UNIVERSITÄT LEIPZIG

Fakultät für Mathematik und Informatik

Institut für Informatik

Diplomarbeit

Konstruktion von Worten mit Matchingeigenschaften und ihre Anwendung auf das RNA Design

Referent: Prof. Dr. Martin Middendorf

Betreuer: Dr. Daniel Merkle

Leipzig, 14. November 2006

vorgelegt von:

Marc Hellmuth

geb. am: 25.06.1980

Studiengang Wirtschaftsmathematik

Zusammenfassung

Bekanntermaßen spielt die RNA eine zentrale Rolle in lebenden Zellen und übt eine Vielzahl von Funktionen in den verschiedensten biologischen Zusammenhängen aus. Da die Funktion der RNA oft eng durch ihre Struktur bestimmt ist, nimmt die Analyse der Struktur einen wichtigen Stellenwert in der Untersuchung über die RNA ein. Das Modell der Sekundärstruktur ist dabei sehr hilfreich.

Wir verschaffen uns zunächst einen Überblick über schon vorhandene Erkenntnisse zu der Frage, unter welchen Voraussetzungen zu gegebenen Sekundärstrukturen Sequenzen existieren, die diese Strukturen realisieren. Das heißt, welche die Möglichkeit haben sich unter bestimmten Paarungsregeln in diese zu falten. Dabei wird deutlich werden, dass die Beantwortung der Frage sehr stark mit den Voraussetzungen an Graphen von Sekundärstrukturen verknüpft ist.

In dem bisherigen Modell werden allerdings nur die entsprechenden Basenpaarungen berücksichtigt, nicht aber Positionen, welche keine Paarung eingehen dürfen. Wir setzen deshalb neue Kanten, so genannte Pseudokanten, in diese Graphen ein, um den Realisierungsbegriff einzuschränken. Die Motivation entspringt dem Wunsch, Sequenzen zu finden, die im Sinne der Paarungsregeln, strukturerhaltender sind. Es wird sich zeigen, dass dabei neue Voraussetzungen an Graphen mit Pseudokanten zu stellen sind, so dass einzelne Sequenzen existieren, die mehrere Sekundärstrukturen realisieren.

Durch diese Sachverhalte motiviert, werden Problemstellungen formuliert und auf ihre Komplexität untersucht. Es geht hierbei um minimale Kanten- bzw. Knotenmengen, die aus Graphen entfernt werden müssen, so dass diese bestimmte Voraussetzungen erfüllen. Insbesondere werden wir zeigen, dass die formulierten Probleme NP-vollständig sind.

Im letzten Teil dieser Arbeit führen wir empirische Studien zu Betrachtungen von Graphen mit und ohne Pseudokanten durch.

Inhaltsverzeichnis

1 Grundlagen

1.1 Die RNA

1.1.1 Einleitung

Betrachtet man die heutigen Lebensformen, so ist allen gemeinsam, dass Desoxyribonukleinsäure (DNA[1]) oder Ribonukleinsäure (RNA[2]) als Träger der Erbinformation fungiert. Die Information wird in allen Organismen, in einem komplexen Prozess, in eine Aminosäuresequenz übersetzt. Die Aminosäurekette faltet sich zu einer dreidimensionalen Struktur, dem fertigen Protein, welches anschließend die wichtigen Lebensfunktionen durchführt [SGM+89].

Lange Zeit wurde davon ausgegangen, dass die einzige Funktion der RNA darin besteht, die in der DNA gespeicherten genetischen Informationen weiterzuleiten, damit diese schließlich in Proteine übersetzt werden können. In diesem Bild stellt die RNA also lediglich eine Zwischenstufe auf dem Weg von der DNA zum Protein dar [Rid05]. Die Annahme, dass dies die einzige Aufgabe der RNA ist, wurde Mitte der achtziger Jahre widerlegt, als Thomas Cech herausfand, dass RNA Moleküle auch als Enzyme fungieren können [Cec86]. Später entdeckte man immer weitere Funktionen, so dass die RNA aus dem Schatten der DNA heraustrat und ein wichtiger Bestandteil der Betrachtungen in der Biologie wurde [Rid05, ZH05].

Die Tatsache, dass RNA nicht nur Informationsträger, sondern eines der wenigen Moleküle ist, welches auch chemische Reaktionen katalysieren kann, führte sogar zu der Idee einer RNA-Welt Hypothese [Moo05, Rid05]. In dieser Hypothese wird davon ausgegangen, dass der Ursprung des Lebens in sich selbst replizierenden RNA Molekülen liegt, aus der sich dann durch molekulare Darwinische Evolution komplexere Systeme entwickelt haben könnten [Gil86, Joy91]

Somit ist klar, dass die RNA eine zentrale Rolle in lebenden Zellen spielt und eine Viel-

[1] aus dem engl.: deoxyribonucleic acid
[2] aus dem engl.: ribonucleic acid

zahl von Aufgaben in den verschiedensten biologischen Zusammenhängen ausübt. Eine Vielzahl von Beispielen findet man hierzu etwa in [BY93, FMS+83, GBP98, HWS+88]. Es ist bekannt, dass die Funktion der RNA sehr oft durch ihre räumliche Struktur festgelegt ist [FHM+01, Fon02]. Die Analyse und Vorhersage der Struktur der RNA ist somit ein wichtiger Bestandteil der Untersuchung über diese. Im Folgenden gehen wir deshalb auf den Aufbau der RNA genauer ein. Wir werden in den nächsten Abschnitten diese Arbeit biologisch motivieren und sehen, warum es spannend und sinnvoll sein kann, sich mit der Struktur der RNA mathematisch auseinander zusetzen.

1.1.2 Chemischer Aufbau der RNA

Die RNA ist ein einzelsträngiges, langkettiges Nukleinsäuremolekül. Sie ist dabei aus vier Bausteinen, den Nukleotiden, zusammengesetzt. Jedes dieser Nukleotide besteht aus einem Ribosemolekül (d.h. einem Zucker mit 5 Kohlenstoffatomen), einem Phosphatrest und einer der vier organischen Basen Adenin, Guanin, Cytosin und Uracil (Abbildung 1.1) [SGM+89].

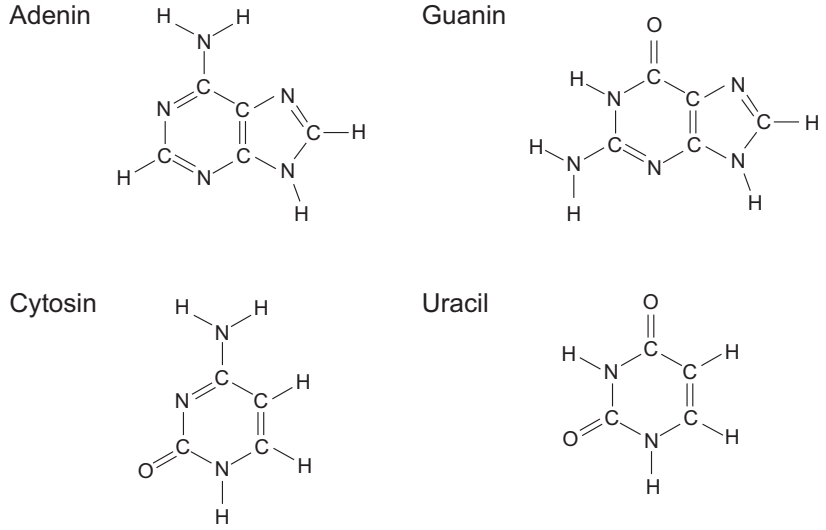

Abbildung 1.1: Der Aufbau der vier Basen Adenin, Guanin, Cytosin und Uracil [SGM+89]

Der Phosphatrest dient als Verbindung zwischen den Zuckermolekülen zweier benachbarter Nukleotide. Er verbindet das 3’ Kohlenstoffatom des Ribosmoleküls eines Nukleotids mit dem 5’ Kohlenstoffatom des Ribosemoleküls des anderen Nukleotids. Dieses Grundgerüst wird auch Rückgrat der RNA genannt. Eine schematische Darstellung findet sich in der Abbildung 1.2.

Allerdings passiert es in der Natur eher selten, dass die RNA als einzelne Kette von Nukleotiden vorkommt. Vielmehr können sich Paarungen zwischen den Basen, die sich auf ein

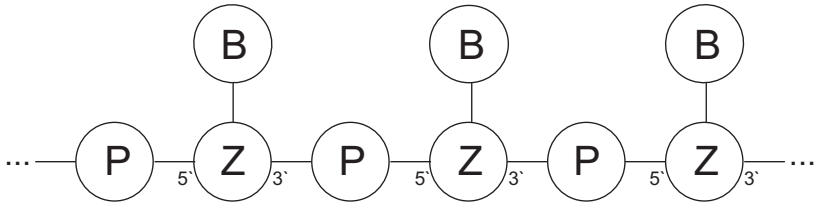

Abbildung 1.2: Schematische Darstellung der Rückgrats der RNA [SGM+89]

und demselben RNA Strang befinden, mittels Wasserstoffbrückenbindungen ausbilden (Abbildung 1.3). Diesen Prozess nennt man RNA Faltung. Jedes Basenpaar besteht aus einer Purinbase und einer Pyrimidinbase. Es können sich hierbei die Basen Guanin und Cytosin, Adenin und Uracil sowie Guanin und Cytosin paaren. Die ersten beiden Verbindungen nennt man auch Watson-Crick Basenpaarungen, die letztere wird Wobble[3] Basenpaarung genannt [SGM+89]. In seltenen Fällen existieren auch andere Paarungen, auf die wir aber nicht weiter eingehen wollen [Mül06].

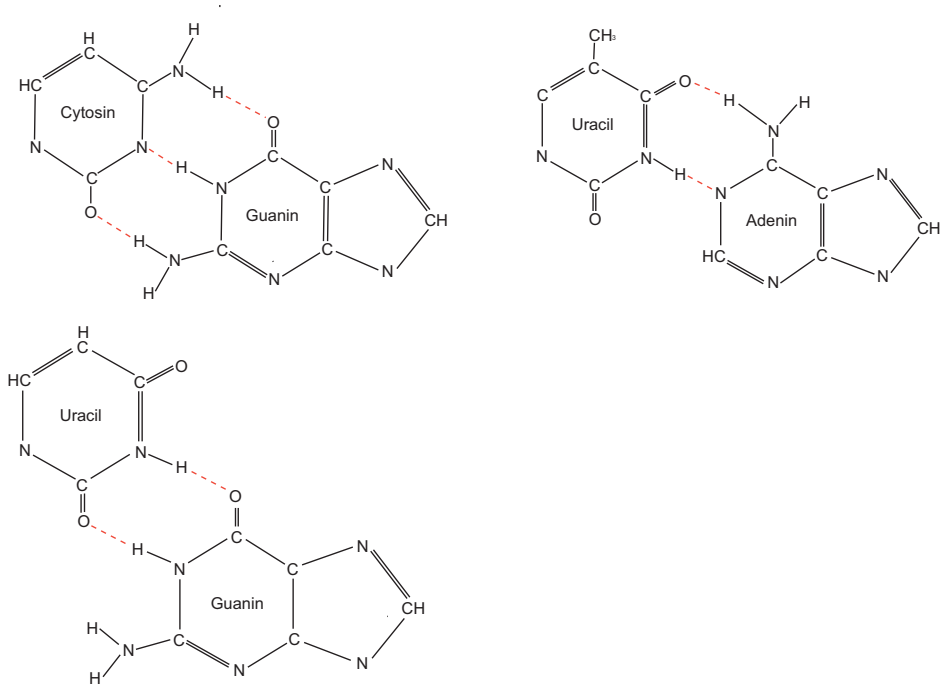

Abbildung 1.3: Die Basenpaarungen Adenin – Uracil, Guanin – Cytosin und Uracil – Guanin [SGM+89]

[3] aus dem engl.: wacklig, schwankend

1.1.3 Primär-, Sekundär und Tertiärstruktur

Seitdem bekannt ist, dass die RNA nicht nur bloßer Informationsträger zwischen DNA und Protein ist, sondern auch andere wichtige Funktionen übernimmt, ist die Analyse und Vorhersage der Struktur der RNA ein wesentlicher Bestandteil der Untersuchungen geworden. Insbesondere wird dieser Sachverhalt dadurch unterstützt, dass die biologische Funktion der RNA sehr oft eng durch ihre räumliche Struktur festgelegt ist. Diese wiederum wird durch die Abfolge der Nukleotide bestimmt [FHM+01, Fon02, Sch99].

Wenn man von der Struktur der RNA spricht, so unterscheidet man in Primär-, Sekundär- und Tertiärstruktur [Fon02, SGM+89]. Als Primärstruktur versteht man die unterste Ebene der Strukturinformation der RNA. Sie bezeichnet die lineare Abfolge der Nukleotide mit den jeweiligen Basen in 5'-3'- Richtung, aus der die RNA besteht. Häufig wird die Primärstruktur auch als Nukleotidsequenz oder einfach nur als Sequenz der RNA bezeichnet. Formal lässt sich die Sequenz einer RNA mit *n* Nukleotiden als ein Wort der Länge *n* aus den Buchstaben A,C,G und U, stellvertretend für die vier Basen, beschreiben.

Definition 1.1. [FHM+01] *Eine* RNA-Nukleotidsequenz *oder kurz* Sequenz *der Länge n ist ein Wort über dem Alphabet* $\mathscr{A} = \{A,C,G,U\}$*, d.h.*

$$s = s_1 \dots s_n \in C_n = \{A,C,G,U\}^n.$$

Wie schon erwähnt, tritt die RNA in der Natur selten als einzelne Kette von Nukleotiden auf. Vielmehr können sich Basenpaarungen bilden. Um die Funktion der RNA zu kennen, muss man wissen, wie die räumliche Struktur der RNA aussieht. Diese wird Tertiärstruktur genannt. Allerdings sind die Beziehungen zwischen Sequenz und räumlicher Struktur sehr komplex [Sch99]. Die räumliche Struktur der RNA ist nur sehr schwer vorherzusagen und der gegenwärtige Wissensstand über dreidimensionale Strukturen von RNA-Molekülen ist rudimentär [Sch99]. Man nutzt deshalb eine grob-aufgelöste Strukturversion, die so genannte Sekundärstruktur. Dieses Modell listet nur Watson-Crick und Wobble Basenpaare. Die RNA-Faltung ist ein Vorgang, bei dem die Sekundärstruktur die Tertiärstruktur wesentlich beeinflusst [Mül06]. Die Sekundärstruktur ist aber konzeptionell viel einfacher und erlaubt die Durchführung einer strengen mathematischen Analyse und großskaliger Berechnungen [FHM+01, Fon02, Sch99]. Zudem sind die Vorhersagen der RNA-Sekundärstruktur verlässlicher als diejenigen von vollen räumlichen Strukturen. Sie stellt somit ein exzellentes Modell zur Untersuchung der RNA zwischen theoretischer Fügsamkeit und empirischer Zugänglichkeit dar. Es sei bemerkt, dass die Sekundärstruktur ein topologisches Konzept ist. Sie sollte daher nicht mit einer Art zweidimensionaler Struktur verwechselt werden [Fon02]. Es folgt nun die formale Definition.

Definition 1.2. [CLK+05] *Sei* $a = a_1 \dots a_n \in C_n$ *und* $\theta \in \mathbb{N}$ *ein fest gewählter Parameter. Eine* Sekundärstruktur Sec *ist eine Menge geordneter Paare* (i, j)*, mit* $1 \leq i < j \leq n$*, so dass folgenden Eigenschaften erfüllt sind:*

1. *Sei* $(i, j) \in Sec \Rightarrow a_i a_j \in \mathscr{B} = \{AU, UA, GC, CG, GU, UG\}$.
2. *Sei* $(i, j), (k, l) \in Sec$*, dann ist* $i < k < j < l$ *nicht erlaubt.*
3. *Sei* $(i, j), (k, l) \in Sec$ *und* $i \in (k, l) \Rightarrow i = k$ *und* $j = l$.
4. *Sei* $(i, j) \in Sec \Rightarrow j > i + \theta$*, mit* θ *als fest gewählten Parameter.*

Die Menge $\mathscr{B}$ beschreibt die Regel, unter denen Paarungen möglich sind. Der zweite Punkt formuliert die Tatsache, dass keine Pseudoknoten erlaubt sind , d.h. Kanten dürfen sich nicht überkreuzen. Aus Punkt drei geht hervor, dass jede Base in nur maximal einer Basenpaarung auftreten darf. Die letzte Forderung ist auf die Starrheit der Moleküle zurückzuführen, d.h. zwischen zwei gepaarten Basen müssen mindestens θ ungepaarte Positionen existieren. Hierbei wird im allgemeinen $\theta = 3$ gesetzt.

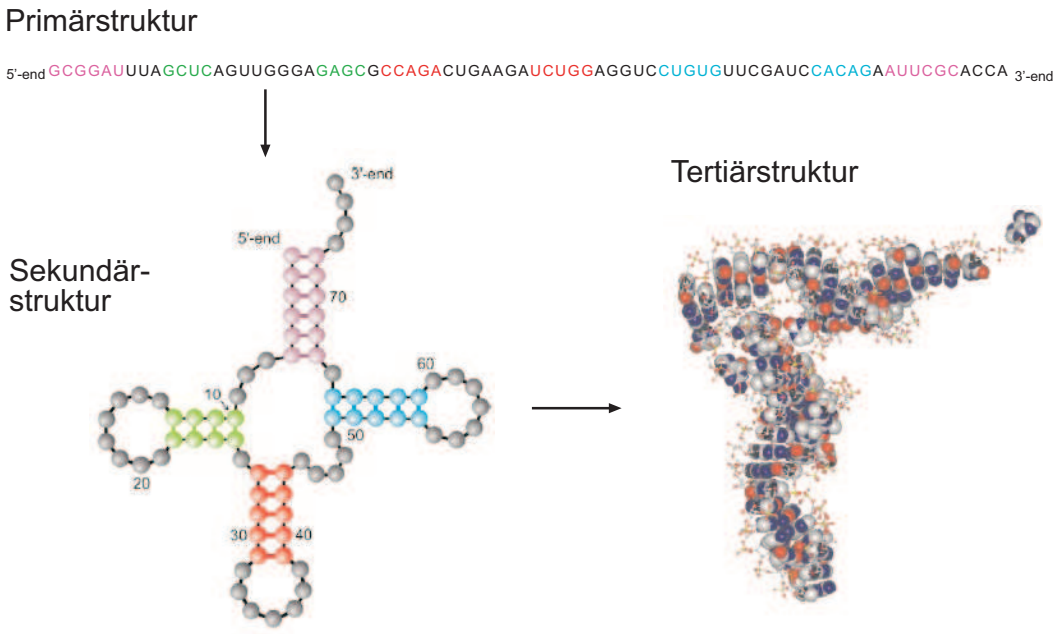

Abbildung 1.4: Faltung der natürlichen t-RNA^{phe}-Sequenz in ihre dreidimensionale Struktur. Zuerst wird die Sekundärstruktur durch die Bildung von Watson-Crick und Wooble Basenpaaren aufgebaut. Dann wird die sekundäre Struktur gefaltet, um die volle dreidimensionale Struktur zu ergeben [Sch99].

1.1.4 Spezielle RNA

Wie im vorangegangenen Abschnitt schon erwähnt, wird die Funktion der RNA durch ihre Struktur bestimmt. Spannenderweise gibt es RNA Sequenzen, die sich in mehr als eine Struktur falten können. Alternative Strukturen von ein und derselben RNA können somit vollständig verschiedene Funktionen festlegen [BSR97, PB98]. Man nennt eine RNA mit diesen Eigenschaften auch RNA switch. Einige Beispiele von diesen speziellen RNA findet man in der Replikation von Viren [GBP98, HWS+88, LSS+91]. Weitere Beispiele in denen die RNA in alternativen Strukturen vorkommen kann sind u.a. trypanosomatid protozoa [HC93, HCU95], sowie in Escherichia coli and Bacillus subtilis [BY93, FMS+83, PGG92]. Mehr über die zunehmende Rolle von RNA switches kann man weiterhin in [CSX+02, GZ97, MBB+03] erfahren.

Betrachten wir auf der anderen Seite Aptamere. Ein Aptamer ist eine kurze RNA Sequenz (i.d.R. 15-60 Nukleotide). Diese werden *in vitro*[4], d.h. künstlich, mittels einem SELEX (Systematic Evolution of Ligands by Exponential enrichment) genannten Verfahren gezielt hergestellt, um bestimmte Moleküle spezifisch zu binden [ES90, TG90]. Die Fähigkeit bestimmte Moleküle, wie zum Beispiel Antibiotika, Peptide, Proteine (z. B. Wachstumsfaktoren und bakterielle Gifte) oder organische Moleküle gezielt zu binden [BKF95, DS02, GBE+96, HP00, SSK+01, WS98], eröffnet für diese molekularen Werkzeuge viele Anwendungsmöglichkeiten in der Diagnostik, in der Medizin und in der Forschung [AMS+99, BK02, BVK+97, KBR99]. So treten Aptamere immer stärker in Konkurrenz zu den klassischen Antikörpern, da sie einige herausragende Eigenschaften besitzen, welche sie zu einer echten Alternative machen [Mül06].

Zusammenfassend können wir sagen, dass im Fall der RNA switches Nukleotidsequenzen existieren, die in der Lage sind, sich in verschiedene Strukturen zu falten. Auf der anderen Seite werden gezielt Sequenzen gesucht, die bestimmte Strukturen annehmen können. Aus diesen Betrachtungen ergibt sich eine natürlich motivierte Fragestellung:

Unter welchen mathematischen Voraussetzungen existiert zu k gegebenen Sekundärstrukturen eine einzelne Sequenz, welche mit diesen kompatibel ist?

Es sei bemerkt, dass wir in dieser Arbeit nicht auf die Betrachtung von Energiewerten eingehen, sondern die Voraussetzungen unter rein mathematischen, insbesondere graphentheoretischen, Aspekten untersuchen werden. Für unsere Betrachtungen stellen Graphen ein sehr

[4] lat.: im (Reagenz)Glas

nützliches Werkzeug dar, um Sekundärstrukturen zu visualisieren. Deshalb werden wir im nächsten Abschnitt genauer auf den Begriff *Graph* eingehen und verschiedene Sachverhalte zu diesen klären.

1.2 Graphen

Wie eben schon erwähnt, werden wir uns in diesem Abschnitt mit Graphen beschäftigen. Im ersten Teil werden grundlegende Definitionen gegeben. Im zweiten Teil werden *mit Vorzeichen markierte Graphen* genauer untersucht, da diese, wie wir später sehen werden, für unsere weiteren Betrachtungen ein sehr nützliches Werkzeug darstellen.

1.2.1 Grundlagen

Definieren wir zunächst formal die benötigten Begrifflichkeiten. Diese sind bis auf wenige Veränderungen aus [Die01] und [IK00] entnommen.

Definition 1.3. *Ein* Graph *ist ein geordnetes Paar $G = (V,E)$ disjunkter Mengen mit $V \neq \emptyset$ und E als Teilmenge der Menge aller 1-und 2-elementigen Teilmengen von V.*

Die Elemente von V werden als Knoten *(engl.* vertex*) und die Elemente von E als* Kanten *(engl.* edges*) des Graphen G bezeichnet. Für V schreiben wir auch $V(G)$ und für E auch $E(G)$, falls der Bezug zum Graphen G wichtig ist. Häufig benutzen wir auch die Bezeichnung V_n für V, wenn wichtig ist, dass $|V| = n$.*

Ein Knoten $v \in V$ heißt mit einer Kante $e \in E$ inzident*, wenn $v \in e$ gilt. Zwei Knoten $u,v \in V$ heißen* adjazent*, falls eine Kante $\{u,v\} \in E$ existiert.*

Mit dem Knotengrad *deg(v), bezeichnen wir die Anzahl aller zu v inzidenten Kanten.*

Ein Graph G heißt endlich *bzw.* unendlich*, falls die Knotenmenge $V(G)$ endlich bzw. unendlich ist. Mit der* Länge *eines endlichen Graphen bezeichnen wir die Kardinalität der Knotenmenge.*

Falls gilt, dass für alle $u,v \in V$ mit $u \neq v$ eine Kante $\{u,v\} \in E$ existiert, d.h. jeder Knoten ist mit jedem anderen Knoten durch eine Kante verbunden, dann heißt der Graph G vollständig*.*

Zwei Graphen G und G' bezeichnet man als isomorph*, falls eine Bijektion ϕ von V(G) in V(G') existiert, so dass gilt: $\{\phi(v),\phi(w)\} \in E(G')$ gdw. $\{u,w\} \in E(G)$*

Da wir im Nachfolgenden nur endliche Graphen betrachten werden, wird auf den Zusatz *endlich* verzichtet. Weiterhin betrachten wir ausschließlich Graphen für die gilt: $|e| = 2$ für

alle $e \in E$. Da wir in den nächsten Abschnitten und Kapiteln sehr häufig die Begriffe *Pfade* und *Zyklen* verwenden werden, definieren wir diese nun.

Definition 1.4. *Sei $G = (V,E)$ ein Graph.*
Eine Folge $[v_1, e_1, v_2, e_2, \ldots, v_n, e_n, v_{n+1}]$ mit $v_i \in V$, $e_i = \{v_i, v_{i+1}\} \in E$ heißt Kantenzug *mit n Kanten. Wahlweise werden wir auch die Darstellung $\{v_1, v_2\} \ldots \{v_n, v_{n+1}\}$ verwenden.*
Ein Kantenzug heißt Weg *gdw. $e_i \neq e_j$ für alle $i \neq j$.*
Ein Weg heißt Pfad *gdw. $v_i \neq v_j$ für alle $i \neq j$.*
Ein Pfad heißt Zyklus *gdw. $v_1 = v_{n+1}$.*

Mit der Länge *eines Pfades bzw. Zyklus bezeichnen wir speziell die Anzahl der Kanten aus denen der Pfad bzw. Zyklus besteht.*

Des Weiteren benötigen wir die nun folgenden erklärten Begriffe *Leiter* und *Schleife*.

Definition 1.5. *Sei $G = (V,E)$ ein Graph mit der Knotenmenge $V = \{1, \ldots, n\}$, $n \in \mathbb{N}$. Wir bezeichnen eine Folge von Kanten $\{i, j\}, \{i+1, j-1\}, \ldots, \{i+n, j-n\}$ als* Leiter *der Länge $n+1$. Eine Folge ungepaarter Knoten, d.h. Knoten die keine Kanten bilden, der Form $\{i, i+1, \ldots, i+n\}$ bezeichnen wir als* Schleife.

Nachdem wir die grundlegende Definition eines Graphen gegeben haben und verschiedene Begrifflichkeiten erklärt haben, schauen wir jetzt, was man unter der Partition eines Graphen versteht. Die Partition von Graphen wird in späteren Betrachtungen von RNA Strukturen einen wichtigen Stellenwert einnehmen.

Definition 1.6. [Die01] *Eine Menge $\{V_1, \ldots, V_n\}$ disjunkter Teilmengen einer Menge V ist eine* Partition *von V, falls $\bigcup_{i=1}^{n} V_i = V$ und $V_i \neq \emptyset$ für jedes $i \in \{1, \ldots, n\}$.*
Sei $r \geq 2$ eine natürliche Zahl. Ein Graph $G = (V,E)$ heißt r-partit, *wenn eine Partition von V in r Teilmengen existiert, so dass die Endknoten einer jeden Kante von G in verschiedenen Teilmengen liegen, d.h. Elemente aus der gleichen Teilmenge dürfen nicht adjazent sein.*
Ein 2-partiter Graph wird auch bipartit *genannt.*

Ein sehr hilfreiches Lemma für die späteren Betrachtungen ist folgendes. Der Beweis kann in [Die01] nachgelesen werden.

Lemma 1.7. [Die01] *Ein beliebiger Graph $G = (V,E)$ ist bipartit $\Leftrightarrow$ G enthält keine Zyklen ungerader Länge.*

1.2.2 Mit Vorzeichen markierte Graphen

Es wird sich in den späteren Kapiteln zeigen, dass mit Vorzeichen markierte Graphen ein sehr nützliches Werkzeug für unsere Betrachtungen von erweiterten Shapes sein werden. Deswegen werden wir auf diese speziellen Graphen im Folgenden genauer eingehen. Es sei kurz erwähnt, dass die Motivation der mit Vorzeichen markierten Graphen der Analyse von sozialen Netzwerken entspringt. Hierbei sind Psychologen nicht nur an einer binären Relation interessiert, d.h. ob sich verschiedene Individuen kennen oder nicht (diese kann mit einfachen ungerichteten Graphen dargestellt werden), sondern auch daran, eine zusätzliche Relation "mögen" oder "nicht-mögen" zwischen den Individuen einzusetzen. Diese Beziehung wird realisiert, in dem die entsprechenden Graphen mit Vorzeichen markiert werden [Har54]. Geben wir nun die formale Definition.

Definition 1.8. [Har54] *Ein* S-Graph $H = (G, \varphi)$ *ist ein mit Vorzeichen markierter Graph, bestehend aus einem Graphen* $G = (V, E)$ *und einer Abbildung* $\varphi : E \to \{+1, -1\}$.

Die Abkürzung *S* steht hier für das englische Wort *signed.*

Definition 1.9. [Har54] *Sei* $P = [v_1, \ldots, e_n, v_{n+1}]$ *ein Pfad und* $C = [v_1, \ldots, e_n, v_1]$ *ein Zyklus. Das* Vorzeichen $\varphi(P)$ *bzw.* $\varphi(C)$ *eines Pfades bzw. Zyklus wird durch das Produkt der Vorzeichen der Kanten auf P bzw. C definiert, d.h.* $\varphi(P) := \prod_{i=1}^{n} \varphi(e_i)$ *bzw.* $\varphi(C) := \prod_{i=1}^{n} \varphi(e_i)$. *Ein Pfad P bzw. Zyklus C wird als* positiv *bzw.* negativ *bezeichnet, falls ihre Vorzeichen* $\varphi(P)$ *bzw.* $\varphi(C)$ *positiv bzw. negativ sind.*

Definition 1.10. *Ein S-Graph* $H = (V, E, \sigma)$ *heißt* balanciert, *falls alle Zyklen in H positiv sind.*

Man sieht aus dieser Definition leicht, dass ein S-Graph balanciert ist, falls keine Zyklen mit einer ungeraden Anzahl von negativen Vorzeichen existieren. Wir werden am Ende dieses Abschnitts den Beweis liefern, dass ein S-Graph $H = (V, E, \sigma)$ balanciert ist gdw. sich die Knotenmenge V in eine Partition aus zwei Teilmengen V_1 und V_2 zerlegen lässt, so dass alle Kanten die aus adjazenten Knoten innerhalb der jeweiligen Teilmengen bestehen positiv markiert sind und Kanten aus adjazenten Knoten zwischen den Teilmengen negativ markiert sind. Um dies zu zeigen benötigen wir zunächst folgende Lemmata und Theoreme. Die entsprechenden Beweise sind aus [Har54] entnommen.

