Analytische und empirische Untersuchungen über abstrakte Shapes von RNA-Sekundärstrukturen

Dissertation zur Erlangung des akademischen Grades eines Doktors der Naturwissenschaften (Dr. rer. nat.) der Technischen Fakultät der Universität Bielefeld

vorgelegt von
Soummaya Abdul-Hak

Bielefeld, im Dezember 2010

Gedruckt auf alterungsbeständigem Papier nach ISO 9706.

Danksagung

An erster Stelle gilt mein größter Dank dem Leiter der Arbeitsgruppe Praktische Informatik, Prof. Dr. Robert Giegerich, der mich zu dieser Promotion anregte und mich in all den Jahren wissenschaftlich hervorragend betreute. Seine stete Gespächsbereitschaft, seine Unterstützung und seine wertvollen Ratschläge und Hinweise haben diese Arbeit erst möglich gemacht.

Ich danke sehr Frau Prof. Dr. Ellen Baake für ihre freundliche Bereitschaft, die vorliegende Arbeit zu begutachten.
Bei Herrn Prof. Dr. Jens Stoye möchte ich mich herzlich für die Bereitschaft bedanken, den Prüfungsvorsitz zu übernehmmen, ebenso bei Herrn Dr. Benjamin Kormeier für die Teilnahme am Prüfungsausschuss.

Großer Dank gilt Dr. Peter Steffen für seine zahlreichen Anstrengungen, seine Hilfsbereitschaft und für die Korrektur dieser Arbeit.
Des weiteren bedanke ich mich besonders bei Dr. Jens Reeder, der mir als guter Arbeitkollege mit Rat und Tat immer hilfsbereit zur Seite stand.
Herzlich bedanken möchte ich mich bei Britta Quisbrok für die Unterstützung in allen Lebenslagen, und ihre Hilfe in LaTeX.

Großer Dank geht an alle meine Kolleginnen und Kollegen (Susanne Konermann, Stefanie Schirmer, Janina Reeder, Jan Krüger, Sven Hartmeier, Jan Reinkensmeier, Stefan Janssen, Marc Rehmsmeier, Georg Sauthoff, Daniel Hagemeier) der AG Praktische Informatik und des BiBiServ für die immer gute und freundschaftliche Atmosphäre. Diese Zeit werde ich immer in bester Erinnerung behalten.

Nicht zu letzt möchte ich mich ganz herzlich bedanken bei meiner Mutter, die mit ihrem Körper gestorben ist, aber ihre Seele ist bei mir, in meiner Seele, in meinem Herzen und in meinem Körper.
Ebenfalls danke ich in besonderer Weise meinem Vater und meinen Geschwistern (Shaffika, Rana und Mustafa) und meiner lieben treuen Freundin Rana Manzalgi. Sie sind zusammen der Sonnenschein meines Lebens.

Inhaltsverzeichnis

1 Einführung und Überblick

In den letzten Jahrzehnten werden viele Anstrengungen unternommen, die grundlegende Ansätze zur Vorhersage von RNA-Sekundärstrukturen zu entwickeln.
Der am weitesten verbreitete Ansatz ist dabei die Minimierung der freien Energie des Moleküls. Hierbei wird angenommen, dass die tatsächliche Struktur diejenige mit der kleinsten freien Energie ist.
Häufig findet sich die tatsächliche Struktur allerdings unter den Strukturen mit einer etwas höheren und nur fast minimalen freien Energie. Viele neuere Faltungsprogramme berechnen daher auch zusätzlich suboptimale Strukturen. Der Nachteil dieses Ansatzes ist, dass der Raum der suboptimalen Strukturen sehr viele ähnliche Strukturen enthält, obwohl wir eigentlich nur an denen mit signifikanten Unterschieden interessiert sind. Aus diesem Grund wurde den Ansatz der abstrakten Shapes eingeführt [4]. Abstrakte Shapes sind Abstraktionen von Strukturen, wobei eine einzelne Shape eine Klasse von ähnlichen Strukturen enthält. Weiterhin enthält eine Shape als repräsentative Struktur diejenige mit der kleinsten freien Energie [9].

Der abstrakte Shape-Ansatz wurde von Giegerich et al. [4] beschrieben. In dem genannten Artikel wird jede Klasse ähnlicher Sekundärstrukturen von einer Shape dargestellt. Dabei werden die nativen Strukturen durch den besten Shape repräsentiert.
In dieser Arbeit leiten wir einige interessante Ergebnisse analytischer und empirischer Untersuchungen über abstrakte Shapes von RNA-Sekundärstrukturen ab.

Der Plan dieser Arbeit wird wie folgt dargestellt:

Kapitel 2: befasst sich mit zusammengefassten Erklärungen über die RNA-Sekundärstrukturen und über die bekannten Ergebnisse von RNA-Sekundärstrukturen. Dann wird die Beschreibung der Baum-Grammatik für RNA-Sekundärstrukturen [12] mit ihrer Knotenmarkierung, die wir in die folgenden Kapiteln 4 und 5 benötigen, erklärt. Danach folgt die Beschreibung des Abstrakt-Shapeansatzes, der eine Abstraktionsabbildung $\pi(x)$ von jeder Sekundärstruktur x ist und Sekundärstrukturen mit identischer Abstraktion zu Shapes zusammenfasst.

Kapitel 3: In diesem Kapitel beginnen wir mit der Rekurrenzenformel für die Anzahl der Shapes der Länge $\leq n$. Diese Formel gehört zum Shape-Typ 3. Wir interessieren uns hauptsächlich für die Rekurrenzenformel der Anzahl der Shapes zu Sequenzen der Länge $\leq n$ des Shape-Typ 5 mit ihrem asymptotischen Wert. Dabei spielt der Shape-Typ 5 in dieser Arbeit eine Schlüsselrolle.

Kapitel $4, 5$: hier geht es darum, die Formel für die Anzahl der Shape-Typen $i\big(i \in \{1...5\}\big)$ mit n Klammerpaaren herzuleiten. Dieses geschieht mit Hilfe der Shapes-Grammatik für die Typen $i\big(i \in \{1...5\}\big)$, die wir von der Abstraktions-Abbildung $\pi_i\big(i \in \{1...5\}\big)$ und vom Homomorphismus $\nu_i\big(i \in \{1...5\}\big)$ herleiten. Danach bestimmen wir die asymptotische Anzahl der Shape-Typen $i\big(i \in \{3, 4, 5\}\big)$ mit n Klammerpaaren.

Um den asymptotischen Wert in dieser Arbeit zu bestimmen, ermitteln wir die erzeugende Funktion $f(z) = \sum_{n\geq 0} f_n z^n$ für die Rekurrenzenformel und wenden ein Theorem von Flajolet-Odlyzko an.

Kapitel 6: befasst sich mit der empirischen Untersuchung bezüglich der Größe des Shaperaums-Typ 5. Der asymptotische Wert für die Anzahl der Shape-Typen aller Sequenzen der Länge n hat die allgemeine Formel $f(n) = a^n \cdot b \cdot n^{-3/2}$. Eine solche Formel ist nicht bekannt für die Erwartungswertberechnung. Die empirische Untersuchung werden wir mit Zufallssequenzen und dem Programm RNAshapes [13] durchführen, und wir werden mit der Anwendung der statistische Analyse die Parameter der Formel $f(n) = a^n \cdot b \cdot n^{-3/2}$ bestimmen.

2 Grundlegende Definitionen und vorliegende Ergebnisse

2.1 RNA-Struktur, Faltungsraum

Die RNA ist eine lineare Abfolge von Basen, genannt Primärstruktur [15]. Die Primärstruktur der RNA ist nahezu identisch mit der von DNA, d.h die Bausteine sind Basen und Phosphate.
Der Hauptunterschied ist die Ribose anstelle der $2'-$Desoxyribose; sie ist die Grundlage für die verschiedenen Helixkonformationen von RNA und DNA und für die chemische Instabilität von RNA im Vergleich zu DNA.

Die vier verschiedenen Basen sind die zwei Purine, Adenin und Guanin, und die zwei Pyrimidine, Cytosin und Uracil. Das Vorkommen von Uracil in RNA anstelle von Thymin in der DNA, die sich nur durch die bei Thymin zusätzliche $5-$ Methylgruppe unterscheiden, ist zum Teil verantwortlich für die höhere thermodynamische Stabilität der doppelsträngigen RNA Basen. Nukleoside (Base plus Ribose) und die Nukleotide (Nukleoside plus Phosphat) werden meist A,G,C und U abgekürzt.
Die offiziellen Abkürzungen sind Ade, Gua, Cyt und Ura für die Basen; A,G,C und U für Nukleotide. Abbildung 1 zeigt die vier Basen.

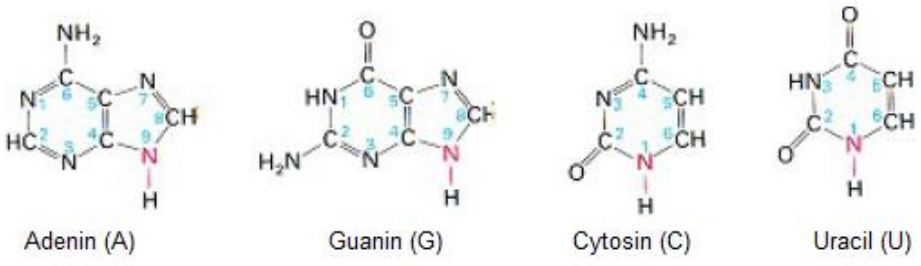


Abbildung 1*: Die vier RNA Basen: A und G sind Purine, U und C sind Pyrimidine.*

RNA kann ähnlich wie DNA in komplexe Strukturen falten. Die Grundlage für alle diese Strukturen höherer Ordnung ist die Fähigkeit der Basen, untereinander Wasserstoffbrücken und Stapelwechselwirkung auszubilden.

Wasserstoffbrücken verbinden nicht-benachbarte Basen. Der Energiegewinn durch Wasserstoffbrücken-Paarung ist relativ gering, da ähnlich gute Wasserstoffbrücken mit der umgebenden Stapelwechselwirkung benachbarte Basen oder Basenpaare verbinden.

Ein RNA-Molekül besteht aus einer Kette von Nukleotiden. Im Unterschied zur DNA liegt die RNA zumeist als Einzelstrang vor, d.h. die RNA besteht nicht aus zwei komplementären Nukleotidketten, die über Wasserstoffbrückenbindungen zwischen komplementären Basen miteinander verbunden sind.

Die räumliche Anordnung des kettenförmigen Moleküls nennt man "Faltung".
Daraus ergibt sich die Möglichkeit, dass sich Komplementäre des gleichen RNA-Strangs zusammenlagern können, und sich so verschiedene Sekundärstrukturen ergeben können.

Die Primärstruktur der RNA beschreibt die Abfolge der Nucleotide in einem RNA-Molekül. Die Darstellung erfolgt als String über dem Alphabet $\sum_{RNA} = \{A, C, G, U\}$.

"RNA ist meistens einzelsträngig. Damit ist die einfachste Sekundärstruktur ein sog. Hairpin (Haarnadel), der aus einer Helix (Helix: ist ein doppelsträngiger komplementärer Bereich desselben RNA-Strangs) und einem Hairpin-Loop (Haarnadel-Schleife) besteht.

Die Größe des Hairpin-Loops muss ausreichend sein, um den Helix-Durchmesser zu überbrücken. Da RNA nur in Ausnahmefällen komplett selbstkomplementär ist, ist der helikale Strukturteil üblicherweise durch nicht-komplementäre Bereiche auf einer oder beiden Seiten der Helix unterbrochen. All diese nicht-komplementären Bereiche werden Loops (Schleifen) genannt." [15]

Hairpin-Loop: Ein Hairpin besteht aus einer Helix und einem Loop, der die Helix überbrückt. Hairpin-Loops bilden sich relativ schnell, aber wachsen mit steigender Loop-Größe.
Die thermodynamische Stabilität von Hairpin-Loops hängt ab von der Loop-Größe, der Loop-Sequenz und vom Typ des den Loop schließenden Basenpaares.

Bulge-Loops: Bulge-Loops besitzen ungepaarte Basen in einem Strang einer doppelsträngigen Region, während der andere Strang durchgehend basengepaart ist. Ein Bulge-Loop kann eine einzige Base groß sein, die Größe ist aber prinzipiell unbegrenzt.

Internal-Loops: Internal-Loops besitzen ungepaarte Basen in beiden Strängen einer doppelsträngigen Region. Bei gleicher Zahl ungepaarter Basen in beiden Strängen werden sie als symmetrisch bezeichnet. Ein symmetrischer Internal-Loop aus zwei Basen wird ,,Mismatch" genannt. Die thermodynamische Stabilität eines internen Loops hängt von der Zahl und der Art der ungepaarten Basen und von der Art der Nachbarbasenpaare ab.

Multi-Loop: Multi-Loops verbinden mehr als zwei Helices. Zwischen den Helices können ungepaarte Basen liegen.

Einfache Sekundärstrukturen ohne Multi-Loops werden manchmal als **Stem-Loop-Strukturen** bezeichnet.

2.2 RNA-Sekundärstruktur

Anders als die DNA, die als Doppelhelix vorliegt, besteht die RNA aus einem einzelnen Strang. Basen dieses Stranges können mit anderen Basen desselben Stranges eine Verbindung eingehen. So entsteht die Sekundärstruktur (x).

Definition: Sei $r_1 r_2 ... r_n$, mit $r_i \in \sum_{RNA} (1 \leq i \leq n)$ die Primärstruktur einer RNA in ihrer Stringdarstellung. Eine Sekundärstruktur der RNA ist dann eine Menge von Paaren von Indizes aus dem Bereich 1 bis n, $x \subseteq \{(i,j) \mid 1 \leq i < j \leq n\}$, mit der Bedeutung, dass die Base r_i mit der Base r_j durch Wasserstoffbrückenbindungen verbunden ist.

- $\forall k \in \{1, ..., n\} : r_k$ ist höchstens mit einer anderen Base gepaart.
- $\forall (i,j) : (r_i, r_j) \in \{(A,U), (U,A), (C,G), (G,C)\}$ Watson-Crick-Paare oder $(r_i, r_j) \in \{(U,G), (G,U)\}$ Wobble Paar. Man spricht von gültigen Basenpaaren.

- $\forall(i,j) : j - i \geq 3$, mindestens 3 Basen in Hairpin-Loop.

2.2.1 Komponenten von RNA-Sekundärstruktur

Die Komponenten einer RNA-Sekundärstruktur sind Stem, Hairpin-Loop, Bulge-Loop, Internal-Loop, und Multi-Loop. Wir definieren kurz diese Komponenten wie folgt:

- Stems: es gilt nach (i,j) auch $(i+1, j-1)$. D.h. mehrere aufeinander folgende Basen.
- Hairpin-Loop: falls von (i,j) kein Basenpaar in der Sekundärstruktur erreichbar ist, so dass durch das Basenpaar (i,j) eine Schleife entsteht, die kein weiteres Basenpaar enthält.
- Bulge-Loop: nach (i,j) kommt $(i+k+1, j-1)$, bzw. $(i+1, j-k-1)$. D.h. auf einer Seite entsteht eine Ausbuchtung mit k Basen.
- Internal-Loop: nach (i,j) folgt $(i+k_1+1, j-k_2-1)$. D.h. es entsteht eine Ausbuchtung auf beiden Seiten.
- Multi-Loop: falls von (i,j) aus mehr als ein weiteres Basenpaar erreichbar ist. Diese Strukturen können viele Komponenten besitzen.

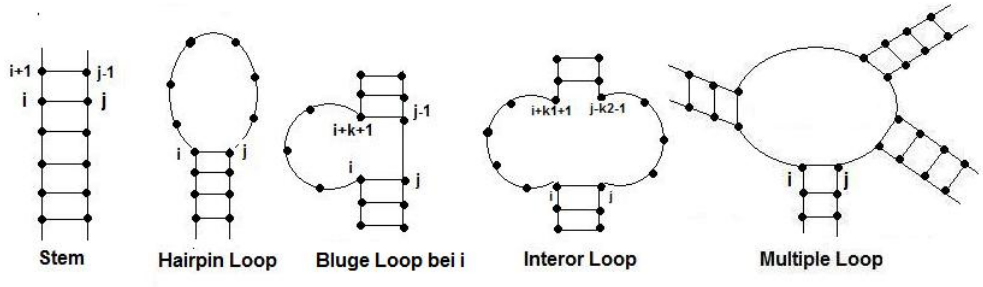


Abbildung 2: *Der Unterschied zwischen den Loops.*

Definition: Der Grad eines Loops wird durch 1 plus die Zahl der schließenden Basenpaare gegeben, die sich direkt neben den ungepaarten Regionen befinden.
Ein Loop des Grades 1 wird Hairpin-Loop genannt. Ein Loop des Grades 2

wird Bulge genannt, wenn das Schlusspaar des Loops und des inneren Basenpaars angrenzend sind; sonst wird ein Loop des Grads 2 Internal-Loop genannt. Ein Loop eines Grades größer als 2 wird Multi-Loop genannt.

2.2.2 Darstellung von Sekundärstrukturen

Die Anzahl der Sekundärstrukturen hängt von der Länge und der Komposition der Sequenz ab.
Sei s eine RNA Sequenz der Länge n und $s_i \in \sum_{RNA}$.
RNA Strukturen werden als Zeichenreihen über $\{(, .,)\}$ repräsentiert, wobei ein Klammerpaar ein Basenpaar symbolisiert wenn (p, q) ein Basenpaar ist und $p < q$, dann ist $S_p^* =$ "(" und $S_q^* =$ ")".
Der Punkt symbolisiert ein ungepaartes Nukleotid i .
Für den Satz von Basenpaaren in unserem Fall gelten folgende Einschränkungen [15]:

1. $j - i \geq 3$ ist die minimale Hairpin-Loop Größe und die Reihenfolge von zwei Basenpaaren $i : j$ und $k : l$ ist beschränkt, durch
2. $i = k$ *und* $j = l$ *oder*
3. $i < j < k < l$ *oder*
4. $i < k < l < j$.

$\mathcal{F}(n)$: die Menge aller Sekundärstruktur-Faltungen ist definiert durch:

$$\mathcal{F}(n) = \underset{s}{\cup} \mathcal{F}(s) \ ; \ s \in \{A, C, G, U\}^n$$

wobei $\mathcal{F}(s)$ ein spezieller Sekundärstruktur-Faltungsraum für eine RNA-Sequenz s ist.

2.3 Bekannte Ergebnisse zur Kombinatorik von Strukturen

In der Analyse der RNA Sekundärstruktur ist es häufig erforderlich, suboptimale Strukturen zu betrachten.

Die Zahl von suboptimalen Strukturen ist exponentiell groß und sogar die Zahl von naheoptimalen Strukturen wächst exponentiell mit der Länge der Sequenz.

Strukturen mit bestimmten Eigenschaften [5]
Eine Sekundärstruktur aus $n+1$ Basen kann aus einer Struktur aus n Basen erhalten werden, indem man entweder ein freies Ende am rechten Ende hinzufügt oder eine Substruktur einfügt $(1, k+2)$.
Im zweiten Fall wird durch dieses Paar eine beliebige Substruktur aus k Basen eingeschlossen. Der restliche Teil der Länge $n-k-1$ ist auch eine beliebige gültige Sekundärstruktur:

$$\begin{aligned} S^*_{n+1} &= S^*_n + \sum_{k=m}^{n-1} S^*_k * S^*_{n-k-1}, n \geq m+1, \\ S^*_0 &= S^*_1 = ... = S^*_{m+1} = 1. \end{aligned} \tag{1}$$

m sei die minimale Anzahl für ungepaarte Basen in Hairpin-Loops.

$J^*_n(b)$ sei die Anzahl von Strukturen aus n Knoten mit genau b Komponenten. Die Ableitung der rekursiven Beziehungen wird analog zu Gleichung (1) geführt:

$$\begin{aligned} &J^*_{n+1}(b) = J^*_n(b) + \sum_{k=m}^{n-1} S^*_k * J^*_{n-k-1}(b-1), \qquad b > 0, n \geq m+1, \\ &J^*_n(b) = 0,\ b > 0,\ n \leq m+1,\ J^*_n(0) = 1,\ n \geq 0. \end{aligned}$$

Das Hinzufügen einer ungepaarten Base zu einer Struktur aus n Basen ändert die Zahl der Komponenten nicht. Das Einführen einer zusätzlichen Klammer macht den eingeklammerten Teil der Länge k zu einer einzelnen Komponente, wobei der Rest der Sequenz nicht betroffen ist.

$H^*_n(b)$ sei die Anzahl von Strukturen mit genau b Basenpaaren aus n Knoten. Die Rekursion:

$$\begin{aligned} &H^*_{n+1}(b) = H^*_n(b) + \sum_{k=m}^{n-1} \sum_{l=0}^{b-1} H^*_k(l) * H^*_{n-k-1}(b-l-1), \quad b > 0, n \geq m+1, \\ &H^*_n(b) = 0,\ b > 0,\ n \leq m+1,\ H^*_n(0) = 1,\ n \geq 0. \end{aligned}$$

folgt auch unmittelbar. Man muss bemerken, dass eine zusätzliche Summe über die Zahl von ungepaarten Basen im zuletzt eingeklammerten Teil der Struktur eingeführt werden muss. Diese Rekursion ist auch in Waterman [17] betrachtet worden und führt zu dem geschlossenen Ausdruck

$$H_n^*(b) = \frac{1}{b}\binom{n-b}{b+1}\binom{n-b-1}{b-1}.$$

$A_n^*(b)$ sei die Anzahl der Strukturen mit genau b Hairpin-Loops. Da die Anzahl der Hairpins unverändert bleibt, wenn eine Substruktur um ein Basenpaar erweitert wird, erhalten wir:

$$A_{n+1}^*(b) = A_n^*(b) + \sum_{k=m}^{n-1}\left[\sum_{l=1}^{b} A_k^*(l) * A_{n-k-1}^*(b-1) + A_{n-k-1}^*(b-1)\right],\ n \geq m+1,$$
$$A_n^*(b) = \delta_{0,b},\ n \leq m+1$$

wobei $\delta_{0,b}$ Kronecker's δ, $\delta_{0,0} = 1$ *und* $\delta_{0,b} = 0,\ b \neq 0$.

2.4 Zahl möglicher Strukturen

Die Anzahl der Strukturen ist definiert durch die Menge aller Faltungen:

$$S^*(n) = |\mathcal{F}(n)|$$

In [15] wird dafür eine Formel hergeleitet.
Annahme: Jede Base kann mit jeder anderen ein Paar bilden.
Die minimale Hairpin-Loop Größe sei ein Nukleotid.
$S^*(n)$ sei die Zahl der Sekundärstrukturen für eine Sequenz der Länge n.
$S^*(0) = 0$; ohne Nukleotid keine Struktur.
$S^*(1) = 1$; mit einem Nukleotid kann man nur eine Struktur bilden.
$S^*(2) = 1$; mit zwei Nukleotiden kann ebenfalls keine Struktur gebildet werden.

1. Wenn das Nukleotid n nicht gepaart ist: $S_1^*(n) = S^*(n-1)$.
2. Wenn n gepaart mit 1 ist: $S_2^*(n) = S^*(n-2)$.

3. Wenn n gepaart mit k ist; $2 \leq k \leq n-1$; (jeder der Teilbereiche $[2,k]$ bzw. $[(k+1),n-1]$ kann noch $S^*(k-1)$ bzw. $S^*(n-k-1)$ Strukturen bilden): $S_3^*(n) = \sum_{k=2}^{n-1} S^*(k-1) * S^*(n-k-1) \Rightarrow$

$$S^*(n) = S_1^*(n) + S_2^*(n) + S_3^*(n)$$

$$S^*(n) = S^*(n-1) + S^*(n-2) + \sum_{k=2}^{n-1} S^*(k-1) * S^*(n-k-1)$$

Die Tabelle zeigt die Zahl der Strukturen $S^*(n)$ für eine Sequenz der Länge n:

n	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
$S^*(n)$	1	1	2	4	8	17	37	82	185	423	978	2283	5373	12735	30372

	16	17	18	19	20	40	80	100
	72832	175502	424748	1032004	2516347	2.1E14	4.1E30	6.8E38

Tabelle 1 : *Anzahl der Sekundärstrukturen für eine Sequenz der Länge n*

Für große Nukleotidzahlen $n \to \infty$ lässt sich zeigen [17], dass

$$S^*(n) \sim \sqrt{\frac{15+7\sqrt{5}}{8\pi}} n^{-3/2} \left(\frac{3+\sqrt{5}}{2}\right)^n .$$

2.5 Strukturgrammatik

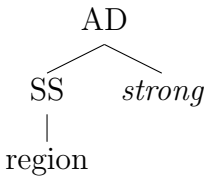
Knotenmakierung eines Strukturbaums

RNA-Sekundärstrukturen können als Strings, Basenpaar-Listen, Grafiken, Bäume und in vielen anderen Formen dargestellt werden.

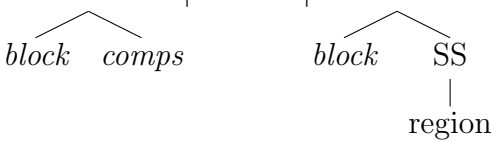
Die verschiedenen strukturellen Komponenten in der RNA, die wir als Knotenmarkierung eines Strukturbaums benutzen, sind: Singlestranded Regionen (SS), Hairpin-Loops (HL), Stacking Regionen (SR), Bulge-Loops (Bulge-Loop links (BL) und Bulge-Loop rechts (BR)), und Internal-Loops (IL).
Darüber hinaus haben wir Listen von benachbarten Strukturen, wie die Komponenten der äußeren Loops (AD). Multi-Loops (ML) umfassen ein schließendes Basenpaar und eine Liste der benachbarten Strukturelemente (AD) innen.
Wir benötigen zusätzlich ein Label für eine leere Liste von benachbarten Komponenten (E).

Baum-Grammatik ($\mathcal{B}$) für RNA Strukturen, aus [12] entnommen:

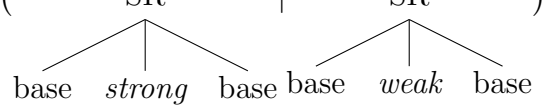
struct $\longrightarrow$ *comps* | E
block $\longrightarrow$ *strong* |



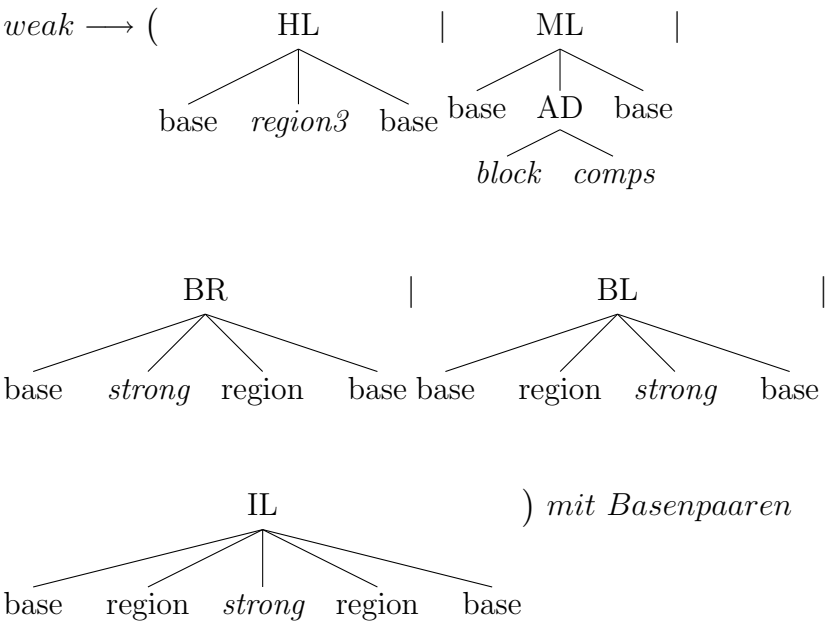
comps $\longrightarrow$



strong $\longrightarrow$ (



) *mit Basenpaaren*



*region*3 ⟶ region mit minsize 3.

Abbildung 3 : *Baum-Grammatik ($\mathcal{B}$) für RNA Strukturen*

In Anlehnung an die Baum-Grammatik ($\mathcal{B}$) für RNA Sekundärstrukturen, können wir jetzt die folgende Struktur für eine gegebene Sequenz als Baum Struktur darstellen (Abbildung 5).

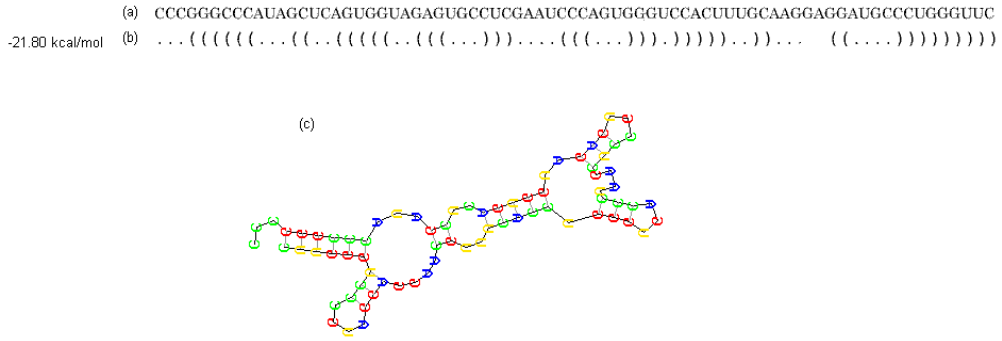


Abbildung 4 : *Darstellung einer RNA-Sekundärstruktur. Gezeigt wird die übliche Darstellung einer RNA-Sekundärstruktur mit der oben angegebenen Sequenz.*
(a)*: Die Primärstruktur der RNA.* **(b)***: Vienna-Dot-Bracket Notation; ein Punkt symbolisiert ein ungepaartes Nukleotid; eine öffnende bzw. schließende Klammer ein Basenpaar.* **(c)***: Eine Darstellung, in der die Loops als gleichwinklige Polygone gezeichnet sind.*

Später werden wir die RNA-Sequenz mit ihrer vorgestellten Sekundärstruktur als Vienna-Dot-Bracket Notation (Abbildung 4) und ihrem Struktur-Baum in Abbildung 5 verwenden, um den Shapes-Baum für den fünf Shape-Typen aus ihrem Struktur-Baum zu bekommen.

