Formalsprachliche Theorie der Haarnadelstrukturen

Steffen Kopecki

Department of Computer Science
The University of Western Ontraio, London, Canada
steffen@csd.uwo.ca

Abstract: Die (berenzte) Haarnadel-Vervollständigung und die Haarnadel-Verlängerung sind Operationen auf formalen Sprachen, welche die Modifikation von DNA Strängen durch Bildung von Haarnadelstrukturen während der Polymerase-Kettenreaktion modellieren. In dieser Arbeit befassen wir uns mit der formalsprachlichen Analyse dieser Operationen. Neben der Untersuchung der Abschlusseigenschaften von Sprachklassen unter den Operationen, beschäftigt sich die Arbeit mit der Lösung von Entscheidungsproblemen, die durch Haarnadel-Operationen gegeben sind.

1 Einführung

Die Haarnadelstruktur ist eine intramolekulare Basenpaarung, die in einsträngiger DNA oder RNA auftreten kann. Inspiriert durch dieses Phänomen wurden Operationen, wie die Haarnadel-Vervollständigung und Haarnadel-Verlängerung, auf formalen Sprachen definiert. Diese Arbeit beschäftigt sich ausschließlich mit der formalsprachlichen Untersuchung der Haarnadelstrukturen. Zunächst werden wir jedoch den biochemischen Ursprung der Haarnadelstruktur darlegen.

1.1 Haarnadelstrukturen in der Biochemie

Einsträngige DNA, im folgenden *DNA Strang* genannt, ist ein Polymer, bestehend aus Nukleotiden, welche sich durch ihre Nukleobasen A (Adenin), T (Thymin), G (Guanin) und C (Cytosin) unterscheiden. Jeder DNA Strang besitzt ein $5'$- und ein $3'$-Ende, welche nach der chemischen Struktur der Nukleotide benannt sind. Üblicherweise werden DNA Stränge in $5'$-nach-$3'$-Orientierung notiert. Ein DNA Strang kann abstrakt als ein Wort über dem vier-Buchstaben Alphabet $\{\mathsf{A}, \mathsf{C}, \mathsf{G}, \mathsf{T}\}$ gesehen werden. Die Basen A und T, bzw. C und G, sind zueinander *Watson-Crick-komplementär*. Zwei DNA Stränge mit unterschiedlicher Orientierung können sich aneinander anlagern, wenn ihre Basen paarweise komplementär sind und können so einen *DNA Doppelstrang* bilden, die wohlbekannte Doppelhelix. In Abb. 1 ist ein Beispiel gegeben.

Für das Watson-Crick-Komplement und sein formalsprachliches Pendant verwenden wir die $\overline{\ }$ Notation, d. h. $\overline{\mathsf{A}} = \mathsf{T}$, $\overline{\mathsf{T}} = \mathsf{A}$, $\overline{\mathsf{C}} = \mathsf{G}$ und $\overline{\mathsf{G}} = \mathsf{C}$. Diese Notation erweitern wir

$$\begin{array}{l} 5'\text{—}\mathsf{C}\text{—}\mathsf{G}\text{—}\mathsf{G}\text{—}\mathsf{T}\text{—}\mathsf{T}\text{—}\mathsf{T}\text{—}\mathsf{C}\text{—}\mathsf{A}\text{—}\mathsf{G}\text{—}\mathsf{G}\text{—}\mathsf{A}\text{—}3' \\ 3'\text{—}\mathsf{G}\text{—}\mathsf{C}\text{—}\mathsf{C}\text{—}\mathsf{A}\text{—}\mathsf{A}\text{—}\mathsf{A}\text{—}\mathsf{G}\text{—}\mathsf{T}\text{—}\mathsf{C}\text{—}\mathsf{C}\text{—}\mathsf{T}\text{—}5' \end{array}$$

Abbildung 1: Anlagerung zweier komplementärer DNA Stränge

auf DNA Stränge in $5'$-nach-$3'$-Orientierung (bzw. Wörter) durch $\overline{a_1 \cdots a_n} = \overline{a_n} \cdots \overline{a_1}$, wobei $a_1, \ldots, a_n$ einzelne Basen (bzw. Buchstaben) sind. Hierdurch wird die chemische Eigenschaft abgebildet, dass sich der DNA Strang $a_1 \cdots a_n$ an den komplementären DNA Strang $\overline{a_1 \cdots a_n}$ anlagern kann. Man beachte hierbei, dass

$$\overline{5'\text{-}a_1 \cdots a_n\text{-}3'} = 3'\text{-}\overline{a_1} \cdots \overline{a_n}\text{-}5' = 5'\text{-}\overline{a_n} \cdots \overline{a_1}\text{-}3'.$$