Theorem 1.11. [Har54] *Ein vollständiger S-Graph $H = (V, E, \sigma)$ ist balanciert $\Leftrightarrow$ V lässt sich in zwei disjunkte Teilmengen V_1, V_2 zerlegen, so dass gilt:*

1. $V_1 \cup V_2 = V$ *und*
2. $\varphi(\{u,v\}) = \begin{cases} -1 & \text{, falls } u \in V_1 \text{ und } v \in V_2 \\ +1 & \text{, sonst} \end{cases}$

   Das heißt, alle Kanten innerhalb einer Teilmenge sind positiv und Kanten zwischen den Teilmengen sind negativ markiert.

Beweis. $\Longrightarrow$: Sei $H = (V, E, \sigma)$ ein balancierter und vollständiger S-Graph, sowie $\hat{v}$ ein beliebiger Knoten aus V.
Wir definieren $V_1 := \{\hat{v}\} \cup \{w \mid \exists\, e = \{\hat{v}, w\} \in E \text{ mit } \varphi(e) = +1\}$, also als die Menge aller zu $\hat{v}$ adjazenten Knoten, deren Kanten positiv sind zuzüglich $\hat{v}$ und die Menge $V_2 := V \backslash V_1$. Offensichtlich ist damit die erste Eigenschaft erfüllt. Um zu zeigen, dass die zweite Eigenschaft gilt, muss man drei Fälle unterscheiden:

- Seien $u, w \in V_1$. Zz.: u, w sind positiv adjazent.
  Falls $u = \hat{v}$, so folgt die Behauptung nach Konstruktion der Menge V_1.
  Seien also nun $u, w \neq \hat{v}$. Da H vollständig ist, existiert der Zyklus $\{\hat{v}, u\}, \{u, w\}, \{w, \hat{v}\}$
  Nach Konstruktion von V_1 gilt, dass $\varphi(\{\hat{v}, u\}) = \varphi(\{\hat{v}, w\}) = +1$. Da weiterhin der S-Graph H balanciert ist, also alle Zyklen positiv sind, folgt dass die Kante {u,w} positiv ist.

- Seien $u, w \in V_2$. Zz.: u, w sind positiv adjazent.
  Auch hier gilt wieder wegen der Eigenschaften des S-Graphen H, dass der Zyklus $\{\hat{v}, u\}, \{u, w\}, \{w, \hat{v}\}$ positiv ist und dass nach Konstruktion von V_1 gelten muss $\varphi(\{\hat{v}, u\}) = \varphi(\{\hat{v}, w\}) = -1$. Da alle Zyklen in H positiv sind, muss die Kante $\{u, w\}$ positiv sein.

- Entsprechend zeigt man für $\{u, w\} \in E$ mit $u \in V_1, w \in V_2$, dass die zweite Eigenschaft erfüllt ist.

$\Longleftarrow$: Sei $H = (V, E, \sigma)$ ein vollständiger S-Graph und V_1, V_2 Zerlegungen von V, so dass alle Kanten zwischen Elementen aus $V_i,\ i = 1, 2$ positiv adjazent sind, sowie Kanten die zwischen Elementen aus V_1 und V_2 existieren negativ markiert sind. Jeder Zyklus innerhalb einer Menge $V_i,\ i = 1, 2$ ist somit positiv. Falls ein Zyklus zwischen Elementen aus V_1 und V_2 existiert, so muss es eine gerade Anzahl von Kanten geben, die zwischen den Teilmengen V_1

und V_2 verlaufen, damit der Zyklus geschlossen wird. Folglich ist jeder Zyklus positiv und somit H balanciert. □

Lemma 1.12. [Har54] *Jeder Teilgraph eines balancierten S-Graphen ist balanciert.*

Beweis. Da jeder Zyklus im Teilgraphen auch ein Zyklus im balancierten S-Graphen ist und somit positiv ist, folgt die Behauptung. □

Theorem 1.13. [Har54] *Ein S-Graph $H = (V, E, \sigma)$ ist balanciert $\Leftrightarrow$ für alle Knoten $u, w \in V$ mit $u \neq w$ gilt: alle Pfade zwischen u und w tragen dasselbe Vorzeichen.*

Beweis. $\Longrightarrow$: Seien P_1, P_2 zwei verschiedene Pfade die u und w beinhalten. Das Entfernen gemeinsamer Kanten (sofern welche vorhanden sind) führt zu einer Menge Kanten disjunkter Zyklen. Jeder der Zyklen C besteht aus einem Teilpfad von P_1, sowie von P_2. Da C positiv ist, muss jeder dieser Teilpfade dasselbe Vorzeichen haben. Betrachtet man nun alle Teilpfade und alle gemeinsamen Kanten zusammen, folgt $\varphi(P_1) = \varphi(P_2)$.

$\Longleftarrow$: Da alle Pfade, die u und w beinhalten, dasselbe Vorzeichen tragen, ist jeder Zyklus, der u und w beinhaltet, positiv. Da weiterhin u und w beliebig gewählt werden können, sind alle Zyklen positiv. □

Theorem 1.14. [Har54] *Ein S-Graph $H = (V, E, \sigma)$ ist balanciert $\Leftrightarrow$ V lässt sich in zwei disjunkte Teilmengen V_1 und V_2 zerlegen, so dass gilt:*

1. $V_1 \cup V_2 = V$ *und*
2. $\varphi(\{u, v\}) = \begin{cases} -1 & \text{, falls } u \in V_1 \text{ und } v \in V_2 \\ +1 & \text{, sonst} \end{cases}$

Beweis. $\Longrightarrow$: Wir erweitern H induktiv zu einem vollständigen S-Graphen H'. O.B.d.A. kann H als zusammenhängend betrachtet werden. Seien jetzt u, w zwei beliebige nicht adjazente Knoten aus V. Nach Theorem 1.13 gilt, dass alle Pfade, die u und w beinhalten, gleiches Vorzeichen tragen. Nun verbinden wir diese Knoten durch eine Kante, welche das Vorzeichen all dieser Pfade trägt. Alle neue Zyklen, die durch das Verbinden der Knoten u, w entstehen, sind somit positiv. Der auf diese Weise entstandene Graph $\tilde{H} = (V, E \cup \{u, w\}, \varphi')$ ist also balanciert. Nun wird mit dem S-Graphen $\tilde{H}$ ebenso verfahren, bis wir einen vollständigen S-Graphen H' erhalten. Induktiv folgt, H' ist balanciert. Da nach Theorem 1.11 die Knotenmenge V von H' wieder in zwei disjunkte Teilmengen zerlegt werden kann und man H aus H' durch einfaches Entfernen von Kanten aus H' erhält, folgt die Behauptung.

$\Longleftarrow$: Wir werden den gegebenen S-Graphen H wieder zu einem vollständigen S-Graphen H' erweitern. Dazu nehmen wir zwei beliebige nicht adjazente Knoten u, w aus V. Falls diese Knoten in der gleichen Teilmenge liegen, verbinden wir die Knoten mit einer positiven Kante. Falls die Knoten aus verschiedenen Teilmengen sind, so verbinden wir sie durch eine negative Kante. Da die Knotenmenge von H' immer noch in zwei disjunkte Teilmengen zerlegt werden kann, gilt nach Theorem 1.11, dass H' balanciert ist. Aus Lemma 1.12 folgt nun die Behauptung. □

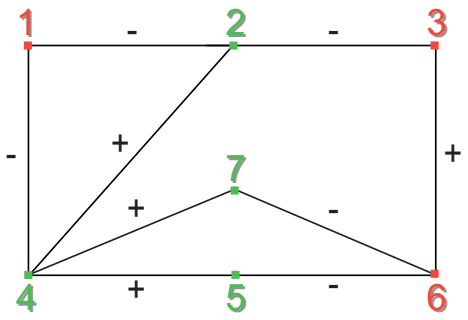

Abbildung 1.5: Beispiel eines balancierten S-Graphen mit der Partition $V_1 = \{1,3,6\}$ und $V_2 = \{2,4,5,7\}$.

Somit haben wir gesehen, dass sich die Knotenmenge balancierter S-Graphen in zwei disjunkte Teilmengen zerlegen lässt, so dass zwischen diesen Teilmengen nur negativ markierte Kanten verlaufen und positiv markierte Kanten nur aus Knoten innerhalb der entsprechenden Teilmengen bestehen. Dieses Wissen wird uns für Beweise in späteren Kapiteln sehr von Nutzen sein.

Am Ende dieses Abschnitts werden wir noch folgendes hilfreiche Lemma beweisen.

Lemma 1.15. *Ein beliebiger Graph $G = (V,E)$ ist bipartit $\Leftrightarrow$ der S-Graph $H = (G, \varphi)$ mit $\varphi(e) = -1 \; \forall \, e \in E$ ist balanciert.*

Beweis. $\Longrightarrow$ Sei $G = (V,E)$ bipartit. Nach Definition 1.6 existiert eine Partition der Knotenmenge V in zwei disjunkte Teilmengen V_1 und V_2, so dass für alle Kanten aus E die Endknoten in verschiedenen Teilmengen liegen und insbesondere Elemente aus der gleichen Teilmenge nicht adjazent sind. Setze also $\varphi(e) = -1$ für alle $e \in E$. Nach Theorem 1.14 ist H balanciert.

$\Longleftarrow$: Da H balanciert ist und somit nach Theorem 1.14 eine Zerlegung von V in zwei disjunkte Teilmengen V_1 und V_2 existiert und weiterhin $\varphi(e) = -1$ für alle $e \in E$ ist, existieren keine adjazenten Knoten innerhalb einer Teilmenge. Somit ist G bipartit. □

1.3 Zusammenfassung

Wir haben in diesem Kapitel nicht nur die notwendigen und benötigten Begrifflichkeiten zu Graphen erklärt, sondern auch etwas über den Aufbau und die Struktur der RNA erfahren. Wir haben die Frage motiviert, unter welchen Voraussetzungen zu mehreren gegebenen Sekundärstrukturen eine einzelne Sequenz existiert, welche zu diesen kompatibel ist. Um diese Frage beantworten zu können, haben wir verschiedene Begriffe der Graphentheorie definiert und sind hierbei u.a. genauer auf die mit Vorzeichen markierten Graphen eingegangen. Wenden wir uns nun im nächsten Kapitel der Darstellung von Sekundärstrukturen, einer erweiterten Darstellung, sowie dem Begriff der *Realisierung* zu.

2 Shapes und realisierende Sequenzen

Wir haben in den vorangegangen Kapiteln die notwendigen Begrifflichkeiten der Sekundärstruktur kennengelernt und sind detailliert auf Graphen, insbesondere mit Vorzeichen markierte Graphen, eingegangen. Es stellt sich natürlich die Frage, wieso wir dies so ausführlich behandelt haben. Die Antwort ist einfach. Wir werden die Sekundärstruktur der RNA mit der Hilfe von Graphen darstellen. Es sei bemerkt, dass es mehr als nur diese eine Repräsentation der Sekundärstruktur gibt. Ein Beispiel dafür ist die Darstellung als Klammerausdruck. Diese ist in der Abbildung 2.1 rechts unten dagestellt. Da wir in den folgenden Kapiteln nur diese beiden Repräsentationen benötigen, wollen wir auf weitere auch nicht eingehen. Zusätzliche Beispiele verschiedener Darstellungen findet man u.a. in [Fon02].

2.1 Repräsentation der Sekundärstruktur durch Shapes

Wir werden uns nun näher mit der Repräsentation von einer bzw. mehreren Sekundärstrukturen als Graphen beschäftigen und hierfür die Begriffe *Shapes* und *Shapegraphen* genauer erläutern. Die folgenden Definitionen sind, bis auf wenige Ausnahmen in leicht abgewandelter Form, aus [CLK+05] und [FHM+01] entnommen.

Definition 2.1. *Ein* Shape[1] *S mit Länge n ist ein Graph $S = (V_n, E)$ mit der Knotenmenge $V_n = \{v_1, \ldots, v_n\}$ und einer Menge E unabhängiger Kanten, so dass für alle $\{v_i, v_j\}_{i<j}, \{v_k, v_l\}_{k<l} \in E$ niemals die Ungleichung $i < k < j < l$ erfüllt ist, d.h. Pseudoknoten sind nicht erlaubt.*

Die Voraussetzung unabhängiger Kanten beschreibt die Tatsache, dass jede Base nur in maximal einer Basenpaarung auftreten darf. Somit ist ein Shape eine graphische Darstellung der Sekundärstruktur. Der Parameter θ spielt in unseren Betrachtungen keine weitere Rolle.

[1] aus dem engl.: Form, Gebilde

Im Folgenden wird vorausgesetzt, dass alle Shapes derselben Größe die gleiche Knotenmenge V_n besitzen. Die Abbildung 2.1 zeigt verschiedene Möglichkeiten die Sekundärstruktur als Graph darzustellen.

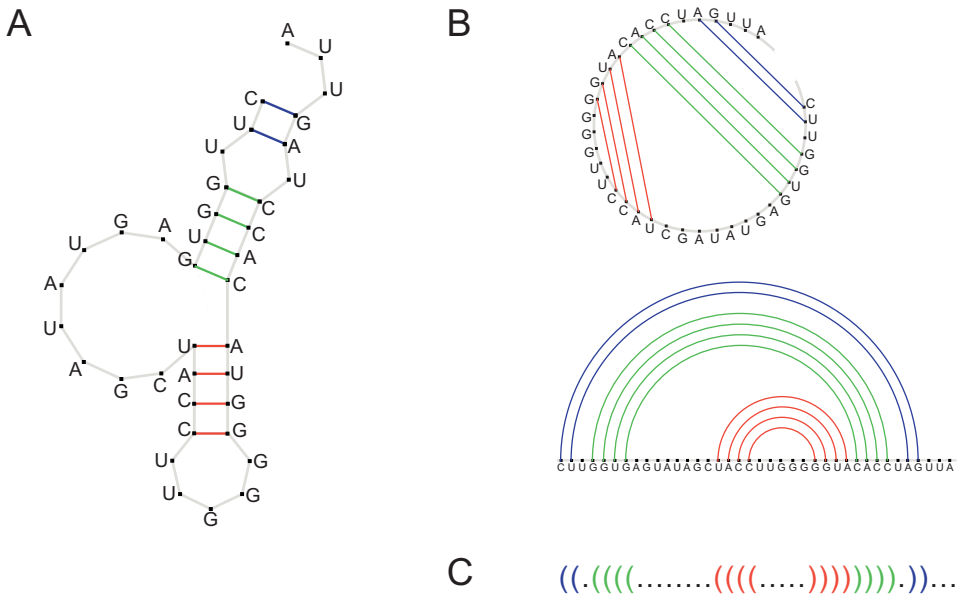

Abbildung 2.1: Mögliche Darstellungen ein und derselben Sekundärstruktur. A: Typische Darstellung der Sekundärstruktur. B: Kreis- und Geradenrepräsentation der Sekundärstruktur. Es sei erwähnt, dass die grauen Kanten zwischen Nukleotiden nicht als Kanten gewertet werden, sondern sie nur zur Visualisierung des Rückgrats der RNA dienen. C: Klammernotation. Hierbei stehen die Punkte für ungepaarte Positionen und zusammengehörige Klammern beschreiben die Positionen, welche gepaart sind.

Definition 2.2. *Mit* $SHAPE_n = \{\text{Shape: Shape der Länge } n\}$ *wird der Raum aller Shapes der Länge n bezeichnet.*

Definition 2.3. *Zu gegebenen Shapes* $S_1, \ldots, S_k \in SHAPE_n,\ k \in \mathbb{N}$ *definiert man den Graphen* $G(S_1, \ldots, S_k) := G(V_n, \cup_{i=1}^{k} E(S_i))$. *Dieser wird als* Shapegraph *bezeichnet.*

Ein erklärendes Beipiel ist in Abbildung 2.2 zu sehen. Bei unserer Betrachtungen werden Mehrfachkanten nur als eine Kante gezählt.

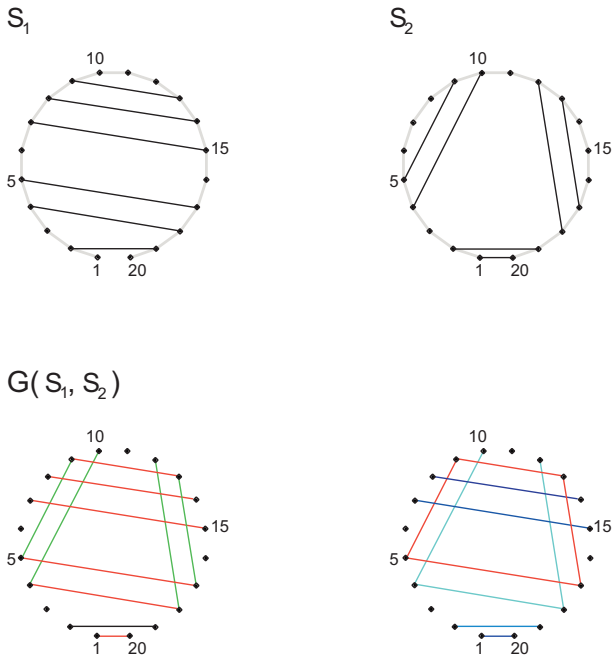

Abbildung 2.2: Oben: Kreisrepräsentation der Shapes S_1 und S_2. Unten: durch das Vereinen der Kantenmenge der Graphen S_1 und S_2 erhält man den Shapegraphen $G(S_1, S_2)$. Kanten, die nur in S_1 liegen, sind rot, Kanten, die nur in S_2 liegen, sind grün und Kanten, die in beiden Shapes liegen, sind schwarz markiert. Mehrfachkanten werden in dieser Betrachtung nur als eine Kante gezählt. Unten rechts: Pfade sind blau und Zyklen sind rot gekennzeichnet [FHM+01].

2.2 Bedingungen für die Realisierung von Shapes

Im letzten Abschnitt haben wir gesehen, wie wir die Sekundärstruktur mittels Graphen visualisieren. In den vorangegangenen Kapiteln haben wir festgestellt, unter welchen Paarungsregeln sich eine RNA-Nukleotidsequenz falten kann. Wir haben die Sekundärstruktur als eine Menge von Basenpaarungen betrachtet, die bestimmte Voraussetzungen erfüllt. Offensichtlich gibt es mehr als eine Sequenz, die sich, unter der Voraussetzung der Paarungsregel $\mathscr{B} = \{AU, UA, GU, UG, CG, GC\}$, in eine gegebenen Struktur falten kann. Wir wollen uns nun der Frage zuwenden, unter welchen mathematischen Voraussetzungen zu gegebenen Shapes, welche entsprechende Sekundärstrukturen visualisieren, eine einzelne RNA Sequenz exisitiert, die sich unter der Regel $\mathscr{B}$ in diese falten könnte. Falls diese Sequenz existiert, so bezeichnen wir die gegebenen Shapes durch diese Sequenz als *realisiert*. Es sei nocheinmal erwähnt, dass wir in unsere Betrachtung keine Energiewerte mit einfließen las-

sen. Das heißt, wenn wir davon sprechen, dass sich eine Sequenz unter der Paarungsregel $\mathscr{B}$ in eine bestimmte Struktur falten könnte, so meinen wir damit im allgemeinen keine Strukturen mit minimalen freien Energien. Formal bezeichnen wir eine Zeichenkette, die gegebene Shapes gleicher Länge *realisiert*, wie folgt.

Definition 2.4. *Eine Zeichenkette oder Sequenz* $s = s_1 \dots s_n \in \{A,C,G,U\}^n$ realisiert *einen Shape S der Länge* $n \Leftrightarrow \forall \{v_i, v_j\} \in E(S)$ *gilt:* $s_i s_j \in \mathscr{B} = \{AU, UA, GU, UG, CG, GC\}$*, d.h. dass die Buchstaben an den zugehörigen Positionen* i *und* j *entsprechend der Paarungsregel gewählt sind.*

Es sei bemerkt, dass wir im Folgenden, wenn wir von der Realisierung gegebener Shapes sprechen, ausschließlich den Begriff der Realisierung aus der der ebengenannten Definition 2.4 meinen, sofern nicht explizit etwas anderes erwähnt wird. Es ergibt sich nun aus dieser Definition das folgende Problem.

Problem 2.5 (Realisierungsproblem)**.** *Konstruiere für beliebige Shapes* $S_1, \dots, S_k$ *der Länge* n *eine Zeichenkette* $s = s_1 \dots s_n \in \{A,C,G,U\}^n$*, so dass alle Shapes durch diese eine Zeichenkette realisiert werden.*

Natürlich stellt sich zunächst die Frage, unter welchen Voraussetzungen eine solche Sequenz überhaupt existieren kann. Beantworten wir diese Frage zunächst für zwei Shapes mit Hilfe des Intersection Theorems[2].

Theorem 2.6 (Intersection Theorem)**.** [RSS97a, RSS97b] *Seien* S_1, S_2 *zwei beliebige Shapes der Länge n. Weiterhin bezeichne* $R(S)$ *die Menge aller Sequenzen, die den Shape S realisieren. Es gilt:*

$$R(S_1) \cap R(S_2) \neq \emptyset$$

Ein kurzer gruppentheoretischer Beweis findet sich in [RSS97a, RSS97b]. Ein alternativer Beweis kann in [FHM+01] nachvollzogen werden. Das Intersection Theorem beschreibt somit die Tatsache, dass zu zwei beliebigen Shapes S_1 und S_2 der Länge n, welche entsprechende Sekundärstrukturen visualisieren, immer eine Zeichenkette existiert, die beide Shapes realisiert. Im allgemeinen ist es jedoch nicht immer möglich drei oder mehr Shapes durch eine einzelne Zeichenkette zu realisieren. Man betrachte als Gegenbeispiel die Shapes S_1, S_2, S_3 der Länge 3 mit $V_3 = \{v_1, v_2, v_3\}$ und ihrer jeweiligen Kantenmenge $E(S_1) = \{\{v_1, v_2\}\}$, $E(S_2) = \{\{v_2, v_3\}\}$, $E(S_3) = \{\{v_1, v_3\}\}$ in Abbildung 2.3.

Es ist klar, dass eine Sequenz, die alle Kanten der Shapes realisiert auch alle Kanten in dem entsprechenden Shapegraphen realisiert.

[2] Schnittmengen Theorem

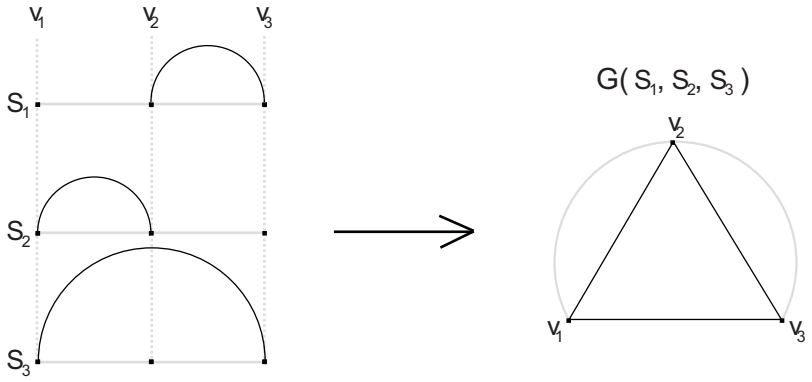

Abbildung 2.3: Drei Shapes, die sich nicht durch eine einzelne Zeichenkette realisieren lassen.

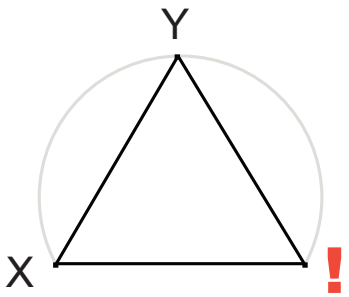

Abbildung 2.4: Um eine realisierende Zeichenkette zu konstruieren beginnt man mit einem beliebigen Buchstaben $X \in \{A,C,G,U\}$, für die darauffolgende Kante wählt man einen weiteren Buchstaben Y, so dass $XY \in \mathscr{B} = \{AU,UA,GU,UG,CG,GC\}$. Offensichtlich gibt es keine Möglichkeit in diesem Beispiel die dritte Kante zu realisieren.

Somit können wir anstatt der Frage:

"Unter welchen Voraussetzungen existiert eine Sequenz, die alle gegebenen Shapes realisiert?"

auch folgende Frage stellen:

"Welche Voraussetzungen müssen an den Shapegraphen gestellt werden, so dass eine Sequenz existiert, die alle gegebenen Shapes realisiert?"

Betrachten wir nun also, welche Eigenschaften der Shapegraph $G(S_1,\ldots,S_k)$ aufweisen muss, damit die entsprechenden Shapes $S_1,\ldots,S_k$ der Länge n durch eine einzelne Zeichen-

kette $s \in \{A,C,G,U\}^n$ realisiert werden können. Wie das Generalized Intersection Theorem[3] zeigt, wird die Realisierung von mehr als zwei Shapes dann und nur dann möglich sein, wenn alle Zyklen, die im Graphen $G(S_1,\ldots,S_k)$ enthalten sind, gerader Länge sind. Es sei erwähnt, dass wir nicht das gesamte Theorem benennen, sondern nur den für unsere weiteren Betrachtungen notwendigen Punkt. Das vollständige Theorem kann in [FHM+01] nachgelesen werden. Der Beweis ist in kurzer Form und aus [FHM+01] entnommen.

Theorem 2.7 (Generalized Intersection Theorem)**.** *Seien $S_1,\ldots,S_k$ beliebige Shapes der Länge n. Weiterhin bezeichne $R(S)$ die Menge aller Sequenzen, die den Shape S realisieren. Es gilt:*

$$\bigcap_{i=1}^{k} R(S_i) \neq \emptyset \iff G(S_1,\ldots,S_k) \textit{ ist bipartit.}$$

Beweis. $\Longleftarrow$: Wenn der Graph $G(S_1,\ldots,S_k)$ bipartit ist, so exisitiert eine Partition der Knotenmenge V in zwei disjunkte Teilmengen V_1 und V_2, so dass innerhalb einer Teilmenge keine zueinander adjazenten Knoten existieren. Demzufolge können wir einen beliebigen Buchstaben $X \in \{A,C,G,U\}$ für entsprechende Positionen in der Zeichenkette für die Knoten der einen Partition wählen. Für die Positionen der Knoten der anderen Partition wählt man einen Buchstaben $Y \in \{A,C,G,U\}$, so dass $XY \in \mathscr{B} = \{AU,UA,GU,UG,CG,GC\}$

$\Longrightarrow$: Nach Lemma 1.7 gilt, dass ein Graph dann und nur dann bipartit ist, wenn in dem Graph keine Zyklen ungerader Länge existieren. Deswegen werden wir zeigen, dass ungerade Zyklen nicht durch das Alphabet $\{A,C,G,U\}$ mittels der Paarungsregel $\mathscr{B}$ realisiert werden können. Betrachten wir den Graphen

$$A - U - G - C.$$

Wenn wir eine Zeichenkette für einen beliebigen Zyklus C konstruieren wollen, so müssen wir den Kanten in dem ebengenannten Graphen folgen. Wenn wir mit einem beliebigen Buchstaben $X \in \{A,C,G,U\}$ beginnen, so kann sich dieser folglich erst nach einer geraden Anzahl von Kanten entlang des Zyklus wiederholen. Dies schließt auch mit ein, dass man X wieder erreicht, nachdem man alle Kanten des Zyklus durchlaufen hat. Ungerade Zyklen können demnach nicht durch eine Zeichenkette realisiert werden, woraus die Behauptung folgt. □

Folgende Schlussfolgerungen können wir aus diesem Theorem ziehen.

[3] Allgemeines Schnittmengen Theorem

Korollar 2.8. *Mit Hilfe des Lemmas 1.7 erhält man:*

- *Beliebige Shapes $S_1,\ldots,S_k$ gleicher Länge können realisiert werden $\Leftrightarrow$ der Graph $G(S_1,\ldots,S_k)$ enthält keine Zyklen mit einer ungeraden Anzahl von Kanten.*

Da bei einer Bipartition der Knotenmenge keine adjazenten Knoten innerhalb einer der beiden Teilmenge existieren, gilt des Weiteren:

- *Beliebige Shapes $S_1,\ldots,S_k$ gleicher Länge n können durch eine Sequenz $s \in \{A,C,G,U\}^n$ mit der Paarungsregel $\mathscr{B}$ realisiert werden $\Leftrightarrow$ die Shapes $S_1,\ldots,S_k$ können durch eine binäre Zeichenkette realisiert werden (Definition 2.9).*

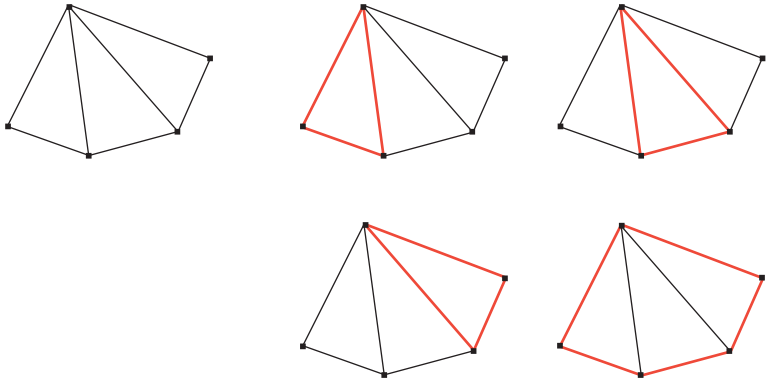

Abbildung 2.5: Darstellung eines Shapegraphen mit sieben Kanten. Die Zyklen ungerader Länge sind jeweils rot gekennzeichnet

Definition 2.9. *Eine binäre Zeichenkette $s = s_1\ldots s_n \in \{0,1\}^n$ realisiert einen Shape S der Länge $n \Leftrightarrow \forall \{v_i,v_j\} \in E(S)$ gilt: $s_i \neq s_j$*

Da eine Zeichenkette über dem Alphabet $\{A,C,G,U\}$ dann und nur dann die Shapes $S_1,\ldots,S_k$ realisieren wird, wenn eine binäre Zeichenkette diese Shapes realisiert, kann man das Realisierungsproblem vereinfachen.

Problem 2.10 (binäres Realisierungsproblem)**.** *Konstruiere für beliebige Shapes $S_1,\ldots,S_k$ der Länge n genau eine binäre Zeichenkette $s = s_1\ldots s_n$, so dass alle Shapes realisiert werden.*

Betrachtet man jetzt den Graphen $G(S_1,\ldots,S_k)$, so gilt auch für diesen, dass die Shapes dann und nur dann von einer binären Zeichenkette s realisiert werden können, wenn für alle Kanten $\{v_i,v_j\} \in G(S_1,\ldots,S_k)$ gilt: $s_i \neq s_j$.