Wir stellen jetzt diese Struktur für die gegebene Sequenz als Baum Struktur dar:

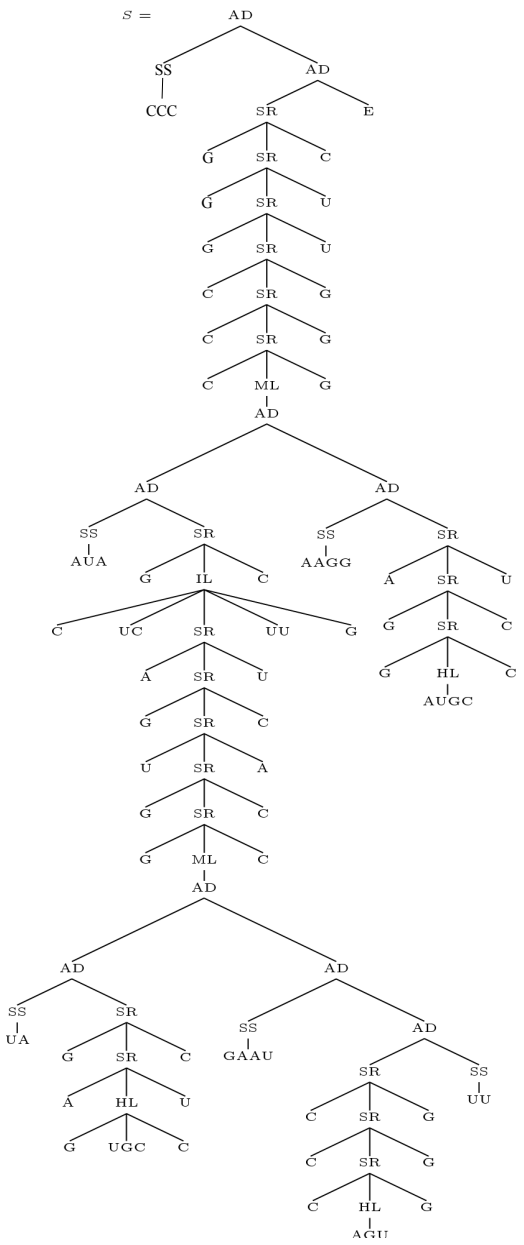


Abbildung 5*: Strukturbaum* (S) *für die RNA-Sequenz mit ihrer Sekundärstruktur als Vienna-Dot-Bracket Notation (Abbildung* 4*)*

2.6 Abstrakte Shapes, Formenräume

Der abstrakte Shape-Ansatz wurde von Giegerich et al. [4] beschrieben. In dem genannten Artikel wird jede Klasse ähnlicher Sekundärstrukturen von einer Shape dargestellt. Dabei werden die nativen Strukturen durch den besten Shape repräsentiert.

Der Abstract Shapes Ansatz bildet eine Abstraktion $\pi(x)$ von jeder Sekundärstruktur x und fasst Sekundärstrukturen mit identischer Abstraktion zu sog. Shape-Klassen oder kurz Shapes zusammen.

Eine RNA Shape ist eine abstrakte Darstellung einer RNA Sekundärstruktur. Sie ist motiviert durch die Klammerdarstellung, bekannt durch die RNA Vienna-Dot-Backet Notation.

Die Notation einer Shape sind Zeichenreihen über $\{[, \_,]\}$. Für einen Abschnitt mit öffnenden Basenpaarungen wird eine eckige, öffnende Klammer "[" benutzt, konsequenter Weise ist die eckige, schließende Klammer "]" das Pendant für einen Abschnitt schließender Basenpaare. Der Unterstrich $''\_''$ symbolisiert einen Bereich von ungepaarten Basen.

Eine Shape fasst in dem von einer RNA Sequenz aufgespannten Raum aller Sekundärstrukturen ein oder meist mehrere Sekundärstrukturen durch Abstraktion zusammen.
Wir interessieren uns für die Zahl der Shapes, solche Analysen müssen effizient berechenbar sein, trotz der Größe des exponentiellen Faltungsraums.
Hier folgt eine Zusammenfassung der grundlegenden Definitionen von abstrakten Shapes:

- Sei $\mathcal{F}$ eine Baum-Struktur und $\mathcal{P}$ ein Baum-Shape.
  RNA Shape-Abstraktion ist eine Abbildung von einer Struktur zu einem Shape: $\pi : \mathcal{F} \longrightarrow \mathcal{P}$

  Sei $\mathcal{P}$ der Faltungsraum der als Bäume modellierten Shapes. Eine Shape-Abstraktion ist dann ein Baum-Homomorphismus π von $\mathcal{F}$ nach $\mathcal{P}$, der Nachbarschaften und Verschachtelungen beibehält. Zwei Sekundärstrukturen x und y des Faltungsraumes $\mathcal{F}$ geben dieselbe Shape p an, wenn $\pi(x) = \pi(y) = p$.

- Der abstrakte Shape-Raum der Sequenz s ist $\mathcal{P}(s)_p i = \{\pi(x)|x \in \mathcal{F}(s)\}$. Die Klasse der p-shaped Strukturen in $\mathcal{F}(s)$ ist: $\{x|x \in \mathcal{F}(s), \pi(x) = p\}$.

$\mathcal{P}_i(n)$: Die Menge aller Shapes ist definiert durch:

$$\mathcal{P}_i(n) = \{\pi_i(x) \mid x \in \mathcal{F}(s), s \in \{A, C, G, U\}^*, |s| \leq n \ , i \in \{1...5\}\}$$

$\mathcal{P}_i(n)$ läßt sich in 2 Teilmengen aufteilen, offene und geschlossene Shapes.
Eine geschlossene Shape hat die Form $[A]$ und A ist eine offene Shape.
Eine offene Shape ist eine Shape, die aus mindestens 2 geschlossenen Shapes besteht.

RNA Shapes können mehr oder weniger abstrakt sein, je nachdem welche Details als relevant erachtet werden.
Wir möchten Typen von Shapes verwenden. Im Allgemeinen behalten Shape Abstraktionen das Verschachteln und Nebeneinander von Helices.
Wir können beschließen, Bulge- und Internal-Loops zu verwerfen, was zu verschiedenen Abstraktionen führt [4]:

Typ 1: Unterscheidet alle Loops und ungepaarte Basen. Entsprechend wird jeder Helixbereich von einem einzigen Paar öffnender und schließender eckigen Klammern dargestellt. Ungepaarte Regionen werden durch einen einzigen Unterstrich dargestellt. Daher tragen alle strukturellen Komponenten zu dieser Shaperepräsentation bei, wobei Verschachtelung und Nachbarschaft von Helices beibehalten werden. Dieser Shapetyp abstrahiert allein von Loop- und Helix-Längen.

Typ 2: Unterscheidet alle Loops und alle ungepaarten Basen in Bulge- und Internal-Loops. Folglich werden alle helikalen Regionen durch ein Paar von öffnenden und schließenden eckigen Klammern dargestellt. Weiterhin werden einzelne Basen, die einen Bulge-Loop unterbrechen und ungepaarte Basen in Internal-Loops durch einen einzelnen Unterstrich repräsentiert. In der Shaperepräsentation bleiben also Verschachtelung und Nachbarschaft von Helices erhalten, aber im Unterschied zu Typ 1 tragen nicht alle strukturellen Komponenten zur Repräsentation bei, da Unterstriche für ungepaarte Regionen

in External- und Multi-Loops weggelassen werden.

Typ 3: Unterscheidet alle Loops, aber keine ungepaarten Basen. Der Shapetyp 3 behält Verschachtelung und Nachbarschaft von Helices bei, da alle Regionen durch ein Paar öffnender und schließender eckiger Klammern repräsentiert werden. Im Unterschied zu den beiden ersten Shapetypen werden allerdings keine ungepaarten Regionen berücksichtigt.

Typ 4: Unterscheidet alle Loops außer Bulge-Loops. Im Vergleich zu Shapetyp 3 ist der einzige Unterschied, dass verschachtelte Helices, die nur durch einen einzelnen Bulge unterbrochen werden, kombiniert und nur durch ein einzelnes Paar eckiger Klammern dargestellt werden.

Typ 5: Die stärkste Abstraktion ignoriert Bulge- und Internal-Loops. In dieser Shaperepräsentation berücksichtigen wir keine Helix-Unterbrechungen (durch einzelne Bulge- oder Internal-Loops). Das bedeutet, dass (unterbrochene) Helix-Regionen stets nur durch ein Paar öffnender und schließender Klammern dargestellt werden, da verschachtelte Helices nun immer kombiniert werden.

In dieser Arbeit benötigen wir später die Grammatik für die Shapes-Typen $i, i \in \{1...5\}$, um die Formel für die Anzahl der Shapes mit n Klammerpaaren $S_i(n), i \in \{1...5\}$ abzuleiten.
Dafür definieren wir die Abstraktions-Abbildung $\pi_i\big(i \in \{1...5\}\big)$ von Strukturen zu Shapes, um dann den Shapestring durch den Homomorphismus $\nu_i\big(i \in \{1...5\}\big)$ zu definieren.

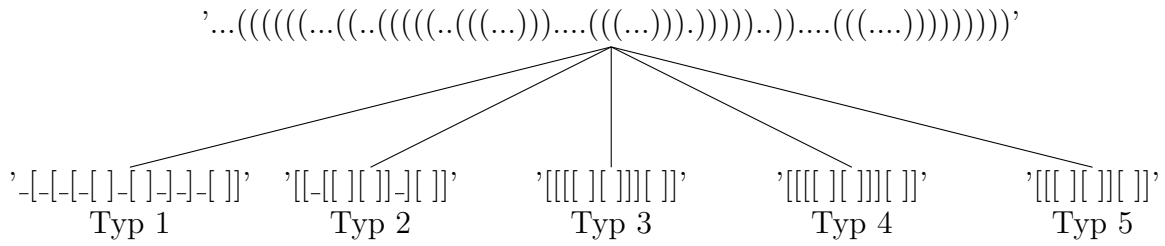


Abbildung 6 : *Die Unterschiede zwischen den Shapes-Typen $i; i \in \{1..5\}$ für die RNA Sequenz mit ihrer vorgestellten Sekundärstruktur als Vienna-Dot-Bracket Notation (Abbildung 4)*

Die folgende Abbildung beschreibt die Unterschiede zwischen den Shape-Typen [14]:

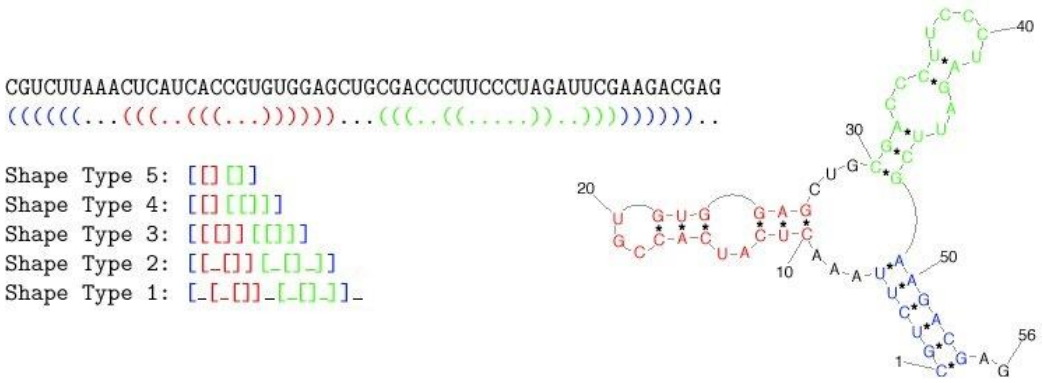


Abbildung 7 : *Eine Beispiel-Sekundärstruktur mit ihren dazugehörigen Shape-Strings der fünf verschiedenen Shape-Typen. Oben links ist die Primärsequenz und die abgebildete Sekundärstruktur in Vienna-Dot-Bracket Notation angegeben. Eckige Klammern stehen für Helices (oder Teile von Helices) und Unterstriche symbolisieren ungepaarte Regionen. Die exakte Bedeutung hängt vom gewählten Abstraktionstyp ab. Typ* 5 *abstrahiert von sämtlichen Helixunterbrechungen und ignoriert alle ungepaarten Bereiche. Die Struktur wird zu einer einzigen helikalen Region, die in zwei weitere Helices verzweigt* [[ ][ ]]. *Typ* 4 *beachtet interne Loops, jedoch keine Bulges, was zu* [[ ][[ ]]] *führt. Typ* 3 *führt dagegen alle Unterbrechungen von Helices an:* [[[ ]][[ ]]]. *Typ* 2 *erweitert Typ* 3*, indem hier zwischen linken und rechten Bulges und Internal-Loops unterschieden wird* [[_[ ]][_[ ]_]]. *In Typ* 1 *werden dann alle Unterbrechungen von ungepaarten Regionen oder Helices aufgeführt:* [_[_[_]]_[_[_]_]]_

2.7 Anwendung von abstrakten Shapes

Abstrakte Shapes werden zur Vorhersage von RNA-Sekundärstrukturen verwendet. Dabei werden ähnliche Strukturen zu Shape-Klassen zusammengefasst. Das Programm RNAshapes [13] stellt eine Reihe darauf basierender Analysen bereit:

1. Berechnung der Shape-Repräsentanten: Herbei werden für einen vorgegebenen Energiebereich alle Sekundärstrukturen berechnet und zu Shape-Klassen zusammengefasst. Dabei wird dann jeweils die beste

Struktur (d.h. mit der geringsten freien Energie) jeder Klasse ermittelt. Gegenüber der Vorhersage mit suboptimalen Strukturen hat dieser Ansatz den Vorteil, dass die Anzahl der Shape-Klassen viel kleiner ist als der Strukturraum.

2. Shape-Wahrscheinlichkeiten: Die Wahrscheinlichkeit einer Shape ist die Summe der Wahrscheinlichkeiten aller Strukturen, die in diese Shape fallen.

3. Konsensus-Shapes: Dieser Ansatz stellt eine Möglichkeit bereit, die Konsensus-Strukturen für mehrere Sequenzen zu berechnen. Dafür wird jeweils für jede Sequenz der Shape-Raum aufgespannt. Als Konsensus wird dann diejenige Shape genommen, die für alle Sequenzen gleich ist.

2.8 Vorliegende Ergebnisse zur Kombinatorik von abstrakten Shapes

In diesem Abschnitt betrachten wir die vorliegenden Ergebnisse, die von Clote et al. [8] und Nebel und Scheid [9] erarbeitet wurden.
Wir beginnen mit den Ergebnissen von Clote et al.:
Sein Ergebnis für die asymptotische Anzahl der Shape-Typ 5 mit n Klammerpaaren mit der Anwendung des Bender-Meir-Moon Theorems ist:

$$[z^{2n}]S(z) = \sqrt{\frac{6}{\pi}} \cdot (2n)^{-3/2} \cdot \sqrt{3}^{2n}$$

Die asymptotische Anzahl der Shapes s_n kompatibler Sekundärstrukturen der Länge n mit der Anwendung des Fljolet-Odlyzko-Theorem (wir wenden dieses Theorem später in unserer Analyse der asymptotischen Anzahl der Shapes an) ist:

$$s_n \sim 2.44251 \cdot n^{-3/2} \cdot 1.32218^n$$

Im folgenden beschreiben wir die kombinatorischen Ergebnisse von Nebel und Scheid für den Shape-Typ $i\big(i \in \{1...5\}\big)$ als erzeugende Funktionen:

Erzeugende Funktion $S_1(z)$:

$$\begin{aligned}
S_1(z) &= AA(z) + z, \\
AA(z) &= CC(z) \cdot z^2 \cdot BB(z) \cdot DD(z), \\
BB(z) &= 1 + CC(z) \cdot z^2 \cdot BB(z) \cdot AA(z) + z \cdot z^2 \cdot BB(z) + z^2 \cdot BB(z) \cdot z + z \cdot z^2 \cdot BB(z) \cdot z, \\
CC(z) &= 1 + z, \\
DD(z) &= 1 + z + AA(z).
\end{aligned}$$

Erzeugende Funktion $S_2(z)$:

$$\begin{aligned}
S_2(z) &= AA(z) + 1, \\
AA(z) &= z^2 \cdot BB(z) \cdot DD(z), \\
BB(z) &= 1 + z^2 \cdot BB(z) \cdot AA(z) + z \cdot z^2 \cdot BB(z) + z^2 \cdot BB(z) \cdot z + z \cdot z^2 \cdot BB(z) \cdot z, \\
DD(z) &= 1 + AA(z).
\end{aligned}$$

Erzeugende Funktion $S_3(z)$ und $S_4(z)$:

$$\begin{aligned}
S_3(z) &= AA(z) + 1, \\
AA(z) &= z^2 \cdot BB(z) \cdot DD(z), \\
BB(z) &= 1 + z^2 \cdot BB(z) \cdot AA(z) + z^2 \cdot BB(z), \\
DD(z) &= 1 + AA(z).
\end{aligned}$$

Erzeugende Funktion $S_5(z)$:

$$\begin{aligned}
S_5(z) &= AA(z) + 1, \\
AA(z) &= z^2 \cdot BB(z) \cdot DD(z), \\
BB(z) &= 1 + z^2 \cdot BB(z) \cdot AA(z), \\
DD(z) &= 1 + AA(z).
\end{aligned}$$

Die genaue asymptotische Anzahl der n-ten Koeffizienten ($n \longrightarrow \infty$) für die fünf erzeugenden Funktionen $S_i(z), i \in \{1...5\}$ mit der Anwendung des Darboux-Theorems nach der Lösung für $S_i(z), i \in \{1...5\}$, sind:

$$\begin{aligned}
s_{1_n} &\sim 2.40591^n \cdot 0.989959 \cdot n^{-3/2} \\
s_{2_n} &\sim 2.0523^n \cdot 0.88639 \cdot n^{-3/2} \\
s_{3_n} = s_{4_n} &\sim ((-2.)^n + 2.^n) \cdot 0.797885 \cdot n^{-3/2} \\
s_{5_n} &\sim 1.73205^n (1. + (-1.)^n) \cdot 0.690988 \cdot n^{-3/2}
\end{aligned}$$

3 Kombinatorische Analysen des Shape Space

In diesem Kapitel werden wir die Rekurrenzformel für die Anzahl der Shapes der Länge $\leq n$ und die Rekurrenzformel für die Anzahl der Shapes zu Sequenzen der Länge $\leq n$ mit ihrem asymptotischen Wert aufstellen.
Wir werden aber im vorhinein zuerst kurz das Folgende erklären:

3.1 Untersuchte Fragestellungen

Zuerst geben wir eine kurze Übersicht über ein paar allgemeine Definitionen:

- Anzahl der Shapes $S_i(n)\big(i \in \{1...5\}\big)$ mit n Klammerpaaren (KP):

  $$S_i(n) = \big|\big\{\pi_i(x)|x \in \mathcal{F}(s), s \in \{A,C,G,U\}^*, \pi_i(x)\text{hat } n\, KP, i \in \{1...5\}\big\}\big|$$

- Anzahl der Shapes $L_i(n)\big(i \in \{1...5\}\big)$ der Länge $\leq n$:

  $$L_i(n) = \big|\big\{\pi_i(x)|x \in \mathcal{F}(s), s \in \{A,C,G,U\}^*, |\pi_i(x)| \leq n, i \in \{1...5\}\big\}\big|$$

- Die Menge aller Shapes $\mathcal{P}_i(n), \big(i \in \{1...5\}\big)$ zu Sequenzen der Länge $\leq n$ ist definiert durch:

  $$\mathcal{P}_i(n) = \big\{\pi_i(x)|x \in \mathcal{F}(s), s \in \{A,C,G,U\}^*, |s| \leq n, i \in \{1...5\}\big\} \Rightarrow$$

  Anzahl der Shapes $R_i(n)$ zu Sequenzen der Länge $\leq n$:

  $$R_i(n) = |\mathcal{P}_i(n)|$$

Im Rest von Kapitel 3 werden wir die Rekurrenzformel für $S_i(n)$ und $L_i(n)$ $(i = 3)$, d.h. für den Shape-Typ 3 bestimmen. Weiterhin werden wir in diesem Kapitel die Rekurrenzformel für die Anzahl der Shapes $R_5(n)$ zu Sequenzen der Länge $\leq n$ mit ihrer asymptotischen Anzahl berechnen, da der Shape-Typ 5 eine Schlüsselrolle spielt.

Der hier gewählte Weg entspricht dem etwa zeitgleich in [8] verfolgten. Allerdings sind unsere Ergebnisse nicht identisch, weil die in [8] verwendeten Shapes nicht ganz mit den tatsächlich in RNAshapes implementierten Shape-Abstraktionen übereinstimmen.
Es wird daher – im Hinblick auf die Analyse der weiteren Shape-Typen ein ,,sicherer" Ansatz gesucht, der direkt die Shape-Abstraktionen π_i benutzt.

Dieser Weg wird in Kapitel 4 für den Shape-Typ 5 entwickelt und in Kapitel 5 auf die weiteren Shape-Typen übertragen. Auf diese Weise wird auch ein Teil der Ergebnisse aus Kapitel 3 auf einem zweiten Weg bestätigt.

Im Detail wird in den Kapiteln $4, 5$ die Rekurrenzformel für die Anzahl der Shapes $S_i(n)\big(i \in \{1...5\}\big)$ mit n Klammerpaaren durch die Shapes-Grammatik herleiten. Wir werden zeigen, dass die mit zwei Methoden in den Kapiteln 3 und 5 erhaltenen $S_3(n)$ das gleiche Ergebnis liefern.

In Kapitel 6 wenden wir uns mit empirischen Mitteln der Frage zu, auf die es nach wie vor keine theoretische Antwort gibt. Eine Formel für den Erwartungswert $E(R(n))$ gibt es nicht. Wir gehen von der Annahme aus, dass sie von ähnlicher Form ist wie $R(n)$, und bestimmen empirisch die Parameter.

Da der asymptotische Wert für die Anzahl der Shape-Typen i ($i \in \{1...5\}$) aller Sequenzen der Länge n die allgemeine Form $f(n) = a^n \cdot b \cdot n^{-3/2}$ hat und diese Formel nicht bekannt für die Erwartungswertberechnung ist, ermitteln wir die Parameter der Formel durch die empirische Untersuchung mit Zufallssequenzen und dem Programm RNAshapes [13].

3.2 Rekurrenzen für die Anzahl der Shapes der Länge $\leq n$

Wir müssen zuerst die Anzahl der Shapes mit n Klammerpaaren bestimmen. Dafür definieren wir eine Funktion S_3.
Sei $S_3(n)$ die Anzahl der Shapes mit n Klammerpaaren.

$$\begin{array}{lll} S_3(0) = 1 & ; & \text{"\_"} \\ S_3(1) = 1 & ; & [\,] \\ S_3(2) = 2 & ; & [\,][\,],\ \big[[\,]\big] \end{array}$$

$$\textit{Behauptung}: \quad S_3(n) = \sum_{i=0}^{n-1} S_3(i) * S_3(n-1-i)$$

Beweis:
Für 0 Klammerpaare gibt es "_" $\Rightarrow S_3(0) = 1$.

Für 1 Klammerpaar gibt es nur eine [ ] $\Rightarrow S_3(1) = 1$.
Für 2 Klammerpaare gibt es nur zwei [ ][ ],[[ ]] $\Rightarrow S_3(2) = 2$.
Für $n \geq 3$ Klammerpaare: wir können die Shape in 2 Teilshapes aufteilen. Wir teilen an der öffnenden Klammer k der letzten schließenden Klammer $2n-1$. Der linke Teil beinhaltet alle Positionen vom Start bis zur öffnenden Klammer. Der rechte Teil beinhaltet alle Positionen innerhalb des Klammerpaares. So kann man noch $n-1$ Klammerpaare auf den linken und rechten Teil verteilen.

Wir fügen ein neues Klammerpaar an Position $2n-1$ und $0 \leq k \leq 2n-1$ ein. Das Klammerpaar trennt die Sequenz in zwei Teile:

$$\overset{0}{.}\overset{.}{.}\overset{.}{.}\overset{.}{.}\overset{.}{.}\overset{k}{.}[\overset{.}{.}\overset{.}{.}\overset{.}{.}\overset{.}{.}]^{2n-1}$$

$$k = 0 \longrightarrow \quad \underset{\downarrow}{-} \quad \underbrace{[\overset{.}{.}\overset{.}{.}\overset{.}{.}\overset{.}{.}\overset{.}{.}]}_{\downarrow} \quad \longrightarrow S_3(n-1)$$
$$S_3(0) \quad S_3(n-1)$$

$$k = 2n-2 \longrightarrow \quad \underbrace{\overset{.}{.}\overset{.}{.}\overset{.}{.}\overset{.}{.}\overset{.}{.}}_{\downarrow} \quad \underbrace{[\overset{.}{.}\overset{.}{.}\overset{.}{.}\overset{.}{.}\overset{.}{.}]}_{\downarrow} \longrightarrow S_3(n-1)$$
$$S_3(n-1) \quad S_3(0)$$

$$0 < k < 2n-1 \longrightarrow \quad \underbrace{\overset{.}{.}\overset{.}{.}\overset{.}{.}\overset{.}{.}\overset{.}{.}}_{\downarrow} \quad \underbrace{[\overset{.}{.}\overset{.}{.}\overset{.}{.}\overset{.}{.}\overset{.}{.}]}_{\downarrow}$$
$$S_3(k) \quad S_3(n-1-k)$$

$$\Longrightarrow S_3(n) = \sum_{\substack{k=0, \\ k \mod 2=0}}^{2n-1} S_3(k \text{ div } 2) * S_3\big(n-1-(k \text{ div } 2)\big)$$

Ungerade k sind ausgeschlossen, weil Teilsequenzen eine ungerade Länge haben.

$$\Longrightarrow S_3(n) = \sum_{i=0}^{n-1} S_3(i) * S_3(n-1-i)$$

□

Die Tabelle zeigt die Zahl der Shapes mit n Klammerpaaren:

n	0	1	2	3	4	5	6	7	8	9	10	15	20
$S_3(n)$	1	1	2	5	14	42	132	429	1430	4862	16796	9694845	2.147483647E9

Tabelle 2 : *Anzahl der Shapes $S_3(n)$ mit n Klammerpaaren.*

Jetzt können wir die Rekurrenzen für die Anzahl der Shapes der Länge $\leq n$ bestimmen. Dafür definieren wir eine Funktion L_3.

Definition: Die Anzahl der Shapes der Länge $\leq n$ wird mit $L_3(n)$ bezeichnet.
Diese Formel berechnet, wie viele Shapes $L_3(n)$ der Länge $\leq n$ existieren:
Wir zeigen:

$$L_3(n) \quad = \begin{cases} 1 & : n = 0 \\ L_3(n-1) + S_3(n \text{ div } 2) & : n \geq 2 \;\; \textit{und} \;\; n \textit{ gerade} \\ L_3(n-1) & : n \geq 2 \;\; \textit{und} \;\; n \textit{ ungerade} \end{cases}$$

Beweis:
$n = 1 \Rightarrow L_3(n) = 1$: Mit einem Symbol gibt es nur einen Shape "_".
$n = 2 \Rightarrow L_3(n) = 2$: Mit $n = 2$ gibt es zwei Möglichkeiten ("_", [ ]).
sowie: $S_3(n \text{ div } 2) = 1([\,])$ nämlich $L_3(2) = L_3(1) + S_3(1)$

Für $n \geq 2$: Es gibt zwei Gruppen von Shapes:

1. Shapes der Länge $< n \rightarrow L_3(n-1)$
2. Shapes der Länge $= n \rightarrow S_3(n \text{ div } 2)$

Dabei gilt es: Für ein ungerades n kann kein neues Klammerpaar erzeugt werden.
Dann ergibt sich die Formel $L(n)$ für alle Shapes der Länge $\leq n$, sowie $n \geq 2$:

$$L_3(n) \quad = \begin{cases} 1 & : n = 1 \\ L_3(n-1) + S_3(n \text{ div } 2) & : n \geq 2 \;\; \textit{und} \;\; n \textit{ gerade} \\ L_3(n-1) & : n \geq 2 \;\; \textit{und} \;\; n \geq 2 \;\; \textit{und } n \textit{ ungerade} \end{cases}$$

□

Die Tabelle zeigt die Zahl der Shapes $L(n)$ für eine Shape der Länge $\leq n$:

n	1	2	3	4	5	6	7	8	10	12	14	16	18	19	20
$L_3(n)$	1	2	2	4	4	9	9	23	65	197	626	2056	6918	6918	23714

30	40	. . .
1.3402697E7	4.570790694E9	. . .

Tabelle 3 : *Anzahl der Shapes $L(n)$ für eine Shape der Länge $\leq n$.*

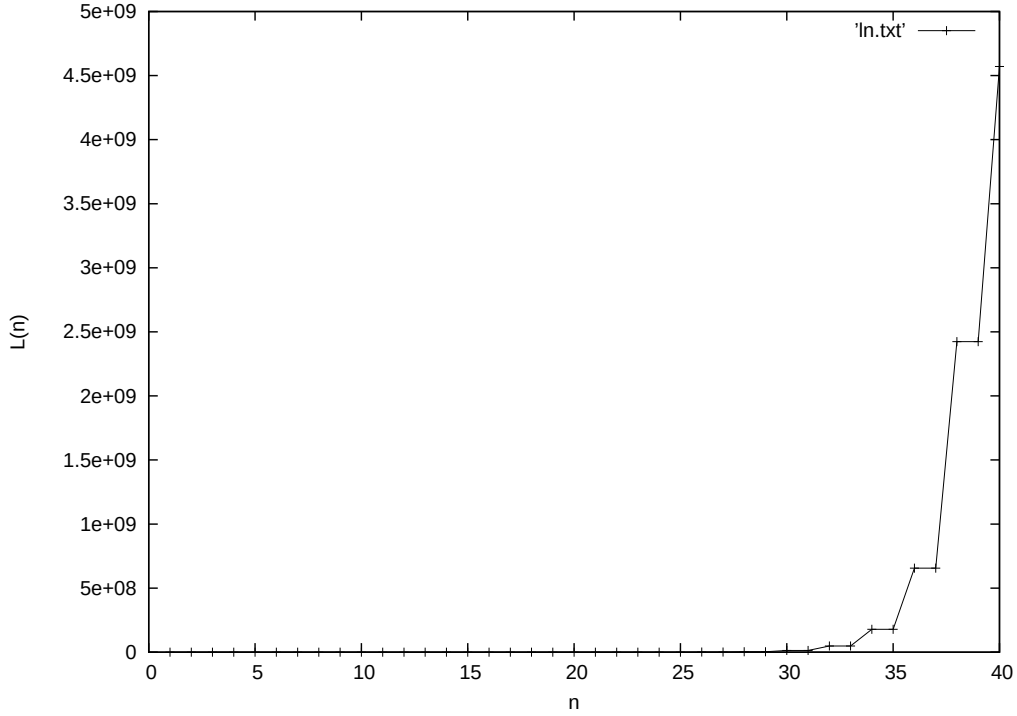


Abbildung 8**:** *Die Anzahl der Shapes $L(n)$ der Länge $\leq n$.*

3.3 Rekurrenzen für die Anzahl der Shapes zu Sequenzen der Länge $\leq n$

Die Anzahl des Shape-Typs 5 ist definiert durch:

$$R_5(n) = |\mathcal{P}_5(n)|$$

In diesem Abschnitt werden wir die Formel für die Anzahl der Shapes zu Sequenzen der Länge $\leq n$ berechnen. Dafür benötigen wir die Anzahl der neuen Shapes zu Sequenz der Länge n, wobei neue Shapes der Sequenz der Länge n bedeutet: die Shapes, die genau für eine Sequenz der Länge n gebildet werden können, nicht für kürzere Sequenzen.

Diese Formel berechnet, wieviele Shapes $R_5(n)$ für alle Sequenzen der Länge $\leq n$ existieren:

$$R_5(n) = \sum_{i=0}^{n} \sum_{k=0}^{i/2} P(i,k)$$

Alle neuen Shapes mit k Klammerpaaren für eine Sequenz der Länge n:

$$P(n,k) = C(n,k) + O(n,k)$$

Alle neuen geschlossenen Shapes mit k Klammerpaaren für Sequenzen der Länge n:

$$C(n,k) \quad = \begin{cases} 0 & : k = 0 \\ 0 & : n < 5 \\ 1 & : (n,k) = (5,1) \\ O(n-2,k-1) & : sonst \end{cases}$$

Alle neuen offenen Shapes mit k Klammerpaaren für Sequenzen der Länge n:

$$O(n,k) \quad = \begin{cases} 0 & : n = 0 \\ 1 & : (n,k) = (1,0) \\ \sum_{i=5}^{n} \sum_{l=1}^{k-1} C(i,l) * P(n-i,k-l) & : sonst \end{cases}$$

Beweis:

Wir setzen voraus, dass jede Base mit jeder anderen ein Paar bilden kann.
Die minimale Hairpin-Loop-Größe sei ein Nukleotid.
Die minimale geschlossene Shape besteht aus 5 Nukleotiden.
z.B. Struktur für 5 Nukleotide (...) $\xrightarrow{\text{gibt}}$ [ ].