Eine Technik, die häufig verwendet wird um DNA Stränge mit bestimmten Eigenschaften und ihre Komplemente exponentiell zu vervielfältigen, ist die *Polymerase-Kettenreaktion* (engl. polymerase chain reaction, PCR). Die PCR wiederholt drei biochemische Prozesse mehrmals hintereinander, siehe Abb. 2. Angenommen ein langer DNA-Strang τ, das *Template*, soll vervielfältigt werden. Hierzu werden kurze DNA Stränge, sogenannte *Primer*, welche komplementär zu einem Suffix des Templates sind, zur Lösung hinzugefügt. (Ein Suffix ist eine Basen-Sequenz, die dem $3'$-Ende vorangeht.) Falls $\tau = \gamma\alpha$, wobei α verhältnismäßig kurz ist, dann ist $\overline{\alpha}$ ein geeigneter Primer. Während der *Hybridisierungsphase* lagert sich einer der Primer an das Template an. Durch freie Nukleobasen wird Base nach Base des Templates komplementiert, beginnend am $3'$-Ende des Primers. Dieser Prozess wird *Elongation* genannt. Nachdem das Template vollständig komplementiert wurde, wird der neu entstandene Doppelstrang *denaturiert* (aufgetrennt) und wir erhalten die Einzelstränge τ und $\overline{\tau}$. Wiederholt man diese drei Schritte und fügt Primer hinzu, welche zu einem Suffix von $\overline{\tau}$ komplementär sind (d. h. ein Präfix von τ sind), verdoppelt sich die Anzahl von Templates und ihrer Komplemente nach jedem Zyklus.

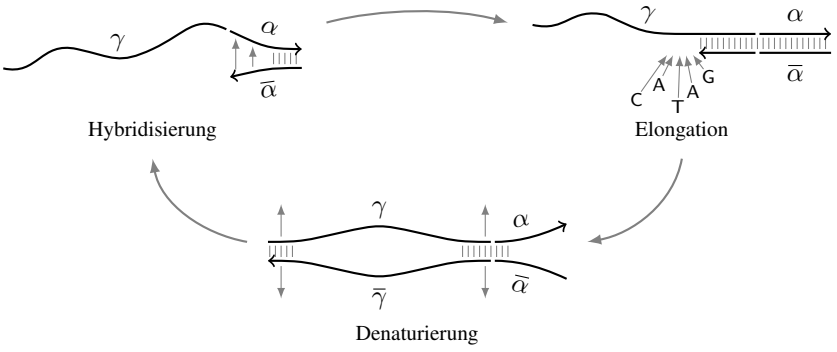

Abbildung 2: Polymerase-Kettenreaktion

Die Haarnadel-Vervollständigung ist die Modifikation eines DNA Strangs, die während der PCR entstehen kann, siehe Abb. 3. Falls ein DNA Strang die Form $\gamma\alpha\beta\overline{\alpha}$ besitzt, so kann der Suffix $\overline{\alpha}$ als Primer auf den DNA-Strang selbst wirken und sich während der Hybridisierung an die DNA Sequenz α anlagern. Eine solche intramolekulare Basenpaarung

wird *Haarnadelstruktur*, oder einfach *Haarnadel*, genannt. Durch Elongation wird der zuvor ungebundene Teil γ des DNA Strangs komplementiert und nach der Denaturierung erhalten wir einen neuen DNA Strang $\gamma\alpha\beta\overline{\alpha}\overline{\gamma}$, welcher als *Haarnadel-Vervollständigung* des DNA Strangs $\gamma\alpha\beta\overline{\alpha}$ bezeichnet wird.

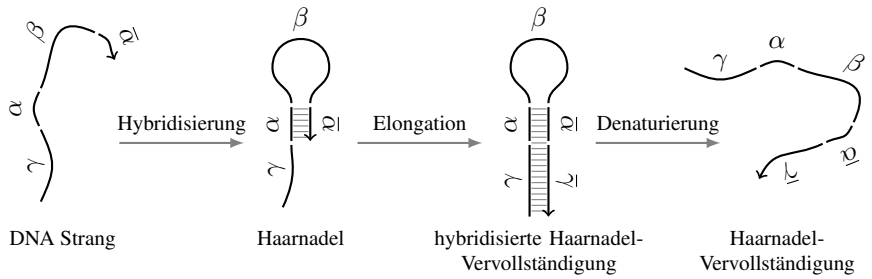

Abbildung 3: Haarnadel-Vervollständigung eines DNA Strangs

Häufig werden Haarnadel-Vervollständigungen als störendes Nebenprodukt von DNA-gestützten Berechnungen gesehen, weshalb in vielen Arbeiten daran geforscht wurde, DNA Bibliotheken (Mengen von DNA Strängen) zu entwerfen, welche keine Haarnadeln oder andere ungewollte Strukturen ausbilden, siehe u. a. [DMG+98, KKL+05, KMT07]. Andererseits wurden auch DNA Algorithmen entworfen, welche Haarnadeln oder Haarnadel-Vervollständigungen nutzen um Berechnungsschritte durchzuführen. Ein Beispiel hierfür ist die *Whiplash-PCR*. Bei dieser Form der PCR werden Haarnadel-Vervollständigungen, welche durch Stopper-Sequenzen kontrolliert sind, verwendet, um zufällige Pfade auf gerichteten Graphen abzulaufen. Da sehr viele dieser zufälligen Pfade parallel berechnet werden können, ist es möglich mithilfe der Whiplash-PCR NP-vollständige Probleme wie das HAMILTON PFAD PROBLEM zu lösen [HAK+97, Win98].