Im Folgenden wird der Graph $G(S_1,\ldots,S_k)$ als *realisiert* oder *realisierbar* bezeichnet, falls die entsprechenden Shapes $S_1,\ldots,S_k$ durch eine einzelne Zeichenkette nach Definition 2.4 bzw. 2.9 realisiert werden können.

2.3 Erweiterte Shapes und Pseudokanten

Wir werden in diesem Abschnitt die Menge der Sequenzen $R(S)$, die einen gegebenen Shape S realisieren, durch das Erstellen neuer Paarungsregeln, spezifizieren. Warum wir dies tun wollen, sollen die Beispiele in Abbildung 2.6 und 2.7 zeigen. In Abbildung 2.6 ist ein sehr einfacher Shape mit sieben Knoten und einer Kanten zu sehen. Eine mögliche Zeichenkette, die diesen Shape realisiert ist etwa $s = AACCCUU$. Allerdings ist zu erkennen, dass diese Zeichenkette auch die unteren beiden Shapes in dieser Abbildung realisiert. Es sei erwähnt, dass ein Shape ohne Kanten von jeder beliebigen Zeichenkette realisiert werden kann, weshalb wir diesen trivialen Fall nicht in dieser Abbildung dargestellt haben.

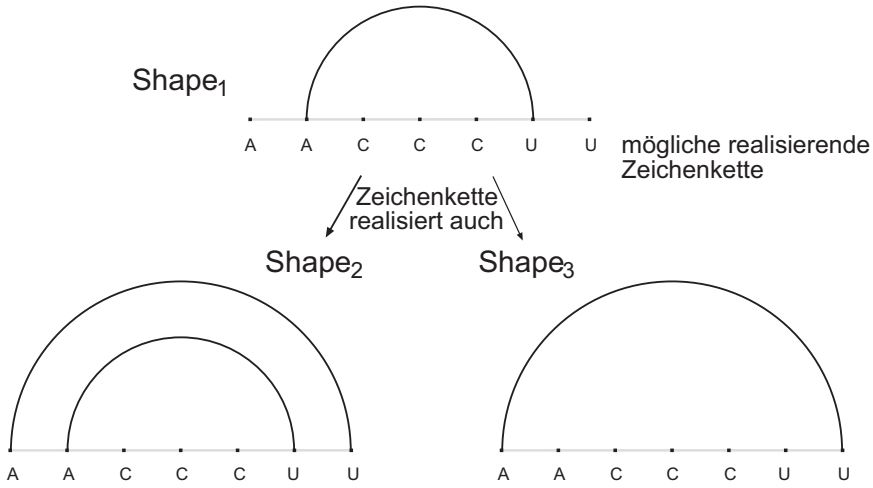

Abbildung 2.6: Beispiel einer Zeichenkette, welche die drei abgebildeten Shapes realisiert.

Sehen wir uns ein etwas komplizierteres Beispiel an, welches in Abbildung 2.7 dargestellt ist. Es ist links ein Shape zu sehen, zu dem eine Zeichenkette gesucht wird, die diesen realisiert. Eine mögliche Sequenz ist in der Mitte der Abbildung zu sehen. Diese Sequenz realisiert aber auch den Shape rechts in der Abbildung 2.7. Wir wir im ersten Kapitel erfahren haben, ist die Funktion der RNA sehr eng mit ihrer Struktur verknüpft. Demzufolge kann es wünschenswert sein, Sequenzen zu finden, die nur bestimmte Strukturen annehmen können bzw. bestimmte Strukturen *nicht* annehmen können. So kann es erstrebenswert sein, Leitern von Kanten zu restriktieren, um zum Beispiel bestimmte Schleifen zu erhalten. In dem bisherigen Realisierungsbegriff wurden nur Kanten, d.h. die gepaarten Positionen der Sekundärstruktur berücksichtigt. Den ungepaarten Knoten ist eine eher passive Rolle in diesem

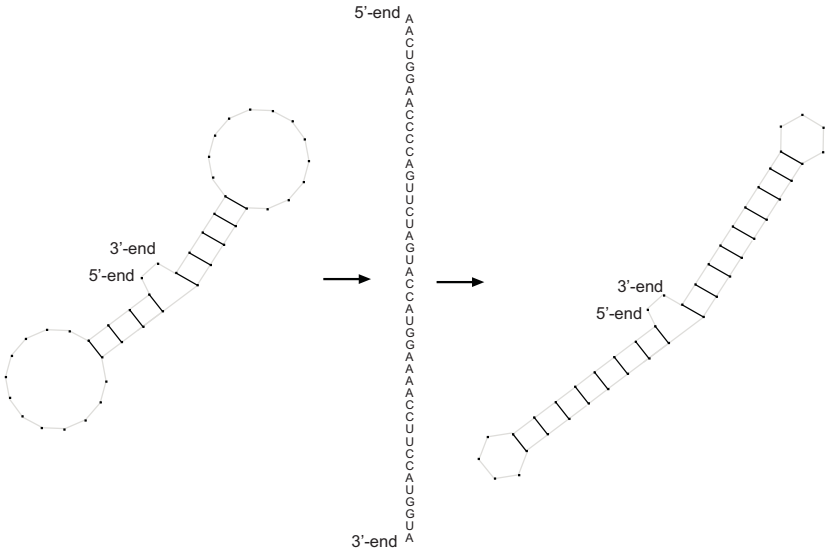

Abbildung 2.7: Links ist ein Shape zu sehen, zu dem eine Zeichenkette gesucht ist, welche diesen realisiert. Ein mögliche Sequenz ist in der Mitte der Abbildung zu sehen. Allerdings realisiert diese Zeichenkette auch den Shape rechts im Bild.

Kontext zugefallen. Nutzen wir also die durch diese Betrachtung resultierende Möglichkeit, auch Bedingungen an Knoten, die bisher keine Kante gebildet haben, zu knüpfen und somit auch diesen Knoten Bedeutung zukommen zu lassen, um letztendlich den Realisierungsbegriff zu restriktieren, mit dem Ziel Sequenzen zu finden, die strukturerhaltender sind, als die bisherigen.

Um dies zu verwirklichen werden wir zunächst neue Kanten, die wir *Pseudokanten* nennen werden, in die Shapes einfügen, um danach einen neuen Realisierungsbegriff zu modellieren. Das Vorgehen kann wie folgt formal beschrieben werden.

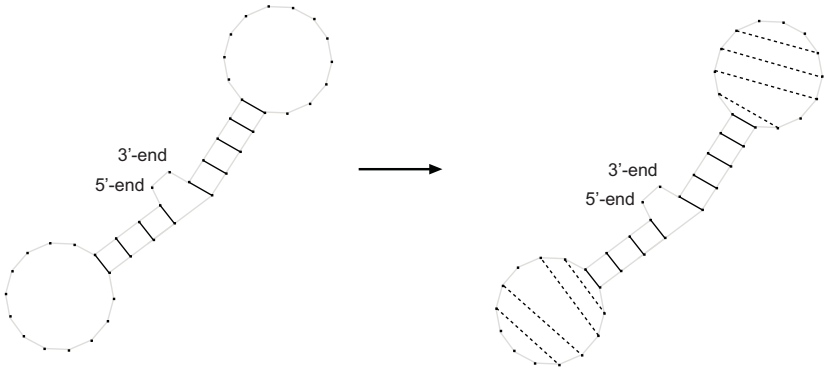

Abbildung 2.8: Beispiel eines Shapes in den Pseudokanten gesetzt wurden (Strichlinien), um bestimmten Positionen zu verbieten, Paarungen einzugehen.

Definition 2.11. *Sei $S = G(V_n, E)$ ein Shape der Länge n mit der Kantenmenge $E(S) := E$ und sei $V'(S) = \{v_i \in V_n \mid \nexists v_j \in V_n,$ so dass $\{v_i, v_j\} \in E(S)\}$, d.h. die Menge aller Knoten, die keine Kanten bilden.*

Jetzt werden beliebige Knoten $v_i, v_j \in V'(S)$, durch neue Kanten verbunden.

Das verbinden der Knoten mit neuen Kanten ist aber eingeschränkt auf die Tatsache, dass die Eigenschaften eines Shapes für S erhalten bleiben müssen, d.h. Pseudoknoten sind nicht erlaubt und für alle Knoten $v \in V_n$ ist der Knotengrad $deg(v) \leq 1$.

Diese eingefügten Kanten werden Pseudokanten *genannt.*

Sei $\hat{E}(S)$ die Menge aller eingefügten Pseudokanten in S. Der aus dem Shape S hervorgehende Graph $\hat{S} = (V_n, E(S), \hat{E}(S)) := (V_n, \tilde{E} := E(S) \cup \hat{E}(S), \gamma)$ mit

$$\gamma : \tilde{E} \to \{0,1\}$$

so dass

$$\gamma(\{v_i, v_j\}) = \begin{cases} 0 & , falls \{v_i, v_j\} \in E(S) \\ 1 & , falls \{v_i, v_j\} \in \hat{E}(S) \end{cases}$$

wird als Shapeerweiterung *von S oder auch* erweiterter Shape *bezeichnet.*

Es sei bemerkt, dass γ keine Gewichtsfunktion ist, sondern nur der Markierung der Kanten als Pseudokanten und regulären Kanten dient.

Definition 2.12. *Seien $\hat{S}_1 = (V_n, E(S_1), \hat{E}(S_1)), \ldots, \hat{S}_k = (V_n, E(S_k), \hat{E}(S_k))$ die aus den Shapes $S_1, \ldots, S_k$ hervorgegangenen Shapeerweiterungen . Der Graph $G(\hat{S}_1, \ldots, \hat{S}_k)$ ist definiert als $(V, E, \hat{E}) := (V_n, \tilde{E} := \cup_{i=1}^{k}(E(S_i) \cup \hat{E}(S_i)), \gamma)$ mit*

$$\gamma : \tilde{E} \to \{0,1,2\}$$

so dass

$$\gamma(\{v_i, v_j\}) = \begin{cases} 0 & , falls \{v_i, v_j\} \in \tilde{E} \setminus \cup_{i=1}^{k}(\hat{E}(S_i)) \\ 1 & , falls \{v_i, v_j\} \in \tilde{E} \setminus \cup_{i=1}^{k}(E(S_i)) \\ 2 & , sonst \end{cases}$$

Falls $\gamma(\{v, w\}) = 2$, d.h. es existieren Kanten $(v, w) \in E(S_l) \cap \hat{E}(S_m)$, so bezeichnen wir diese als Überlagerung. *Des Weiteren werden wir diesen Graphen als* Shapegraphen erweiterter Shapes *bezeichnen.*

Zur Vereinfachung werden wir von *regulären Kanten* sprechen, wenn gilt $\gamma(e) = 0$, sowie weiterhin von *Pseudokanten*, falls $\gamma(e) = 1$. In den folgenden Abbildungen werden für die vereinfachte Darstellung reguläre Kanten mit durchgezogenen Linien und Pseudokanten mit Strichlinien dargestellt (Abbildung 2.9).

Bisher wurden Mehrfachkanten als eine Kante betrachtet. Dies wird auch im Weiteren so bleiben, ebenso werden Mehrfach-Pseudokanten als eine gezählt. Kanten in der Schnittmenge von regulären und Pseudokanten werden als zwei verschiedene betrachtet. Wir bezeichnen diese Form von Kanten als *Überlagerung*.

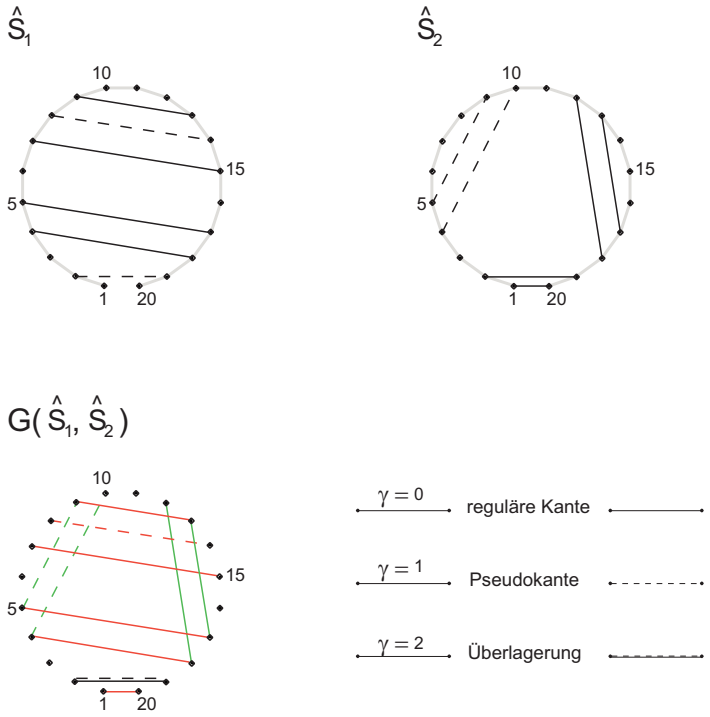

Abbildung 2.9: Darstellung von regulären Kanten, Pseudokanten und Überlagerungen in Shapes und Shapegraphen erweiterter Shapes

2.4 Realisierung erweiterter Shapes

Nachdem wir nun die Shapes mit Hilfe von Pseudokanten erweitert haben, können wir den Realisierungsbegriff restriktieren. Der Einfachheithalber wollen wir diese Spezifizierung zunächst für binäre Zeichenketten vornehmen.

Definition 2.13. *Eine binäre Zeichenkette* $s = s_1 \dots s_n \in \{0,1\}^n$ *realisiert einen erweiterten Shape* $\hat{S}$ *der Länge* $n \iff$

1. $\forall \{v_i, v_j\} \in E(S)$ *gilt:* $s_i \neq s_j$, *d.h.* $s_i s_j \in \mathscr{B} := \{01, 10\}$.

2. $\forall \{v_i, v_j\} \in \hat{E}(S)$ *gilt:* $s_i = s_j$, *d.h.* $s_i s_j \in \mathscr{B}_{PsK} := \{00, 11\}$.

Wie folgendes Beispiel zeigt, gilt das Intersection Theorem (Theorem 2.6) für zwei Shapeerweiterungen entsprechend nicht mehr. Man betrachte die beiden erweiterten Shapes $\hat{S}_1$ und $\hat{S}_2$ in Abbildung 2.10 mit der Menge regulärer Kanten $E(S_1) = \{\{v_i, v_j\}\}$, sowie der Menge von Pseudokanten $\hat{E}(S_2) = \{\{v_i, v_j\}\}$.

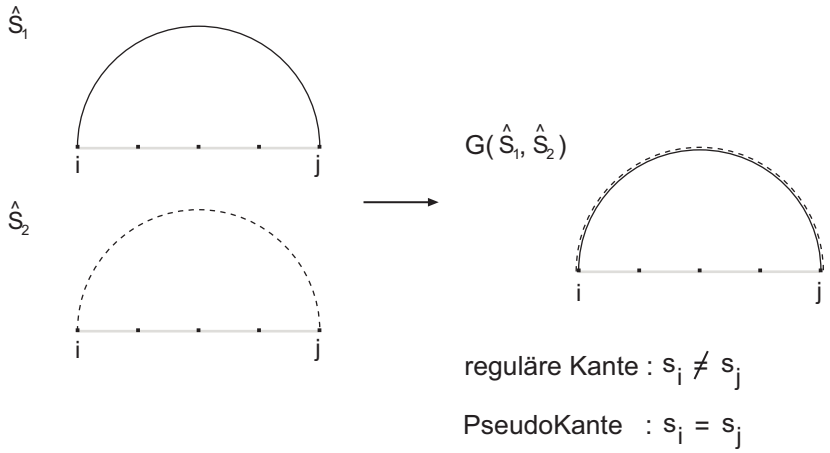

Abbildung 2.10: Beispiel, in dem für zwei erweiterte Shapes keine Sequenz gefunden werden kann, die beide realisiert.

Auch hier wird schnell klar, wenn eine binäre Zeichenkette die Kanten aller erweiterten Shapes realisiert, so realisiert sie auch alle Kanten in dem Shapegraphen erweiterter Shapes. Allerdings benötigen wir neue Voraussetzungen an den Shapegraphen erweiterter Shapes, so dass die entsprechenden Shapeerweiterungen durch eine einzelne binäre Zeichenkette realisiert werden können. Dafür benötigen wir die folgende Definition und die folgenden Lemmata.

Definition 2.14. *Ein* Zeichenwechsel *findet in einer binären Zeichenkette* $s_1 \dots s_n$ *statt, wenn gilt:* $s_i \neq s_{i+1},\ i \in \{1, \dots, n-1\}$.

Lemma 2.15. *Für einen Pfad* $[v_1, e_1, \dots, v_n, e_n, v_{n+1}]$ *ohne Pseudokanten und eine binäre Zeichenkette* $s = s_1 \dots s_{n+1}$, *die den Pfad realisiert, gilt:*

1. *In* $s = s_1 \dots s_{n+1}$ *finden n Zeichenwechsel statt.*

2. *Der Pfad ist*

   - gerader *Länge n* $\Leftrightarrow\ \ s_1 = s_{n+1}$.

- ungerader *Länge* $n \Leftrightarrow \quad s_1 \neq s_{n+1}$.

Beweis. Da per Definition gilt: $s_i \neq s_{i+1}\ \forall i \in 1,\ldots,n$, folgt Behauptung 1.
Da gilt $s_1 \neq s_2$ und $s_1 = s_3$, sowie $s_i \in \{0,1\} \forall i = 1,\ldots,n$, folgt per Induktion: $s_1 \neq s_{2k}$ und $s_1 = s_{2k+1}$ für $k \in \mathbb{N}$. Insbesondere besteht ein Pfad mit $n+1 = 2k$ Knoten aus einer ungeraden Anzahl regulärer Kanten, sowie ein Pfad mit $n+1 = 2k+1$ Knoten aus einer geraden Anzahl regulärer Kanten. Somit folgt die zweite Behauptung. □

Lemma 2.16. *Sei $s = s_1 \ldots s_n$ eine binäre Zeichenkette, die die erweiterten Shapes $\hat{S}_1,\ldots,\hat{S}_k$ des Graphen $G(\hat{S}_1,\ldots,\hat{S}_k)$ realisiert, so gilt: $s^c = s_1^c \ldots s_n^c$ mit*

$$s_i^c = \begin{cases} 0 & \text{, falls } \ s_i = 1 \\ 1 & \text{, falls } \ s_i = 0 \end{cases}$$

realisiert die erweiterten Shapes $\hat{S}_1,\ldots,\hat{S}_k$.

Beweis. Da durch die Vertauschung von $'1'$ und $'0'$ die Bedingungen in s^c für die Realisierung erhalten bleiben, folgt die Behauptung. □

Unser Ziel wird es sein herauszufinden, unter welchen Voraussetzungen mehrere erweiterte Shapes $\hat{S}_1,\ldots,\hat{S}_k$ durch eine einzelne binäre Zeichenkette realisiert werden können. Um dies zu zeigen benötigen wir jedoch vorher das nächste Lemma.

Lemma 2.17. *Sei $G(\hat{S}_1,\hat{S}_2)$ der Graph zweier erweiterter Shapes und sei $K = \{K_1,\ldots,K_l\}$ die Menge aller zusammenhängenden Komponenten des Graphen sowie E_i die Menge der Kanten des Elementes $K_i \in K,\ i = 1,\ldots,l$. Seien weiterhin $\sigma_i : E_i \to E_i$ beliebige Permutationen auf der Menge der Kanten des Elementes $K_i \in K,\ i = 1,\ldots,l$ und G' der durch diese Permutationen aus $G(\hat{S}_1,\hat{S}_2)$ hervorgehende Graph, so gilt:*

$$G(\hat{S}_1,\hat{S}_2) \text{ lässt sich realisieren} \Leftrightarrow G' \text{ lässt sich realisieren}$$

Beweis. Sei $s = s_1 \ldots s_n$ eine binäre Zeichenkette, die den Graphen $G(\hat{S}_1,\hat{S}_2)$ realisiert. Falls für alle $i = 1,\ldots,l$ die Permutationen σ_i ausschließlich reguläre Kanten auf reguläre Kanten bzw. Pseudokanten auf Pseudokanten abbilden, so realisiert die Zeichenkette s auch G', da für alle $v_i, v_j \in V$, die vorher durch reguläre Kanten bzw. Pseudokanten verbunden waren, auch jetzt noch gilt, dass sie durch reguläre Kanten bzw. Pseudokanten verbunden sind, demzufolge muss vorher wie nachher gelten $s_i \neq s_j$ bzw. $s_i = s_j$.

Also bleibt der Fall zu betrachten, dass reguläre Kanten mit Pseudokanten vertauscht werden. Da nach Definition 2.11 der Grad der Knoten in $G(\hat{S}_1,\hat{S}_2)$ maximal 2 ist, folgt, dass die

Elemente aus K entweder Pfade oder Zyklen sind. Insbesondere sind alle Elemente aus K paarweise verschieden, d.h. sie besitzen keine gemeinsamen Knoten oder Kanten. Es genügt also die Behauptung für ein beliebiges $K_i = [v_{i_1}, e_{i_1}, \ldots, e_{i_m}, v_{i_m}] \in K$ zu zeigen, da sich jede Permutation der Kanten aus $E_i = \{e_{i_1}, \ldots, e_{i_m}\}$ ausschließlich auf die Umgestaltung der Teilsequenz $\tilde{s} = s_{i_1} \ldots s_{i_m}$ der ursprünglichen binären Zeichenkette s auswirkt, um $G(\hat{S}_1, \hat{S}_2)$ nach der Permutation mit der ersetzten neuen Teilsequenz $\tilde{s}' = s'_{i_1} \ldots s'_{i_m}$ zu realisieren.

Sei also $K_i = [v_{i_1}, e_{i_1}, \ldots, e_{i_m}, v_{i_m}]$ ein beliebiges Element aus K und $\tilde{s} = s_{i_1} \ldots s_{i_m}$ die zugehörige Teilsequenz der realisierenden Zeichenkette s. Seien weiterhin $e_{i_j} = \{v_{i_j}, v_{i_{j+1}}\}$, $e_{i_{j'}} = \{v_{i_{j'}}, v_{i_{j'+1}}\} \in E_i$ zwei zu vertauschende Kanten (o.B.d.A. sei $i_j < i_{j'}$). Im Folgenden wird gezeigt, dass die aus der ursprünglichen Teilsequenz $\tilde{s}$ hervorgehende Zeichenkette

$$\tilde{s}' = s_{i_1} \ldots s_{i_j} s^c_{i_{j+1}} \ldots s^c_{i_{j'}} s_{i_{j'+1}} \ldots s_{i_m}$$

mit

$$s^c_l = \begin{cases} 0 & \text{, falls } s_l = 1 \\ 1 & \text{, falls } s_l = 0 \end{cases}$$

, für alle $l = i_{j+1}, \ldots, i_{j'}$ eine Realisierung für $K_i \in K$ nach dem Vertauschen der beiden Kanten e_{i_j} und $e_{i_{j'}}$ ist.
Sei nun $e_{i_j} = \{v_{i_j}, v_{i_{j+1}}\}$ eine reguläre Kante (bzw. Pseudokante). Demzufolge gilt, dass $s_{i_j} \neq s_{i_{j+1}}$ (bzw. $s_{i_j} = s_{i_{j+1}}$) . Nach dem Vertauschen der regulären Kante mit der Pseudokante (bzw. der Pseudokante mit der regulären Kante) muss gelten $s_{i_j} = s_{i_{j+1}}$ (bzw. $s_{i_j} \neq s_{i_{j+1}}$), ersetze also $s_{i_{j+1}}$ durch $s^c_{i_{j+1}}$. Daraus folgend müssen nun auch für den unveränderten Teilpfad $[v_{i_{j+1}}, \ldots, v_{i_{j'}}]$ alle s_l $(l = i_{j+2}, \ldots, i_{j'})$ durch s^c_l ersetzt werden, damit die Bedingungen an die binäre Zeichenkette, die diesen Teilpfad realisiert, erhalten bleiben (Lemma 2.16). Da vor der Vertauschung galt, $s_{i_{j'}} = s_{i_{j'+1}}$ (bzw. $s_{i_{j'}} \neq s_{i_{j'+1}}$) und dann $s_{i_{j'}}$ durch $s^c_{i_{j'}}$ ersetzt wurde, gilt nun $s^c_{i_{j'}} \neq s_{i_{j'+1}}$ (bzw. $s^c_{i_{j'}} = s_{i_{j'+1}}$). Dies ist die gewünschte Forderung an die nun ersetzte reguläre Kante (bzw. Pseudokante) $e_{i_{j'}} = \{v_{i_{j'}}, v_{i_{j'+1}}\}$. Also realisiert $\tilde{s}' = s_{i_1} \ldots s_{i_j} s^c_{i_{j+1}} \ldots s^c_{i_{j'}} s_{i_{j'+1}} \ldots s_{i_m}$ das gewählte $K_i \in K$. Da jede Permutation σ_i der Kanten eines beliebigen $K_i \in K$ durch die endliche Hintereinanderausführung von Vertauschungen erzeugt werden kann, folgt die Behauptung. Die Rückrichtung zeigt man analog. □

Dies gilt im allgemeinen nicht für den Graphen $G(\hat{S}_1, \ldots, \hat{S}_k)$ falls der maximale Knotengrad $\max deg(v) > 2$, wie das Beispiel in Abbildung 2.11 zeigt.

Schauen wir jetzt unter welchen Voraussetzungen mehr als zwei erweiterte Shapes realisiert werden können.

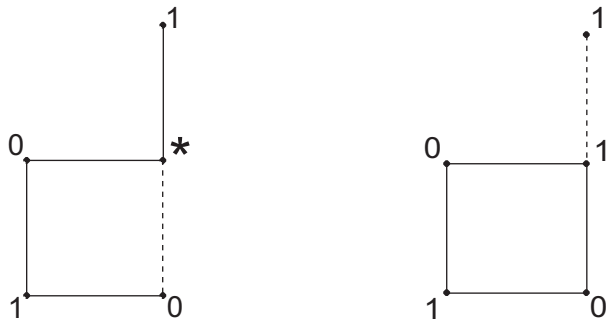

Abbildung 2.11: Links: Shapegraph erweiterter Shapes, dessen Kanten nicht durch eine einzelne binäre Zeichenkette realisiert werden können. Rechts: Nach der Permutation der Kanten sind die erweiterten Shapes durch eine binäre Zeichenkette realisierbar.

Theorem 2.18. *Beliebige erweiterte Shapes $\hat{S}_1,\ldots,\hat{S}_k$ der Länge n können von einer binären Zeichenkette realisiert werden* $\Leftrightarrow$

1. $E(S_l)\cap\hat{E}(S_m)=\emptyset$ *mit* $l\neq m\in\{1,\ldots,k\}$*, d.h. es exitieren keine Überlagerungen und*
2. *der Graph* $G(\hat{S}_1,\ldots,\hat{S}_k)$ *enthält keine Zyklen mit einer ungeraden Anzahl von regulären Kanten.*

Beweis. $\Longrightarrow$ (*Beweis durch Widerspruch*)
Seien $\hat{S}_1,\ldots,\hat{S}_k$ realisiert durch die binäre Zeichenkette $s=s_1\ldots\ s_n$.
Nehmen wir an, es existiert ein Kante $\{v_i,v_j\}\in E(S_l)\cap\hat{E}(S_m)$ mit $l\neq m\in\{1,\ldots,k\}$, dann muss gelten: $s_i\neq s_j$ für die reguläre Kante $\{v_i,v_j\}\in E(S_k)$ und auch $s_i=s_j$ für die Pseudokante $\{v_i,v_j\}\in\hat{E}(S_l)$, woraus der Widerspruch folgt.

Nehmen wir nun an, es existiert ein Zyklus $C=[v_{i_1},e_{i_1}\ldots,v_{i_m},e_{i_m},v_{i_1}]$ ohne Überlagerungen der Länge m mit einer ungeraden Anzahl $0<m_1\leq m$ von regulären Kanten und eine binäre Zeichenkette s, welche die Shapes $\hat{S}_1,\ldots,\hat{S}_k$ realisiert.
Falls m ungerade ist und $m_1=m$, d.h. der Zyklus beinhaltet ausschließlich reguläre Kanten, so folgt aus Theorem 2.7, sowie dem Korollar 2.8 der Widerspruch.
Sei also $0<m_1<m$ und $s'=s_{i_1}\ldots s_{i_m}$ die Teilsequenz von s, die diesen Zyklus realisiert. Nach Lemma 2.17 wissen wir, dass beliebige Permutationen der Kanten innerhalb zusammenhängender Komponenten bei einem Graphen mit einem Knotengrad $deg(v)\leq 2$ keine Auswirkung auf die Realisierbarkeit haben. Es ist klar, dass dieser Fall für C nicht gelten muss. Wenn aber die Teilsequenz $s'=s_{i_1}\ldots s_{i_m}$ den Zyklus $[v_{i_1},e_{i_1}\ldots,v_{i_m},e_{i_m},v_{i_1}]$ realisiert, so realisiert s' folglich auch den Zyklus isoliert, d.h. s' muss den Zyklus C auch für den Fall

$deg(v) = 2\ \forall v \in C$ realisieren. Somit genügt es C isoliert zu betrachten und dies mit Hilfe des Lemmas 2.17 zum Widerspruch zu führen.

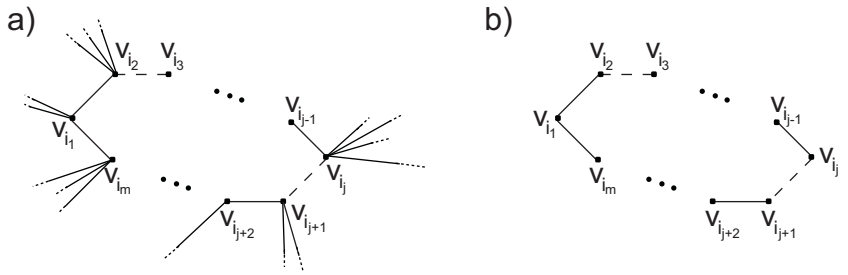

Abbildung 2.12: Wenn $s' = s_{i_1} \dots s_{i_m}$ den Zyklus in a) realisiert, dann realisiert s' auch den Zyklus in b)

Sei nun σ eine Permutation auf der Menge der Kanten des Zyklus C. Sei weiterhin C' der aus C durch die Permutationen hevorgehende Zyklus, so dass der erste Teilpfad von C' $[v'_{i_1}, \dots, v'_{i_{m_1+1}}]$ nur aus regulären Kanten, sowie der zweite Teilpfad $[v'_{i_{m_1+1}}, \dots, v'_{i_1}]$ nur aus Pseudokanten besteht. Sei weiterhin $s_\sigma = s'_{i_1} \dots s'_{i_m}$ die veränderte Teilseqenz, die den den Zyklus C' realisiert. In Lemma 2.17 haben wir gesehen, wie wir diese konstruieren können. Da der erste Teilpfad regulärer Kanten ungerader Länge ist, gilt nach Lemma 2.15, dass $s'_{i_1} \neq s'_{i_{m_1+1}}$.

