5773	G	A	C	3	WC	WC	WC	90,838	151,856			137,163	6,081	7,875			7,923
1n8x	5773	G	A	C	19	nWC	nWC	WC	89,233	153,276			138,671	6,117	8,267			8,283
1na2	5632	G	A	C	19	nWC	WC	nWC	90,63	151,242			135,22	6,081	7,919			7,878
1ow9	5852	G	A	C	9	WC	nWC	nWC	93,12	153,877			139,434	5,897	7,723			7,701
1r7w	6076	G	A	C	6	WC	WC	nWC						6,017	7,841			7,86
1r7w	6076	G	A	C	22	nWC	nWC	WC						6,169	8,291			8,297
1r7z	6077	G	A	C	6	WC	nWC	WC						5,986	7,796			7,824
1r7z	6077	G	A	C	22	nWC	nWC	WC						6,123	8,25			8,285
1s9s	6094	G	A	C	39	-	-	-	89,023	153,825			138,469	6,084	8,171			8,222
1t28	6115	G	A	C	26	nWC	nWC	nWC						5,93	7,66			7,8
1tjz	6239	G	A	C	5	nWC	nWC	WC	90,2	141,6			138,2	5,87	7,66			8,43
1ymo	6477	G	A	C	19	nWC	WC	WC	92,615	153,866			139,186	5,946	7,743			7,831
2fdt	10018	G	A	C	34	nWC	WC	WC						5,914	7,862			7,884
2tpk	4253	G	A	C	5	nWC	WC	WC	89	151,19			136,35	6,07	7,86			8,01

3.3 Vorgehensweise bei der Datenanalyse

Um einen groben Überblick über die Vorgehensweise in dieser Arbeit zu geben, wurde sie in einem Workflow in Abbildung 25 verbildlicht. Die gesammelten Daten werden im Kapitel 4 analysiert und ausgewertet.

In der ersten „Ebene“ der Abbildung sind in den farbig umrahmten Kästen die einzelnen untersuchten Kreuzsignale von chemischen Verschiebungen veranschaulicht. Dabei zeigen die vier kleineren Kästen links in der Abbildung die Kreuzsignale nur innerhalb der Hauptbase, also entweder zwischen den chemischen Verschiebungen in der Nukleinbase oder im Nucleosid zwischen Nukleinbase und dem dazugehörigen Zucker (siehe Kapitel 2.3.1). Die beiden Kästen rechts in der Abbildung veranschaulichten einerseits die Signale zwischen dem Atom eines Vorgängers ($i-1$) und dem der Hauptbase (i) und andererseits die Kreuzsignale zwischen einer chemischen Verschiebung der Hauptbase und einer Resonanz mit dem Nachfolger ($i+1$).

Die mittlere „Ebene“ der Abbildung stellt die Schrittfolge zur Analyse der Kreuzsignale dar. Auf der linken Seite war zuerst (blau) nur die Hauptbase (XHX) von Bedeutung. Vorgänger und Nachfolger waren daher alle Nucleotide. Somit war es möglich allgemeine Aussagen über die Kreuzsignale der chemischen Verschiebungen zu treffen (siehe Kapitel 4.2.2). Außerdem war von Interesse ob sich das Nucleotid, bzw. die chemische Verschiebung der Hauptbase im „stem-Bereich“ und somit „WC“ oder im „loop-Bereich“ und somit „nWC“ liegt (siehe Kapitel 4.2.4). Als nächstes wurden Dupletts (grün), also zwei aufeinanderfolgende Nucleotide untersucht. Dabei waren die Kriterien dieselben, wie die eben beschriebenen. Schließlich wurden nun noch Tripletts (orange), drei aufeinanderfolgende Nucleotide, mit den beiden Kriterien analysiert.

Auf der rechten Seite wurden ebenfalls zwei Kriterien sowohl mit Dupletts (rosa) als auch mit Tripletts (gelb) analysiert. Ebenso waren die Sekundärstruktur der Nucleotide und die Kreuzsignale der chemischen Verschiebungen zwischen Vorgänger und Hauptbase und die zwischen Hauptbase und Nachfolger von Interesse.

In der untersten „Ebene“ sind zwei entstandene Streudiagramme abgebildet. Diese wurden mithilfe des in Kapitel 3.4 beschriebenen Programms erstellt. Dabei wurden die Mittelwerte, Standardabweichungen, Kovarianzen und Korrelationskoeffizienten der einzelnen chemischen Verschiebungen und den daraus entstandenen Kreuzsignalen berechnet mit denen es schließlich möglich war die Ellipsen zu zeichnen. In Kapitel 4 werden diese Streudiagramme näher beschrieben und ausgewertet.

3 Data-mining

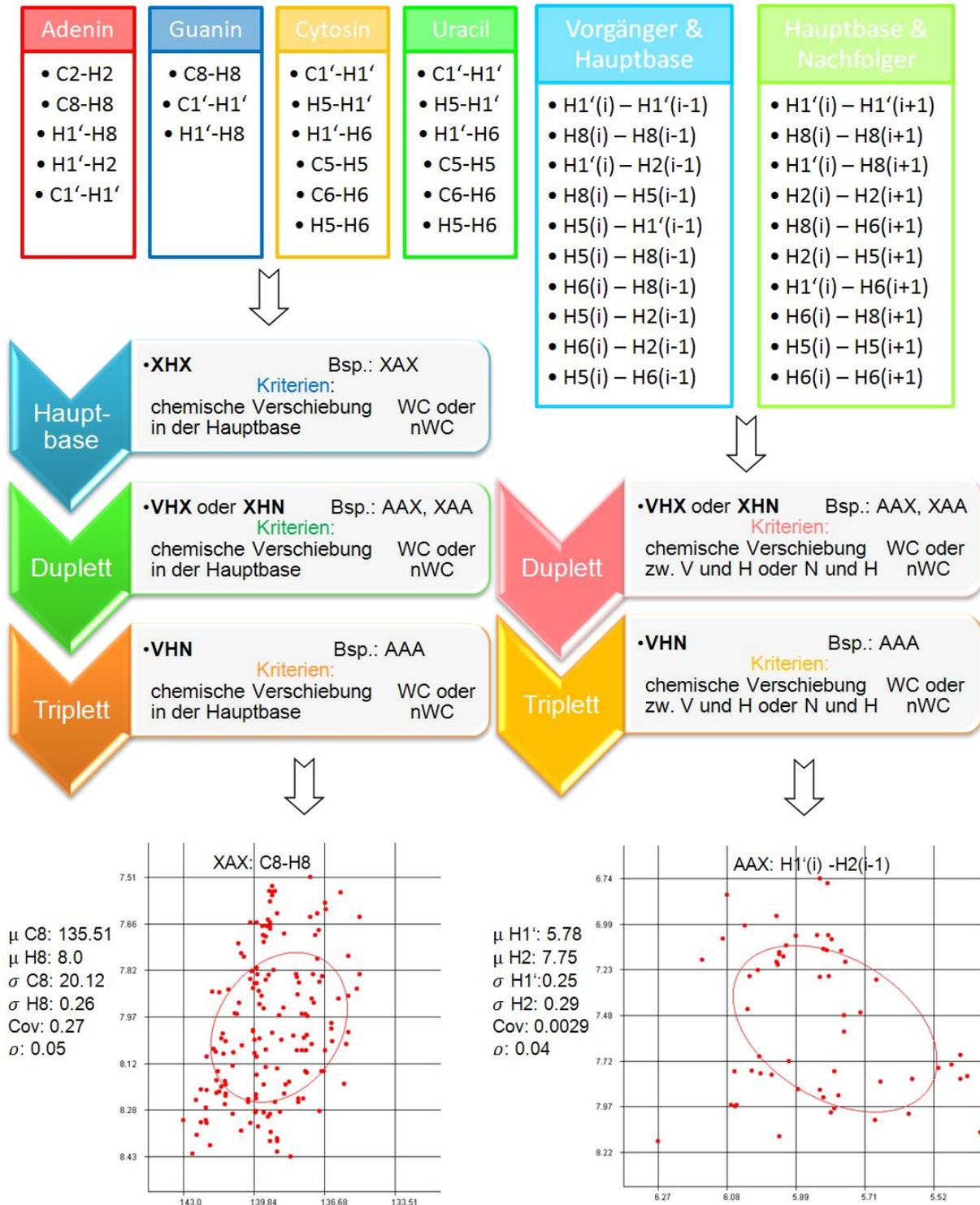


Abbildung 25:Workflow

- oben: Kreuzsignale von chemischen Verschiebungen
- links: Signale innerhalb der Hauptbase
- rechts: Signale zwischen zwei Nukleotiden
- mitte: Schrittfolge der Analyse der Kreuzsignale
- links: nur Hauptbase (blau), Dupletts (grün), Tripletts (orange)
- rechts: Dupletts (rosa), Tripletts (gelb)
- unten: Streudiagramme mit Kovarianzellipsen und berechneten Werten: Mittelwert, Standardabweichung, Kovarianz und Korrelationskoeffizient
- Abkürzungen: V - Vorgänger (i-1), H - Hauptbase (i), N - Nachfolger (i+1), X - alle Nukleotide

3.4 WCStats

Das Programm „WCStats“ dient zur Analyse und Auflistung der einzelnen Informationen, zur Visualisierung der Datenpunkte in Diagrammen und zum Erstellen von Statistiken. Es wurde in der Programmiersprache Java geschrieben und soll in diesem Kapitel kurz erläutert werden.

3.4.1 Einlesen der Dateien

Die 77 Excel-Dateien (siehe Kapitel 3.2) wurden nacheinander geöffnet und der Inhalt jeweils Zeile für Zeile ausgelesen. Diese Zeilen wurden gleichzeitig in eine ArrayList geschrieben, welche nach dem Durchlauf aller Dateien eine komplette Liste des Datensatzes beinhaltet.

3.4.2 Aufteilung nach Kriterien

Die komplette Liste mit allen Informationen wurde im Anschluss sowohl nach Basentriplets (Vorgänger, Hauptbase, Nachfolger) als auch nach Brückenbindungen (WC oder nWC) sortiert. Als Ausgabe dienten, wie in Abbildung 26 ersichtlich, tabstop-getrennte Textdateien.

PDB ID	BMRB ID	G	A	C	Vorgänger	Hauptbase	Nachfolger	Nummer der Base		
1n8x	5773	G	A	C	19	nWC	nWC	WC	89.233	153
1r7w	6076	G	A	C	22	nWC	nWC	WC		
1r7z	6077	G	A	C	22	nWC	nWC	WC		
1tjz	6239	G	A	C	5	nWC	nWC	WC	90.2	141.6 73.
1ymo	6477	G	A	C	19	nWC	WC	WC	92.615	153
2fdt	10018	G	A	C	34	nWC	WC	WC		
2tpk	4253	G	A	C	5	nWC	WC	WC	89.0	151.19

C8:
Mittelwert: 138.10175
Standardabweichung: 1.235303572674695

C1':
Mittelwert: 90.262
Standardabweichung: 1.652466641115632

H1':
Mittelwert: 6.029857142857144
Standardabweichung: 0.1177927962870924

Abbildung 26: Tabstop-getrennte Textdatei (G_A_C # nWC_X_WC.txt)

Im oberen Teil der Abbildung ist die Auftrennung nach dem Basentriplett GAC und der WC-Bedingung „nWC_X_WC“ ersichtlich. Das X steht dabei sowohl für WC als auch für nWC. Im unteren Abschnitt wurden die Berechnungen anhand der jeweiligen chemischen Verschiebung veranschaulicht.

Der Dateiname jeder dieser tabstop-getrennten Textdateien wurde nach folgendem Schema angegeben: „V_H_N # WC_WC_WC.txt“, wobei „V“ für den Vorgänger, „H“ für die Hauptbase und „N“ für die Nachfolgerbase steht. Danach kommt die Information über die Wasserstoffbrückenbindungen des Basentriplets. Anstelle von „WC_WC_WC“ könnte auch „WC_nWC_X“ stehen. Das „X“ steht für beide Bindungsmöglichkeiten, egal ob WC oder nWC. Abschließend steht ein „.txt“, welche für eine Textdatei steht.

Neben Basentriplets und Brückenbindungen wurden außerdem die chemischen Verschiebungen und deren Werte analysiert. Dazu wurde eine Klasse erstellt, welche Berechnungen zum Mittelwert, zur Standardabweichung, zur Kovarianz und zum Korrelationskoeffizienten durchführt. Zuerst wurden alle zu untersuchende Zeilen dieser Klasse und die entsprechenden Verschiebungen, die untersucht werden sollen, übergeben. Da aus den vorhandenen experimentellen Daten nicht immer Werte der chemischen Verschiebungen vorhanden waren, wurden diese Daten bereinigt, was so viel bedeutet, dass Zeilen mit fehlenden Einträgen verworfen wurden. Die jeweiligen berechneten Werte wurden dann an die erzeugten Textdateien angehängt um einen schnellen Überblick zu erhalten (siehe Abbildung 26).

3.4.3 Visualisierung der Datenpunkte

Zur Darstellung der gesammelten und berechneten Daten wurden zusätzlich einige Klassen erzeugt, welche zuerst die darzustellenden Daten erfassen und normieren um sie anschließend in ein Diagramm einzutragen und abzuspeichern.

Abbildung 27 zeigt eines dieser Diagramme, welche jeweils zwei oder mehr Datengruppen darstellen. Dabei wurden immer Wertespaaare betrachtet, wobei eine chemische Verschiebung auf der x-Achse und die andere auf der y-Achse visualisiert wurde. Das entstehende Diagramm bildet eine Punktwolke ab. Desweiteren wurde die Möglichkeit implementiert eine Ellipse über den Schwerpunkt der meisten Datenpunkte zu legen. Mit 50%iger Wahrscheinlichkeit gehören dann diese Datenpunkte mindestens zu der jeweiligen Datengruppe, wenn sie innerhalb der Ellipse liegen. Unter einer Datengruppe versteht man, dass die jeweiligen Datenpunkte dieselben Eigenschaften haben, beispielsweise gehören in Abbildung 27 alle grünen Datenpunkte zu der Datengruppe „UAC_X_X_X“.

Den einzelnen Datengruppen wurden feste Farbwerte zugeordnet um ein einheitliches Muster für alle Diagramme zu erhalten. Falls für eine Datengruppe weniger als 3 Punkte vorhanden waren, wurde das Diagramm verworfen.

3 Data-mining

Die Dateinamen der Diagramme richten sich nach den Namen der Textdateien, aus denen das Diagramm erstellt wurde. Der Dateiname „P_A_X # C # X_X_X # C8-H8.png“ der Abbildung 27 entsteht exemplarisch durch folgende tabstop-getrennte Textdateien: „C_A_C # X_X_X.txt“, „U_A_C # X_X_X.txt“, „A_A_C # X_X_X.txt“, und „G_A_C # X_X_X.txt“. Das „P“ steht für alle möglichen Partner, also für A, G, C, U und -, das A für die Hauptbase A, das „X“ ist immer die Base, welche zwischen den zwei ersten Rauten steht, demnach C. Danach kommt die Information über die Wasserstoffbrückenbindungen, wobei „X_X_X“ für WC als auch für nWC steht. Das Diagramm veranschaulicht die Datenpunkte, die aus zwei chemischen Verschiebungen, nämlich C8 und H8 von der Hauptbase Adenin entstehen. Im Dateinamen ist diese Information nach der letzten Raute abzulesen.

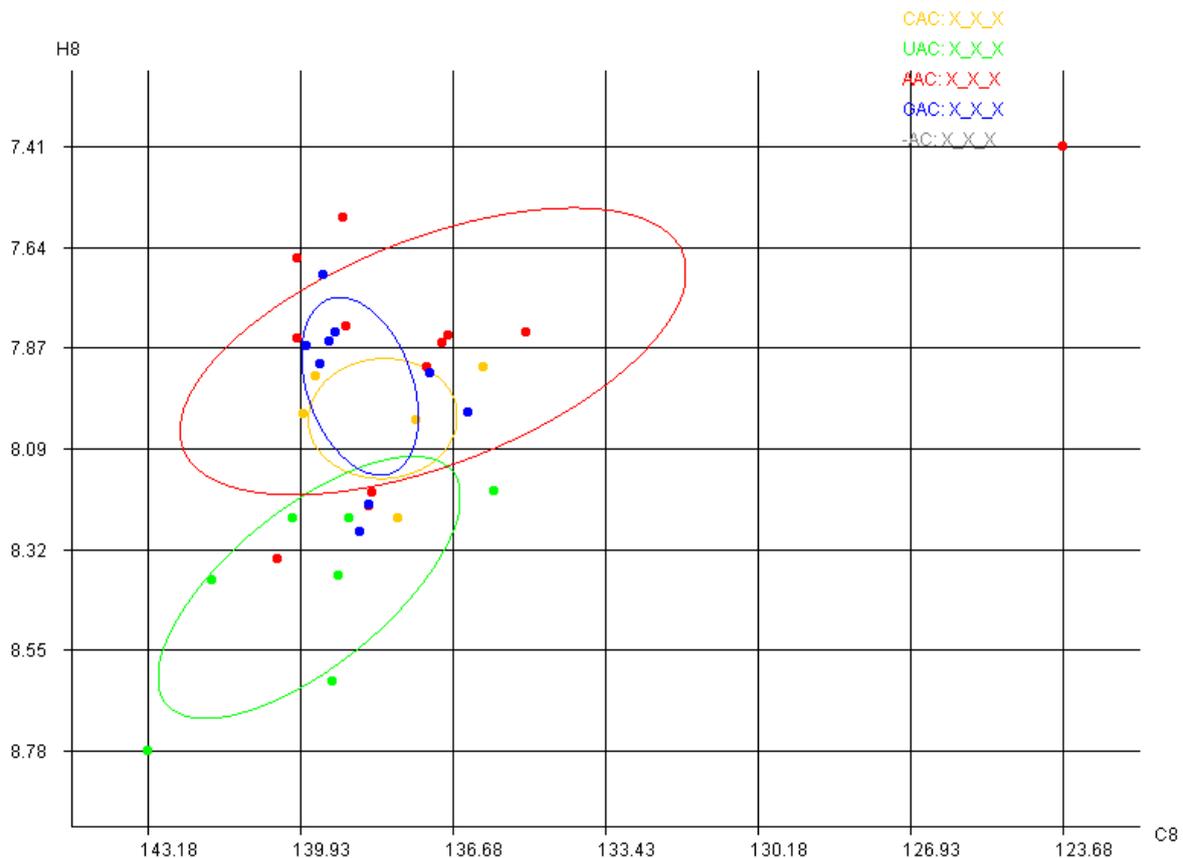


Abbildung 27: Erstelltes Streudiagramm (P_A_X # C # X_X_X # C8-H8.png) mit der Hauptbase A und dem Nachfolger C. Exemplarisch zeigt die blaue Ellipse das Triplet GAC. „X_X_X“ bedeutet, dass die Watson-Crick-Bedingung kein Kriterium war. Laut den Bedingungen rechts oben in der Abbildung hätte ebenfalls eine Ellipse mit „-AC“ entstehen müssen, jedoch waren dafür zu wenig Datenpunkte vorhanden.

3.4.4 Erstellen der TOCSY-, NOESY- und HSQC-Distributionsdatei

Abbildung 28 veranschaulicht eine sogenannte Distributionsdatei. Diese sind notwendig um statistische Analysen, d.h. die prozentuale Zuordnung einer Base, unbekannter RNA-Strukturen durchzuführen. Die durch die Distributionsdatei entstandene Statistik wird in Kapitel 4.4 anhand einer noch unbekanntes RNA-Sequenz und deren chemischen Verschiebungen überprüft.

In diesen Dateien sind alle ermittelten Werte wie folgt aufgeführt. Zuerst beginnt jede Zeile mit einem „X“. Dies dient zur Verdeutlichung, dass die auf der x-Achse berechneten Werte (x-Werte) folgen. Daran angehängt steht die Bedingung, d.h. um welches Basentriplett, welche Art der Wasserstoffbrückenbindungen und um welche chemische Verschiebung es sich handelt. Im Anschluss folgen der Mittelwert und die Standardabweichung der x-Werte. Das „Y“ gibt an, dass nun die auf der y-Achse berechneten Werte (y-Werte) erfasst sind.

```
X XUX_X_X_X(H5X-H6Y) 5.48543438914027 0.2725829840170693 Y XUX_X_X_X(H5X-H6Y) 7.77827601809955
X XCX_X_X_X(H5X-H6Y) 5.498832775919734 0.252304445041778377 Y XCX_X_X_X(H5X-H6Y) 7.69334444816053
X XUX_X_X_WC(H5X-H6Y) 5.481648648648648 0.27580072379297144 Y XUX_X_X_WC(H5X-H6Y) 7.82955405405
X XCX_X_X_WC(H5X-H6Y) 5.484630434782608 0.24315909490944837 Y XCX_X_X_WC(H5X-H6Y) 7.74903260869
X XUX_X_X_nWC(H5X-H6Y) 5.502977777777778 0.26722321510605573 Y XUX_X_X_nWC(H5X-H6Y) 7.74376666666
X XCX_X_X_nWC(H5X-H6Y) 5.507146551724139 0.2937707847761589 Y XCX_X_X_nWC(H5X-H6Y) 7.6337586206
X XUX_X_WC_X(H5X-H6Y) 5.396030769230769 0.25849117303778074 Y XUX_X_WC_X(H5X-H6Y) 7.82269230769
X XCX_X_WC_X(H5X-H6Y) 5.451610294117645 0.237890643186113 Y XCX_X_WC_X(H5X-H6Y) 7.6852794117647
X XUX_X_WC_WC(H5X-H6Y) 5.4325 0.27591888440120754 Y XUX_X_WC_WC(H5X-H6Y) 7.862175000000001 0.17
X XCX_X_WC_WC(H5X-H6Y) 5.450763157894736 0.2423794748417665 Y XCX_X_WC_WC(H5X-H6Y) 7.7248289473
X XUX_X_WC_nWC(H5X-H6Y) 5.349695652173913 0.21553789685398672 Y XUX_X_WC_nWC(H5X-H6Y) 7.7782608
```

Abbildung 28: Aufbau der Distributionsdatei für das TOCSY-Experiment

Die nächsten drei Informationen der y-Werte unterscheiden sich nicht von denen der x-Werte. Zusätzlich werden am Schluss jeder Zeile die Kovarianz und der Korrelationskoeffizient angehängt.

Die Distributionsdateien wurden nur erstellt, wenn mindestens 10 Einträge in den tabstop-getrennten Textdateien vorlagen.

3.4.5 Implementieren der Wahrscheinlichkeitsformel zum Auswerten der Statistik

Zur Berechnung der Wahrscheinlichkeit der Zugehörigkeit eines Wertepaares X_1 und X_2 wurden weitere Klassen im Programm „WCStats“ implementiert. Grundlage bildeten die Distributionsdateien der Experimente und die Wertepaare X_1 und X_2 , die chemischen Verschiebungen. Anhand Gleichung (10) wurde die Wahrscheinlichkeit berechnet.

(10)

$$P(X_1, X_2) = \exp \left\{ -\frac{1}{2} \left[\left(\frac{x_1 - \mu_1}{\sigma_1} \right)^2 - 2\rho \left(\frac{x_1 - \mu_1}{\sigma_1} \right) \left(\frac{x_2 - \mu_2}{\sigma_2} \right) + \left(\frac{x_2 - \mu_2}{\sigma_2} \right)^2 \right] \right\}$$

Abbildung 29 verdeutlicht eine der möglichen Ausgabedateien. In den ersten beiden Spalten der Textdatei stehen die Werte der chemischen Verschiebungen X_1 und X_2 . Im Anschluss folgen die fünf bestmöglichen Ergebnisse der Statistik. An dem bekannten Schema ist zu erkennen, um welche Datengruppe es sich dabei handelt. In Klammern ist angegeben zu wie viel Prozent ein Wertepaar zu einer bestimmten Datengruppe gehört.

```

5.066 7.714:      PuUX_WC_WC_WC(H5X-H6Y) (95.22)      GUC_X_X_X(H5X-H6Y) (78.87)
5.675 7.802:      XUC_nWC_nWC_nWC(H5X-H6Y) (99.99)      GUX_nWC_nWC_nWC(H5X-H6Y) (99.
4.932 7.597:      XUX_WC_WC_nWC(H5X-H6Y) (58.49)      XUX_WC_WC_X(H5X-H6Y) (56.07)
5.378 7.561:      PuCX_nWC_nWC_nWC(H5X-H6Y) (99.97)      XCX_WC_WC_nWC(H5X-H6Y) (99.6)
5.753 7.726:      XCG_nWC_nWC_nWC(H5X-H6Y) (94.79)      XCA_nWC_nWC_nWC(H5X-H6Y) (93.
4.979 7.445:      XCX_nWC_WC_WC(H5X-H6Y) (80.59)      GCG_X_X_X(H5X-H6Y) (77.9)
5.16 7.452:      GCG_X_X_X(H5X-H6Y) (97.78)      AUX_nWC_nWC_nWC(H5X-H6Y) (91.6)
5.407 7.733:      XCC_WC_WC_WC(H5X-H6Y) (99.93)      XUX_WC_X_nWC(H5X-H6Y) (99.4)
5.466 7.65:      XUPu_nWC_nWC_nWC(H5X-H6Y) (99.88)      XCX_X_WC_nWC(H5X-H6Y) (99.36)

```

Abbildung 29: Aufbau einer Statistik-Datei der RNA FZL4 für das TOCSY-Experiment.

4 Resultate und Diskussion

4.1 Kovarianzellipsen

In Kapitel 2.4 wurde beschrieben, wie man eine zweidimensionale Normalverteilung und somit auch die Wahrscheinlichkeitsdichtefunktion durch Konturen in Form von Ellipsen visualisieren kann. Sie dienen zur Auswertung von Statistiken, geben aber auch eine visuelle Hilfestellung zum Zuordnen einzelner Kreuzsignale zu einem Datencluster.

In Abbildung 30 wurden die Kovarianzellipsen mit der einfachen Standardabweichung der dazugehörigen Datenpunkte einer Gruppe in einem Streudiagramm veranschaulicht. Dabei gehören die in den Ellipsen liegenden Datenpunkte mit einer Wahrscheinlichkeit von mindestens 50% auch zu der bestimmten Datengruppe, welche die Ellipse veranschaulicht. Laut der zweidimensionalen Normalverteilung liegen 63% der Kreuzsignale einer Datengruppe in der dazugehörigen Kovarianzellipse.

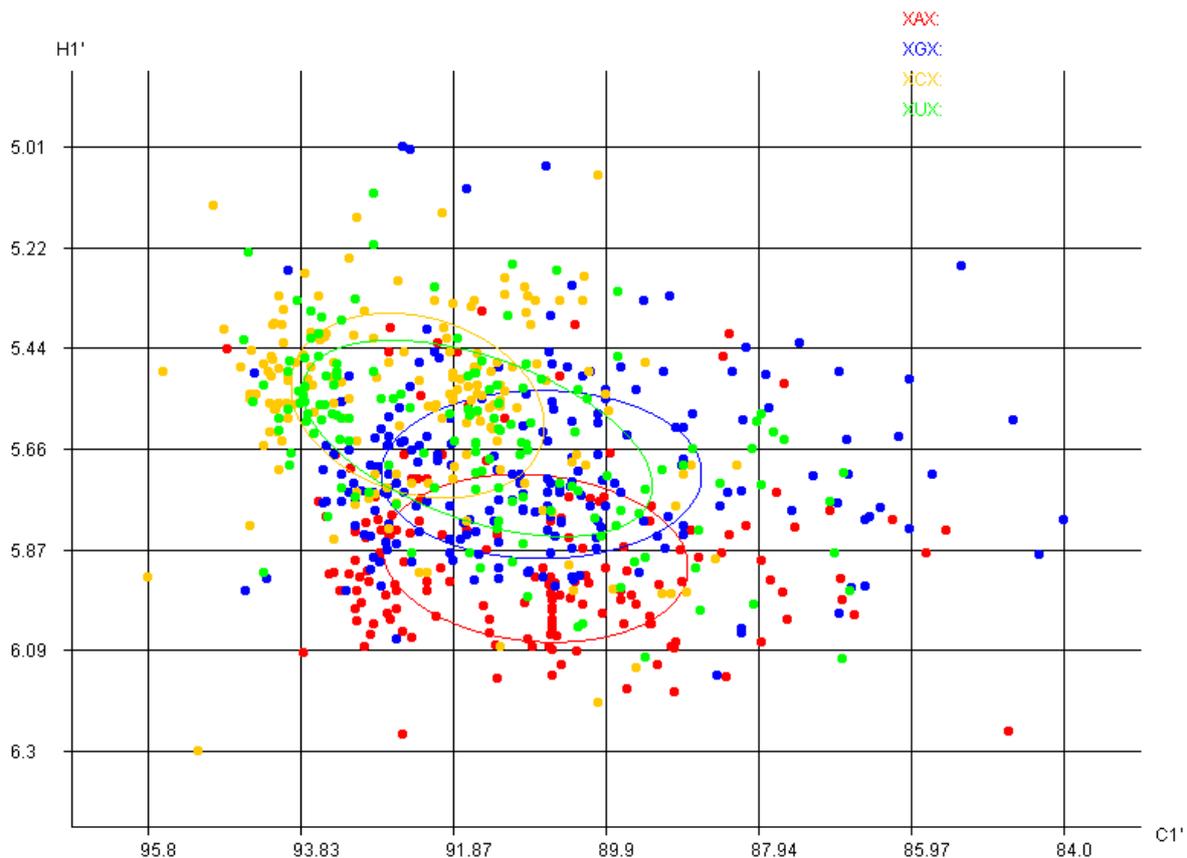


Abbildung 30: Streudiagramm mit den chemischen Verschiebungen C1' und H1' der Hauptbasen A, G, C und U. Vorgänger und Nachfolger sind alle Nucleotide, angegeben mit X.

4 Resultate und Diskussion

Durch die bei der Visualisierung der Datenpunkte entstehende Cluster, wurde nach einer mathematischen Formel gesucht, um diese automatisch zu analysieren, unterscheiden und auswerten zu können. Jedoch stellte sich dies als ein größeres Problem heraus. Da sich die meisten Kovarianzellipsen verschiedener Cluster überschneiden kam man auf die Idee, die Überschneidungsflächen näher zu untersuchen. Dabei wurde festgestellt, dass sich die einzelnen Überschneidungsflächen arg von Umfang unterscheiden. Abbildung 31 veranschaulicht verschiedenen Cluster und deren Überschneidungen.