1.2 Haarnadelstrukturen in Formalen Sprachen

In der restlichen Arbeit betrachten wir Haarnadel-Vervollständigungen als Operation auf formalen Sprachen. Der Leser, der nicht mit den Grundlagen der formalen Sprachen vertraut ist, sei auf [HU79] verwiesen.

Wir betrachten Wörter und Sprachen über einem fest gewählten Alphabet Σ. Die Menge aller Wörter wird mit Σ^* bezeichnet und das leere Wort mit 1. Das Alphabet Σ sei mit einer Involution $\overline{\ }$ ausgestattet, d. h. $\overline{\overline{a}} = a$ für alle $a \in \Sigma$. Wie zuvor erweitern wir diese Notation auf Wörter $\overline{a_1 \cdots a_n} = \overline{a_n} \cdots \overline{a_1}$, wobei $a_1, \ldots, a_n \in \Sigma$. Für $w \in \Sigma^*$, bezeichnet $|w|$ die Länge des Wortes. Lässt sich $w = xyz$ schreiben, mit $x, y, z \in \Sigma^*$, so werden x, y, z als *Präfix*, *Faktor*, bzw. *Suffix* von w bezeichnet.

Sei $w = \gamma\alpha\beta\overline{\alpha}$ ein Wort, so wird $\gamma\alpha\beta\overline{\alpha}\overline{\gamma}$ eine *(rechte) Haarnadel-Vervollständigung* von w genannt. Da Haarnadeln in der Biochemie nur stabil sind, wenn die Bindung zwischen α und $\overline{\alpha}$ stark genug ist, definieren wir eine Konstante k und fordern $|\alpha| = k$. (Wohlgemerkt änderte sich die Definition nicht, würden wir $|\alpha| \geq k$ fordern.) Sei $w = \alpha\beta\overline{\alpha}\overline{\gamma}$, mit

$|\alpha| = k$, so bezeichnen wir $\gamma\alpha\beta\overline{\alpha}\overline{\gamma}$ als eine *(linke) Haarnadel-Vervollständigung* von w.

Seien nun L_1 und L_2 Sprachen, so ist

$$\mathcal{H}_k(L_1, L_2) = \{\gamma\alpha\beta\overline{\alpha}\overline{\gamma} \mid |\alpha| = k \wedge (\gamma\alpha\beta\overline{\alpha} \in L_1 \vee \alpha\beta\overline{\alpha}\overline{\gamma} \in L_2)\}$$

die *Haarnadel-Vervollständigung* von L_1 und L_2, d. h. wir vereinigen alle rechten Haarnadel-Vervollständigungen von Wörtern aus L_1 mit den linken Haarnadel-Vervollständigungen von Wörtern aus L_2. Häufig wird in der Literatur zwischen der rechten, linken und beidseitigen Haarnadel-Vervollständigung unterschieden. Wir erhalten die rechte Haarnadel-Vervollständigung, falls $L_2 = \emptyset$, die linke Haarnadel-Vervollständigung, falls $L_1 = \emptyset$ und die beidseitige Haarnadel-Vervollständigung, falls $L_1 = L_2$. Unsere Definition erlaubt es alle drei Varianten gleichzeitig zu untersuchen.

Da die PCR in der Biochemie selten nach einem Zyklus gestoppt wird, stellt sich die Frage, wie sich DNA Stränge, die durch Haarnadel-Vervollständigungen entstanden sind, weiter entwickeln. Dies inspiriert die Untersuchung der *iterierten Haarnadel-Vervollständigung*. Im iterierten Fall, ergibt es keinen Sinn mit zwei Sprachen zu starten. Es sei L eine formale Sprache. Die iterierte Haarnadel-Vervollständigung von L ist der reflexive und transitive Abschluss der beidseitigen Haarnadel-Vervollständigung

$$\mathcal{H}_k^*(L) = \bigcup_{i \geq 0} \mathcal{H}_k^i(L),$$

wobei

$$\mathcal{H}_k(L) = \mathcal{H}_k(L, L), \qquad \mathcal{H}_k^0(L) = L, \qquad \mathcal{H}_k^{i+1}(L) = \mathcal{H}_k(\mathcal{H}_k^i(L)) \quad \text{für } i \geq 0.$$