Für die Realisierung der nun folgenden Pseudokante gilt, dass $s'_{i_{m_1+1}} = s'_{i_{m_1+2}}$. Insbesondere muss für den gesamten Teilpfad aus Pseudokanten gelten, dass $s'_{i_{m_1+1}} = s'_{i_{m_1+2}} = \dots = s'_{i_m} = s'_{i_1}$ und somit $s'_{i_1} = s'_{i_{m_1+1}}$, woraus der Widerspruch folgt.

$\Longleftarrow$: Um zu zeigen, dass die erweiterten Shapes $\hat{S}_1, \dots, \hat{S}_k$ durch eine binäre Zeichenkette realisierbar sind, wenn der Graph $G(\hat{S}_1, \dots, \hat{S}_k)$ die gewünschten Eigenschaften 1. und 2. besitzt, werden wir als erstes $G(\hat{S}_1, \dots, \hat{S}_k)$ zu einem S-Graphen $H = (G, \varphi)$ erweitern und zeigen, dass dieser balanciert ist. Dazu setzen wir

$$\varphi(e) = \begin{cases} +1 & \text{, falls } e \in \hat{E}(G) \\ -1 & \text{, sonst} \end{cases}$$

Das heißt Pseudokanten sind jetzt positiv und reguläre Kanten negativ markiert. Da nach Voraussetzung, falls Zyklen vorhanden sind, nur Zyklen mit einer geraden Anzahl von regulären Kanten existieren, sind alle Zyklen positiv und somit ist H balanciert. Nach Theorem 1.14 existiert also eine Zerlegung von $V(H)$ in zwei disjunkte Teilmenge V_1 und V_2, so dass Knoten aus V_1 und V_2 nur reguläre Kanten verbinden und weiterhin innerhalb einer jeden

Teilmenge ausschließlich Pseudokanten liegen. Es genügt jetzt also für alle Elemente $v_i \in V_1$ das entsprechende $s_i = 0$, sowie für alle $v_j \in V_2$ das entsprechende $s_j = 1$ zu setzen, um eine binäre Zeichenkette zu konstruieren, die alle erweiterten Shapes realisiert. □

Wir haben mit diesem Theorem nicht nur gezeigt, unter welchen Voraussetzungen an den Shapegraphen die entsprechenden erweiterten Shapes durch eine einzelne Zeichenkette realisierbar sind, sondern auch, dass die Länge der enthaltenen Zyklen keine Rolle spielt und somit nur die Anzahl der regulären Kanten in den Zyklen entscheidend ist. Ein sich innerhalb dieser Betrachtungen ergebendes Lemma, welches die Voraussetzung für die Realisierung von zwei Shapes beschreibt, ist das folgende.

Lemma 2.19. *Sei $G(\hat{S}_1, \hat{S}_2)$ der Shapegraph zweier erweiterter Shapes. Sei weiterhin G' der Graph, der aus $G(\hat{S}_1, \hat{S}_2)$ durch das Ersetzen aller Pseudokanten $e = \{u, w\}$ mit zwei neuen regulären Kanten $\{u, e\}, \{e, w\}$ und einem neuen Knoten e, der nicht in $V(G)$ enthalten ist, hervorgeht (Abbildung 2.13).*

Zwei beliebige erweiterte Shapes $\hat{S}_1, \hat{S}_2$ der Länge n können von einer binären Zeichenkette realisiert werden $\Longleftrightarrow$ der Graph G' ist bipartit.

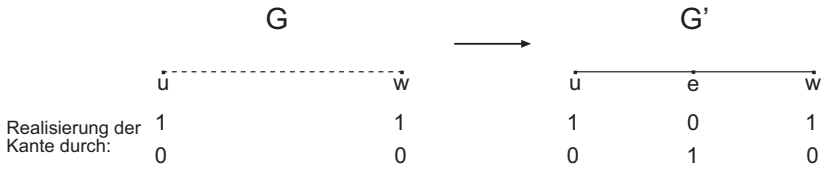

Abbildung 2.13: Vereinfachte Darstellung des Graphen G mit Pseudokanten und des resultierenden Graphen G' nach der Ersetzung der Pseudokanten durch zwei neue reguläre Kanten.

Beweis. $\Longrightarrow$: Es genügt nach Lemma 1.7 zu zeigen, dass in G' nach der Ersetzung keine Zyklen ungerader Länge existieren.
Wir wissen, dass nach Definition 2.1 und 2.11 die Kanten und Pseudokanten in den jeweiligen Shapes unabhängig sind, also der Grad der Knoten in $G(\hat{S}_1, \hat{S}_2)$ maximal 2 ist. Demzufolge gilt, dass jede zusammenhängende Komponente K in $G(\hat{S}_1, \hat{S}_2)$ entweder ein Pfad oder ein Zyklus ist. Es genügt also, die Behauptung für die jeweiligen Komponenten getrennt zu betrachten.
Nach der Ersetzung ändert sich der Grad der Knoten nicht. Daraus folgend werden nichtgeschlossene Pfade in $G(\hat{S}_1, \hat{S}_2)$ auch nach der Ersetzung keine Zyklen bilden.

Bleibt also der Fall zu betrachten, dass K ein Zyklus ist. Da wir nach Theorem 2.18 wissen, dass im Graphen $G(\hat{S}_1,\hat{S}_2)$ nur Zyklen mit einer geraden Anzahl von regulären Kanten existieren und für jede Pseudokante nach der Ersetzung zwei neue reguläre Kanten hinzukommen, folgt dass K auch nach der Ersetzung gerader Länge ist. Demzufolge existieren keine Zyklen ungerader Länge und somit ist G' bipartit.

$\Longleftarrow$: Auch in G' gilt nach Konstruktion, dass der Grad der Knoten maximal 2 ist und somit können auch in G' Pfade und Zyklen wieder getrennt betrachtet werden.
Sei K ein Pfad. Nach der Ersetzung beliebiger regulärer Kanten durch Pseudokanten bleibt K ein Pfad im Graphen $G(\hat{S}_1,\hat{S}_2)$ und stellt nach Theorem 2.18 somit für die Realisierung der Shapes $\hat{S}_1,\hat{S}_2$ durch eine binäre Zeichenkette kein Problem dar.
Sei K ein beliebiger Zyklus. Da der Graph G' bipartit ist, muss K gerader Länge sein. Da zwei reguläre Kanten in G' entfernt werden müssen, um durch eine Pseudokante ersetzt zu werden, also insbesondere die Ersetzung von Pseudokanten nur durch das Entfernen einer geraden Anzahl regulärer Kanten möglich ist, sind auch nach der Substitution noch eine gerade Anzahl von regulären Kanten in diesem Zyklus. Dies gilt für alle Zyklen in G'. Demzufolge existieren im Graphen $G(\hat{S}_1,\hat{S}_2)$ keine Zyklen ungerader Länge (und insbesondere keine Überlagerungen). Somit gilt nach Theorem 2.18, dass $\hat{S}_1,\hat{S}_2$ durch eine einzelne binäre Zeichenkette realisisiert werden können. □

Wir haben in diesem Abschnitt gezeigt, unter welchen Voraussetzungen es möglich ist zwei bzw. mehr als zwei gegebene Shapes durch eine einzelne binäre Zeichenkette zu realisieren. Wir haben gesehen, dass diese Form der Realisierung dann und nur dann möglich ist, wenn in den entsprechenden Shapegraphen keine Überlagerungen und keine Zyklen mit einer ungeraden Anzahl von Pseudokanten existieren.

Wir wollen nun im nächsten Abschnitt einen Realisierungsbegriff, der auch das RNA-Alphabet $\mathscr{A} = \{A,C,G,U\}$ berücksichtigt, formulieren.

2.5 Realisierungsbegriff erweiterter Shapes für $\mathscr{A} = \{A,C,G,U\}$

Wir haben im letzen Abschnitt gezeigt, unter welchen Voraussetzungen die Möglichkeit der Realisierung für eine binäre Zeichenkette exisitiert.

Da die Arbeit durch die Faltung der RNA und insbesondere durch die Fragestellung an das RNA-Alphabet motiviert wurde, wollen wir uns den Realisierungsbegriff für das Al-

phabet $\{A,C,G,U\}$ erweiterter Shapes genauer ansehen. Wie bisher werden wir fordern, dass für reguläre Kanten $\{v_i,v_j\} \in E(S)$ eines erweiterterten Shapes $\hat{S}$ gelten muss, dass die entsprechenden Buchstaben an den zugehörigen Positionen Elemente der Menge $\mathscr{B}$ möglicher Basenpaarungen sind, d.h. $s_i s_j \in \{AU,UA,CG,GC,GU,UG\}$. Allerdings gestaltet sich die Bedingung an die Realisierung der Pseudokanten schwieriger. Wir haben die Möglichkeit Pseudokanten in Shapes zu setzen, solange die Eigenschaften eines Shapes erhalten bleiben. Insbesondere sollte für Realisierung von Pseudokanten $\{v_i,v_j\} \in \hat{E}(S)$ gelten, dass $s_i s_j \notin \{AU,UA,CG,GC,GU,UG\}$. Somit hätten wir die Möglichkeit, folgende zusätzliche Forderungen an die Zeichenkette, die den Shape realisieren soll, zu stellen. Für alle Pseudokanten $\{v_i,v_j\} \in \hat{E}(S)$ und die zugehörige Sequenz, welche den erweiterten Shape $\hat{S}$ realisiert, muss gelten:

$$s_i s_j \in \{A,C,G,U\}^2 \backslash \{AU,UA,CG,GC,GU,UG\} =$$

$$\{AA,AC,AG,CC,CA,CU,GG,GA,UU,UC\}.$$

Mit Hilfe dieser zusätzlichen Annahme gelten die Voraussetzungen an den Shapegraphen, die wir für die Realisierung durch eine binäre Zeichenkette gezeigt haben, nicht mehr, wie Abbildung 2.14 zeigt.

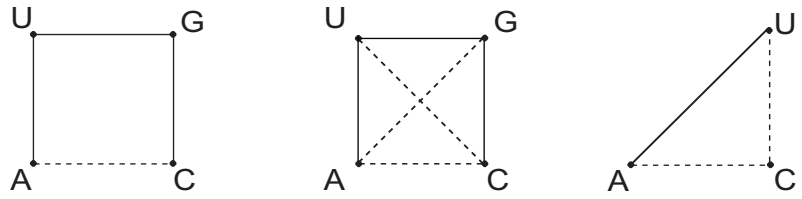

Abbildung 2.14: Darstellung von Shapegraphen und die mögliche Realisierung der Pseudokanten. Bei beliebiger Realisierung der Pseudokanten mit den Buchstaben A,C,G und U sind offensichtlich auch Zyklen mit einer ungeraden Anzahl von regulären Kanten realisierbar.

Wir werden deshalb zusätzliche Einschränkungen an die Regeln für die Realisierung stellen und betrachten im Folgenden nur eine Teilmenge aus der Menge
$\{AA,\ AC,\ AG,\ CC,\ CA,\ CU,\ GG,\ GA,\ UU,\ UC\}$. Sehen wir uns die folgende Teilmenge $\mathscr{B}_{PsK} = \{AA,UU,GG,CC\}$ und den entsprechenden Realisierungsbegriff an.

Definition 2.20. *Eine Zeichenkette $s = s_1 \dots s_n \in \{A,C,G,U\}^n$ realisiert einen erweiterten Shape $\hat{S}$ der Länge n $\Longleftrightarrow$*

1. $\forall \{v_i, v_j\} \in E(S)$ *gilt:* $s_i s_j \in \mathscr{B} := \{AU, UA, CG, GC, GU, UG\}$.
2. $\forall \{v_i, v_j\} \in \hat{E}(S)$ *gilt:* $s_i = s_j$, *d.h.* $s_i s_j \in \mathscr{B}_{PsK} := \{AA, UU, GG, CC\}$

Wir werden jetzt zeigen, dass die Voraussetzungen für die Realisierung erweiterter Shapes aus Theorem 2.18 auch für diesen speziellen Realisierungsbegriff gelten. Dazu sind einige kurze Vorbetrachtungen notwendig, mit denen wir einen alternativen Beweis für diesen Realisierungsbegiff liefern wollen. Betrachten wir zuerst die Definition eines Zeichenwechsels und das darauf folgende Lemma.

Definition 2.21. *Ein* Zeichenwechsel *findet in einer Zeichenkette $s_1 \dots s_n \in \{A,C,G,U\}^n$ statt, wenn gilt: $s_i \neq s_{i+1}, i \in \{1, \dots, n-1\}$.*

Lemma 2.22.

1. *Für einen Pfad $[v_1, e_1, \dots, v_n, e_n, v_{n+1}]$ ohne Pseudokanten und einer Zeichenkette $s = s_1 \dots s_{n+1}$ die den Pfad mittels Definition 2.20 realisiert gilt:*
   a) *In $s = s_1 \dots s_{n+1}$ finden n Zeichenwechsel statt.*
   b) *Wenn gilt: $s_1 = s_{n+1}$, dann ist die Anzahl n der Zeichenwechsel gerade.*
2. *Für einen Pfad $[v_1, e_1, \dots, v_n, e_n, v_{n+1}]$ der ausschließlich aus Pseudokanten besteht und einer Zeichenkette $s = s_1 \dots s_{n+1}$ die den Pfad mittels Definition 2.20 realisiert gilt: In $s = s_1 \dots s_{n+1}$ finden keine Zeichenwechsel statt.*
3. *Für einen Pfad $[v_1, e_1, \dots, v_n, e_n, v_{n+1}]$ der aus m_1 regulären Kanten und m_2 Pseudokanten besteht und einer Zeichenkette $s = s_1 \dots s_{n+1}$ die den Pfad mittels Definition 2.20 realisiert gilt:*
   a) *In $s = s_1 \dots s_{n+1}$ finden m_1 Zeichenwechsel statt.*
   b) *Wenn gilt: $s_1 = s_{n+1}$, dann ist die Anzahl m_1 der Zeichenwechsel gerade.*

Beweis. 1. a) Da für alle Kanten $\{v_i, v_{i+1}\}$ des Pfades gelten muss, dass $s_i s_{i+1} \in \mathscr{B} = \{AU, UA, CG, GC, GU, UG\}$, gilt $s_i \neq s_{i+1}$ $\forall i = 1, \dots, n$, woraus die Behauptung folgt.

b) Aus dem Beweis des Generalized Intersection Theorems 2.7 in [FHM+01] geht hervor, dass ein beliebiger Buchstabe $X \in \{A,C,G,U\}$ nur nach einer geraden Anzahl von Kanten wieder auftaucht. Da die Anzahl der Kanten aber nun nach 1.a) gerade der Anzahl der Zeichenwechsel entspricht, folgt die Behauptung. Dabei sei bemerkt, dass die Rückrichtung im Allgemeinen nicht gilt.

2. Nach Voraussetzung muss für alle Kanten $\{v_i, v_{i+1}\}$ des Pfades, der nur aus Pseudokanten besteht, gelten: $s_i s_{i+1} \in \mathscr{B}_{PsK} = \{AA, CC, GG, UU\}$. Beginnend mit einem beliebigen Buchstaben $X \in \{A, C, G, U\}$ für s_1, muss nun für jedes weitere s_i, auf Grund der Regel $\mathscr{B}_{PsK}$, immer derselbe Buchstabe X gewählt werden. Demzufolge gilt $s_i = s_{i+1}$ für alle $i = 1, \ldots, n$ und somit folgt die Behauptung.

3. a) Betrachten wir den gesamten Pfad $[v_1, e_1, \ldots, v_n, e_n, v_{n+1}]$. Nehmen wir an, es existieren k_1 Teilpfade $P_1, \ldots, P_{k_1}$, die nur aus regulären Kanten bestehen und k_2 Teilpfade $\hat{P}_1, \ldots, \hat{P}_{k_2}$, die nur aus Pseudokanten bestehen. Sei weiterhin l_i mit $i = 1, \ldots, k_1$ die Länge des entsprechenden Teilpfades P_i. Wir wissen, dass in den Teilzeichenketten der Teilpfade $\hat{P}_1, \ldots, \hat{P}_{k_2}$ keine Zeichenwechsel statt finden. In den Teilzeichenketten für ein $P_i \in \{P_1, \ldots, P_{k_1}\}$ finden aber jeweils l_i Zeichenwechsel statt. Demzufolge entspricht die Gesamtanzahl der Zeichenwechsel in der Zeichenkette, die den Pfad $[v_1, e_1, \ldots, v_n, e_n, v_{n+1}]$ realisiert, genau $\sum_{i=1}^{k_1} l_i = m_1$, woraus die Behauptung folgt.

   b) Sei $s_1 = s_{n+1}$. Wir wissen, dass Zeichenwechsel bei zwei aufeinanderfolgenden Zeichen s_i und s_{i+1} nur auftreten, falls die zugehörige Kante $\{v_i, v_{i+1}\}$ eine reguläre Kante ist. Sei jetzt s_1 ein beliebiger Buchstabe $X \in \{A, C, G, U\}$. Da dieser spezielle Buchstabe, wie wir dem Beweis des Theorems 2.7 [FHM+01] entnehmen können, nur nach einer geraden Anzahl von regulären Kanten wieder auftreten kann und die Anzahl der regulären Kanten in einem beliebigen Pfad gerade der Anzahl der Zeichenwechsel entspricht, folgt die Behauptung.

□

Theorem 2.23. *Beliebige erweiterte Shapes $\hat{S}_1, \ldots, \hat{S}_k$ der Länge n können von einer Zeichenkette aus dem Alphabet $\mathscr{A}$ entsprechend der Definition 2.20 realisiert werden $\Leftrightarrow$*

1. *$E(S_l) \cap \hat{E}(S_m) = \emptyset$ mit $l \neq m \in \{1, \ldots, k\}$ (d.h. es exitieren keine Überlagerungen von Pseudokanten des einen erweiterten Shapes mit 'regulären' Kanten eines anderen Shapes) und*

2. *der Graph $G(\hat{S}_1, \ldots, \hat{S}_k)$ enthält keine Zyklen mit einer ungeraden Anzahl von regulären Kanten.*

Beweis. $\Longrightarrow$: (*Beweis durch Widerspruch*)
Seien die erweiterterten Shapes $\hat{S}_1, \ldots, \hat{S}_k$ der Länge n durch die Zeichenkette $s = s_1 \ldots s_n$ realisierbar.
Annahme $\exists \{v_i, v_j\} \in E(S_l) \cap \hat{E}(S_m)$ mit $l \neq m$, dann muss gelten: $s_i \neq s_j$ für $\{v_i, v_j\} \in E(S_l)$,

als auch $s_i = s_j$ für $\{v_i, v_j\} \in \hat{E}(S_m) \Rightarrow$ WIDERSPRUCH.
Annahme es existiert ein Zyklus $[v_{i_1}, e_{i_1}, \ldots, v_{i_m}, e_{i_m}, v_{i_1}]$ (ohne Überlagerungen) der Länge m mit einer ungeraden Anzahl m_1 von regulären Kanten. Bezeichne $s_{i_1} \ldots s_{i_m}$ die Teilsequenz aus s, die diesen Zyklus realisiert. Falls der Zyklus keine Pseudokanten enthält und somit $m_1 = m$ gilt, ist nach Lemma 2.22 in der Teilzeichenkette $s_{i_1} \ldots s_{i_m}$ die Anzahl der Zeichenwechsel m. Da aber m ungerade ist gilt weiterhin, dass $s_{i_1} \neq s_{i_1}$. $\Rightarrow$WIDERSPRUCH.
Nehmen wir nun an, dass der Zyklus Pseudokanten enthält und betrachten folgende Zerlegung des Zyklus in zwei Pfade. Der eine Pfad $P_1 = [v_{i_l}, e_{i_l}, \ldots, e_{i_{l'}}, v_{i_{l'+1}}]$ sei ein Pfad, der nur aus Pseudokanten besteht, so dass die zusätzliche Eigenschaft, dass $e_{i_{l-1}}$ und $e_{i_{l'+1}}$ reguläre Kanten sind, erfüllt ist. Es sei bemerkt, dass der Pfad auch aus nur einer Pseudokante bestehen darf, d.h. $l = l'$ ist erlaubt. Der andere Pfad P_2 bestehe aus allen Kanten des Zyklus ohne die Kanten des Pfades P_1. Sei nun $s_{i_{l'+1}} \ldots s_{i_l}$ die Teilsequenz aus $s_{i_1} \ldots s_{i_m}$, die den Pfad P_2 realisiert. Nach Konstruktion enthält dieser Pfad alle im Zyklus vorkommenden regulären Kanten. Da die Anzahl der regulären Kanten ungerade ist und somit die Anzahl der Zeichenwechsel ungerade ist, folgt nach Lemma 2.22, dass $s_{i_{l'+1}} \neq s_{i_l}$. Da aber der Pfad P_1 nur aus Pseudokanten besteht und somit gelten muss $s_{i_l} = s_{i_{l'+1}} \Rightarrow$WIDERSPRUCH.

$\Longleftarrow$: Da das Alphabet $\mathscr{A} = \{A, C, G, U\}$ mit den Regeln $\mathscr{B}, \mathscr{B}_{PsK}$ aus Definition 2.20 eine Erweiterung des Realisierungsbegriffes durch eine binäre Zeichenkette aus Definition 2.13 darstellt, folgt aus Theorem 2.18 die Behauptung. □

Da die Voraussetzungen dieses Theorems genau denen des Theorems 2.18 entsprechen, können wir folgende zusätzliche Schlussfolgerung ziehen.

Korollar 2.24. *Beliebige erweiterte Shapes $\hat{S}_1, \ldots, \hat{S}_k$ der Länge n können von einer Zeichenkette aus dem Alphabet $\mathscr{A}$ entsprechend der Definition 2.20 realisiert werden $\Leftrightarrow$ die Shapes $\hat{S}_1, \ldots, \hat{S}_k$ können von einer binären Zeichenkette entsprechend der Definition 2.13 realisiert werden.*

Natürlich exisitieren noch offene Fragen an weitere Realisierungsbegriffe mittels anderer Paarungsregeln für Pseudokanten, die wir aber im Zuge dieser Arbeit nicht näher untersuchen werden.

Für den betrachteten Realisierungsbegriff erweiterter Shapes können wir zum Schluss noch festhalten, wenn eine Sequenz existiert, die gegebene erweitertete Shapes realisiert, so realisiert diese die entsprechenden Shapes auch ohne Pseudokanten. Dies sei noch einmal kurz erklärt. Wir haben gesehen, dass eine einzelne Zeichenkette, die gegebene erweiterte Shapes realisiert, dann und nur dann existiert, wenn in dem entsprechenden Shapegraphen

erweiterter Shapes keine Überlagerungen und keine Zyklen mit einer ungeraden Anzahl regulärer Kanten existieren. Es ist klar, dass es keine Überlagerungen in dem Shapegraphen der entsprechenden "nicht erweiterten" Shapes gibt, da in diesem keine Pseudokanten existieren. Des Weiteren existieren in den Shapegraphen erweiterter Shapes keine Zyklen mit einer ungeraden Anzahl von regulären Kanten. Dieses wiederum bedeutet, dass in dem Graphen nach der Entfernung aller Pseudokanten keine Zyklen ungerader Länge existieren und somit der entsprechende Graph bipartit ist.

Allerdings besitzen diese Sequenzen jetzt die zusätzliche Eigenschaft, das auch nicht gepaarte Positionen, insbesondere Positionen die unter den Voraussetzungen der Basenpaarungsregeln $\mathscr{B}$ keine Paarung eingehen sollen, berücksichtigt wurden.

2.6 Zusammenfassung

Wir haben in diesem Kapitel gesehen, unter welchen Voraussetzungen es möglich ist, eine einzelne Sequenz zu finden, so dass gegebene Shapes durch diese realisiert werden können. Dabei wurde die Möglichkeit der Realisierung ganz eng mit den Eigenschaften des entsprechenden Shapegraphen verknüpft. Allerdings wurden in dem bisherigen Realisierungsbegriff nur gepaarte Positionen berücksichtigt. Wir erweiterten die Shapes daraufhin mittels Pseudokanten, um den Begriff der Realisierung einzuschränken und somit Sequenzen zu finden, in denen auch Positionen berücksichtigt werden, die keine Paarungen eingehen dürfen. Dabei definierten wir zunächst den Realisierungsbegriff für eine binäre Zeichenkette. Wir haben gesehen, unter welchen Voraussetzungen gegebene erweiterte Shapes durch eine einzelne binäre Zeichenkette realisiert werden können. Im letzten Abschnitt definierten wir einen entsprechenden Realisierungsbegriff für Zeichenketten $s \in \{A,C,G,U\}^n$ und fanden heraus, dass dieser Realisierungsbegriff äquivalent zu dem Begriff der Realisierung durch eine binäre Zeichenkette ist.

3 Komplexitätsbetrachtungen

In diesem Kapitel werden wir Komplexitätsbetrachtungen zu verschieden Problemen vornehmen. Insbesondere wird uns die NP-Vollständigkeit dieser Probleme interessieren. Zunächst werden wir die Probleme formulieren und erklären, wodurch die Betrachtung dieser Sachverhalte motiviert ist. In den darauf folgenden Abschnitten zeigen wir die NP-Vollständigkeit der genannten Probleme. Dazu ist es notwendig, den Begriff von *homöomorphen Erweiterungen* zu definieren, sowie einen Algorithmus vorzustellen, der überprüpft, ob gegebene (erweiterte) Shapes durch eine einzelne Zeichenkette realisiert werden können. Wenn wir im Folgenden von der Realisierung "nicht erweiterter" Shapes sprechen, so meinen wir ausschließlich den Begriff aus Definition 2.4. Wenn wir von der Realisierung erweiterter Shapes sprechen, so meinen wir die Definition 2.13 bzw. 2.20. Beginnen wir nun im nächsten Abschnitt mit der Motivation der zu formulierenden Probleme.

3.1 Problemformulierungen

In dem vorangegangenem Kapitel haben wir gesehen, dass beliebige Shapes bzw. erweiterte Shapes gleicher Länge nicht durch eine einzelne Zeichenkette realisiert werden können, falls bestimmte Voraussetzungen in den zugehörigen Shapegraphen nicht erfüllt sind. Demzufolge stellt sich die Frage, welche Wege es gibt, den Shapegraphen so umzugestalten, so dass die Shapes bzw. erweiterten Shapes durch eine einzelne Zeichenkette realisiert werden können. Eine Möglichkeit besteht darin, bestimmte Knoten zu entfernen. Ein weiterer Weg ergibt sich aus der Möglichkeit, bestimmte Kanten zu entfernen, bis die Eigenschaften der Realisierbarkeit für diesen Graphen erfüllt sind. Trivialerweise könnte man alle Knoten bzw. alle Kanten entfernen, wodurch aber auch alle Eigenschaften des Graphen verloren gingen. Ziel ist es deshalb, die minimale Anzahl von Kanten bzw. Knoten zu entfernen, so dass die Eigenschaften im Shapegraphen für die Realisierung der Shapes erfüllt sind. Wir wollen nun folgende Probleme betrachten:

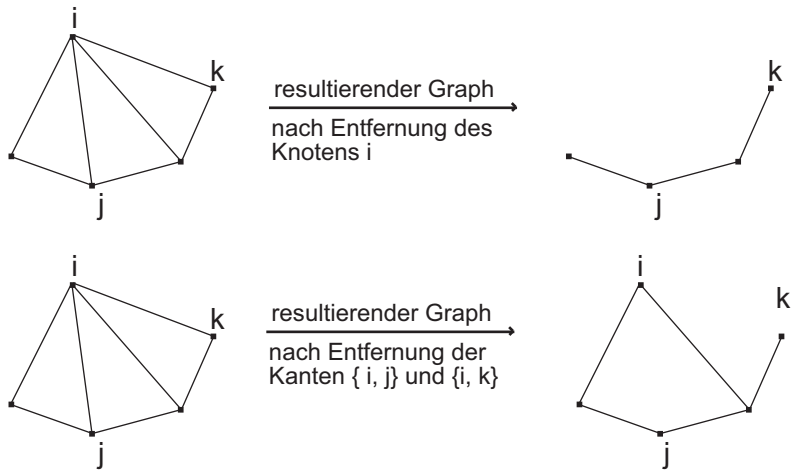

Abbildung 3.1: Links ist ein Shapegraph zu sehen, dessen Shapes nicht durch eine einzelne Zeichenkette realisiert werden können. Nach dem Entfernen von Knoten bzw. Kanten sind die Eigenschaften der Realisierbarkeit erfüllt.

Problem 3.1. (MinKN_S) *Berechne die minimale Anzahl von* KNOTEN*, die aus dem Graphen $G(S_1,\ldots,S_k)$ entfernt werden müssen, so dass die Shapes $S_1,\ldots,S_k$ durch eine Zeichenkette realisiert werden können.*

Problem 3.2. (MinKA_S) *Berechne die minimale Anzahl von* KANTEN*, die aus dem Graphen $G(S_1,\ldots,S_k)$ entfernt werden müssen, so dass die Shapes $S_1,\ldots,S_k$ durch eine Zeichenkette realisiert werden können.*

Problem 3.3. (MinKN_eS) *Berechne die minimale Anzahl von* KNOTEN*, die aus dem Graphen $G(\hat{S}_1,\ldots,\hat{S}_k)$ entfernt werden müssen, so dass die erweiterten Shapes $\hat{S}_1,\ldots,\hat{S}_k$ durch eine Zeichenkette realisiert werden können.*

Problem 3.4. (MinKA_eS) *Berechne die minimale Anzahl von* REGULÄREN UND PSEUDOKANTEN*, die aus dem Graphen $G(\hat{S}_1,\ldots,\hat{S}_k)$ entfernt werden müssen, so dass die erweiterten Shapes $\hat{S}_1,\ldots,\hat{S}_k$ durch eine Zeichenkette realisiert werden können.*

Bevor wir Algorithmen suchen, welche die Probleme in polynomieller Zeit lösen könnten, sollten wir uns mit der NP[1]-Vollständigkeit der Probleme beschäftigen. Es ist bekannt, dass zu solchen Problemen kein polynomieller Lösungsalgorithmus existiert, sofern die Komplexitätsklasse polynomieller Probleme P nicht identisch mit der Komplexitätsklasse der NP Probleme ist, d.h. solange nicht gilt: $P = NP$ [GJ79]. Die folgende Definition ist aus dem Lehrbuch von Garey und Johnson [GJ79] entnommen.

[1] aus dem engl.: Non-deterministic Polynomial time

Definition 3.5. *Ein Problem P ist NP-vollständig, falls gilt:*

1. $P \in NP$.