Separierten sich die Cluster bei manchen Abbildungen, konnte mit dieser Idee nicht weitergearbeitet werden, da nun keine Überschneidungsflächen oder weniger als bei anderen auftauchten.

Weiterhin stellte sich die Frage ob man über die Anzahl an Überlappungen einen Faktor finden könnte. Auch diese Idee konnte nicht umgesetzt werden.

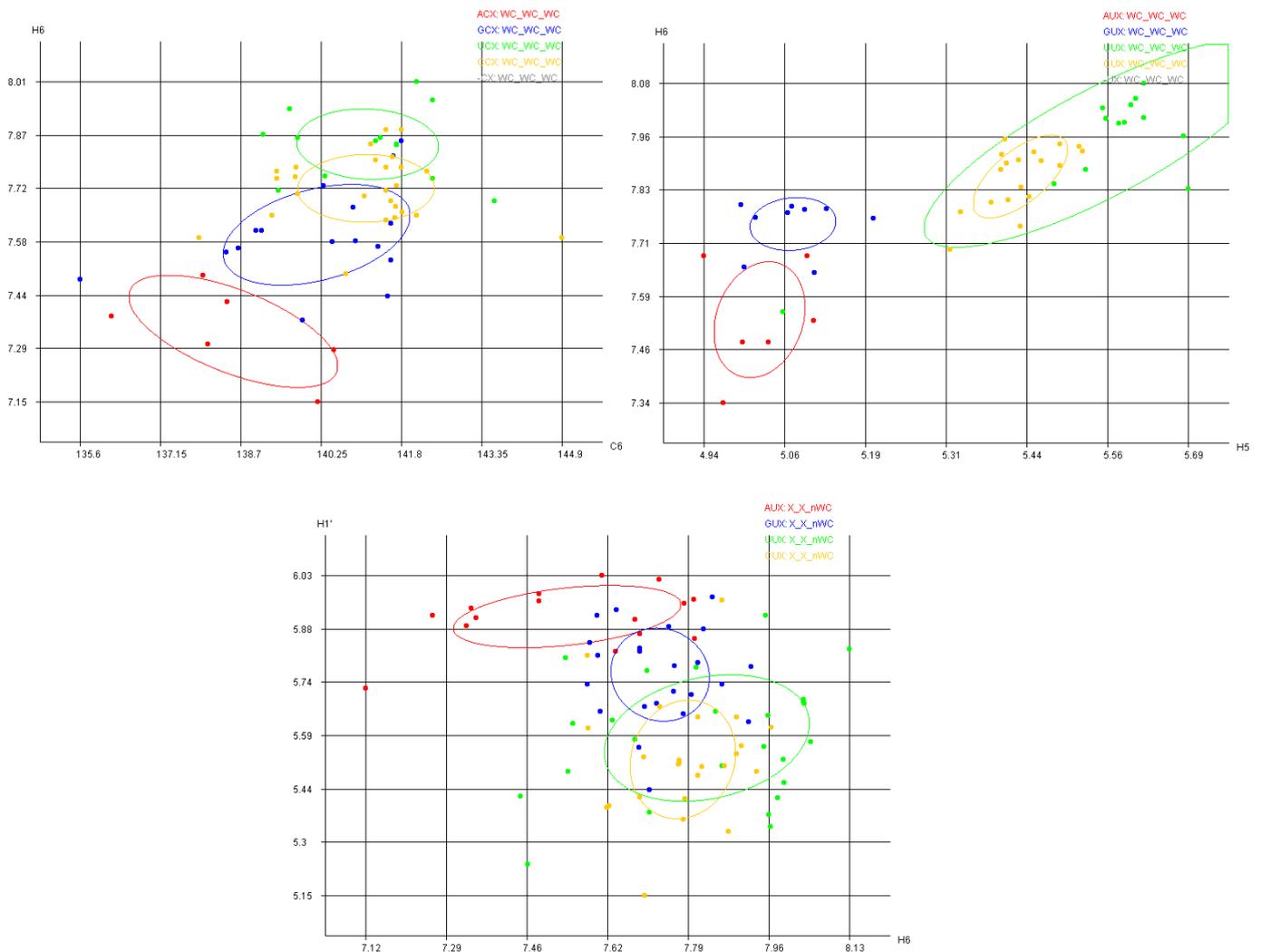


Abbildung 31: Verschiedene Streudiagramme mit verschiedenen Clustergrößen und Überschneidungsflächen

4.2 Statistik der ^1H - und ^{13}C -chemischen Verschiebungen in einem Nukleotid

Für die Zuordnungen der chemischen Verschiebungen wurden die Resonanzen der H1', H2, H5, H6, und H8 Protonen sowie der C1', C2, C5, C6 und C8 Kohlenstoffatomen analysiert. Abbildung 25 oben links veranschaulicht die untersuchten Kreuzsignale von chemischen Verschiebungen in den Hauptbasen.

In vorangegangenen Studien wie die von Cromsigt & Hilbers 2001 wurde erkannt, dass die chemischen Verschiebungen der Nukleinbasen mit ihrem 5'-Nachbar korrelieren und nicht mit der 3'-Nachbarnukleinbase. Im Gegensatz dazu fand man heraus, dass die C2 und H2 Resonanzen des Adenosins von beiden Nachbarn beeinflusst werden. Diese Abhängigkeiten wurden in dieser Arbeit nochmals überprüft.

Ein Großteil der RNA-Strukturen beginnt mit zwei Guanosinen am 5'-Ende der RNA und enden mit zwei Cytidinen am 3'-Ende der RNA. Die chemischen Verschiebungen in diesen Nukleotiden wurden ferner untersucht. Desweiteren wurde die chemischen Verschiebungen dahingegen analysiert, inwieweit ihre Resonanzen eine Kriterium für eine Watson-Crick-Bindung oder keine Watson-Crick-Bindung sind.

4.2.1 2D-Statistik

Eine zweidimensionale Statistik ist eine Statistik, bei der sowohl auf der x- als auch auf der y-Achse je eine chemische Verschiebung aufgetragen wird und beide zusammen Kreuzsignale ausbilden. Die dabei entstandenen Cluster enthalten genügend Werte um statistische Aussagen treffen zu können. Die Anzahl der Datenpunkte in einem Diagramm veranschaulicht dabei die statistische Relevanz der einzelnen Cluster.

Abbildung 32 zeigt ein Streudiagramm der Kreuzsignale aus H5- und H6-chemischen Verschiebungen von Cytidin und Uridin mit den vier möglichen Vorgängern Adenosin, Guanosin, Cytidin und Uridin. Zu erkennen sind verschiedene Cluster der einzelnen Datengruppen. Vor allem die Datenpunkte von CCX, CUX, GUX, UUX und GCX grenzen sich von den anderen Clustern ab.

Abbildung 33 veranschaulicht ein Streudiagramm von den H1'-H8 chemischen Verschiebungspaaren. Diese Kombination hilft Adenosine und Guanosine zu unterscheiden, da sie sich, wie ersichtlich, in fast unterschiedlichen Regionen befinden. Die H8-Signale des Adenosin verlaufen etwa zwischen 7,6 ppm und 8,4 ppm und die von Guanosin zwischen 6,8 ppm und 7,8 ppm. Sehr gut zu unterscheiden sind GGX und CGX, deren Datenpunkte nicht sehr gestreut und damit komprimiert in einem Bereich liegen.

4 Resultate und Diskussion

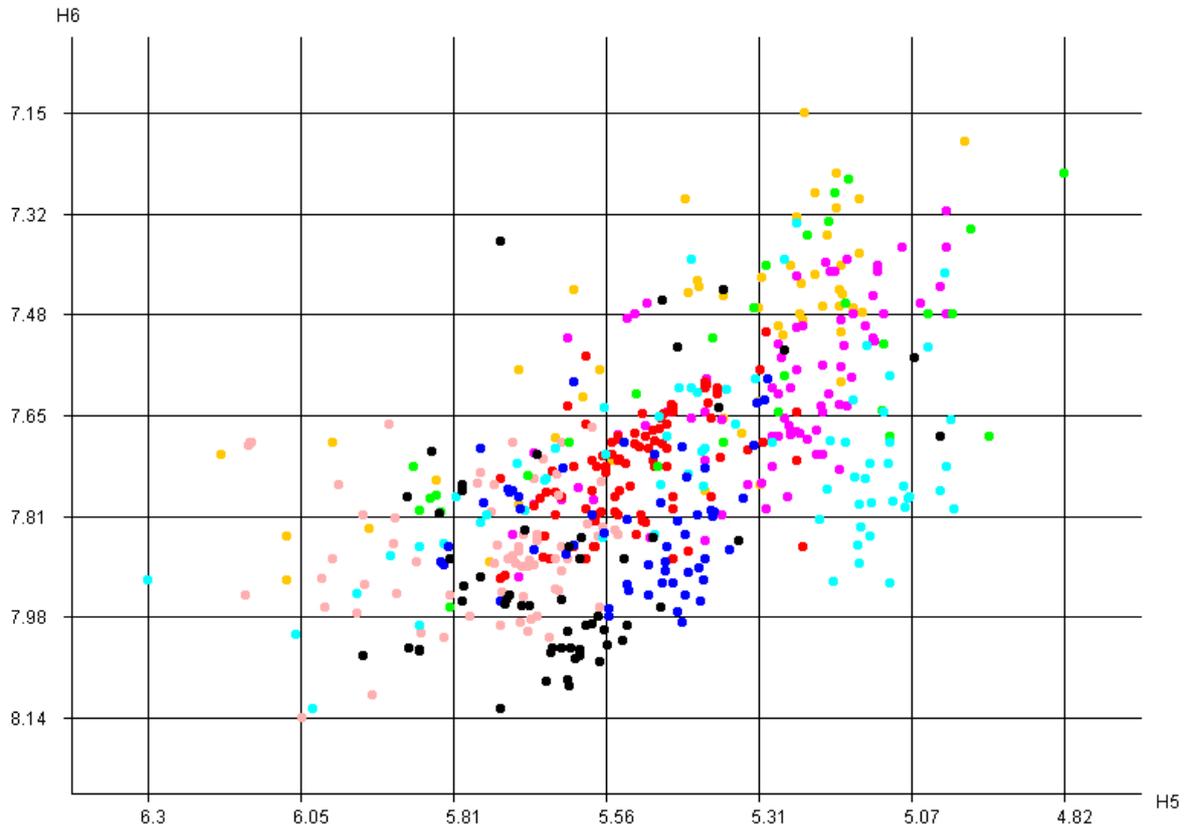


Abbildung 32: Streudiagramm zeigt Kreuzsignale aus der H5-chemischen Verschiebung mit der H6-chemischen Verschiebung von Uracil und Cytosin als Hauptbasen mit den vier möglichen Vorgängern A, C, G und U.

ACX bedeutet die chemischen Verschiebungen von Cytidin, als Vorgänger ein Adenosin und als Nachfolger ein beliebiges Nukleotid, angegeben mit X. Die weiteren Datengruppen sind GCX, CCX, UCX, AUX, GUX, CUX und UUX.

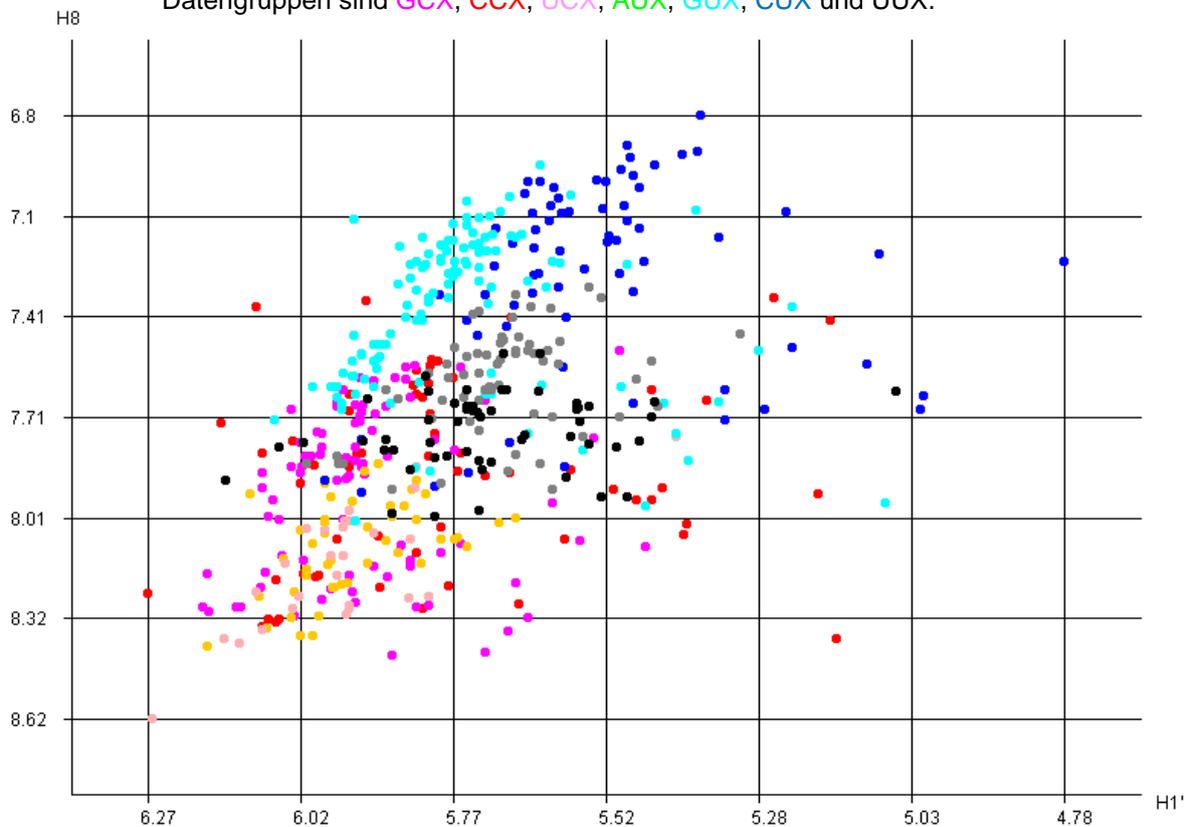


Abbildung 33: Streudiagramm zeigt Kreuzsignale aus der H1'-chemischen Verschiebung mit der H8-chemischen Verschiebung von Adenin und Guanin als Hauptbasen mit den 4 möglichen Vorgängern A, C, G und U. Das X kennzeichnet ein beliebiges Nukleotid als Nachfolger. AAX, GAX, CAX, UAX, AGX, GGX, CGX, UGX

4 Resultate und Diskussion

In Abbildung 34 ist ein Streudiagramm mit dem Kreuzsignal aus H5 und H1' von Cytidin und Uridin gezeigt. H5-chemische Verschiebungen von Cytidin mit einem Purin als Vorgänger sind von den H5 chemischen Verschiebungen von Cytidin mit einem Pyrimidin als Vorgänger gut zu unterscheiden. ACX und GCX liegen etwa 0,2 ppm bis 0,3 ppm voneinander entfernt. Bei Uridin ist dies ähnlich, wie in Abbildung 35 ersichtlich.

Laut Cromsigt & Hilbers korrelieren die H2-chemischen Verschiebungen des Adenosins natürlicherweise sowohl mit dem 5'-und dem 3'-Nukleotid wohingegen das H1'-Proton vom Vorgängernukleotid abhängig ist. In der Studie von Cromsigt & Hilbers wurden für die H1'-H2 Statistiken verschiedenen Cluster untersucht, mit dem Resultat, dass die Cluster Pu-A-Pu, Pu-A-Py, Py-A-Pu und Py-A-Py die beste Trennung zeigten. Die Erkenntnisse wurden in dieser Arbeit nochmal überprüft. Das dabei entstandene Streudiagramm ist in Abbildung 36 veranschaulicht. Es ist eindeutig zu sehen, dass die vier Cluster nicht sehr gut voneinander unterschieden werden können. Einzelne Tendenzen sind zu erkennen, aber keine Trennung der einzelnen Datengruppen. Die Datenpunkte der Gruppe PuAPy liegen als einzige komprimiert beieinander und könnten am besten von den anderen Gruppen abgegrenzt werden.

Abbildung 37 zeigt ein Streudiagramm aller C8-H8-chemischen Verschiebungspaaren von Adenosin und Guanosin einer RNA. Eindeutig zu erkennen sind die unterschiedlichen Cluster von Adenosin und Guanosin. Durch die chemischen Verschiebungen des C8 ist es möglich beide sehr gut von einander zu unterscheiden. Die C8-chemische Verschiebung von Adenosin liegt etwa zwischen 137 ppm und 143 ppm und die von Guanosin zwischen 137 ppm und 130 ppm. Außerdem ist ebenfalls wieder ein kleiner Unterschied zwischen Purinen und Pyrimidinen als Vorgänger zu erkennen, denn das H8-Signal der Purine liegt etwa 0,3 ppm tiefer als das der Pyrimidine.

Im Gegensatz dazu kann man mit Hilfe des C6-H6 Kreuzsignals von Uridin und Cytidin keine Trennung der Cluster erkennen. Abbildung 38 veranschaulicht dies in einem Streudiagramm. Die verschiedenen Gruppen von Cytidinen und Uridinen überschneiden sich zu sehr.

In Abbildung 39 sind die C5-H5 chemischen Verschiebungspaare von Uridin und Cytidin veranschaulicht. Das Streudiagramm zeigt sehr gut getrennte Cluster. Sie könnten später verwendet werden, um Startpunkte für die Zuordnungsstrategie zu erhalten. Da es möglich ist, die H5-H6 Verschiebungen zu verknüpfen ist es wichtig von der C5-chemischen Verschiebung die Informationen zu erhalten, ob das jeweilige Signal von einem C oder einem U stammt. Dies dürfte anhand dieser Abbildung kein Problem sein.

4 Resultate und Diskussion

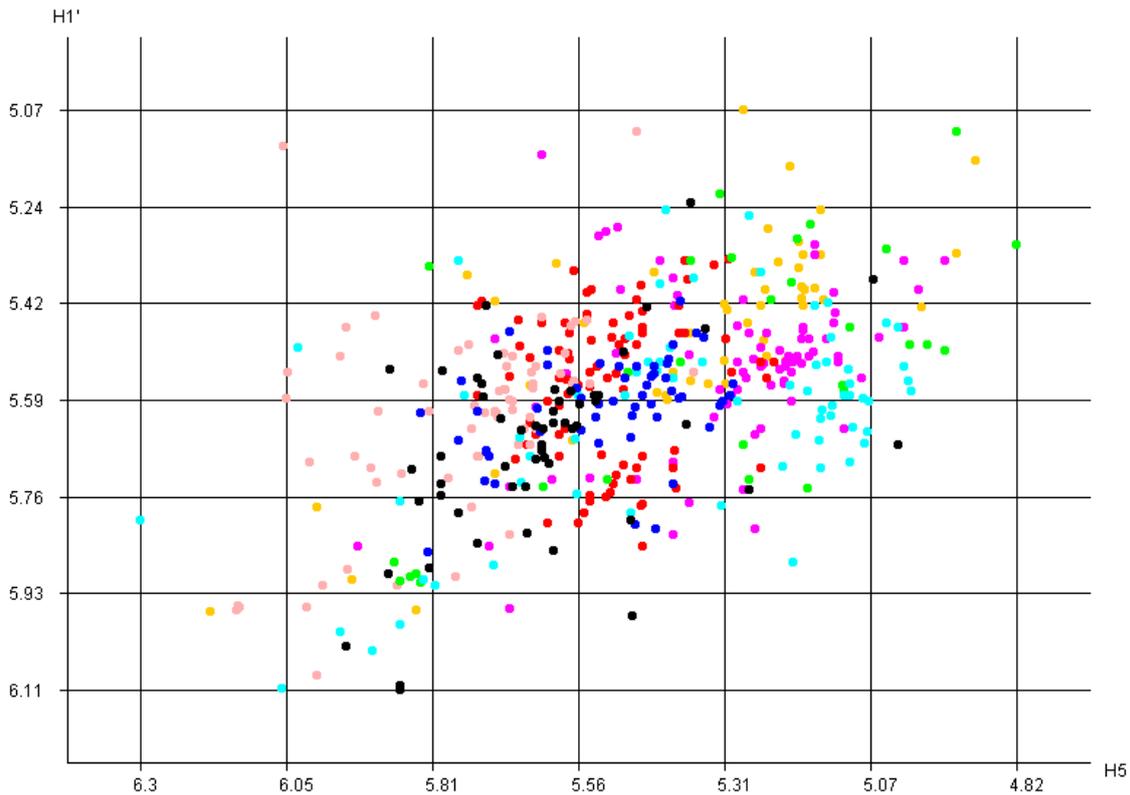


Abbildung 34: Streudiagramm zeigt Kreuzsignale aus dem H5-H1'-chemischen Verschiebungspaar von Uridin und Cytidin. Das X kennzeichnet ein beliebiges Nukleotid als Nachfolger.
 ACX, GCX, CCX, UCX, AUX, GUX, CUX, UUX

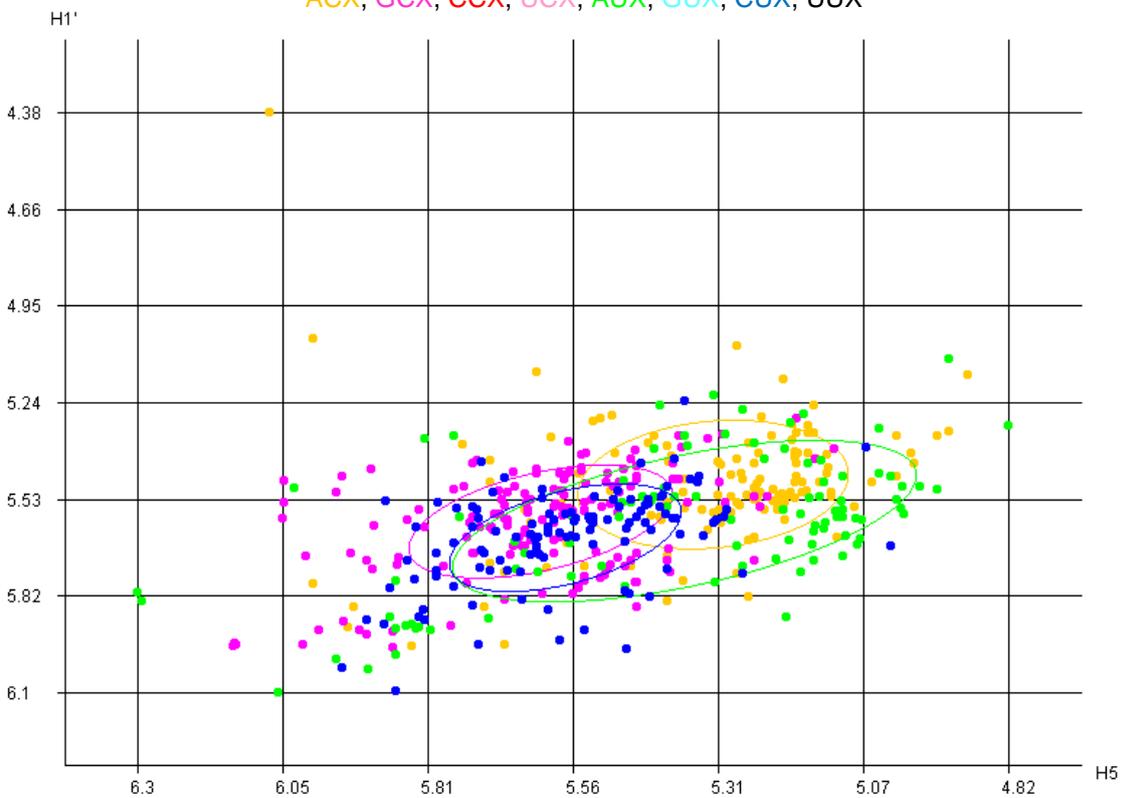


Abbildung 35: Streudiagramm mit der H5 und H1' chemischen Verschiebung veranschaulicht den Unterschied zwischen Purinen und Pyrimidinen als Vorgänger und Uridin und Cytidin als Hauptbase. Das X kennzeichnet ein beliebiges Nukleotid als Nachfolger.
 PuCX, PyCX, PuUX, PyUX

4 Resultate und Diskussion

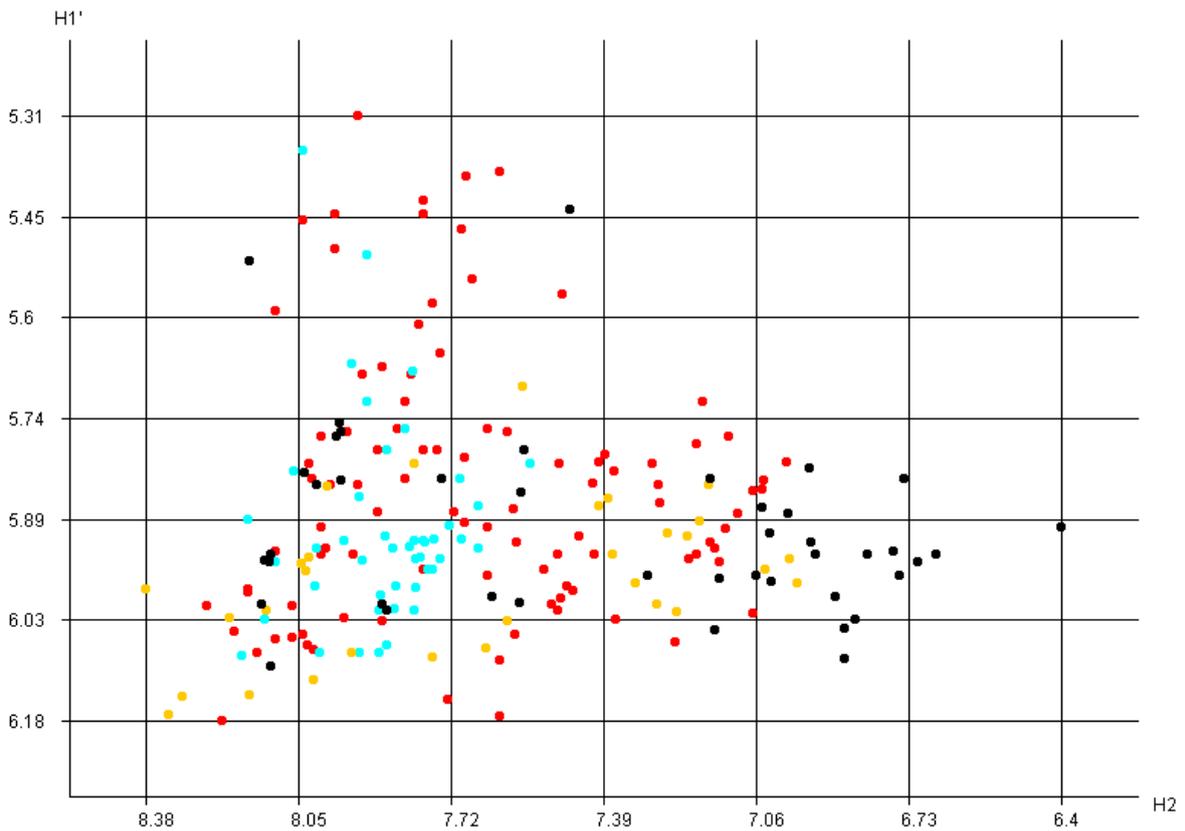


Abbildung 36: Streudiagramm der Kreuzsignale H2-H1' von Adenosin mit Purin (Pu) oder Pyrimidin (Py) als Vorgänger und Purin oder Pyrimidin als Nachfolger. Die einzelnen Datengruppen sind wie folgt bezeichnet: PuAPu, PuAPy, PyAPy, PyAPu.