Die Haarnadel-Vervollständigung wurde erstmals in einer Arbeit von Cheptea, Martín-Vide und Mitrana 2006 definiert [CMVM06]. In den darauffolgenden Jahren wurde die Haarnadel-Vervollständigung aus formalsprachlichen und algorithmischen Gesichtspunkten untersucht, siehe u. a. [CMVM06, MMY08, MMY09, MMVM09]. In dieser Arbeit betrachten wir neben der Haarnadel-Vervollständigung zwei weitere, verwandte Operationen, die begrenzte Haarnadel-Vervollständigung und die Haarnadel-Verlängerung. Bei der begrenzten Variante ist die Länge des γ-Faktors durch eine Konstante m begrenzt. Diese Variante der Haarnadel-Vervollständigung wurde in [ILM09,ILMM11] untersucht. Bei der Haarnadel-Verlängerung erlauben wir, dass nur ein Teil des ungebunden γ-Faktors komplementiert wird, siehe [MMVM10]. Formale Definitionen beider Operationen werden wir später angeben.

Die folgenden Kapitel fassen die Ergebnisse zusammen, die im Rahmen meiner Dissertation [Kop11] erarbeitet wurden. Aufgrund der Seitenbegrenzung wurde allerdings auf Beweise verzichtet.

2 Haarnadel-Vervollständigungen regulärer Sprachen

In diesem Kapitel seinen L_1 und L_2 reguläre Sprachen. Wir untersuchen ihre Haarnadel-Vervollständigung $\mathcal{H}_k(L_1, L_2)$. In Kapitel 2.1 beschäftigen wir uns mit der Klasse der

Sprachen, in der $\mathcal{H}_k(L_1, L_2)$ liegt und wir zeigen, dass es Entscheidbar ist, ob $\mathcal{H}_k(L_1, L_2)$ regulär ist. Kapitel 2.2 befasst sich mit Haarnadel-Vervollständigungen von Sprachen in bestimmten Varietäten, d. h. Unterklassen der regulären Sprachen.

2.1 Eindeutigkeit und Regularität

In [CMVM06] wurden die Abschlusseigenschaften verschiedener Sprachklassen unter Haarnadel-Vervollständigung untersucht. Es wurde gezeigt, dass weder die regulären, noch die kontextfreien Sprachen unter Haarnadel-Vervollständigung abgeschlossen sind. Hingegen sind die kontextsensitiven Sprachen unter dieser Operation abgeschlossen. Weiterhin, ist bekannt, dass Haarnadel-Vervollständigung regulärer Sprachen eine linear kontextfreie Sprache ist.

Beispiel 2.1. Sei $\Sigma = \{a, \overline{a}\}$ und $L_1 = a^*\overline{a}^k$. Die (rechte) Haarnadel-Vervollständigung von L_1 ist

$$\mathcal{H}_k(L_1, \emptyset) = \left\{ a^i\overline{a}^j \;\middle|\; i \geq j \geq k \right\}$$

und somit nicht regulär. Wählen wir allerdings $L_2 = \overline{L_1} = a^k\overline{a}^*$, so ist die Haarnadel-Vervollständigung

$$\mathcal{H}_k(L_1, L_2) = \left\{ a^i\overline{a}^j \;\middle|\; i, j \geq k \right\}$$

wieder eine reguläre Sprache.

In diesem Kapitel betrachten wir das Entscheidungsproblem, ob zwei gegebene reguläre Sprachen eine reguläre Haarnadel-Vervollständigung besitzen. In [CMVM06] wurde dies als offenes Problem gestellt. Da unentscheidbar ist, ob eine gegebene lineare Grammatik eine reguläre Sprache erzeugt [Gre68], kann kein allgemeiner Ansatz gewählt werden, um das Problem zu lösen. Eine erste Lösung für das Problem haben wir in [DKM09] gegeben, wo wir zeigten, dass das Problem in polynomieller Zeit entscheidbar ist. Der Grad des Polynoms wurde allerdings nur abgeschätzt auf 20. In dieser Arbeit wird zum einen ein verbesserter Entscheidungsalgorithmus präsentiert, und zum anderen werden wir die Sprachklasse, in der Haarnadel-Vervollständigungen regulärer Sprachen liegen, weiter einschränken.

Theorem 2.2. *Es seien L_1 und L_2 regulär, dann ist die Haarnadel-Vervollständigung $\mathcal{H}_k(L_1, L_2)$ eindeutig linear kontextfrei.*

Eindeutig bedeutet hier, dass eine lineare Grammatik existiert, welche für jedes Wort in der Sprache genau einen Ableitungspfad besitzt. Dieses neue Resultat ermöglicht es, die Wachstumsfunktion

$$g_{\mathcal{H}_k(L_1,L_2)}(n) = \left|\Sigma^{\leq n} \cap \mathcal{H}_k(L_1, L_2)\right|$$

der Haarnadel-Vervollständigung zu berechnen und sie mit den Wachstumsfunktionen der zugrundeliegenden Sprachen L_1 und L_2 zu vergleichen. Somit lässt sich zum Beispiel berechnen, wie groß die erwartete Anzahl fehlerhafter DNA Stränge ist, die durch Haarnadel-Vervollständigungen während eines PCR Zyklus erzeugt werden (angenommen, dass die Haarnadelbildung ungewollt ist).