2. *P ist NP-schwer.*

Um den ersten Punkt zu beweisen, muss gezeigt werden, dass ein nichtdeterministischer Algorithmus existiert, der eine Lösung rät und in polynomieller Zeit verifizieren kann, ob eine gültige Lösung gefunden wurde.

Der zweite Punkt beschreibt die Tatsache, dass sich alle Probleme $P' \in NP$ in polynomieller Zeit auf das Problem P reduzieren lassen. Um die NP-Schwere eines Problems P zu zeigen, nimmt man in der Regel ein als NP-vollständig bekanntes Problem P' und reduziert dieses in polynomieller Zeit auf das Problem P. Aus der Transitivität von Polynomialzeitreduktionen folgern wir, dass alle Probleme aus der Klasse NP auch auf das betrachtete reduzierbar sind [GJ79].

Von Peter Clote et al. wurde bewiesen, dass das Problem 3.1 für $k \geq 4$ NP-vollständig ist [CLK+05].

Wir werden in den nächsten Abschnitten zeigen, dass zudem auch die Probleme *MinKA_S*, *MinKN_eS* und *MinKA_eS* NP-vollständig sind.

Falls wir nicht fordern, dass Pseudokanten gesetzt werden müssen, so ist ersichtlich, dass das Problem *MinKN_S* eine Teilinstanz des Problems *MinKN_eS* ist, sowie das Problem *MinKA_S* eine Teilinstanz des Problems *MinKA_eS* ist. Wir könnten somit den Fall betrachten, dass keine Pseudokanten existieren. Für diese Fälle müssen wir für die Probleme *MinKA_eS* und *MinKN_eS* nichts weiter zeigen. Fordern wir demzufolge, dass Pseudokanten gesetzt werden müssen.

Bevor wir uns nun den Beweisen der NP-Vollständigkeit der genannten Probleme zuwenden, werden wir ein für die Beweise nützliches Werkzeug, sogenannte *homöomorphe Erweiterungen*, vorstellen.

3.2 Vorbetrachtungen

3.2.1 Homöomorphe Erweiterungen

Die Beweise der NP-Schwere der Probleme 3.2 (MinKA_S), 3.3 (MinKN_eS) und 3.4 (MinKA_eS) beruhen auf der Reduktion bzw. der Äquivalenz anderer bekannter NP-vollständiger Probleme. Die Schwierigkeit, welche zusätzlich entsteht, ist die folgende:

Falls wir die Reduktion bzw. Äquivalenz von schon bekannten NP-vollständigen Problemen zeigen, so haben wir in diesen Fällen, wie wir später sehen werden, die NP-Schwere nur für einen beliebigen Graphen $G' = (V,E)$ bzw. $G'' = (V,E,\hat{E})$ mit maximalen Knotengrad k bewiesen. Es stellt sich somit natürlich die Frage, wie man zu einem gegebenen Graphen $G' = (V,E)$ bzw. $G' = (V,E,\hat{E})$ mit maximalen Knotengrad k die Shapes $S_1,\ldots,S_k$ bzw. $\hat{S}_1,\ldots,\hat{S}_k$ konstruiert, so dass $G(S_1,\ldots,S_k)$ bzw. $G(\hat{S}_1,\ldots,\hat{S}_k)$ isomorph zu einem gegebenen Graphen $G' = (V,E)$ bzw. $G'' = (V,E,\hat{E})$ ist. Es ist offensichtlich, dass man für jede Kante einen Shape konstruieren kann. Man benötigt in diesem Fall $|E|$ bzw. $|E \cup \hat{E}|$ Shapes. Ein anderer Weg über sogenannte *homöomorphe Erweiterungen*, der die Anzahl der Shapes deutlich reduziert, wird in [CLK+05] vorgestellt.

Definition 3.6. [CLK+05] *Seien $G = (V,E)$, $G' = (V',E')$ zwei Graphen.*
Wir bezeichnen G' homöomorphe Erweiterung *von G, falls man G' durch das Ersetzen beliebiger Kanten $\{v',v''\} \in E(G)$ durch einen Pfad $P(\{v',v''\}) =$*
$\{v',u_1\}\{u_1,u_2\}\ldots\{u_{k-1},u_k\}\{u_k,v''\}$ mit $deg(u_i) = 2$ und $u_i \notin V(G)$ für alle $i = 1,\ldots,k$ aus G erhält.
G' bezeichnet man als ein ungerade homöomorphe Erweiterung *von G, falls alle Pfade $P(\{v',v''\})$ ungerader Länge sind. Für alle u_i eines solchen Pfades $P(\{v',v''\})$ bezeichnet man die Knoten v' und v'' als* Endknoten *von u_i.*

Dabei darf die Anzahl der ersetzten Kanten natürlich nicht exponentiell sein, da eine solche Ersetzung nicht in polynomieller Laufzeit möglich wäre. Folgendes Theorem wurde in [CLK+05] bewiesen.

Theorem 3.7. [CLK+05] *Sei $k \geq 3$ und G' ein Graph mit maximalen Knotengrad k. Es existieren k Shapes $S_1,\ldots,S_k$, so dass $G(S_1,\ldots,S_k)$ isomorph zu einer ungeraden homöomorphen Erweiterung von G' ist. Die Shapes können in polynomieller Zeit in Abhängigkeit der Anzahl von Kanten konstruiert werden.*

Der technisch sehr aufwändige Beweis zu diesem Theorem beruht auf der Konstruktion von k Shapes zu einem gegebenen Graphen $G' = (V,E)$. Zunächst werden, unter der Vorgabe des Graphen G' mit einem maximalen Knotengrad k, Shapes $S'_1,\ldots,S'_k$ konstruiert. Es wird danach bewiesen, dass der entsprechende Shapegraph $G(S'_1,\ldots,S'_k)$ isomorph zu einer homöomorphen Erweiterung von G' ist. Daraufhin werden die Shapes und somit der Shapegraph $G(S'_1,\ldots,S'_k)$ zu einem Graphen $G(S_1,\ldots,S_k)$ erweitert. Zuletzt wird gezeigt, dass dieser Graph $G(S_1,\ldots,S_k)$ isomorph zu einer ungeraden homöomorphen Erweiterung von G' ist.

Wenn wir zeigen, dass die minimalen zu entfernenden Kanten- bzw. Knotenmengen der entsprechenden Probleme in einem Graphen G und einer ungeraden homöomorphen Erweiterung $\tilde{G}$ von G die gleiche Kardinalität besitzen, so dass der Graph $\tilde{G}$ bzw. G die Voraussetzungen für die Realisierbarkeit der entsprechenden Shapes erfüllt, können wir mittels dieses Theorems die Anzahl der Shapes nicht nur deutlich reduzieren, sondern auch Rückschlüsse auf die Anzahl der Shapes bei gegebenen Knotengrad ziehen.

Wir werden nun die Definition homöomorpher Erweiterungen für Graphen erweiterter Shapes, d.h. Graphen mit Pseudokanten, modifizieren.

Definition 3.8. *Sei $G(\hat{S}_1,\ldots,\hat{S}_k)=(V(G),E(G),\gamma)$ ein Graph erweiterter Shapes ohne Überlagerungen, d.h. $\nexists\, e\in E(G)$ mit der Eigenschaft $\gamma(e)=2$.*
Sei weiterhin $G'=(V(G'),E(G'),\gamma')$ ein Graph, den man aus diesem Graphen G erhält, indem man beliebige Kanten $\{v',v''\}\in E(G)$ durch einen Pfad $P(\{v',v''\})=$ $\{v',u_1\}\{u_1,u_2\}\ldots\{u_{k-1},u_k\}\{u_k,v''\}$ mit $deg(u_i)=2$ und $u_i\notin V(G)\ \forall\ i=1,\ldots,k$ ersetzt. Man bezeichnet G' als homöomorphe Erweiterung *eines Graphen G erweiterter Shapes falls zusätzlich gilt, dass*

$$\gamma'(\{u,w\})=\begin{cases}\gamma(\{v',v''\}) & \text{, falls } \{u,w\}\in P(\{v',v''\}) \text{ und } \{u,w\}\notin E(G)\cap E(G')\\ \gamma(\{u,w\}) & \text{, sonst}\end{cases}$$

Das heißt, Pseudokanten werden im Fall der Ersetzung durch Pfade aus Pseudokanten bzw. reguläre Kanten durch Pfade regulärer Kanten erweitert. Man bezeichnet G' als ungerade homöomorphe Erweiterung*, wenn alle so ersetzten Pfade ungerader Länge sind. Für alle u_i eines solchen Pfades $P(\{v',v''\})$ bezeichnet man die Knoten v' und v'' als* Endknoten *von u_i.*

Theorem 3.7 gilt auch für den Graphen erweiterter Shapes, da die grundlegende Eigenschaft eines Shapes nach Definition 2.1 auch in erweiterten Shapes nach Definition 2.11 erhalten bleiben, d.h. die Kanten sind unabhängig und es existieren keine Pseudoknoten. Bei der Konstruktion der Shapes müssen somit vorher alle Pseudokanten durch reguläre Kanten ersetzt werden. Danach ersetzt man die eben modifizierten regulären Kanten wieder durch die entsprechenden Pseudokanten.

3.2.2 Algorithmus zum Test auf Realisierbarkeit

Um zu zeigen, dass ein Problem P in der Klasse NP liegt, ist es notwendig, einen nichtdeterministischen Algorithmus anzugeben, der eine Lösung rät und in polynomieller Zeit verifizieren kann, ob eine gültige Lösung gefunden wurde. Wir werden deshalb zunächst

einen Algorithmus polynomieller Laufzeit vorstellen, der Shapegraphen von Shapes bzw. erweiterten Shapes daraufhin überprüft, ob in diesem Graphen kritische Zyklen bzw. Überlagerungen existieren. Bevor uns diesem Algorithmus zuwenden, werden wir mit Vorzeichen markierte Graphen in Graphen mit Pseudokanten und reguläre Kanten überführen.

Lemma 3.9. *Sei $H = (V,E,\varphi)$ ein mit Vorzeichen markierter Graph. Sei weiterhin $G = (V,E,\hat{E}) = (V,E,\gamma)$ der aus H resultierende Graph, indem man $V(G) = V(H)$ setzt und*

$$\gamma(e) = \begin{cases} 0 & \text{, falls } \varphi(e) = -1 \\ 1 & \text{, falls } \varphi(e) = +1 \end{cases},$$

d.h. Kanten aus H mit negativen Vorzeichen werden als reguläre Kanten und Kanten aus H mit positiven Vorzeichen werden als Pseudokanten betrachtet.
Es gilt:

H ist balanciert. $\Longleftrightarrow$ G enthält keine Zyklen mit einer ungeraden Anzahl regulärer Kanten, sowie keine Überlagerungen.

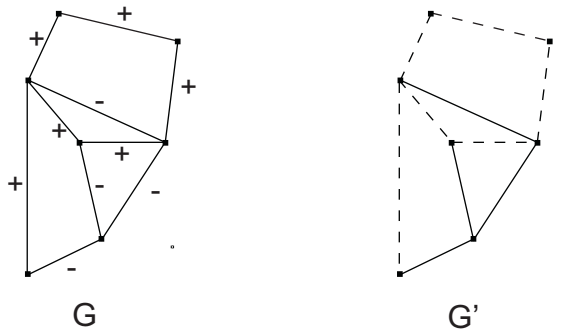

Abbildung 3.2: Konstruktion des Graphen *G* aus *H*

Beweis. $\Longrightarrow$: Sei *H* balanciert. Dies ist der Fall, gdw. in *H* keine Zyklen mit einer ungeraden Anzahl von negativen Kanten existieren und nach Konstruktion gilt somit für *G*, dass keine Zyklen mit einer ungeraden Anzahl von regulären Kanten existieren. Insbesondere befinden sich auch keine Überlagerungen in *G*, da es in *H* keine Mehrfachkanten gibt.
$\Longleftarrow$: Sei $G = (V,E,\hat{E})$ ein Graph, indem keine Zyklen mit ungerader Anzahl regulärer Kanten und keine Überlagerungen existieren, so folgt nach Konstruktion, dass auch in *H* keine Zyklen mit einer ungeraden Anzahl von negativen Kanten existieren, also ist *H* balanciert. □

Ein dynamischer Algorithmus mit einer Zeitkomplexität $O(|E|+|V|)$ der überprüft ob ein S-Graph $H = (V,E,\varphi)$ balanciert ist, wurde in [Lou03] vorgestellt und ist bis auf wenige Veränderungen in Algorithmus 1 dargestellt. Diesen können wir nach einigen wenigen Modifikationen benutzen, um zu testen ob ein Graph $G = (V,E,\hat{E})$ die notwendigen Eigenschaften der Realisierbarkeit erfüllt. Zunächst werden alle Pseudokanten mit '+1' und alle regulären Kanten mit einer '-1' gekennzeichnet. Dass diese Transformation in polynomieller Zeit in Abhängigkeit von der Kantenanzahl $|E \cup \hat{E}|$ geschieht, ist offensichtlich. Ein solch veränderter Shapegraph wird im Algorithmus *modifizierter* Shapegraph genannt. Eine weitere notwendige Veränderung die wir vornehmen müssen, so dass das Programm auch abbricht, falls Überlagerungen existieren, ist in der achten und neunten, sowie in der neunzehnten und zwanzigsten Zeile des Algorithmus 1 festgehalten. Die Grundidee des Algorithmus besteht auf dem Aufstellen eines Breitensuchbaumes und der Erkenntnis, dass alle Pfade zwischen zwei Knoten dasselbe Vorzeichen besitzen, wie wir mit Theorem 1.13 bewiesen haben. Es sei bemerkt, dass in dem Pseudocode $\sigma[v]$ das Vorzeichen eines Knoten *v* kennzeichnet. Die Funktion $\sigma[v,w]$ bezeichnet das Vorzeichen der Kante $\{v,w\}$. Weitere Informationen und Erklärungen dazu findet man in [Lou03] und [HK80].

Algorithm 1 Test_Realisierbarkeit

1: **INPUT:** modifizierter Shapegraph $(V, E, \hat{E})$ und Startknoten s;
2: besucht[s]:=true; balance:=true; σ[s]:=+1 oder -1;
3: **for all** x $\in$ V(G)$\backslash\{s\}$ **do**
4: besucht[x]:=false;
5: **end for**
6: first:=last:=1;
7: **for all** x $\in$ Adj[s] **do**
8: **if** ((x,s) ist Überlagerung) und (balance = true) **then**
9: balance:=false;
10: **else if** balance = true **then**
11: σ[x]:=σ[s]$\cdot\sigma$[(s,x)]; last:=last+1;
12: Q[last]:=x; besucht[x]:=tru;e
13: **end if**
14: **end for**
15: **while** (first $\leq$ last) und (balance = true) **do**
16: x:=Q[first]; first:=first+1;
17: besucht[x]:=true;
18: **for all** y$\in$ Adj[x] **do**
19: **if** ((x,y) ist Überlagerung) und (balance = true) **then**
20: balance:=false;
21: **else**
22: **if** besucht[y] = false **then**
23: besucht[y]:=true; last=last+1; Q[last]:=y;
24: σ[y]:=σ[x]$\cdot\sigma$[(x,y)];
25: **else**
26: **if** (σ[x]$\cdot\sigma$[(x,y)]$\neq$ σ[y]) **then**
27: balance:=false;
28: **end if**
29: **end if**
30: **end if**
31: **end for**
32: **end while**

3.3 MinKA_S ist NP-vollständig.

Wir werden uns nun dem Beweis der NP-Vollständigkeit des Problems MinKA_S zuwenden. Dieses wurde wie folgt formuliert:

Problem 3.2. (MinKA_S) *Berechne die minimale Anzahl von* KANTEN*, die aus dem Graphen $G(S_1,\ldots,S_k)$ entfernt werden müssen, so dass die Shapes $S_1,\ldots,S_k$ durch eine Zeichenkette realisiert werden können.*

Wir wissen nach Theorem 2.7, dass beliebige Shapes $S_1,\ldots,S_k$ gleicher Länge realisiert werden können gdw. der Graph $G(S_1,\ldots,S_k)$ keine ungeraden Zyklen enthält und nach Lemma 1.7 ist dies der Fall gdw. der Graph $G(S_1,\ldots,S_k)$ bipartit ist. Dieses Problem wurde von Yannakakis in [Yan81] als NP-vollständig bewiesen.

Theorem 3.10. [Yan81] *Sei $k \geq 3$ eine natürliche Zahl und $G = (V,E)$ ein Graph mit maximalen Knotengrad k. Folgendes Problem ist NP-vollständig:* Entferne die minimale Anzahl von Kanten in *G,* so dass *G* bipartit ist.

Im Folgenden werden wir zeigen, dass die minimale Anzahl zu entfernender Kanten aus *G*, damit *G* bipartit ist, der minimale Anzahl zu entfernender Kanten einer ungeraden homöomorphe Erweiterung G' von *G*, damit G' bipartit ist, entspricht. Nach Theorem 3.7 wissen wir somit, dass auf Grund der Isomorphie der Graphen G' und $G(S_1,\ldots,S_k)$ auch in dem entsprechenden Shapegraph die minimale Anzahl der Kanten, die entfernt müssen, um die Eigenschaft der Bipartition zu bekommen, gleich sind. Damit wäre die NP-Schwere des Problems MinKA_S für $k \geq 3$ bewiesen.

Lemma 3.11. *Sei $G' = (V',E')$ eine ungerade homöomorphe Erweiterung von $G = (V,E)$ und seien $E^*(G) \subseteq E(G)$ und $E^*(G') \subseteq E(G')$ die minimalen Kantenmengen, die aus G bzw. G' entfernt werden müssen, so dass G und G' bipartit sind, so gilt:*

$$|E^*(G)| = |E^*(G')|$$

Beweis. Halten wir zunächst fest, dass ungerade bzw. gerade Zyklen in *G* auch ungerade bzw. gerade Zyklen in G' sind. Diese Eigenschaft ist leicht zu erkennen, wenn man sich überlegt, dass in einem Zyklus ungerader Länge zuerst eine Kante entfernt wird und somit der Restpfad gerader Länge ist. Danach ersetzt man die entfernte Kante durch einen ungeraden Pfad. Schlussfolgernd ist der Zyklus nach der Ersetzung immer noch ungerade. Entsprechend gilt dies auch für Zyklen gerader Länge.

Des Weiteren existiert in jeder ungeraden homöomorphen Erweiterung G' von G eine eineindeutige Zuordnung von Kanten $\{v',v''\} \in E(G)$ durch einen Pfad $P(\{v',v''\})$, falls dieser vorhanden ist, ansonsten durch die Kante $\{v',v''\} \in E(G')$ selbst. Man kann also für jede Kante $\{v',v''\} \in E^*(G)$ in G' entweder die Kante $\{v',u_1\} \in E(G')$ entfernen, falls $P(\{v',v''\})$ existiert oder die Kante $\{v',v''\}$ selbst, damit auch G' bipartit ist.
Umgekehrt kann man für alle Kanten aus $\{u,w\} \in (E(G')\backslash E(G)) \cap E^*(G')$, d.h. Kanten die Teil eines Pfades $P(\{v',v''\})$ sind, die Kante $\{v',v''\}$ in G entfernen und für alle $\{u,w\} \in E(G) \cap E^*(G')$ die Kante {u,w} selbst. □

Somit haben wir die NP-Schwere unseres Problems bewiesen. Bleibt zu zeigen, dass das Problem *MinKA_S* in der Klasse NP liegt. Einen nichtdeterministischen Algorithmus für unser Problem erhalten wir, indem eine Lösung des Problems geraten wird, d.h. eine Teilmenge der Kantenmenge und nach Entfernung dieser Teilmenge mittels des Algorithmus *Test_Realisierbarkeit* verifiziert wird, ob die so erhaltende Lösung gültig ist. Aus dem im ersten Kapitel gezeigten Lemma 1.15 folgt die Möglichkeit, diesen Algorithmus auch zu benutzen, um zu überprüfen ob ein Graph bipartit ist. Somit haben wir das folgende Theorem bewiesen.

Theorem 3.12. *MinKA_S ist NP-vollständig für $k \geq 3$.*

3.4 MinKN_eS ist NP-vollständig

Wir werden nun zeigen, dass das Problem MinKN_eS NP-vollständig ist. Rekapitulieren wir noch einmal.

Problem 3.3. (MinKN_eS) *Berechne die minimale Anzahl von* KNOTEN*, die aus dem Graphen $G(\hat{S}_1,\ldots,\hat{S}_k)$ entfernt werden müssen, so dass die erweiterten Shapes $\hat{S}_1,\ldots,\hat{S}_k$ durch eine Zeichenkette realisiert werden können.*

Als *Vertex Set*[2] *für kritische Komponenten (VSkK)* bezeichen wir des Weiteren die Menge von Knoten, die aus dem Graphen $G(\hat{S}_1,\ldots,\hat{S}_k)$ erweiterter Shapes entfernt werden muss, so dass die entsprechenden Shapes $\hat{S}_1,\ldots,\hat{S}_k$ durch eine einzelne Zeichenkette realisiert werden können. Es folgt nun eine formale Definition.

Definition 3.13. *Sei $G = (V,E,\hat{E})$ ein Graph erweiterter Shapes und V' eine Teilmenge von V. Wir bezeichnen V' als* Vertex Set für kritische Komponenten (VSkK)*, falls V' mindestens*

[2] aus dem engl.: Knotenmenge

einen Knoten aus jeder Überlagerung und jedem Zyklus mit einer ungeraden Anzahl von regulären Kanten enthält.

Betrachten wir nun die Definition eines *Vertex Cocer Sets*[3] und das daraus resultierende, als NP-vollständig bekannte Problem.

Definition 3.14. *Sei $G = (V,E)$ ein Graph und V' eine Teilmenge von V. Man bezeichnet V' als* Vertex Cover Set (VCS)*, falls für alle $e \in E$ gilt: $\exists\, v \in V'$, so dass $v \in e$.*

Natürlich ist es einfach, dieser Definition zu genügen, indem man die gesamte Knotenmenge V des Graphen als Vertex Cover Set begreift. Somit ergibt sich folgendes Problem aus dieser Betrachtung.

Problem 3.15 (Vertex Cover Set Problem VCSP). *Sei $G = (V,E)$ ein Graph mit maximalem Knotengrad k. Finde ein Vertex Cover Set von G mit minimaler Kardinalität.*

Von Garey und Johnson wissen wir, dass dieses Problem für einen Knotengrad $k \geq 3$ NP-vollständig ist [GJ79]. Bevor wir uns dem Beweis der NP-Vollständigkeit unseres Problems MinKN_eS zuwenden, betrachten wir das folgende Lemma.

Lemma 3.16. *Sei $G' = (V', E', \gamma')$ eine ungerade homöomorphe Erweiterung eines Graphen $G = (V,E,\gamma)$ erweiterter Shapes, so haben die minimalen Vertex Sets für kritische Komponenten in G und G' die gleiche Kardinalität.*

Beweis. Durch die Definition von homöomorphen Erweiterungen erhalten wir eine eineindeutige Zuordnung von Zyklen aus G und G', so dass jeder Zyklus C in G einem Zyklus in G' entspricht, der alle Knoten von C beinhaltet. Demnach ist jedes Vertex Set für kritische Komponenten in G auch ein Vertex Set für kritische Komponenten in G'.
Sei nun $V^* \subseteq V'$ das minimale Vertex Set für kritische Komponenten in G'. Da jeder Knoten $v \in (V' \backslash V) \cap V^*$ durch seinen Endknoten ersetzt werden kann, ohne dass V^* die Eigenschaften eines minimalen Vertex Set für kritische Komponenten in G' verliert, kann man annehmen, dass V^* nur aus Knoten aus V besteht. Also ist V^* ein Vertex Set für kritische Komponenten aus G. □

Wenden wir uns nun dem Beweis der NP-Vollständigkeit des Problems MinKN_eS zu.

[3] aus dem engl.: Knotenüberdeckungsmenge

Theorem 3.17. *MinKN_eS ist NP-vollständig für $k \geq 4$.*

Beweis. Sei $G = (V,E)$ ein beliebiger Graph.
Wir werden einen neuen Graphen $G' = (V', E', \hat{E}')$ mit Pseudokanten aus G wie folgt konstruieren. V' besteht aus

1. der Menge V und
2. einem neuen Knoten v' für alle $v \in V$.

Wir werden die Menge $V' \backslash V$ mit $\tilde{V}$ bezeichen. Die Menge regulärer Kanten E' besteht aus

1. der Menge E und
2. einer neuen regulären Kante $\{v, v'\}$ für alle $v \in V$.

Die Menge der Pseudokanten $\hat{E}'$ besteht aus den Pseudokanten $\{a', b'\}$, falls die Kante $\{a,b\} \in E(G)$ existiert.

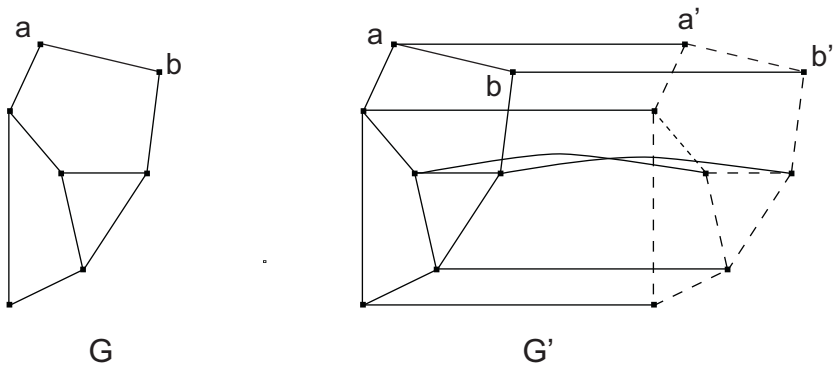

Abbildung 3.3: Konstruktion des Graphen G' aus dem Graphen G

Es ist ersichtlich, dass der Graph G' keine Überlagerungen enthält. Wir zeigen nun, dass die minimale Anzahl m von Knoten in einem Vertex Cover von G gleich der minimalen Anzahl m' von Knoten in einem Vertex Set für kritische Komponenten von G' ist.
Sei U ein Vertex Cover Set in G und C ein Zyklus mit einer ungeraden Anzahl von regulären Kanten in G'. Offensichtlich kann C nicht vollständig in $(\tilde{V}, \hat{E}')$ liegen, da dieser Teilgraph ausschließlich aus Pseudokanten besteht. Also enthält C mindestens eine Kante aus $E(G)$. Demzufolge beinhaltet C mindestens einen Knoten aus U. Somit ist U ein Vertex Set für kritische Komponenten in G', also gilt $m' \leq m$.
Nehmen wir nun an, dass gilt $m' < m$. Sei U' ein Vertex Set für kritische Komponenten in G', so dass $|U'| = m'$. Wir wissen, dass U' eine Teilmenge der Menge $V \cup \tilde{V}$ ist. Betrachten wir nun $U'' \subseteq V$, bestehend aus allen Knoten $v \in V$ mit der Eigenschaft, dass einer der Knoten v oder v' in U' liegt. Da $|U''| \leq |U'| < m$, folgt U'' ist kein Vertex Cover Set in G. Also

existiert mindestens eine Kante $\{v_i, v_j\} \in E(G)$, so dass weder v_i, noch v_j zu U'' gehören. Somit beinhaltet U' keinen der Knoten v_i, v_j, v'_i, v'_j und somit auch keinen Knoten des kritischen Zyklus $\{v_i, v_j\}, \{v_j, v'_j\}, \{v'_j, v'_i\}, \{v'_i, v_i\}$ $\Rightarrow$ WIDERSPRUCH $\Rightarrow m' = m$, demzufolge haben VSkK und VCS die gleiche Kardinalität. Wenn wir also eine Lösung vom Problem MinKN_eS finden, so ergibt sich auch eine Lösung für das VCS-Problem und umgekehrt.

Den Graphen G' können wir in linearer Zeit in Abhängigkeit der Anzahl von Knoten des Graphen G konstruieren. Des Weiteren ist ersichtlich, dass der maximale Knotengrad in G' dem maximalen Knotengrad in G plus eins entspricht. Da das *VCS*-Problem für einen Knotengrad $deg(v) \geq 3$ NP-vollständig ist, folgt die NP-Schwere in Verbindung mit dem Lemma 3.16 unseres Problems für einen Knotengrad $deg(v) \geq 4$.

Bleibt zu zeigen, dass gilt: $MinKN\_S \in NP$. Einen nichtdeterministischen Algorithmus für unser Problem erhalten wir, indem eine Lösung des Problems geraten wird, d.h. eine Teilmenge der Knotenmenge und dann, nach Entfernung dieser Knoten, mittels des Algorithmus *Test_Realisierbarkeit* verifiziert wird, ob die Lösung gültig ist. Somit haben wir bewiesen, dass das Problem *MinKN_eS* für $k \geq 3$ NP-vollständig ist. □

3.5 MinKA_eS ist NP-vollständig

Abschließend wenden wir uns nun der NP-Vollständigkeit des Problems MinKA_eS zu.

Problem 3.4. (MinKA_eS) *Berechne die minimale Anzahl von* REGULÄREN UND PSEUDOKANTEN*, die aus dem Graphen $G(\hat{S}_1, \ldots, \hat{S}_k)$ entfernt werden müssen, so dass die erweiterten Shapes $\hat{S}_1, \ldots, \hat{S}_k$ durch eine Zeichenkette realisiert werden können.*

Ein Problem, welches von Harary in [Har59] vorgestellt wurde, ist das folgende:

Problem 3.18. (BKI-Problem) *Bestimme den* Balance-Kanten-Index *für einen gegebenen S-Graphen $H = (V, E, \varphi)$, d.h. die minimale Anzahl von Kanten, die entfernt werden müssen, so dass H balanciert ist.*

Barahona bewies in [Bar82] das folgende Theorem.

Theorem 3.19. *Das BKI-Problem ist NP-vollständig für $k \geq 5$.*

Es liegt auf der Hand, dass eine Teilinstanz des Problems 3.4 das BKI Problem ist. Wir werden dennoch den formalen Beweis der NP-Vollständigkeit des Problems MinKA_eS führen. Bevor wir uns diesem Beweis zuwenden, benötigen wir das folgende Lemma.

Lemma 3.20. *Sei $G' = (V', E', \gamma')$ eine ungerade homöomorphe Erweiterung eines Graphen $G = (V, E, \gamma)$ erweiterter Shapes, so gilt: Die minimalen Kantenmengen, die in G und G' entfernt werden müssen, so dass G und G' realisiert werden können, haben die gleiche Kardinalität.*

Beweis. In jeder ungeraden homöomorphen Erweiterung G' von G existiert eine eineindeutige Zuordnung von Kanten $\{v', v''\} \in E(G)$ durch einen Pfad $P(\{v', v''\})$, falls dieser vorhanden ist, sonst durch die Kante $\{v', v''\} \in E(G')$ selbst. Insbesondere sind Zyklen mit einer ungeraden Anzahl regulärer Kanten in G auch Zyklen mit einer ungeraden Anzahl regulärer Kanten in G'.