Die obere Grenze $i/2$ ergibt sich, weil ein Klammerpaar zwei Buchstaben der Sequenz verbraucht. Damit sind maximal $i/2$ Klammerpaare für eine Sequenz der Länge i möglich.

Alle neuen Shapes mit k Klammerpaaren für eine Sequenz der Länge n, wobei neue Shapes bedeutet: eine Shape ist neu, wenn sie mit einem kleineren Wert von n nicht gebildet werden kann:

$$P(n,k) = C(n,k) + O(n,k)$$

$C(n,k)$: Alle neuen geschlossenen Shapes mit k Klammerpaaren für Sequenzen der Länge n der Form [A]:
Es ist offensichtlich, dass wenn wir kein Klammerpaar ($k = 0$) oder kein Nukleotid haben, dann gibt es keine neue geschlossene Shape, also $C(n,0) = 0, C(0,k) = 0$.
Für eine Sequenz, die aus weniger als 5 Nukleotiden besteht, ist offensichtlich, dass es keine geschlossene Shape gibt (wir haben gezeigt, dass die minimale geschlossene Shape aus 5 Nukleotiden besteht).
Wenn wir aber eine Sequenz der Länge 5 mit einem Klammerpaar haben, haben wir eine Shape vom Typ [ ], also $C(5,1) = 1$.
Sonst:

$$C(n,k) \widehat{=} [O(n-2,k-1)]$$

Weil das äußere Klammerpaar zwei Nukleotiden entspricht, bleiben $n-2$ Nukleotide und $k-1$ Klammerpaare für die innere Struktur A.
$\big($ A beinhaltet alle Shapes von $O(n-2,k-1)\big)$.

$$Insgesamt \Rightarrow C(n,k) \quad = \begin{cases} 0 & : k = 0 \\ 0 & : n < 5 \\ 1 & : (n,k) = (5,1) \\ O(n-2,k-1) & : sonst \end{cases}$$

$O(n,k)$: Alle neuen offenen Shapes mit k Klammerpaaren für Sequenzen der Länge n, die aus mindestens 2 geschlossenen Shapes bestehen.
Es ist klar: Wenn wir keine Nukleotide haben, gibt es keine Shape.
Mit einem Nukleotid gibt es nur eine Shape vom Typ "_" und diese hat kein

Klammerpaar, also $O(1,0) =$ "_" .
Sonst:

$$O(n,k) = \sum_{i=3}^{n} \sum_{l=1}^{k-1} C(i,l) * P(n-i, k-l)$$

Wir wollen zeigen: Jede offene Shape $o \in O(n,k)$ lässt sich eindeutig aus 2 kleineren Shapes kombinieren.
Jede Shape $o \in O$ besteht aus mindestens 2 geschlossenen Shapes.
Wir zerlegen eindeutig, indem wir die erste geschlossene Shape abtrennen.
Der Rest ist entweder aus O oder aus C, also aus P.

$$\underbrace{[...]}_{\downarrow \in C} \quad \underbrace{[...]...[...]}_{\downarrow \in P}$$

□

Das folgende Beispiel mit der folgenden Tabelle erklärt die vorherigen Ausführungen für $n = 17$: Die Anzahl der Shapes für eine Sequenz der Länge ≤ 17 ist:
$R_5(17) = \big(\underbrace{\_}_{P(1,0)}, \underbrace{[\,]}_{P(5,1)}, \underbrace{[\,][\,]}_{P(10,2)}, \underbrace{[[\,][\,]]}_{P(12,3)}, \underbrace{[\,][\,][\,]}_{P(15,3)}, \underbrace{[\,][[\,][\,]], [[\,][\,]][\,], [[\,][\,][\,]]}_{P(17,3)} \big)$

C	0	1	2	3	4	5
0	0	0	0	0	0	0
1	0	0	0	0	0	0
2	0	0	0	0	0	0
3	0	0	0	0	0	0
4	0	0	0	0	0	0
5	0	**1**	0	0	0	0
6	0	0	0	0	0	0
7	0	0	0	0	0	0
8	0	0	0	0	0	0
9	0	0	0	0	0	0
10	0	0	0	0	0	0
11	0	0	0	0	0	0
12	0	0	0	**1**	0	0
13	0	0	0	0	0	0
14	0	0	0	0	0	0
15	0	0	0	0	0	0
16	0	0	0	0	0	0
17	0	0	0	0	**1**	0

O	0	1	2	3	4	5
0	0	0	0	0	0	0
1	0	**1**	0	0	0	0
2	0	0	0	0	0	0
3	0	0	0	0	0	0
4	0	0	0	0	0	0
5	0	0	0	0	0	0
6	0	0	0	0	0	0
7	0	0	0	0	0	0
8	0	0	0	0	0	0
9	0	0	0	0	0	0
10	0	0	**1**	0	0	0
11	0	0	0	0	0	0
12	0	0	0	0	0	0
13	0	0	0	0	0	0
14	0	0	0	0	0	0
15	0	0	0	**1**	0	0
16	0	0	0	0	0	0
17	0	0	0	0	**2**	0

P	0	1	2	3	4	5
0	0	0	0	0	0	0
1	**1**	0	0	0	0	0
2	0	0	0	0	0	0
3	0	0	0	0	0	0
4	0	0	0	0	0	0
5	0	**1**	0	0	0	0
6	0	0	0	0	0	0
7	0	0	0	0	0	0
8	0	0	0	0	0	0
9	0	0	0	0	0	0
10	0	0	**1**	0	0	0
11	0	0	0	0	0	0
12	0	0	0	**1**	0	0
13	0	0	0	0	0	0
14	0	0	0	0	0	0
15	0	0	0	**1**	0	0
16	0	0	0	0	0	0
17	0	0	0	0	**3**	0

Aus der ganz rechtsliegenden Tabelle können wir nun $R(n)$ berechnen:

$$\begin{array}{lllllll} R_5(17) & = P(1,0) & +P(5,1) & +P(10,2) & +P(12,3) & +P(15,3) & +P(17,3) \\ & = 1 & +1 & +1 & +1 & +1 & +3 \qquad = 8 \end{array}$$

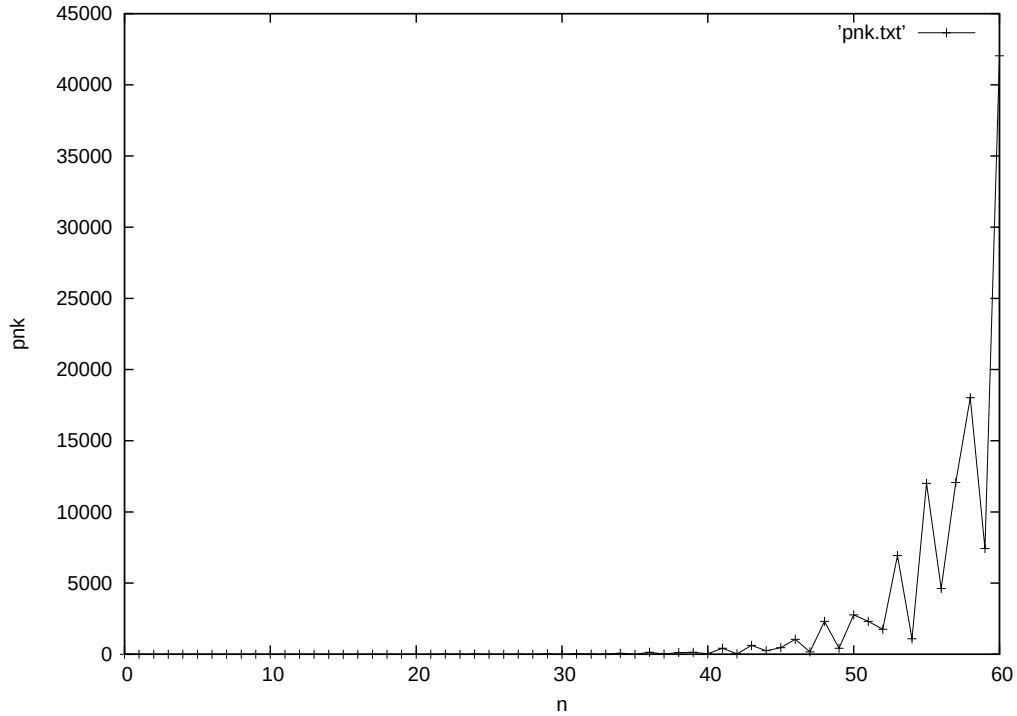


Abbildung 9**:** *Anzahl aller neuen Shapes $P(n,k)$ mit k Klammerpaaren für eine Sequenz der Länge n.*

Die Tabelle zeigt die Zahl der Shapes $R(n)$ für eine Sequenz der Länge $\leq n$:

n	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
$R_5(n)$	0	1	1	1	1	2	2	2	2	2	3	3	4	4	4	5	5	8

	18	19	20	30	40	50	60	80	100
	8	10	11	74	679	9193	117469	17218689	919530613

Tabelle 4 : *Anzahl der Shapes $R_5(n)$ für eine Sequenz der Länge $\leq n$.*

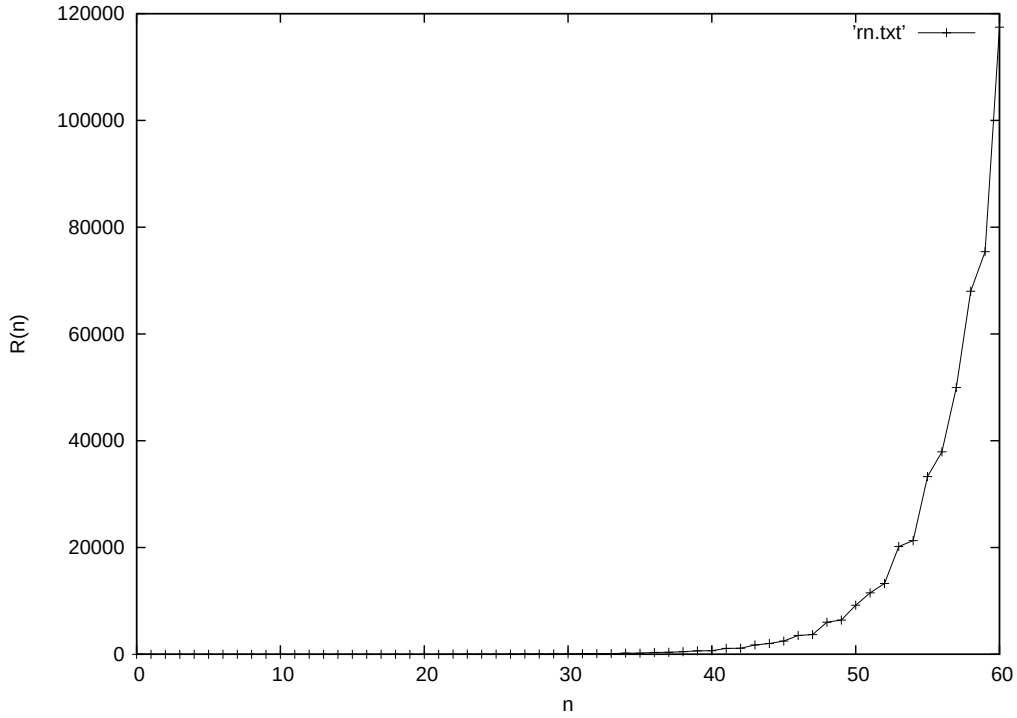


Abbildung 10 : *Anzahl der Shapes $R_5(n)$ zu Sequenzen der Länge $\leq n$.*

3.4 Asymptotische Zahl der Shapes zu Sequenzen der Länge $\leq n$

In diesem Abschnitt interessieren wir uns für die asymptotische Anzahl der Shapes für Sequenzen der Länge $\leq n$ und für die Länge n (3.3), bestimmt mit der Hilfe der erzeugenden Funktionen und eines Theorems von Flajolet-Odlyzko.

$$R_5(n) = |\mathcal{P}(n)|$$

$$R_5(n) = \sum_{i=0}^{n} \sum_{k=0}^{i/2} P_5(i,k) \tag{2}$$

$$P(n,k) = C(n,k) + O(n,k) \tag{3}$$

$$C(n,k) \;=\; \begin{cases} 0 & : k = 0 \quad (4.1) \\ 0 & : n < 5 \quad (4.2) \\ 1 & : (n,k) = (5,1) \quad (4.3) \\ O(n-2,k-1) & : sonst \quad (4.4) \end{cases} \tag{4}$$

$$O(n,k) \;=\; \begin{cases} 0 & : n = 0 \quad (5.1) \\ 1 & : (n,k) = (1,0) \quad (5.2) \\ \sum_{i=5}^{n} \sum_{l=1}^{k-1} C(i,l) * P(n-i,k-l) & : \text{sonst} \quad (5.3) \end{cases} \tag{5}$$

Aus (4.4) bei $C(n,k) = O(n-2,k-1) : sonst$
ergibt beim Einsetzen in (3) und (5.3) bei $O(n,k)$ folgendes:

$$P(n,k) = O(n-2,k-1) + O(n,k) \tag{6}$$

$$O(n,k) \;=\; \begin{cases} 0 & : n = 0 \\ 1 & : (n,k) = (1,0) \\ \sum_{i=5}^{n} \sum_{l=1}^{k-1} O(i-2,l-1) * P(n-i,k-l) & : sonst \end{cases} \tag{7}$$

Aus (6) $P(n,k) = O(n-2,k-1) + O(n,k)$ folgt:

$$P(n-i,k-l) = O(n-i-2,k-l-1) + O(n-i,k-l) \tag{8}$$

Insgesamt erhalten wir für $O(n,k)$:

$$O(n,k) \;=\; \begin{cases} 0 & : n = 0 \quad (9.1) \\ 1 & : (n,k) = (1,0) \quad (9.2) \\ \sum_{i=5}^{n} \sum_{l=1}^{k-1} O(i-2,l-1) * O(n-i-2,k-l-1) & : sonst \quad (9.3) \\ + \sum_{i=5}^{n} \sum_{l=1}^{k-1} O(i-2,l-1) * O(n-i,k-l) & \end{cases} \tag{9}$$

Aus (9.3) bekommen wir also:

$$\begin{aligned} O(n,k) &= \sum_{i=5}^{n}\sum_{l=1}^{k-1} O(i-2,l-1) * O(n-i-2,k-l-1) \\ &\quad + \sum_{i=5}^{n}\sum_{l=1}^{k-1} O(i-2,l-1) * O(n-i,k-l) \end{aligned}$$

nach Indexverschiebung:

$$O(n,k) = \underbrace{\sum_{i=0}^{n-4}\sum_{l=0}^{k-2} O(i,l) * O(n-4-i,k-2-l)}_{O_1(n,k)} + \underbrace{\sum_{i=0}^{n-2}\sum_{l=0}^{k-2} O(i,l) * O(n-2-i,k-1-l)}_{O_2(n,k)} \tag{10}$$

Wir werden $O_1(n,k), O_2(n,k)$ einzeln untersuchen, um unser Ziel (asymptotische Anzahl der Shapes zu Sequenzen der Länge n) zu berechnen.

Als erstes berechnen wir die erzeugende Funktion von $O_1(n,k)$. Für die erzeugende Funktion gilt folgende Definition:

Definition: Für eine Doppelfolge $A_{n,k}, (n,k) \in \mathbb{N}_0^2$ heißt die bivariate formale Potenzreihe $A(z,u) = \sum_{n,k} A_{n,k} z^n u^k$ die erzeugende Funktion von $A_{n,k}$.
Wir bezeichnen für eine bivariate Potenzreihe $A(z,u)$ den Koeffizienten $A_{n,k}$ mit $[z^n][u^k]A(z,u)$.

Angewendet auf $O(n,k)$ lautet unsere erzeugende Funktion:

$$\mathcal{O}(x,y) = \sum_{n,k} O(n,k) x^n y^k \Rightarrow$$

$$\begin{aligned} O_1(n,k) &= \sum_{i=0}^{n-4}\sum_{l=0}^{k-2} O(i,l) * O(n-4-i,k-2-l) \\ &= [x^{n-4}][y^{k-2}]\mathcal{O}(x,y)^2 \end{aligned}$$

$$O_1(n,k) = [x^n][y^k] x^4 y^2 \mathcal{O}(x,y)^2 \tag{11}$$

Für $O_2(n,k)$ müssen wir eine Indexänderung bei $k-2$ vornehmen. Das geschieht, indem wir rechts bei den Termen T_1 und T_2 $(k-1)$ erzeugen:

$$
\begin{aligned}
O_2(n,k) &= \sum_{i=0}^{n-2}\sum_{l=0}^{k-2} O(i,l) * O(n-2-i,k-1-l) \\
&= \underbrace{\sum_{i=0}^{n-2}\sum_{l=0}^{k-1} O(i,l) * O(n-2-i,k-1-l)}_{T_1} - \underbrace{\sum_{i=0}^{n-2} O(i,k-1) * O(n-2-i,0)}_{T_2}
\end{aligned}
$$

Auf ähnliche Art und Weise wie bei $O_1(n,k)$ erhalten wir für T_1 :

$$
\begin{aligned}
T_1 &= \sum_{i=0}^{n-2}\sum_{l=0}^{k-1} O(i,l) * O(n-2-i,k-1-l) \\
&= [x^{n-2}][y^{k-1}]\mathcal{O}(x,y)^2
\end{aligned}
$$

$$
T_1 = [x^n][y^k]x^2y\mathcal{O}(x,y)^2 \tag{12}
$$

$$
\begin{aligned}
Nun\ zu\ \ T_2 \quad : \quad T_2 &= \sum_{i=0}^{n-2} O(i,k-1).O(n-2-i,0) \\
\iff \qquad T_2 &= O(0,k-1).\underbrace{O(n-2,0)}_{=0}{}^{1} + O(1,k-1).\underbrace{O(n-3,0)}_{=0} \\
\iff \qquad &\quad +\ldots + O(n-3,k-1).\underbrace{O(1,0)}_{=1} + O(n-2,k-1).\underbrace{O(0,0)}_{=0} \\
\iff \qquad T_2 &= O(n-3,k-1)
\end{aligned}
$$

Auf ähnliche Art und Weise wie bei $O_1(n,k)$ erhalten wir für T_2:

$$
\begin{aligned}
T_2 &= O(n-3,k-1) \\
&= [x^{n-3}][y^{k-1}]\mathcal{O}(x,y)
\end{aligned}
$$

$$
T_2 = [x^n][y^k]x^3y\mathcal{O}(x,y) \tag{13}
$$

[1] $O(t,0) = 0\ ;\ t \neq 1$

Bei $O(n,k)$ in $(9)(9.1-9.3)$ setzen wir für $O(n,k)$ in (9.3) die entsprechende erzeugende Funktionen ein $\Rightarrow$

$$\mathcal{O}(x,y) = x^1y^0 + x^4y^2\mathcal{O}(x,y)^2 + x^2y\mathcal{O}(x,y)^2 - x^3y\mathcal{O}(x,y)$$

Daraus entsteht eine quadratische Gleichung in $\mathcal{O}(x,y)$:

$$(x^4y^2 + x^2y)\mathcal{O}(x,y)^2 - (x^3y+1)\mathcal{O}(x,y) + x = 0$$

Die Lösungen dieser quadratischen Gleichung sind die Nullstellen von $\mathcal{O}(x,y)$

$$\mathcal{O}(x,y)_+ = \frac{1}{2}\cdot\frac{x^3y+1+\sqrt{1+x^6y^2-4x^5y^2-2x^3y}}{x^2y(x^2y+1)}$$

$$\mathcal{O}(x,y)_- = \frac{1}{2}\cdot\frac{x^3y+1-\sqrt{1+x^6y^2-4x^5y^2-2x^3y}}{x^2y(x^2y+1)} \tag{14}$$

wir müssen wählen: $\mathcal{O}(x,y) = \mathcal{O}(x,y)_- = \frac{1}{2}\cdot\frac{x^3y+1-\sqrt{1+x^6y^2-4x^5y^2-2x^3y}}{x^2y(x^2y+1)}$ [1], weil der Grenzwert dieser Lösung existiert, d.h, $\lim\limits_{(x,y)\to(0,0)} \mathcal{O}(x,y)_-$ existiert.

Aus (6) haben wir: $P(n,k) = O(n-2,k-1) + O(n,k)$
Die erzeugende Funktion für $P(n,k)$ in (6) lautet:

$\mathcal{P}(x,y) = \mathcal{O}(x,y) + x^2y\mathcal{O}(x,y) \Longrightarrow$

$$\mathcal{P}(x,y) = (1+x^2y)\mathcal{O}(x,y) \tag{15}$$

Wir setzen (14) in (15) ein:

$$\mathcal{P}(x,y) = \frac{1}{2x^2y}(1+x^3y-\sqrt{1+x^6y^2-4x^5y^2-2x^3y}) \tag{16}$$

[1]Beachte, dass die Taylorreihe (Taylor-Entwicklung) von $\sqrt{1+x^6y^2-4x^5y^2-2x^3y}$ über $(x,y)=(0,0)$ ist: $1-x^3y-2x^5y^2-2x^8y^3-2x^{10}y^4-2x^{11}y^4+...$, wobei die Koeffizienten von (x,y) negativ sind. Deswegen hat $\mathcal{O}(x,y)_-$ ein negatives Vorzeichen vor dem Term $\sqrt{1+x^6y^2-4x^5y^2-2x^3y}$, seine Taylorreihe bei $(0,0)$ hat positive Koeffizienten für jeden Term $(x,y)^n$, wie für die erzeugende Funktion benötigt.

Aus (2) folgt:

$$\begin{aligned} R_5(n) &= \sum_{i=0}^{n} \sum_{k=0}^{i/2} P(i,k) \\ &= \sum_{i=0}^{n} \Big(\underbrace{\sum_{k=0}^{i/2} P(i,k) + \underbrace{\sum_{k=i/2}^{\infty} P(i,k)}_{0}}_{\sum_{k=0}^{\infty} P(i,k)} \Big) \end{aligned}$$

$$\Longrightarrow R_5(n) = \sum_{i=0}^{n} \underbrace{\sum_{k=0}^{\infty} P(i,k)}_{t(i)} = \sum_{i=0}^{n} t(i) \longrightarrow \textit{gesucht !} \tag{17}$$

Nach der Definition lautet die erzeugende Funktion für $P(i,k)$ wie folgt:

$$\begin{aligned} \mathcal{P}(x,y) &= \sum_{i} \sum_{k} P(i,k).x^i.y^k \\ \Longleftrightarrow \quad &= \sum_{i} x^i \sum_{k} P(i,k).y^k \end{aligned}$$

Für $y = 1 \Longrightarrow \mathcal{P}(x,1) = \sum_{i} \Big(\underbrace{\sum_{k} P(i,k)}_{t(i)} \Big) \cdot x^i$

$$\mathcal{P}(x,1) = \sum_{i=0}^{\infty} t(i) \cdot x^i \tag{18}$$

Um die Asymptotik für $R_5(n)$ zu berechnen, müssen wir zuerst die erzeugende Funktion für $R_5(n)$ nach der Variablen x aufstellen:

$$\mathcal{R}_5(x) = \sum_{n=0}^{\infty} R_5(n) \cdot x^n \overset{nach(17)}{=} \sum_{n=0}^{\infty} \Big(\sum_{i=0}^{n} t(i) \Big) x^n \tag{19}$$

Nach der Umformung ergibt sich:

$$\sum_{n=0}^{\infty}\big(\sum_{i=0}^{n} t(i)\big)x^n \;=\; \underbrace{\big(\sum_{i=0}^{\infty} t(i)\cdot x^i\big)}_{\mathcal{P}(x,1)}\cdot\underbrace{(\sum_{n=0}^{\infty} x^n)}_{\frac{1}{1-x}};$$ (Summe der unendlichen geometrischen Reihe für $|x|<1$)

Nach (16) , für $y = 1$ folgt daraus:

$$\Rightarrow \mathcal{R}_5(x) = \mathcal{P}(x,1)\cdot\frac{1}{1-x} = \frac{1}{1-x}\cdot\frac{1}{2x^2}(1+x^3-\sqrt{1+x^6-4x^5-2x^3})$$

$$\mathcal{R}_5(x) = \frac{1}{2x^2\cdot(1-x)}(1+x^3-\sqrt{1+x^6-4x^5-2x^3})$$

Die dominierende[2]Singularität[3] $x = \rho$ wird die Lösung von $D(x) = 1 + x^6 - 4x^5 - 2x^3$ mit kleinsten Betrag.
Das Polynom $D(x)$ ist eine Gleichung komplexer Variablen $6-ten$ Grads und besitzt genau 6 Lösungen (vier imaginäre Lösungen und zwei reale Lösungen):
$(4,117142763806521, 0,119280397900636 \pm 0,857272882739130i,$
$-0,508444936068360 \pm 0,481500996444461i, 0,661186312528936)$
Wir finden, dass die dominiernde Singularität in unserem Fall ist: $\rho := 0.66118631252893$.
Definieren: $A(x) = \frac{1+x^3}{2x^2\cdot(1-x)}$, *und* $U(x) = -\frac{\sqrt{1+x^6-4x^5-2x^3}}{2x^2\cdot(1-x)}$, so dass

$$\mathcal{R}_5(x) = A(x) + U(x) \tag{20}$$

$A(x)$ ist analytisch bei ρ. Für das asymptotische Verhalten der Koeffizienten von $R_5(n)$ reicht es, das singuläre Verhalten von $U(x)$ zu studieren.

Um eine explizite Formel für den asymptotischen Wert zu erhalten, verwenden wir das folgende Theorem:

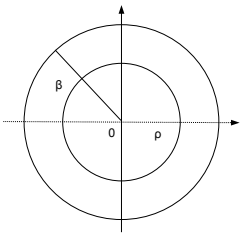
[2]dominierende Sing.: Sing. mit kleinstem Betrag.

[3]Singularität: Stellen, an der die Funktion nicht komplex differenzierbar ist, d.h $\lim_{z\to\ z_0}\frac{f(z)-f(z_0)}{z-z_0}$ existiert nicht

Theorem 1 (Flajolet und Odlyzko):

- *$f(z)$[4] sei analytisch in einer ,,geschlitzten Kreisscheibe"*

$$\{z \in \mathbb{C} \mid |z| < \beta\} \setminus [\rho, \beta)$$



- *$f(z)$ habe eine Singularität bei $z = \rho$, genauer*

$$f(z) \sim K \cdot (1 - z/\rho)^{\alpha} \quad (z \to \rho),$$

mit $\alpha \neq 0, 1, 2, ...$ und einer Konstanten K.
Dann gilt für die Koeffizienten

$$[z^n]f(z) = f_n \sim \frac{K}{\Gamma(-\alpha)} \cdot n^{-\alpha-1} \cdot \left(\frac{1}{\rho}\right)^n$$

ρ ist eine Nullstelle von $D(x)$ und daher ist $D(x) = (1 - x/\rho) \cdot Q(x)$ mit einem Polynom $Q(x)$.

$$\begin{aligned} D(x) &= 1 + x^6 - 4x^5 - 2x^3 \\ &= (1 - x/\rho) \cdot Q(x) \end{aligned}$$

$$D(x) = \underbrace{(1 - 1,512433002x)}_{1-x/\rho}(-0,661186313x^5 + 2,207577912x^4 + 1,459620299x^3 + 2,287453589x^2 + 1,512433004x + 0,761972928)$$

[4]Es sei $\mathbb{D} \subseteq \mathbb{C}$ eine offene Teilmenge. Eine Funktion $f : \mathbb{D} \to \mathbb{C}$ heißt analytisch in Punkt $x_0 \in \mathbb{D}$, wenn es eine Potenzreihe $\sum_{i \geq 0} a_n(x - x_0)^n$ gibt, die auf einer Umgebung von x_0 gegen $f(x)$ konvergiert. Ist f in jedem Punkt von $\mathbb{D}$ analytisch, so heißt f analytisch.

Damit haben wir für $x \longrightarrow \rho$

$$\begin{aligned}\mathcal{R}_5(x) &\sim A(\rho) + U(x) \\ &= A(\rho) - \frac{\sqrt{Q(x)}\sqrt{1-x/\rho}}{2x^2\cdot(1-x)} \\ &\sim A(\rho) - \frac{\sqrt{Q(\rho)}}{2\rho^2\cdot(1-\rho)}\sqrt{1-x/\rho}\end{aligned}$$

Also gilt

$$\mathcal{R}_5(x) \sim A(\rho) + K \cdot (1 - x/\rho)^{1/2},\ (x \to \rho)$$

mit $K = -\frac{\sqrt{Q(\rho)}}{2\rho^2(1-\rho)} \approx -6,335341642$

Mit $\Gamma(-\alpha) = \Gamma(-1/2) = -2\sqrt{\pi}$ erhalten wir aus dem Theorem 1 von Flajolet und Odlyzko:

$$R_5(n) = [x^n]\mathcal{R}_5(x) \sim \frac{K}{\Gamma(-1/2)} \cdot n^{-3/2} \cdot \left(\frac{1}{\rho}\right)^n$$

$$R_5(n) \sim 1,787166881 \cdot n^{-3/2} \cdot 1,512433002^n$$

Die letzte Gleichung gibt die asymptotische Zahl der Shapes kompatibler Sequenzen (Sekundärenstrukturen) der Länge $\leq n$; d.h. die Shapes der Sekundärstrukturen von RNA Sequenzen der Länge $\leq n$, unter der Annahme, dass die minimale Hairpin-Loop-Größe 3 Nukleotide sei, und die mimimale geschlossene Shape aus 5 Nukleotiden besteht.

4 Abstrakte Shapes der Stufe 5

4.1 Entwicklung der systematischen Vorgehensweise

In diesem Kapitel werden wir die Formel für die Anzahl der Shapes mit n Klammerpaaren herleiten. Um die dafür notwendige Formel zu ermitteln, benötigen wir die Shape-Grammatik für den Typ 5, die wir von der Abstraktions-Abbildung π_5 von Struktur zur Shape [4], und dann vom Homomorphismus ν_5 (um die Notation für Shapes zu definieren), herleiten.
Danach interessieren wir uns für die asymptotische Anzahl der Shapes, bestimmt durch die erzeugenden Funktionen für die Anzahl der Shapes mit n Klammerpaaren und der Anwendung des Theorems von Flajolet-Odlyzko.

4.2 Herleitung der Typ 5 Shape Grammatik

Zusammengefasst erklären wir den Shape-Typ 5 wie folgt:
Shape-Typ5 ist die stärkste Abstraktion, die Bulge und Internal-Loops ignoriert.

Wir benutzen die Variablen a und b für Nukleotide, l und l' für Loop-Sequenzen, c für eine Liste von benachbarten Komponenten und x für beliebige Strukturen. Die verschiedenen strukturellen Komponenten in der RNA, die wir als Knotenmarkierung eines Strukturbaums $\mathcal{B}$ benutzt haben, bezeichnen wir mit $SS, SR, BL, BR, IL, ML, AD$ und E. Im folgenden beschreiben wir die Knotenmarkierungen im Einzelnen [4].

$SS(l)$: Einzelstrang Regionen l
$HL(a, l, b)$: Hairpin-Loop mit Singlestranded Regionen l, geschlossen von Basenpaaren (a, b)
$SR(a, x, b)$: Stacking Regionen, geschlossen von Basenpaaren (a, b); x ist eine geschlossene Struktur.
$BL(a, l, x, b)$: Bulge-Loop links mit Singlestranded Regionen l, geschlossen von Basenpaaren (a, b); x ist eine geschlossene Struktur.
$BR(a, x, l, b)$: Bulge-Loop rechts mit Singlestranded Regionen l, geschlossen von Basenpaaren (a, b); x ist eine geschlossene Struktur.
$IL(a, l, x, l')$: Internal-Loop mit Singlestranded Regionen l und l', von Basenpaaren (a, b) geschlossen; x ist eine geschlossene Struktur.
$ML(a, c, b)$: Multi-Loop, geschlossen von Basenpaar (a, b).