Für das Entscheidungsproblem ob $\mathcal{H}_k(L_1, L_2)$ regulär ist, gehen wir davon aus, dass die Sprachen L_1 und $\overline{L_2}$ als deterministische endliche Automaten (DEAs) gegeben sind. Folgende Komplexitäten konnten wir beweisen.

Theorem 2.3. *Es seien L_1 und $\overline{L_2}$ gegeben als DEAs deren Zustandszahl durch n beschränkt ist. Das Problem, ob $\mathcal{H}_k(L_1, L_2)$ regulär ist, ist entscheidbar in*

i.) $\mathcal{O}(n^2)$ *Zeit, falls $L_1 = \emptyset$ oder $L_2 = \emptyset$.*

ii.) $\mathcal{O}(n^6)$ *Zeit, falls $L_1 = \overline{L_2}$.*

iii.) $\mathcal{O}(n^8)$ *Zeit, im Allgemeinen.*

Theorem 2.4. *Es seien L_1 und $\overline{L_2}$ gegeben als DEAs. Das Problem, ob $\mathcal{H}_k(L_1, L_2)$ regulär ist, ist* NL*-vollständig.*

Die Sprachklasse NL enthält diejenigen Probleme, die von einer nicht-deterministischen Turingmaschine in logarithmischem Platz entschieden werden können. Hierbei sei angemerkt, dass die Zugehörigkeit zu NL impliziert, dass das Problem in *Nick's Class* NC_2 liegt und damit effizient parallelisierbar ist, siehe z. B. [Pap94].

2.2 Varietäten

Varietäten sind Unterklassen der regulären Sprachen, welche durch Eigenschaften ihrer syntaktischen Monoide definiert sind. Eine Varietät über endlichen Monoiden ist eine Menge von Monoiden, welche unter Division und direktem Produkt abgeschlossen ist. Für eine ausführliche Einführung in Varietäten über formalen Sprachen sei auf [Pin86] verwiesen. Einige Varietäten besitzen weitere algebraische, kombinatorische und logische Charakterisierungen. So entspricht die Varietät der *aperiodischen* Sprachen $\mathbf{A}$ unter anderem den *Stern-freien* Sprachen und der Klasse von Sprachen, die durch Sätze in *Logik erster Stufe* (first order logic) $\mathrm{FO}[<]$ spezifiziert werden können. Stern-frei bedeutet, dass eine Sprache durch einen regulären Ausdruck ohne Kleene-Stern, dafür aber mit mengentheoretischem Komplement $L^c = \Sigma^* \setminus L$, beschrieben werden kann. Die zweite Varietät, die wir betrachten werden, ist die Varietät $\mathbf{LDA}$, welche der Klasse von Sprachen entspricht, die durch Sätze in *Logik erster Stufe mit zwei Variablen und Nachfolger-Prädikat* $\mathrm{FO}^2[<, +1]$ spezifiziert werden können.

Man beachte, dass die Sprache L_1 in Beispiel 2.1 in der Varietät $\mathbf{LDA} \subsetneq \mathbf{A}$ liegt und dass $\mathcal{H}_k(L_1, \emptyset)$ nicht regulär ist. Sofern die Haarnadel-Vervollständigung $\mathcal{H}_k(L_1, L_2)$ von zwei Sprachen allerdings regulär ist, bleibt die Zugehörigkeit zu den Varietäten $\mathbf{A}$ und $\mathbf{LDA}$ erhalten.

Theorem 2.5. *Seien L_1 und L_2 Sprachen in $\mathbf{A}$ (bzw. $\mathbf{LDA}$). Die Haarnadel-Vervollständigung $\mathcal{H}_k(L_1, L_2)$ ist entweder nicht regulär oder sie gehört ebenfalls zur Varietät $\mathbf{A}$ (bzw. $\mathbf{LDA}$).*

3 Haarnadel-Verlängerung

Die Haarnadel-Verlängerung kann während eines PCR-Schrittes entstehen, wenn der Elongationsprozess abgebrochen wird, bevor der ungebundene γ-Teil des des DNA Strangs vollständig komplementiert ist. Sei $w = \gamma_1\alpha\beta\overline{\alpha}$ ein Wort mit $|\alpha| = k$, dann ist $\gamma_1\alpha\beta\overline{\alpha}\overline{\gamma_2}$ eine *(rechte) Haarnadel-Verlängerung* von w, falls γ_2 ein Suffix von γ_1 ist. Des weiteren ist $\gamma_1\alpha\beta\overline{\alpha}\overline{\gamma_2}$ eine *(linke) Haarnadel-Verlängerung* von $\alpha\beta\overline{\alpha}\overline{\gamma_2}$ falls $|\alpha| = k$ und $\overline{\gamma_1}$ ein Präfix von $\overline{\gamma_2}$ (bzw. γ_1 ein Suffix von γ_2) ist, siehe Abb. 4.