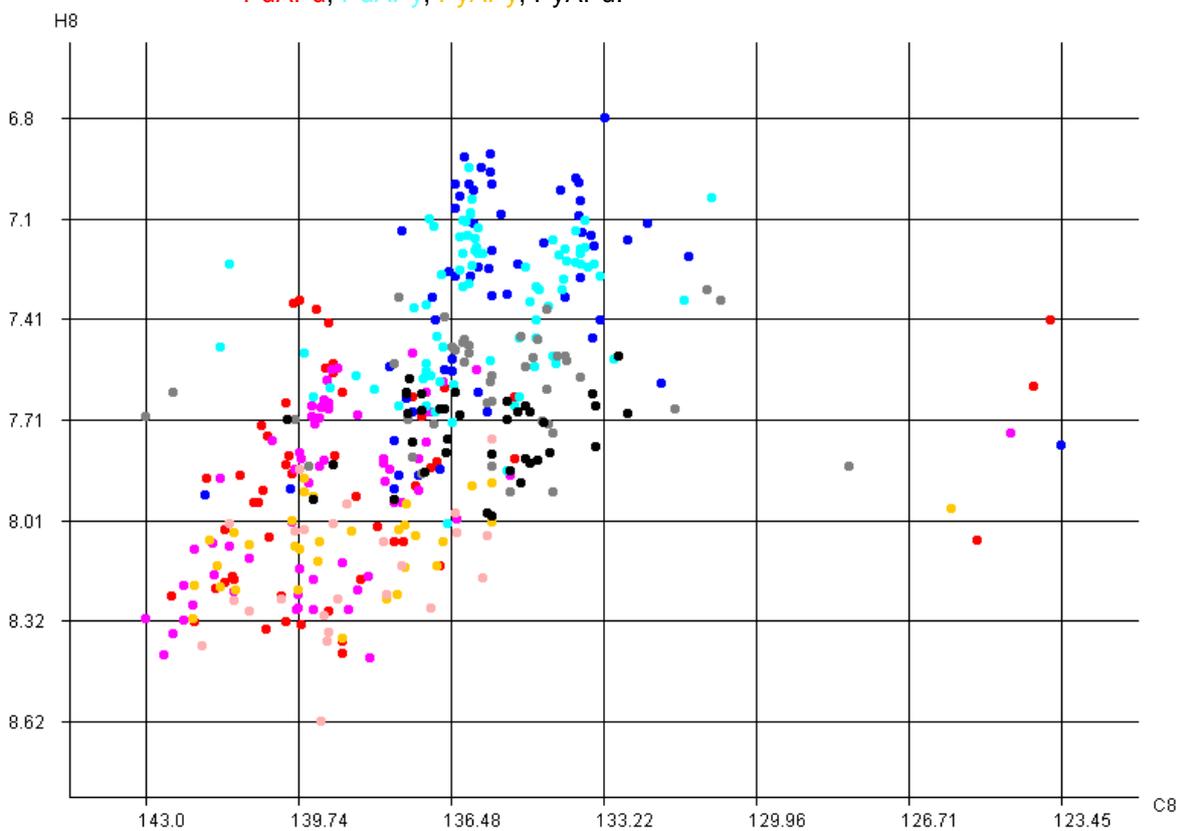


Abbildung 37: Streudiagramm zeigt Kreuzsignale aus der C8-chemischen Verschiebung mit der H8-chemischen Verschiebung von Adenin und Guanin als Hauptbasen mit den vier möglichen Vorgängern A, C, G und U. Das X kennzeichnet ein beliebiges Nucleotid als Nachfolger. AAX, GAX, CAX, UAX, AGX, GGX, CGX, UGX

4 Resultate und Diskussion

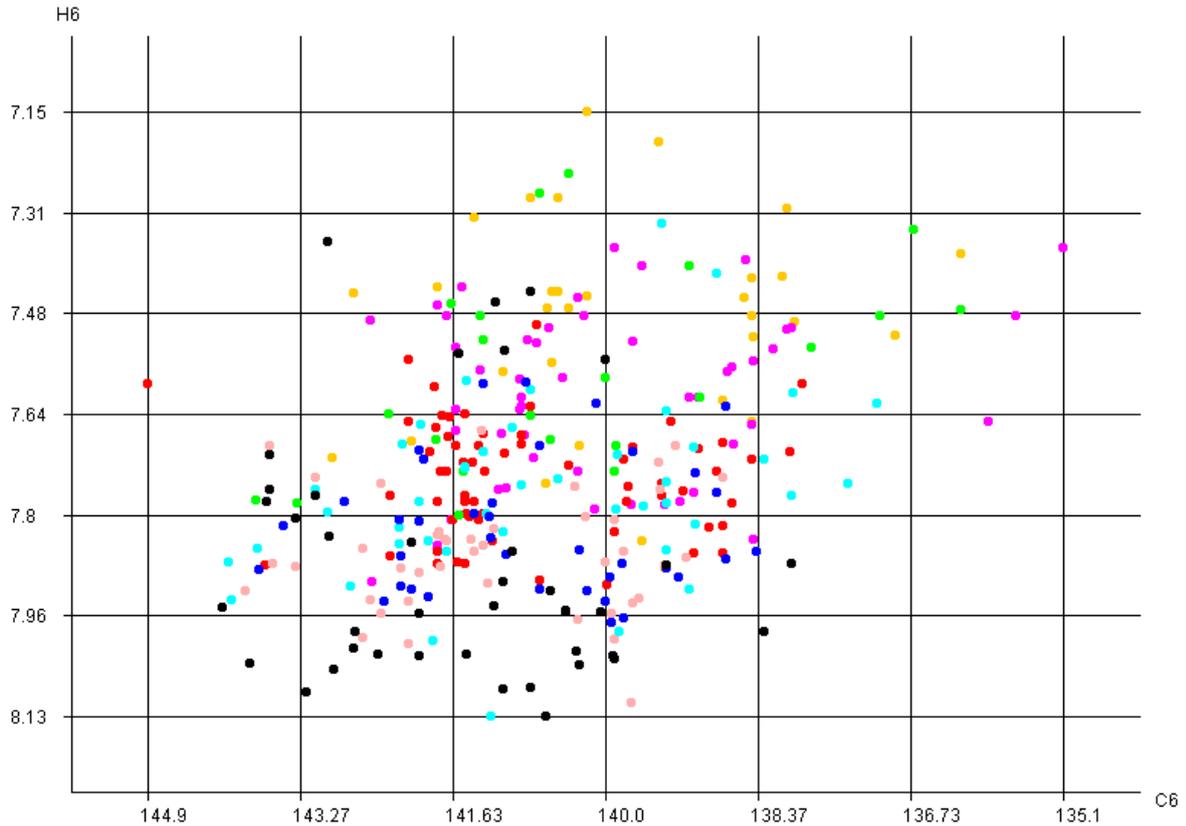


Abbildung 38: Streudiagramm mit den Kreuzsignalen von den chemischen Verschiebungen C6 und H6 von Cytidin und Uridin. Vorgänger sind A, G, C und U. Nachfolger sind beliebige Nukleotide, angegeben mit X.
 ACX, GCX, CCX, UCX, AUX, GUX, CUX, UUX

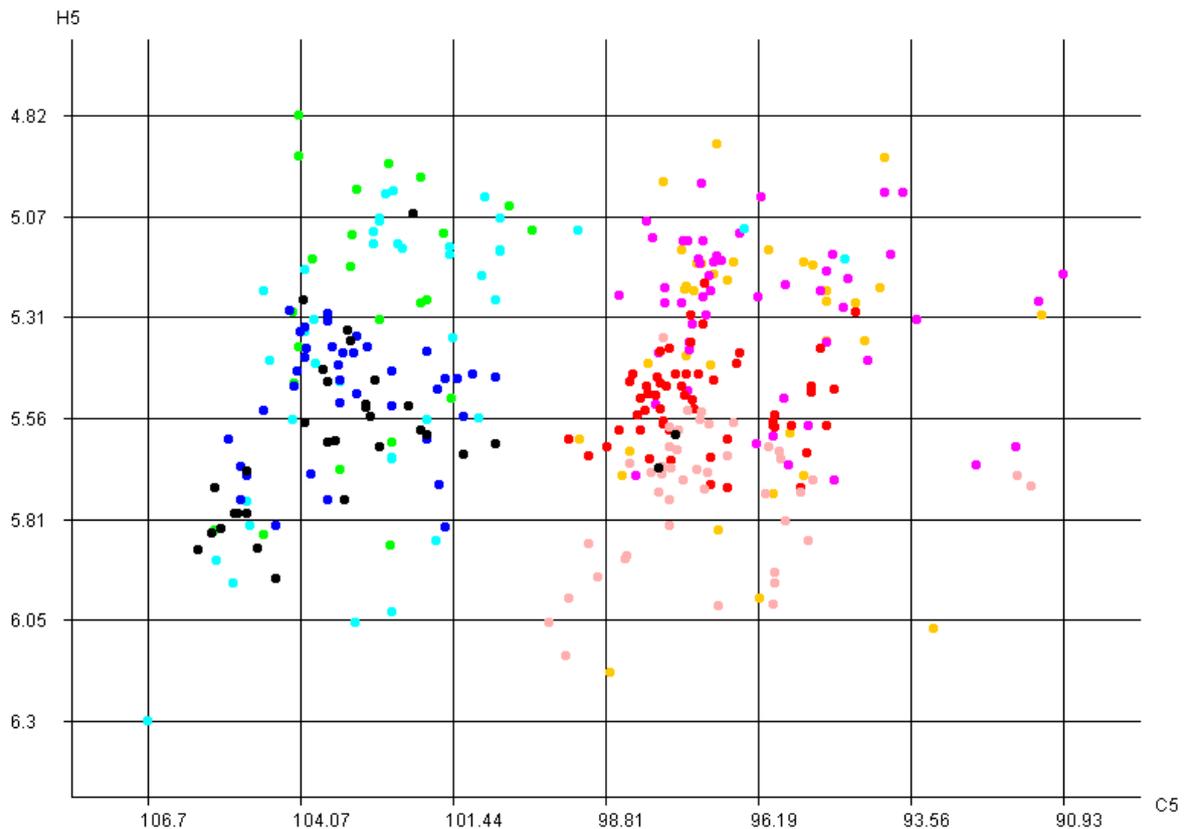


Abbildung 39: Streudiagramm des Kreuzsignals C5-H5 der Hauptbasen Uracil und Cytosin mit den Vorgängern A, G, C und U. Das X kennzeichnet ein beliebiges Nucleotid als Nachfolger.
 ACX, GCX, CCX, UCX, AUX, GUX, CUX, UUX

4.2.2 Abhängigkeit der chemischen Verschiebung in der Hauptbase

In früheren Studien [25] wurde verdeutlicht, dass der Vorgänger vermutlich einen größeren Einfluss auf die chemische Verschiebung der Hauptbase hat als der Nachfolger. Jedoch gibt es zwei Ausnahmen, das C2 und H2. Abbildung 40 veranschaulicht das H2-Resonanzsignal. Erkennbar ist in beiden Streudiagrammen, dass sich die einzelnen Cluster nicht sehr gut trennen lassen. Folglich wird das H2 offenbar von beiden Nachbarn beeinflusst.

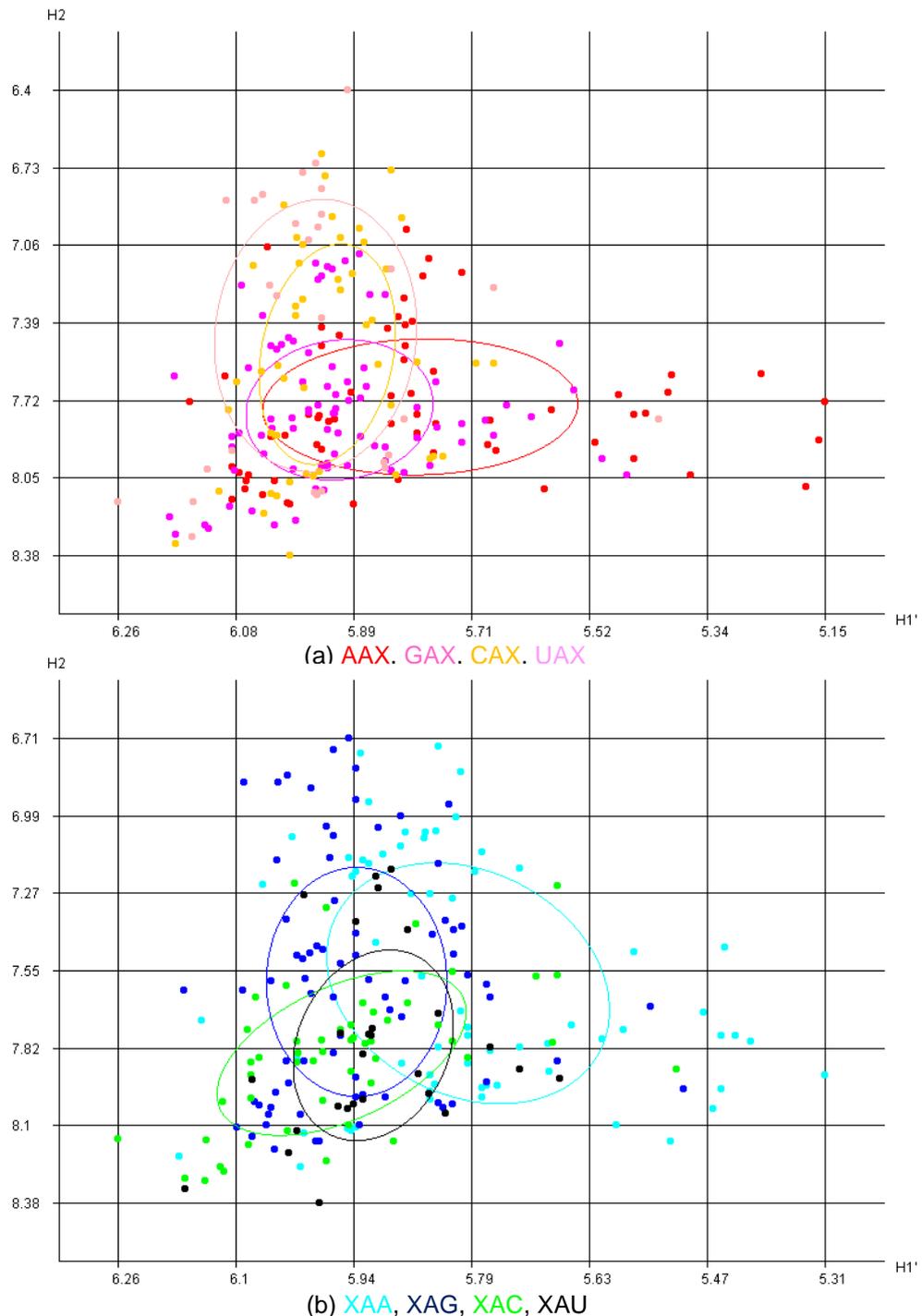


Abbildung 40: Streudiagramme veranschaulichen Datenpaare mit dem Kreuzsignal H1'-H2. Die Datenpaare wurden hierbei nach der jeweiligen Hauptbase und der Vorgängerbase (a) bzw. der Nachfolgerbase (b) in Gruppen zusammengefasst. X – beliebiges Nukleotid

4 Resultate und Diskussion

In Abbildung 41 (a) zeichnen sich eindeutige Datencluster ab. Die einzelnen Datenmengen liegen im Gegensatz zu (b) komprimierter beieinander und die Cluster sind ebenfalls in (a) besser voneinander zu unterscheiden als in (b). Zwischen der Vorgängerbase, der Art der Base und der chemischen Verschiebung des H5 und H6 scheint demzufolge ein Zusammenhang zu existieren. Eine Beziehung zur Nachfolgerbase ist hingegen nicht gegeben.

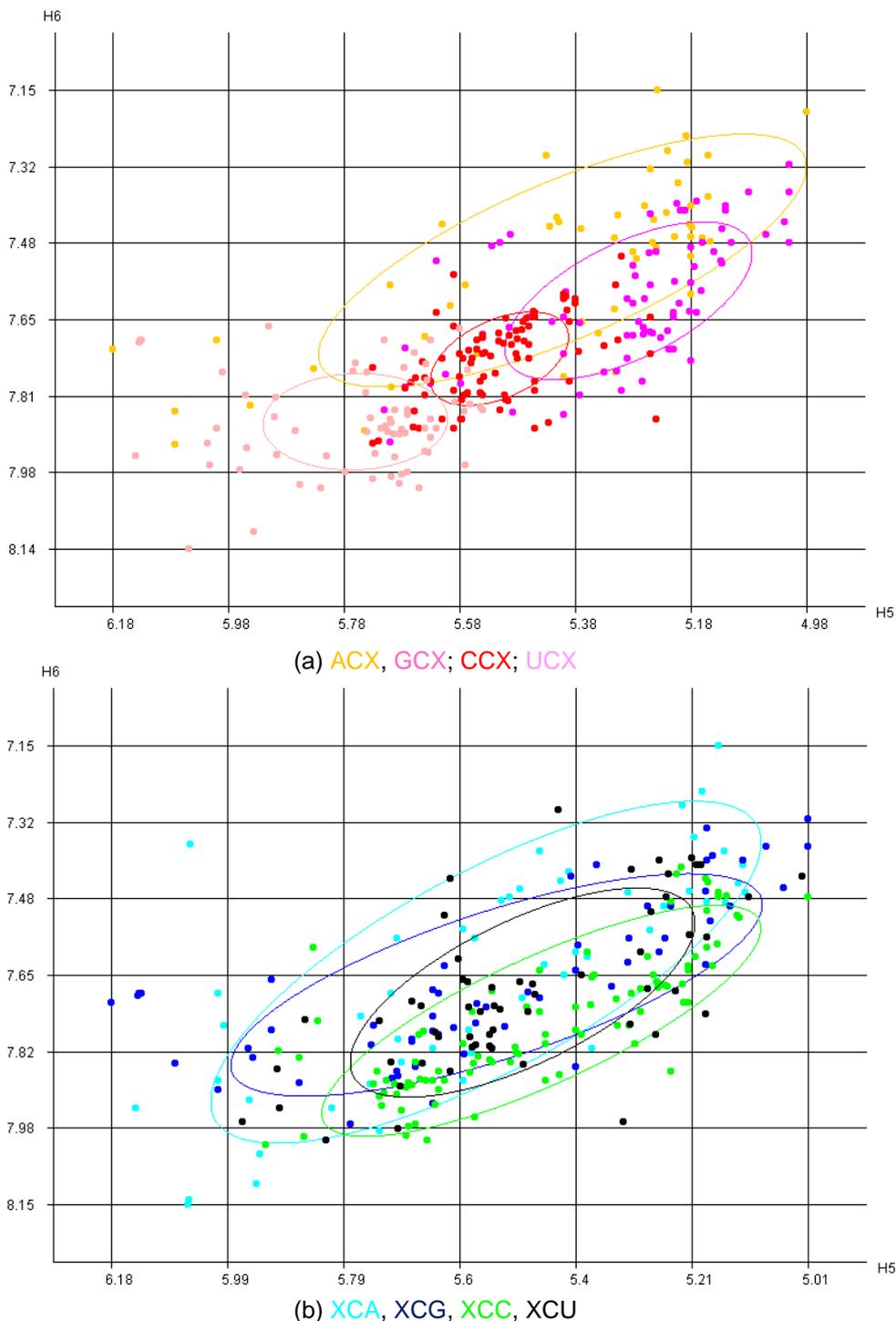


Abbildung 41: Streudiagramme veranschaulichen Datenpaare mit dem Kreuzsignal H5-H6. Die Datenpaare wurden hierbei nach der jeweiligen Hauptbase und der Vorgängerbase (a) bzw. der Nachfolgerbase (b) in Gruppen zusammengefasst. X – beliebiges Nukleotid

4 Resultate und Diskussion

Wie auch schon in Abbildung 41 (a) ersichtlich war, zeichnen sich auch in Abbildung 42 (a) eindeutige Cluster ab. Die H1'-chemischen Verschiebungswerte unterscheiden sich in beiden Streudiagramme nur kaum von den anderen. Jedoch ist in (a) deutlich ein Unterschied zwischen den H8-Werten zu erkennen, was in (b) nicht der Fall ist. Insofern ist davon auszugehen, dass der Vorgänger einen größeren Einfluss auf die chemische Verschiebung hat als der Nachfolger.

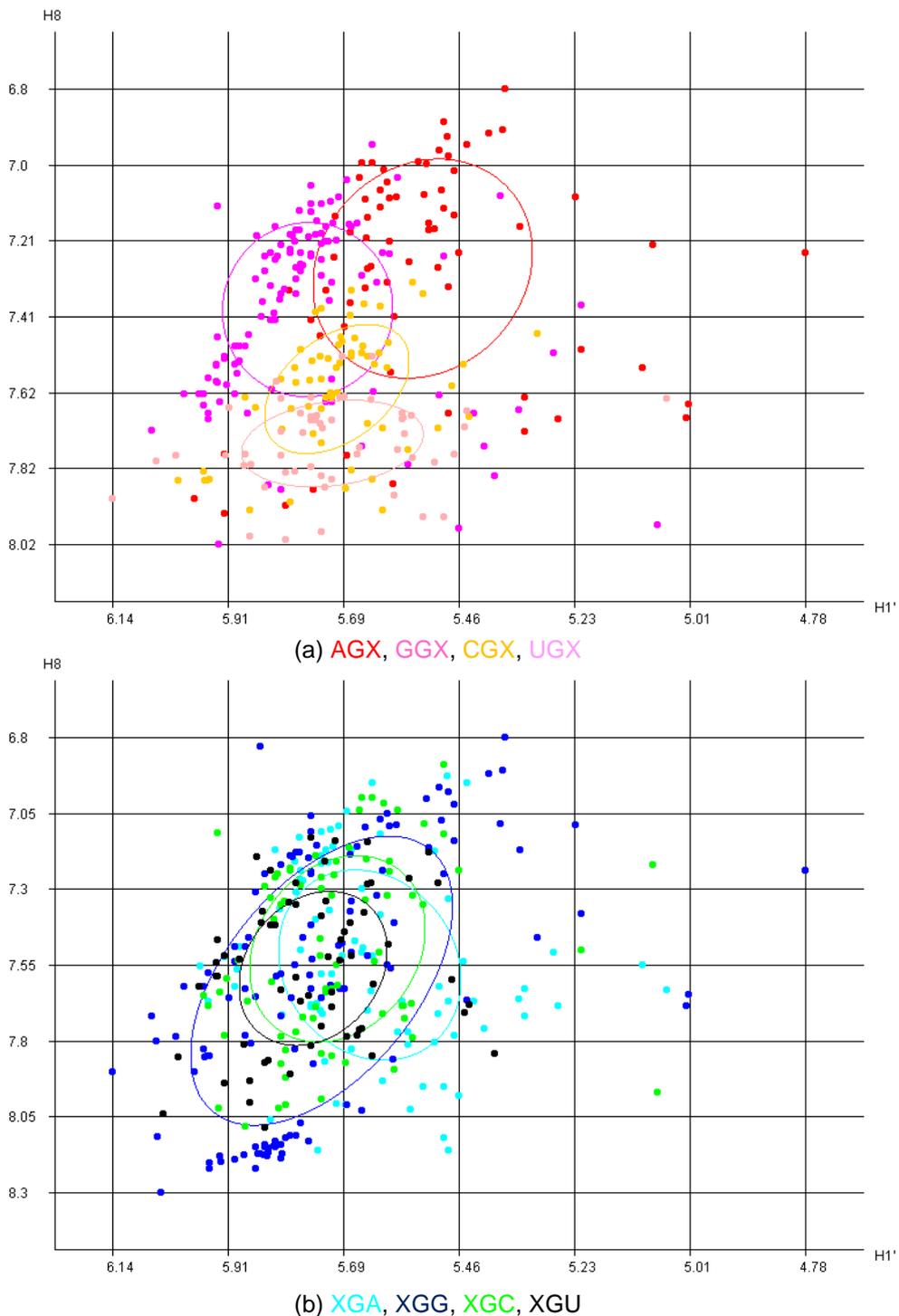


Abbildung 42: Streudiagramme veranschaulichen Datenpaare mit dem Kreuzsignal H1'-H8. Die Datenpaare wurden hierbei nach der jeweiligen Hauptbase und der Vorgängerbase (a) bzw. der Nachfolgerbase (b) in Gruppen zusammengefasst. X – beliebiges Nukleotid

4 Resultate und Diskussion

Abbildung 43 veranschaulicht zwei Streudiagramme mit dem C6-H6 chemischen Verschiebungspaar der Hauptbase Uracil. In Abbildung (a) zeichnen sich die Cluster gut ab, da die H6-Werte der einzelnen Cluster nicht genau im selben Bereich liegen. In (b) hingegen sind die H6-Werte aller Datenpunkte verstreut und nicht genau zuordenbar.

Auch anhand dieser Abbildung würde man Cromsigt & Hilbers, 2001 zustimmen, dass die chemischen Verschiebungen eher vom Vorgänger als vom Nachfolger abhängig sind.

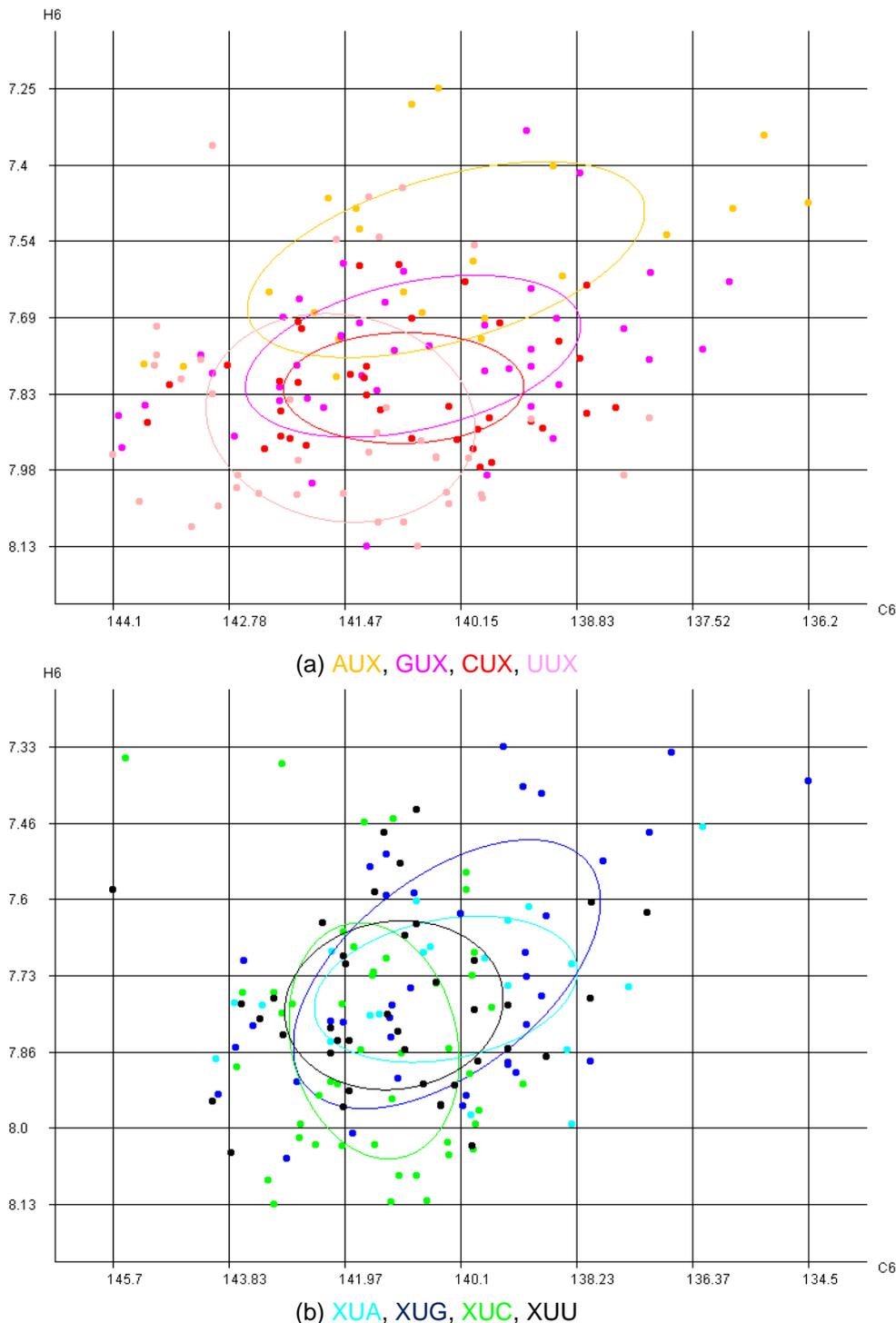


Abbildung 43: Streudiagramme veranschaulichen Datenpaare mit dem Kreuzsignal C6-H6. Die Datenpaare wurden hierbei nach der jeweiligen Hauptbase und der Vorgängerbase (a) bzw. der Nachfolgerbase (b) in Gruppen zusammengefasst. X – beliebiges Nukleotid

4.2.2 Statistik von chemischen Verschiebungen mit allen Variationen

Statistiken, in denen alle möglichen Variationen enthalten sind, dienen zum groben Überblick über die Verteilung der Signale im Streudiagramm.

Abbildung 45 zeigt ein Streudiagramm von den C1'-H1' chemischen Verschiebungspaaren. Die Signale des entsprechenden Streudiagramms überlappen stark. Trotz allem sind deutliche Verteilungstendenzen zu erkennen. Die H1'-Werte von Uridin und Cytidin liegen zwischen 5,2 ppm und 5,6 ppm und die von Guanosin und Adenosin zwischen 5,6 ppm und 6 ppm. Dadurch ist lediglich eine direkte Unterscheidung in Purine und Pyrimidine möglich. Bei Guanosin und Adenosin könnte man dennoch eine Trennung der Cluster durchführen.

Abbildung 44 veranschaulicht die Kreuzsignale zwischen H1'-H8 von den Purinbasen Adenin und Guanin und H1'-H6 von den Pyrimidinbasen Uracil und Cytosin. Deutlich zu erkennen ist, dass die Werte von H6 und H8 etwa im selben ppm-Bereich liegen. Die H1'-chemischen Verschiebungen aber unterscheiden sich besser. Eindeutig kann man 3 bis 4 Cluster erkennen. Uridin und Cytidin sind schlecht voneinander unterscheidbar, jedoch aber Guanosin von Adenosin und Uridin/Cytidin, als auch Adenosin von Uridin/Cytidin ist gut zu unterscheiden.

Die Cluster der chemischen Verschiebungspaare C5-H5 von Uridin und Cytidin und C8-H8 von Adenosin und Guanosin lassen sich sehr gut voneinander trennen, wie in Abbildung 47 ersichtlich ist. Die chemischen Verschiebungen des Kohlenstoff unterscheiden sich sehr stark, die des H8 unterscheiden sich noch gut, aber die des H5 eher weniger. Klar zu erkennen sind vier gut voneinander trennbare Cluster, wobei Guanosin und Adenosin leicht miteinander überlappen, Uridin und Cytidin hingegen überlappen sich nicht. Wie auch schon durch Abbildung 39 darauf zu schließen war, ist davon auszugehen, dass das C5 ein guter Startpunkt für die Zuordnungsstrategie sein könnte.

Abbildung 46 zeigt ein Streudiagramm ähnlich wie in Abbildung 47. Hier wurde lediglich das Kreuzsignal C5-H5 durch das Signal aus der C6- und H6-chemischen Verschiebung ausgetauscht. Deutlich erkennbar ist hierbei die mäßige Clusterung der Datenpunkte von Uridin und Cytidin, da diese sehr stark überlappen. Tendenzen sind jedoch auch in diesem Diagramm ersichtlich.