$AD(x, c)$: Liste der benachbarten Strukturelemente; x ist eine Struktur, c ist eine (möglicherweise leere) Liste von benachbarten Strukturen.
E: eine leere Liste von benachbarten Strukturen.

Bei dem Shape-Typ 5 müssen wir die geschlossen Strukturen, Multi-Loops, benachbarte Regionen und die leere Struktur unterscheiden. Diese Strukturelemente werden in der Reihenfolge durch die Knoten-Labels OP, CL, FK, AD und E repräsentiert.

Die Abstraktions-Abbildung π_5 von Struktur zur Shape wird durch die folgende Gleichung definiert.
1)$\pi_5 : \mathit{Strukturbaum} \rightarrow \mathit{Shapebaum}$

$$
\begin{array}{lll}
1. & \pi_5(SS(l)) & = E \\
2. & \pi_5\big(HL(a,b,l)\big) & = CL \\
3. & \pi_5\big(SR(a,x,b)\big) & = \pi_5(x) \\
4. & \pi_5\big(BL(a,l,x,b)\big) & = \pi_5(x) \\
5. & \pi_5\big(BR(a,x,l,b)\big) & = \pi_5(x) \\
6. & \pi_5\big(IL(a,l,x,l',b)\big) & = \pi_5(x) \\
7. & \pi_5\big(ML(a,c,b)\big) & = FK\big(\pi_5(c)\big) \\
8. & \pi_5\big(AD(SS(l),c)\big) & = \pi_5(c) \\
9. & \pi_5\big(AD(c,SS(l))\big) & = \pi_5(c) \\
10. & \pi_5\big(AD(x,c)\big) & = AD\big(\pi_5(x),\pi_5(c)\big); x \neq SS(l) \\
11. & \pi_5(E) & = E
\end{array}
\tag{21}
$$

Hier können wir sehen, dass diese Abstraktion Hairpin-Loops (HL) und Multi-Loops (ML) erhält, aber von Stacking Regionen (SR), Bulges (BL,BR), und Internal-Loops (IL) abstrahiert.

Auf Grund der Gleichungen (21) können wir nun den Shapebaum für den Shape-Typ 5 zeichnen (Abbildung 11), dabei stützen wir uns auf den Strukturbaum in Abbildung 5.

$\pi_5(S) =$

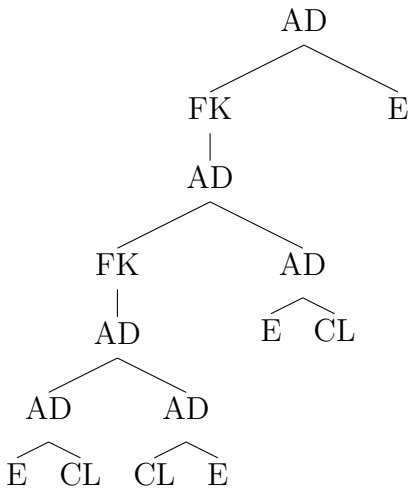


Abbildung 11**:** *Shapebaum* $\pi_5(S)$ *des Shape-Typs* 5 *für den Strukturbaum in Abbildung* 5

Um eine Notation für Shapes zu definieren, benutzen wir den Homomorphismus ν_5 [4]; wobei ε für die leere Sequenz steht.
2)ν_5 : *Shapebaum* $\rightarrow$ *Shapestring*

$$\begin{array}{lcl} 1.\ \nu_5(CL) & = & [\ ] \\ 2.\ \nu_5(FK(c)) & = & [\nu_5(c)] \\ 3.\ \nu_5(AD(x,c)) & = & \nu_5(x)\nu_5(c) \\ 4.\ \nu_5(E) & = & \varepsilon \end{array} \tag{22}$$

Wir wenden die Gleichungen (22) auf den Shapebaum $\pi_5(S)$ (Abbildung 11) an, um den Shapestring herzuleiten.

$$\begin{array}{ll} \nu_5(\pi_5(S_1)) & = \nu_5\big(AD(FK(AD(FK(AD(AD(E,CL),AD(CL,E))),AD(E,CL))),E)\big) \\ & = \nu_5\big(FK(AD(FK(AD(AD(E,CL),AD(CL,E))),AD(E,CL)))\big).\nu_5(E); \quad \nu_5(E)=\varepsilon \\ & = [\ \nu_5(AD(FK(AD(AD(E,CL),AD(CL,E))),AD(E,CL)))\ ].\varepsilon \\ & = [\ \nu_5(FK(AD(AD(E,CL),AD(CL,E))))\nu_5(AD(E,CL))\ ] \\ & = [\ [\nu_5(AD(E,CL))\nu_5(AD(E,CL))]\nu_5(E)\nu_5(CL)\ ]; \quad \nu_5(CL)=[\ ] \\ & = [\ [\nu_5(E)\nu_5(CL)\ \nu_5(E)\nu_5(CL)][\ ]\ ] \\ & = [\ [\ [\ ][\ ]\ ][\ ]\ ] \end{array}$$

Dieses entspricht dem Shape-Typ 5 für die gegebene Sequenz in Abbildung 6.

Im folgenden werden wir die Grammatik für den Typ 5 definieren.
Wir brauchen dafür:

1. Strukturen, die die Baum-Grammatik $\mathcal{B}$ in Abbildung 3 beschreiben.

2. Grammatik $\mathcal{G}$, die durch die Anwendung der Verkettung $(\nu_5 \circ \pi_5)$ auf Regeln von $\mathcal{B}$ entsteht.

Um die Grammatik zu finden, berechnen wir zuerst $\nu_5 \circ \pi_5$. Wir wissen, dass $f \circ g(x) = f\big(g(x)\big)$ ist.

Wir wenden nun die Verkettung $(\nu_5 \circ \pi_5)$ auf die Baum-Grammatik an, um die Grammatik des Typs 5 zu erzeugen.

Dafür wenden wir als erstes die Verkettungsabbildung auf die Baum-Grammatik, die mit *struct* beginnt, an:

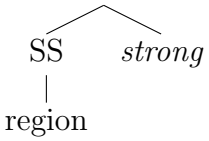
$(\nu_5 \circ \pi_5)(struct) = \nu_5\big(\pi_5(struct)\big)$ wobei: $struct \rightarrow \overbrace{comps}^{1} \mid \overbrace{E}^{2}$

1. $\nu_5\big(\pi_5(struct)\big) \rightarrow \nu_5\big(\pi_5(comps)\big)$.
2. $\nu_5\big(\pi_5(struct)\big) \rightarrow \nu_5\big(\pi_5(E)\big) \rightarrow \nu_5(E)$ $\big[$ergibt sich aus $\pi_5(E) = E$ in Gleichung (21)$\big] = \varepsilon$ $\big[$ $\nu_5(E) = \varepsilon$ aus Gleichung(22)$\big]$.

Insgesamt bekommen wir: $\nu_5\big(\pi_5(struct)\big) \rightarrow \begin{cases} \nu_5\big(\pi_5(comps)\big) \\ \varepsilon \end{cases}$

Nun wenden wir die Verkettungsabbildung auf *block* an, wobei:

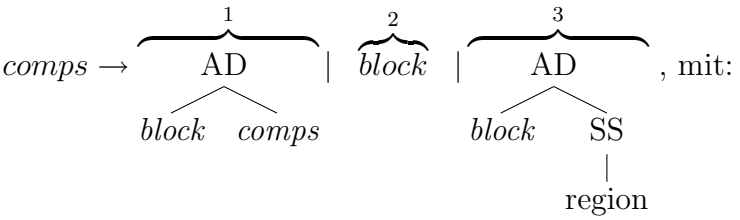
$block \rightarrow \overbrace{strong}^{1} \mid \overbrace{\text{AD}}^{2}$, mit:



1. $\nu_5\big(\pi_5(block)\big) \rightarrow \nu_5\big(\pi_5(strong)\big)$.
2. $\nu_5\big(\pi_5(block)\big) \rightarrow \nu_5\big(\pi_5(AD(SS(\text{region}), strong))\big) \rightarrow \nu_5\big(\pi_5(strong)\big)$ $\big[$nach Anwendung 8 aus Gleichung (21).

Dann bekommen wir $\nu_5\big(\pi_5(block)\big) \rightarrow \nu_5\big(\pi_5(strong)\big)$.

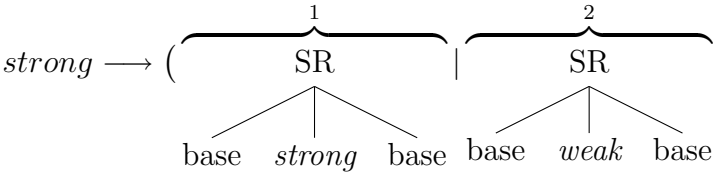
Auch wenden wir die Verkettungsabbildung auf *comps* an, wobei:



1. $\nu_5\big(\pi_5(comps)\big) \rightarrow \nu_5\big(\pi_5(AD(block, comps))\big) \rightarrow \nu_5\big(AD(\pi_5(block), \pi_5(comps))\big)$ $\big[$nach Anwendung 10 aus Gleichung (21) $\big] \rightarrow \nu_5\big(\pi_5(block)\big).\nu_5\big(\pi_5(comps)\big)$ $\big[$nach Anwendung 3 aus Gleichung (22) $\big]$.
2. $\nu_5\big(\pi_5(comps)\big) \rightarrow \nu_5\big(\pi_5(block)\big)$.
3. $\nu_5\big(\pi_5(comps)\big) \rightarrow \nu_5\big(\pi_5(AD(block, SS(\text{region})))\big) \rightarrow \nu_5\big(\pi_5(bock)\big)$ $\big[$nach Anwendung 10 aus Gleichung (21) $\big]$.

Insgesamt bekommen wir: $\nu_5\big(\pi_5(comps)\big) \;\; = \begin{cases} \nu_5\big(\pi_5(block)\big) \\ \nu_5\big(\pi_5(block)\big).\nu_5\big(\pi_5(comps)\big) \end{cases}$

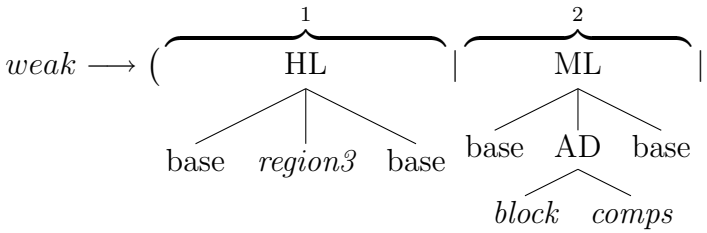
Wir wenden die Verkettungsabbildung auf *strong* an, wobei:

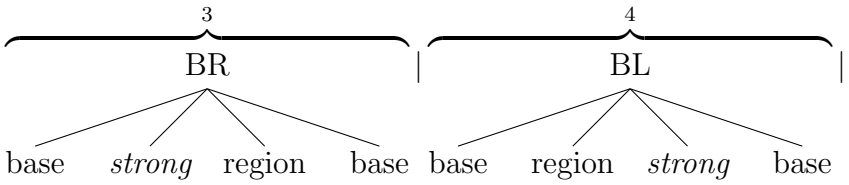


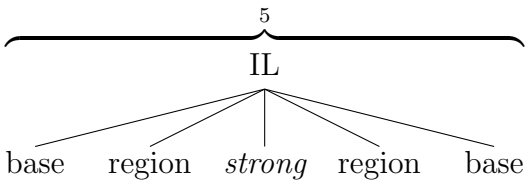
1. $\nu_5\big(\pi_5(strong)\big) \;\rightarrow\; \nu_5\big(\pi_5(SR(\text{base}, strong, \text{base}))\big) \;\rightarrow\; \nu_5\big(\pi_5(strong)\big)$ $\big[$nach Anwendung 3 aus Gleichung (21) $\big]$.
2. $\nu_5\big(\pi_5(strong)\big) \rightarrow \nu_5\big(\pi_5(SR(\text{base}, weak, \text{base}))\big) \rightarrow \nu_5\big(\pi_5(weak)\big)$ $\big[$nach Anwendung 3 aus Gleichung (21) $\big]$.

Dann bekommen wir: $\nu_5\big(\pi_5(strong)\big) = \nu_5\big(\pi_5(weak)\big)$

Jetzt wenden wir die Verkettungsabbildung auf *weak* an, wobei:







1. $\nu_5\big(\pi_5(weak)\big) \to \nu_5\big(\pi_5(HL(\text{base}, region3, \text{base}))\big) \to \nu_5(CL)$ [nach Anwendung 2 aus Gleichung (21)] $\to$ [ ] [nach Anwendung 1 aus Gleichung (22)].

2. $\nu_5\big(\pi_5(weak)\big) \to \nu_5\big(\pi_5(ML(\text{base}, AD(block, comps), \text{base}))\big) \to \nu_5\big(FK(\pi_5(AD(block, comps)))\big)$ [nach Anwendung 7 aus Gleichung (21)] $\to [\nu_5\big(\pi_5(AD(block, comps))\big)]$ [nach Anwendung 2 aus Gleichung (22)] $\to [\nu_5\big(AD(\pi_5(block), \pi_5(comps))\big)]$ [nach Anwendung 10 aus Gleichung (21)] $\to [\nu_5\big(\pi_5(block)\big)\nu_5\big(\pi_5(comps)\big)]$ [nach Anwendung 3 aus Gleichung (22)].

3. $\nu_5\big(\pi_5(weak)\big) \to \nu_5\big(\pi_5(BR(\text{base}, strong, \text{region}, \text{base}))\big) \to \nu_5\big(\pi_5(strong)\big)$ [nach Anwendung 5 aus Gleichung (21)].

4. $\nu_5\big(\pi_5(weak)\big) \to \nu_5\big(\pi_5(BL(\text{base}, \text{region}, strong, \text{base}))\big) \to \nu_5\big(\pi_5(strong)\big)$ [nach Anwendung 4 aus Gleichung (21)].

5. $\nu_5\big(\pi_5(weak)\big) \to \nu_5\big(\pi_5(IL(\text{base}, \text{region}, strong, \text{region}, \text{base}))\big) \to \nu_5\big(\pi_5(strong)\big)$ [nach Anwendung 6 aus Gleichung (21)].

Insgsammt bekommen wir: $\nu_5\big(\pi_5(weak)\big) \;\rightarrow \begin{cases} [\,] \\ \nu_5\big(\pi_5(strong)\big) \\ [\nu_5\big(\pi_5(block)\big).\nu_5\big(\pi_5(comps)\big)] \end{cases}$

Nach der bisherigen Anwendung der Verkettungsabbildungs auf $struct, block$, $comps, strong, weak$ aus der Baum-Grammatik $\mathcal{B}$, erhalten wir insgesamt die folgende Grammatik des Typs 5, nach Ersetzung $\nu_5\big(\pi_5(block)\big) = \nu_5\big(\pi_5(strong)\big)$:

$$\nu_5\big(\pi_5(struct)\big) \;\rightarrow \begin{cases} \nu_5\big(\pi_5(strong)\big) \\ \nu_5\big(\pi_5(strong)\big).\nu_5\big(\pi_5(comps)\big) \\ \varepsilon \end{cases} \tag{23}$$

$$\nu_5\big(\pi_5(comps)\big) \;\rightarrow \begin{cases} \nu_5\big(\pi_5(strong)\big) \\ \nu_5\big(\pi_5(strong)\big).\nu_5\big(\pi_5(comps)\big) \end{cases} \tag{24}$$

$$\nu_5\big(\pi_5(strong)\big) \rightarrow \nu_5\big(\pi_5(weak)\big) \tag{25}$$

$$\nu_5\big(\pi_5(weak)\big) \;\rightarrow \begin{cases} [\,] \\ \nu_5\big(\pi_5(strong)\big) \\ [\nu_5\big(\pi_5(strong)\big).\nu_5\big(\pi_5(comps)\big)] \end{cases} \tag{26}$$

Wir setzen für

$$\begin{array}{ll} \nu_5\big(\pi_5(struct)\big) & := S_5 \\ \nu_5\big(\pi_5(strong)\big) & := P \\ \nu_5\big(\pi_5(comps)\big) & := C \end{array}$$

Wir möchten nun die Grammatik für den Typ 5 aus (23,24,25,26) herleiten:

$$\begin{array}{lllll} Aus\ (23): & S_5 & \longrightarrow & \varepsilon \mid P \mid PC \\ Aus\ (24): & C & \longrightarrow & P \mid PC \\ Aus\ (25,26): & P & \longrightarrow & [\,] \mid [PC] \end{array}$$

vereinfacht sieht unsere Grammatik für den Typ 5 so aus:

$$\begin{array}{lll} S_5 & \longrightarrow & \varepsilon \mid C \\ C & \longrightarrow & P \mid PC \\ P & \longrightarrow & [\,] \mid [PC] \end{array} \tag{27}$$

Wir können aus (27) die folgende Formel $S_5(n)$ aufstellen: Anzahl der Shapes mit n Klammerpaaren:

$$S_5(n) = \begin{cases} 0 & : n = 0 \\ C(n) & : sonst \end{cases}$$

$$C(n) = \begin{cases} 0 & : n = 0 \\ P(n) + \sum_{i=1}^{n-1} P(i) * C(n-i) & : sonst \end{cases}$$

$$P(n) = \begin{cases} 0 & : n = 0 \\ 1 & : n = 1 \\ \sum_{i=1}^{n-2} P(i)C(n-i-1) & : sonst \end{cases}$$

4.3 Die asymptotische Anzahl der Shapes $S_5(n)$ mit n Klammerpaaren:

In diesem Abschnitt bestimmen wir die asymptotische Anzahl der Shapes durch die erzeugende Funktionen für die Anzahl der Shapes mit n Klammerpaaren und eines Theorems von Flajolet-Odlyzko.

Wir haben:

$$S_5(n) = \begin{cases} 0 & : n = 0 \\ C(n) & : sonst \end{cases} \tag{28}$$

$$C(n) = \begin{cases} 0 & : n = 0 \\ P(n) + \sum_{i=1}^{n-1} P(i) * C(n-i) & : sonst \end{cases} \tag{29}$$

$$P(n) = \begin{cases} 0 & : n = 0 \\ 1 & : n = 1 \\ \sum_{i=1}^{n-2} P(i)C(n-i-1) & : sonst \end{cases} \tag{30}$$

Um unser Ziel (die asymptotische Anzahl der Shape-Typ 5 mit n Klammerpaaren) zu erhalten, müssen wir zuerst die erzeugende Funktion der Rekurrenzformel für $S_5(n)$ ermitteln. Dafür ermitteln wir die erzeugende Funktion der Rekurrenzformel für $P(n), C(n)$. Für erzeugende Funktionen gilt die folgende Definition:

Definition: Für eine Folge $f_n, n \in \mathbb{N}_0$ heißt die formale Potenzreihe $f(z) = \sum_{n=0}^{\infty} f_n z^n$ die erzeugende Funktion von f_n.
wir bezeichnen für eine Potenzreihe $f(z)$ den Koeffizienten f_n mit $[z^n]f(z)$.

Angewendet auf $P(n)$, müssen wir die erzeugende Funktion der Rekurrenzformel (30) ermitteln, und es gilt für die erzeugende Funktion:

$$P(z) = \sum_{n\geq 0} P(n)z^n = P(0).z^0 + P(1).z^1 + \underbrace{\sum_{n\geq 2} z^n \sum_{i=0}^{n-1} P(i) * C(n-i-1)}_{T} \quad (31)$$

$$\begin{aligned} T &= \sum_{n\geq 2} z^n \sum_{i=0}^{n-1} P(i) * C(n-i-1) \\ &= z \sum_{n\geq 2} z^{n-1} \sum_{i=0}^{n-1} P(i) * C(n-i-1) \\ &= z(z^1[P(0)C(1) + P(1)C(0)] + z^2[P(0)C(2) + P(1)C(1) + P(2)C(0)] + ...) \\ &= z(z^0[P(0)C(0)] + z^1[P(0)C(1) + P(1)C(0)] + z^2[P(0)C(2) + P(1)C(1) + P(2)C(0)] + ...) \\ &= zP(z)C(z) \end{aligned}$$

Die Gleichnug (31) ergibt sich zu:

$$P(z) = z + zP(z)C(z) \quad (32)$$

Auf ähnliche Art und Weise wie bei der erzeugenden Funktion für $P(n)$ erhalten wir die erzeugende Funktion für $C(n)$:

$$C(z) = \sum_{n\geq 0} C(n)z^n = C(0).z^0 + \underbrace{\sum_{n\geq 1} P(n)z^n}_{P(z)} + \underbrace{\sum_{n\geq 1} z^n \sum_{i=0}^{n} P(i) * C(n-i)}_{P(z)C(z)}$$

$$\Longrightarrow C(z) = P(z) + P(z)C(z) \tag{33}$$

Auflösen aus (32) und nach $P(z)$ liefert:

$$P(z)\big(1 - zC(z)\big) = z \Rightarrow P(z) = \frac{z}{1 - zC(z)} \tag{34}$$

(33) einsetzen in (34) $\Longrightarrow C(z) = \frac{z}{1-zC(z)} + \frac{zC(z)}{1-zC(z)}$

$$\Longrightarrow C(z)^2 + (z-1)C(z) + z = 0$$

Die Wurzeln dieser quadratischen Gleichung sind:

$$C(z)_+ = \frac{1 - z + \sqrt{-3z^2 - 2z + 1}}{2z}$$
$$C(z)_- = \frac{1 - z - \sqrt{-3z^2 - 2z + 1}}{2z}$$

Wir müssen $C(z) = C(z)_- = \frac{1-z-\sqrt{-3z^2-2z+1}}{2z}$ [5] wählen, weil der $\lim_{z\to o} C(z)$ nur für die zweite Lösung existiert.

Im vorliegenden Fall ist die dominierende Singularität ρ eine Lösung der Gleichung $(-3z^2 - 2z + 1 = 0)$, die Wurzeln sind $-1, \frac{1}{3}$, wir wählen ρ mit dem kleinsten $\rho = \frac{1}{3}$.

Wir definieren: $W(z) = \frac{1-z}{2z}$, und $H(z) = -\frac{\sqrt{-3z^2-2z+1}}{2z}$, so dass:

$$C(z) = W(z) + H(z)$$

[5] Beachte, dass die Taylorreihe (Taylor-Entwicklung) von $\sqrt{-3z^2 - 2z + 1}$ über $z = 0$ ist: $1 - z - 2z^2 - 2z^3 - 4z^4 - 8z^5 - 18z^6 + ...$, wobei die Koeffizienten von z negativ sind. Deswegen hat $\mathcal{C}(z)_-$ ein negatives Vorzeichen vor dem Term $\sqrt{-3z^2 - 2z + 1}$, seine Taylorreihe bei 0 hat positive Koeffizienten für jeden Term z^n, wie für die erzeugende Funktion benötigt.

.

ρ ist eine Nullstelle von $I(z) = -3z^2 - 2z + 1$. Rechnen wir $(1 - z/\rho)$ aus dem Polynom $I(z)$ heraus, um $Q(z)$ zu erhalten, folgt:

$$\begin{aligned} I(z) &= -3z^2 - 2z + 1 \\ &= Q(z) \cdot (1 - z/\rho),\ \rho = 1/3 \Rightarrow I(z) = (z+1)\underbrace{(1-3z)}_{1-z/\rho} \end{aligned}$$

Damit haben wir für $z \longrightarrow \rho$

$$\begin{aligned} C(z) &\sim W(\rho) + H(z) \\ &= W(\rho) - \tfrac{\sqrt{Q(z)}\sqrt{1-z/\rho}}{2z};\ Q(z) = (z+1) \\ &\sim W(\rho) - \tfrac{\sqrt{Q(\rho)}}{2\rho}\sqrt{1-z/\rho} \\ &= W(\rho) - \tfrac{\sqrt{\rho+1}}{2\rho}\sqrt{1-z/\rho} \end{aligned}$$

Also gilt

$$C(z) \sim W(\rho) + K(1 - z/\rho)^{1/2},\ (z \to \rho)$$

mit $K = -\frac{\sqrt{\rho+1}}{2\rho} = -\frac{\sqrt{\frac{4}{3}}}{2(\frac{1}{3})} = -\sqrt{3} \Rightarrow K = -\sqrt{3}$

Mit $\Gamma(-\alpha) = \Gamma(-1/2) = -2\sqrt{\pi}$ erhalten wir aus dem Theorem 1 von Flajolet und Odlyzko:

$$\begin{aligned} C(n) = [x^n]C(z) &\sim \tfrac{K}{\Gamma(-\alpha)} \cdot n^{-\alpha-1} \cdot \left(\tfrac{1}{\rho}\right)^n = \tfrac{\sqrt{3}}{2\sqrt{\pi}} \cdot n^{-3/2} \cdot 3^n \\ &= \tfrac{1}{2}\sqrt{\tfrac{3}{\pi}} \cdot n^{-3/2} \cdot 3^n \end{aligned}$$

$$S_5(n) = C(n) \sim 0,488602511 \cdot n^{-3/2} \cdot 3^n$$

Die letzte Gleichung gibt die asymptotische Anzahl $S_5(n)$ der Shapes mit n Klammerpaaren an.
Im Vergleich mit den Ergebnissen von Nebel und Scheid [9], bekommen wir die gleichen Ergebnisse. Allerdings findet sich in deren Ergebnissen die genaue Asymptotik der Anzahl $S_5(n)$ von Typ 5 Shapes der Größe n, und in unseren Ergebnissen findet sich der genaue asymptotische Wert für die Anzahl der Shapes mit n Klammerpaaren.

Ergebniss von Nebel und Scheid:
$S_5(n) \sim 3^{3/2}(1+(-1)^n)\cdot\sqrt{\frac{3}{2\pi}}(\frac{1}{n})^{3/2} \approx 1,73205^n(1.+(-1.)^n)\cdot 0,690988\cdot n^{-3/2}$

(n) Klammerpaare$(Kp) \mathrel{\widehat{=}}$ Größe $(2n)$, dann setzen wir in die Ergebnis von Nebel und Scheid $(2n)$ statt (n) ein; wobei n gerade ist.

$$\begin{aligned} S_5\underbrace{(2n)}_{\text{Größe}} = S_5\underbrace{(n)}_{Kp} &\approx (\sqrt{3})^{2n}\cdot 2\cdot 0,690988\cdot(2n)^{-3/2} \\ &\approx 3^n\cdot 2\cdot 0,690988\cdot 2^{-3/2}\cdot n^{-3/2}; 2^{-3/2}=\tfrac{1}{2\sqrt{2}} \\ &\approx 0,488602511\cdot n^{-3/2}\cdot 3^n(unsere\ Ergebnisse) \end{aligned}$$

Im Vergleich mit den Ergebnissen von Clote et al. [8] $[z^{2n}]S(z) = S_5(n) = \sqrt{\frac{6}{\pi}}(2n)^{-3/2}(\sqrt{3})^{2n}$ [6], erhalten das gleiche Ergebnis.

[6] $\sqrt{\frac{6}{\pi}}(2n)^{-3/2}(\sqrt{3})^{2n} = 2^{-3/2}\cdot\sqrt{\frac{6}{\pi}}\cdot n^{-3/2}\cdot 3^n = 0,488602511\cdot n^{-3/2}\cdot 3^n$

5 Abstrakte Shapes der Stufen 1, 2, 3, 4

In diesem Kapitel werden wir die Formel für die Anzahl der Shapes mit n Klammerpaaren herleiten. Um die benötigte Formel zu ermitteln, brauchen wir die Shape-Grammatik für Typen $i, \big(i \in \{1,2,3,4\}\big)$, die von der Abstraktions-Abbildung $\pi_i, i \in \{1..4\}$ von Struktur nach Shape operiert, und dann vom Homomorphismus $\nu_i, \big(i \in \{1,2,3,4\}\big)$ (Um die Notation für Shapes zu definieren)

Danach bestimmen wir die asymptotische Anzahl der Shapes-Typen 3 und 4 mit n Klammerpaaren. Das geschieht mit Hilfe der erzeugenden Funktionen und der Anwendung des Theorems von Flajolet-Odlyszko.

Wir werden feststellen, dass die asymptotische Anzahl der Shape-Typen 3 und 4 gleich sind.
Außerdem zeichnen wir den Graphen für die Anzahl der Shape-Typen $S_i(n)$, $\big(i \in \{1...5\}\big)$ mit n Klammerpaaren in ein geeignetes Koordinatensystem.

5.1 Herleitung der Shape-Grammatik von Typ 1

Kurzgefasst erklären wir den Shape-Typ 1 wie folgt:
Shape-Typ 1 unterscheidet alle Loops und alle ungepaarten Basen.

Wir benutzen die gleichen Daten wie in (4.2) für die Variablen a, b, l, l', c, x und für die Knotenmarkierung auch in diesem Abschnitt.

Bei dem Shape-Typ 1 müssen wir die offene und geschlossene Strukturen, Bulge-Loops (links und rechts), Internal-Loops, Multi-Loops, benachbarte Regionen und die leere Struktur unterscheiden. Diese Strukturelemente werden in der Reihenfolge durch die Knoten-Labels $OP, CL, CCBL, CCBR, CCIL, FK, AD$ und E repräsentiert.

Die Abstraktions-Abbildung π_1 von Struktur nach Shape wird durch die folgende Gleichung definiert.
1)$\pi_1 : \mathit{Strukturbaum} \rightarrow \mathit{Shapebaum}$

$$
\begin{array}{lcl}
\pi_1(SS(l)) & = & OP \\
\pi_1\big(HL(a,b,l)\big) & = & CL \\
\pi_1\big(SR(a,x,b)\big) & = & \pi_1(x) \\
\pi_1\big(BL(a,l,x,b)\big) & = & CCBL\big(OP,\pi_1(x)\big) \\
\pi_1\big(BR(a,x,l,b)\big) & = & CCBR\big(\pi_1(x),OP\big) \\
\pi_1\big(IL(a,l,x,l',b)\big) & = & CCIL\big(OP,\pi_1(x),OP\big) \\
\pi_1\big(ML(a,c,b)\big) & = & FK\big(\pi_1(c)\big) \\
\pi_1\big(AD(x,c)\big) & = & AD\big(\pi_1(x),\pi_1(c)\big) \\
\pi_1(E) & = & E
\end{array}
\tag{35}
$$

Auf Grund der Gleichungen (35) können wir nun den Shapebaum für den Shape-Typ 1 zeichnen (Abbildung 12). Dabei stützen wir uns auf den Strukturbaum in Abbildung 5.