rechte Haarnadel-Verlängerung linke Haarnadel-Verlängerung

Abbildung 4: Haarnadel-Verlängerung

Analog zur Haarnadel-Vervollständigung, definieren wir für Sprachen L_1 und L_2 die Haarnadel-Verlängerung

$$\mathcal{HL}_k(L_1, L_2) = \{\gamma_1\alpha\beta\overline{\alpha}\overline{\gamma_2} \mid |\alpha| = k \wedge ((\gamma_1\alpha\beta\overline{\alpha} \in L_1 \wedge \gamma_2 \text{ ist Suffix von } \gamma_1) \vee (\alpha\beta\overline{\alpha}\overline{\gamma_2} \in L_2 \wedge \gamma_1 \text{ ist Suffix von } \gamma_2))\}.$$

Es ist bekannt, dass die Haarnadel-Verlängerung regulärer Sprachen linear kontextfrei und nicht zwingend regulär ist [MMVM10]. Auf den ersten Blick scheint es, als würden sich die Haarnadel-Vervollständigung und die Haarnadel-Verlängerung, angewandt auf reguläre Sprachen, sehr ähnlich verhalten. Betrachten wir erneut die Frage, ob entscheidbar ist, ob reguläre Sprachen eine reguläre Haarnadel-Verlängerung besitzen, so können wir zumindest im einseitigen Fall analoge Resultate beweisen.

Theorem 3.1. *Sei L eine reguläre Sprache. Das Problem, ob die rechte Haarnadel-Verlängerung $\mathcal{HL}_k(L, \emptyset)$ (bzw. linke Haarnadel-Verlängerung $\mathcal{HL}_k(\emptyset, L)$) regulär ist, ist*

i.) NL*-vollständig,*

ii.) entscheidbar in $\mathcal{O}(n^2)$ Zeit,

falls L (bzw. $\overline{L}$) als DEA mit n Zuständen gegeben ist.

Allerdings war es uns nicht möglich ein analoges Resultat für den beidseitigen oder allgemeinen Fall zu beweisen. Das Problem, ob die Regularität einer Haarnadel-Verlängerung regulärer Sprachen entscheidbar ist, bleibt weiterhin offen. Wir können allerdings ein Indiz nennen, warum dieses Problem schwieriger zu lösen sein könnte, als das Regularitätsproblem der Haarnadel-Vervollständigung. Während im Beweis der Theoreme 2.3 und 2.4 ausnutzt wird, dass eine eindeutige lineare Darstellung der Haarnadel-Vervollständigung konstruiert werden kann (vgl. Theorem 2.2), können wir zeigen, dass keine eindeutig lineare Grammatik für die rechte oder beidseitige Haarnadel-Verlängerung der

Sprache $L = (b^+\alpha)^+\overline{\alpha}$ existiert, wobei $\Sigma = \{a, \bar{a}, b, \bar{b}\}$ und $\alpha = a^k$. Die Sprachen $\mathcal{HL}_k(L, \emptyset)$ und $\mathcal{HL}_k(L, L)$ sind also *inhärent mehrdeutig*. Wir können sogar zeigen, dass für jedes $m \in \mathbb{N}$ ein Wort $w \in \mathcal{HL}_k(L, \emptyset)$ (bzw. $w \in \mathcal{HL}_k(L, L)$) existiert, sodass in jeder Grammatik, die $\mathcal{HL}_k(L, \emptyset)$ (bzw. $\mathcal{HL}_k(L, L)$) generiert, das Wort w mindestens m verschiedene Ableitungen besitzt, d. h. jede Grammatik, die $\mathcal{HL}_k(L, \emptyset)$ oder $\mathcal{HL}_k(L, L)$ erzeugt, besitzt einen *unbeschränkten Grad der Mehrdeutigkeit*.

Theorem 3.2. *Die Haarnadel-Verlängerung $\mathcal{HL}_k(L_1, L_2)$ zweier regulärer Sprachen L_1 und L_2 kann inhärent Mehrdeutig sein, selbst wenn $L_1 = \emptyset$ oder $L_2 = \emptyset$.*

4 Iterierte begrenzte Haarnadel-Vervollständigung

Die begrenzte Haarnadel-Vervollständigung ist eine Variante der Haarnadel-Vervollständigung, bei der wir fordern, dass die Länge des γ-Faktors einer Haarnadel-Vervollständigung durch eine Konstante m begrenzt ist. Sei L eine Sprache und $m \geq 1$, so definieren wir die *begrenzte Haarnadel-Vervollständigung* als

$$\mathcal{H}_{k,m}(L) = \{\gamma\alpha\beta\overline{\alpha}\overline{\gamma} \mid |\alpha| = k \wedge |\gamma| \leq m \wedge (\gamma\alpha\beta\overline{\alpha} \in L \vee \alpha\beta\overline{\alpha}\overline{\gamma} \in L)\}.$$