4 Resultate und Diskussion

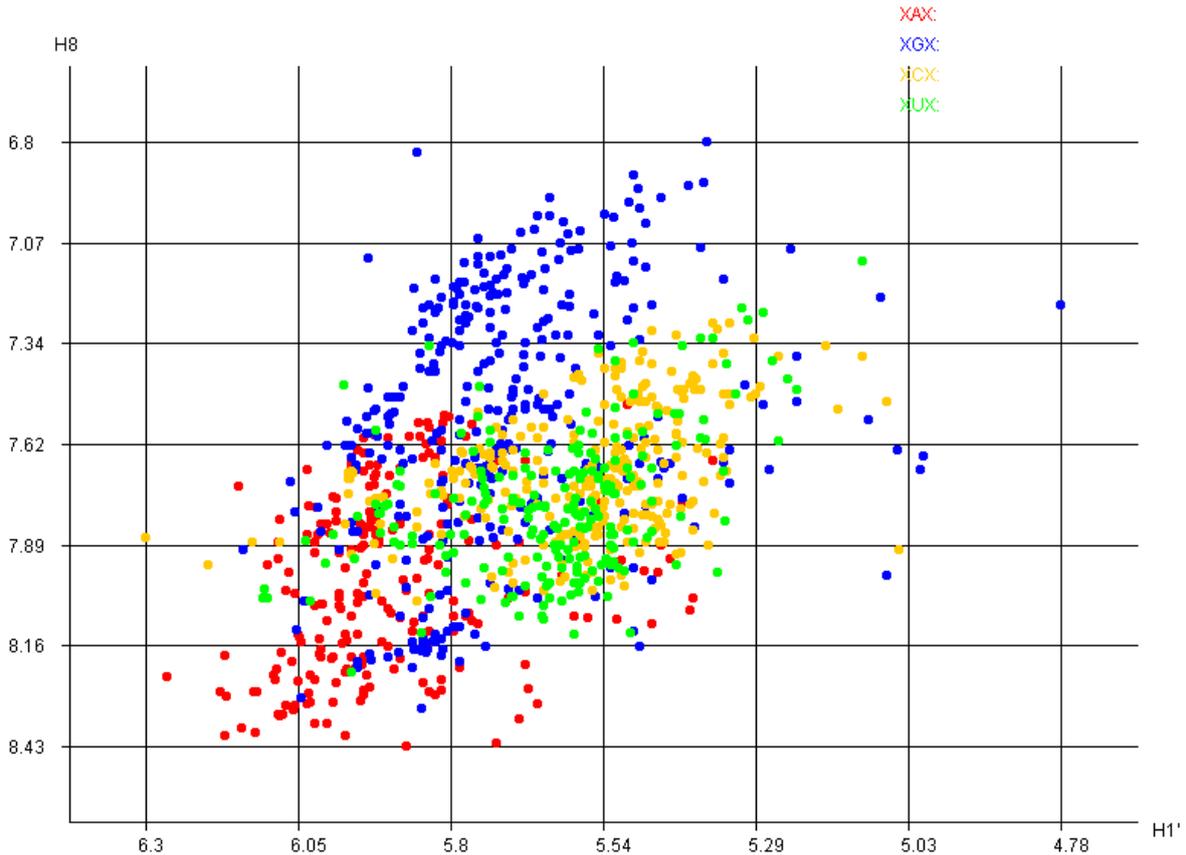


Abbildung 44: Streudiagramm mit den Kreuzsignalen aus dem chemischen Verschiebungspaar H1'-H8 der Hauptbasen A, G, C und U. Das X kennzeichnet ein beliebiges Nukleotid als Vorgänger und Nachfolger.

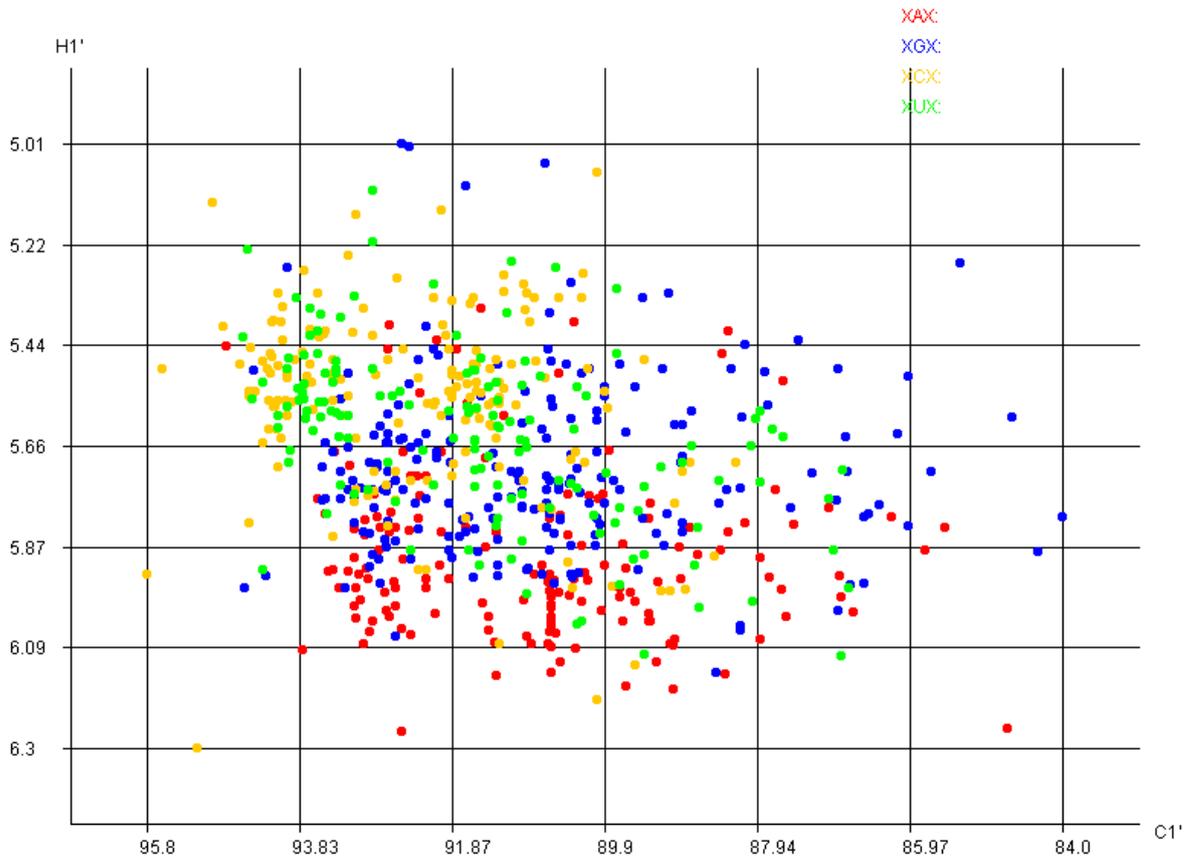


Abbildung 45: Streudiagramm mit der chemischen Verschiebung C1' und H1' der Hauptbasen Adenin, Guanin, Cytosin und Uracil. Das X kennzeichnet ein beliebiges Nukleotid als Vorgänger und Nachfolger.

4 Resultate und Diskussion

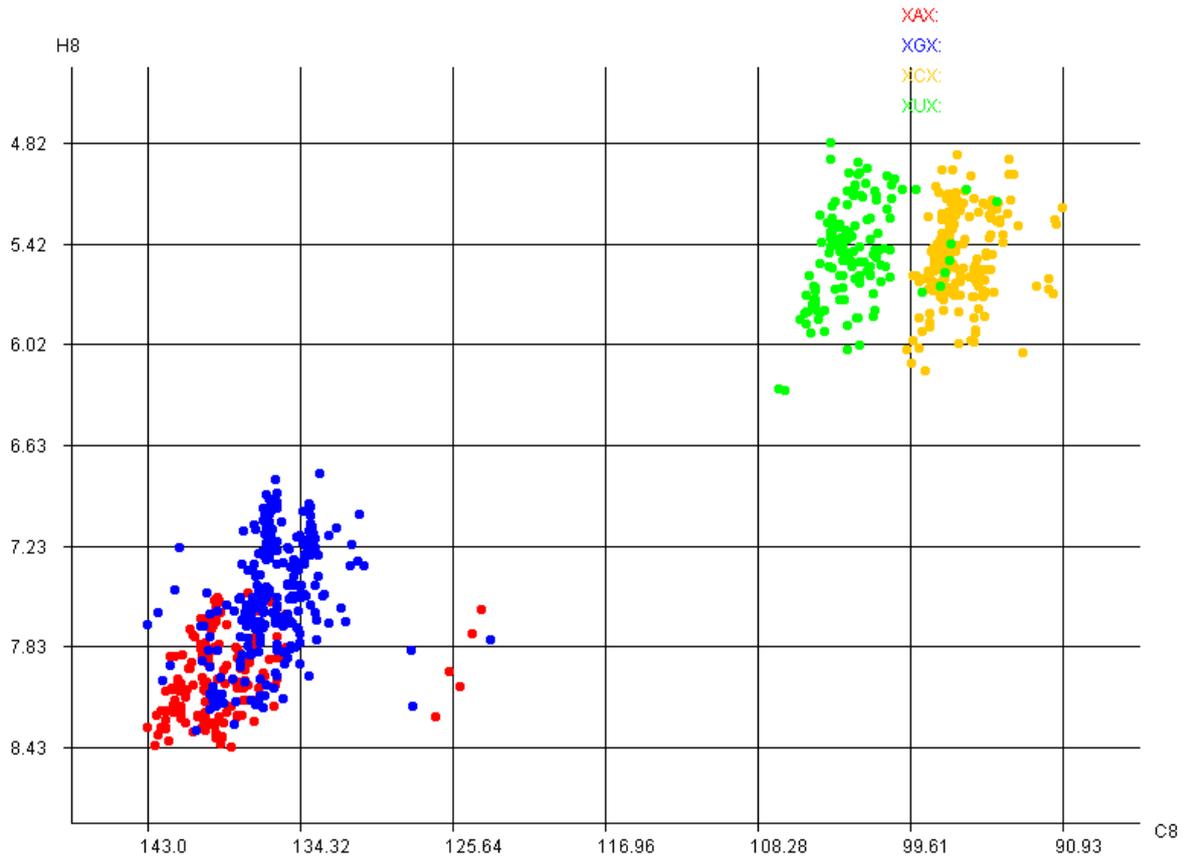


Abbildung 47: Streudiagramm mit der chemischen Verschiebung C8 und H8 der Hauptbasen A und G, sowie C5 und H5 der Hauptbasen C und U. Das X kennzeichnet ein beliebiges Nucleotid als Vorgänger und Nachfolger.

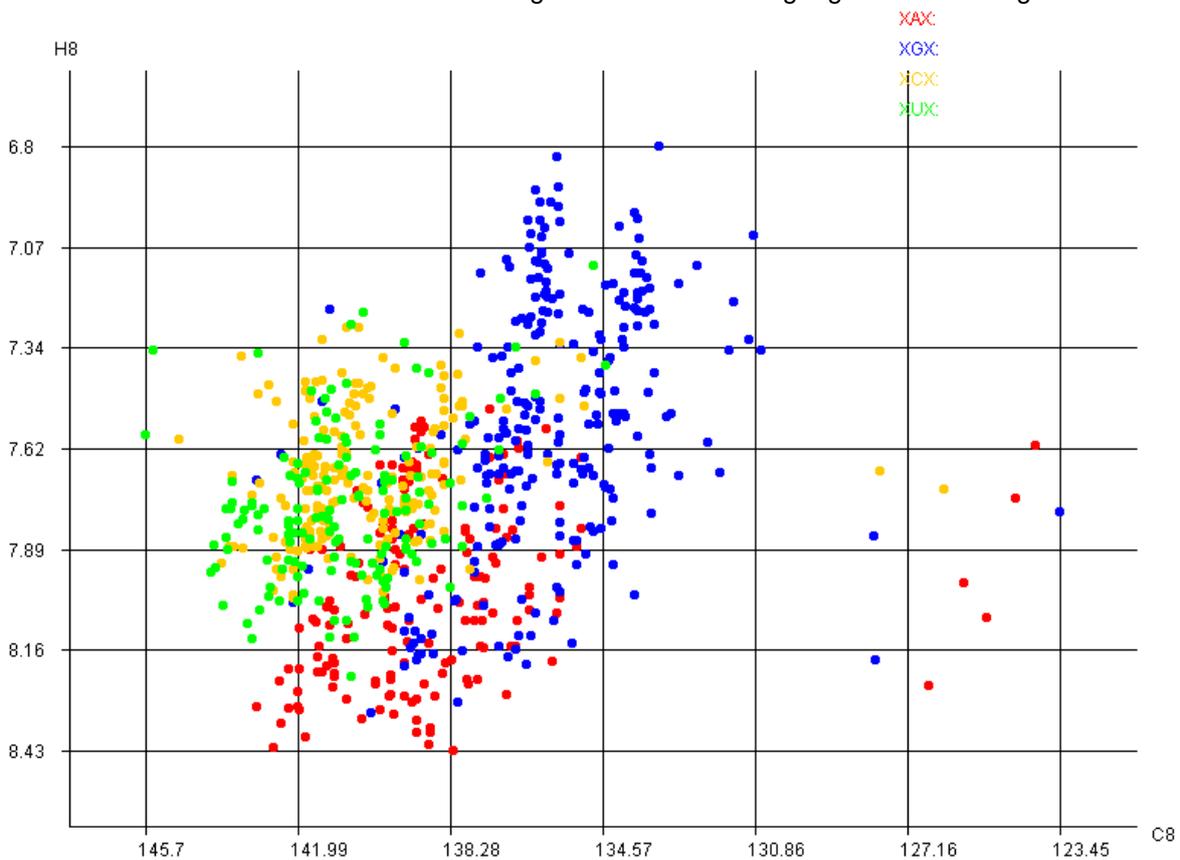


Abbildung 46: Streudiagramm mit der chemischen Verschiebung C8 und H8 der Hauptbasen A und G, sowie C6 und H6 der Hauptbasen C und U. Das X kennzeichnet ein beliebiges Nucleotid als Vorgänger und Nachfolger.

4.2.3 5'GG-Start und 3'CC-Ende

Der 5'GG-Start zeigt ein charakteristisches Muster von Kreuzsignalen. Fast alle untersuchten RNA-Strukturen besaßen am Anfang zwei Guanosine.

Abbildung 48 veranschaulicht das Kreuzsignal von H1'-H8 im ersten Guanosin einer RNA. Dieses Signal ist ein guter Ausgangspunkt für die Zuordnung der einzelnen Signale. Der Wert liegt etwa zwischen 8 ppm und 8,3 ppm auf der H8-Achse und zwischen 5,6 ppm und 6 ppm auf der H1'-Achse.

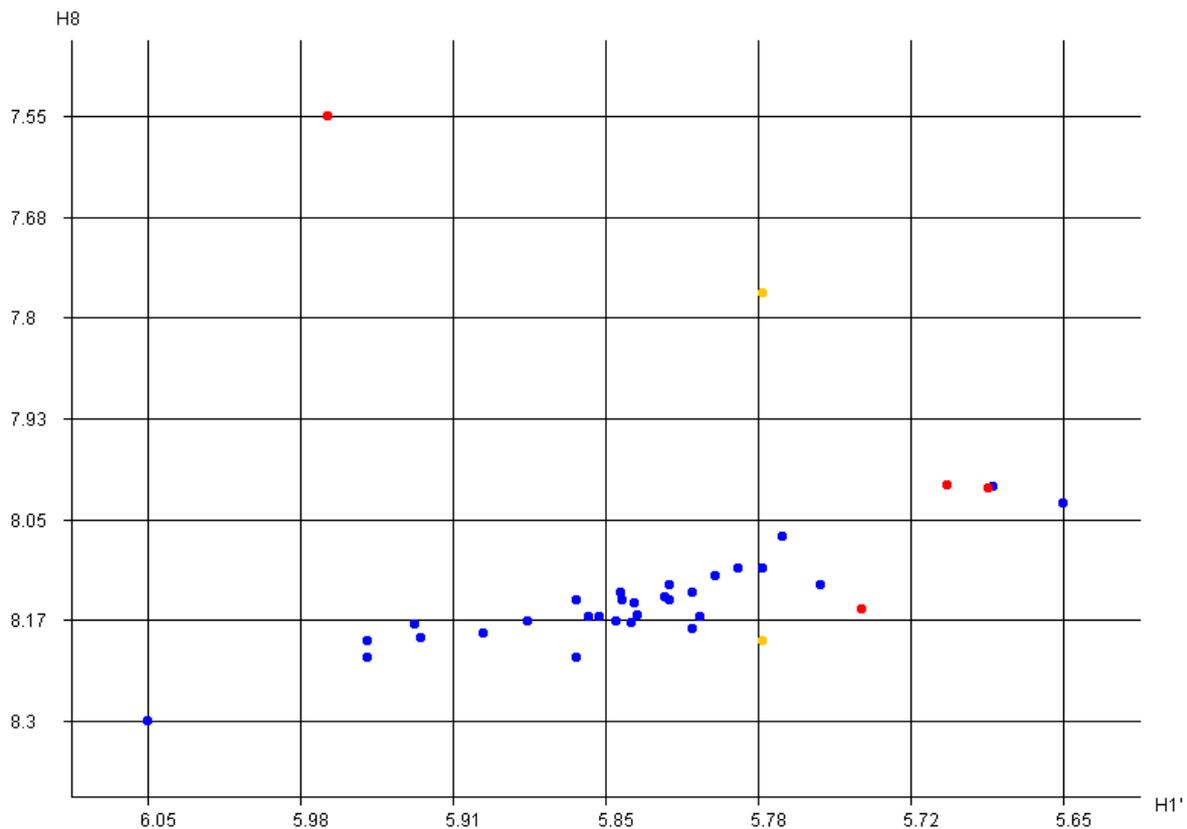


Abbildung 48: Streudiagramm mit der chemischen Verschiebung H1' und H8 der Hauptbase Guanin, welche auch gleichzeitig die erste Base der RNA ist. Nachfolger sind A, G, C und U.
 -GG, -GA, -GC, -GU

Abbildung 49 zeigt ein Guanosin mit einem weiteren Guanosin als Nachfolger und als Vorgänger alle Nukleotide A, G, C und U als auch der Anfang „-“.

Abbildung 49 oben mit den chemischen Verschiebungen H1' und H8 zeigt deutlich die Abgrenzung der Datenpunkte von -GG zu den anderen Datengruppen. Die Signale überschneiden sich nicht mit anderen Kreuzsignalen. Auch in Abbildung 49 unten ist -GG deutlich von den anderen Datengruppen unterscheidbar.

4 Resultate und Diskussion

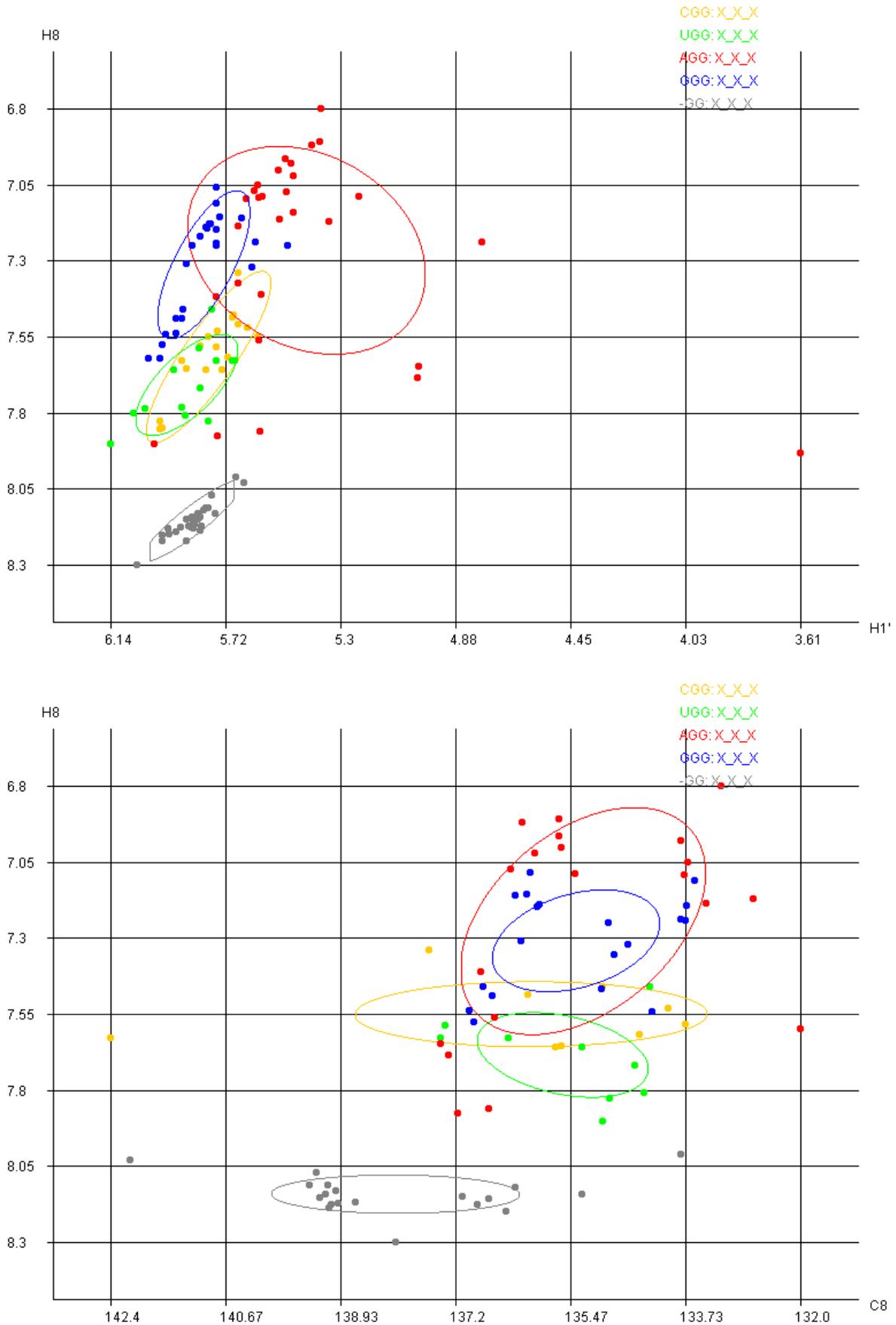


Abbildung 49: Streudiagramme mit den chemischen Verschiebungen H1' und H8 (oben) sowie C8 und H8 (unten) der Hauptbase Guanin. Vorgänger sind A, G, C, U und „-“, also der Anfang der RNA. Nachfolger ist ein Guanosin. X_X_X bedeutet, dass die 2D-Struktur in dieser Abbildung nicht von Bedeutung ist.

4 Resultate und Diskussion

Das 3'-CC-Ende der RNA ist in Abbildung 50 sehr gut von den anderen Datenclustern abgrenzbar. Die H1'-chemische Verschiebung des Cytidins bei CC- unterscheidet sich sehr stark zu den Cytidinen der anderen Cluster. Die H1'-Werte liegen bei CC- etwa zwischen 5,7 ppm und 5,8 ppm. Alle anderen liegen tiefer. Durch diesen Fakt, kann man CC- sehr gut zuordnen.

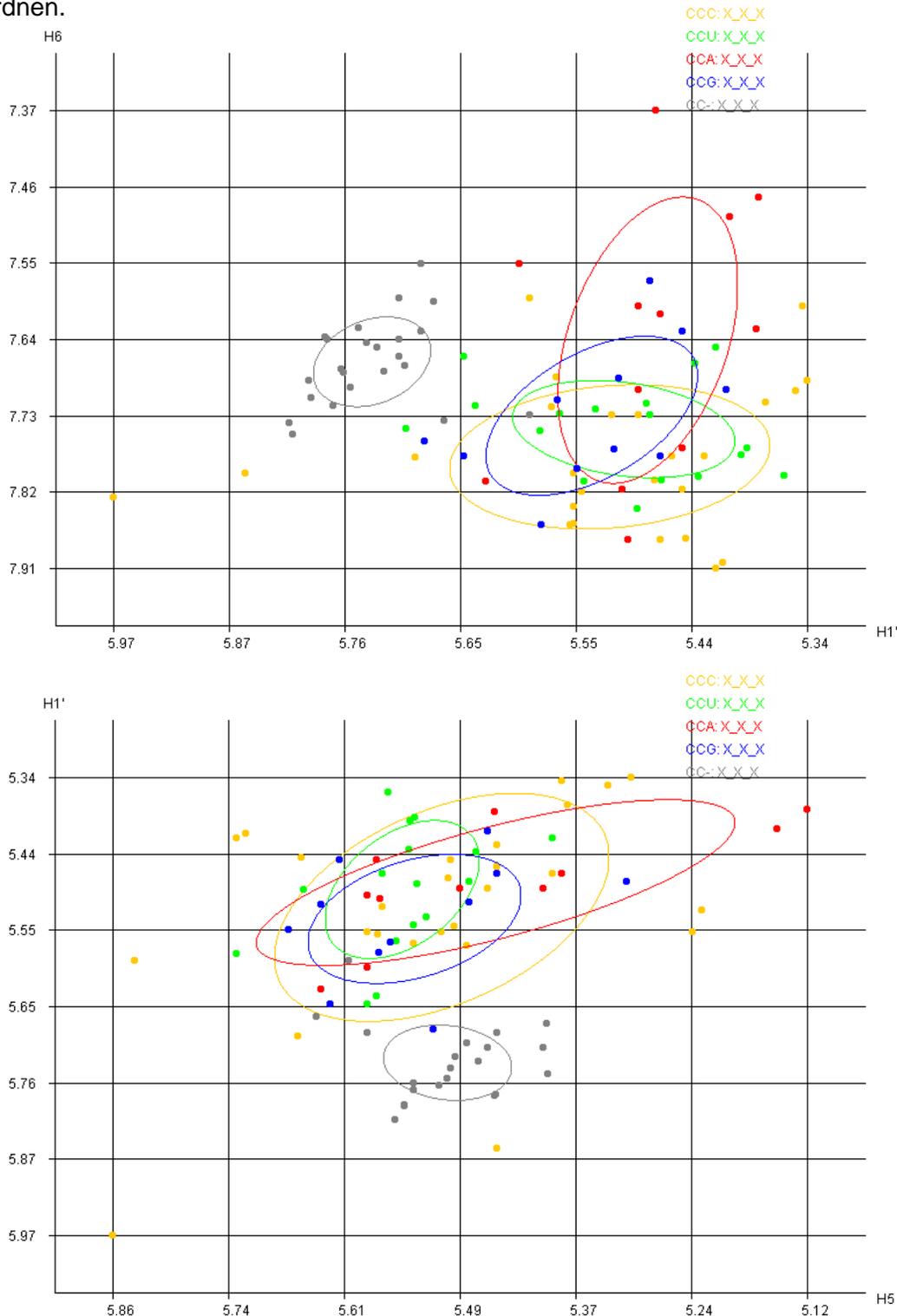


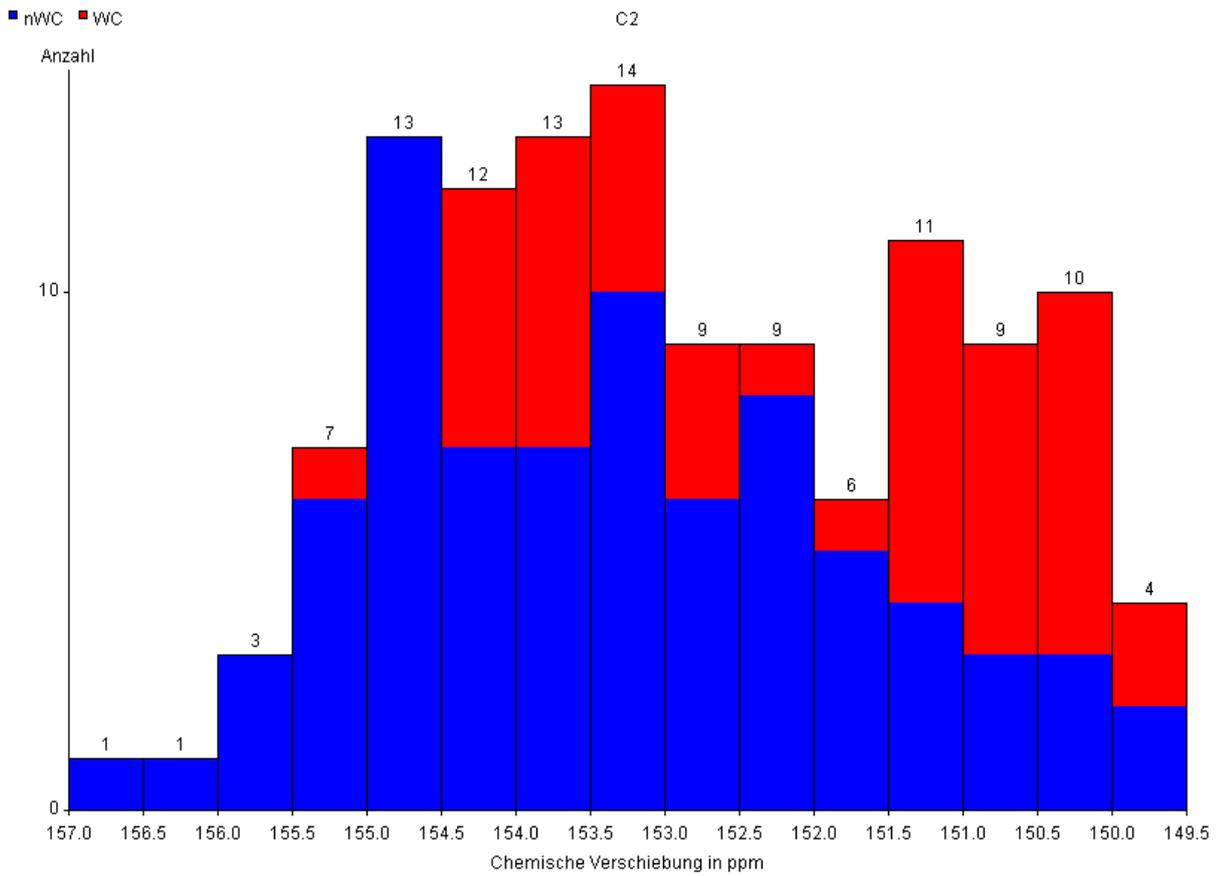
Abbildung 50: Streudiagramme mit den chemischen Verschiebungen H1' und H6 (oben) und H5-H1' (unten) der Hauptbase Cytosin. Vorgänger ist ein Cytidin und Nachfolger sind A, G, C, U und „-“, also das Ende der RNA. X_X_X bedeutet, dass die 2D-Struktur in dieser Abbildung nicht von Bedeutung ist.