Man kann also für jede Kante $\{v', v''\} \in E^*(G)$ in G' entweder die Kante $\{v', u_1\} \in E(G')$ entfernen, falls $P(\{v', v''\})$ existiert und sonst die Kante $\{v', v''\}$ selbst, damit auch G' realisierbar ist.

Umgekehrt ist es möglich, für alle Kanten aus $\{u, w\} \in (E(G') \backslash E(G)) \cap E^*(G')$, d.h. Kanten die Teil eines Pfades $P(\{v', v''\})$ sind, die Kante $\{v', v''\}$ in G zu entfernen und für alle $\{u, w\} \in E(G) \cap E^*(G')$ die Kante $\{u, w\}$ selbst. □

Theorem 3.21. *MinKA_eS ist NP-vollständig für $k \geq 5$.*

Beweis. Wir werden die NP-Schwere des Problems durch die Reduktion vom BKI-Problem zeigen. Sei eine beliebige Instanz vom BKI-Problem durch einen S-Graphen $H = (V, E, \varphi)$ gegeben. Wir konstruieren nun einen Graphen $G = (V, E, \hat{E}) = (V, E, \gamma)$ und zeigen, dass der Balance-Kanten-Index gerade der minimalen Anzahl k von Kanten aus $E \cup \hat{E}$ entspricht, die entfernt werden müssen, so dass in G keine Zyklen mit ungerader Anzahl regulärer Kanten und keine Überlagerungen existieren.
Sei $H = (V, E, \varphi)$ gegeben. Des Weiteren sei $V(G) = V(H)$ und

$$\gamma(e) = \begin{cases} 0 & \text{, falls } \varphi(e) = -1 \\ 1 & \text{, falls } \varphi(e) = +1 \end{cases},$$

d.h. Kanten aus H mit negativen Vorzeichen werden als reguläre Kanten und Kanten aus H mit positiven Vorzeichen werden als Pseudokanten betrachtet. Insbesondere entstehen bei dieser Konstruktion keine Überlagerungen. Es ist weiterhin ersichtlich, dass die Reduktion in polynomieller Zeit in Abhängigkeit von der Kantenanzahl geschieht.

Wir wissen, dass *BKI* Kanten aus $E(H)$ entfernt werden müssen, so dass H balanciert ist. Dies wiederum ist der Fall, gdw. in H keine Zyklen mit einer ungeraden Anzahl von negativen Kanten existieren und nach Konstruktion gilt somit für G, dass keine Zyklen mit einer

ungeraden Anzahl von regulären Kanten existieren.
Sei umgekehrt k die Anzahl von Kanten aus $E \cup \hat{E}$, die entfernt werden müssen, so dass in G keine Zyklen mit ungerader Anzahl regulärer Kanten existieren, so folgt nach Konstruktion, dass auch in H keine Zyklen mit einer ungeraden Anzahl von negativen Kanten existieren, also ist H balanciert. $\Rightarrow k = BKI$. In Verbindung mit dem Lemma 3.20 folgt die NP-Schwere des Problems.

Bleibt zu zeigen, dass gilt: $MinKA\_eS \in NP$. Einen nichtdeterministischen Algorithmus erhalten wir für unser Problem, indem eine Lösung geraten wird, d.h. eine Teilmenge der Kantenmenge und danach mittels des Algorithmus *Test_Realisierbarkeit* überprüft wird, ob die, nach der Entfernung dieser Teilmenge, erhaltene Lösung gültig ist. □

Somit haben wir gezeigt, dass das Problem MinKA_eS für $k \geq 5$ NP-vollständig ist.

3.6 Zusammenfassung

Wir haben in diesem Kapitel die vier Probleme MinKA_S, MinKN_S, MinKA_eS und MinKN_eS der minimalen Kanten- bzw. Knotenmenge, die aus einem Shapegraphen entfernt werden muss, so dass die entsprechenden Shapes bzw. erweiterten Shapes durch eine einzelne Zeichenkette realisiert werden können, motiviert und formuliert. In [CLK+05] wurde gezeigt, dass das Problem MinKN_S NP-vollständig ist. Es stellte sich somit die Frage, ob die anderen genannten Probleme auch NP-vollständig sind. Um diese Frage zu beantworten, führten wir den Begriff der homöomorphen Erweiterung ein und stellten einen Algorithmus vor, der die entsprechenden Shapegraphen auf ihre Realisierbarkeit testet. Wir haben daraufhin bewiesen, dass das Probleme MinKA_S, als auch die Probleme für erweiterte Shapes MinKA_eS und MinKN_eS NP-vollständig sind.

4 Kombinatorische Betrachtungen

Im letzten Kapitel werden wir uns der empirischen Betrachtung von Shapes und Shapegraphen zuwenden. Zur Vereinfachung werden wir Shapegraphen $G(S_1 \dots S_k)$ bzw. $G(\hat{S}_1 \dots \hat{S}_k)$ als realisierbar bezeichnen, falls die entsprechenden Shapes bzw. erweiterten Shapes nach den Definitionen 2.4 und 2.20 durch eine einzelne Zeichenkette realisierbar sind.

Wir werden im Folgenden genauer untersuchen, wie der Zusammenhang von Shapegraphen und der Realisierbarkeit der entsprechenden Shapes bzw. erweiterten Shapes in Abhängigkeit von der Knotenanzahl, der Anzahl von regulären Kanten und Pseudokanten ist. Des Weiteren werden wir Sequenzen von Aptameren und den daraus resultierenden Sekundärstrukturen mit zufällig erzeugten Sequenzen und deren Sekundärstrukturen vergleichen.

Im ersten Abschnitt dieses Kapitels werden das Vorgehen und die implementierten Algorithmen erläutert. Im zweiten Abschnitt sehen wir uns die erhobenen Daten an und analysieren diese.

Im Folgenden wird häufig der Begriff der *Länge* eines Shape bzw. Shapegraphen benutzt. Wir bezeichnen mit diesem Begriff die Anzahl der Knoten.

4.1 Vorgehen und Algorithmen

Geben wir nun einen Überblick über unser Vorgehen und die implementierten Algorithmen.

Alle Algorithmen wurden in der Programmiersprache C implementiert. Ein wichtiges Werkzeug, welches uns bei unseren Untersuchungen zur Verfügung steht, ist *RNAfold* vom *Vienna RNA Package*[1] [Hof03]. *RNAfold* liest RNA Sequenzen ein und berechnet die Struktur der eingelesenen Sequenzen, welche minimale freie Energien aufweisen. Diese wiederum werden in Klammernotation in einer Textdatei ausgegeben. Das Programm erstellt außerdem eine PostScript Datei, in der die resultierende Struktur als Graph dargestellt ist. Somit haben wir ein sehr komfortables Werkzeug zur Erzeugung der Sekundärstrukuren [Hof03, HFS+94].

[1] `http://www.tbi.univie.ac.at/~ivo/RNA/` Stand: 14.11.06

Die Klammernotationen können wir dann nutzen, um die Adjazenzmatrix der Shapes, welche die entsprechenden Sekundärstrukturen repräsentieren, zu erzeugen. In der Adajazenzmatrix werden hierbei reguläre Kanten, Pseudokanten und nicht adjazente Knoten entsprechend gekennzeichnet. Mehrfachkanten gleicher Arten von Kanten werden als eine gezählt. Überlagerungen, d.h. Mehrfachkanten aus regulären Kanten und Pseudokanten, wurden als zwei verschiedene Kanten gewertet und in der Adjazenzmatix entsprechend gekennzeichnet. Für die Überprüfung der Realisierbarkeit verwendeten wir den im dritten Kapitel vorgestellten Algorithmus 1 (*Test_Realisierbarkeit*).

Des Weiteren sei erwähnt, dass wir die erhobenen Daten mittels des Programms *SPSS*[2] *9.0.1* analysiert und visualisiert haben. Dieses Programm ist ein modular aufgebautes Programmpaket zur statistischen Analyse von Daten. Das Basismodul ermöglicht das grundlegende Datenmanagement und umfangreiche statistische und grafische Datenanalysen mit den gängigsten statistischen Verfahren [JW05].

Erklären wir nun das weitere Vorgehen. Wir werden bei unseren Betrachtungen zuerst untersuchen, ob die intuitive Annahme, dass es einen Zusammenhang zwischen der Anzahl von Knoten und regulärer Kanten gibt, experimentell bestätigt werden kann. Unser Vorgehen ist dabei wie folgt:

1. Wir werden je 200 Zufallssequenzen aus dem Alphabet $\{A,C,G,U\}$ der Länge 5 bis 800 erstellen. Die Buchstaben kommen hierbei mit gleicher Wahrscheinlichkeit vor.

2. Diese Zufallssequenzen werden wir mit *RNAfold* falten lassen,

3. um dann die Anzahl regulärer Kanten in den jeweiligen Shapes durch einfaches Durchzählen zu bestimmen.

Im nächsten Schritt wollen wir prüfen, ob es einen Zusammenhang zwischen der Anzahl regulärer Kanten und der Anzahl von Knoten in den Shapegraphen $G(S_1,S_2,S_3)$ und $G(S_1,S_2,S_3,S_4)$ gibt. Für diese Untersuchung werden wir aus den vorangegangenen 200 zufällig erstellten Sequenzen jeweiliger Länge und den daraus gefalteten Shapes, 100 mal drei bzw. vier Shapes zufällig auswählen, um danach aus diesen die entsprechenden Shapegraphen $G(S_1,S_2,S_3)$ bzw. $G(S_1,S_2,S_3,S_4)$ zu erstellen. Wir werden uns hierbei auf die Längen 5 bis 100 beschränken. Weiterhin sei zu beachten, dass bei der zufälligen Auswahl der Shapes keine gleichen Shapes erlaubt sind. Auch hier werden wir durch einfaches Durchzählen die Anzahl der regulären Kanten bestimmen. Das Vorgehen lässt sich wie folgt zusammenfassen:

[2] Statistical Product and Service Solution

1. Aus den mit *RNAfold* gefalteten Strukturen der Länge 5 bis 100 werden je 100 mal drei bzw. vier Shapes zufällig ausgewählt.

2. Daraus erstellen wir die Graphen der Shapes $G(S_1,S_2,S_3)$ bzw. $G(S_1,S_2,S_3,S_4)$.

3. Die Anzahl der regulären Kanten in $G(S_1,S_2,S_3)$ und $G(S_1,S_2,S_3,S_4)$ werden wir wieder durch einfaches Durchzählen bestimmen.

Danach werden wir uns empirisch mit der Realisierbarkeit der Graphen $G(S_1,S_2,S_3)$ und $G(S_1,S_2,S_3,S_4)$ beschäftigen. Das heißt, wir werden schauen, wie sich die Eigenschaft der Realisierbarkeit aus den Definitionen 2.4 und 2.20 mit zunehmender Anzahl von Knoten, sowie nach dem Setzen von Pseudokanten in den Graphen $G(S_1,S_2,S_3)$ und $G(S_1,S_2,S_3,S_4)$ verhält. Bei dieser Betrachtung werden wir als erstes die Realisierbarkeit der Shapegraphen dreier bzw. vierer Shapes ohne Pseudokanten betrachten und dabei untersuchen, ob es einen Zusammenhang zwischen der Anzahl von Knoten, sowie der Anzahl regulärer Kanten und der Eigenschaft der Realisierbarkeit gibt.

Als zweites werden wir die drei bzw. vier Shapes der Shapegraphen $G(S_1,S_2,S_3)$ und $G(S_1,S_2,S_3,S_4)$ der jeweiligen Längen 5 bis 100 mit Pseudokanten füllen. Wir setzen dabei solange wie möglich zufällig Pseudokanten in die Shapes, d.h. solange wie die Eigenschaften der Shapes nach Definition 2.1 bzw. 2.11 erhalten bleiben. Insbesondere berücksichtigen wir im Folgenden beim Setzen der Pseudokanten den im Kapitel *2* formulierten Realisierungsbegriff. Das heißt, Pseudokanten können nur an Positionen gesetzt werden an denen die zugehörige gefaltete Sequenz die Buchstaben AA, UU, GG oder CC enthält.

Danach werden wir die relative Häufigkeit der Anzahl gesetzter Pseudokanten und die Eigenschaft der Realisierbarkeit in den Shapegraphen untersuchen.

Nach diesen Betrachtungen werden wir die Shapes der Shapegraphen $G(S_1,S_2,S_3)$ und $G(S_1,S_2,S_3,S_4)$ zufällig mit Pseudokanten besetzen und zwar diesmal solange die Eigenschaft der Realisierbarkeit erhalten bleibt, um uns danach erneut die relative Häufigkeit der Anzahl von Pseudokanten in den Shapegraphen anzusehen. Das Vorgehen kann kurz wie folgt zusammengefasst werden.

1. Die 100 Shapegraphen $G(S_1,S_2,S_3)$ und $G(S_1,S_2,S_3,S_4)$ der Längen 5 bis 100 werden auf die Eigenschaft der Realisierbarkeit getestet.

2. Ist die Eigenschaft der Realisierbarkeit erhalten, dann werden

a) werden die Shapes durch zufälliges Setzen von Pseudokanten mit diesen gefüllt. Danach wird erneut getestet, ob die Shapegraphen erweiterter Shapes der Eigenschaft der Realisierbarkeit genügen und die relative Häufigkeit realisierbarer Shapegraphen, sowie die relative Häufigkeit der Anzahl gesetzter Pseudokanten bestimmt.

b) zufällig Pseudokanten in die Shapes der entsprechenden Shapegraphen eingesetzt und zwar solange die Eigenschaft der Realisierbarkeit erhalten bleibt. Danach wird die relative Häufigkeit der Anzahl der eingesetzen Pseudokanten bestimmt.

Zum Ende der empirischen Betrachtungen werden wir uns Shapes fester Länge zuwenden. Wir werden dafür die Shapegraphen dreier Shapes $G(S_1, S_2, S_3)$ der Länge 40 untersuchen. Dabei werden wir versuchen uns der maximal möglichen Anzahl von Pseudokanten, welche in die Shapes gesetzt werden können, so dass die Eigenschaft der Realisierbarkeit noch erhalten bleibt, anzunähern. Bei diesem Vorgehen werden wir uns nicht nur auf zufällig erzeugte Sequenzen und deren Strukturen beschränken, sondern für diese Betrachtungen auch Aptamere mit einbeziehen. Diese sind aus der Ellington Labor Datenbank[3] der Universität Texas entnommen. Wir werden demzufolge schauen, ob sich Unterschiede in der durchschnittlichen Anzahl gesetzter Pseudokanten zwischen den Shapes der Aptamere und den von Zufallsequenzen feststellen lassen.

Wir werden bei diesen Untersuchungen zwei Verfahren benutzen, um Pseudokanten in die Shapes zu setzen. Das eine setzt zufällig Pseudokanten in die Shapes. Das andere Verfahren besetzt die Shapes nach einer festen Regel mit Pseudokanten. Wir wollen dieses Verfahren *"Sättige Shape"* nennen. Es ist im Algorithmus 2 am Ende des Abschnitts festgehalten und wird an einem einfachen Beispiel genauer erklärt.

Das Beispiel zeigt die Repräsentation einer Sekundärstruktur der Länge 30 in Klammernotation und der zugehörigen Sequenz. Wir werden in dieser Notation Pseudokanten mit eckigen Klammern darstellen.

. . . . . (((((. . . . .))) .)) . .
A U U A C U A G U C A C A A U A A U U G A G G U G U A C A C

Die farblich gekennzeichneten Knoten stellen diejenigen Stufen dar, auf denen Knoten durch Pseudokanten verbunden werden könnten, so dass die grundlegenden Eigenschaften eines Shapes erhalten bleiben. Es sei erwähnt, dass wir für unsere Betrachtungen den festen Parameter $\theta = 3$ gewählt haben.

[3] `http://aptamer.icmb.utexas.edu` Stand: 12.10.06

```
[ . . . . . ( ( ( ( ( . . . . . . . . . . ) ) ) . ) ) ] .
A U U A C U A G U C A C A A U A A U U G A G G U G U A C A C

[ [ . . . ] . ( ( ( ( ( . . . . . . . . . . ) ) ) . ) ) ] .
A U U A C U A G U C A C A A U A A U U G A G G U G U A C A C

[ [ . . . ] . ( ( ( ( ( [ . . . . . . . ] . ) ) ) . ) ) ] .
A U U A C U A G U C A C A A U A A U U G A G G U G U A C A C

[ [ . . . ] . ( ( ( ( ( [ . [ . . . ] . ] . ) ) ) . ) ) ] .
A U U A C U A G U C A C A A U A A U U G A G G U G U A C A C
```

Im ersten Schritt wird der kleinste Knoten v und der Knoten w mit dem grössten Abstand zu v auf der entsprechenden Stufe gewählt. Es wird daraufhin geprüft, ob diese Knoten durch eine Pseudokante verbunden werden können. Ist dies der Fall, wird die Pseudokante ausgebildet und dieser Schritt für die anderen Knoten dieser Stufe wiederholt, sofern welche existieren. Ist dies nicht der Fall, so wird der nächstkleinere Knoten w mit grösstem Abstand zu v gewählt und erneut geprüft ob Pseudokanten ausgebildet werden können. Dies wird solange wiederholt, wie Stufen freier Knoten mit den erforderlichen Voraussetzungen existieren.

Um eine genauere Aussage treffen zu können, wieviel Pseudokanten maximal in einen Shapegraphen $G(S_1,S_2,S_3)$ fester Länge gesetzt werden können, ohne die genaue Lösung zu kennen, werden wir im Folgenden 10 Instanzen von Shapegraphen dreier Shapes aus Aptameren bzw. aus gefalteten Zufallssequenzen entnehmen und in jede dieser Instanzen 50 mal auf verschiedene Arten Pseudokanten setzen, um uns der Lösung der maximale Anzahl von Pseudokanten anzunähern. Fassen wir das Vorgehen zusammen.

1. Es werden zunächst 100 Sequenzen der Länge 40 aus der Datenbank von Aptameren entnommen. Aus diesen werden zufällig 10 mal drei ausgewählt, um aus diesen dann die entsprechenden Shapegraphen $G(S_1,S_2,S_3)$ zu erstellen. Gleiche Sequenzen bei der zufälligen Wahl sind nicht erlaubt.

2. Des Weiteren werden aus dem Pool von Shapegrahen dreier Shapes $G(S_1,S_2,S_3)$ zufälliger Sequenzen 10 zufällig ausgewählt. Gleiche Shapegraphen bei der zufälligen Wahl sind nicht erlaubt.

3. Wir werden dann die Shapegraphen $G(S_1,S_2,S_3)$ von Aptameren bzw. der Shapes aus Zufallssequenzen 50 mal auf je drei verschiedene Arten mit Pseudokanten besetzen, so dass die Eigenschaft der Realisierbarkeit erhalten bleibt. Diese Verfahren seien im Folgenden kurz erklärt. Falls die Shapegraphen $G(S_1,S_2,S_3)$ realisierbar sind, dann werden wir

a) die Shapes zufällig mit Pseudokanten besetzen, bis entweder keine Pseudokanten mehr gesetzt werden können oder die erweiterten Shapes nicht mehr durch eine einzelne Zeichenkette realisiert werden können, somit nähern wir uns der maximalen Anzahl von Pseudokanten von unten an. Dieses Verfahren ist im Algorithmus 3 dargestellt und wird im Folgenden mit *PsK_rein* bezeichnet.

b) die Shapes mit dem Verfahren 'Sättigen der Shapes' füllen, um danach

   i. zufällig Pseudokanten zu entfernen, bis die Shapes wieder durch eine einzelne Zeichenkette realisiert werden können (Algorithmus 4), bzw.

   ii. zufällig Knoten zu entfernen, an denen Pseudokanten liegen, bis die Shapes wieder durch eine Zeichenkette realisierbar sind (Algorithmus 5).

   Das heißt wir nähern uns in den letzten beiden Verfahren der maximalen Anzahl von Pseudokanten, so dass die Eigenschaft der Realisierbarkeit erhalten bleibt, von oben an. Diese Verfahren werden wir im Folgenden zur Vereinfachung mit *PsK_raus* und *Knoten_raus* bzeichnen.

Am Ende diesen Abschnitts sind die Pseudocodes der genannten Algorithmen aufgeführt.

4.1.1 Pseudocodes der Algorithmen

Algorithm 2 Sättige Shape

1: **INPUT:** Shape
2: {Stufen freier Knoten bezeichnen die Mengen der Knoten die durch Pseudokanten verbunden werden können}
3: **while** Stufen freier Knoten existieren **do**
4: **for all** Stufen freier Knoten **do**
5: v:=0; w:=0;
6: last:=Anzahl freier Knoten der aktuellen Stufe;
7: **for** $k = 0$ to *Anzahl freier Knoten der aktuellen Stufe* **do**
8: **for** $m = last$ to 0 **do**
9: v = k.ter freier Knoten in gewählter Sequenz;
10: w = m.ter freier Knoten in gewählter Sequenz;
11: **if** $s_v s_w \in \{AA, UU, GG, CC\}$ **then**
12: **if** (w > v) UND (w − v > θ) **then**
13: $k = k+1$; $last = m-1$;
14: verbinde Knoten v und w durch Pseudokante;
15: **end if**
16: **end if**
17: **end for**
18: **end for**
19: **end for**
20: Ermittle Sequenz freier Knoten
21: **end while**

Algorithm 3 PsK_rein

INPUT: Shape S_1, S_2, S_3
Anzahl_PseudoKanten:=0;
for $i := 1$ to 3 **do**
    gefüllt[i]:=false;
end for
while ($G(\hat{S}_1, \hat{S}_2, \hat{S}_3)$ ist realisierbar) UND
        ($\exists\, i \in \{1,2,3\}$, so dass gefüllt[i] := false) **do**
    wähle zufällig Shape S_i, für den gilt gefüllt[i]:=false;
    if Möglichkeit besteht in S_i PseudoKanten zu setzen **then**
        setze 1 PseudoKanten zufällig in gewählten Shape S_i;
    else
        gefüllt[i]:=true
    end if
end while
Zähle PseudoKanten;
if Anzahl_PseudoKanten = 0 **then**
    gib aus: Anzahl_PseudoKanten;
else
    gib aus: Anzahl_PseudoKanten-1;
end if

Algorithm 4 PsK_raus

INPUT: Shape S_1, S_2, S_3
Anzahl_PseudoKanten:=0;
if $G(S_1, S_2, S_3)$ ist realisierbar **then**
    Sättige die Shapes S_1, S_2, S_3;
    Erstelle $G(\hat{S}_1, \hat{S}_2, \hat{S}_3)$;
    while $G(\hat{S}_1, \hat{S}_2, \hat{S}_3)$ ist nicht realisierbar **do**
        entferne zufällig gewählte PseudoKante;
    end while
    zähle PseudoKanten;
end if
gib aus: Anzahl_PseudoKanten;

Algorithm 5 Knoten_raus

1: **INPUT:** Shape S_1, S_2, S_3
2: Anzahl_PseudoKanten:=0;
3: **if** $G(S_1, S_2, S_3)$ ist realisierbar **then**
4: Sättige die Shapes S_1, S_2, S_3;
5: Erstelle $G(\hat{S}_1, \hat{S}_2, \hat{S}_3)$;
6: **while** $G(\hat{S}_1, \hat{S}_2, \hat{S}_3)$ ist nicht realisierbar **do**
7: entferne zufällig gewählten Knoten an dem PseudoKante liegt
8: **end while**
9: zähle PseudoKanten;
10: **end if**
11: gib aus: Anzahl_PseudoKanten;

Nachdem wir nun erklärt haben, wie wir vorgehen werden, schauen wir uns im nächsten Abschnitt die erhobenen Daten an und analysieren diese.

4.2 Datenanalyse

Es sei kurz erwähnt, dass wir im Folgenden keine Annahmen über die Verteilung der Daten machen werden und wollen. Deswegen werden wir nichtparametrische Tests nutzen, um Zusammenhänge zu analysieren und zu beschreiben. Ein häufig verwendetes Maß, um monotone Zusammenhänge ohne die Annahme einer Verteilung zu beschreiben ist der Rangkorrelationskoeffizient nach Spearman [Wae57]. Dieser nichtparametrische Test prüft die nach Rängen transformierten Variablen auf einen monotonen Zusammenhang, d.h. auf einen Zusammenhang, der sich durch eine monotone Funktion beschreiben lässt. Zudem reagiert dieser Korrelationskoeffizient weniger stark auf Ausreißer.

Definition 4.1. [Wae57] *Berechnung des Rangkorrelationskoeffizienten nach Spearman: Seinen N Datenpunkte* $(x_1,y_1),\ldots,(x_N,y_N)$ *gegeben.*

1. *Weise jedem beobachteten* x_i *einen Rang* X_i *wie folgt zu:*
   - $X_i = 1$ *für den kleinsten Wert* x_i
   - $X_i = N$ *für den grössten Wert* x_i
   - *Sind zwei oder mehr Werte gleich, so wird der Durchschnittsrang zugewiesen.*
2. *Wiederhole den ersten Schritt für die Variable* Y.
3. *Berechne den Rangkorrelationskoeffizienten* $\rho = 1 - 6\sum_{i=1}^{N} \frac{(X_i - Y_i)^2}{N(N^2-1)}$.

Es sei erwähnt, dass der Korrelationskoeffizient nach Spearman, genau wie der bekannte Korrelationskoeffizient nach Pearson, immer zwischen -1 und 1 liegt. Somit gilt, dass je näher ρ an 1 bzw. je näher ρ an -1 liegt, desto höher ist der entsprechende positive bzw. negative monotone Zusammenhang der getesteten Variablen, d.h. desto besser lässt sich der entsprechende Zusammenhang mittels einer monoton wachsenden bzw. fallenden Funktion beschreiben.

4.2.1 Anzahl Knoten vs. Anzahl reguläre Kanten

Schauen wir uns nun an, ob es einen Zusammenhang zwischen der Anzahl der Knoten und der relativen Häufigkeit der Anzahl regulärer Kanten in den einzelnen Shapes gibt. Das entsprechende Streudiagramm ist in Abbildung 4.1 zu sehen.

Bei der Betrachtung der Abbildung 4.1 wird schnell klar, dass es einen monotonen postitiven Zusammenhang geben muss. Die Analyse mit dem Korrelationskoeffizient nach Spearman bestätigt diese Vermutung. So hat der berechnete Korrelationskoeffizient mit $\rho = 1,00$

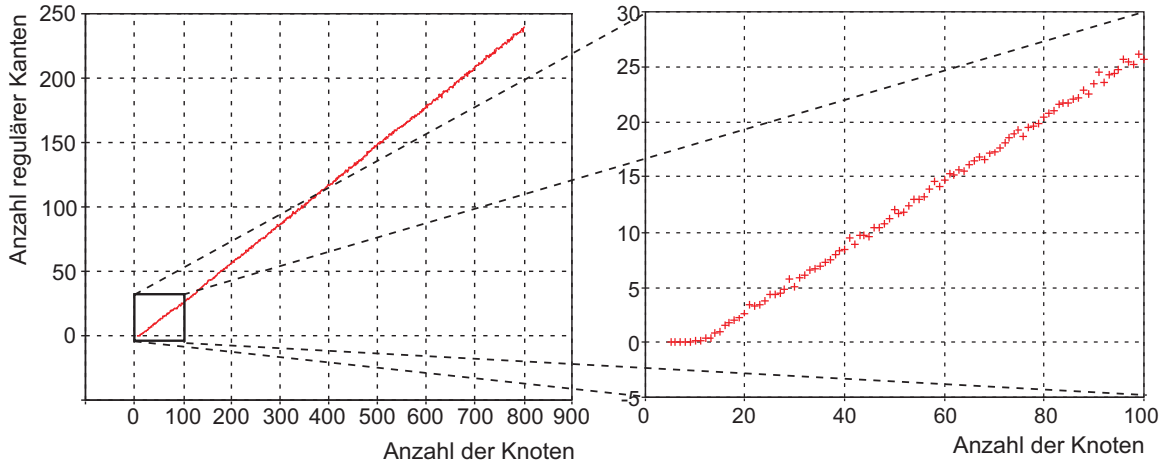

Abbildung 4.1: Streudiagramm des Zusammenhangs zwischen der Anzahl regulärer Kanten in den einzelnen Shapes verschiedener Längen in Abhängigkeit von der Anzahl der Knoten.

auf dem 0,01 Signifikanzniveau sein Maximum erreicht. Dies bedeutet dass sich der Zusammenhang zwischen der Anzahl von regulären Kanten in Abhängigkeit von der Anzahl der Knoten mit einer Irrtumswahrscheinlichkeit von 0,01 ausgesprochen gut durch eine monotone Funktion beschreiben lässt. Offensichtlich handelt es sich hierbei um eine affine Funktion. Diese werden wir später mittels der Regressionsanalyse [Wae57] bestimmen. Schlussfolgernd können wir sagen, dass mit zunehmender Anzahl der Knoten auch die Anzahl regulärer Kanten steigt.

Man sieht weiterhin, dass die relative Häufigkeit regulärer Kanten in den Shapes der Längen 5 bis 9 nahezu bei 0 ist. Dies bedeuted, dass die entsprechenden Sequenzen mit Hilfe von *RNAfold* nicht gefaltet wurden. Biologisch könnte das auf die Starrheit kurzer Moleküle zurückgeführt werden.

Schauen wir nun, ob es auch einen Zusammenhang zwischen der Anzahl der Knoten und der Anzahl regulärer Kanten in den Shapegraphen $G(S_1, S_2, S_3)$ von drei und den Shapegraphen $G(S_1, S_2, S_3, S_4)$ von vier Shapes gibt. Die relativen Häufigkeiten der Anzahl der regulären Kanten in Abhängigkeit von der Anzahl der Knoten sind im Streudiagramm in Abbildung 4.2 dargestellt.