Im Gegensatz zur unbeschränkten Variante sind alle Klassen der Chomsky Hierarchie unter begrenzter Haarnadel-Vervollständigung abgeschlossen [ILM09, ILMM11]. Weiterhin wurde gezeigt, dass die Klassen der kontextfreien, kontextsensitiven und rekursiv aufzählbaren Sprachen unter *iterierte begrenzte Haarnadel-Vervollständigung* $\mathcal{H}^*_{k,m}(L) = \bigcup_{i\geq 0} \mathcal{H}^i_{k,m}(L)$ abgeschlossen sind. Allerdings blieb in den Arbeiten unbeantwortet, ob die regulären Sprachen unter iterierter begrenzter Haarnadel-Vervollständigung abgeschlossen ist. Wir zeigen, dass diese tatsächlich unter iterierter begrenzter Haarnadel-Vervollständigung abgeschlossen sind. Unser Ergebnis ist sogar allgemeiner und lässt sich auf allen Klassen der Chomsky Hierarchie sowie auf alle „klassischen“ Komplexitätsklassen anwenden.

Theorem 4.1. *Sei $\mathcal{C}$ eine Sprachklasse, welche (effektiv) unter Vereinigung, Durchschnitt mit regulären Sprachen und Konkatenation mit regulären Sprachen abgeschlossen ist, dann ist $\mathcal{C}$ auch (effektiv) abgeschlossen unter iterierter begrenzter Haarnadel-Vervollständigung.*

Insbesondere ist die Klasse der regulären Sprachen effektiv unter iterierter begrenzter Haarnadel-Vervollständigung abgeschlossen.

Im Beweis des Theorems wird die Sprache der iterierten begrenzten Haarnadel-Vervollständigung konstruiert. Dies erlaubt die Untersuchung, wie groß ein nicht-deterministischer endlicher Automat (NEA) ist, der die iterierte Haarnadel-Vervollständigung einer regulären Sprache akzeptiert. Wir geben eine untere und obere Schranke für einen solchen NEA an, welche beide exponentiell in der Konstante m sind.

Theorem 4.2.

i.) Es existiert eine reguläre Sprache L, sodass für alle $m \geq 1$ weder $\mathcal{H}_{k,m}(L)$ noch $\mathcal{H}^*_{k,m}(L)$ durch einen NEA mit weniger als 2^m Zuständen akzeptiert werden kann.

ii.) Sei L eine reguläre Sprache, welche durch einen NEA mit n Zuständen akzeptiert wird und $m \geq 1$. Es existiert ein NEA mit $2^{\mathcal{O}(m^2)}n$ Zuständen, welcher die iterierte begrenzte Haarnadel-Vervollständigung $\mathcal{H}_{k,m}(L)$ akzeptiert.

5 Iterierte Haarnadel-Vervollständigungen einelementiger Sprachen

Die Klasse der Sprachen, die durch iterierte Haarnadel-Vervollständigung einelementiger Sprachen (oder Wörter) erzeugt werden, ist gegeben durch

$$\mathrm{HCS}_k = \{\mathcal{H}_k^*(\{w\}) \mid w \in \Sigma^*\}.$$

Diese Sprachklasse wurde erstmals in [MMY08] untersucht. Da die Klasse NL unter iterierter Haarnadel-Vervollständigung abgeschlossen ist [CMVM06], ist HCS_k eine Teilmenge von NL und somit in den kontextsensitiven Sprachen enthalten. Dennoch wurde die Frage, ob HCS_k nicht-reguläre oder nicht-kontextfreie Sprachen enthält, nicht beantwortet und als offenes Problem in [MMY08] gestellt. Wir lösen dieses Problem, indem wir zeigen, dass die iterierte Haarnadel-Vervollständigung $\mathcal{H}_k^*(\{\alpha b \alpha \overline{\alpha} \alpha c \overline{\alpha}\})$ nicht kontextfrei ist, wobei $\Sigma = \left\{a, \bar{a}, b, \bar{b}, c, \bar{c}\right\}$ und $\alpha = a^k$.

Theorem 5.1. *Die iterierte Haarnadel-Vervollständigung einer einelementigen Sprache ist nicht im Allgemeinen kontextfrei.*

Literatur

[CMVM06] Daniela Cheptea, Carlos Martín-Vide und Victor Mitrana. A new operation on words suggested by DNA biochemistry: Hairpin completion. *Transgressive Computing*, Seiten 216–228, 2006.

[DKM09] Volker Diekert, Steffen Kopecki und Victor Mitrana. On the Hairpin Completion of Regular Languages. In Martin Leucker und Carroll Morgan, Hrsg., *ICTAC*, Jgg. 5684 of *LNCS*, Seiten 170–184. Springer, 2009.