4.2.4 Unterscheidung der Datengruppen nach ihrer Sekundärstruktur

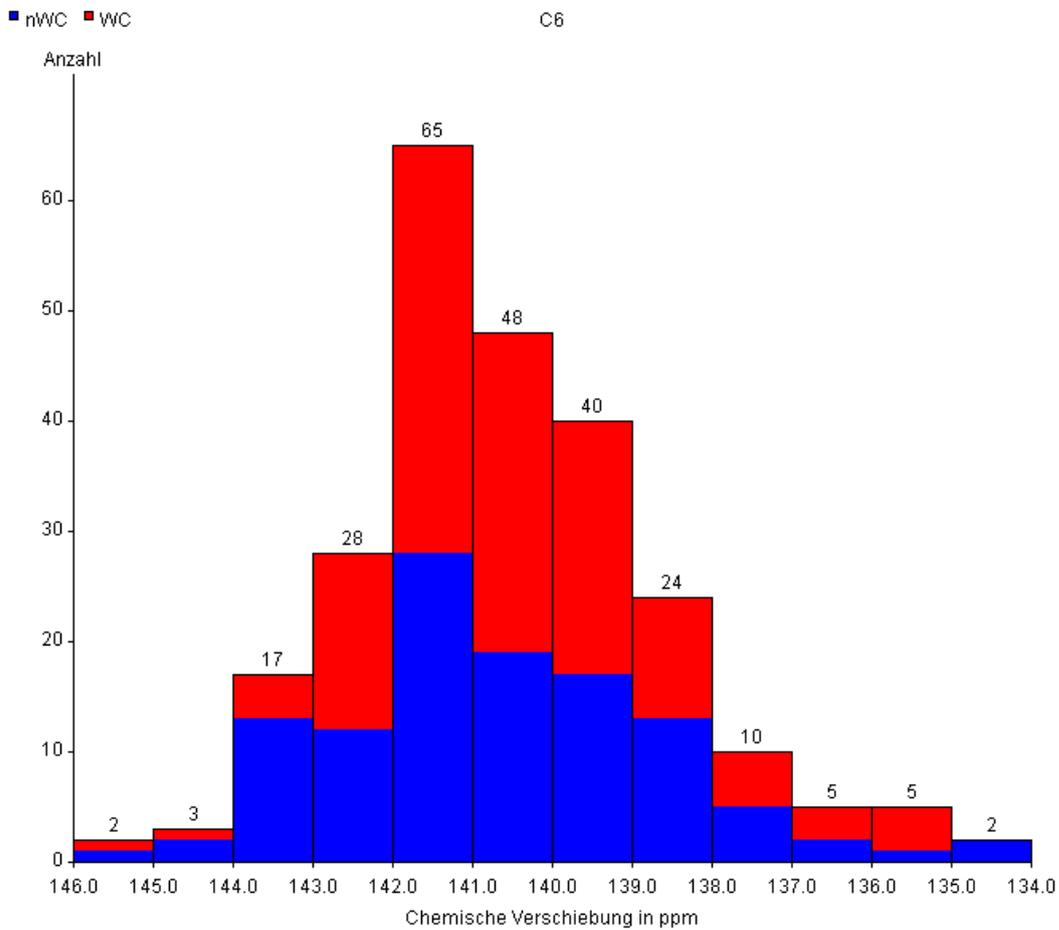
Durch die Kenntnis über die Wasserstoffbrückenbindungen in einer RNA könnte vorhergesagt werden wie die Beschaffenheit ihrer Sekundärstruktur ist. Diese wurde in die Statistik eingebaut und anhand der RNA FZL4 (siehe Kapitel 4.4) überprüft.

Zunächst wurde eine mögliche Trennung der Basen in Watson-Crick- und non-Watson-Crick auf eindimensionaler Ebene untersucht. Wie anhand Abbildung 51 und Abbildung 52 zu sehen, ist es leider nur schwer möglich die Basen, die in Watson-Crick-Paarung von Basen, die nicht in Watson-Crick-Paarung sind auseinanderzuhalten. Bei der chemischen Verschiebung von C2 ist deutlich zu erkennen, dass diese öfter in non-Watson-Crick-Paarung als in Watson-Crick-Paarung vorliegt. Zwischen 151,5 ppm und 157 ppm liegt sie eher im „nWC-Bereich“ und zwischen 151,5 ppm und 149,5 ppm liegt sie öfter im „WC-Bereich“ vor. Eine generelle Trennung ist jedoch nicht durchführbar. Bei der chemischen Verschiebung des C6 ist im Gegensatz zu der des C2 gar keine Unterscheidung möglich. Abbildung 52 (a) veranschaulicht die Verteilung der chemischen Verschiebung des H2. Wie auch schon bei der C2-chemischen Verschiebung ist zu erkennen, dass sie öfter im „nWC-Bereich“ vorkommt. Somit kann man darauf schließen, dass im verwendeten Datensatz Adenin öfter im „loop“ liegt als im „stem“. Zwischen 7,75 ppm und 8,5 ppm liegt die chemische Verschiebung des H2 in non-Watson-Crick-Paarung vor und zwischen 7,15 ppm und 6,7 ppm kommt sie öfter in einer Watson-Crick-Paarung vor. Die chemische Verschiebung des H6 wiederum ist schlecht in Watson-Crick und non-Watson-Crick zu unterteilen, da diese nahezu komplett miteinander übereinstimmen. Auch die anderen chemischen Verschiebungen können nicht gut in non-Watson-Crick-Paarung und Watson-Crick-Paarung eingeteilt werden.

4 Resultate und Diskussion



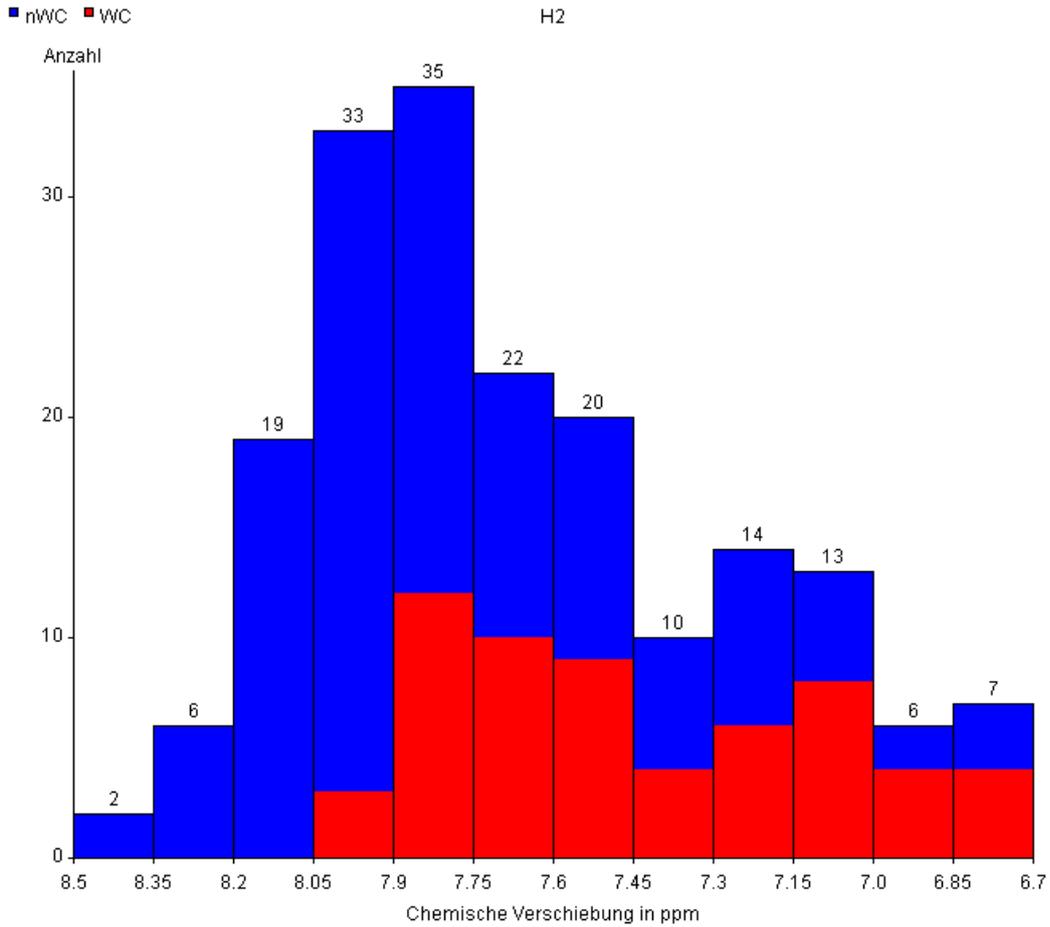
(a) Verteilung der chemischen Verschiebung des C2



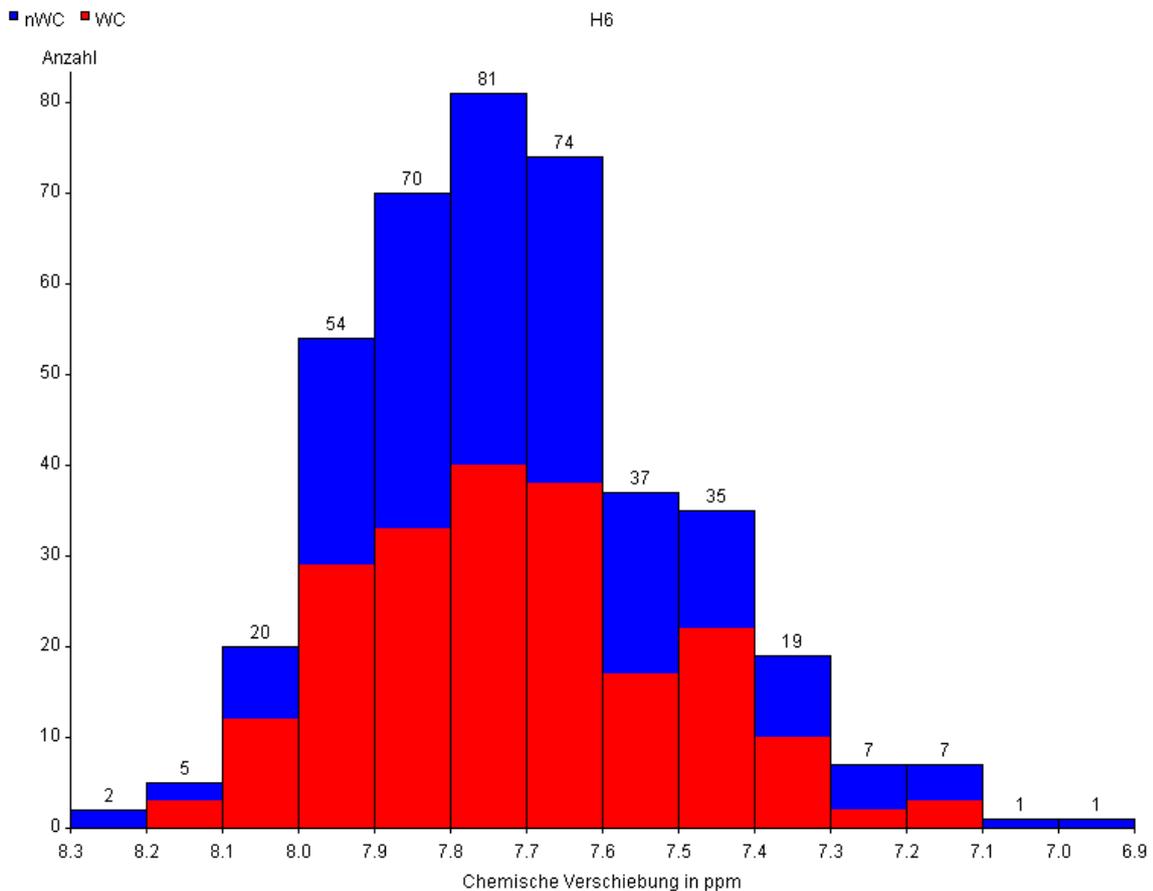
(b) Verteilung der chemischen Verschiebung des C6

Abbildung 51: Watson-Crick-Basen versus non-Watson-Crick-Basen

4 Resultate und Diskussion



(a) Verteilung der chemischen Verschiebung des H2



(b) Verteilung der chemischen Verschiebung des H6

Abbildung 52: Watson-Crick-Basen versus non-Watson-Crick-Basen

4 Resultate und Diskussion

Auf der anderen Seite ist die Unterteilung auf zweidimensionaler Ebene um einiges besser. Abbildung 53 und Abbildung 54 veranschaulichen die Kreuzsignale von C1' und H1' von den Hauptbasen mit allen möglichen Variationen, mit dem Zusatz, dass in Abbildung 53 alle drei Nukleotide „WC-Nukleotide“ und in Abbildung 54 alle drei Nukleotide „nWC-Nukleotide“ sind. Dabei ist deutlich zu erkennen, dass wenn sich die einzelnen Nukleotide im „stem“ befinden, Cluster klar zu erkennen sind. Wohingegen bei den Nukleotiden, welche im „loop“ liegen, keine eindeutigen Datencluster zu erkennen und die einzelnen Datenpunkte verstreuter sind.

In Abbildung 55 ist veranschaulicht, dass man bei der Hauptbase Adenin mit beliebigen Vorgänger und Nachfolger eindeutig zwischen „WC-Triplett“ und „nWC-Triplett“ unterscheiden kann. Die H2-chemische Verschiebung von Adenosin als „WC-Nukleotid“ liegt etwa zwischen 6,7 ppm und 7,7 ppm und von Adenosin als „nWC-Nukleotid“ zwischen 7,7 ppm und 8,2 ppm. Die zwei entstandenen Cluster kann man sehr gut voneinander unterscheiden und somit auch mit einer sehr hohen Wahrscheinlichkeit voraussagen, ob sich ein Adenosin im „stem-Bereich“ oder eher im „loop-Bereich“ befindet. Dies war auf der Ebene der eindimensionalen Trennung nicht so gut zu erkennen.

Abbildung 56 zeigt ebenfalls zwei Cluster, wobei die Hauptbase Guanin ist und die eine Datenmenge aus „WC-Nukleotiden“ und die andere aus „nWC-Nukleotiden“ besteht. Angegeben sind die chemischen Verschiebungen von H1' und H8. Deutlich ist zu erkennen, dass sich die zwei Datengruppen leicht überlappen, wodurch eine genaue Trennung nicht möglich ist, Vermutungen jedoch aufgestellt werden können, ob das besagte Triplett im „stem“ oder im „loop“ liegt.

Das chemische Verschiebungspaar H1'-H8 von Adenosin (rot und gelb) und Guanosin (blau und cyan) einmal als „WC-Triplets“ und das andere mal als „nWC-Triplets“ wird in Abbildung 57 veranschaulicht. Klar zu erkennen sind die unterschiedlichen Regionen, in denen sich die einzelnen Datenmengen befinden. Aus diesem Grund ist es sehr gut möglich, sowohl Guanosin von Adenosin, als auch „WC-Nukleotide“ von „nWC-Nukleotiden“ zu unterscheiden.

Die H8 chemische Verschiebung der „nWC-Nukleotide“ liegt im höheren ppm-Bereich als die der „WC-Nukleotide“. Ebenfalls zu erkennen ist, dass die Datenpunkte der „WC-Nukleotide“ komprimierter beieinander liegen, als die der „nWC-Nukleotide“. Exemplarisch liegen die H1'-chemischen Verschiebungen von Guanosin in WC-Paarung zwischen 5,6 ppm und 5,9 ppm und die der nWC-Paarung zwischen 5,4 ppm und 5,8 ppm.

4 Resultate und Diskussion

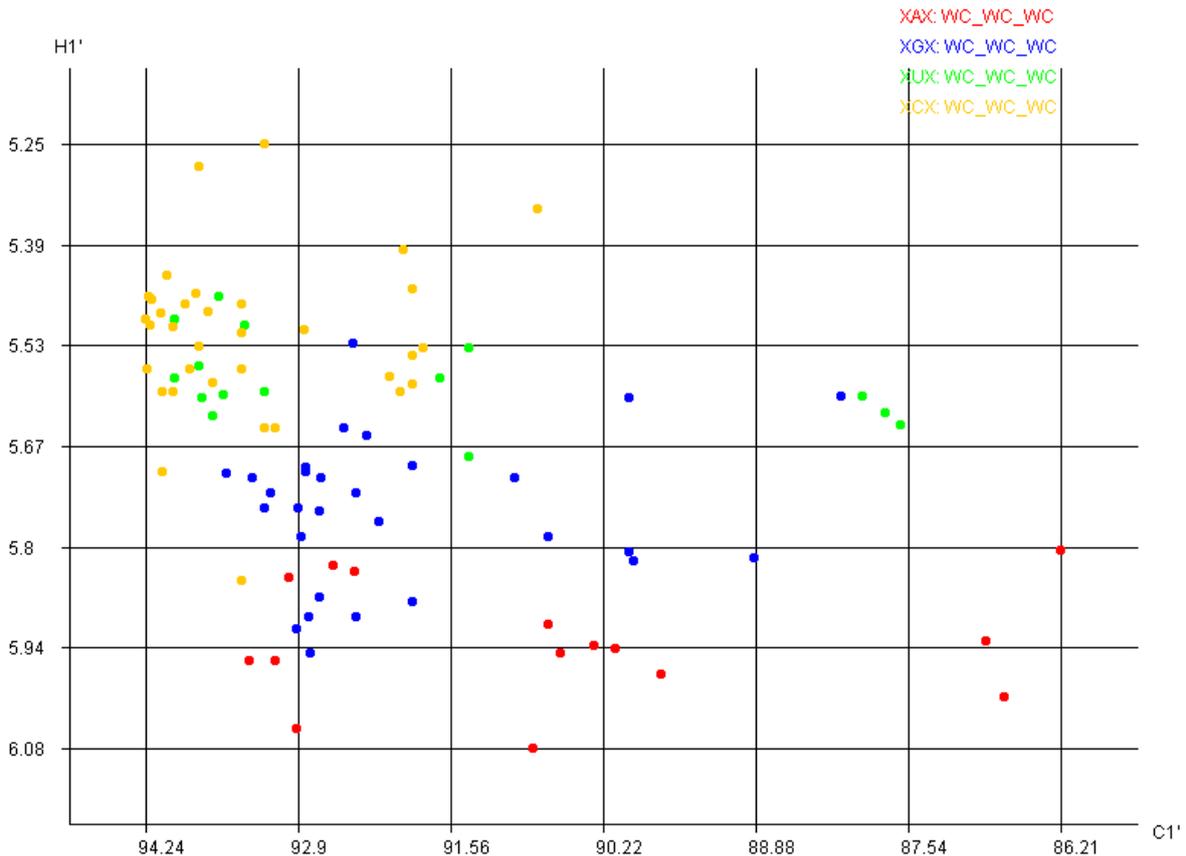


Abbildung 53: Streudiagramm mit den Kreuzsignalen aus dem chemischen Verschiebungspaar C1'-H1' der Hauptbasen A, G, C und U. Das X kennzeichnet ein beliebiges Nukleotid als Vorgänger und Nachfolger. WC_WC_WC bedeutet, dass sich die Basen im „stem“ bzw. im „WC-Bereich“ befinden und mit ihrem Partner Wasserstoffbrückenbindungen ausbilden.

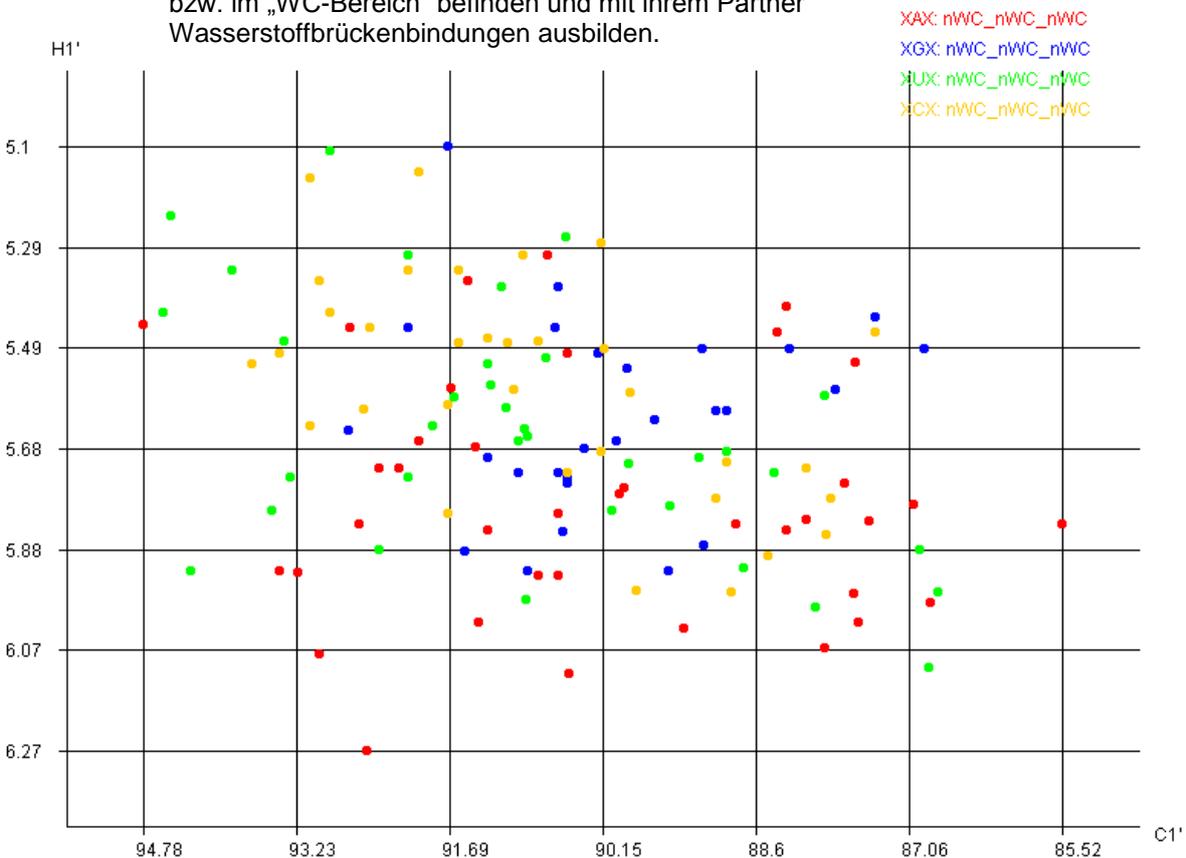


Abbildung 54: Streudiagramm mit den Kreuzsignalen aus dem chemischen Verschiebungspaar C1'-H1' der Hauptbasen A, G, C und U. Das X kennzeichnet ein beliebiges Nukleotid als Vorgänger und Nachfolger. nWC_nWC_nWC bedeutet, dass sich die Basen im „loop“ bzw. im „nWC-Bereich“ befinden und somit keine oder unzureichende Wasserstoffbrückenbindungen mit ihrem Partner ausbilden.

4 Resultate und Diskussion

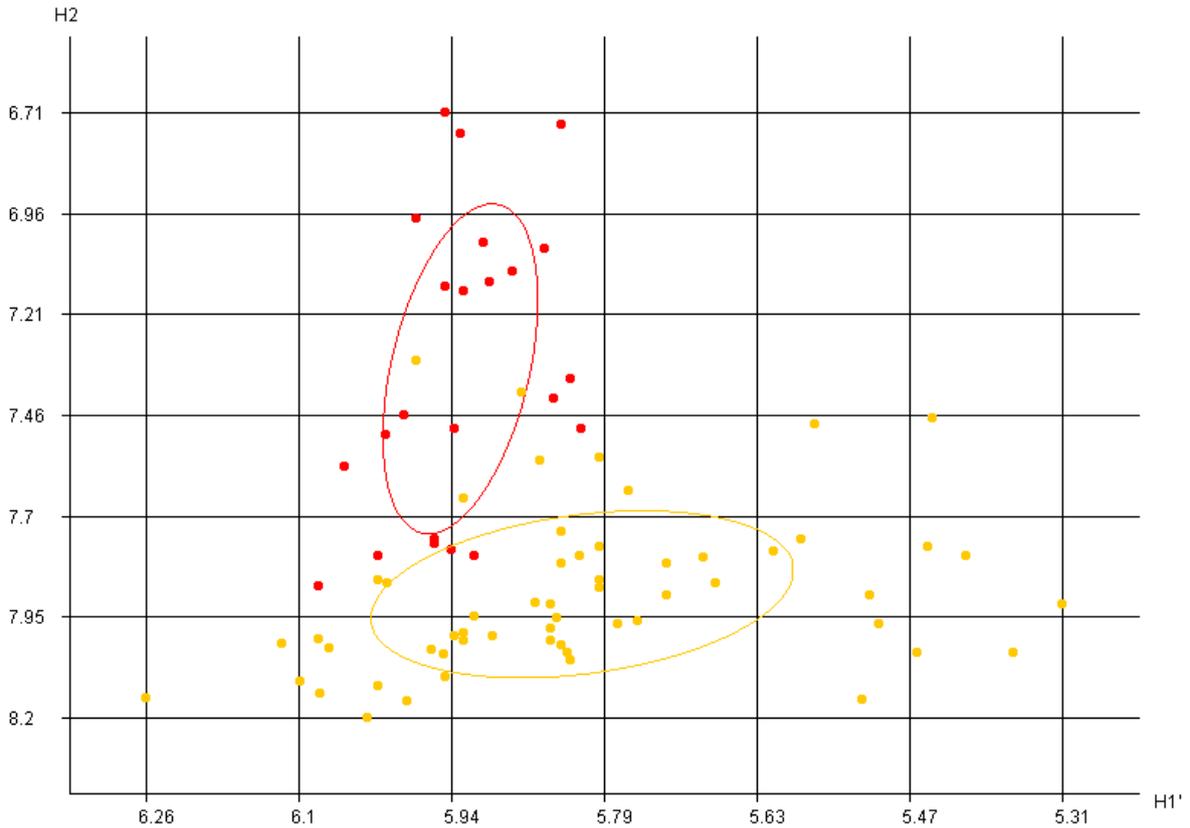


Abbildung 55: Streudiagramm mit den Kreuzsignalen aus dem chemischen Verschiebungspaar H1'-H2 der Hauptbase Adenin. XAX sind Nukleotide im „WC-Bereich“, XAX im „nWC-Bereich“ Das X kennzeichnet ein beliebiges Nukleotid als Vorgänger und Nachfolger.

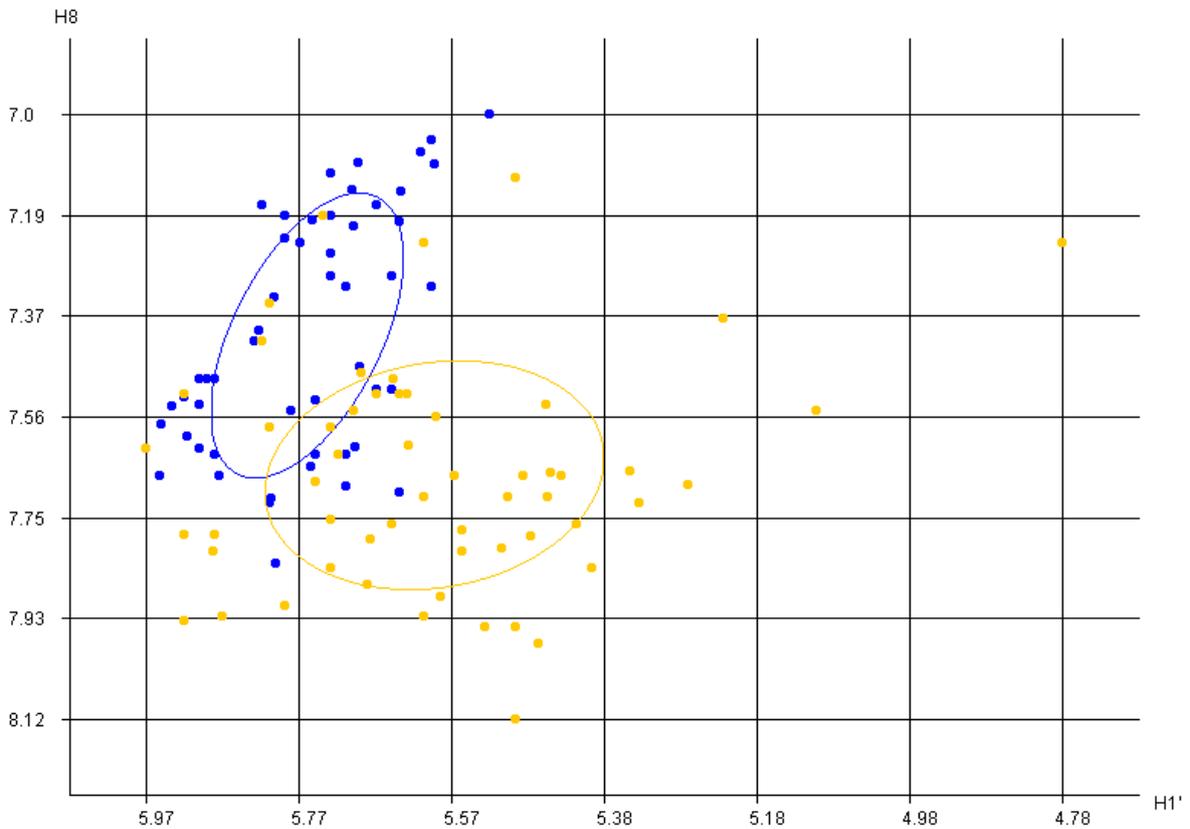


Abbildung 56: Streudiagramm mit den Kreuzsignalen aus dem chemischen Verschiebungspaar H1'-H8 der Hauptbase Guanin. XGX sind Nukleotide im „WC-Bereich“, XGX im „nWC-Bereich“ Das X kennzeichnet ein beliebiges Nukleotid als Vorgänger und Nachfolger.