Bei der Betrachtung der Abbildung 4.2 ist erneut ein starker positiver monotoner Zusammenhang erkennbar. Das Maß des Zusammenhangs lässt sich für den Shapegraphen dreier, als auch für den vierer Shapes, nach dem Korrelationstest nach Spearman auf dem 0,01 Signifikanzniveau mit $\rho = 0,999$ beziffern. Dies bestätigt die Annahme, dass es einen positiven Zusammenhang zwischen der Anzahl der Knoten und regulärer Kanten gibt. Offensichtlich

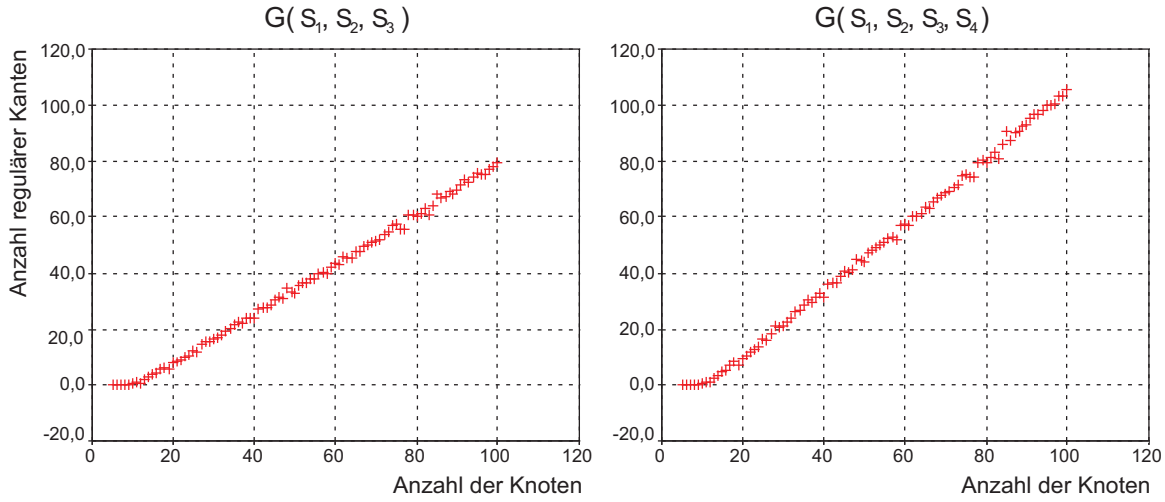

Abbildung 4.2: Dargestellt ist die durchschnittliche Anzahl regulärer Kanten in Abhängigkeit, von der Anzahl der Knoten in den Shapegraphen $G(S_1, S_2, S_3)$ und $G(S_1, S_2, S_3, S_4)$.

gilt auch hier die Annahme, dass mit zunehmender Anzahl von Knoten die Möglichkeit reguläre Kanten zu bilden steigt. Es ist allerdings überraschend, dass die monotonen Funktionen, die diesen Zusammenhang beschreiben könnten, ausgesprochen affin aussehen.

Man hätte erwarten können, dass der Anstieg bei kleiner Knotenanzahl langsamer steigt und mit zunehmender Anzahl von Knoten steiler wird, da bei Strukturen kleinerer Längen, die Möglichkeit reguläre Kanten an verschiedenen Stellen auszubilden geringer ist als bei Instanzen grösserer Länge. Dies hätte zur Folge, dass sich reguläre Kanten verschiedener Sekundärstrukturen kleiner Instanzen eher überlagern, als in denen mit einer größeren Anzahl von Knoten. Das wiederum bedeutet, dass sie im Shapegraphen nur als eine Kante gezählt werden und somit die Anzahl der Kanten am Anfang der Funktion langsamer steigt und die Kurve mit einer zunehmenden Anzahl von Knoten steiler wird.

Zwar ist ersichtlich, dass der Anstieg bei einer Knotenanzahl zwischen 5 und 9 konstant und insbesondere wesentlich kleiner ist als bei einer grösseren Anzahl von Knoten, dies ist aber darauf zurückzuführen, dass die einzelnen Zufallssequenzen der Längen 5 bis 9 gar nicht erst gefaltet wurden, d.h. es existieren keine regulären Kanten in den Shapes, wie wir in der Abbildung 4.1 gesehen haben. Nachdem aber die Möglichkeit besteht, dass Sequenzen mit *RNAfold* gefaltet wurden, bleibt der Anstieg konstant, wie in der Abbildung 4.2 zu erkennen ist. Somit können wir schlussfolgern: Wenn es zu einer Überlagerung regulärer Kanten kommt, passiert dies in den Shapegraphen immer in gleichem Maß.

Untersuchen wir also nun die Frage nach dem Maß der Überlagerungen regulärer Kanten

genauer. Wir werden dafür als erstes die monotonen Funktionen, welche den Zusammenhang der Streudiagramme in den Abbildungen 4.1 und 4.2 beschreiben, mit Hilfe der Regressionsanalyse bestimmen. Wir erhalten folgende affine Funktionen, welche in Abbildung 4.3 dargestellt sind:

- für einen Shape: $g_1(x) := 0,294x - 3,081$,
- für $G(S_1,S_2,S_3)$: $g_3(x) := 0,877x - 9,274$
- und für $G(S_1,S_2,S_3,S_4)$: $g_4(x) := 1,167x - 12,535$.

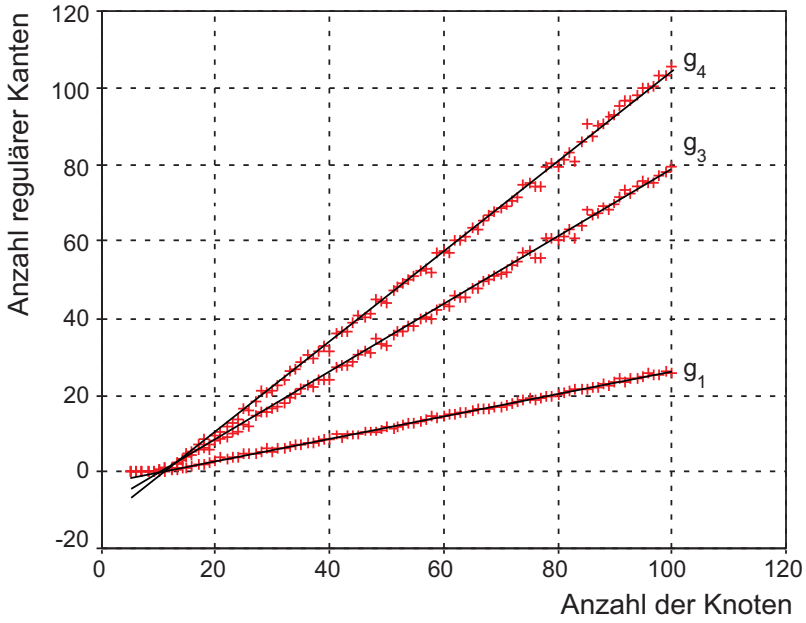

Abbildung 4.3: Dargestellt ist die Anzahl regulärer Kanten in Abhängigkeit von der Knotenanzahl von einem Shape, sowie von den Shapegraphen dreier und vierer Shapes. Zusätzlich sind die, durch die Regressionsanalyse bestimmten, affinen Funktionen g_1, g_3 und g_4 abgebildet.

Bei genauerer Betrachtung der Funktionen g_1, g_3 und g_4 fällt folgendes auf:

- $g_3 \approx 3 * g_1 = 0,882x - 9,243$ und
- $g_4 \approx 4 * g_1 = 1,176x - 12,324$.

Man sieht, dass die Anzahl regulärer Kanten in einem Shapegraphen dreier bzw. vierer Shapes etwa das Drei- bzw. Vierfache der Anzahl der regulären Kanten in einem Shape

beträgt. Somit lässt sich für die gefalteten Sequenzen und den daraus resultierenden Shapegraphen dreier bzw. vierer Shapes die Anzahl regulärer Kanten sehr gut durch die Anzahl regulärer Kanten in einem Shape approximieren. Aus dieser Erkenntnis können wir nur einen Schluss ziehen: Es tritt bei der Bildung der Shapegraphen aus den entsprechenden Shapes nur sehr selten die Bildung von Mehrfachkanten auf.

Des Weiteren stellen wir somit fest, dass es einen Unterschied zwischen der Anzahl der regulären Kanten zwischen den Graphen $G(S_1,S_2,S_3)$ und $G(S_1,S_2,S_3,S_4)$ gibt. Offensichtlich ist die Anzahl der Kanten in dem Shapegraphen $G(S_1,S_2,S_3,S_4)$ bei entsprechender Anzahl von Knoten höher, als in dem Graphen dreier Shapes. Dieser Punkt ist aber einfach zu erklären, wenn man sich überlegt, dass einmal drei und einmal vier Shapes in einem Graphen zusammengefasst werden und den zusätzlich ebengenannten Schluss, dass es selten zu Mehrfachkantenbildung kommt, in diese Überlegung mit einbezieht.

4.2.2 Anzahl Knoten vs. Eigenschaft "Realisierbar"

Wenden wir uns nun der Betrachtung der relativen Häufigkeiten der Eigenschaft der Realisierbarkeit der Shapegraphen ohne Pseudokanten in Abhängigkeit von der Knotenanzahl zu. Die entsprechenden Streudiagramme sind in Abbildung 4.4 dargestellt.

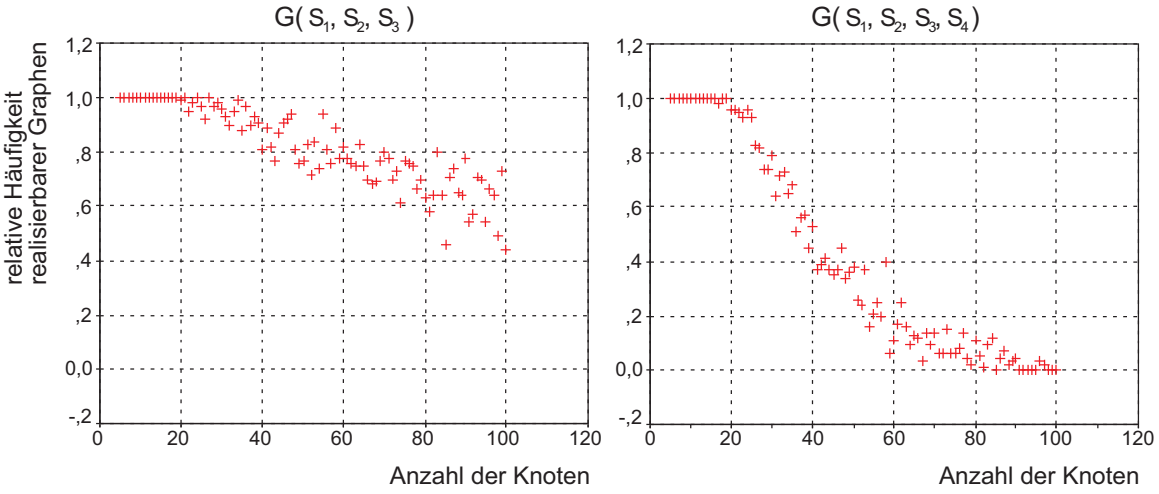

Abbildung 4.4: Dargestellt ist die relative Häufigkeit der Eigenschaft der Realisierbarkeit von den Shapegraphen $G(S_1,S_2,S_3)$ und $G(S_1,S_2,S_3,S_4)$ ohne Pseudokanten in Abhängigkeit von der Knotenanzahl.

In beiden Fällen sieht man, dass die Eigenschaft der Realisierbarkeit mit zunehmender Knotenanzahl abnimmt. Dass es einen monotonen Zusammenhang gibt, wird durch die ge-

messenen hohen negativen Rangkorrelationen mit $\rho = -0,927$ für Shapegraphen dreier Shapes und $\rho = -0,976$ für den vierer Shapes, auf dem 0,01 Signifikanzniveau, bestätigt. Wir haben in den vorangegangenen Abbildungen gesehen, dass mit zunehmender Knotenanzahl die Anzahl der regulären Kanten stieg. Aus diesem und den vorangegangenen Ergebnissen können wir schlussfolgern, dass mit zunehmender Anzahl von Knoten und daraus resultierend mit zunehmender Anzahl regulärer Kanten, die Strukturen in den Graphen komplizierter werden, d.h. die Möglichkeit Zyklen ungerader Länge auszubilden und demzufolge die Eigenschaft der Realisierbarkeit zu verlieren, nimmt mit zunehmender Knotenanzahl und demzufolge mit zunehmender Anzahl von regulären Kanten zu.

Diese Annahme lässt sich auch dadurch bekräftigen, dass im Shapegraphen vierer Shapes $G(S_1,S_2,S_3,S_4)$ die relative Häufigkeit der Eigenschaft der Realisierbarkeit deutlich schneller abnimmt als im Shapegraphen dreier Shapes. Die Eigenschaft der Realisierbarkeit im Shapegraphen $G(S_1,S_2,S_3)$ ist selbst bei einer großen Knotenanzahl noch zu etwa 50% vorhanden, während im Graphen $G(S_1,S_2,S_3,S_4)$ die Eigenschaft der Realisierbarkeit deutlich schneller abnimmt und bei Instanzen der Länge grösser als 60 schon deutlich unter 20% liegt. Wie wir in Abbildung 4.3 gesehen haben, existieren im Graphen $G(S_1,S_2,S_3,S_4)$ mit zunehmender Knotenanzahl auch mehr reguläre Kanten, als in dem Shapegraphen $G(S_1,S_2,S_3)$ mit derselben Anzahl von Knoten. Somit kommen im Shapegraphen dreier Shapes weniger reguläre Kanten vor, als in dem Graphen vierer Shapes, obwohl die Eigenschaft der Realisierbarkeit länger erhalten bleibt. Daraus können wir schlussfolgern: Bei gleicher Anzahl von Knoten im Graphen $G(S_1,S_2,S_3)$ entstehen seltener Zyklen ungerader Länge als in dem Shapegraphen $G(S_1,S_2,S_3,S_4)$. Demzufolge ist die Eigenschaft der Realisierbarkeit für die Shapegraphen gefalteten Zufallssequenzen abhängig von der Anzahl der Knoten und daraus resultierend von der Anzahl regulärer Kanten.

Bevor wir uns nun mit der Ausprägungen der Realisierbarkeit, nach dem zufälligen Setzen von Pseudokanten beschäftigen, schauen wir uns zuerst das Streudiagramm in Abbildung 4.5 an. Dieses zeigt die Anzahl der Pseudokanten in Abhängigkeit von der Anzahl der Knoten. Wie schon erwähnt, wurden hierbei alle entsprechenden Shapes zufällig mit Pseudokanten gefüllt. Die grundlegenden Forderungen an einen Shape dürfen dabei nicht verletzt werden und auf die Eigenschaft der Realisierbarkeit wurde keine Rücksicht genommen.

Man erkennt wieder, dass mit zunehmender Anzahl von Knoten auch die Anzahl der Pseudokanten steigt. Die Annahme, dass es einen monotonen Zusammenhang gibt, wird durch die aus unseren Daten berechneten ausgesprochen hohen Korrelationskoeffizienten bestätigt. In unseren Fällen ist $\rho = 0.995$ für den Shapegraphen dreier Shapes und $\rho = 0,996$ für den

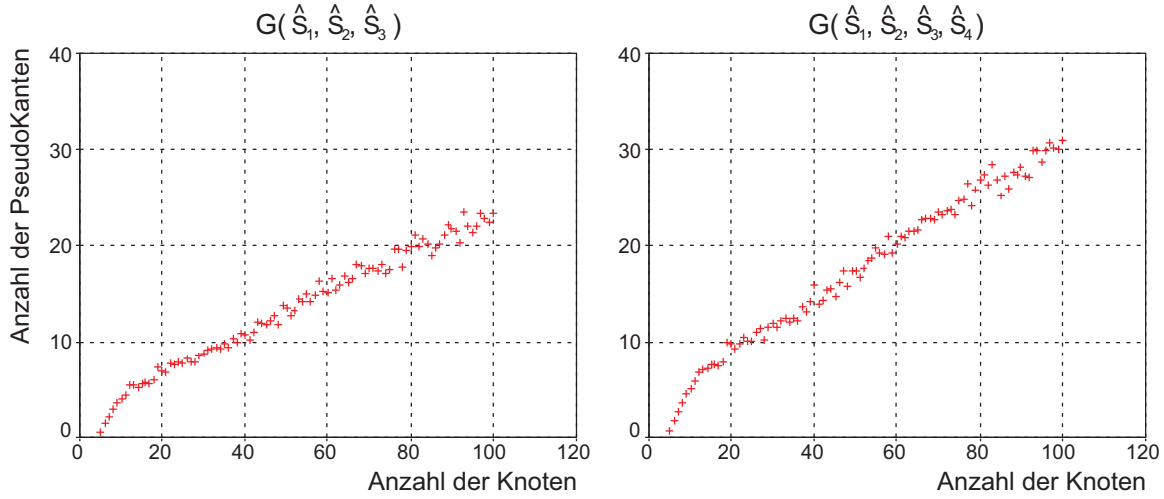

Abbildung 4.5: Dargestellt ist die durchschnittliche Anzahl der Pseudokanten in Abhängigkeit von der Knotenanzahl, ohne der Voraussetzung zu genügen, dass die entprechenen erweiterten Shapes durch eine einzelne Zeichenkette realisiert werden können.

Shapegraphen vierer Shapes. Beide Koeffizienten wurden auf einem 0,01 Signifikanzniveau berechnet. Es lässt sich weiterhin beobachten, dass durchschnittlich mehr Pseudokanten in den Shapegraphen $G(S_1, S_2, S_3, S_4)$ gesetzt wurden, als in den Graphen $G(S_1, S_2, S_3)$ dreier Shapes gleicher Länge. Auf eine Regressionsanalyse, um ein Verständnis zu bekommen, wie hoch der Anteil der Bildung von Mehrfachkanten ist, wollen wir im Zuge dieser Betrachtungen verzichten. Sehen wir uns stattdessen die Ausprägungen der Realisierbarkeit, nach dem zufälligen Setzen von Pseudokanten in die Shapes der entsprechenden Shapegraphen in Abbildung 4.6 an.

Wir erkennen wieder einen deutlichen Unterschied zwischen den Graphen $G(\hat{S}_1, \hat{S}_2, \hat{S}_3)$ und $G(\hat{S}_1, \hat{S}_2, \hat{S}_3, \hat{S}_4)$. Die Realisierbarkeit im Shapegraphen $G(\hat{S}_1, \hat{S}_2, \hat{S}_3)$ bei gleicher Anzahl von Knoten bleibt häufiger nach dem Setzen von Pseudokanten erhalten, als in dem Graphen $G(\hat{S}_1, \hat{S}_2, \hat{S}_3, \hat{S}_4)$. Im Shapegraphen vier erweiterter Shapes ist die relative Häufigkeit realisierbarer Graphen bei einer Knotenanzahl größer als 37 bei 0, d.h. keiner der 100 Shapegraphen $G(\hat{S}_1, \hat{S}_2, \hat{S}_3, \hat{S}_4)$ der jeweiligen Länge größer als 37 behielt nach dem Setzen von Pseudokanten noch die Eigenschaft, dass die zugehörigen Shapes durch eine einzelne Zeichenkette realisierbar sind. Bei dem Shapegraphen dreier erweiterter Shapes $G(\hat{S}_1, \hat{S}_2, \hat{S}_3)$ nimmt die Anzahl nicht realisierbarer Graphen deutlich langsamer ab. So sind etwa die relativen Häufigkeiten zwischen den Instanzen der Längen zwischen 20 und 40 beim Graphen dreier Shapes durchschnittlich bei 0,429 und beim Graphen vierer Shapes nur noch bei 0,033.

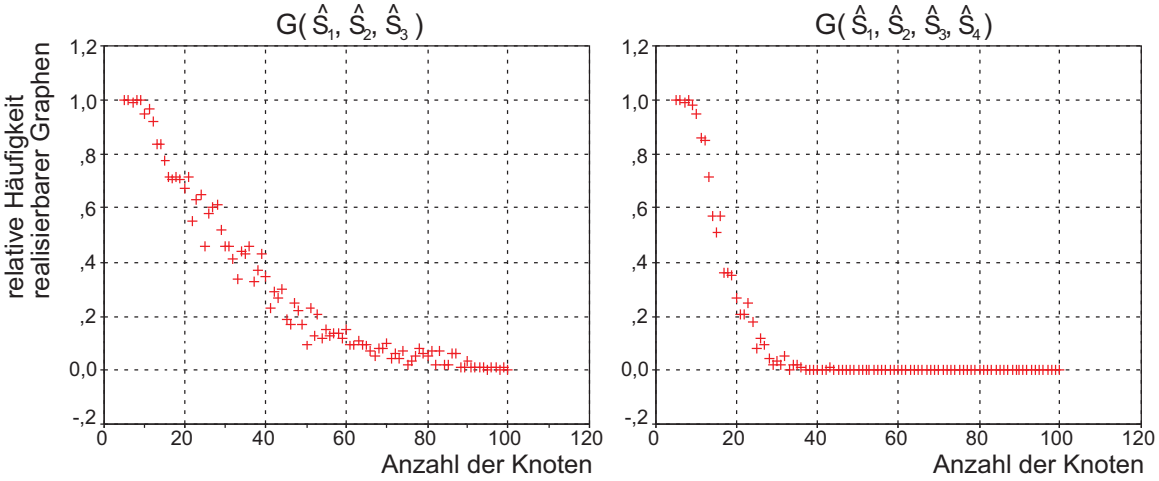

Abbildung 4.6: Dargestellt ist die relative Häufigkeit realisierbarer Shapegraphen erweiterter Shapes in Abhängigkeit von der Knotenanzahl nach dem zufälligen Setzen von Pseudokanten.

Selbst bei Instanzen der Länge zwischen 80 und 100 gibt es im Graphen $G(\hat{S}_1, \hat{S}_2, \hat{S}_3)$ bei unseren Instanzen noch realisierbare Strukturen.

Dies lässt sich auf der einen Seite darauf zurückführen, dass der Graph vierer Shapes $G(S_1, S_2, S_3, S_4)$ schon ohne Pseudokanten die Eigenschaft der Realisierbarkeit deutlich weniger häufig bei entsprechender Länge besitzt als der Graph $G(S_1, S_2, S_3)$,wie in Abbildung 4.4 erkennbar ist. Auf der anderen Seite haben wir eben gesehen, dass mit zunehmender Anzahl von Knoten auch die Anzahl von regulären Kanten und Pseudokanten steigt. Insbesondere ist diese Zunahme in dem Shapegraphen vierer Shapes höher als in dem Graphen dreier Shapes. Somit können wir auch hier wieder schlussfolgern, dass mit der zunehmenden Anzahl von Knoten und somit mit zunehmender Anzahl regulärer Kanten die Möglichkeit, dass kritische Zyklen bzw. Überlagerungen von Pseudo- mit regulären Kanten nach dem Setzen von Pseudokanten entstehen, steigt.

Schauen wir nun, wie hoch die durchschnittliche Anzahl von Pseudokanten in den Shapegraphen nach dem zufälligen Setzen von Pseudokanten ist, so dass die Eigenschaft der Realisierbarkeit in den Shapegraphen erhalten bleibt. Wir werden dies später noch genauer an Shapegraphen fester Länge untersuchen. Die entsprechenden Streudiagramme können der Abbildung 4.7 entnommen werden.

Es sind auch hier wieder, wie vermutet, deutliche Unterschiede zwischen den Graphen $G(\hat{S}_1, \hat{S}_2, \hat{S}_3)$ und $G(\hat{S}_1, \hat{S}_2, \hat{S}_3, \hat{S}_4)$ zu sehen. Die durchschnittliche Anzahl von gesetzten Pseudokanten in den Shapegraphen $G(\hat{S}_1, \hat{S}_2, \hat{S}_3)$ und $G(\hat{S}_1, \hat{S}_2, \hat{S}_3, \hat{S}_4)$ über alle Längen sind in der

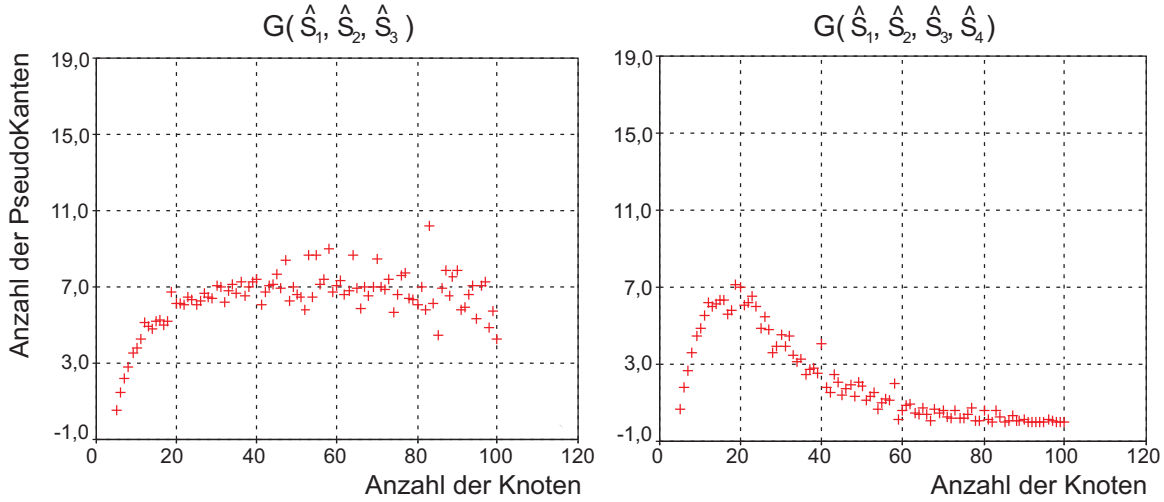

Abbildung 4.7: Dargestellt ist die durchschnittliche Anzahl der Pseudokanten in Abhängigkeit von der Anzahl der Knoten unter Voraussetzung, dass die entprechenen erweiterten Shapes durch eine einzelne Zeichenkette realisiert werden können.

Graph	Mittelwert Pseudokanten
$G(\hat{S}_1, \hat{S}_2, \hat{S}_3)$	6,3984
$G(\hat{S}_1, \hat{S}_2, \hat{S}_3, \hat{S}_4)$	2,1326

Tabelle 4.1: Mittelwert der durchschnittlichen Anzahl gesetzter Pseudokanten über die Anzahl aller Knoten der jeweiligen Graphen $G(\hat{S}_1, \hat{S}_2, \hat{S}_3)$ und $G(\hat{S}_1, \hat{S}_2, \hat{S}_3, \hat{S}_4)$.

Tabelle 4.1 festgehalten.

Man erkennt einen deutlichen Unterschied bei diesen Werten. So ist die durchschnittliche Anzahl gesetzter Pseudokanten, mit einem Wert von etwa 6,4, in die Shapegraphen dreier Shapes fast drei mal so hoch, wie in den Graphen vierer Shapes, in denen ein durchschnittlicher Wert von etwa 2,1 berechnet wurde.

Bei der Betrachtung der Streudiagramme wird deutlich, dass am Anfang, d.h. bei einer Knotenanzahl zwischen 5 und 10, die Kurve sehr steil steigt. Dieser steile Anstieg ist einfach zu erklären, wenn man sich vor Augen führt, dass in diese Shapes keine bzw. ausgesprochen wenige regulären Kanten gesetzt wurden. Somit lassen sich Shapegraphen dieser Länge immer realisieren. Demzufolge ist die einzige Einschränkung an das Setzen der Pseudokanten die Anzahl der Knoten, sowie die zugehörige Sequenz und nicht der Realisierungsbegriff als solcher. Mit zunehmender Anzahl von Knoten im Shapegraphen $G(\hat{S}_1, \hat{S}_2, \hat{S}_3, \hat{S}_4)$ ist ein deutlicher Abfall der Kurve zu sehen, welcher sich im Graphen dreier erweiterter Shapes nur

vermuten lässt. Bringen wir die beiden schon betrachteten Komponenten "Anzahl reguläre Kanten in Abhängigkeit der Anzahl von Knoten" und "Eigenschaft der Realisierbarkeit in Abhängigkeit der Anzahl von Knoten" zusammen, so lassen sich die eben gemachten Beobachtungen einfach erklären.

Wir haben gesehen, dass mit zunehmender Anzahl von Knoten die Anzahl von regulären Kanten steigt und die Eigenschaft der Realisierbarkeit in den Shapegraphen ohne Pseudokanten abnimmt. Dies geschah in den Shapegraphen vierer Shapes deutlich schneller als in den Graphen dreier Shapes. Da in die Shapegraphen, deren Shapes nicht durch eine einzelne Zeichenkette realisiert werden können, keinen Pseudokanten gesetzt wurden, können wir folgern, dass in die Shapegraphen $G(S_1,S_2,S_3,S_4)$ vierer Shapes entsprechen häufiger *keine* Pseudokanten gesetzt wurden, als in die Graphen $G(S_1,S_2,S_3)$. Dies erklärt den schnellen Abfall der Kurve im linken Streudiagramm der Abbildung 4.7.

Des Weiteren lässt sich vermuten, dass selbst bei den noch realisierbaren Strukturen nach dem Setzen von Pseudokanten bei einer höheren Anzahl regulärer Kanten, die Möglichkeit, bisher unkritische Zyklen kritisch zu machen bzw. Überlagerungen zu bilden, entsprechend steigt. Da in den Shapegraphen $G(S_1,S_2,S_3,S_4)$ die durchschnittliche Anzahl regulärer Kanten bei entsprechender Länge höher ist, als in den Graphen dreier Shapes, nimmt somit auch die Anzahl gesetzter Pseudokanten mit zunehmender Anzahl von Knoten ab.

4.2.3 Aptamer vs. Zufall

Zum Schluss unserer empirischen Betrachtungen werden wir uns Shapegraphen von drei Shapes $G(S_1,S_2,S_3)$ mit einer Anzahl von 40 Knoten genauer ansehen und untersuchen, ob wir eine Aussage darüber treffen können, wieviel Pseudokanten in die Shapes unter der Voraussetzung der Realisierbarkeit gesetzt werden können.

Die Länge 40 wurde gewählt, da in dem Graphen $G(S_1,S_2,S_3)$ von Shapes ohne Pseudokanten die relative Häufigkeit von realisierbaren Strukturen ohne Pseudokanten mit einem Wert von 0,8 noch relativ hoch ist. Auf der anderen Seite liegt diese Häufigkeit nach dem Setzen von Pseudokanten deutlich unter dem Wert von 0,4. Somit haben wir auf der einen Seite genug Instanzen, in die noch Pseudokanten gesetzt werden können. Auf der anderen Seite ist das Setzen von Pseudokanten begrenzt und zwar nicht nur auf Grund der Anzahl von Knoten, sondern auch auf Grund des Realisierungsbegriffs, so dass die Verfahren *PsK_raus* und *Knoten_raus* eine durchaus sinnvolle Verwendung finden.

Um nicht ausschließlich von den bisherigen gefalteten Zufallssequenzen auszugehen, werden wir zusätzlich noch Aptamere betrachten und diese mit unseren Zufallssequenzen ver-

gleichen.

Schauen wir uns zunächst die Boxplots der durchschnittlichen Anzahl gesetzter Pseudokanten, welche mittels der drei Verfahren *PsK_rein*, *Psk_raus* und *Knoten_raus* in die jeweiligen Instanzen von Shapes aus Zufallssequenzen bzw. denen der Aptamere gesetzt wurden, in Abbildung 4.8 an.