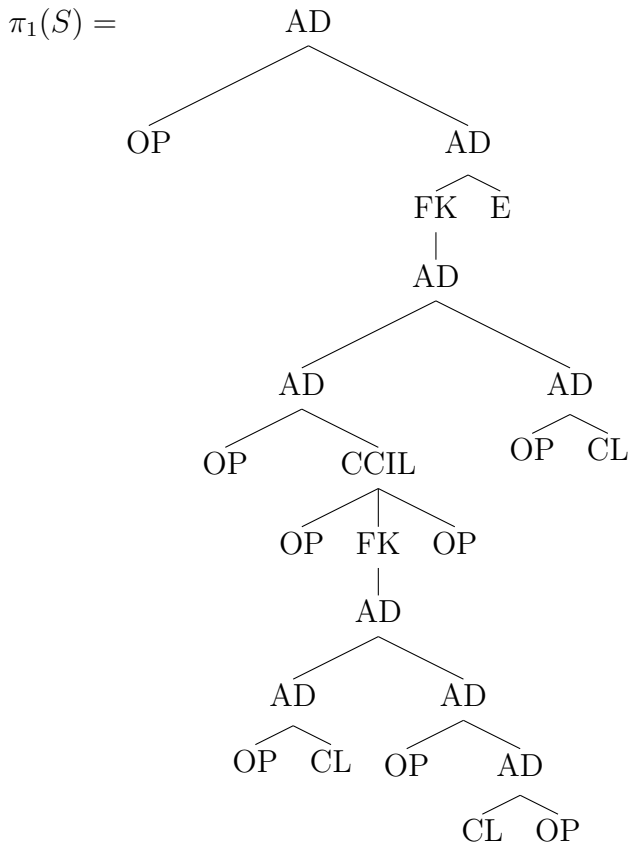


Abbildung 12**:** *Shapebaum* $\pi_1(S)$ *des Shape-Typs* 1 *für den Strukturbaum in*

Abbildung 5

Um eine Notation für Shapes zu definieren, benutzen wir den Homomorphismus ν_1; wobei ε für die leere Sequenz steht.
2)$\nu_1 : Shapebaum \rightarrow Shapestring$

$$\begin{array}{lcl}
\nu_1(OP) & = & _ \\
\nu_1(CL) & = & [\ ] \\
\nu_1(FK(c)) & = & [\nu_1(c)] \\
\nu_1(CCBL(OP,x)) & = & [_\ \nu_1(x)] \\
\nu_1(CCBR(x,OP)) & = & [\nu_1(x)_] \\
\nu_1(CCIL(OP,x,OP)) & = & [_\ \nu_1(x)_] \\
\nu_1(AD(x,c)) & = & \nu_1(x)\nu_1(c) \\
\nu_1(E) & = & \varepsilon
\end{array} \tag{36}$$

Wir wenden die Gleichungen(36) auf dem Shapebaum $\pi_1(S)$ (Abbildung 12) an, um den Shapestring herzuleiten.

$$\begin{array}{ll}
\nu_1(\pi_1(S)) & = \nu_1\big(AD(OP, AD(FK(AD(AD(OP, CCIL(OP, FK(AD(AD(OP, CL), \\
& \quad AD(OP, AD(CL, OP)))), OP))), AD(OP, CL))), E)\big) \\
& = \nu_1(OP)\nu_1\big(AD(FK(AD(AD(AD(OP, CCIL(OP, FK(AD(AD(OP, CL), \\
& \quad AD(OP, AD(CL, OP)))), OP))), AD(OP, CL))), E)\big) \\
& = _\ \nu_1\big(FK(AD(AD(OP, CCIL(OP, FK(AD(AD(OP, CL), \\
& \quad AD(OP, AD(CL, OP)))), OP))), AD(OP, CL)))\big)\ \nu_1(E)\ ;\ \nu_1(E) = \varepsilon \\
& = _\ [\ \nu_1(AD(AD(OP, CCIL(OP, FK(AD(AD(OP, CL), AD(OP, AD(CL, OP)))), OP)))) \\
& \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \nu_1(AD(OP, CL))\ ] \\
& = _\ [\ \nu_1(AD(OP, CCIL(OP, FK(AD(AD(OP, CL), AD(OP, AD(CL, OP)))), OP))) \\
& \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \nu_1(OP)\nu_1(CL)\ ] \\
& = _\ [\nu_1(OP, CCIL(OP, FK(AD(AD(OP, CL), AD(OP, AD(CL, OP))\ _\ [\ ]\ ] \\
& = _\ [\ \nu_1(OP)\ [_\ \nu_1(FK(AD(AD(OP, CL), AD(OP, AD(CL, OP)))))\ _\ ]\ _\ [\ ]\ ] \\
& = _\ [_[_[\ \nu_1(OP)\nu_1(CL)\nu_1(OP)\nu_1(CL)\nu_1(OP)]_\ ]_[\ ]\ ] \\
& = _\ [_[_[\ _[\ ]_[\ ]_\ ]_[\ ]\ ]
\end{array}$$

Dieses entspricht dem Shape-Typ 1 für die gegebene Sequenz in Abbildung 6.

Im folgenden werden wir die Grammatik für den Typ 1 definieren.
Wir brauchen dafür:

1. Strukturen, die die Baum-Grammatik $\mathcal{B}$ in Abbildung 3 beschreiben.
2. Finde die Grammatik $\mathcal{G}$ durch die Anwendung der Verkettung $(\nu_1 \circ \pi_1)$ auf Regeln von $\mathcal{B}$.

Mit den gleichen Herleitungsmethoden für die Grammatik des Shape-Typs 5 in (4.2) bekommen wir die Grammatik des Typs 1.

$$\nu_1\big(\pi_1(struct)\big) \;\rightarrow \begin{cases} \nu_1\big(\pi_1(comps)\big) \\ \varepsilon \end{cases} \tag{37}$$

$$\nu_1\big(\pi_1(comps)\big) \;\rightarrow \begin{cases} \nu_1\big(\pi_1(block)\big).\nu_1\big(\pi_1(comps)\big) \\ \nu_1\big(\pi_1(block)\big) \\ \nu_1\big(\pi_1(block)\big)\ \_ \end{cases} \tag{38}$$

$$\nu_1\big(\pi_1(strong)\big) \rightarrow \nu_1\big(\pi_1(weak)\big) \tag{39}$$

$$\nu_1\big(\pi_1(weak)\big) \;\rightarrow \begin{cases} [\_] \\ [\nu_1\big(\pi_1(block)\big).\nu_1\big(\pi_1(comps)\big)] \\ [\nu_1\big(\pi_1(strong)\big)\ \_\ ] \\ [\ \_\ \nu_1\big(\pi_1(strong)\big)] \\ [\ \_\ \nu_1\big(\pi_1(strong)\big)\ \_\ ] \end{cases} \tag{40}$$

$$\nu_1\big(\pi_1(block)\big) \;\rightarrow \begin{cases} \nu_1\big(\pi_1(strong)\big) \\ \_\ \nu_1\big(\pi_1(strong)\big) \end{cases} \tag{41}$$

Wir setzen für

$$\begin{array}{ll} \nu_1\big(\pi_1(struct)\big) & := S_1 \\ \nu_1\big(\pi_1(block)\big) & := B \\ \nu_1\big(\pi_1(strong)\big) & := P \\ \nu_1\big(\pi_1(comps)\big) & := C \end{array}$$

Wir möchten nun die Grammatik für den Typ 1 aus (37-41) herleiten:

$$\begin{array}{lrcl} Aus\ (37): & S_1 & \longrightarrow & \varepsilon \mid C \\ Aus\ (38): & C & \longrightarrow & B \mid B\_ \mid BC \\ Aus\ (41): & B & \longrightarrow & P \mid \_P \\ Aus\ (39,40): & P & \longrightarrow & [\ \_\ ] \mid [P\_] \mid [\_P] \mid [\_P\_] \mid [BC] \end{array} \tag{42}$$

Wir können aus (42) die folgende Formel $S_1(n,m)$ aufstellen: Anzahl der Shapes mit n Klammerpaaren und m Unterstrichen

$$S_1(n,m) \;= \begin{cases} 1 & : n = 0, m = 0 \\ C(n,m) & : sonst \end{cases}$$

$$C(n,m) = \begin{cases} 0 & : n = 0, m = 0 \\ B(n,m) + B(n,m-1) + & : sonst \\ \sum_{i=1}^{n} \sum_{j=1}^{m} B(i,j) * C(n-i, m-j) & \end{cases}$$

$$B(n,m) = P(n,m) + P(n,m-1)$$

$$P(n,m) = \begin{cases} 0 & : n = 0, m = 0 \\ 1 & : n = 1, m = 1 \\ 2 * P(n-1,m-1) + P(n-1,m-2) + & : sonst \\ \sum_{i=2}^{n-2} \sum_{j=2}^{m} B(i,j) * C(n-i, m-j) & \end{cases}$$

5.2 Herleitung des Typs 2 Shape-Grammatik

Shape-Typ 2 unterscheidet alle Loops und alle ungepaaarten Basen in External-Loops und Multi-Loops.

Wie beim Shape-Typ 1 werden die gleichen Variablen und Knotenmarkierung aus (4.2) auch hier eingesetzt.

Bei dem Shape-Typ 2 müssen wir die geschlossen Strukturen, Bulge-Loops (links und rechts), Internal-Loops, Multi-Loops, benachbarte Regionen und die leere Struktur unterscheiden. Diese Strukturelemente werden in der Reihenfolge durch die Knoten-Labels $CL, CCBL, CCBR, CCIL, FK, AD$ und E repräsentiert.

Die Abstraktions-Abbildung π_2 von Struktur zu Shape wird durch die folgende Gleichung definiert.

1)$\pi_2 : Strukturbaum \rightarrow Shapebaum$

$$
\begin{array}{lcl}
\pi_2(SS(l)) & = & E \\
\pi_2\big(HL(a,b,l)\big) & = & CL \\
\pi_2\big(SR(a,x,b)\big) & = & \pi_2(x) \\
\pi_2\big(BL(a,l,x,b)\big) & = & CCBL\big(OP,\pi_2(x)\big) \\
\pi_2\big(BR(a,x,l,b)\big) & = & CCBR\big(\pi_2(x),OP\big) \\
\pi_2\big(IL(a,l,x,l',b)\big) & = & CCIL\big(OP,\pi_2(x)OP\big) \\
\pi_2\big(ML(a,c,b)\big) & = & FK\big(\pi_2(c)\big) \\
\pi_2\big(AD(x,c)\big) & = & AD\big(\pi_2(x),\pi_2(c)\big) \\
\pi_2(E) & = & E
\end{array}
\tag{43}
$$

Auf Grund der Gleichungen (43) können wir nun den Shapebaum für den Shape-Typ 2 zeichnen (Abbildung 13), dabei stützen wir uns auf den Strukturbaum in Abbildung 5.

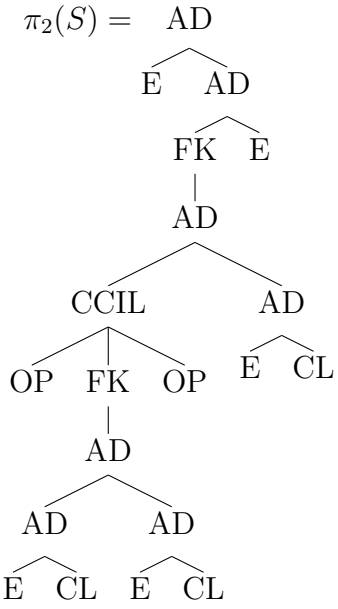


Abbildung 13**:** *Shapebaum* $\pi_2(S)$ *des Shape-Typs* 2 *für den Strukturbaum in Abbildung* 5

Um eine Notation für Shapes zu definieren, benutzen wir den Homomorphismus ν_2 ; wobei ε für die leere Sequenz steht.

2)$\nu_2 : Shapebaum \rightarrow Shapestring$

$$
\begin{array}{lcl}
\nu_2(OP) & = & \_ \\
\nu_2(CL) & = & [\ ] \\
\nu_2(FK(c)) & = & [\nu_2(c)] \\
\nu_2\big(CCBL(OP,x)\big) & = & [\_\nu_2(x)] \\
\nu_2\big(CCBR(x,OP)\big) & = & [\nu_2(x)\_] \\
\nu_2\big(CCIL(OP,x,OP)\big) & = & [\_\nu_2(x)\_] \\
\nu_2(AD(x,c)) & = & \nu_2(x)\nu_2(c) \\
\nu_2(E) & = & \varepsilon
\end{array}
\tag{44}
$$

Wir wenden die Gleichungen(44) auf den Shapebaum $\pi_1(S)$ (Abbildung 12) an, um den Shapestring herzuleiten.

$$
\begin{array}{ll}
\nu_2(\pi_2(S)) & = \nu_2(E)\nu_2(AD(FK(AD(CCIL(OP,FK(AD(AD(E,CL),AD(E,CL))),OP),AD(E,CL))),E)) \\
& = \varepsilon.\nu_2(FK(AD(CCIL(OP,FK(AD(AD(E,CL),AD(E,CL))),OP),AD(E,CL))))\nu_2(E) \\
& = [\nu_2(AD(CCIL(OP,FK(AD(AD(E,CL),AD(E,CL))),OP),AD(E,CL)))].\varepsilon \\
& = [\nu_2(CCIL(OP,FK(AD(AD(E,CL),AD(E,CL))),OP))\nu_2(AD(E,CL))] \\
& = [\ [\nu_2(OP)\nu_2(FK(AD(AD(E,CL),AD(E,CL))))\nu_2(OP)]\ \nu_2(E)\nu_2(CL)] \\
& = [\ [\_\ [\nu_2(E)\nu_2(CL)\nu_2(E)\nu_2(CL)\ ]\ \_\ ]\ \varepsilon\ [\ ]\ ] \\
& = [\ [\_\ [[\ ][\ ]\ ]\ \_\ ]\ [\ ]\ ]
\end{array}
$$

Dieses enspricht dem Shape-Typ 2 für die gegebene Sequenz in Abbildung 6.

Im folgenden definieren wir die Grammatik für den Typ 2.

Wir brauchen dafür:

1. Strukturen, die die Baum-Grammatik $\mathcal{B}$ in Abbildung 3 beschreiben.
2. Finde die Grammatik $\mathcal{G}$ durch die Anwendung der Verkettung $(\nu_2 \circ \pi_2)$ auf Regeln von $\mathcal{B}$.

$$
\nu_2\big(\pi_2(struct)\big) \ \rightarrow \left\{ \begin{array}{l} \nu_2\big(\pi_2(comps)\big) \\ \varepsilon \end{array} \right. \tag{45}
$$

$$
\nu_2\big(\pi_2(comps)\big) \ \rightarrow \left\{ \begin{array}{l} \nu_2\big(\pi_2(block)\big).\nu_2\big(\pi_2(comps)\big) \\ \nu_2\big(\pi_2(block)\big) \end{array} \right. \tag{46}
$$

$$
\nu_2\big(\pi_2(strong)\big) \rightarrow \nu_2\big(\pi_2(weak)\big) \tag{47}
$$

$$\nu_2\big(\pi_2(weak)\big) \;\rightarrow \begin{cases} [\ ] \\ [\nu_2\big(\pi_2(strong)\big)\ \_\ ] \\ [\ \_\ \nu_2\big(\pi_2(strong)\big)] \\ [\ \_\ \nu_2\big(\pi_2(strong)\big)\ \_\ ] \\ [\nu_2\big(\pi_2(block)\big).\nu_2\big(\pi_2(comps)\big)] \end{cases} \tag{48}$$

$$\nu_2\big(\pi_2(block)\big) \rightarrow\ \nu_2\big(\pi_2(strong)\big) \tag{49}$$

Wir setzen für:

$$\begin{array}{ll} \nu_2\big(\pi_1(struct)\big) & := S_2 \\ \nu_2\big(\pi_1(block)\big) & := B \\ \nu_2\big(\pi_1(strong)\big) & := P \\ \nu_2\big(\pi_1(comps)\big) & := C \end{array}$$

Wir möchten nun die Grammatik für den Typ 2 aus $(45-49)$ herleiten.

$$\begin{array}{lllll} Aus\ (45): & S_2 & \longrightarrow & \varepsilon \mid C \\ Aus\ (46): & C & \longrightarrow & B \mid BC \\ Aus\ (49): & B & \longrightarrow & P \\ Aus\ (47,48): & P & \longrightarrow & [\ ] \mid [BC] \mid [P\_] \mid [\_P] \mid [\_P\_] \end{array}$$

Vereinfacht sieht unsere Grammatik für den Typ 2 so aus:

$$\begin{array}{lll} S_2 & \longrightarrow & \varepsilon \mid C \\ C & \longrightarrow & P \mid PC \\ P & \longrightarrow & [\ ] \mid [PC] \mid [P\_] \mid [\_P] \mid [\_P\_] \end{array} \tag{50}$$

Wir können aus (50) die folgende Formel $S_2(n,m)$ aufstellen: Anzahl der Shapes mit n Klammerpaaren und m Unterstrichen.

$$S_2(n,m) \;= \begin{cases} 1 & : n=0, m=0 \\ C(n,m) & : sonst \end{cases}$$

$$C(n,m) \;= \begin{cases} 0 & : n=0, m=0 \\ P(n,m) + \displaystyle\sum_{i=1}^{n}\sum_{j=1}^{m} P(i,j) * C(n-i, m-j) & : sonst \end{cases}$$

$$P(n,m) \quad = \begin{cases} o & : n = 0, m = 0 \\ 1 & : n = 1 \\ 2 * P(n-1, m-1) + P(n-1, m-2) + \sum_{i=2}^{n-2} \sum_{j=2}^{m} P(i,j) * C(n-i, m-j) & : sonst \end{cases}$$

5.3 Herleitung des Typs 3 Shape-Grammatik

Zusammengefasst erhält die Abstraktion Hairpin-Loops (HL), Multi-Loops (ML), Bulges (BL,BR), und Internal-Loops (IL), abstrahiert aber von Stacking Regionen (SR).

Wir benutzen die gleichen Daten aus $(4, 2)$ für die Variablen und für die Knotenmarkierung auch in diesem Abschnitt.

Bei dem Shape-Typ 3 müssen wir die geschlossen Strukturen, Bulge-Loops (links und rechts), Internal-Loops, Multi-Loops, benachbarte Regionen und die leere Struktur unterscheiden. Diese Strukturelemente werden in der Reihenfolge durch die Knoten-Labels $CL, CCBL, CCBR, CCIL, FK, AD$ und E repräsentiert.
Die Abstraktions-Abbildung π_3 von Struktur zu Shape wird durch die folgende Gleichung definiert.

$1) \pi_3 : Strukturbaum \rightarrow Shapebaum$

$$\begin{array}{lcl} \pi_3(SS(l)) & = & E \\ \pi_3\big(HL(a,b,l)\big) & = & CL \\ \pi_3\big(SR(a,x,b)\big) & = & \pi_3(x) \\ \pi_3\big(BL(a,l,x,b)\big) & = & CCBL\big(\pi_3(x)\big) \\ \pi_3\big(BR(a,x,l,b)\big) & = & CCBR\big(\pi_3(x)\big) \\ \pi_3\big(IL(a,l,x,l',b)\big) & = & CCIL\big(\pi_3(x)\big) \\ \pi_3\big(ML(a,c,b)\big) & = & FK\big(\pi_3(c)\big) \\ \pi_3\big(AD(SS(l),c)\big) & = & \pi_3(c) \\ \pi_3\big(AD(x,c)\big) & = & AD\big(\pi_3(x), \pi_3(c)\big); x \neq SS(l) \\ \pi_3(E) & = & E \end{array} \tag{51}$$

Auf Grund der Gleichungen (51) können wir nun den Shapebaum für den Shape-Typ 3 zeichnen (Abbildung 14). Dabei stützen wir uns auf den Strukturbaum in Abbildung 5.

$\pi_3(S) =$

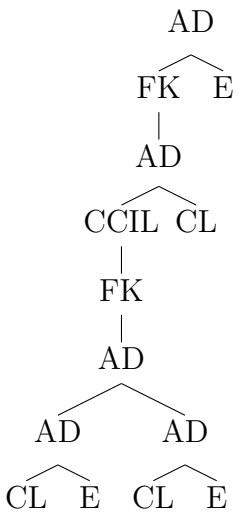


Abbildung 14*: Shapebaum* $\pi_3(S)$ *des Shape-Typs* 3 *für den Strukturbaum in Abbildung* 5

Um eine Notation für Shapes zu definieren, benutzen wir den Homomorphismus ν_3; wobei ε für die leere Sequenz steht.

2)$\nu_3 : Shapebaum \rightarrow Shapestring$
Hier benutzen wir auch den Homomorphismus ν_3. ε steht für die leere Sequenz.

$$\begin{array}{lcll} \nu_3(CL) & = & [\,] & \\ \nu_3(FK(c)) & = & [\nu_3(c)] & \\ \nu_3(AD(x,c)) & = & \nu_3(x)\nu_3(c) & (52) \\ \nu_3(CC(x)) & = & [\nu_3(x)] & \\ \nu_3(E) & = & \varepsilon & \end{array}$$

Wir wenden die Gleichungen(52) auf den Shapebaum $\nu_3(S)$ (Abbildung 14) an, um den Shapestring herzuleiten.

$$
\begin{aligned}
\nu_3(\pi_3(S)) &= \nu_3\big(AD(FK(AD(CCIL(FK(AD(AD(E,CL),AD(CL,E)))),CL)),E)\big)\\
&= \nu_3(FK(AD(CCIL(FK(AD(AD(E,CL),AD(CL,E)))),CL)))\nu_5(E)\\
&= [\nu_3(AD(CCIL(FK(AD(AD(E,CL),AD(CL,E)))),CL))]\ \varepsilon\\
&= [\nu_3(CCIL(FK(AD(AD(E,CL),AD(CL,E)))))\nu_3(CL)] \qquad (53)\\
&= [\ [\nu_3(FK(AD(AD(E,CL),AD(CL,E))))\ ]\ [\ ]\ ]\\
&= [\ [\ [\nu_3(E)\nu_3(CL)\nu_3(CL)\nu_3(E)\ ]\ ]\ [\ ]\ ]\\
&= [\ [\ [\ [\ ]\ [\ ]\ ]\ ]\ [\ ]\ ]
\end{aligned}
$$

Das enspricht dem Shape-Typ 3 für die gegebene Sequenz in Abbildung 6.

Jetzt möchten wir die Grammatik für den Typ 3 in Abbildung 3 definieren. Wir brauchen dafür:

1. Beschreibe Strukturen durch Baum-Grammatik $\mathcal{B}$.
2. Finde die Grammatik $\mathcal{G}$ durch die Anwendung der Verkettung $(\nu_3 \circ \pi_3)$ auf Regeln von $\mathcal{B}$.

Mit den gleichen Herleitungsmethoden für die Grammatik des Shape-Typs 5 in (4.2) bekommen wir die Grammatik des Typs 3

$$
\nu_3\big(\pi_3(struct)\big) \rightarrow \begin{cases} \nu_3\big(\pi_3(strong)\big) \\ \nu_3\big(\pi_3(strong)\big).\nu_3\big(\pi_3(comps)\big) \\ \varepsilon \end{cases} \qquad (54)
$$

$$
\nu_3\big(\pi_3(comps)\big) \rightarrow \begin{cases} \nu_3\big(\pi_3(strong)\big) \\ \nu_3\big(\pi_3(strong)\big).\nu_3\big(\pi_3(comps)\big) \end{cases} \qquad (55)
$$

$$
\nu_3\big(\pi_3(strong)\big) \rightarrow \nu_3\big(\pi_3(weak)\big) \qquad (56)
$$

$$
\nu_3\big(\pi_3(weak)\big) \rightarrow \begin{cases} [\ ] \\ [\nu_3\big(\pi_3(strong)\big)] \\ [\nu_3\big(\pi_3(block)\big).\nu_3\big(\pi_3(comps)\big)] \end{cases} \qquad (57)
$$

$$
\nu_3\big(\pi_3(block)\big) \rightarrow \nu_3\big(\pi_5(strong)\big) \qquad (58)
$$

Wir setzen für:

$$
\begin{aligned}
\nu_3\big(\pi_1(struct)\big) &:= S_3\\
\nu_3\big(\pi_1(block)\big) &:= B\\
\nu_3\big(\pi_1(strong)\big) &:= P\\
\nu_3\big(\pi_1(comps)\big) &:= C
\end{aligned}
$$

Wir möchten nun die Grammatik für den Typ 3 aus $(54-58)$ herleiten.

$$\begin{array}{rll} & Aus\ (54): & S_3 \longrightarrow \varepsilon \mid P \mid PC \\ G_1: & Aus\ (56,57,58): & P \longrightarrow [\,] \mid [P] \mid [PC] \\ & Aus\ (55): & C \longrightarrow P \mid PC \end{array}$$

Mit Ausnahme des ε sind S_3 und C identisch. Wir können C durch S_3 ersetzen, weil:
$P\varepsilon = P\ und\ [P\varepsilon] = [P]$.

$$\begin{array}{lll} S_3 & \longrightarrow & \varepsilon \mid P \mid PS_3 \\ P & \longrightarrow & [\,] \mid [P] \mid [PS_3] \end{array}$$

Wenn $S_3 \to \varepsilon$, dann $PS_3 \to P$, in diesem Fall können wir diese Grammatik so vereinfachen:

$$\begin{array}{lll} S_3 & \longrightarrow & \varepsilon \mid PS_3 \\ P & \longrightarrow & [\,] \mid [PS_3] \end{array}$$

Die geschlossene Klammer enthält entweder (ε) und ist somit leer, oder enthält PS_3. Dies ist genau S_3.
$P \to [\,] \mid [PS_3] \Rightarrow P \to [\varepsilon \mid PS_3] \Rightarrow P \to [S_3]$

$$G_2: \begin{array}{lll} S_3 & \longrightarrow & \varepsilon \mid PS_3 \\ P & \longrightarrow & [S_3] \end{array}$$

Wir müssen $\mathcal{L}(G_1) = \mathcal{L}(G_2)$ beweisen.
Beweis:

$$\begin{array}{lll} (a). & w \in \mathcal{L}(G_1) & \Rightarrow \quad w \in \mathcal{L}(G_2) \\ (b). & w \in \mathcal{L}(G_2) & \Rightarrow \quad w \in \mathcal{L}(G_1) \end{array}$$

$a)$

G_1			G_2		
S_3	$\longrightarrow$	P	S_3	$\longrightarrow$	$PS_3 \to P$
S_3	$\longrightarrow$	ε	S_3	$\longrightarrow$	ε
S_3	$\longrightarrow$	PC	S_3	$\longrightarrow$	PS_3
P	$\longrightarrow$	$[\,]$	P	$\longrightarrow$	$[S_3] \to [\,]$
P	$\longrightarrow$	$[P]$	P	$\longrightarrow$	$[S_3] \to [PS_3] \to$ [P]
P	$\longrightarrow$		P	$\longrightarrow$	$[S_3] \to [PS_3]$
C	$\longrightarrow$	P	S_3	$\longrightarrow$	$PS_3 \to P$
C	$\longrightarrow$	PC	S_3	$\longrightarrow$	PS_3

$\Rightarrow \forall w \in \mathcal{L}(G_1) \Rightarrow w \in \mathcal{L}(G_2)$

$b)$

	G_1				G_2	
	S_3	$\longrightarrow$	ε		S_3	$\longrightarrow$ ε
	S_3	$\longrightarrow$	PC		S_3	$\longrightarrow$ PS_3
$Fall_1:$	S_3	$\longrightarrow$	P	$Fall_1:$	S_3	$\longrightarrow$ $PS_3 \rightarrow P$
$Fall_2:$	S_3	$\longrightarrow$	$PC \rightarrow PPC$	$Fall_2:$	S_3	$\longrightarrow$ PPS_3
					P	$\longrightarrow$ $[S_3]$
	P	$\longrightarrow$	$[\,]$	$Fall_1:$	P	$\longrightarrow$ $[S_3] \rightarrow [\,]$
	P	$\longrightarrow$		$Fall_2:$	P	$\longrightarrow$ $[S_3] \rightarrow [PS_3]$

$\Rightarrow \forall w \in \mathcal{L}(G_2) \Rightarrow w \in \mathcal{L}(G_1)$
$\Rightarrow \mathcal{L}(G_1) = \mathcal{L}(G_2)$

Aus dieser Grammatik vom Typ 3:

$$\begin{array}{lcl} S_3 & \longrightarrow & \varepsilon \mid PS_3 \\ P & \longrightarrow & [S_3] \end{array}$$

können wir die Formel $S_3(n)$: Anzahl der Shapes mit n Klammerpaaren ableiten:
Wenn wir keine Klammerpaare haben, haben wir eine Shape vom Typ "_", dann folgt $S_3(0) = 1$.

$$\begin{array}{lcll} S_3(0) & = & 1 & : \text{keine Klammerpaare} \\ S_3(n) & = & \displaystyle\sum_{i=0}^{n-1} S_3(i)P(n-i) & : \text{mit } n \text{ Klammerpaaren} \end{array}$$

$P \rightarrow [S_3]$, $S_3(n-1)$ ergibt sich, weil das äußere Klammerpaar einem Klammerpaar entspricht, bleiben somit $(n-1)$ Klammerpaare.

$$\begin{array}{lcl} P(0) & = & 0 \\ P(n) & = & S_3(n-1) \end{array}$$

Mit P in S_3 eingesetzt, bekommen wir:

$$S_3(n) = \begin{cases} 1 & : n = 0 \\ \displaystyle\sum_{i=0}^{n-1} S_3(i) * S_3(n-i-1) & : sonst \end{cases}$$

Diese Formel haben wir in (3.2) bewiesen.

□

5.3.1 Die asymptotische Form $S_3(n)$: Anzahl der Shapes mit n Klammerpaaren

Wir bestimmen die asymptotische Anzahl für die Anzahl der Shape-Typen 3 mit n Klammerpaaren. Das geschieht mit Hilfe der erzeugenden Funktionen und der Anwendung des Theorems von Flajolet-Odlyzko. Wir haben:

$$S_3(n) \quad = \begin{cases} 1 & : n = 0 \\ \sum_{i=0}^{n-1} S_3(i) * S_3(n-i-1) & : sonst \end{cases}$$

Als erstes berechnen wir die erzeugende Funktion von $S_3(n)$. Für die erzeugende Funktion (EZF) gilt die folgende Definition:

Definition: Für eine Folge $f_n, n \in \mathbb{N}_0$ heißt die formale Potenzreihe $f(z) = \sum_{n=0}^{\infty} f_n z^n$ die erzeugende Funktion von f_n.
wir bezeichnen für eine Potenzreihe $f(z)$ den Koeffizienten f_n mit $[z^n]f(z)$.