4 Resultate und Diskussion

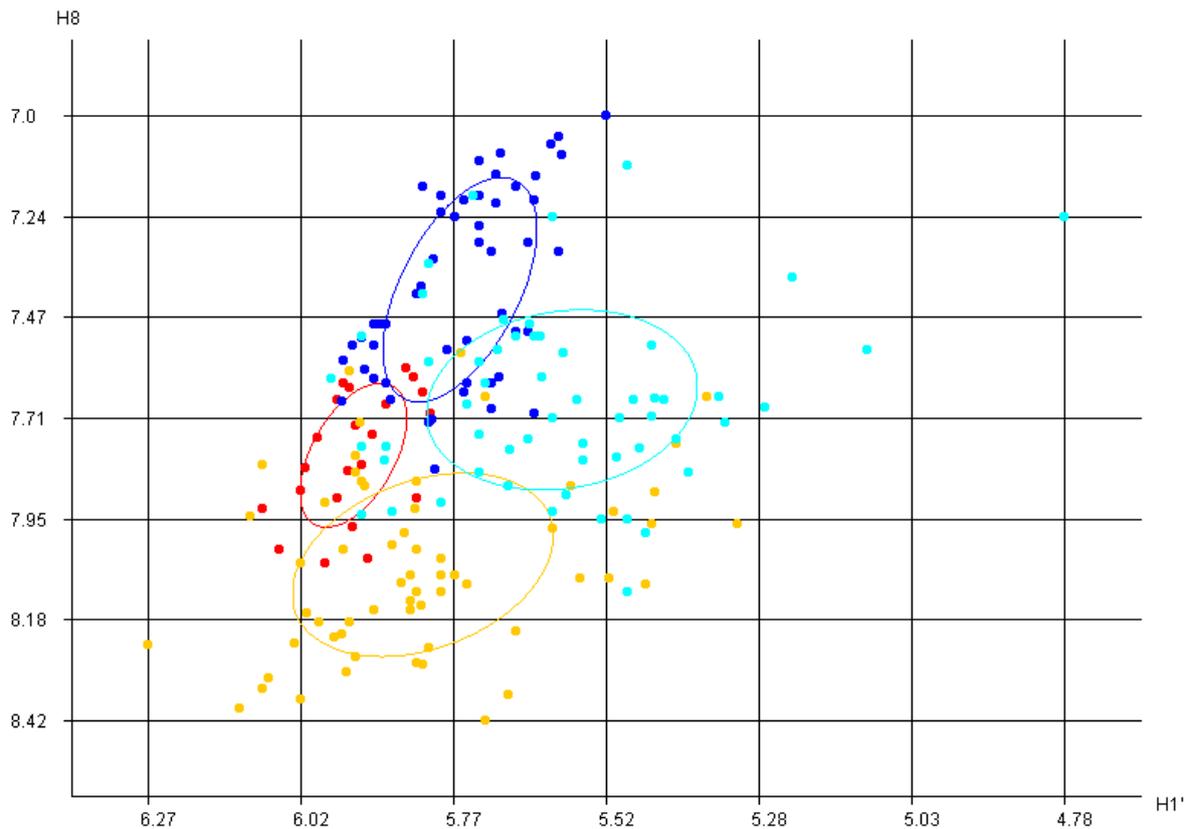


Abbildung 57: Streudiagramm mit den Kreuzsignalen aus dem chemischen Verschiebungspaar $H1'$ - $H8$ der Hauptbasen Adenin und Guanin. **XAX** und **XGX** sind Nukleotide im „WC-Bereich“, **XGX** und **XAX** im „nWC-Bereich“ Das X kennzeichnet ein beliebiges Nukleotid als Vorgänger und Nachfolger.

Abbildung 59 und Abbildung 58 zeigen ebenso den Unterschied der chemischen Verschiebungen zwischen einem „WC-Triplett“ und einem „nWC-Triplett“. Jedoch überlappen die Ellipsen leicht, wodurch eine exakte Trennung beider Datenmengen nicht möglich ist. In Abbildung 58 kann man die Cluster durch die zweidimensionale Trennung sehr gut voneinander unterscheiden. Anhand der Trennung auf eindimensionaler Ebene wäre dies nicht möglich gewesen. Abbildung 52 (b) veranschaulicht die Verteilung der chemischen Verschiebung des $H6$ und deutlich zu erkennen ist, dass man Basen in WC-Paarung nicht von Basen in nWC-Paarung unterscheiden kann.

Somit kann man klar sagen, dass man Nukleotide im „WC-Bereich“ gut von Nukleotiden im „nWC-Bereich“ anhand der zweidimensionalen Trennung voneinander unterscheiden kann.

4 Resultate und Diskussion

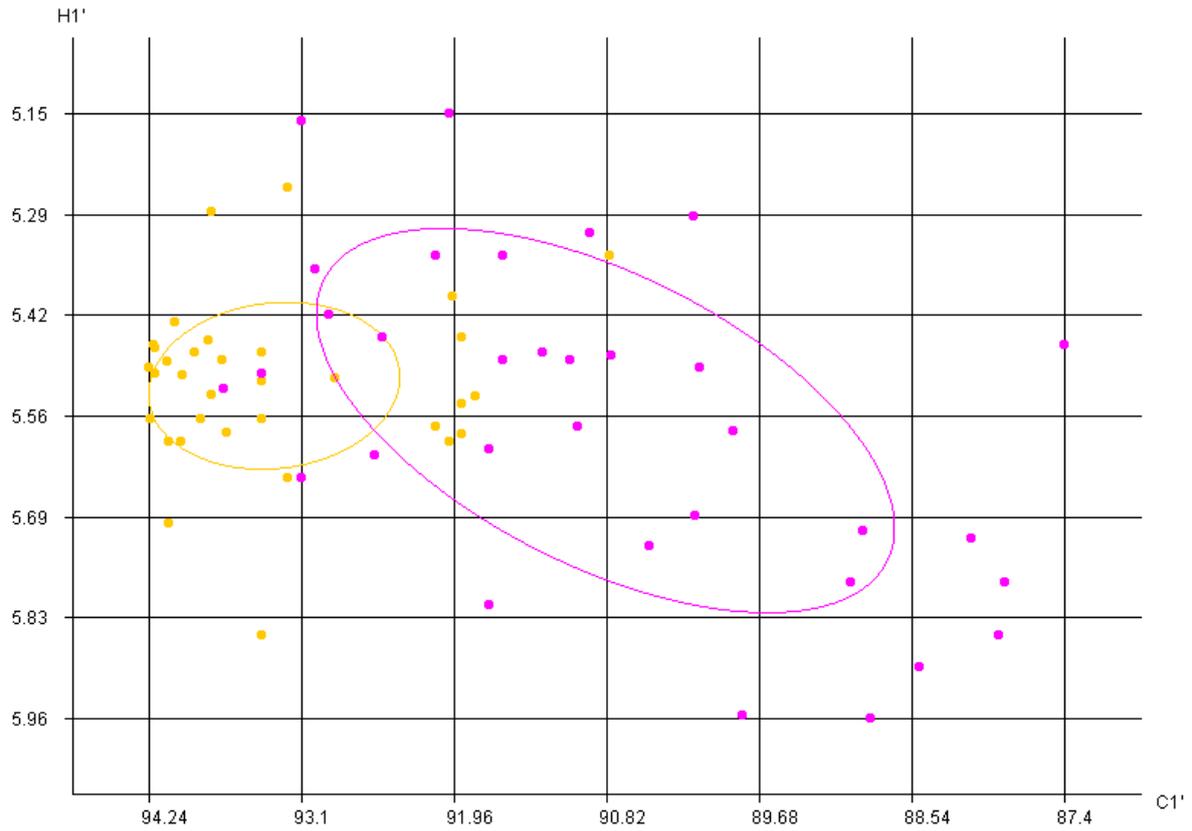


Abbildung 59: Streudiagramm mit den Kreuzsignalen aus dem chemischen Verschiebungspaar C1'-H1' der Hauptbase Cytosin. XCX sind Nukleotide im „stem“, XCX im „loop“
Das X kennzeichnet ein beliebiges Nukleotid als Vorgänger und Nachfolger.

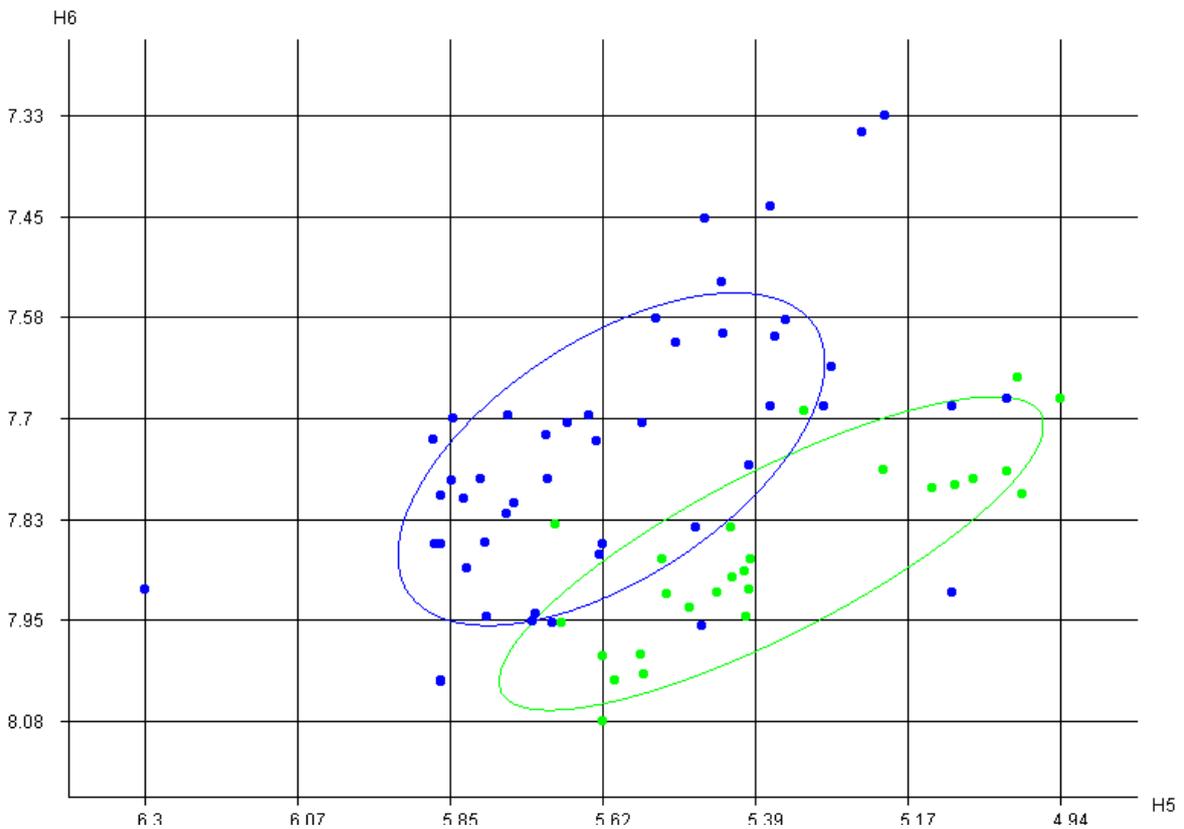


Abbildung 58: Streudiagramm mit den Kreuzsignalen aus dem chemischen Verschiebungspaar H5-H6 der Hauptbase Uracil. XUX sind Nukleotide im „stem“, XUX im „loop“
Das X kennzeichnet ein beliebiges Nukleotid als Vorgänger und Nachfolger.

4.3 Statistik der ^1H - chemischen Verschiebungen zwischen zwei Nukleotiden

Für die Zuordnungen der ^1H - chemischen Verschiebungen wurden nur die Resonanzen analysiert, welche in einem NOESY-Spektrum vorkommen. Dabei unterscheidet man zwischen einem Kreuzsignal mit der Hauptbase (i) und dessen Vorgänger (i-1) oder Hauptbase und dessen Nachfolger (i+1). Untersucht wurden die Kreuzsignale zwischen folgenden chemischen Verschiebungspaaren:

Hauptbase und Vorgänger:

H1'(i) – H1'(i-1)
 H8(i) – H8(i-1)
 H1'(i) – H2(i-1)
 H8(i) – H5(i-1)
 H5(i) – H1'(i-1)
 H5(i) – H8(i-1)
 H6(i) – H8(i-1)
 H5(i) – H2(i-1)
 H6(i) – H2(i-1)
 H5(i) – H6(i-1)

Hauptbase und Nachfolger:

H6(i) – H8(i+1)
 H1'(i) – H1'(i+1)
 H8(i) – H8(i+1)
 H1'(i) – H8(i+1)
 H2(i) – H2(i+1)
 H8(i) – H6(i+1)
 H2(i) – H5(i+1)
 H1'(i) – H6(i+1)
 H5(i) – H5(i+1)
 H6(i) – H6(i+1)

Einige dieser Kreuzsignale wurden in diesem Kapitel näher beschrieben und analysiert.

4.3.1 Kreuzsignal mit dem Vorgänger

Abbildung 60 veranschaulicht das Kreuzsignal der H8-chemischen Verschiebung von der Hauptbase Adenin oder Guanin (x-Achse) mit deren H8-chemischen Verschiebung des Vorgängers Adenosin oder Guanosin (y-Achse). Die Datencluster sind sehr gut voneinander zu unterscheiden. Vor allem die Datenpunkte von AAX und AGX kann man gut trennen. Außerdem zeigt diese Abbildung, dass sich die H8-chemischen Verschiebungen der Vorgänger sehr unterscheiden. Die H8-Werte des vorangegangenen Adenosin liegen etwa zwischen 7,6 ppm und 8,3 ppm und die des Guanosin zwischen 7 ppm und 8,1 ppm. Somit kann man die beschriebenen Dupletts nicht nur vom Vorgänger, sondern auch von der Hauptbase unterscheiden.

4 Resultate und Diskussion

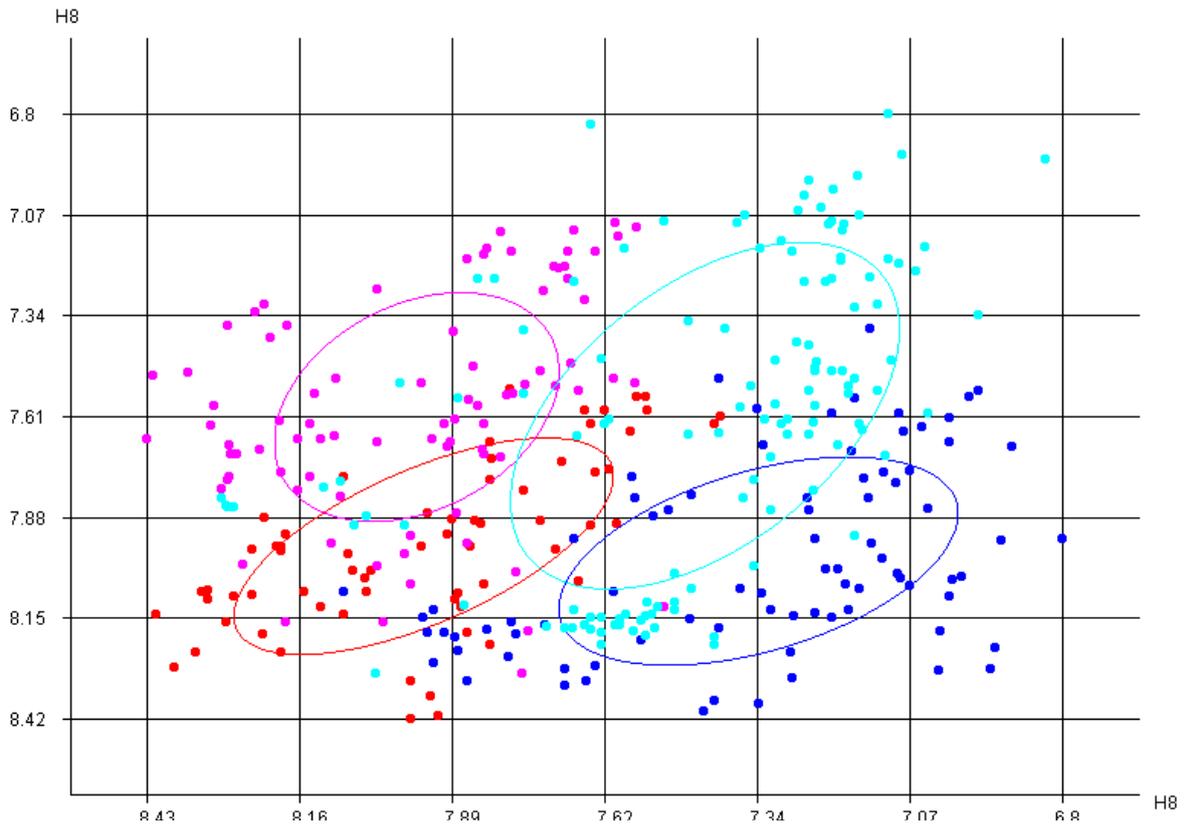


Abbildung 60: Streudiagramm mit der H8-chemischen Verschiebung der Hauptbasen Adenin oder Guanin (x-Achse) und der H8-chemischen Verschiebung des Vorgängers Adenosin oder Guanosin (y-Achse). Das X kennzeichnet ein beliebiges Nucleotid als Nachfolger.
AAX, GAX, AGX, GGX

Im Gegensatz dazu ist in Abbildung 61 ersichtlich, dass man die entstandenen Cluster nur nach dem Vorgänger unterscheiden kann. Die H5-chemischen Verschiebungswerte liegen bei allen vier Cluster nahezu im selben ppm-Bereich. Dadurch können sie nicht unterscheiden werden. Die H8-chemische Verschiebung jedoch teilt die vier Cluster in zwei Gruppen auf, diese mit einem Adenosin als Vorgänger und in die mit einem Guanosin als Vorgänger. Infolgedessen ist klar verdeutlicht, dass das Kreuzsignal aus dem H5 der Hauptbase und dem H8 des Vorgängers lediglich durch das Signal des Vorgängers zu unterscheiden ist.

Abbildung 62 zeigt folgende vier Dupletts: UA, UG, UC und UU. Dabei werden die H1'-Signale beider Nucleotide betrachtet. Deutlich sichtbar ist, dass sich die Datenpunkte des UA von den anderen Datenclustern unterscheiden. Das H1' der Hauptbase Adenin liegt in einer anderen Umgebung, nämlich etwa 0,2 ppm höher als die anderen Datenpunkte. Die H1'-chemischen Verschiebungen der Vorgänger hingegen unterscheiden sich überhaupt nicht in ihrem ppm-Bereich. Folglich kann man sagen, dass man die Cluster so gut wie gar nicht voneinander trennen kann, da sich nur UA-Datenpunkte von den anderen abgrenzen lassen. Aus diesem Grund ist es nicht möglich eine Aussage darüber zu treffen, ob die Dupletts durch die chemische Verschiebung des Vorgängers oder der Hauptbase unterschieden werden können.

4 Resultate und Diskussion

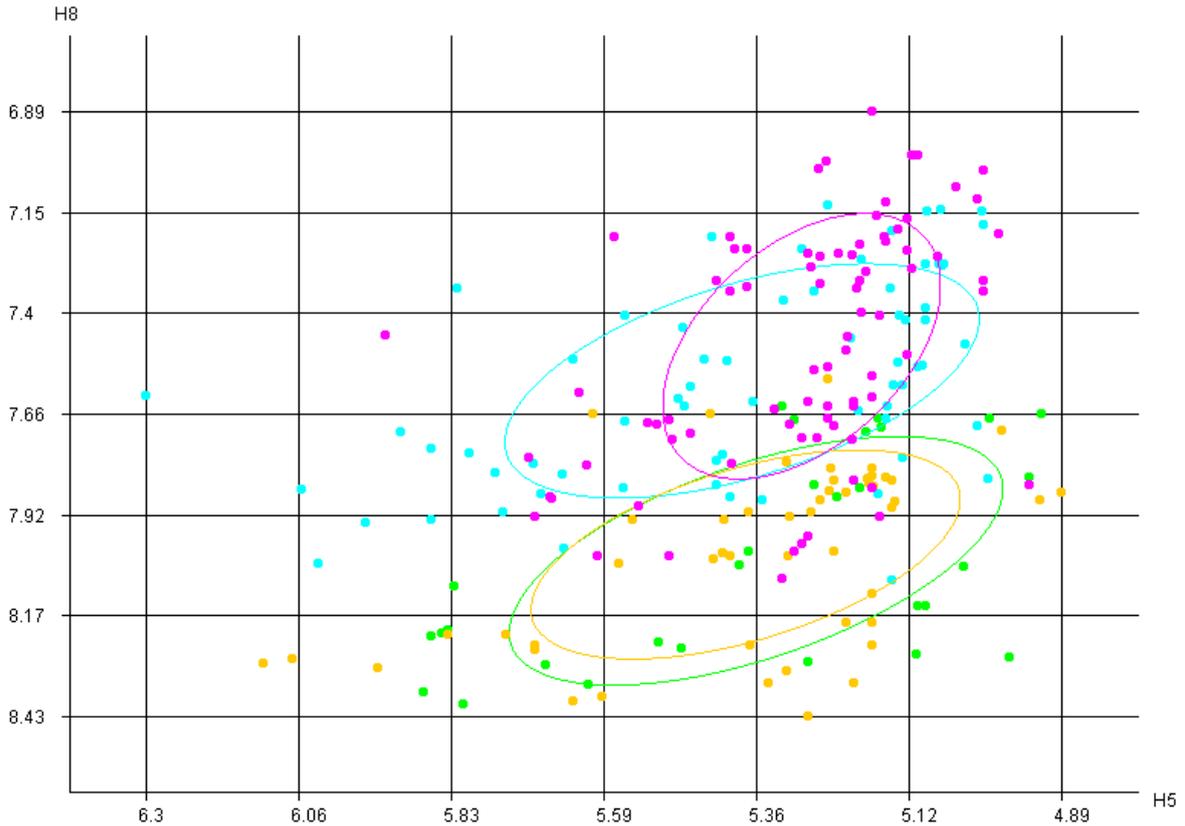


Abbildung 61: Streudiagramm mit der H5-chemischen Verschiebung der Hauptbasen Cytosin oder Uracil (x-Achse) und der H8-chemischen Verschiebung des Vorgängers Adenosin oder Guanosen (y-Achse). Das X kennzeichnet ein beliebiges Nukleotid als Nachfolger.

AUX, GUX, ACX, GCX

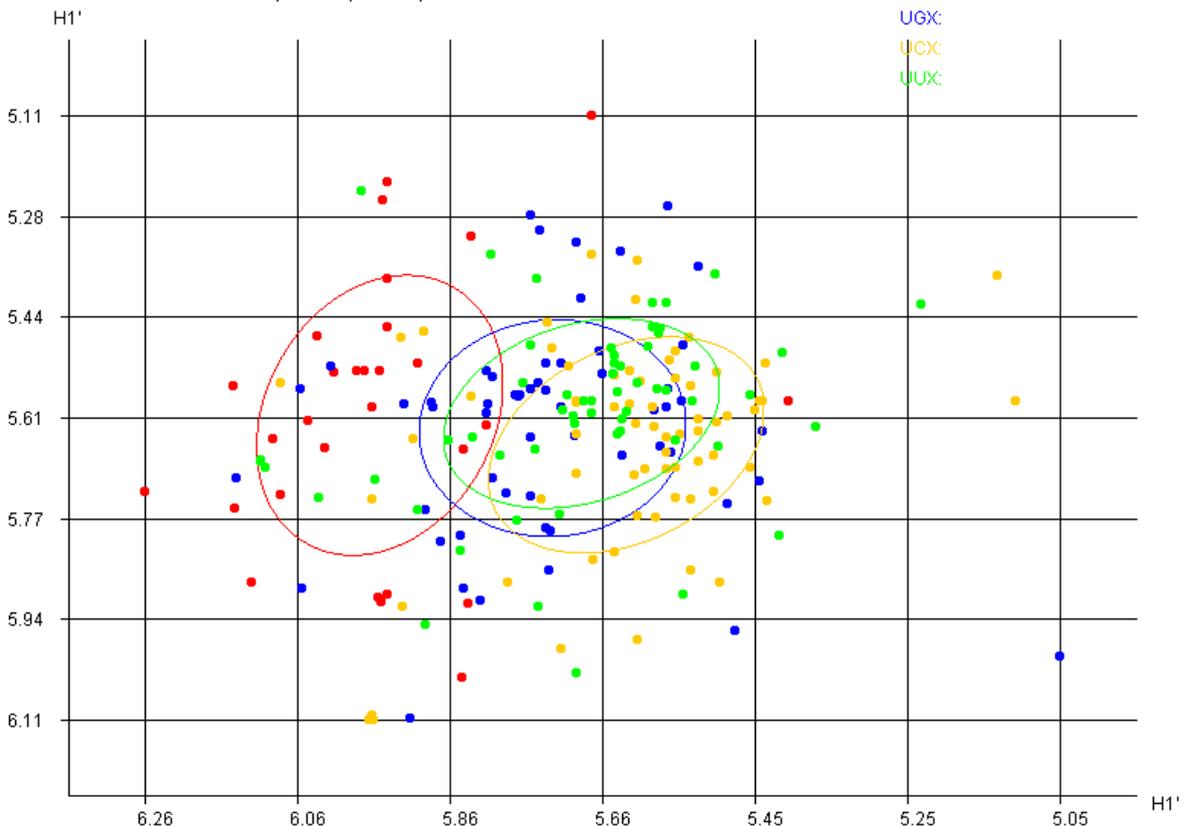


Abbildung 62: Streudiagramm mit der H1'-chemischen Verschiebung der Hauptbasen A, G, C und U (x-Achse) und der H1'-chemischen Verschiebung des Vorgängers Uridin (y-Achse). Das X kennzeichnet ein beliebiges Nukleotid als Nachfolger.

4.3.2 Kreuzsignal mit dem Nachfolger

Abbildung 63 veranschaulicht das Kreuzsignal der H8-chemischen Verschiebung von der Hauptbase Adenin oder Guanin (x-Achse) mit deren H8-chemischen Verschiebung des Nachfolgers Adenosin oder Guanosin (y-Achse). Die Datencluster sind sehr gut voneinander zu unterscheiden. Vor allem die Datenpunkte von XGA und XAG kann man gut trennen. Aus der Abbildung ist klar zu erkennen, dass man mithilfe der H8-chemischen Verschiebung der Hauptbase und des Nachfolgers die Cluster voneinander unterscheiden kann.

Abbildung 64 zeigt Kreuzsignale aus der chemischen Verschiebung des H8 von Guanosin oder Adenosin mit der H6-chemischen Verschiebung von Uridin oder Cytidin. Die H8-Werte von Guanosin und Adenosin unterscheiden sich stark voneinander, wodurch man die vier Cluster in 2 Gruppen einteilen kann (rot und orange mit Hauptbase Adenin; grün und blau mit Hauptbase Guanin). Die H6-chemischen Verschiebungswerte unterscheiden sich nur geringfügig. Zu erkennen ist dennoch, dass die H6-Werte des Uridin in einem höheren ppm-Bereich liegen, als die des Cytidin. Eine genaue Trennung lässt sich dadurch jedoch trotzdem nicht erzielen.

Das Kreuzsignal H1'-H1' von Hauptbase und Nachfolger wird in Abbildung 65 veranschaulicht. Erkennbar ist wie schon in Abbildung 62, dass das Duplett mit einem Adenin als Hauptbase sich von den anderen Datenpunkten in der Abbildung abgrenzt. Dem ungeachtet ist es nicht möglich die anderen drei Datencluster voneinander zu unterscheiden.

Die sehr gute Unterscheidung des Triplett –GG vom Triplett –GA zeigt das Streudiagramm in Abbildung 66. Auf den ersten Blick ist schon ersichtlich, dass es viel mehr Datenpunkte zu –GG, als zu –GA gibt, wodurch deutlich wird, dass –GG viel häufiger am Anfang einer RNA zu finden ist. Die Abbildung veranschaulicht das chemische Verschiebungspaar zwischen dem H8 der Hauptbase und dem H8 des Nachfolgers. Die H8-Werte des zweiten Guanosin liegen etwa zwischen 7,4 ppm und 7,7 ppm und bilden einen großen Cluster. Die H8-Werte des Adenins hingegen sind relativ verstreut, was aber auf die geringe Datenmenge zurückzuführen ist. Sie befinden sich etwa im ppm-Bereich zwischen 8 und 8,3.

4 Resultate und Diskussion

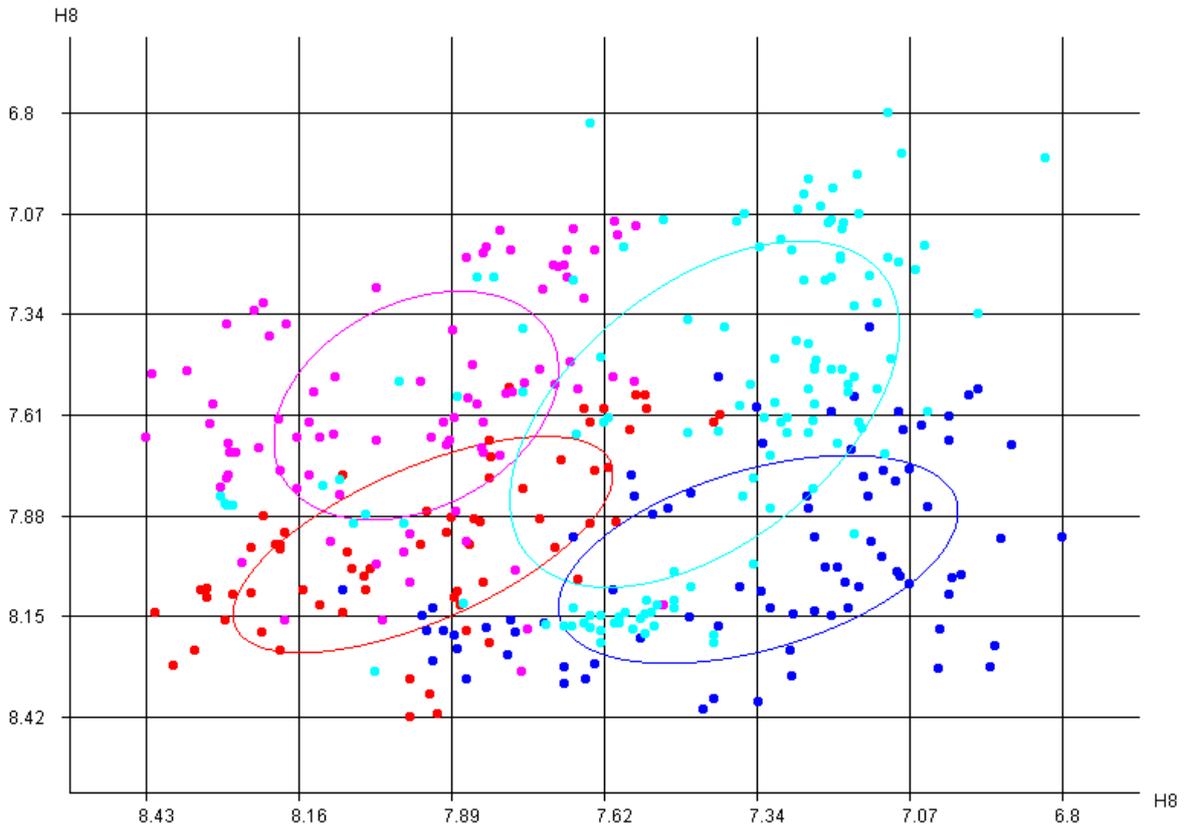


Abbildung 63: Streudiagramm mit der H8-chemischen Verschiebung der Hauptbasen Adenin oder Guanin (x-Achse) und der H8-chemischen Verschiebung des Nachfolgers Adenosin oder Guanodin (y-Achse). Das X kennzeichnet ein beliebiges Nukleotid als Vorgänger.