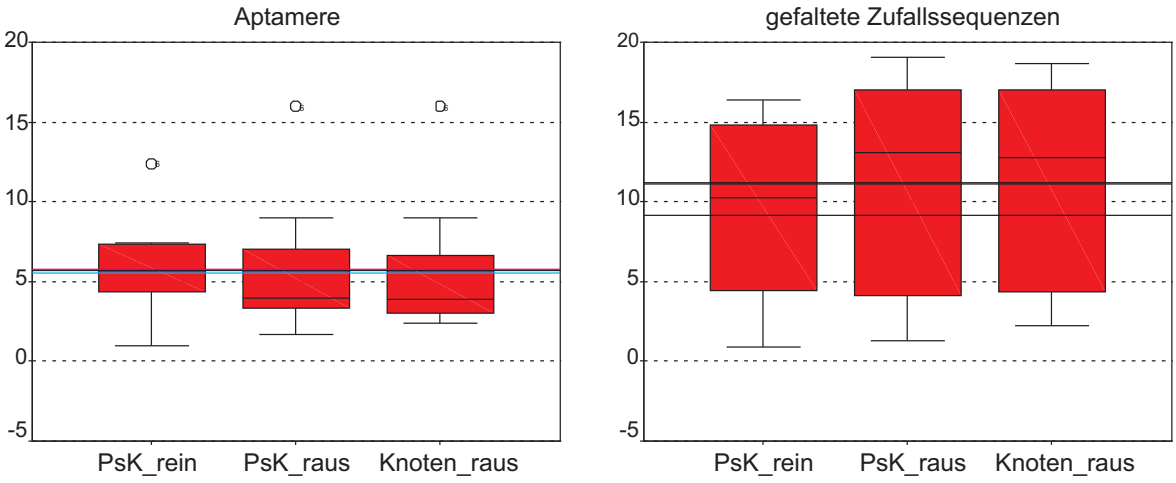

Abbildung 4.8: Boxplots der durchschnittlichen Anzahl gesetzter Pseudokanten mit den drei Verfahren *PsK_rein*, *PsK_raus* und *Knoten_raus*. Links sind die Boxplots der Aptamere und rechts die der Shapes aus Zufallsesequenzen zu sehen.

Man erkennt sofort einen deutlich Unterschied zwischen den Boxplots der Shapegraphen der Aptamere im Vergleich zu den Shapegraphen der gefalteten Zufallssequenzen. Man sieht, dass in die Shapegraphen der Aptamere bei unseren Instanzen durchschnittlich weniger Pseudokanten gesetzt wurden als in die Shapegraphen der Zufallssequenzen. Dies wird auch deutlich, wenn man sich die Mittelwerte bzw. Mediane der gesetzten Pseudokanten in Tabelle 4.2 anschaut. So sieht man, dass sich die Mittelwerte bzw. Mediane der gesetzten Pseudokanten in den den Shapegraphen aus Aptameren deutlich von denen der Zufallssequenzen unterscheiden. So liegen die durchschnittlichen Mittelwerte gesetzter Pseudokanten bei den Shapegraphen der Aptamarer bei 5,67 und denen der Zufallssequenzen bei 10,45.

Bei weiterer Betrachtung der Boxplots erkennt man, dass die Anzahl der Pseudokanten bei den Shapegraphen aus Aptameren deutlich dichter um den Mittelwert bzw. Median verteilt sind als bei den Graphen aus den Zufallssequenzen. Um diese Beobachtungen zu verstehen, sehen wir uns zusätzlich die Streudiagramme in Abbildung 4.9 an. Diese zeigen die durchschnittliche Anzahl der mit Hilfe der drei Verfahren gesetzten Pseudokanten in Abhängigkeit von der Anzahl regulärer Kanten.

Art	Verfahren	Median	arith. Mittelwert
Aptamer	PsK_rein	5,68	5,67
	PsK_raus	3,96	5,57
	Knoten_raus	3,91	5,76
ZUF	PsK_rein	10,25	9,13
	PsK_raus	13,14	11,13
	Knoten_raus	12,8	11,08

Tabelle 4.2: Mittelwerte und Mediane der Anzahl gesetzter Pseudokanten mittels der drei Verfahren *PsK_rein*, *PsK_raus* und *Knoten_raus*.

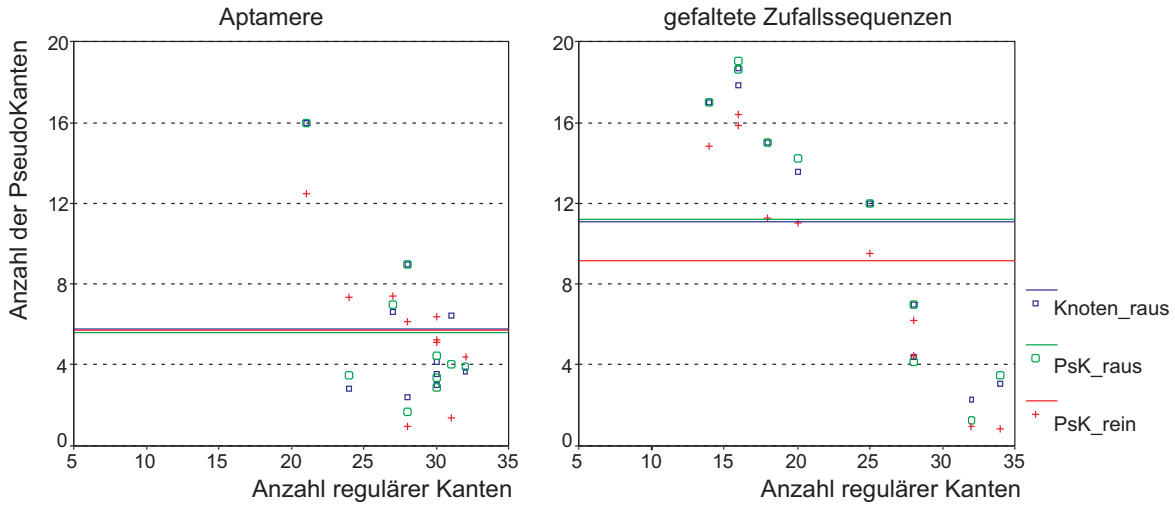

Abbildung 4.9: Darstellung der Anzahl von Pseudokanten in Abhängigkeit von der Anzahl regulärer Kanten. Links ist das Streudiagramm der Aptamere und rechts das der Shapes aus Zufallsesequenzen zu sehen.

In den Diagrammen ist zu erkennen, dass die Anzahl der regulären Kanten in den Shapegraphen gefalteter Zufallssequenzen deutlich weiter gestreut ist, als in denen aus Aptameren. So liegt die Anzahl der regulären Kanten in den Shapegraphen aus Aptameren zwischen 21 und 32 (Spannweite 11) und in den Shapegraphen zufällig erstellter Sequenzen zwischen 14 und 34 (Spannweite 20). Des Weiteren ist der Mittelwert der Anzahl von regulären Kanten bei den Shapegraphen der Aptamere mit 28,1 deutlich höher als in den Shapegraphen gefalteter Zufallssequenzen. Hier beträgt der Mittelwert 19,7.

In dem Streudiagramm der Shapegraphen aus Zufallssequenzen sieht man weiterhin, dass mit zunehmender Anzahl von regulären Kanten die Anzahl der Pseudokanten sinkt. Auch in dem Diagramm der Apamershapegraphen ist dieser Zusammenhang, trotz einiger Ausreisser, zu erkennen.

Bei der zusätzlichen Betrachtung der maximalen Anzahl gesetzter Pseudokanten in Tabelle 4.3 wird des Weiteren sichtbar:

1. Die maximale Anzahl von Pseudokanten ist sowohl bei den Aptameren als auch bei den Zufallsequenzen in den Shapegraphen mit geringerer Anzahl regulärer Kanten zu finden.

2. Die minimalen Maxima aller Instanzen sind bei den Graphen mit einer hohen Anzahl regulärer Kanten zu finden.

Art	Verfahren	PsK: grösstes Max	regK	PsK: kleinstes Max	regK
Aptamer	PsK_rein	16	21	5	28
	PsK_raus	16	21	6	30
	Knoten_raus	16	21	6	32
ZUF	PsK_rein	20	16	3	34
	PsK_raus	20	16	4	32
	Knoten_raus	20	16	5	34

Tabelle 4.3: In der Tabelle ist die grösste Anzahl der gesetzten Pseudokanten, als auch die kleinsten, durch die drei Verfahren ermittelten, Maxima zu entnehmen. Es ist zu erkennen, dass sowohl bei den Aptameren, als auch bei den Zufallsequenzen eine hohe Anzahl von Pseudokanten in Shapegraphen mit einer geringen Anzahl regulärer Kanten gesetzt wurden und umgekehrt

Diese Erkenntnisse erklären nicht nur die hohe bzw. geringe Streuung der Werte in den Graphen gefalteter Zufallssequenzen bzw. der Aptamere in den Boxplots der Abbildung 4.8,

sondern bestätigen auch die bisher gewonnenen Erkenntnisse über die Zusammenhänge zwischen der Anzahl der regulären Kanten, sowie der Realisierbarkeit und der Anzahl von Pseudokanten. Es ist somit anzunehmen, dass mit einer grösseren Anzahl von regulären Kanten die Wahrscheinlichkeit steigt, Zyklen mit einer ungeraden Anzahl regulärer Kanten bzw. Überlagerungen, nach dem Setzen von Pseudokanten, auszubilden.

Es ist somit deutlich geworden, dass der Unterschied der Mittelwerte und Mediane gesetzter Pseudokanten, sowie die ersichtlichen Unterschiede in den Boxplots, sehr stark durch die Anzahl regulärer Kanten zwischen den Shapes der Aptamere und denen der Zufallssequenzen bestimmt ist, welche in den Apatamere unserer Instanzen deutlich höher lag, als in denen gefalteter Zufallssequenzen. Leider können wir auf Grund zu weniger Daten, keine quantitative Aussage darüber treffen ob in den Shapegraphen aus Aptameren im allgemeinen die Anzahl regulärer Kanten im Vergleich zu den Shapegraphen aus Zufallssequenzen höher ist oder ob dies auf die zufällige Wahl der Sequenzen zurückzuführen ist.

Zuletzt sei erwähnt: Um eine genauere Aussage darüber treffen zu können, wieviel Pseudokanten maximal in entsprechende Shapes gesetzt werden können, sollten nicht nur Shapes gleicher Länge, sondern auch Shapes mit der zusätzlichen Eigenschaft, dass die Anzahl regulärer Kanten übereinstimmt, getestet werden. Dies würde jedoch den Rahmen dieser Arbeit sprengen.

4.3 Zusammenfassung

Fassen wir die Ergebnisse nocheinmal kurz zusammen. Zuerst haben wir untersucht, ob es einen Zusammenhang zwischen der Anzahl von Knoten und der Anzahl regulärer Kanten in einzelnen Shapes, als auch in den Shapegraphen dreier bzw. vierer Shapes gibt. Wir haben einen deutlichen Zusammenhang feststellen können. So nahm in allen Fällen mit zunehmender Anzahl von Knoten auch die Anzahl regulärer Kanten zu. Insbesondere konnten wir sehen, dass es sehr selten zu der Bildung von Mehrfachkanten kommt und sich somit die Anzahl regulärer Kanten in den Shapegraphen dreier bzw. vierer Shapes durch die Anzahl der regulären Kanten in einem Shape approximieren lässt.

Wir haben daraufhin überprüft, ob es einen Zusammenhang zwischen der Anzahl von Knoten und der Eigenschaft der Realisierbarkeit gibt. Wir haben hier feststellen können, dass mit zunehmender Anzahl von Knoten und demzufolge mit steigender Anzahl regulärer Kanten die Anzahl realisierbarer Graphen abnimmt. Dies lies sich darauf zurückführen, dass mit zunehmender Anzahl von regulären Kanten die Wahrscheinlichkeit steigt, Zyklen mit einer ungeraden Anzahl regulärer Kanten auszubilden, insbesondere unter der Erkenntnis,

dass es selten zu einer Bildung von Mehrfachkanten kommt.

Danach untersuchten wir den Zusammenhang zwichen der Anzahl von Pseudokanten bzw. die Eigenschaft der Realisiserbarkeit erweiterter Shapes und der Anzahl von Knoten. Es ließ sich zunächst zeigen, dass mit zunehmender Anzahl von Knoten die Anzahl von gesetzten Pseudokanten steigt und gleichzeitig die Anzahl realisierbarer Shapegraphen erweiterter Shapes deutlich schneller abnahm, als in denen ohne Pseudokanten. Offensichtlich stieg somit die Wahrscheinlichkeit nach dem Setzen von Pseudokanten kritsiche Zyklen bzw. Überlagerungen zu bilden. Bei der Untersuchung der Abhängigkeit der Anzahl von Knoten und der Anzahl von Pseudokanten, so dass die Eigenschaft der Realisierbarkeit erweiterter Shapes erhalten bleibt, konnten wir feststellen, dass die Anzahl gesetzter Pseudokanten mit zunehmender Knotenanzahl abnimmt.

Zuletzt untersuchten wir die Shapegraphen dreier Shapes von gefalteten Zufallssequenzen, sowie die von Aptameren.Wir haben hierbei einen deutlichen Unterschied zwischen gefalteten Zufallssequenzen und Aptameren feststellen können. Eine quantitative Aussage, ob dieser Unterschied im allgemeinen gilt und wieviel Pseudokanten maximal in diese Shapes gesetzt werden können, konnten wir auf Grund zu weniger getester Instanzen nicht treffen. Letztendlich bestätigten aber die gewonnenen Erkenntnisse aus diesem Vergleich die bisherigen Ergebnisse.

5 Schlussbetrachtung

5.1 Zusammenfassung

Im ersten Teil dieser Arbeit wurde die Struktur und der Aufbau der RNA erklärt. Es wurde die Frage motiviert, unter welchen Voraussetzungen zu mehreren gegebenen Sekundärstrukturen eine einzelne Sequenz existiert, welche zu diesen kompatibel ist. Um diesen Sachverhalt beantworten zu können, definierten wir verschiedene Begriffe der Graphentheorie und führten den Begriff der *Realisierung*. Dabei wurde die Möglichkeit der Realisierung ganz eng mit den Eigenschaften des entsprechenden Shapegraphen verknüpft. Allerdings wurden in dem bisherigen Realisierungsbegriff nur gepaarte Positionen berücksichtigt.

Wir erweiterten die Shapes daraufhin mittels *Pseudokanten*, um den Begriff der Realisierung einzuschränken und somit Sequenzen zu finden, in denen auch Positionen beachtet werden, die keine Paarungen eingehen dürfen. Dabei definierten wir zunächst den Realisierungsbegriff für eine binäre Zeichenkette $s = s_1 \dots s_n \in \{0,1\}^n$. Hierbei war die Forderung an reguläre Kanten $\{v_i, v_j\}$, dass $s_i \neq s_j$ und für Pseudokanten $\{v_i, v_j\}$, dass $s_i = s_j$. Wir haben festgestellt, dass gegebene erweiterte Shapes dann und nur dann durch eine einzelne binäre Zeichenkette realisiert werden können, wenn in dem entsprechenden Shapegraphen keine Überlagerungen, d.h. Mehrfachkanten aus regulären Kanten und Pseudokanten, sowie keine Zyklen mir einer ungeraden Anzahl regulärer Kanten existieren. Danach definierten wir einen entsprechenden Realisierungsbegriff für Zeichenketten $s \in \{A,C,G,U\}^n$ und fanden heraus, dass dieser spezielle Realisierungsbegriff äquivalent zu dem Begriff der Realisierung durch eine binäre Zeichenkette ist.

Durch diese Sachverhalte motiviert, wurden die Probleme MinKA_S, MinKN_S, MinKA_eS und MinKN_eS der minimalen Kanten bzw. Knotenmenge, die aus einem Shapegraphen entfernt werden müssen, so dass die entsprechenden Shapes bzw. erweiterten Shapes durch eine einzelne Zeichenkette realisiert werden können, formuliert und auf ihre Komplexität hin untersucht. Wir haben bewiesen, dass neben dem als NP-vollständigen bekannten Problem MinKN_S, auch das Probleme MinKA_S, sowie die Probleme für erweiterte Shapes MinKA_eS und MinKN_eS NP-vollständig sind.

Im letzten Teil dieser Arbeit führten wir empirische Studien zu Betrachtungen von Graphen mit und ohne Pseudokanten durch. Wir stellten hierbei u.a. fest, dass es einen starken Zusammenhang zwischen der Anzahl von Knoten und der Anzahl von regulären Kanten, sowie Pseudokanten gibt, welcher sich wiederum deutlich auf den Begriff der Realisierbarkeit auswirkt. Die zuletzt untersuchten Unterschiede zwischen gefalteten Zufallssequenzen und Aptameren bestätigten die vorangegangenen Erkenntnisse.

5.2 Ausblick

Im Fokus zukünftiger Betrachtungen sollte die Untersuchung anderer Realisierungsbegriffe an Pseudokanten stehen. Hierbei ist es möglich neue Forderungen an Pseudokanten $\{v_i, v_j\}$ zu stellen, so dass $s_i s_j \in \mathscr{B}'_{PsK} \subseteq \{A,C,G,U\}^2 \setminus \{AU, UA, CG, GC, GU, UG\}$. Daraus resultierend müssen vermutlich neue Voraussetzungen an Shapegraphen geknüpft werden, so dass die entsprechenden, erweiterten Shapes durch eine einzelne Zeichenkette mittels des neuen Begriffs realisiert werden können.

Zudem erscheint es interessant, Pseudokanten nicht nur an Positionen zu setzen, so dass die Eigenschaften eines Shapes erhalten bleiben, sondern auch hier ein anderes Modell zu entwickeln, welches erlaubt Positionen mit Pseudokanten zu besetzen, die nicht an die Eigenschaften eines Shapes gebunden sind.

Des Weiteren wissen wir, dass die Probleme MinKN_S, MinKN_eS für $k \geq 4$ und MinKA_eS für $k \geq 5$ NP-vollständig sind. Offen ist die Betrachtung der NP-Vollständigkeit für $k = 3$ bzw. $k = 4$.

Zuletzt scheint es auch lohnenswert, bessere Heuristiken bei der empirischen Betrachtung zu nutzen, um genauere Aussagen darüber zu treffen, wie viele Pseudokanten maximal in die entsprechenden Shapes gesetzt werden können, so dass die Eigenschaft der Realisierbarkeit erhalten bleibt.

A Verzeichnisse

Abbildungsverzeichnis

Tabellenverzeichnis

Literaturverzeichnis

[AMS+99] R.G. Amado, R.T. Mitsuyasu, G. Symonds, J.D. Rosenblatt, J. Zack, L.Q. Sun, M. Miller, J. Ely, W. Gerlach. A phase I trial of autologous CD34+ hematopoietic progenitor cells transduced with an anti-HIV ribozyme. Hum Gene Ther 10, 2255-2270. 1999

[Bar82] F. Barahona. On the Computational Complexity of Ising Spin Glass Models. J. Phys. A.: Math Gen., 15, 3241-3253. 1982

[BK02] J.G. Bruno, J.L. Kiel. Use of magnetic beads in selection and detection of biotoxin aptamers by electrochemiluminescence and enzymatic methods. Biotechniques 32, 178-180, 182-173. 2002

[BKF95] P. Burgstaller, M. Kochoyan, M. Famulok. Structural probing and damage selection of citrulline- and arginine-specific RNA aptamers identify base positions required for binding. Nucl Acids Res 23, 4769-4776. 1995

[BSR97] T. Baumstark, A.R. Schroder, D. Riesner. Viroid processing: Switch from cleavage to ligation is driven by a change from a tetraloop to a loop E conformation. EMBO J., 16, 599-610. 1997

[BVK+97] G. Bauer, P. Valdez, K. Kearns, I. Bahner, S.F. Wen, J.A. Zaia, D.B. Kohn. Inhibition of human immunodeficiency virus-1 (HIV-1) replication after transduction of granulocyte colony-stimulating factor-mobilized CD34+ cells from HIV-1-infected donors using retroviral vectors containing anti-HIV-1 genes. Blood 89, 2259-2267. 1997

[BY93] P. Babitzke, C. Yanofsky. Reconstitution of Bacillus subtilis Trp attenuation in vitro with TRAP, the Trp RNA-binding attenuation protein. Proc. Natl. Acad. Sci. USA, 90, 133-137. 1993

[Cec86] T. Cech. RNA as an enzyme. Scientific American, 11, 76-84. 1986

[CLK+05] P. Clote, G. Leszek, R. Kolpakov, E. Kranakis, D. Krizanc. On realizing shapes in the theory of RNA neutral networks. Journal of Theoretical Biology, 236, 216–227. 2005

[CLR+01] T.H. Cormen, C.E. Leiserson, R.L. Rivest, C. Stein. Introduction to Algorithms. second ed. MIT Press, Cambridge, MA. 2001

[CSX+02] L.R. Comolli, I. Smirnov, L. Xu, E.H. Blackburn,T.L. James. A molecular switch underlies a human telomerase disease. Proc. Natl Acad. Sci. USA 99, 26, 16998–17003. 2002

[Die01] R. Diestel. Graphentheorie. Elektronische Ausgabe 2000. Springer-Verlag Heidelberg. 2000

[DS02] J.H. Davis, J.W. Szostak. Isolation of high-affinity GTP aptamers from partially structured RNA libraries. PNAS, 99, 11616-11621. 2002

[ES90] A.D. Ellington, J.W. Szostak. In vitro selection of RNA molecules that bind specific ligands. Nature 346, 818-822. 1990

[FHM+01] C. Flamm, I.L. Hofacker, S. Maurer-Stroh, P.F. Stadler, M. Zehl. Design of multistable RNA Molecules. 2001

[FMS+83] G. Fayat, F.J. Mayaux, C. Sacerdot, M. Fromant, M. Springer, M. Grunberg-Manago, S. Blanquet. Escherichia coli phenylalanyl-tRNA synthetase operon region. Evidence for an attenuation mechanism. Identifcation of the gene for the ribosomal protein L20. J. Mol. Biol., 171, 239-261. 1983

[Fon02] W. Fontana. Modelling 'evo-devo' with RNA. BioEssays, 24, 1164-1177. 2002

[GBE+96] A. Geiger, P. Burgstaller, H. von der Eltz, A. Roeder, M. Famulok. RNA aptamers that bind L-arginine with sub-micromolar dissociation constants and high enantioselectivity. Nucl Acids Res 24, 1029-1036. 1996

[GBP98] A.P. Gultyaev, F.H. Batenburg, C.W. Pleij. Dynamic competition between alternative structures in viroid RNAs simulated by an RNA folding algorithm. J. Mol. Biol., 276, 43-55. 1998

[GGS+96] W. Grüner, R. Giegerich, D. Strothmann, C. Reidys, J. Weber, I.L. Hofacker, P.F. Stadler, P. Schuster. Analysis of RNA Sequence Structure Maps by Exhaustive Enumeration. Monatshefte für Chemie 1996, 127, 355-374. 1996

[Gil86] W. Gilbert. Origin of life: The RNA world. Nature, 319, 618. 1986

[GJ79] M.R. Garey, D.S. Johnson. Computers and Intractability: A Guide to the Theory of NP-Completeness. W.H. Freeman and Company, New York. 1979

[GZ97] K. Gowda, C. Zwieb. Determinants of a protein-induced rna switch in the large domain of signal recognition particle identified by systematic-site directed mutagenesis. Nucleic Acids Res. 25, 14, 2835–2840. 1997

[Har54] F. Harary. On the Notion of Balance of a Signed Graph. Michigan Mathematical Journal, 2, 143-146. 1954

[Har59] F. Harary. On the Measurement of Structural Balance. Behavioral Sci., 4, 316-323. 1959

[HC93] K.A. Harris, D.M. Crothers. The Leptomonas collosoma spliced leader RNA can switch between two alternate structural forms. Biochemistry 32, 20, 5301–5311. 1993

[HCU95] K.A. Harris, D.M. Crothers, E. Ullu. In vivo structural analysis of spliced leader RNAs in Trypanosoma brucei and Leptomonas collosoma: a flexible structure that is independent of cap4 methylations. RNA 1, 4, 351–362. 1995

[HFS+94] I.L. Hofacker, W. Fontana, P.F. Stadler, S. Bonhoeffer, M. Tacker, P. Schuster. Fast folding and comparison of RNA secondary structures (the Vienna RNA Package). Monatsh. Chem., 125(2):167–188. 1994

[HK80] F. Harary, J.A. Kabell. A Simple Algorithm to Detect Balance in Signed Graphs. Mathematical Social Science, 1, 131-136. 1980

[Hof03] Ivo L. Hofacker. RNA secondary structure analysis using the Vienna RNA Package. John Wiley and Sons. In A.D. Baxevanis and D.B. Davison, Current Protocols in Bioinformatics, volume 1. 2003.

[HP00] T. Hermann, D.J. Patel. Adaptive Recognition by Nucleic Acid Aptamers. Science, 287, 820-825. 2000

[HWS+88] R. Hecker, Z.M. Wang, G. Steger, D. Riesner. Analysis of RNA structures by temperarure-gradient gel electrophoresis: viroid replication and processing. Gene, 72, 59-74. 1988

[IK00] W. Imrich, S. Klavzar. Product Graphs: Structure and Recognition. Inc. John Wiley and Sons. 2000

[Joy91] G.F. Joyce. The rise and fall of the RNA world. The New Biologist, 3, 399-407. 1991

[JW05] T. Janssen, W. Laatz: Statistische Datenanalyse mit SPSS für Windows. Springer-Verlag Berlin. 2005

[Lou03] E. Loukakis. A Dynamic Programming Algorithm to Test an signed Graph for Balance. Intern. J. Computer Math., 80(4), 499-507. 2003

[LSS+91] P. Loss, M. Schmitz, G. Steger, D. Riesner. Formation of a thermodynamically metastable structure containing hairpin II is critical for infectivity of potato spindle tuber viroid RNA. EMBO J., 10, 719-727. 1991.

[KBR99] D.B. Kohn, G. Bauer, C.R. Rice. A clinical trial of retroviral-mediated transfer of a rev-responsive element decoy gene into CD34(+) cells from the bone marrow of human immunodeficiency virus-1-infected children. Blood 94, 368-371. 1999

[Mat05] J.S. Mattick. The Functional Genomics of Noncoding RNA: Science 309 (5740), 1527-1528. 2005

[MBB+03] M. Mandal, B. Boese, J.E. Barrick, W.C. Winkler, R.R. Breaker. Riboswitches control fundamental biochemical pathways in Bacillus subtilis and other bacteria. Cell 113 (5), 577–586. 2003

[Moo05] M.J. Moore. From Birth to Death: The Complex Lives of Eukaryotic mRNAs. Science 309 (5740), 1514-1518. 2005

[Mül06] M. Müller. Molekulare Analyse eines synthetischen Tetracyclin-abhängigen RNA-Regulators. Dissertation. Friedrich-Alexander-Universität Erlangen-Nürnberg. 2006

[Nol05] H.F. Noller. RNA Structure: Reading the Ribosome. Science 309 (5740), 1508-1514. 2005

[PB98] A.T. Perrotta, M.D. Been. A toggle duplex in hepatitis delta virus self-cleaving RNA that stabilizes an inactive and a salt-dependent pro-active ribozyme conformation. J. Mol. Biol., 279, 361-373. 1998.

[PGG92] H. Putzer, N. Gendron, M. Grunberg-Manago. Co-ordinate expression of the two threonyl-tRNA synthetase genes in Bacillus subtilis: Control by transcriptional antitermination involving a conserved regulatory sequence. EMBO J., 11, 3117-3127. 1992

[Rid05] G. Riddihough. In the Forests of RNA Dark Matter. Science 309 (5740), 1507. 2005

[RSS97a] C. Reidys, P.F. Stadler, P. Schuster. Generic properties of combinatory maps: neutral networks of RNA secondary structures. Bull. Biol. 59, 2, 339–397. 1997a

[RSS97b] C. Reidys, P.F. Stadler, P. Schuster. Generic properties of combinatory maps: neutral networks of RNA secondary structures. Bull. Biol. 59, 2, 339–397. 1997b

[Sch99] P. Schuster. Beherrschung von Komplexität in der molekularen Evolution. 117-145. 1999

[SGM+89] D.T. Suzuki, A.J.F. Griffiths, J.H. Miller, R.C. Lewontin. An Introduction to Genetic Analysis. W.H. Freeman and Company. 1989

[Spi71] S. Spiegelman. An approach to the experimental analysis of precellular evolution. Quarterly Reviews of Biophysics, 4, 213-253. 1971

[SSK+01] H. Schurer, K. Stembera, D. Knoll, G. Mayer, M. Blind, H.H. Forster, M. Famulok, P. Welzel, U. Hahn. Aptamers that bind to the antibiotic moenomycin A. Bioorganic and Medicinal Chemistry 9, 2557. 2001

[TG90] C. Tuerk, L. Gold. Systematic evolution of ligands by exponential enrichment: RNA ligands to bacteriophage T4 DNA polymerase. Science 249, 505-510. 1990

[Wae57] B.L. van der Waerden. Mathematische Statistik. Springer Verlag. 1957

[Wat78] M.S. Waterman. Secondary Structure of Single-Stranded Nucleic Acids. Studies in Foundations and Combinatorics, Advances in mathematics supplementary studies, Academic Press N.Y., 1, 167-212. 1978

[WS98] S.T. Wallace, R. Schroeder. In vitro selection and characterization of streptomycin-binding RNAs: recognition discrimination between antibiotics. RNA 4, 112-123. 1998

[Yan81] M. Yannakakis. Edge Deletion Problems. SIAM Journal on Computing, 10, 297-309. 1981

[ZH05] P.D. Zamore, B. Haley. Ribo-gnome: The Big World of Small RNAs. Science 309 (5740), 1519-1524. 2005

B Anhang

DANKE...!

Allen voran möchte ich Prof. Dr. Martin Middendorf und Dr. Daniel Merkle danken, die immer wieder offen für Fragen und spontane Diskussionen waren und die mir den nötigen Freiraum gaben, dieses Thema zu bearbeiten. Und nochmals vielen Dank für die Möglichkeit nach Tschechien zu fahren - ein Prost auf die Zukunft!
Danke auch an meine Familie, ohne deren finanzieller Unterstützung dieses Studium nicht möglich gewesen wäre.
Vielen Dank an Sabrina, die tolle Frau an meiner Seite, für ihre Zeit, Geduld und Liebe, sowie für die Abbildungen 1.2 und 1.3.
Und Danke natürlich an alle Freunde, die mir in dieser Zeit durch viele Biere und Gespräche, Skatabende, Spaziergänge, Partys und Tanztees, Koch- und Filmabende, im Garten Kastanien essen und so viel mehr geholfen haben, meinen Kopf wieder frei zu bekommen und Gehirnknoten zu lösen.

Erklärung

Ich versichere, dass ich die vorliegende Arbeit selbstständig und nur unter Verwendung der angegebenen Quellen und Hilfsmittel angefertigt habe.

Marc Hellmuth Leipzig, 14. November 2006