Angewendet auf $S_3(n)$ lautet unsere erzeugende Funktion:

$$EZF: \; S_3(z) = \sum_{n\geq 0} S_3(n) z^n = 1.z^0 + \underbrace{\sum_{n\geq 1} z^n \sum_{i=0}^{n-1} S_3(i) S_3(n-1-i)}_{z \sum_{n\geq 1} z^{n-1} \sum_{i=0}^{n-1} S_3(i) S_3(n-1-i)} \tag{59}$$

Nach der Formel für die Koeffizienten des Produktes:

$$A_1(z) * B_1(z) = C_1(z) \; ; \quad C_{1_n} = [z^n]C_1(z) = \sum_{k=0}^{n} a_k * b_{n-k}$$

Die Gleichung (59) ergibt sich wie folgt:

$$S_3(z) = 1 + zS_3(z)^2 \Rightarrow zS_3(z)^2 - S_3(z) + 1 = 0$$

Die Wurzeln dieser quadratischen Gleichung sind:

$$S_3(z)_+ = \frac{1+\sqrt{1-4z}}{2z}$$

$$S_3(z)_- = \frac{1-\sqrt{1-4z}}{2z}$$

Wir müssen $S_3(z) = S_3(z)_- = \frac{1-\sqrt{1-4z}}{2z}$ [7] wählen , weil der $\lim\limits_{z\to o} S_3(z)$ nur für die zweite Lösung existiert.

$$\Rightarrow S_3(z) = \frac{1-\sqrt{1-4z}}{2z}$$

$[z^n]S_3(z) = S_3(n) = \frac{1}{n+1}\binom{2n}{n}$ = Catalan-Zahlen [8]

Im vorliegenden Fall ist die dominierende Singularität ρ eine Lösung der Gleichung $1-4z=0$. Wir finden $\rho = \frac{1}{4}$
Frage: $[z^n]S_3(n) \sim$?

$$[z^n]S_3(z) = [z^{n+1}]\big(S_3(z).z\big) = [z^{n+1}](\tfrac{1}{2} - \tfrac{\sqrt{1-4z}}{2}) \overset{n\geq 0}{=} [z^{n+1}](-\tfrac{\sqrt{1-4z}}{2})$$

Um eine explizite Formel für den asymptotischen Wert zu erhalten, verwenden wir wieder das Theorem 1 von Flajolet-Odlyzko.
$\Rightarrow [z^n]S_3(z) = [z^{n+1}](-\frac{\sqrt{1-4z}}{2}) \sim -\frac{1}{2}.\frac{(n+1)^{-\frac{3}{2}}}{\Gamma(-\frac{1}{2})}.4^{n+1};\ \Gamma(-\frac{1}{2}) = -2\sqrt{\pi}$
$\Rightarrow [z^n]S_3(z) \sim -\frac{1}{2}.\frac{(n+1)^{-\frac{3}{2}}}{-2\sqrt{\pi}}.4^n.4 = \frac{4^n}{\sqrt{\pi}}(n+1)^{-\frac{3}{2}} \sim \frac{4^n}{\sqrt{\pi}}n^{-\frac{3}{2}}$

$$S_3(n) = [z^n]S_3(z) \sim \frac{4^n}{\sqrt{\pi}}n^{-\frac{3}{2}} = 0,564189583 * 4^n * n^{-\frac{3}{2}}$$

[7] Beachte, dass die Taylorreihe (Taylor-Entwicklung) von $\sqrt{1-4z}$ über $z=0$ ist: $1-2z-2z^2-4z^3-10z^4-28z^5-84z^6+...$, wobei die Koeffizienten von z negativ sind. Deswegen hat $S_3(z)_-$ ein negatives Vorzeichen vor dem Term $\sqrt{1-4z}$, seine Taylorreihe bei 0 hat positive Koeffizienten für jeden Term z^n, wie für die erzeugende Funktion benötigt.

[8] Die ersten Catalan-Zahlen sind: $1,1,2,5,14,42,132,429,....$ Für die Catalan-Zahlen gilt für $n \geq 0 : C_n^* = \frac{1}{n+1}\binom{2n}{n}$; wobei $\binom{n}{k}$ den Binomialkoeffizienten bezeichnet. Eine Rekursionsformel lautet:$C_{n+1}^* = \sum\limits_{k=0}^{n} C_k^* C_{n-k}^*$ und die erzeugende Funktion ist: $\sum\limits_{n\geq 0} C_n^* z^n = \frac{1-\sqrt{1-4z}}{2z}$

Die letzte Gleichung gibt die asymptotische Anzahl $S_3(n)$ der Shapes mit n Klammerpaaren an.

Wieder ergeben sich beim Vergleich meiner Ergebnissen für $S_3(n)$ mit n Klammerpaaren mit den Ergebnisse von Nebel und Scheid [9] die gleichen Werte für die Asymptotik für $S_3(n)$ der Größe n, wenn wir $n := 2n$ in ihre Ergebnisse einsetzen.
Ergebnis von Nebel und Scheid:
$S_3(n) \sim ((-2)^n + 2^n) \cdot \sqrt{\frac{2}{\pi}} \left(\frac{1}{n}\right) \approx ((-2.)^n + 2.^n) \cdot 0,797885 \cdot n^{-3/2}$
$(-2.)^n + 2.^n = 2 \cdot 2^n \ : n \ gerade$
setze $n := 2n$ ein:

$$S_3 \underbrace{(2n)}_{\text{Größe}} = S_3 \underbrace{(n)}_{KP} \approx 2 \cdot 4^n \cdot 0.797885 \cdot 2^{-3/2} n^{-3/2}$$
$$\approx \underbrace{0,564189 \cdot 4^n \cdot n^{-3/2}}_{unser \ Ergebnis}$$

5.4 Herleitung des Typs 4 Shape-Grammatik

Shape-Typ 4 erhält Hairpin-Loops und die ungepaarte Basen in Multi-Loops und External-Loops.
Wir benutzen die gleichen Daten von (4.2) für die Variablen und für die Knotenmarkierung auch in diesem Abschnitt.

Bei dem Shape-Typ 3 müssen wir die geschlossen Strukturen, Internal-Loops, Multi-Loops, benachbarte Regionen und die leere Struktur unterscheiden. Diese Strukturelemente werden in der Reihenfolge durch die Knoten-Labels $CL, CCIL, FK, AD$ und E repräsentiert.
In dem Gebiet des Shape-Typs 4 müssen wir geschlossene Strukturen, Internal-Loops, Multi-Loops und benachbarte Strukturen unterscheiden. Diese Situationen sind repräsentiert durch die Knoten-Labels $CL, CCIL, FK, AD$ und E.

Die Abstraktions-Abbildung π_4 von Struktur zur Shape wird durch die folgende Gleichung definiert.

1)$\pi_4 : \mathit{Strukturbaum} \rightarrow \mathit{Shapebaum}$

$$\begin{array}{lcl}
\pi_4(SS(l)) & = & E \\
\pi_4\big(HL(a,b,l)\big) & = & CL \\
\pi_4\big(SR(a,x,b)\big) & = & \pi_4(x) \\
\pi_4\big(BL(a,l,x,b)\big) & = & \pi_4(x) \\
\pi_4\big(BR(a,x,l,b)\big) & = & \pi_4(x) \\
\pi_4\big(IL(a,l,x,l',b)\big) & = & CCIL\big(\pi_4(x)\big) \\
\pi_4\big(ML(a,c,b)\big) & = & FK\big(\pi_4(c)\big) \\
\pi_4\big(AD(x,c)\big) & = & AD\big(\pi_4(x),\pi_4(c)\big) \\
\pi_4(E) & = & E
\end{array} \tag{60}$$

Auf Grund der Gleichungen (60) können wir nun den Shapebaum für den Shape-Typ 4 zeichnen (Abbildung 15), dabei stützen wir uns auf den Strukturbaum in Abbildung 5.

$\pi_4(S) =$

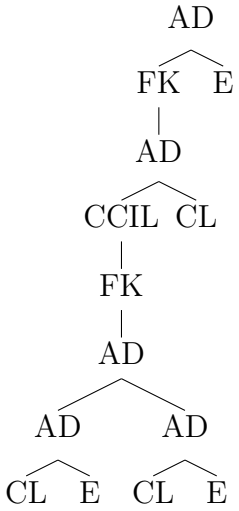


Abbildung 15*: Shapebaum $\pi_4(S)$ des Shape-Typs* 3 *für den Strukturbaum in Abbildung* 5

Um eine Notation für Shapes zu definieren, benutzen wir den Homomorphismus ν_4; wobei ε für die leere Sequenz steht.

2)$\nu_4 : \mathit{Shapebaum} \rightarrow \mathit{Shapestring}$

Hier benutzen wir auch den Homomorphismus ν_4. ε steht für die leere Sequenz.

$$\begin{array}{lcl} \nu_4(CL) & = & [\,] \\ \nu_4(FK(c)) & = & [\nu_4(c)] \\ \nu_4(CC(c)) & = & [\nu_4(x)] \\ \nu_4(AD(x,c)) & = & \nu_4(x)\nu_4(c) \\ \nu_4(E) & = & \varepsilon \end{array} \tag{61}$$

Wir wenden die Gleichungen(61) auf den Shapebaum $\pi_4(S)$ (Abbildung 15) an, so bekommen wir den gleichen Shapestring wie beim Typ 3 in (53).

Im folgenden definieren wir die Grammatik für den Shape-Typ 4.
Wir brauchen dafür:

1. Strukturen, die die Baum-Grammatik $\mathcal{B}$ in Abbildung 3 beschreiben.
2. Grammatik $\mathcal{G}$, die durch die Anwendung der Verkettung $(\nu_4 \circ \pi_4)$ auf Regeln von $\mathcal{B}$, entsteht:

Mit den gleichen Herleitungsmethoden für die Grammatik des Typs 5 in (4.2) bekommen wir die Grammatik des Typs 4.

$$\nu_4\big(\pi_4(struct)\big) \;\rightarrow \left\{ \begin{array}{l} \nu_4\big(\pi_4(comps)\big) \\ \varepsilon \end{array} \right. \tag{62}$$

$$\nu_4\big(\pi_4(comps)\big) \;\rightarrow \left\{ \begin{array}{l} \nu_4\big(\pi_4(block)\big).\nu_2\big(\pi_4(comps)\big) \\ \nu_4\big(\pi_4(block)\big) \end{array} \right. \tag{63}$$

$$\nu_4\big(\pi_4(strong)\big) \;\rightarrow \left\{ \begin{array}{l} \nu_4\big(\pi_4(strong)\big) \\ \nu_4\big(\pi_4(weak)\big) \end{array} \right. \tag{64}$$

$$\nu_4\big(\pi_4(weak)\big) \;\rightarrow \left\{ \begin{array}{l} [\,] \\ \nu_4\big(\pi_4(strong)\big) \\ [\nu_4\big(\pi_4(strong)\big)] \\ [\nu_4\big(\pi_4(block)\big).\nu_4\big(\pi_4(comps)\big)] \end{array} \right. \tag{65}$$

$$\nu_4\big(\pi_4(block)\big) \;\rightarrow \nu_4\big(\pi_4(strong)\big) \tag{66}$$

Wir setzen für:

$$\begin{aligned}
\nu_4\big(\pi_1(struct)\big) &:= S_4 \\
\nu_4\big(\pi_1(block)\big) &:= B \\
\nu_4\big(\pi_1(strong)\big) &:= P \\
\nu_4\big(\pi_1(comps)\big) &:= C
\end{aligned}$$

Wir möchten nun die Grammatik für den Typ 3 aus $(62, 63, 64, 65, 66)$ herleiten.

$$\begin{array}{lllll}
Aus\ (62): & S_4 & \longrightarrow & \varepsilon \mid C \\
Aus\ (63): & C & \longrightarrow & B \mid BC \\
Aus\ (66): & B & \longrightarrow & P \\
Aus\ (64,65): & P & \longrightarrow & [\,] \mid [P] \mid [BC]
\end{array}$$

vereinfacht sieht die Grammatik für Typ 4 so aus:

$$\begin{array}{lll}
S_4 & \longrightarrow & \varepsilon \mid C \\
C & \longrightarrow & P \mid PC \\
P & \longrightarrow & [\,] \mid [P] \mid [PC]
\end{array} \tag{67}$$

Nun können wir aus (67) die Formel $S_4(n)$, die die Anzahl der Shapes mit n Klammerpaaren angibt, herleiten:

$$S_4(n) = \begin{cases} 1 & : n = 0 \\ C(n) & : sonst \end{cases}$$

$$C(n) = \begin{cases} 0 & : n = 0 \\ P(n) + \displaystyle\sum_{i=1}^{n-1} P(i) * C(n-i) & : sonst \end{cases}$$

$$P(n) = \begin{cases} 0 & n = 0 \\ 1 & n = 1 \\ P(n-1) + \displaystyle\sum_{i=1}^{n-2} P(i) * C(n-i-1) & sonst \end{cases}$$

5.4.1 Die asymptotische Form $S_4(n)$: Anzahl der Shapes mit n Klammerpaaren

Wir bestimmen die asymptotische Anzahl der Shapes des Typs 4 mit n Klammerpaaren. Das geschieht mit Hilfe der erzeugenden Funktionen und der Anwendung des Theorems von Flajolet-Odlyzko.

Wir haben:

$$S_4(n) = \begin{cases} 1 & : n = 0 \\ C(n) & : sonst \end{cases} \tag{68}$$

$$C(n) = \begin{cases} 0 & : n = 0 \\ P(n) + \sum_{i=1}^{n-1} P(i) * C(n-i) & : sonst \end{cases} \tag{69}$$

$$P(n) = \begin{cases} 0 & : n = 0 \\ 1 & : n = 1 \\ P(n-1) + \sum_{i=1}^{n-2} P(i) * C(n-i-1) & : sonst \end{cases} \tag{70}$$

Bei einem Vergleich zwischen den Gleichungen (69,70) stellen wir fest, dass:

$$P(n) = C(n-1) \tag{71}$$

Um die Anzahl des Shape-Typs 4 mit n Klammerpaaren zu erhalten, müssen wir die erzeugende Funktion für $S_4(n)$ berechnen, dafür berechnen wir zuerst die erzeugende Funktionen für $P(n)$ und $C(n)$.

Erzeugende Funktion für $P(n)$:

$$P(z) = \sum_{n\geq 0} P(n)z^n = 0.z^0 + 1.z^1 + \underbrace{\sum_{n\geq 2} P(n-1)z^n}_{T_1} + \sum_{n\geq 2} z^n \underbrace{\sum_{i=0}^{n-1} P(i) * C(n-i-1)}_{T_2} \tag{72}$$

Nebenrechnung:

$$\begin{aligned} T_1 &= z\sum_{n\geq 2} P(n-1)z^{n-1} \\ &= z(P(1)z^1 + P(2)z^2 + ...) \\ &= z(P(0)z^0 + P(1)z^1 + P(2)z^2 + ...) \\ &= zP(z) \end{aligned}$$

$$\begin{aligned} T_2 &= z\sum_{n\geq 2} z^{n-1}\sum_{i=0}^{n-1} P(i) * C(n-i-1) \\ &= z(z^1[P(0)C(1) + P(1)C(0)] + z^2[P(0)C(2) + P(1)C(1) + P(2)C(0)] + ...) \\ &= z(z^0[P(0)C(0)] + z^1[P(0)C(1) + P(1)C(0)] + z^2[P(0)C(2) + P(1)C(1) + P(2)C(0)] + ...) \\ &= zP(z)C(z) \end{aligned}$$

Wir setzen die Werte von T_1 und T_2 in die Gleichung (72) ein:

$$P(z) = z + zP(z) + zP(z)C(z) \tag{73}$$

Erzeugende Funktion für $C(n)$:

$$C(z) = \sum_{n\geq 0} C(n)z^n = 0.z^0 + \underbrace{\sum_{n\geq 1} P(n)z^n}_{P(z)} + \underbrace{\sum_{n\geq 1} z^n \sum_{i=0}^{n} P(i) * C(n-i)}_{P(z)C(z)}$$

$$C(z) = P(z) + P(z)C(z) \tag{74}$$

von (74) $\Rightarrow C(z) = \frac{P(z)}{1-P(z)}$ Einsetzen in (73): $\Rightarrow P(z) = z + zP(z) + \frac{zP(z)^2}{1-P(z)}$

$$\Rightarrow P(z)^2 - P(z) + z = 0$$

Die Wurzeln dieser quadratischen Gleischung sind:

$$P(z)_+ = \frac{1+\sqrt{1-4z}}{2}$$

$$P(z)_- = \frac{1-\sqrt{1-4z}}{2}$$

Wir müssen wählen: $P(z) = P_{-}(z) = \frac{1-\sqrt{1-4z}}{2}$, weil der $\lim_{z \to o} P(z)$ nur für - existiert.

$$P(z) = \frac{1 - \sqrt{1-4z}}{2}$$

$[z^n]P(z) = [z^n](\frac{1}{2} - \frac{\sqrt{1-4z}}{2}) \overset{n>0}{=} [z^n](-\frac{1}{2}\sqrt{1-4z})$

Im vorliegenden Fall ist die dominierende Singularität ρ eine Lösung der Gleichung $1 - 4z = 0$. Wir finden $\rho = \frac{1}{4}$.
Wir verwenden wieder das Theorem 1 von Fljolet-Odlyzko wie in (5.3.1), um die explizite Formel für den asymptotischen Wert zu erhalten.

$$[z^n]P(z) \sim (-\tfrac{1}{2})\frac{n^{-\frac{3}{2}}}{\Gamma(-\frac{1}{2})}; \quad \Gamma(-\tfrac{1}{2}) = -2\sqrt{\pi} \; und \; \alpha = \tfrac{1}{2}$$

$$\Rightarrow P(n) = [z^n]P(z) \sim \frac{4^n}{4\sqrt{\pi}}.n^{-\frac{3}{2}}$$

Aus Gleichung (71): $P(n) = C(n-1) \Rightarrow C(n) = P(n+1)$

$$\Rightarrow C(n) \sim \frac{4^{n+1}}{4\sqrt{\pi}}.(n+1)^{-\frac{3}{2}} \sim \frac{4^n}{\sqrt{\pi}}.n^{-\frac{3}{2}}$$

Aus Gleichung (68):$S_4(n) = C(n)$

$$\Rightarrow S_4(n) = C(n) \sim \frac{4^n}{\sqrt{\pi}}.n^{-\frac{3}{2}} = 0,564189583 * 4^n * .n^{-\frac{3}{2}}$$

Beim Vergleich der Shape-Typen 3 und 4 ergeben sich ebenfalls die gleichen Ergebnisse.

Wir zeichnen jetzt den Graphen für die Anzahl der Shapes $S_i(n), i \in \{1...5\}$ mit n Klammerpaaren in ein geeignetes Koordinatensystem.

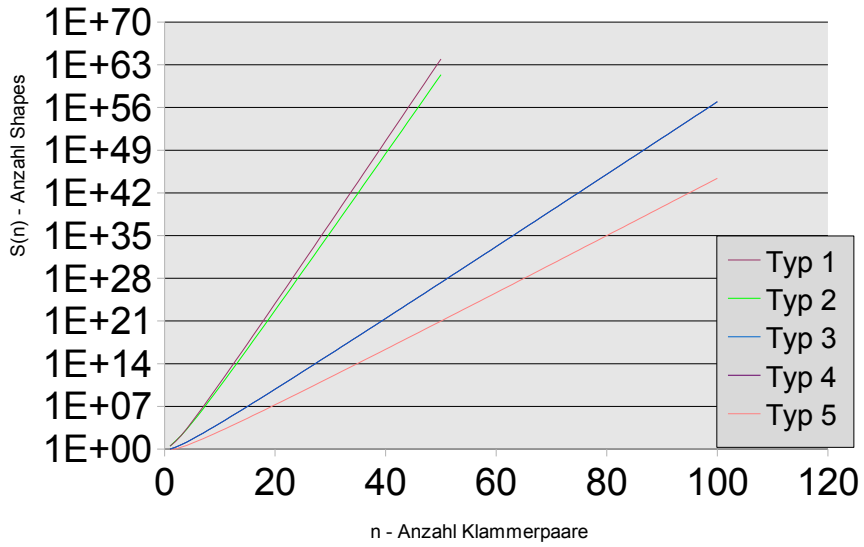


Abbildung 16*: Vergleich der Anzahl der Shape-Typen* i $\left(i \in \{1...5\}\right)$ *mit* n *Klammerpaaren.*

Die Messwerte aus diesem Graph werden in einer Tabelle dargestellt (siehe Anhang).

6 Empirische Untersuchungen zum Erwartungswert der Anzahl der Shapes für den Typ 5

Der asymptotische Wert für die Anzahl der Shapes-Typen i $\left(i \in \{1...5\}\right)$ aller Sequenzen der Länge n [8, 9] hat die allgemeine Formel $f(n) = a^n \cdot b \cdot n^{-3/2}$. Eine solche Formel wurde noch nicht für die Erwartungswertberechnung ermittelt, deswegen machen wir die empirische Untersuchung mit Zufallssequenzen und dem Programm RNAshapes [13].

Wir untersuchen im folgenden nur die Verteilung für den Shape-Typ 5.
Wir haben ein Programm entwickelt, welches für eine große Menge Sequenzen bestimmter Länge die jeweilige Anzahl des Shape-Typs 5 im Shape space bestimmt. Um einen Überblick über die Verteilung zu bekommen, haben wir die Daten in einem Histogramm visualisiert.
Wir werden eine statistische Analyse der Verteilung (Erwartungswert) erhalten, um die Bestimmung der Parameter von der Formel $f(n) = a^n \cdot b \cdot n^{-3/2}$ vorzunehmen. Wir haben in (2.6) den abstrakten Shape-Raum der Sequenz s so definiert: $\mathcal{P}_\pi(s) = \{\pi(x) | x \in \mathcal{F}(s), s \in \{A, C, G, U\}^*\}$.

Dann: $f(n) = a^n \cdot b \cdot n^{-3/2} \widehat{=} E_s(|\mathcal{P}_\pi(s)|)$

```
#!/vol/perl-5.8/bin/perl -w

# Dieses Skript erzeugt eine grosse Menge an Zufallssequenzen und ermittelt die Anzahl
# der Shapes, die mit diesen Sequenzen gefaltet werden koennen.

use strict;

my $n;
my $i;
my $randomseq;
#my @histogramm;


# aussere Schleife gibt die Sequenzlaengen vor
for ($n =20; $n <= 120; $n +=20){
    my %histogramm = ();
    print "Sequenzlaenge: $n\n";
    # $i Anzahl der Zufallssequenzen
    for($i=0; $i<1000; $i++){

        # erzeuge Zufallssequenz der Laenge $n
        $randomseq = `/vol/biotools/bin/randomseq -dna -uniform -min $n -q | tail -1`;
        #newline der randomseq abschneiden
        chomp $randomseq;
        # rufe RNAshapes auf und zaehle Shapes mit wc
        my $num = `/vol/biotools/bin/RNAshapes64 -t 5 -e 1000 $randomseq | wc -l`;
```

```
        #print $num;
        $histogramm{$num-1} += 1;
    }
    # histogramm sortiert ausgeben
    foreach my $key ( sort {$a <=>$ b} keys %histogramm){
        print "$key: $histogramm{$key}\n";
    }
}
```

Aus diesem Programm erhalten wir für die Sequenzlängen 20 und 40 folgende Ergebnisse, die wir in folgender Tabelle aufführen:

Sequenzlänge 20:

X:# Shapes	Y:# Sequenzen
1	3
2	313
3	640
4	44

Sequenzlänge 40:

X:# Shapes	Y:# Sequenzen
3	2
4	13
5	34
6	53
7	69
8	114
9	154
10	132
11	106
12	87
13	74
14	62
15	45
16	34
17	13
18	3
19	3
20	2

Weiterhin werden aus diesem Programm folgende Ergebnisse für die Sequenzlängen $60, 80, ..., 120$ generiert (siehe Anhang)

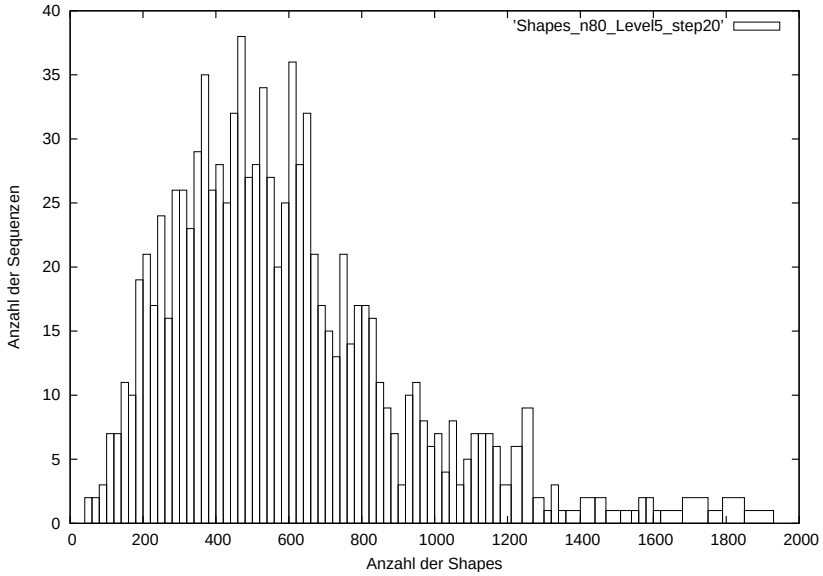


Abbildung 17**:** *Darstellung zwischen Anzahl der Shapes und Sequenzen der Länge* 80*.*

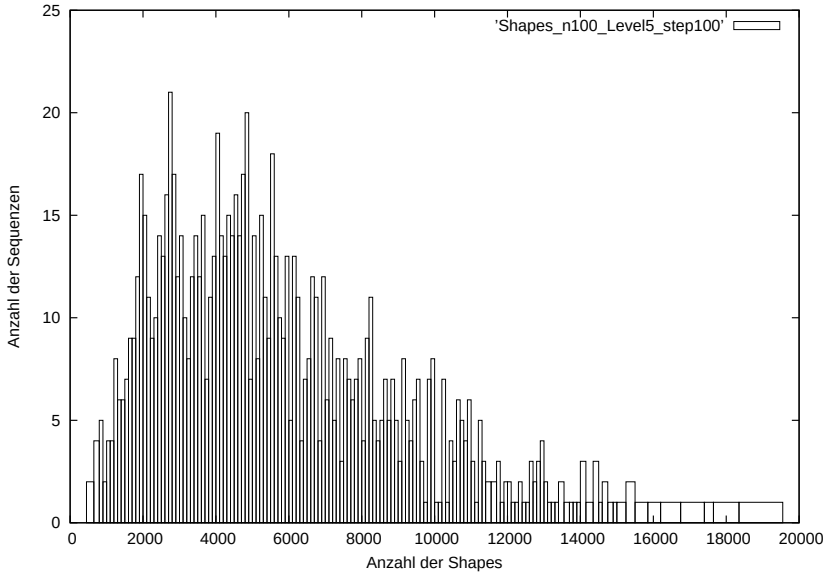


Abbildung 18**:** *Darstellung zwischen Anzahl der Shapes und Sequenzen der Länge* 100*.*

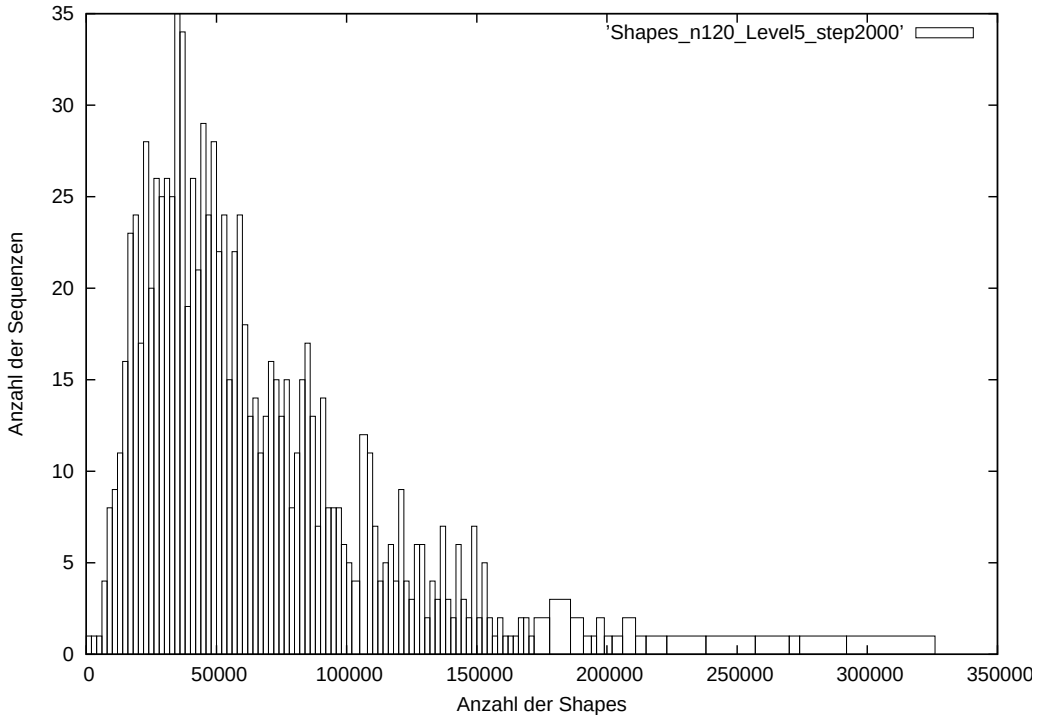


Abbildung 19**:** *Darstellung zwischen Anzahl der Shapes und Sequenzen der Länge* 120*.*

Wir ermitteln den arithmetischen Erwartungswert für die Anzahl der Shapes pro jeweiliger Sequenzlänge, und wir berechnen exemplarisch den Erwartungswert für die Sequenzlänge 20, nach der Formel:

$$f(n) = \frac{\sum_{s} |\mathcal{P}_\pi(s)|}{1000}; n = 20, 40, ..., 120 \tag{75}$$

Wir wenden die Formel (75) für die Sequenzlängen $20, 40, 60, ..., 120$ an, um die jeweiligen Erwartungswerte zu berechnen.
Dadurch bekommen wir folgende Tabelle für die Erwartungswerte:

$\underbrace{n}$ Sequenzlänge	$\underbrace{f(n)}$ Erwartungswert für Shapes$-n-$T$yp5$	$log(f(n))$
20	2,725	1,002
40	10,278	2,330
60	67,77	4,216
80	582,927	6,368
100	5733,398	8,6541
120	63618,943	11,061

Tabelle 5**:** *Berechnung des Erwartungswerts der Anzahl der Shapes Typ* 5 $f(n)$ *für eine Menge Sequenzen bestimmter Länge* $n = 20, 40, 60, 80, 100, 120$ *mit der logarithmischen Basis* $log(f(n))$.