XAA, XAG, XGA, XGG

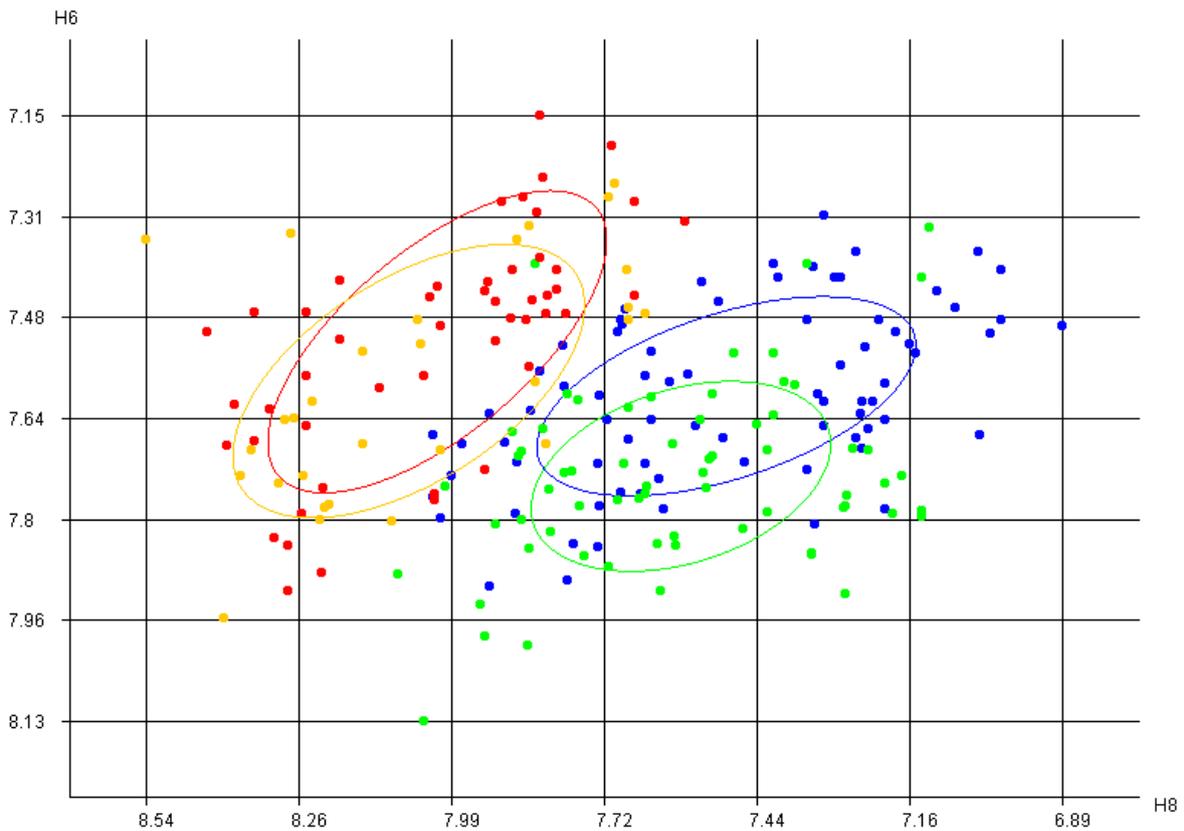


Abbildung 64: Streudiagramm mit der H8-chemischen Verschiebung der Hauptbasen Adenin oder Guanin (x-Achse) und der H6-chemischen Verschiebung der Nachfolger Cytidin und Uridin (y-Achse). Das X kennzeichnet ein beliebiges Nukleotid als Vorgänger.

XGC, XGU, XAC, XAU

4 Resultate und Diskussion

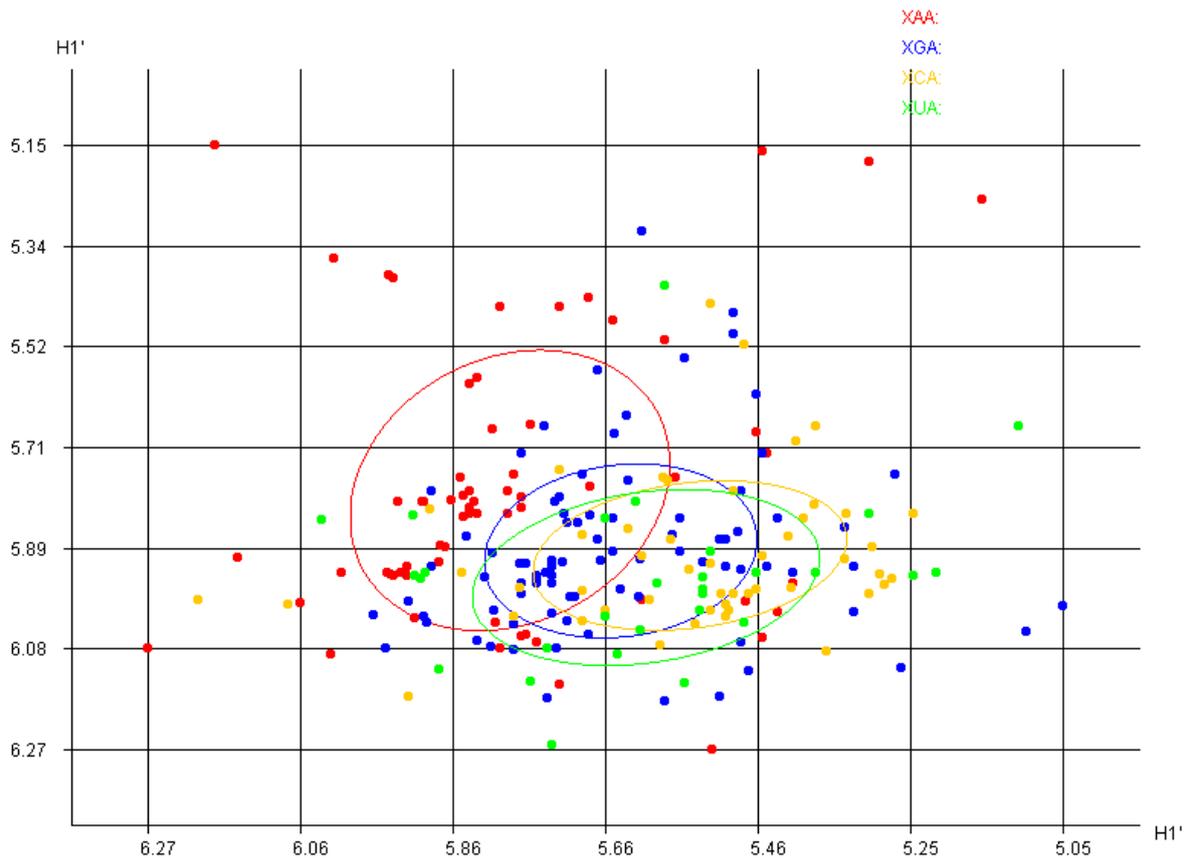


Abbildung 65: Streudiagramm mit der H1'-chemischen Verschiebung der Hauptbasen A, G, C und U (x-Achse) und der H1'-chemischen Verschiebung des Nachfolgers Adenosin (y-Achse). Das X kennzeichnet ein beliebiges Nukleotid als Vorgänger.

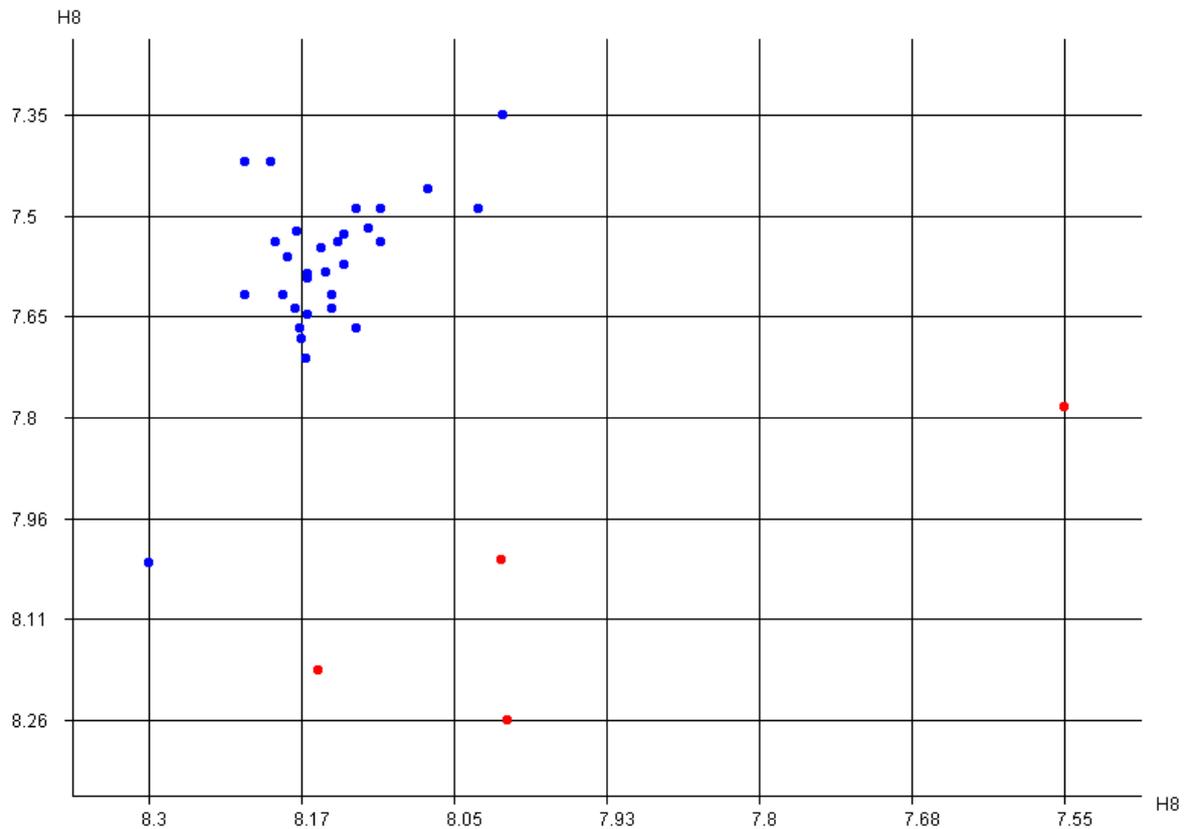


Abbildung 66: Streudiagramm mit der H8-chemischen Verschiebung der Hauptbase Guanosisin (x-Achse), welche auch gleichzeitig den Anfang der RNA darstellt und der H8-chemischen Verschiebung der Nachfolger Adenosin oder Guanosisin (y-Achse). Das X kennzeichnet ein beliebiges Nukleotid als Vorgänger.
-GG, -GA

4.3.2 Unterscheidung der Datengruppen nach ihrer Sekundärstruktur

Abbildung 68 zeigt ein Streudiagramm mit vier Clustern, die sich voneinander sehr gut unterscheiden lassen. Zwei dieser Cluster (rot und blau) befinden sich im „WC-Bereich“. Ihre H8-chemische Verschiebung der Hauptbase liegt tiefer, als die deren Nukleotide sich im „nWC-Bereich“ befinden. Die zwei AAX-Gruppen lassen sich ebenfalls sehr gut durch die H8-chemische Verschiebung des Vorgängers trennen. Im „WC-Bereich“ liegt diese etwa zwischen 7,5 ppm und 7,9 ppm und im „nWC-Bereich“ zwischen 7,9 ppm und 8,3 ppm.

Ist jedoch die Hauptbase Guanin, so lassen sich der „WC-Bereich“ und der „nWC-Bereich“ nicht durch den Vorgänger bestimmen, da sich dessen H8-chemische Verschiebung in derselben Umgebung befindet. Demzufolge ist es nur durch die H8-chemische Verschiebung der Hauptbase möglich die beiden AGX-Gruppen zu bestimmen.

Die Datenpunkte und Cluster aus der H8-chemischen Verschiebung des Nachfolgers und der Hauptbase sind in Abbildung 67 aufgezeigt. Wie schon in Abbildung 68 ist zu erkennen, dass sich die vier Cluster sehr gut voneinander unterscheiden lassen. Die H8-chemische Verschiebung der Hauptbase der Gruppe XAA im „WC-Bereich“ lässt sich von der im „nWC-Bereich“ sehr gut unterscheiden, genauso wie durch das H8 des Nachfolgers. Bei XGA ist dies ebenfalls zu erkennen.

Schließlich kann man sagen, dass durch das H8-H8 chemische Verschiebungspaar sowohl zwischen Vorgänger und Hauptbase, als auch zwischen Hauptbase und Nachfolger, eine Unterscheidung nach der Sekundärstruktur möglich ist.

Abbildung 70 und Abbildung 69 veranschaulichen das H5-H5-chemische Verschiebungspaar zwischen der Hauptbase und dem Nachfolger. In beiden Streudiagrammen ist zu erkennen, dass man die Datenpunkte eindeutig als WC oder nWC klassifizieren kann.

4 Resultate und Diskussion

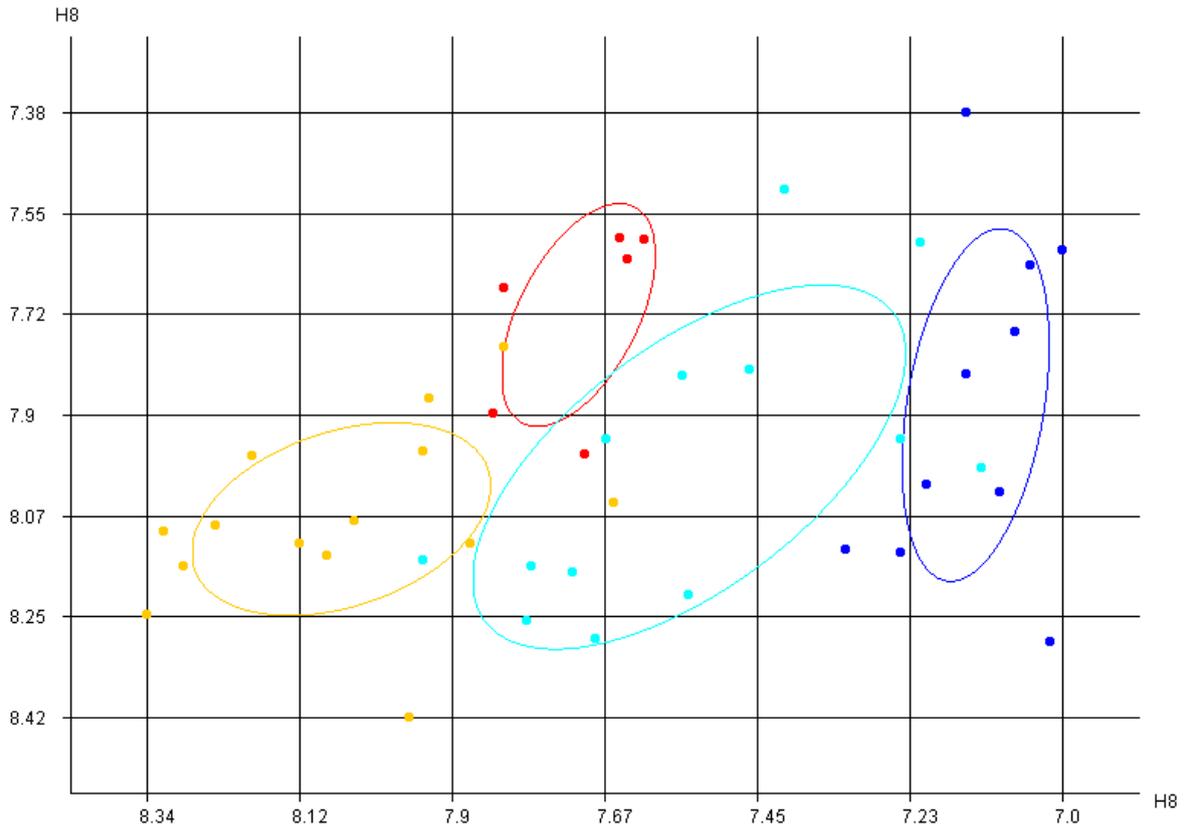


Abbildung 68: Streudiagramm mit den Kreuzsignalen aus der H8-chemischen Verschiebung der Hauptbasen Adenin und Guanin (x-Achse) mit der H8-chemischen Verschiebung des Vorgängers Adenosin (y-Achse).
AAX und **AGX** sind Nukleotide im stem, **AAX** und **AGX** im loop.
 Das X kennzeichnet ein beliebiges Nukleotid als Nachfolger.

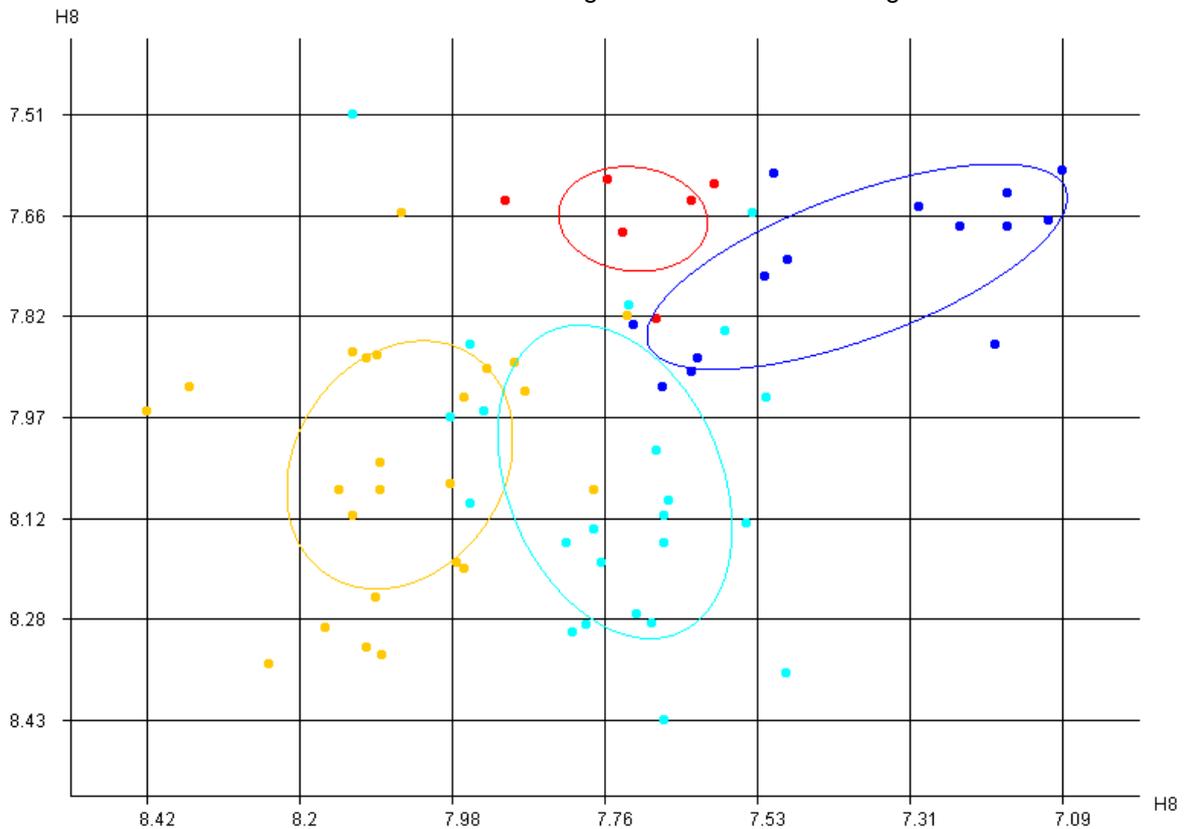


Abbildung 67: Streudiagramm mit den Kreuzsignalen aus der H8-chemischen Verschiebung der Hauptbasen Adenin und Guanin (x-Achse) mit der H8-chemischen Verschiebung des Nachfolgers Adenosin (y-Achse).
XAA und **XGA** sind Nukleotide im stem, **XAA** und **XGA** im loop.
 Das X kennzeichnet ein beliebiges Nukleotid als Vorgänger.

4 Resultate und Diskussion

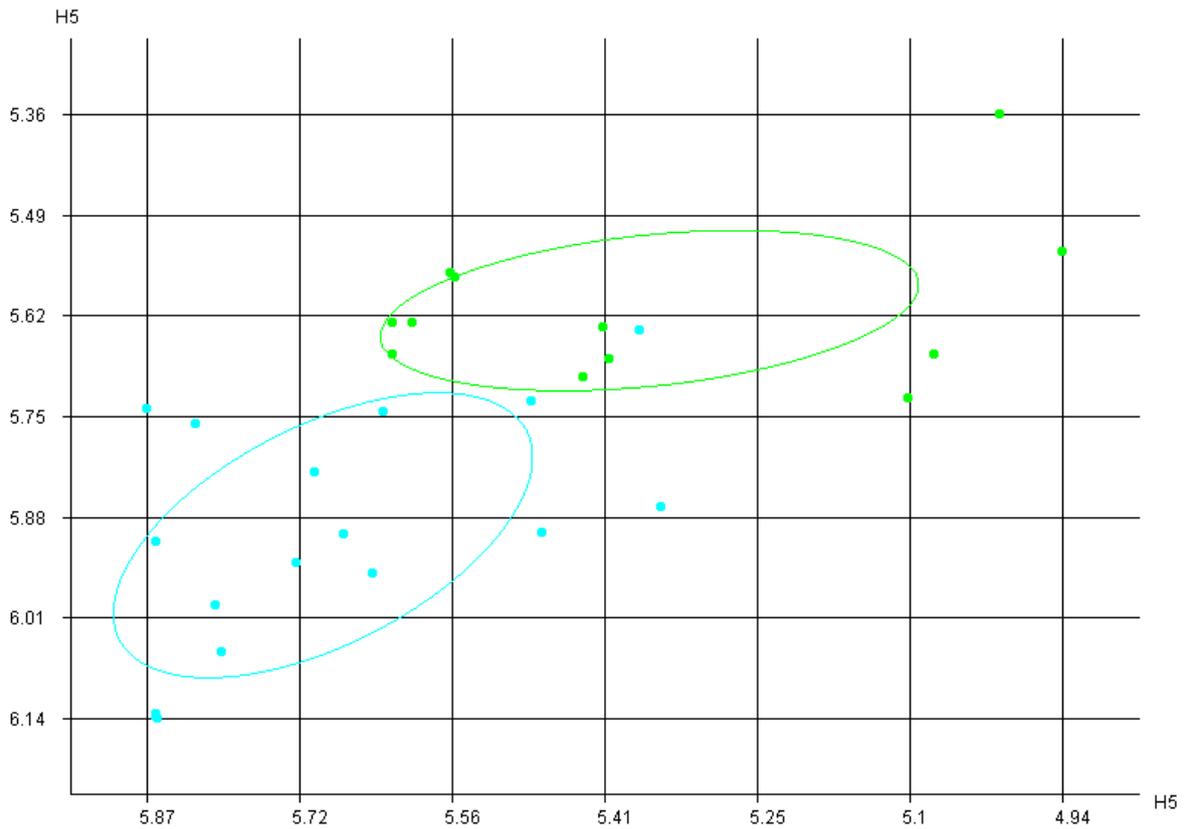


Abbildung 70: Streudiagramm mit den Kreuzsignalen aus der H5-chemischen Verschiebung der Hauptbase Uracil (x-Achse) mit der H5-chemischen Verschiebung des Nachfolgers Cytidin (y-Achse). XUC sind Nukleotide im stem, XUC im loop. Das X kennzeichnet ein beliebiges Nukleotid als Vorgänger.

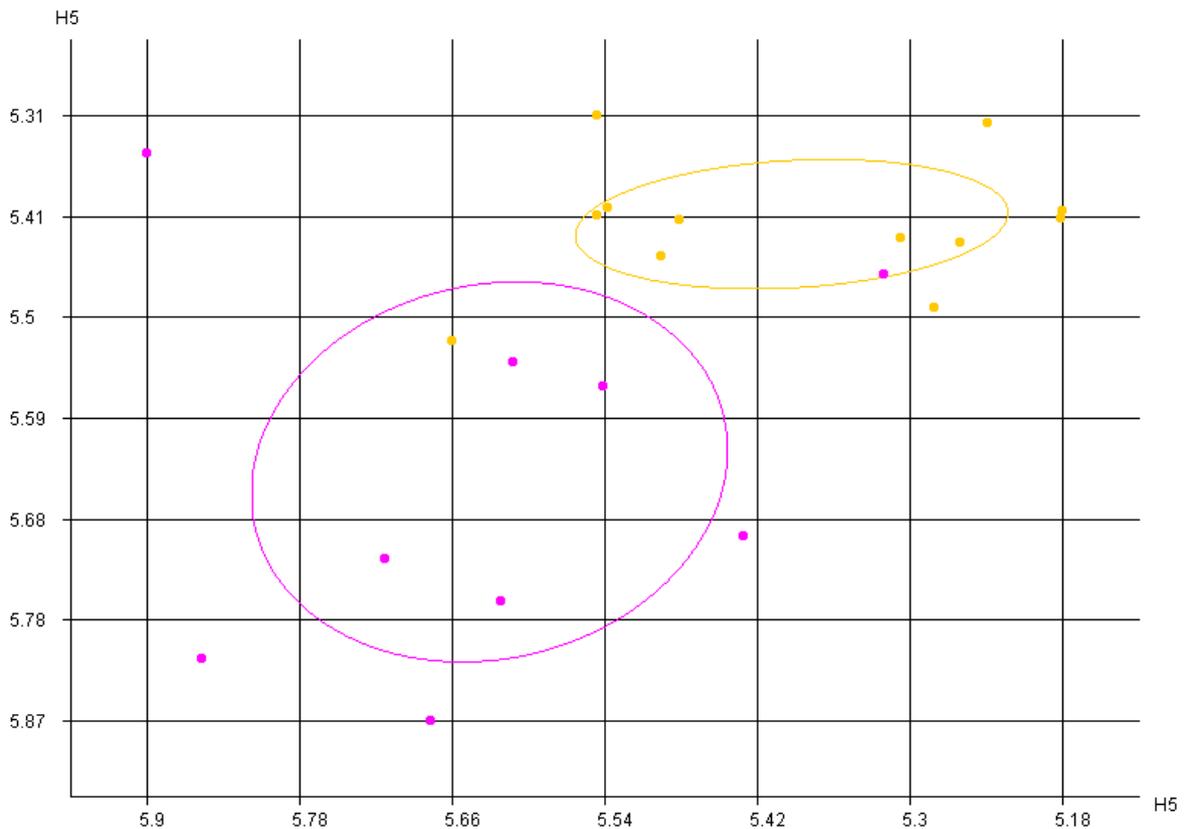


Abbildung 69: Streudiagramm mit den Kreuzsignalen aus der H5-chemischen Verschiebung der Hauptbase Cytosin (x-Achse) mit der H5-chemischen Verschiebung des Nachfolgers Uracil (y-Achse). XCU sind Nukleotide im stem, XCU im loop. Das X kennzeichnet ein beliebiges Nukleotid als Vorgänger.

4.4 Statistische Auswertung von TOCSY-, NOESY- und HSQC-Experimenten anhand der RNA FZL4

In diesem Abschnitt werden die in dieser Arbeit erstellten Statistiken für die TOCSY-, NOESY- und HSQC-Experimente, mithilfe einer unbekanntes, noch nicht veröffentlichten Beispielsequenz analysiert und bewertet.

Die verwendete RNA-Sequenz ist ein sogenannter „stemloop“ oder „hairpin loop“ und wird als FZL4 bezeichnet. Ihre Sekundärstruktur ist in Abbildung 71 veranschaulicht. Die experimentell bestimmten chemischen Verschiebungen dieser RNA wurden mit der selbstaufgestellten Statistik den Triplets zugeordnet.