[DMG+98] R. Deaton, R. Murphy, M. Garzon, D.R. Franceschetti und S.E. Stevens. Good encodings for DNA-based solutions to combinatorial problems. *Proc. of DNA-based computers DIMACS Series*, 44:247–258, 1998.

[Gre68] Sheila A. Greibach. A Note on Undecidable Properties of Formal Languages. *Mathematical Systems Theory*, 2(1):1–6, 1968.

[HAK+97] Masami Hagiya, Masanori Arita, Daisuke Kiga, Kensaku Sakamoto und Shigeyuki Yokoyama. Towards Parallel Evaluation and Learning of Boolean μ-Formulas with Molecules. In *Second Annual Genetic Programming Conf.*, Seiten 105–114, 1997.

[HU79] J. E. Hopcroft und J. D. Ulman. *Introduction to Automata Theory, Languages and Computation*. Addison-Wesley, 1979.

[ILM09] Masami Ito, Peter Leupold und Victor Mitrana. Bounded Hairpin Completion. In *LATA '09: Proceedings of the 3rd International Conference on Language and Automata Theory and Applications*, Seiten 434–445, Berlin, Heidelberg, 2009. Springer-Verlag.

[ILMM11] Masami Ito, Peter Leupold, Florin Manea und Victor Mitrana. Bounded hairpin completion. *Inf. Comput.*, 209:471–485, March 2011.

[KKL+05] Lila Kari, Stavros Konstantinidis, Elena Losseva, Petr Sosík und Gabriel Thierrin. Hairpin Structures in DNA Words. In Alessandra Carbone und Niles A. Pierce, Hrsg., *DNA*, Jgg. 3892 of *LNCS*, Seiten 158–170. Springer, 2005.

[KMT07] Lila Kari, Kalpana Mahalingam und Gabriel Thierrin. The syntactic monoid of hairpin-free languages. *Acta Inf.*, 44(3-4):153–166, 2007.

[Kop11] Steffen Kopecki. *Formal language theory of hairpin formations*. Dissertation, University of Stuttgart, 2011.
`http://elib.uni-stuttgart.de/opus/volltexte/2011/63780`

[MMVM09] Florin Manea, Carlos Martín-Vide und Victor Mitrana. On some algorithmic problems regarding the hairpin completion. *Discrete Applied Mathematics*, 157(9):2143–2152, 2009.

[MMVM10] Florin Manea, Carlos Martín-Vide und Victor Mitrana. Hairpin Lengthening. In Fernando Ferreira, Benedikt Löwe, Elvira Mayordomo und Luís Mendes Gomes, Hrsg., *CiE*, Jgg. 6158 of *LNCS*, Seiten 296–306. Springer, 2010.

[MMY08] Florin Manea, Victor Mitrana und Takashi Yokomori. Some Remarks on the Hairpin Completion. In Erzsebet Csuhaj-Varju und Zoltan Esik, Hrsg., *12th International Conference AFL 2008 Proceedings*, Seiten 302–312, 2008.

[MMY09] Florin Manea, Victor Mitrana und Takashi Yokomori. Two complementary operations inspired by the DNA hairpin formation: Completion and reduction. *Theor. Comput. Sci.*, 410(4-5):417–425, 2009.

[Pap94] Christos H. Papadimitriou. *Computatational Complexity*. Addison Wesley, 1994.

[Pin86] Jean-Éric Pin. *Varieties of Formal Languages*. North Oxford Academic, London, 1986.

[Win98] Erik Winfree. Whiplash PCR for O(1) Computing. In *University of Pennsylvania*, Seiten 175–188, 1998.

Steffen Kopecki wurde am 15. Januar 1983 in Stuttgart geboren. Bis zu seinem 12. Lebensjahr wohnte er in Stuttgart Kaltental, wo er die Grundschule Kaltental und bis zur sechsten Klasse das Fanny-Leicht-Gymnasium Stuttgart besuchte. 1995 zog er mit seiner Familie nach Nürtingen und beendete dort seine Schulausbildung am Hölderlin-Gymnasium Nürtingen im Jahr 2002 mit Abitur. Bevor er zurück nach Stuttgart zum Studieren zog, leistete er im Kreiskrankenhaus Nürtingen Zivildienst als OP-Pfleger. An der Universität Stuttgart begann er den Diplomstudiengang Informatik mit Nebenfach Physik, welchen er im Juni 2009 erfolgreich abschloss. Im gleichen Jahr begann er als Doktorand an der Universität Stuttgart im Institut für formale Methoden der Informatik unter seinem Doktorvater Volker Diekert. Zwei Jahre später, im Juni 2011, verteidigte er seine Dissertation erfolgreich und bekam für die Arbeit die Gesamtnote „mit Auszeichnung bestanden“. Im September 2011 zog Steffen Kopecki nach London in Kanada, wo er heute als Post-Doktorand and der Univeristy of Western Ontario in der Arbeitsgruppe von Lila Kari arbeitet.