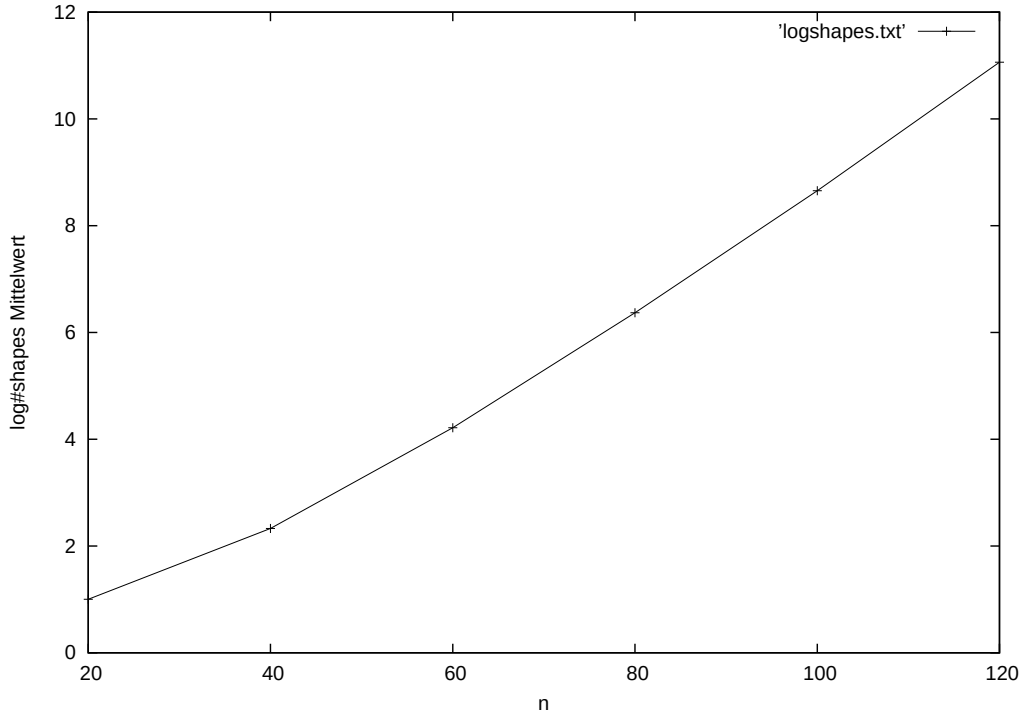


Abbildung 20**:** *Darstellung zwischen Sequenzenlänge n und dem Erwartungswert der Anzahl der Shapes.*

Wir wenden die Formel $f(n) = a^n * b * n^{-3/2}$ auf verschiedene Sequenzlängen n_1, n_2 an, um die Parameter a, b zu bestimmen; wir zerlegen die folgende Tabelle in $5-$Blöcke: $B_{20}, B_{40}, B_{60}, B_{80}, B_{100}$

$$
\begin{array}{lll}
B_{20} : n_1 = 20 & , & n_2 = 40, ..., 120 \\
B_{40} : n_1 = 40 & , & n_2 = 60, ..., 120 \\
B_{60} : n_1 = 60 & , & n_2 = 80, 100, 120 \\
B_{80} : n_1 = 80 & , & n_2 = 100, 120 \\
B_{100} : n_1 = 100 & , & n_2 = 120
\end{array}
$$

(n_1 bleibt in jedem Block konstant, n_2 wächst)

Aus der Tabelle mit den Erwartungswerten können wir nun die Parameter a, b aus der Gleichung $f(n) = a^n * b * n^{-3/2}$ bestimmen:

$$f(n) = a^n * b * n^{-3/2} \longrightarrow a =?, b =?$$

Wir lösen nach a, b auf und erhalten:

$$\log(a) = \frac{\log f(n_2) - \log f(n_1) + 1.5(\log(n_2) - \log(n_1))}{n_2 - n_1}$$

$$\log(b) = \log f(n_1) + 1.5 * \log(n_1) - n_1 \log(a)$$

$$\Longrightarrow a = \log^{-1}(a), b = \log^{-1}(b)$$

$f(n_1)$	$f(n_2)$	n_1	n_2	a	b
2,725	10,278	20	40	1,1256525383209	22,846748443556
2,725	67,77	20	60	1,1292336768637	21,440507924464
2,725	582,927	20	80	1,1321108714490	20,376626399384
2,725	5733,398	20	100	1,1340801719452	19,680510179187
2,725	63618,943	20	120	1,1359375818856	19,046805040968
10,278	67,77	40	60	1,1328262084011	17,720137137091
10,278	582,927	40	80	1,1353539200503	16,208694946876
10,278	5733,398	40	100	1,1369033813958	15,348158429561
10,278	63618,943	40	120	1,1385234928140	14,498355396726
67,77	582,927	60	80	1,1378872718640	13,561570597582
67,77	5733,398	60	100	1,1389474674288	12,824568663560
67,77	63618,943	60	120	1,1404289478857	11,862340995645
582,927	5733,398	80	100	1,1400086508020	11,684157355949
582,927	63618,943	80	120	1,1417019140886	10,376020453121
5733.398	63618.943	100	120	1,1433976923918	8,683222950367

Tabelle 6**:** *Bestimmung der Parameter a, b aus der Gleichung:*
$f(n) = a^n \cdot b \cdot n^{-3/2}$ auf verschiedenen Sequenzlängen n_1, n_2 mit ihrem Er-

wartungswert der Anzahl der Shapes-Typ 5 *(die aus Tabelle* 5 *genommen wurden).*

Wir vergleichen nun mit verschiedenen Exponenten von n. Wir setzen den Exponenten auf die Werte $(-1, -1/2, 0)$, und vergleichen die Werte der Parameter a, b.
Die zweite Annahme:
$f(n) = a^n * b * \underbrace{n^{-3/2}}_{n^{-1}} \Rightarrow f(n) = a^n * b * n^{-1} \longrightarrow a =?, b =?$

$$log(a) = \frac{(\log f(n_2) - \log f(n_1)) + (\log(n_2) - \log(n_1))}{n_2 - n_1}$$

$$\log(b) = \log f(n_1) + \log(n_1) - n_1 \log(a)$$

$f(n_1)$	$f(n_2)$	n_1	n_2	a	b
2,725	10,278	20	40	1,1063145014666	7,2247762210546
2,725	67,77	20	60	1,1138322948976	6,3095790440681
2,725	582,927	20	80	1,1191074690213	5,7406440199402
2,725	5733,398	20	100	1,1227296573381	5,3813669256658
2,725	63618,943	20	120	1,1258063966699	5,0947427935355
10,278	67,77	40	60	1,1214011743608	4,2026995352807
10,278	582,927	40	80	1,1255593210831	3,6243743727773
10,278	5733,398	40	100	1,1282553237155	3,2936158760068
10,278	63618,943	40	120	1,1307327957827	3,0169565766664
67,77	582,927	60	80	1,1297328861807	2,6955194433650
67,77	5733,398	60	100	1,1316980924694	2,4285872136882
67,77	63618,943	60	120	1,1338605604796	2,1657572493777
582,927	5733,398	80	100	1,1336667172969	2,0411382896647
582,927	63618,943	80	120	1,1359300496292	1,7401115301831
5733.398	63618,943	100	120	1,1381979006381	1,3697271595477

Tabelle 7**:** *Bestimmung der Parameter* a, b *aus der Gleichung:* $f(n) = a^n \cdot b \cdot n^{-1}$ *auf verschiedenen Sequenzlängen* n_1, n_2 *mit ihrem Erwartungswert der Anzahl der Shapes-Typ* 5 *(die aus Tabelle* 5 *genommen wurden).*

Die dritte Annahme:

$$f(n) = a^n * b * \underbrace{n^{-3/2}}_{n^{-1/2}} \Rightarrow f(n) = a^n * b * n^{-1/2} \longrightarrow a =?, b =?$$

$$log(a) = \frac{(logf(n_2) - logf(n_1)) + 0.5(log(n_2) - log(n_1))}{n_2 - n_1}$$

$$log(b) = logf(n_1) + 0.5 * log(n_1) - n_1 log(a)$$

$f(n_1)$	$f(n_2)$	n_1	n_2	a	b
2,725	10,278	20	40	1,087308680510	2,28467484435569
2,725	67,77	20	60	1,098640969159	1,85680245326274
2,725	582,927	20	80	1,106253423409	1,61729390909729
2,725	5733,398	20	100	1,111492745089	1,47146134551301
2,725	63618,943	20	120	1,115765569336	1,36276945537331
10,278	67,77	40	60	1,110091366647	0,99675771396137
10,278	582,927	40	80	1,115849219264	0,81043474734383
10,278	5733,398	40	100	1,119673049023	0,70678873875782
10,278	63618,943	40	120	1,122995408991	0,62779720433299
67,77	582,927	60	80	1,121636936871	0,53576575200325
67,77	5733,398	60	100	1,124494859618	0,45990130422466
67,77	63618,943	60	120	1,127330004201	0,39541136652151
582,927	5733,398	80	100	1,127360064331	0,35657218493499
582,927	63618,943	80	120	1,130187364782	0,29182557524404
5733,398	63618.943	100	120	1,133021755804	0,21606637331858

Tabelle 8**:** *Bestimmung der Parameter a, b aus der Gleichung: $f(n) = a^n \cdot b \cdot n^{-1/2}$ auf verschiedenen Sequenzlängen n_1, n_2 mit ihrem Erwartungswert der Anzahl der Shapes-Typ* 5 *(die aus Tabelle* 5 *genommen wurden).*

Die vierte Annahme:
$f(n) = a^n * b * \underbrace{n^{-3/2}}_{n^0} \Rightarrow f(n) = a^n * b \longrightarrow a =?, b =?$

$$\log(a) = \frac{\log f(n_2) - \log f(n_1)}{n_2 - n_1}$$

$$\log(b) = \log f(n_1) - n_1 \log(a)$$

$f(n_1)$	$f(n_2)$	n_1	n_2	a	b
2,725	10,278	20	40	1,0686293681829	0,72247762210546
2,725	67,77	20	60	1,0836568347359	0,54642557393489
2,725	582,927	20	80	1,0935470191034	0,45563521781140
2,725	5733,398	20	100	1,1003682982017	0,40235102367993
2,725	63618,943	20	120	1,1058142940016	0,36452097068666
10,278	67,77	40	60	1,0988956231539	0,23640184885954
10,278	582,927	40	80	1,1062228856451	0,18121871863886
10,278	5733,398	40	100	1,1111560569294	0,15167230789539
10,278	63618,943	40	120	1,1153109676472	0,13063805187538
67,77	582,927	60	80	1,1135990051655	0,10648965702182
67,77	5733,398	60	100	1,1173374751817	0,08709146142062
67,77	63618,943	60	120	1,1208370611593	0,07219190831259
582,927	5733,398	80	100	1,1210884956385	0,06229059721877
582,927	63618,943	80	120	1,1244737120302	0,04894063678639
5733.398	63618,943	100	120	1,1278691503536	0,03408319485645

Tabelle 9**:** *Bestimmung der Parameter a, b aus der Gleichung: $f(n) = a^n \cdot b \cdot n^0$ auf verschiedenen Sequenzlängen n_1, n_2 mit ihrem Erwartungswert der Anzahl der Shapes-Typ* 5 *(die aus Tabelle* 5 *genommen wurden).*

Wir sehen, dass die Werte von dem Parameter a in jedem einzelnen Block monoton wachsen, und die Werte für b in jedem einzelnen Block monoton fallen. Die Werte von Parameter b werden kleiner, wenn der Exponent von n größer wird.

Wir möchten nun den Wert für die Parameter a bestimmen, da der Exponent n nur bei a vorkommt.
Als erstes brauchen wir einen Schätzwert für die Standardabweichung für den Parameter a aus unserer Stichprobe vom Umfang $n = 15$.
Wir streben einen möglichst erwartungstreuen Schätzer (s) für die Standardabweichung (σ) an.
Statistisch berechnen wir die Varianz s^2 durch [1]:

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (a_i - \bar{a})^2 \tag{76}$$

wobei

$$\bar{a} = \frac{1}{n} \sum_{i=1}^{n} a_i.$$

Wir wenden die Gleichung (76) auf alle Werte für den Parameter a in den letzten vier Tabellen an und erhalten für s^2 jeweils folgendes:

$$\begin{array}{lll} f(n) = a^n \cdot b \cdot n^{-3/2} & : s^2 & = 2,159875133570e-05 \\ f(n) = a^n \cdot b \cdot n^{-1} & : s^2 & = 6,970418419126e-05 \\ f(n) = a^n \cdot b \cdot n^{-1/2} & : s^2 & = 1,557051311287e-04 \\ f(n) = a^n \cdot b \cdot n^{0} & : s^2 & = 2,626037916835e-04 \end{array}$$

Wir erhalten den kleinsten Wert von s^2 für a, wenn wir in der Formel $f(n) = a^n \cdot b \cdot n^{-3/2}$, $n^{-3/2}$ beibehalten, da wir für $n^{-1}, n^{-1/2}, n^0$ größere Werte für s^2 bekommen haben.
Für $n^{-3/2}$ in $f(n) = a^n \cdot b \cdot n^{-3/2}$ bekommen wir: $s^2 = 2,15987513357017e-05$

Danach berechnen wir die exakten Werte für a und b mit Hilfe der Methode der kleinsten Quadrate [7].

Mithilfe der Methode der kleinsten Quadrate bestimmen wir nun die Gleichung der Ausgleichsgeraden, um die allgemeinen Werte bei a und b funktional zu ermitteln.
Im folgenden wird die Methode der kleinsten Quadrate einer Ausgleichsgeraden (Regressionsgerade) kurz ausgeführt:
Im zweidimensionalen Fall lautet die Gleichung der Ausgleichsgeraden:
$y(t) = x_0 + x_1 t$.
Man erhält in Matrixschreibweise:

$$\min_{x_0,x_1} \left\| \begin{pmatrix} 1 & t_1 \\ . & . \\ . & . \\ . & . \\ 1 & t_n \end{pmatrix} \begin{pmatrix} x_0 \\ x_1 \end{pmatrix} - \begin{pmatrix} y_1 \\ . \\ . \\ . \\ y_n \end{pmatrix} \right\|_2 = \min_x \|Ax - b\|_2 \qquad (77)$$

Für die resultierende Ausgleichsgerade dieses einfachen, aber relevanten Beispiels lassen sich die Lösungen für die Parameter direkt angeben als:

$$x_1 = \frac{(\sum_{i=1}^{n} t_i y_i) - n.\bar{t}.\bar{y}}{(\sum_{i=1}^{n} t_i^2) - n.(\bar{t})^2}, \quad x_0 = \bar{y} - x_1\bar{t}$$

mit $\bar{t} = \frac{1}{n}\sum_{i=1}^{n} t$ als arithmetischem Mittel der t-Werte, $\bar{y}$ entsprechend.

Für meine Untersuchung werden fünf Sequenzen unterschiedlicher Längen, sowie die entsprechende erwartete Anzahl der Shapes (logarithmisch berechnet) eingesetzt.

$f(n) = a^n * b * n^{-3/2} \longrightarrow a =?, b =?$

$$n^{3/2} * f(n) = a^n * b \xrightarrow{\log} \underbrace{\log(n^{3/2} * f(n))}_{y(n)} = \underbrace{\log(b)}_{\tilde{b}} + n * \underbrace{\log(a)}_{\tilde{a}}$$

Wir setzen in (77) x_1 und x_0 durch $\tilde{a}$ und $\tilde{b}$, dann ergibt die Gleichung
$y(n) = \tilde{b} + n * \tilde{a}$
In der folgenden Tabelle werden alle benötigten Punkte eingetragen:

Nummer i	Sequenzlänge n	Erwartung für Shapes_n $f(n)$	$\underbrace{\log(n^{3/2} * f(n))}_{y}$	$n * y$	n^2
1	40	10,278	7,86332487	314,5329948	1600
2	60	67,77	10,35763646	621,4581877	3600
3	80	582,927	12,94110192	1035,288154	6400
4	100	5733,398	15,56181893	1556,181893	10000
5	120	63618,943	18,24190471	2189,0285	14400
$\sum$	400	70013,316	64,96578635	5716,489729	36000

Nun setzen wir die Punkte der Untersuchung in (77) ein:

$$\min_{\tilde{b},\tilde{a}} \left\| \begin{pmatrix} 1 & 40 \\ 1 & 60 \\ 1 & 80 \\ 1 & 100 \\ 1 & 120 \end{pmatrix} \begin{pmatrix} \tilde{b} \\ \tilde{a} \end{pmatrix} - \begin{pmatrix} 7,86332487 \\ 10,35763646 \\ 12,94110192 \\ 15,56181893 \\ 18,24190471 \end{pmatrix} \right\|_2$$

Man erhält nun analog zum oben angegebenen Fall zunächst:

$$\bar{t} = \bar{n} = \tfrac{1}{n}\sum_{i=1}^{5} n = \frac{1}{5}(40 + 60 + 80 + 100 + 120) \longrightarrow \bar{n} = 80$$

und entsprechend $\bar{y} = \tfrac{1}{n}\sum_{i=1}^{5} y = \frac{1}{5}(7,86332487+10,35763646+12,94110192+$
$15,56181893 + 18,24190471) \longrightarrow \bar{y} = 12,99315727$

$$\tilde{a} = \log(a) = \frac{(\sum_{i=1}^{5} n_i y_i) - n.\bar{n}.\bar{y}}{(\sum_{i=1}^{5} n_i^2) - n.(\bar{n})^2} = 0,129806705 \Rightarrow a = 1,13860064$$

$$\tilde{b} = \log(b) = \bar{y} - \tilde{a}\bar{n} = 12,99315727 - 12980705 * 80 = 2,608620841$$
$$\Rightarrow b = 13,58030853$$

Als Ergebnis der Berechnung der Ausgleichsgeraden unserer Untersuchung wird folgendes gezeigt: Die Anzahl der Shapes wächst monoton bezüglich der Anzahl der Sequenzlängen in der Größenordnung von $1,13860064$ (auf logarithmischer Basis $\log(n^{3/2} * f(n))$) berechnet.

7 Zusammenfassung und Schluss

In dieser Arbeit haben wir einige Ergebnisse analytischer und empirischer Untersuchungen über abstrakte Shapes von RNA-Sekundärstrukturen abgeleitet.

Wir haben die Anzahl der Shapes $S_i(n)$ mit n Klammerpaaren, die Anzahl der Shapes $L_i(n)$ der Länge n und die Anzahl der Shapes $R_i(n)$ zu Sequenzen der Länge $\leq n$ $\big(i \in \{1...5\}\big)$ definiert. Danach haben wir die Rekurrenzformel für $S_i(n)$ und $L_i(n), i = 3$ (Shape-Typ 3) bestimmt. In den darauffolgenden Kapiteln haben wir die Rekurrenzformel durch die Shape-Grammatik hergeleitet.

Weiterhin haben wir die Rekurrenzformel für die Anzahl der Shapes $R_5(n)$ zu Sequenzen der Länge $\leq n$ mit ihrem asymptotischen Wert berechnet.

Danach haben wir die Formel für die Anzahl der Shape-Typen $i\big(i \in \{1...5\}\big)$ mit n Klammerpaar durch die Shape-Grammatik hergeleitet, die von der Abstraktions-Abbildung $\pi_i\big(i \in \{1...5\}\big)$ und vom Homomorphismus $\nu_i\big(i \in \{1...5\}\big)$ hergeleitet wird. Weiterhin haben wir den asymptotischen Wert der Shape-Typen $i(i = 3, 4, 5)$ mit n Klammerpaaren berechnet.

Um die asymptotische Anzahl zu bekommen, haben wir die erzeugende Funktione ermittelt und ein Theorem von Flajolet-Odlyszko angewendet.

Der asymptotische Wert für die Anzahl der Shape-Typen $i\big(i \in \{1...5\}\big)$ aller Sequenzen der Länge n hat die allgemeine Formel $f(n) = a^n \cdot b \cdot n^{-3/2}$ und diese Formel war nicht bekannt für die Erwartungswertberechnung. Da der Shape-Typ 5 eine große Rolle spielt, haben wir eine emiprische Untersuchung mit Zufallssequenzen und dem Programm RNAshapes durchgeführt. Danach haben wir eine statistische Analyse (Erwartungswert) erhalten, um die Parameter der Formel $f(n) = a^n \cdot b \cdot n^{-3/2}$ zu ermitteln, wobei $f(n) \widehat{=} E_s(|\mathcal{P}(s)|) < |\mathcal{P}(|s|)|$

Danach haben wir den Exponenten n auf die Werte $(-1, -1/2, 0)$ gesetzt und die daraus resultiereneden Werte für a verglichen.
Im folgenden haben wir einen möglichst erwartungstreuen Schätzer für die

Standardabweichung angewendet. Mit der Varianz s^2 für alle Wert für den Parameter a in den vier Formeln haben wir den kleinsten Wert von s^2 für a in der Formel $f(n) = a^n \cdot b \cdot n^{-3/2}$ erhalten.

Eine mathematisch exakte Darstellung von $E(R(n))$ bleibt damit weiter offen. Allerdings haben unsere empirischen Untersuchungen gezeigt, dass die Varianz recht hoch ist, so dass der Erwartungswert mehr theoretische Bedeutung hat als praktische Relevanz für die Laufzeit der Algorithmen zur Shape-Analyse.

Ein weiteres offenes Problem bleibt die maximale Anzahl von Shapes zu einer Sequenz der Länge n. Sie wäre wichtig für eine Bestimmung der worst-case Laufzeit. Diese Frage dürfte sich als schwierig erweisen, wenn man bedenkt: Betrachtet für Shape-Typ 0, also für konkrete Strukturen, ist diese Frage bereits mehr als zwei Jahrzehnte offen geblieben.

Literatur

[1] Wikipedia: Variance. http://en.wikipedia.org/wiki/Variance, 2010.

[2] Hans-Joachim Böckenhauer and Dirk Bongartz. *Algorithmische Grundlagen der Bioinformatik: Modelle, Methoden und Komplexität*. Teubner, 2003.

[3] Philippe Flajolet and Robert Sedgewick. *Analytic Combinatorics*. Cambridge University Press, 2009.

[4] Robert Giegerich, Björn Voß, and Marc Rehmsmeier. Abstract shapes of RNA. *Nucleic Acids Research*, 32(16):4843, September 2004.

[5] Ivo L. Hofacker, Peter Schuster, and Peter F. Stadler. Combinatorics of RNA secondary structures. *Discrete Applied Mathematics*, 88(Issues 1-3):207–237, 9 November 1998.

[6] S. Janssen, J. Reeder, and R. Giegerich. Shape based indexing for faster search of RNA family databases. *BMC Bioinformatics*, 9:131, 2008.

[7] Wolfgang Kohn. *Statistik: Datenanalysis und Wahrscheinlichkeitsrechnung*. Springer-Verlag Berlin Heidelberg, 2005.

[8] W. A. Lorenz, Yann Ponty, and Peter Clote. Asymptotics of RNA shapes. *Journal of Computational Biology*, 15(1):31–63, 2008.

[9] Markus E. Nebel and Anika Scheid. On quantitative effects of RNA shape abstraction. *Theory in Biosciences*, 128(4):211–225, 2009.

[10] Jens Reeder. *Algorithms for RNA Secondary Structure Analysis: Prediction of Pseudoknots and the Consensus Shapes Approach*. PhD thesis, Universität Bielefeld, 2007.

[11] Jens Reeder and Robert Giegerich. Consensus shapes: an alternative to the sankoff algorithm for RNA consensus structure prediction. *Bioinformatics*, 21(17 2005):3516–3523, 2005.

[12] Peter Steffen and Robert Giegerich. Versatile and declarative dynamic programming using pair algebras. *BMC Bioinformatics*, 6(1):224, September 2005.

[13] Peter Steffen, Björn Voß, Marc Rehmsmeier, Jens Reeder, and Robert Giegerich. RNAshapes: an integrated RNA analysis package based on abstract shapes. *Bioinformatics*, 22(2006):500–503, 2005.

[14] Peter Steffen, Björn Voß, Marc Rehmsmeier, Jens Reeder, and Robert Giegerich. RNAshapes 2.1.6 manual. http://bibiserv.techfak.uni-bielefeld.de/rnashapes/manual.html, 2006.

[15] Gerhard Steger. *Bioinformatik, Methoden zur Vorhersage von RNA- und Proteinstrukturen*. Birkhäuser Verlag, Basel-Boston-Berlin, 2003.

[16] Björn Voß, Robert Giegerich, and Marc Rehmsmeier. Complete probabilistic analysis of RNA shapes. *BMC Biology*, 4(1):5, February 2006.

[17] M. S. Waterman. *Introduction to Computational Biology*. Chapman & Hall, 1995.

[18] Herbert S. Wilf. *generatingfunctionology*. Academic Press, Inc., Second Edition, 1994.

Anhang

Anhang A: Die Messwerte des Graph in Abbildung 16 für die Anzahl der Shape-Typen i $S_i(n), i \in \{1...5\}$ mit n Klammerpaaren.

n	$S_1(n)$	$S_2(n)$	$S_3(n)$	$S_4(n)$	$S_5(n)$
1	3	3	1	1	1
2	18	21	2	2	1
3	164	171	5	5	2
4	2111	1771	14	14	4
5	31998	21368	42	42	9
6	528563	289449	132	132	21
7	9219873	4233738	429	429	51
8	167125422	65315109	1430	1430	127
9	3117794350	1047039807	4862	4862	323
10	59478678671	17277944942	16796	16796	835
11	1,15516E+12	2,91674E+11	58786	58786	2188
12	2,27656E+13	5,01526E+12	208012	208012	5798
13	4,54163E+14	8,75597E+13	742900	742900	15511
14	9,15432E+15	1,54842E+15	2674440	2674440	41835
15	1,86158E+17	2,76848E+16	9694845	9694845	113634
16	3,81474E+18	4,99708E+17	35357670	35357670	310572
17	7,86974E+19	9,09488E+18	129644790	129644790	853467
18	1,63314E+21	1,66746E+20	477638700	477638700	2356779
19	3,40699E+22	3,07707E+21	1767263190	1767263190	6536382
20	7,141E+23	5,71143E+22	6564120420	6564120420	18199284
21	1,50309E+25	1,06567E+24	24466267020	24466267020	50852019
22	3,17594E+26	1,99784E+25	91482563640	91482563640	142547559
23	6,73395E+27	3,76152E+26	3,4306E+11	3,4306E+11	400763223
24	1,43234E+29	7,11007E+27	1,2899E+12	1,2899E+12	1129760415
25	3,05552E+30	1,3488E+29	4,86195E+12	4,86195E+12	3192727797
26	6,53561E+31	2,5672E+30	1,83674E+13	1,83674E+13	9043402501
27	1,40139E+33	4,90119E+31	6,95336E+13	6,95336E+13	25669818476
28	3,01179E+34	9,38365E+32	2,63748E+14	2,63748E+14	73007772802
29	6,48651E+35	1,80129E+34	1,00224E+15	1,00224E+15	2,08023E+11
30	1,39976E+37	3,46626E+35	3,81499E+15	3,81499E+15	5,93743E+11
31	3,02621E+38	6,68543E+36	1,45446E+16	1,45446E+16	1,69739E+12
32	6,55376E+39	1,29219E+38	5,55341E+16	5,55341E+16	4,85976E+12

n	$S_1(n)$	$S_2(n)$	$S_3(n)$	$S_4(n)$	$S_5(n)$
33	1,42162E+41	2,50262E+39	2,12336E+17	2,12336E+17	1,39336E+13
34	3,08841E+42	4,85601E+40	8,12944E+17	8,12944E+17	4,00025E+13
35	6,71903E+43	9,4392E+41	3,11629E+18	3,11629E+18	1,14989E+14
36	1,46373E+45	1,83788E+43	1,19598E+19	1,19598E+19	3,30931E+14
37	3,19277E+46	3,58413E+44	4,59508E+19	4,59508E+19	9,53468E+14
38	6,97262E+47	7,00003E+45	1,76734E+20	1,76734E+20	2,75002E+15
39	1,52447E+49	1,36908E+47	6,80425E+20	6,80425E+20	7,93966E+15
40	3,33663E+50	2,68129E+48	2,62213E+21	2,62213E+21	2,29447E+16
41	7,31044E+51	5,25788E+49	1,01139E+22	1,01139E+22	6,63682E+16
42	1,60325E+53	1,0323E+51	3,90444E+22	3,90444E+22	1,92138E+17
43	3,51935E+54	2,02909E+52	1,50853E+23	1,50853E+23	5,56705E+17
44	7,73228E+55	3,99279E+53	5,833E+23	5,833E+23	1,61428E+18
45	1,70028E+57	7,86518E+54	2,25712E+24	2,25712E+24	4,68448E+18
46	3,7418E+58	1,55088E+56	8,74033E+24	8,74033E+24	1,36037E+19
47	8,24095E+59	3,061E+57	3,38688E+25	3,38688E+25	3,95322E+19
48	1,81633E+61	6,04714E+58	1,31328E+26	1,31328E+26	1,14956E+20
49	4,00606E+62	1,19569E+60	5,09552E+26	5,09552E+26	3,34496E+20
50	8,84171E+63	2,36623E+61	1,97826E+27	1,97826E+27	9,739E+20
51			7,68479E+27	7,68479E+27	2,83721E+21
52			2,98692E+28	2,98692E+28	8,27014E+21
53			1,16158E+29	1,16158E+29	2,41196E+22
54			4,5196E+29	4,5196E+29	7,03807E+22
55			1,75941E+30	1,75941E+30	2,05473E+23
56			6,85246E+30	6,85246E+30	6,00162E+23
57			2,6701E+31	2,6701E+31	1,75382E+24
58			1,04088E+32	1,04088E+32	5,12739E+24
59			4,05945E+32	4,05945E+32	1,49968E+25
60			1,58385E+33	1,58385E+33	4,38817E+25
61			6,18213E+33	6,18213E+33	1,28454E+26
62			2,41397E+34	2,41397E+34	3,76167E+26
63			9,42959E+34	9,42959E+34	1,102E+27
64			3,68479E+35	3,68479E+35	3,22955E+27
65			1,44042E+36	1,44042E+36	9,46802E+27
66			5,63268E+36	5,63268E+36	2,77669E+28
67			2,20337E+37	2,20337E+37	8,14598E+28
68			8,62189E+37	8,62189E+37	2,39057E+29

n	$S_1(n)$	$S_2(n)$	$S_3(n)$	$S_4(n)$	$S_5(n)$
69			3,37486E+38	3,37486E+38	7,01774E+29
70			1,32142E+39	1,32142E+39	2,06076E+30
71			5,17557E+39	5,17557E+39	6,05326E+30
72			2,02769E+40	2,02769E+40	1,7786E+31
73			7,94635E+40	7,94635E+40	5,22745E+31
74			3,11497E+41	3,11497E+41	1,53682E+32
75			1,2214E+42	1,2214E+42	4,5193E+32
76			4,79041E+42	4,79041E+42	1,32933E+33
77			1,87931E+43	1,87931E+43	3,91118E+33
78			7,37452E+43	7,37452E+43	1,15104E+34
79			2,8945E+44	2,8945E+44	3,38827E+34
80			1,13636E+45	1,13636E+45	9,97628E+34
81			4,46229E+45	4,46229E+45	2,93805E+35
82			1,75266E+46	1,75266E+46	8,65461E+35
83			6,88544E+46	6,88544E+46	2,54995E+36
84			2,70557E+47	2,70557E+47	7,51465E+36
85			1,06335E+48	1,06335E+48	2,21501E+37
86			4,18008E+48	4,18008E+48	6,53031E+37
87			1,64353E+49	1,64353E+49	1,92565E+38
88			6,46333E+49	6,46333E+49	5,67944E+38
89			2,54224E+50	2,54224E+50	1,6754E+39
90			1,00013E+51	1,00013E+51	4,94322E+39
91			3,93531E+51	3,93531E+51	1,45875E+40
92			1,54874E+52	1,54874E+52	4,30558E+40
93			6,09609E+52	6,09609E+52	1,27103E+41
94			2,39993E+53	2,39993E+53	3,75282E+41
95			9,44974E+53	9,44974E+53	1,10823E+42
96			3,72144E+54	3,72144E+54	3,27321E+42
97			1,46579E+55	1,46579E+55	9,66913E+42
98			5,77434E+55	5,77434E+55	2,85673E+43
99			2,27509E+56	2,27509E+56	8,44148E+43
100			8,9652E+56	8,9652E+56	2,49479E+44

Anhang B: Die folgenden Tabellen zeigen die empirisch ermittelte Zahl der Shapes für den Typ 5 für unterschidliche Sequenzlängen.