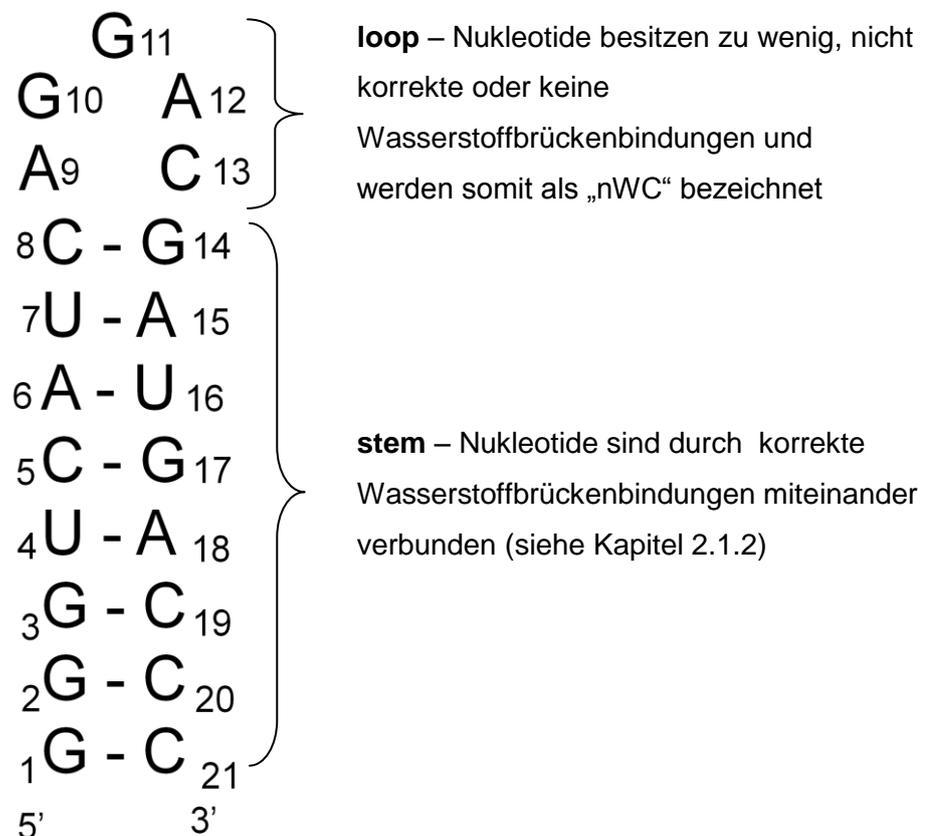


Abbildung 71: Sekundärstruktur der RNA FZL4
[Allain F. H.-T. ,nicht veröffentlicht]

Jedes Nukleotid konnte durch mindestens eins der drei Experimente ermittelt werden. Außerdem war es durch das Wissen über vorhandene Wasserstoffbrückenbindungen möglich die Sekundärstruktur nahezu komplett vorherzusagen. Nur das erste Guanosin der RNA konnte weder als WC noch als nWC vorhergesagt werden.

4 Resultate und Diskussion

Die HSQC-Statistik der RNA FZL4 sagte 27 der 55 möglichen chemischen Verschiebungen korrekt voraus. Für jede chemische Verschiebung wurden die fünf wahrscheinlichsten Zuordnungen analysiert. Am besten gelang die Vorhersage mit dem Kreuzsignal aus C5 und H5. Dabei wurden acht von neun Zuordnungen richtig erkannt, wobei der Großteil mit einer Wahrscheinlichkeit von über 70% vorhergesagt wurde. Außerdem wurden mehr als die Hälfte des Kreuzsignals von C1'-H1' vorhergesagt. Fünf von diesen lagen bei mehr als 90% Wahrscheinlichkeit.

Von den 27 erkannten chemischen Verschiebungen wurden neun gleich an erster Stelle, sechs an zweiter Stelle, sieben an dritter Stelle, vier an vierter Stelle und eine an fünfter Stelle vorhergesagt. Somit wurden immerhin ein Drittel der gefundenen chemischen Verschiebungen auch an erster Stelle der fünf möglichen erkannt.

Die HSQC-Statistik konnte folgende Nukleotide der RNA FZL4 einer Sekundärstruktur zuordnen: G2, G3, U4, C5, A6, U7, C8, A9, G10, G11, A12, C13, G14, A15, U16, G17, A18, C19 und C20. Lediglich bei G1 und C21 konnte keine Wasserstoffbrückenbindung bei der bestmöglichen richtigen Zuordnung gefunden werden.

Zu etwa 78% konnte die Statistik der TOCSY- Werte alle chemischen Verschiebungen von H5-H6 richtig vorhersagen. Fünf von diesen wurden mit über 90% Wahrscheinlichkeit vorhergesagt. Lediglich zwei Nukleotide konnten nicht unter den ersten fünf bestmöglichen Nukleotiden erkannt werden.

Von den sieben gefundenen Verschiebungen lagen zwei an erster und zweiter Stelle und jeweils eine an dritter, vierter und fünfter Stelle.

Ebenfalls konnten einigen Nukleotiden der RNA FZL4 eine Sekundärstruktur vorhergesagt werden. Den Basen G3, U4, C5, A6, U7, C8, C19, C20 und C21 wurde eine WC-Bindung zugeordnet und der Base A9 eine nWC-Bindung. Den Nukleotiden von G10 bis A18 konnte mithilfe der TOCSY-Statistik keine Sekundärstruktur vorhergesagt werden.

Das NOESY-Spektrum der RNA umfasst 133 Signale. Schon im „stem“-Bereich gibt es 24 verschiedene mögliche Signale der RNA FZL4. Allein durch diese große Anzahl war damit zu rechnen, dass die Statistik nicht so gut ist, wie die der anderen Experimente. Im „loop“-Bereich existieren häufig Signale zwischen der Hauptbase und dem zweiten Vorgänger oder Nachfolger, aber auch noch mit weiter entfernten Nukleotiden. Außerdem entstehen Signale zwischen unterschiedlichsten Atomen, die demzufolge nicht erkannt werden konnten. Im TOCSY-Spektrum gibt es nur Signale zwischen H5 und H6 und auch im HSQC gab es nur ein Signal, welches nicht in der Statistik eingebaut ist.

Lediglich 17 chemische Verschiebungen wurden richtig vorhergesagt, wovon acht an erster, zwei an zweiter, zwei an dritter, drei an vierter und zwei an fünfter Stelle lagen.

4 Resultate und Diskussion

Am besten gelang die Vorhersage bei der chemischen Verschiebung von H8 mit dem H8 des Vorgängers. Hier wurden zwei mit mehr als 95 prozentiger Wahrscheinlichkeit richtig erkannt und lagen dabei an Stelle eins und zwei. Ein weiteres Kreuzsignal welches zufriedenstellend vorhergesagt wurde war H5-H1' eines Uracil oder Cytosin. Vier von sechs möglichen wurden erkannt und von diesen lagen zwei an erster Stelle. Weniger gut gelang die Vorhersage bei H1'-H1' und bei H6-H8. Diese Kreuzsignale wurden überhaupt nicht erkannt.

Durch die extrem große Anzahl an Signalen war es kaum möglich eine gute Vorhersage zu erhalten. Viele Signale der ^1H -Atome überlappen mit weiteren ^1H -Atomen. Auch aus diesem Grund war es nur möglich 13 der 21 Nukleotide vorherzusagen. Diese waren G1, G2, G3, U4, U7, A9, G10, A12, U16, G17, A18, C19, C20. Genauso wenig gut vorhergesagt wurde die Sekundärstruktur der RNA. Nur 11 der 21 Nukleotide wurde eine bestimmte 2D-Struktur zugeschrieben. Dabei liegen G2, G3, U4, C5, A18, C19, C20 und C21 im „stem“ und sind demzufolge WC und G11, A12 und C13 wurden als nWC vorhergesagt und liegen somit im „loop“-Bereich.

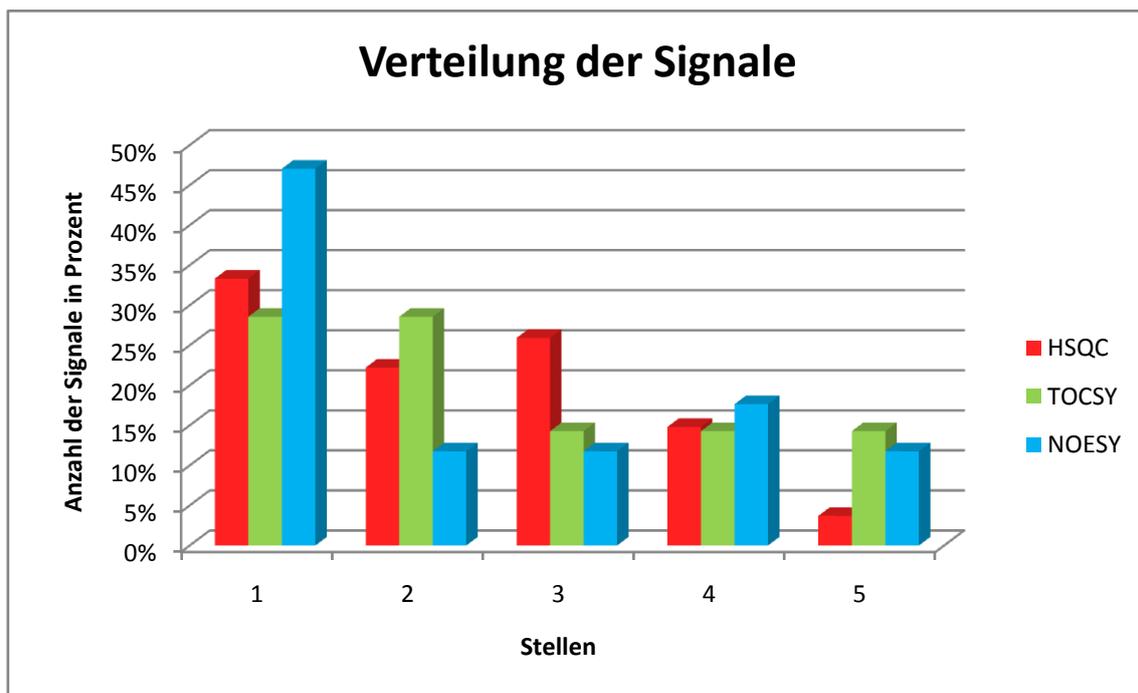


Abbildung 72: Prozentuale Verteilung der Signale an den ersten fünf Stellen.

Abbildung 72 veranschaulicht die prozentuale Verteilung der vorhergesagten Signale an ihren Stellen in der Statistik. Gut zu erkennen ist dabei, dass fast die Hälfte der im NOESY-Experiment erkannten Signale an erster Stelle liegen. Ebenfalls sind die meisten Signale von TOCSY und HSQC an erster Stelle zu finden. An fünfter Stelle sind prozentual die wenigsten Signale erkannt worden.

4 Resultate und Diskussion

Abschließend ist zu sagen, dass die aufgestellte Statistik für die TOCSY- und HSQC-Experimente zufriedenstellend ist. Es konnten zwar nicht alle chemischen Verschiebungen erkannt, aber alle Nukleotide vorhergesagt werden. Nicht immer liegen die richtigen Triplets an erster Stelle, was die Vorhersage erschwert, da vorher nicht bekannt ist, welches Signal genau zu welchem Triplet bzw. zu welchem Nukleotid passt. Jedoch bietet diese Statistik eine gute Basis für weitere Erforschungen und Verbesserungen. Für die NOESY-Statistik ist es um einiges komplizierter. Durch die Fülle an Signalen ist es kaum möglich eine richtige Vorhersage zu erstellen.

5 Ausblick

5.1 Data-mining

Zur besseren Analyse und Zuordnung der Datenpunkte von Ribonukleinsäuren wäre ein größerer Datensatz und somit eine umfangreichere Informationsmenge von Vorteil. Erhöht sich die Anzahl der Einträge an RNA-Strukturen in der BMRB, so könnten die Daten in die Statistik eingebaut werden und diese somit verbessern.

Des Weiteren könnte untersucht werden, wie sich Fehlpaarungen verschiedener Nukleinbasen auf die chemischen Verschiebungen auswirken. Auch chemische Verschiebungen modifizierter Nukleinbasen könnten statistisch ausgewertet werden.

5.2 Erweiterung der Statistiken für die Experimente

Eine Möglichkeit zum Ausbau der Statistik ist es weitere chemische Verschiebungen in die Statistik einzubauen. Vor allem interresiduale Verschiebungen wären interessant zu untersuchen. Beispielsweise erkennt man im NOESY-Spektrum ein Signal zwischen dem H2-Atom eines Adenosin mit dem H1'-Atoms der sich schräg auf dem gegenüberliegenden Strang befindenden Nukleotids, kurz: H2[A] – H1'[k+1] (Zeichenerklärung siehe Abbildung 73).

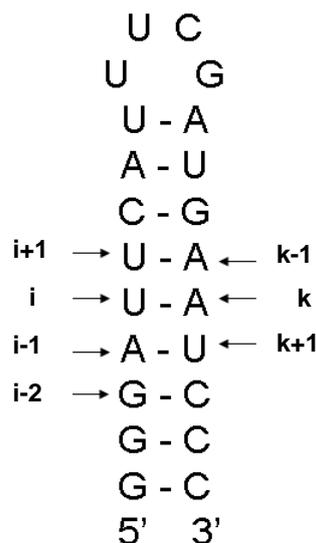


Abbildung 73: Sekundärstruktur der RNA TASL1.

i ist die betrachtete Base, $i+1$ ihr Nachfolger, $i-1$ ihr Vorgänger, k ist die gegenüberliegende Base von i und jeweils wieder der Vorgänger ($k-1$) und Nachfolger ($k+1$) von k , $i-2$ bedeutet der zweite Vorgänger der Base i
 [Allain F. H.-T., nicht veröffentlicht]

Weiterhin eröffnete sich die Möglichkeit, nicht nur den ersten Vorgänger eines Nukleotids, sondern wie in Abbildung 73 ersichtlich auch den zweiten Vorgänger ($i-2$) zu analysieren. Demzufolge wäre es denkbar die Messwerte noch eindeutiger zu einem Nukleotid zuordnen zu können.

5.3 Prüfen der Statistik

Um die Aussagekraft der erstellten Statistik zu stärken wäre es notwendig mehrere unbekannte RNA-Strukturen zu überprüfen. Durch die Analyse jeder unbekanntes RNA könnte die Statistik lernen um dadurch besser zu werden.

Bei der NOESY-Statistik könnte man versuchen den Datensatz an chemischen Verschiebungen zu verkleinern oder die Wahrscheinlichkeiten anders zu berechnen, beispielsweise durch eine $\frac{1}{2}$ Standardabweichung.

Desweiteren wäre es von Vorteil eine Art Gewichtung einzubauen, da im Moment nur die höchste Prozentzahl zählt und nicht das was es ausgibt, sprich ein einzelnes Nukleotid (beispielsweise XAX), Duplett (beispielsweise XAG) oder ganzes Triplet (beispielsweise GAC) an Nukleotiden. Das gleiche gilt dabei auch für die Sekundärstruktur. Wird ein Triplet als WC_WC_WC vorhergesagt, müsste es besser gestellt sein, als eine Vorhersage, in der kein Nukleotid, eins oder nur zwei Nukleotide zugeordnet wurden, beispielsweise WC_X_X.

5.4 RASP

Das von Sebastian Eube weiterentwickelte Programm RASP (**RNA Sequence assignment Prediction**) dient bisher zur statistischen Auswertung von H5-H6-Wertepaaren von Uridin und Cytidin im TOCSY-Experiment (siehe Abbildung 75).

Es visualisiert diese Wertepaare und ihre Verteilung zu den 8 Dinukleotidgruppen AC, UC, GC, CC AU, GU, CU und UU. Außerdem ist das Programm in der Lage die drei besten Wahrscheinlichkeiten für die Zuordnung der Wertepaar zu einer Datengruppe auszugeben und ihrer Primärstruktur zuzuordnen. Dafür benötigt es die Sequenz im FASTA-Format und die chemischen Verschiebungen in einer sogenannten Peakliste.

5 Ausblick

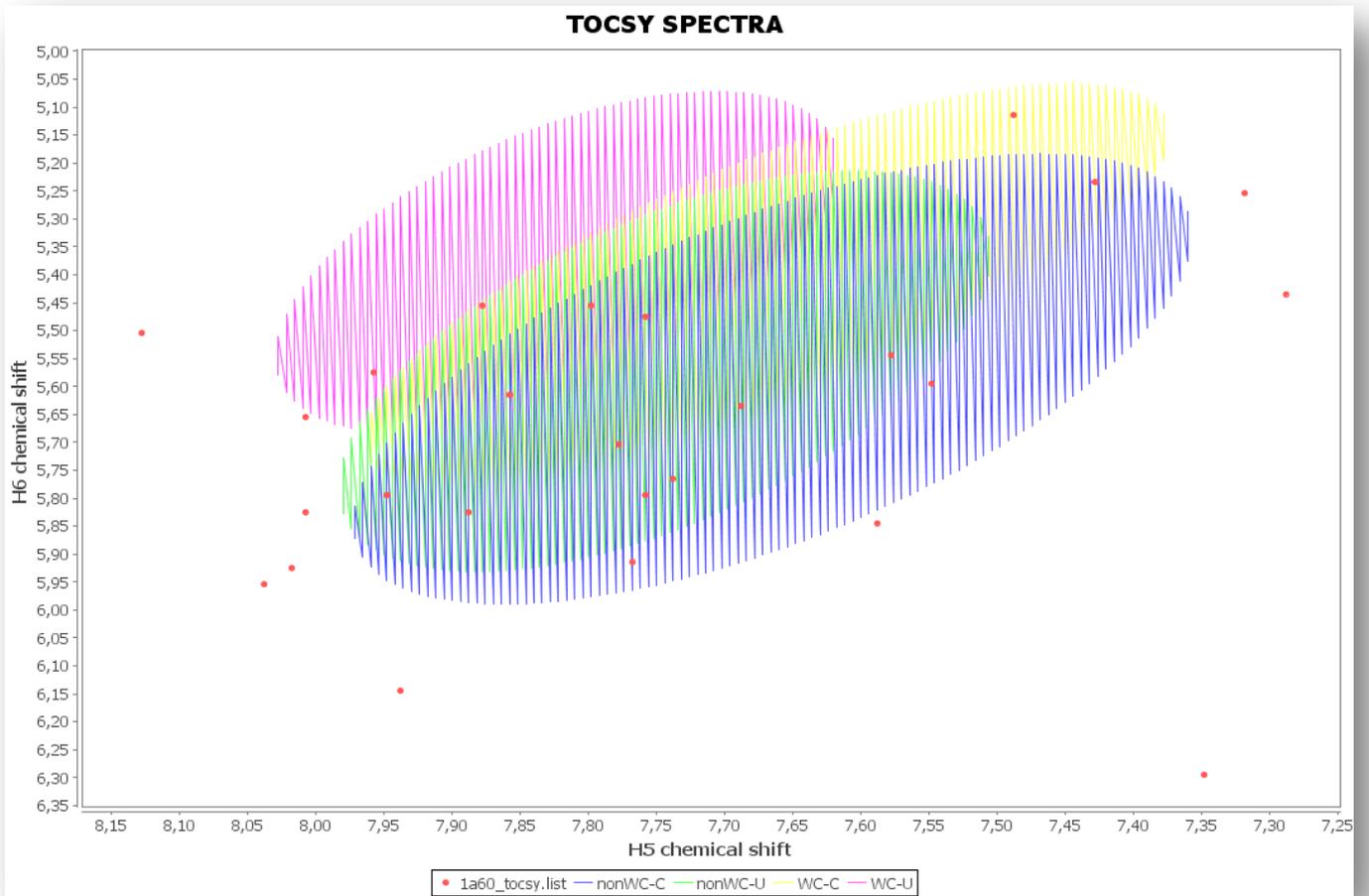


Abbildung 75: H5-H6 Diagramm der RNA 1A60. Eingezeichnet sind die Verteilungen der vier Gruppen: non Watson-Crick-Base Cytosin, Watson-Crick-Base Cytosin, non Watson-Crick-Base Uracil und Watson-Crick-Base Uracil. [aus dem Programm RASP]

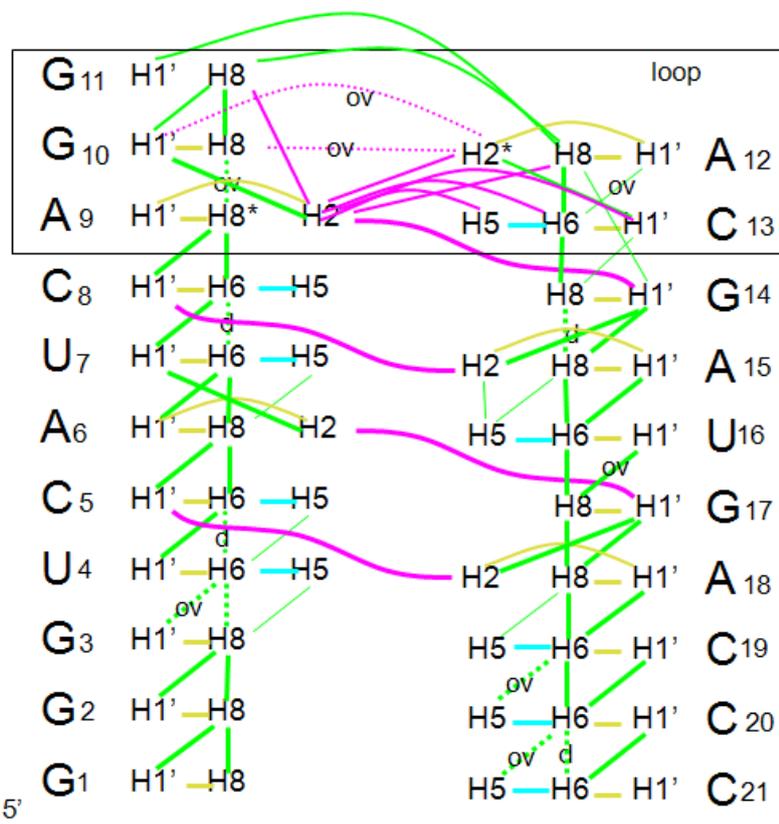


Abbildung 74: Bindungen innerhalb der Sekundärstruktur eines RNA-stems [Allain F. H.-T. nicht veröffentlicht]

Auf der Grundlage der in dieser Arbeit erstellten Statistik soll das Programm RASP weiter aufgebaut und umstrukturiert werden. Unter anderem soll es basierend auf Triplets und nicht wie zuletzt auf Dupletts die Zugehörigkeit eines Wertepaares zu einer Datengruppe zuordnen. Dabei sollen alle drei beschriebenen und in der Statistik vorkommenden Experimente verwendet werden. Desweiteren soll eine Anzeige der 2D-Struktur, als auch der 3D-Struktur realisiert werden. Innerhalb der 2D-Struktur wäre es von Vorteil, wie in Abbildung 74 erkennbar, Bindungen anzeigen zu lassen und ToolTip-Informationen dieser ausgeben zu lassen.

5.5 Zuordnungsstrategie

In weiteren Arbeiten könnte auf der Grundlage der erstellten Statistik eine neuartige Zuordnungsstrategie entwickelt werden. Das Ziel wäre mit Hilfe der Informationen aus Sequenz und chemischen Verschiebungen die Sekundärstruktur der unbekannteren RNA vorherzusagen.

Literaturverzeichnis

- [1] Dingermann, Theo; Zündorf Ilse: RNA als Arzneistoff: Faszinierendes Konzept, aber schwer zu realisieren. In: Pharmazeutische Zeitung. – Eschborn/Ts. : Govi Verlag Pharmazeutischer Verlag GmbH. - 15/2010 S. 18-27
- [2] Mortimer, Charles; Müller, Ulrich: Chemie – 9.Auflage – Stuttgart: Thieme Verlag, 2007
- [3] Mülhardt, Cornel: Der Experimentator: Molekularbiologie/ Genomics – 6.Auflage – Heidelberg: Springer Akademischer Verlag, 2009
- [4] Fuchs, Georg: Allgemeine Mikrobiologie – 8.Auflage – Stuttgart: Thieme Verlag, 2007
- [5] Gaedecke, Nicola: Biowissenschaftlich recherchieren – 1.Auflage – Basel: Birkhäuser Verlag, 2007
- [6] Rashidi, Hooman; Bühler, Lukas: Grundriss der Bioinformatik: Anwendungen in den Biowissenschaften und der Medizin – 1.Auflage – Heidelberg: Spektrum Akademischer Verlag, 2001
- [7] Lesk, Arthur: Bioinformatik: Eine Einführung – 1.Auflage – Heidelberg: Spektrum Akademischer Verlag, 2003
- [8] Friebolin, Horst: Ein- und zweidimensionale NMR- Spektroskopie: Eine Einführung – 4.Auflage – Weinheim: WILEY-VCH Verlag, 2006
- [9] Holtzhauer, Martin: Methoden in der Proteinanalytik – 1.Auflage – Heidelberg: Springer- Verlag, 1996
- [10] Lottspeich, Friedrich; Haralabos, Zorbas: Bioanalytik – 1.Auflage – Heidelberg: Spektrum Akademischer Verlag, 1998

Literaturverzeichnis

- [11] Rutgers, the State University of New Jersey Department of Chemistry and Chemical Biology <info@rcsb.org>: Protein Data Bank, URL: <<http://www.pdb.org>>, Stand: 26.11.2009
- [12] Departement of Biochemistry University Wisconsin- Madison <bmrhelp@bmr.wisc.edu>: Biological Magnetic Resonance Data Bank, URL: <<http://www.bmr.wisc.edu/bmr/about/policies.html>>, Stand: 26.11.2009
- [13] Hildebrandt, Juliana: Automatische Auswertung von TOCSY- Experimenten an RNA-Molekülen. Bachelorarbeit, TU Dresden, 2008
- [14] Bader, Reto <unbekannt>: Pharmazeutische Chemie. URL:< <http://e-collection.ethbib.ethz.ch/eserv/eth:25261/eth-25261-01.pdf>>, Stand: 12.10.09
- [15] Kühlewein, Angelika <unbekannt>: Anwendung NMR-spektroskopischer Methoden zur Strukturaufklärung von Proteinen und Peptiden. URL:< http://deposit.ddb.de/cgi-bin/dokserv?idn=96342808x&dok_var=d1&dok_ext=pdf&filename=96342808x.pdf>, Stand: 12.10.09
- [16] Hecht, Oliver <unbekannt>: Aufklärung der dreidimensionalen Struktur des Proteins Amoebapore A mit Hilfe der mehrdimensionalen NMR-Spektroskopie. URL:< http://deposit.d-nb.de/cgi-bin/dokserv?idn=971990042&dok_var=d1&dok_ext=pdf&filename=971990042.pdf>, Stand: 12.10.09
- [17] Bayer, Peter <unbekannt>: Strukturaufklärung von Biomolekülen mit Hilfe der mehrdimensionalen Kernspinresonanz: Proteine aus der Signaltransduktion und der Regulation von Zellkernprozessen. URL:< http://deposit.ddb.de/cgi-bin/dokserv?idn=961682191&dok_var=d1&dok_ext=pdf&filename=961682191.pdf>, Stand: 12.10.09
- [18] Eube, Sebastian: Programm zur Auswertung von NMR-Experimenten am RNA-Stem, Bachelorarbeit, HS-Mittweida, 2009
- [19] Crick, Francis <unbekannt>: Ideas on the protein synthesis (Oct.1956). URL: <http://profiles.nlm.nih.gov/SC/B/B/F/T/_/scbbft.pdf>, Stand: 14.11.2009

Literaturverzeichnis

- [20] Salzer, Reiner <info@fiz-chemie.de>: Kernresonanzspektroskopie (NMR).URL: <http://www.chemgapedia.de/vsengine/topics/de/vlu/Chemie/Analytische_00032Chemie/NMR_00045Spektroskopie/index.html>, Stand: 29.06.2010
- [21] Zangger, Klaus <klaus.zangger@kfunigraz.ac.at>: 1D-2D-NMR. URL: <<http://www.uni-graz.at/klaus.zangger/1D+2DNMR.pdf>>, Stand: 29.06.2010
- [22] Aeschbacher,Thomas: Resonance assignment strategies based on chemical shift statistics, Diplomarbeit, Institut of Molecular Biology and Biophysiks ETH Zürich, 2009
- [23] Clark, David P.: Molecular Biology: Understanding the Genetic Revolution – 1.Auflage – München: Spektrum Akademischer Verlag, 2006
- [24] Renneberg, Reinhard: Biotechnologie für Einsteiger – 2.Auflage – München: Spektrum Akademischer Verlag, 2007
- [25] Cromsigt, Jenny; Hilbers, Cees; Wijmenga, Sybren: Prediction of proton chemical shifts in RNA: Their use in structure refinement and validation. In: Journal of Biomolecular NMR. – The Netherlands.: Kluwer Academic Publishers. – 21/2001 S. 11-29
- [26] Kremer, Andreas: Crashkurs Biochemie – 1.Auflage – München: Urban & Fischer Verlag, 2005
- [27] Igo-Kemes, Tibor.: Seminar der Biochemie – 1.Auflage – Putzbrunn: Rukker Verlag, 1999
- [28] OriginLab Corporation <info@originlab.com>: Box Plots. URL: <<http://www.originlab.com/index.aspx?go=Products/Origin/Graphing/2D&pid=959>>, Stand: 25.07.2010

Selbstständigkeitserklärung

Hiermit erkläre ich, dass ich die von mir am heutigen Tag eingereichte Bachelorarbeit mit dem Thema

„Statistische Analyse der chemischen Verschiebungen von RNA-Molekülen“

vollkommen selbstständig verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel benutzt sowie Zitate kenntlich gemacht habe.

Mittweida, den 22.08.2010

Unterschrift

Anhang

Tabelle 2: PDB und BMRB ID der für die Statistik verwendeten reinen RNA-Strukturen. Hinter allen kursiv gedruckten IDs stehen Strukturen, bei denen unzureichende Informationen zu den Wasserstoffbrückenbindungen gegeben waren.

PDB	BMRB	PDB	BMRB	PDB	BMRB	PDB	BMRB
1a60	4120	1ldz	4226	<i>1r2p</i>	<i>5962</i>	1ysv	6485
1bn0	4175	1m82	5528	<i>1r4h</i>	<i>5980</i>	<i>1z2j</i>	<i>6543</i>
1esy	4780	1mfj	5559	1r7w	6076	1z30	6562
1hwq	5007	1mfy	5553	1r7z	6077	1zc5	6633
1l4C	5023	1mnx	5919	1s34	6062	28sr	4750
<i>1idv</i>	<i>5046</i>	1n8x	5773	<i>1s9s</i>	<i>6094</i>	<i>2aht</i>	<i>6756</i>
1jo7	4816	1na2	5632	1t28	6115	<i>2f87</i>	<i>10014</i>
<i>1ju7</i>	<i>5170</i>	<i>1nc0</i>	<i>5655</i>	1tjz	6299	2fdt	10018
1kka	5256	1ow9	5852	1xhp	6320	2tpk	4253
1l1w	5321	1pjy	5834	1ymo	6477	3php	4250
1lc6	5371	1q75	5932				