Sequenzlänge 60:

# Sh	# Sq
13	1
14	2
15	1
17	3
19	1
20	4
21	2
22	1
23	4
24	5
25	4
26	5
27	3
28	4
29	11
30	8
31	4
32	6
33	10
34	6
35	12
36	8
37	15
38	11
39	12
40	8
41	3
42	15
43	13
44	15
45	16
46	20
47	19
48	16
49	14

# Sh	# Sq
50	15
51	11
52	18
53	13
54	16
55	15
56	17
57	21
58	16
59	15
60	18
61	15
62	17
63	14
64	15
65	11
66	13
67	15
68	21
69	8
70	12
71	9
72	17
73	11
74	12
75	11
76	8
77	14
78	13
79	14
80	12
81	10
82	12
83	12
84	8

# Sh	# Sq
85	10
86	4
87	10
88	11
89	6
90	11
91	2
92	11
93	7
94	11
95	7
96	5
97	8
98	6
99	5
100	9
101	2
102	4
103	4
104	3
105	2
106	8
107	2
108	7
109	3
110	2
111	1
112	5
113	6
114	1
116	3
117	3
118	4
119	2
120	6

# Sh	# Sq
121	2
122	1
123	1
124	1
125	1
126	1
127	3
128	2
129	4
130	2
131	1
132	2
133	2
134	2
135	2
136	1
138	3
139	1
140	2
143	1
146	1
149	1
153	2
156	1
157	1
160	1
161	1
167	1
170	1
172	1
184	1
191	1

Sequenzlänge 80:

# Sh	# Sq	# Sh	# Sq	# Sh	# Sq	# Sh	# Sq
41	1	184	1	235	1	288	1
43	1	185	1	236	1	290	2
75	1	186	1	237	2	291	1
76	1	187	1	239	1	292	4
89	1	188	2	240	1	293	1
95	1	190	1	241	2	294	1
97	1	191	1	242	3	296	1
102	1	193	1	244	1	297	1
104	2	194	1	245	2	298	4
109	1	195	2	248	4	299	1
110	2	198	3	250	3	300	3
114	1	199	1	251	1	301	4
127	1	200	2	252	1	303	2
130	1	202	1	253	1	304	1
132	1	203	1	255	2	306	1
135	2	204	1	256	1	307	1
139	2	205	1	257	1	308	2
141	1	206	1	258	1	310	1
144	1	207	1	259	1	311	2
146	1	208	1	261	1	312	3
150	1	209	2	262	1	313	1
151	2	210	3	263	1	314	2
157	1	213	2	265	1	315	2
158	1	214	1	266	1	316	1
159	3	216	1	271	1	318	1
162	1	217	2	273	1	320	2
167	1	218	1	274	1	321	1
168	1	219	1	276	3	322	2
172	1	220	1	277	2	323	2
173	2	221	1	278	1	325	2
174	1	230	2	279	2	327	2
178	1	231	4	283	3	328	3
179	1	232	1	284	1	332	1
180	1	233	1	285	1	333	1
181	1	234	2	286	1	334	3

# Sh	# Sq
335	1
338	2
339	1
340	2
343	2
344	1
345	4
346	4
348	3
351	2
352	1
355	2
356	5
357	2
358	1
359	2
361	5
362	1
365	3
366	4
367	3
368	1
369	2
370	3
371	2
372	1
375	3
376	2
377	3
379	1
380	1
381	4
382	3
385	2

# Sh	# Sq
386	1
387	2
388	1
390	2
391	2
392	4
393	1
397	1
398	1
399	1
400	1
401	1
403	5
404	4
405	2
407	4
408	1
409	3
410	1
412	2
416	1
417	1
418	2
419	1
422	1
423	1
424	2
426	1
427	1
428	3
429	3
430	1
431	1
432	2

# Sh	# Sq
433	1
434	4
436	2
437	1
439	1
441	1
442	3
443	2
444	2
445	2
446	1
447	1
448	2
449	1
450	3
451	1
452	2
453	1
454	1
455	1
457	2
458	2
459	3
460	1
461	4
462	2
463	1
464	3
465	2
466	1
468	3
470	3
471	1
472	1

# Sh	# Sq
473	1
474	3
475	1
476	1
477	1
478	2
479	3
480	5
482	2
483	1
484	1
485	3
486	5
488	1
490	1
492	1
493	1
494	2
495	3
497	2
498	1
499	1
500	2
501	2
502	1
504	3
505	3
508	1
509	1
511	3
512	4
513	1
514	1
515	1

# Sh	# Sq
517	2
518	4
519	1
521	1
522	4
523	2
525	2
526	2
527	2
528	2
529	2
530	2
531	1
532	3
533	1
534	2
535	1
536	1
537	1
538	2
539	2
540	1
543	3
544	1
545	1
546	3
547	2
548	2
551	2
552	1
553	3
555	1
557	3
559	2
560	3

# Sh	# Sq
561	1
562	1
563	2
564	1
565	2
568	1
569	1
570	1
572	4
573	2
574	1
576	1
579	1
580	1
581	1
582	2
583	1
585	2
586	2
587	2
588	1
589	3
592	3
593	2
594	2
596	1
598	2
599	1
602	1
603	3
604	1
605	1
606	2
607	3
608	2

# Sh	# Sq
609	1
610	1
611	4
613	4
614	3
615	1
616	2
617	3
618	2
619	1
620	1
621	1
623	4
624	2
626	1
627	1
629	2
630	1
631	1
632	3
633	2
634	1
636	4
637	1
638	1
639	1
640	2
641	1
642	2
643	1
645	2
646	2
647	3
648	3
649	1

# Sh	# Sq
650	1
651	1
652	2
653	2
654	2
655	1
656	2
657	1
658	2
659	3
661	2
662	1
663	2
666	2
667	2
669	2
670	2
671	1
674	1
676	1
677	1
678	1
679	1
680	2
681	1
683	2
684	2
685	2
686	1
687	1
688	1
690	3
691	1
692	1
700	2

# Sh	# Sq	# Sh	# Sq	# Sh	# Sq	# Sh	# Sq
701	1	760	1	827	2	936	1
702	2	761	2	828	2	937	1
703	2	762	1	833	1	938	1
704	1	764	1	834	1	940	1
706	2	765	2	835	3	942	1
707	1	766	1	838	2	945	1
708	1	768	1	840	2	946	1
714	2	770	1	841	2	949	1
716	1	771	1	842	1	951	1
717	1	775	2	843	1	952	1
719	1	776	2	846	1	954	1
721	2	782	2	847	1	955	1
722	2	783	1	849	1	956	1
723	1	784	1	852	2	958	1
724	1	785	2	854	1	959	1
728	2	786	2	858	1	962	1
730	1	787	2	865	2	964	1
732	1	791	2	866	1	968	1
733	1	796	2	867	2	970	2
734	1	798	1	870	1	975	1
739	1	800	2	873	2	977	1
742	1	802	1	877	1	980	1
743	2	805	3	882	1	981	1
744	2	806	1	884	1	982	1
745	1	809	1	885	1	986	1
746	1	810	1	889	1	990	1
747	1	811	1	893	1	993	1
748	2	814	2	897	1	995	1
749	1	815	3	900	1	1004	1
750	1	817	1	901	1	1010	1
754	1	818	1	907	1	1011	1
755	1	819	1	919	1	1013	1
756	3	820	1	922	2	1014	1
757	1	823	2	933	2	1016	1
759	2	825	1	935	1	1017	1

# Sh	# Sq
1021	1
1029	1
1030	1
1040	1
1045	2
1048	1
1051	1
1053	4
1072	1
1078	1
1079	1
1082	1
1086	1
1087	1
1096	1
1100	1
1103	1
1105	2
1110	1
1112	1
1113	1
1120	1
1130	1
1131	1
1132	2
1133	2
1136	1
1141	1

# Sh	# Sq
1146	2
1148	1
1152	2
1159	1
1162	1
1164	1
1168	1
1170	1
1179	1
1180	1
1186	1
1187	1
1194	1
1222	1
1229	1
1230	1
1231	1
1233	1
1237	1
1241	1
1244	1
1245	1
1246	1
1248	1
1250	1
1253	1
1256	1
1258	1

# Sh	# Sq
1282	1
1287	1
1318	1
1328	1
1332	1
1333	1
1359	1
1368	1
1426	1
1438	1
1449	1
1455	1
1487	1
1527	1
1543	1
1567	1
1575	1
1587	1
1599	1
1619	1
1624	1
1732	1
1738	1
1780	1
1804	1
1810	1
1897	1
1974	1

Sequenzlaenge: 100

# Sh	# Sq	# Sh	# Sq	# Sh	# Sq	# Sh	# Sq
523	1	1429	1	1859	1	2101	1
539	1	1450	2	1873	2	2118	1
728	1	1464	1	1877	1	2122	1
739	1	1469	1	1893	1	2144	1
768	1	1503	1	1902	1	2150	1
783	1	1520	1	1925	1	2152	1
805	1	1536	1	1934	1	2173	1
814	1	1560	1	1945	1	2176	1
833	1	1566	1	1952	1	2180	1
855	1	1571	1	1955	2	2186	1
899	1	1589	1	1956	1	2192	1
901	1	1607	1	1963	1	2228	1
913	1	1610	1	1972	1	2235	1
1003	1	1621	1	1974	1	2240	1
1034	1	1623	1	1980	1	2253	1
1035	1	1643	1	1982	1	2255	1
1089	1	1647	1	1984	1	2289	1
1109	1	1658	1	1986	1	2292	1
1142	1	1684	1	1991	1	2296	1
1144	1	1688	1	1995	1	2299	1
1146	1	1704	1	2016	1	2317	1
1222	1	1732	1	2047	1	2325	1
1228	1	1735	1	2049	1	2335	1
1249	1	1754	1	2052	1	2350	1
1269	1	1758	1	2053	1	2357	3
1277	1	1766	1	2055	1	2360	1
1286	2	1771	1	2056	1	2379	1
1290	1	1783	1	2057	1	2397	1
1320	1	1784	1	2059	1	2407	1
1331	1	1809	1	2061	1	2409	1
1339	1	1823	1	2073	1	2416	1
1357	1	1838	1	2077	1	2432	1
1394	1	1839	2	2078	1	2433	1
1395	1	1846	1	2085	1	2444	1
1427	1	1847	1	2096	1	2465	1

#Sh	#Sq	#Sh	#Sq	#Sh	#Sq	#Sh	#Sq
2467	1	2718	1	2958	1	3248	1
2469	1	2723	1	2963	1	3291	1
2479	1	2731	1	2965	1	3300	1
2482	1	2736	1	2966	1	3326	2
2484	1	2740	1	2971	1	3329	1
2493	1	2746	1	2973	1	3335	1
2497	1	2750	1	3000	1	3337	1
2501	2	2769	1	3007	1	3338	1
2503	1	2772	2	3026	1	3347	1
2507	1	2775	1	3032	1	3371	1
2535	1	2777	1	3035	1	3372	1
2552	1	2779	1	3050	1	3382	1
2556	1	2790	1	3060	1	3385	1
2561	1	2792	3	3066	1	3400	1
2568	1	2806	1	3069	1	3405	1
2570	1	2812	1	3074	1	3433	2
2584	1	2825	2	3075	1	3435	1
2585	1	2828	1	3077	1	3439	1
2591	1	2830	1	3082	1	3442	1
2617	1	2831	1	3096	1	3452	1
2619	1	2837	1	3099	1	3456	1
2627	1	2841	1	3108	1	3461	1
2632	1	2843	1	3114	1	3477	1
2643	2	2847	1	3145	1	3486	1
2647	1	2849	1	3155	1	3496	1
2655	3	2859	1	3162	1	3498	2
2675	1	2866	1	3173	1	3503	1
2683	1	2892	1	3176	1	3505	1
2688	1	2898	1	3178	1	3535	1
2689	1	2899	1	3184	1	3544	1
2693	1	2910	1	3189	1	3557	2
2697	1	2917	1	3206	1	3560	1
2704	1	2923	1	3208	1	3564	1
2705	1	2944	1	3215	2	3566	1
2712	1	2946	1	3241	1	3583	1

#Sh	#Sq	#Sh	#Sq	#Sh	#Sq	#Sh	#Sq
3587	1	3939	1	4193	1	4456	1
3594	1	3940	1	4194	1	4465	1
3603	1	3947	1	4196	2	4466	1
3604	1	3953	1	4215	1	4472	1
3620	1	3970	1	4220	1	4484	1
3632	1	3975	1	4234	1	4488	1
3642	1	3980	1	4244	1	4493	1
3644	1	3988	1	4247	1	4497	1
3651	1	3990	1	4253	1	4499	1
3658	1	3996	1	4261	1	4501	1
3662	1	4001	2	4265	1	4503	1
3672	1	4009	1	4267	1	4505	1
3675	1	4014	1	4284	1	4510	1
3676	1	4016	1	4288	1	4511	1
3677	1	4019	1	4294	1	4512	1
3683	1	4020	1	4295	1	4517	1
3689	1	4022	1	4304	1	4525	1
3727	1	4027	1	4305	1	4548	1
3729	1	4036	1	4321	1	4552	1
3745	1	4040	1	4324	1	4564	1
3748	1	4041	1	4332	1	4569	1
3751	1	4042	1	4336	1	4574	1
3769	1	4044	1	4347	1	4579	1
3773	1	4048	1	4349	1	4588	1
3801	1	4064	2	4355	1	4594	1
3804	1	4082	1	4362	1	4601	1
3818	1	4098	1	4370	1	4609	2
3824	1	4126	1	4377	1	4615	1
3845	1	4127	1	4396	1	4623	1
3849	1	4144	1	4398	1	4628	1
3866	1	4156	2	4399	1	4638	1
3869	2	4173	1	4401	1	4654	1
3891	2	4175	1	4426	1	4670	1
3915	1	4178	1	4437	1	4676	2
3922	1	4189	1	4445	1	4690	1
3928	1	4192	1	4451	1	4692	2

#Sh	#Sq	#Sh	#Sq	#Sh	#Sq	#Sh	#Sq
4701	1	4947	1	5254	1	5575	1
4704	1	4947	1	5265	2	5580	1
4716	1	4947	1	5268	1	5591	1
4727	1	4982	1	5296	1	5592	1
4747	1	4985	1	5298	1	5593	1
4748	1	4999	1	5303	1	5596	1
4751	1	5008	1	5305	1	5598	1
4758	1	5011	1	5306	1	5601	1
4768	1	5021	1	5315	1	5604	1
4770	1	5031	1	5330	1	5625	1
4774	1	5037	1	5353	1	5628	1
4775	1	5040	1	5357	1	5629	1
4779	1	5042	1	5365	1	5640	1
4785	1	5046	1	5371	1	5653	1
4787	1	5056	1	5377	1	5657	1
4791	1	5070	1	5381	1	5658	1
4799	1	5071	1	5404	1	5669	1
4806	1	5083	1	5433	1	5672	1
4808	1	5093	1	5434	1	5677	1
4821	2	5095	1	5437	2	5680	1
4842	1	5117	1	5464	1	5702	1
4847	1	5129	1	5486	1	5716	1
4857	1	5131	1	5491	1	5719	1
4864	1	5150	1	5493	1	5721	1
4867	1	5158	1	5502	1	5726	1
4874	1	5176	1	5507	1	5729	1
4877	2	5188	2	5518	1	5753	1
4887	3	5213	1	5520	1	5765	1
4888	1	5218	1	5522	1	5770	1
4889	1	5221	1	5528	1	5783	1
4895	1	5222	1	5531	1	5822	1
4896	2	5233	1	5545	1	5827	1
4913	1	5239	1	5550	1	5841	1
4923	1	5244	1	5554	1	5842	1
4938	1	5247	2	5555	1	5851	1

#Sh	#Sq	#Sh	#Sq	#Sh	#Sq	#Sh	#Sq
5888	1	6204	1	6663	1	7128	1
5891	1	6212	1	6667	2	7129	1
5893	1	6220	1	6674	1	7138	1
5900	1	6249	1	6696	1	7144	1
5903	1	6259	1	6712	1	7149	1
5905	1	6268	1	6736	1	7161	1
5925	1	6281	1	6743	1	7182	1
5935	1	6283	1	6749	1	7209	1
5942	1	6290	1	6758	1	7216	1
5948	1	6294	1	6762	1	7225	1
5959	1	6298	1	6766	1	7289	1
5960	1	6317	1	6778	1	7300	1
5966	1	6336	1	6783	1	7314	1
5971	1	6378	1	6798	1	7317	1
5973	1	6400	1	6799	1	7330	1
5979	1	6437	1	6809	1	7337	1
5994	1	6444	1	6816	1	7345	1
6012	1	6450	1	6872	1	7387	1
6016	1	6482	1	6892	1	7389	1
6035	1	6484	2	6918	1	7390	1
6039	1	6494	1	6920	1	7402	1
6072	1	6521	1	6968	3	7473	1
6112	1	6541	1	6970	1	7480	1
6120	1	6565	1	6975	2	7503	2
6122	1	6584	1	6977	1	7547	2
6124	1	6585	1	6992	1	7555	1
6128	1	6592	2	6993	1	7567	1
6147	1	6597	1	6997	1	7579	1
6161	1	6602	1	7023	2	7588	1
6162	1	6610	1	7049	1	7621	1
6166	1	6623	1	7056	1	7635	1
6169	1	6624	1	7089	1	7637	1
6176	1	6639	1	7090	1	7659	1
6187	1	6642	1	7117	1	7660	1
6198	1	6657	1	7122	1	7686	1

#Sh	#Sq	#Sh	#Sq	#Sh	#Sq	#Sh	#Sq
7693	1	8216	1	8803	1	9475	1
7703	1	8219	1	8812	1	9485	1
7737	1	8221	1	8832	1	9537	1
7738	1	8222	1	8840	1	9557	1
7744	1	8233	1	8865	1	9562	1
7782	1	8241	1	8876	1	9569	1
7800	1	8246	1	8891	1	9586	2
7809	1	8258	1	8926	1	9599	1
7835	1	8281	1	8928	1	9641	1
7851	1	8296	2	8932	1	9644	1
7854	1	8303	1	8934	1	9654	1
7856	1	8355	1	8955	1	9702	1
7861	1	8358	1	9004	1	9803	1
7877	1	8377	1	9063	1	9808	1
7907	1	8391	1	9076	1	9845	2
7915	1	8407	1	9118	1	9852	1
7929	1	8408	1	9120	1	9860	1
7936	1	8436	1	9124	1	9873	1
7956	1	8462	1	9128	1	9904	1
7964	1	8513	1	9148	1	9931	1
7993	1	8548	1	9163	1	9945	1
7996	1	8562	1	9182	1	9958	1
8007	1	8563	1	9189	1	9962	1
8024	1	8564	1	9217	1	9985	1
8070	1	8607	1	9219	1	9988	1
8072	1	8625	1	9220	1	9991	1
8101	1	8669	1	9292	1	10077	1
8121	1	8675	1	9295	1	10155	1
8124	1	8693	1	9321	1	10205	1
8132	1	8698	2	9342	1	10218	1
8142	1	8704	1	9368	1	10222	1
8177	1	8720	1	9369	1	10250	1
8183	1	8722	1	9410	1	10263	1
8186	1	8736	1	9415	1	10271	1
8194	1	8742	1	9444	2	10293	1

#Sh	#Sq
10345	1
10403	1
10429	1
10444	1
10450	1
10536	1
10551	1
10553	1
10627	1
10649	1
10678	1
10684	2
10697	1
10704	1
10750	1
10757	1
10772	1
10795	1
10801	1
10848	1
10860	1
10897	1
10916	1
10918	1
10925	1
10970	2
10975	1
11018	1
11022	1
11032	1
11144	1
11223	1
11238	1
11270	1
11294	1

#Sh	#Sq
11297	1
11323	1
11365	1
11380	1
11411	1
11438	1
11603	1
11628	1
11706	1
11774	1
11776	1
11864	1
11926	1
11990	1
12020	1
12059	1
12165	1
12238	1
12301	1
12366	1
12458	1
12577	1
12645	1
12659	1
12689	1
12714	1
12737	1
12814	1
12860	1
12865	1
12915	1
12925	1
12926	1
12932	1
13041	1

#Sh	#Sq
13045	1
13195	1
13231	1
13357	1
13466	1
13467	1
13617	1
13729	1
13844	1
13924	1
14040	1
14068	1
14092	1
14201	1
14417	1
14422	1
14475	1
14536	1
14636	1
14684	1
14864	1
14997	1
15006	1
15475	1
15483	1
15593	1
16196	1
16233	1
17236	1
17517	1
17767	1
18958	1
22304	1

Sequenzlaenge: 120

# Sh	# Sq	#Sh	#Sq	#Sh	#Sq	#Sh	#Sq
1394	1	14088	1	17637	1	20924	1
3802	1	14115	1	17642	1	20983	1
4075	1	14208	1	17824	1	21030	1
6828	1	14236	1	17858	1	21304	1
7557	1	14449	1	18207	1	21395	1
7709	1	14563	1	18246	1	21466	1
7774	1	14646	1	18365	1	21614	1
8132	1	14670	1	18476	1	21776	1
8458	1	14797	1	18495	1	21881	1
8524	1	15090	1	18584	1	21966	1
8765	1	15527	1	18632	1	22001	1
9256	1	15570	1	18635	1	22009	1
9448	1	15605	1	18677	1	22251	1
9760	1	15830	1	18722	1	22355	1
9861	1	15895	1	18772	1	22402	1
10375	1	15975	1	18872	1	22432	1
10697	1	16073	1	19194	1	22522	1
10927	1	16105	1	19249	1	22528	1
11036	1	16181	1	19267	1	22562	1
11311	1	16224	1	19307	1	22638	1
11351	1	16277	1	19344	1	22765	1
11707	1	16297	1	19391	1	22916	1
11843	1	16417	1	19399	1	22920	1
11894	1	16769	1	19497	1	23040	1
12025	1	16772	1	19509	1	23168	1
12061	1	16799	1	19628	1	23242	1
12188	1	16818	1	19672	1	23333	1
12400	1	16827	1	19767	1	23536	1
12588	1	17035	1	20035	1	23547	1
12639	1	17104	1	20117	1	23569	1
12680	1	17118	1	20184	1	23598	1
12726	1	17206	1	20312	1	23660	1
12967	1	17244	1	20317	1	23736	1
13602	1	17391	1	20469	1	23757	1
13986	1	17427	1	20529	1	23762	1

# Sh	# Sq	#Sh	#Sq	#Sh	#Sq	#Sh	#Sq
23824	1	27179	1	29890	1	32863	1
23841	1	27197	1	29921	1	32946	1
23983	1	27208	1	30100	1	32949	1
24027	1	27340	1	30148	1	33053	1
24224	1	27425	1	30199	1	33143	1
24313	1	27472	1	30297	1	33148	1
24325	1	27518	1	30312	1	33161	1
24333	1	27551	1	30314	1	33186	1
24403	1	27590	1	30343	1	33251	1
24547	1	27615	1	30495	1	33330	1
24586	1	27741	1	30540	1	33368	1
24659	1	27798	1	30605	1	33436	1
25029	1	27880	1	30822	1	33502	1
25065	1	27981	1	30855	1	33555	1
25225	1	28011	1	30924	1	33698	1
25251	1	28015	1	31009	1	33762	1
25347	1	28088	1	31097	1	33922	1
25549	1	28190	1	31189	1	33977	1
25671	1	28310	1	31333	1	34034	1
25722	1	28417	1	31337	1	34043	1
25813	1	28594	1	31367	1	34180	1
25895	1	28648	1	31539	1	34290	1
25899	1	28696	1	31637	1	34299	1
26074	1	28762	1	31765	1	34312	1
26109	1	28769	1	31773	1	34438	1
26254	1	28864	1	31775	1	34495	1
26427	1	28895	1	31812	1	34588	1
26494	1	28958	2	31835	1	34636	1
26638	1	29290	2	32043	1	34781	1
26642	1	29339	1	32070	1	34804	1
26648	1	29386	1	32105	1	34806	1
26871	1	29538	1	32756	1	34816	1
27015	1	29587	1	32764	1	34843	1
27043	1	29645	1	32818	1	34944	1
27145	1	29668	1	32833	1	34989	1

# Sh	# Sq	#Sh	#Sq	#Sh	#Sq	#Sh	#Sq
35120	1	37576	1	40127	1	43393	1
35190	1	37580	1	40374	1	43477	1
35209	1	37637	1	40388	1	43551	1
35289	1	37681	1	40432	1	43561	1
35324	1	37729	1	40486	1	43627	1
35329	1	37766	1	40532	1	43690	1
35369	1	37790	1	40544	1	43754	1
35441	1	37792	1	40552	1	43785	1
35449	1	37803	1	40618	2	43863	1
35555	1	37898	1	40624	1	44041	1
35590	1	37900	1	40683	1	44050	1
35678	2	37953	1	40791	1	44085	1
35770	1	37967	1	40882	1	44214	1
35888	1	37984	1	40894	1	44278	1
35904	1	38221	1	41011	1	44416	1
35916	1	38222	1	41156	1	44471	1
35921	1	38240	1	41239	1	44479	1
36002	1	38344	1	41263	1	44548	1
36196	2	38404	1	41314	1	44671	1
36270	1	38407	1	41316	1	44690	1
36383	1	38562	1	41512	1	44758	1
36491	2	38654	1	41869	1	44770	1
36515	1	38722	1	41967	1	44773	1
36639	1	38737	1	42014	1	44976	1
36746	1	38831	1	42015	1	44990	1
36794	1	39045	1	42084	1	45003	1
36843	1	39065	1	42290	1	45018	1
36851	1	39149	1	42357	1	45093	1
37101	1	39257	1	42404	1	45173	1
37112	1	39544	1	42426	1	45365	1
37197	1	39570	1	42601	1	45529	1
37199	1	39739	1	42994	1	45560	1
37225	1	39847	1	43140	1	45573	1
37549	1	40027	1	43141	1	45684	1
37571	1	40104	1	43253	1	45742	1

# Sh	# Sq	#Sh	#Sq	#Sh	#Sq	#Sh	#Sq
45867	1	48479	1	51318	1	54798	1
45885	1	48602	1	51390	1	54876	1
45988	1	48806	1	51434	1	54989	1
46171	1	48822	1	51602	1	55105	1
46198	1	48839	1	51639	1	55315	1
46367	1	49047	1	51941	1	55556	1
46372	1	49139	1	52016	1	55573	1
46615	1	49190	1	52123	1	55815	1
46623	1	49264	1	52174	1	55855	1
46681	1	49282	1	52236	1	55858	1
46772	1	49305	1	52371	1	56172	1
46858	1	49524	1	52681	1	56317	1
46871	1	49565	1	52710	1	56403	1
47254	1	49578	1	53055	1	56422	1
47311	1	49613	1	53193	1	56505	1
47343	1	49692	1	53334	1	56564	1
47458	1	49814	1	53412	1	56711	1
47489	1	49918	2	53423	1	56767	1
47571	1	49975	1	53425	1	56846	1
47575	1	50103	1	53584	1	56879	1
47689	1	50256	1	53612	1	56916	1
47708	1	50343	1	53635	1	56944	1
47730	1	50400	1	53725	1	57093	1
47751	1	50554	1	53808	1	57107	1
47768	1	50666	1	53815	1	57235	1
47775	1	50683	1	53863	1	57274	1
47940	1	50745	1	53893	1	57302	1
48061	1	50815	1	53898	1	57378	1
48192	1	50826	1	53908	1	57558	1
48196	1	50827	1	53961	1	57658	1
48309	1	50875	1	54061	1	57723	1
48369	1	50889	1	54280	1	57746	1
48388	1	50951	1	54465	1	58010	1
48456	1	51044	1	54585	1	58024	1
48468	1	51299	1	54675	1	58056	1

# Sh	# Sq	#Sh	#Sq	#Sh	#Sq	#Sh	#Sq
58078	1	61596	1	66973	1	71621	1
58119	1	61663	1	67222	1	72187	1
58158	1	61690	1	67344	1	72210	2
58191	1	61696	1	67397	1	72250	1
58202	1	62135	1	67891	1	72309	1
58208	1	62524	1	67898	1	72407	1
58294	1	62869	1	67969	1	72519	1
58384	1	62904	1	68349	1	72818	1
58635	1	63229	1	68412	1	72906	1
58763	1	63232	1	68499	1	72983	1
58820	1	63286	1	68717	1	73018	1
58832	1	63316	1	68777	1	73539	1
59110	1	63542	1	69037	1	73541	1
59171	1	63606	1	69190	1	73715	1
59196	1	63708	1	69225	1	73949	1
59452	1	63813	1	69284	1	74032	1
59477	1	63904	1	69342	1	74255	1
59492	1	64070	1	69427	1	74259	1
59758	1	64167	1	69540	1	74287	1
59883	1	64327	1	69929	1	74496	1
59930	1	64419	1	70038	1	74519	1
60206	1	64423	1	70259	1	74786	1
60415	1	64553	1	70317	1	74894	1
60554	1	64561	1	70347	1	74943	1
60606	1	64671	1	70432	1	74996	1
60709	1	64924	1	70464	1	75217	1
60784	1	64941	1	70533	1	75525	1
60901	1	65295	1	70547	1	75597	1
61003	1	65339	1	70703	1	76010	1
61010	1	65896	1	70903	1	76039	1
61106	1	65975	1	71189	1	76262	1
61245	1	66164	1	71329	1	76439	1
61355	1	66242	1	71365	1	76456	1
61441	1	66245	1	71411	1	76495	1
61511	1	66949	1	71517	1	76515	1

# Sh	# Sq	#Sh	#Sq	#Sh	#Sq	#Sh	#Sq
76549	1	83329	1	87857	1	94628	1
76758	1	83362	1	87986	1	95523	1
76760	1	83444	1	88300	1	95695	1
77087	1	83596	1	88531	1	95817	1
77226	1	83597	1	88868	1	96155	1
77364	1	83664	1	88883	1	96177	1
77701	1	83784	1	89031	1	96804	1
77715	1	84015	1	89662	1	97331	1
78543	1	84097	1	89847	1	97348	1
78772	1	84145	1	90401	1	97351	1
78920	1	84198	1	90412	1	97366	1
78946	1	84538	1	90443	1	97740	1
78971	1	84583	1	90571	1	98007	1
79203	1	84637	1	90629	1	98307	1
79530	1	84756	1	90693	1	98456	1
79945	1	84776	1	91087	1	99307	1
80144	1	84912	1	91304	1	99469	1
80277	1	84983	1	91332	1	99806	1
80386	1	85177	1	91365	1	100541	1
80439	1	85245	1	91416	1	100777	1
80509	1	85398	1	91462	1	100936	1
80585	1	85491	1	91480	1	101638	1
80600	1	85788	1	91524	1	101950	1
80705	1	86000	1	92366	1	102168	1
80790	1	86697	1	92386	1	102324	1
81625	1	86715	1	92739	1	103158	1
81928	1	86933	1	92936	1	103791	1
82002	1	87015	1	93020	1	106061	1
82081	1	87080	1	93121	1	106286	1
82199	1	87348	1	93634	1	106391	1
82263	1	87421	1	93772	1	106525	1
82797	1	87477	1	94043	1	106690	1
82818	1	87503	1	94404	1	106799	1
82891	1	87644	1	94498	1	106825	1
83243	1	87782	1	94518	1	107017	1

# Sh	# Sq	#Sh	#Sq	#Sh	#Sq	#Sh	#Sq
107069	1	117762	1	131304	1	149152	1
107437	1	117899	1	132127	1	149229	1
107815	1	118275	1	132167	1	149377	1
107818	1	118427	1	133226	1	150589	1
108356	1	118850	1	133690	1	151638	1
108382	1	119369	1	134386	1	152248	1
109217	1	120289	1	135602	1	152635	1
109286	1	120539	1	135606	1	152806	1
109590	1	120561	1	136620	1	153794	1
109653	1	120627	1	136754	1	153939	1
109679	1	121170	1	136877	1	154714	1
109767	1	121390	1	137086	1	155476	1
109843	1	121404	1	137359	1	156088	1
109862	1	121819	1	137423	1	158544	1
109937	1	121914	1	137715	1	159875	1
110084	1	122010	1	139229	1	161077	1
110207	1	122075	1	139387	1	162770	1
110492	1	123157	1	139651	1	165498	1
110792	1	123412	1	141735	1	166093	1
110897	1	124287	1	141831	1	167419	1
111588	1	124659	1	142212	1	169485	1
111618	1	125314	1	142566	1	169802	1
112207	1	126599	1	142575	1	170340	1
112516	1	126751	1	142864	1	172644	1
113615	1	126978	1	143420	1	173134	1
113713	1	127716	1	143878	1	182906	1
114297	1	127803	1	144218	1	182957	1
114597	1	127886	1	144470	1	183616	1
114631	1	128210	1	144910	1	188722	1
115172	1	128367	1	146727	1	189370	1
115788	1	128748	1	147629	1	193322	1
116572	1	129384	1	148018	1	194002	1
116836	1	129662	1	148578	1	196570	1
117504	1	129717	1	148752	1	197352	1
117679	1	130939	1	148895	1	201777	1

#Sh	#Sq
202565	1
208784	1
209457	1
213203	1
217223	1
229323	1
247486	1
266475	1
272608	1
275570	1
309547	1
340631	1