DNA Computing
Allgemeine Sticker Systeme

Markus Eich
Universität Bremen
eich@informatik.uni-bremen.de

SS 1999

Zusammenfassung

In diesem Artikel wird gezeigt, daß die Allgemeine Stickersprachen in den kontextsensitiven Sprachen enthalten sind und daß es Allgemeine Stickersprachen gibt, die nicht kontextfrei sind.

1 Einleitung

1.1 Einführung in die Stickersysteme

In der Theorie der formalen Sprachen betrachtet man Zeichenketten über einem bestimmten Alphabet. Die formalen Sprachen lassen sich nach Chomsky in 4 Klassen einteilen [2]. Das Wissen über formale Sprachen und Chomsky-Hierarchien wird in diesem Artikel als bekannt vorrausgesetzt und es soll auch nicht näher darauf eingegangen werden. Betrachtet man bei einem doppelten DNS-Strang die 4 Nukleotide Adenin, Guanin, Thymin und Cytosin (im folgenden als A, G, T und C bezeichnet), so kann man diesen auch als eine Zeichenkette über dem Alphabet A, G, T, C betrachten. Die doppelte Helixstruktur von DNS-Strängen hat eine wichtige Eigenschaft, die man sich beim Rechnen mit DNS-Strängen zunutze machen kann, nämlich die Komplementarität der sich verbindenden Nukleotide. So bindet sich A nur an T und G nur an C. Im folgenden Abschnitt werden zuerst die Konzepte von sogenannten Sticker-Systemen vorgestellt. Im 3. Abschnitt wird gezeigt, daß die von Allgemeinen Stickersystemen erzeugte Sprache in der Klasse der kontextsensitiven Sprachen enthalten ist. Dazu wird ein linear beschränkter Automat mit Hilfe von Sticker Systemen gebaut, der bekanntlich die kontextsensitiven Sprachen erkennt [2]. Im weiteren wird gezeigt, daß es eine von Allgemeinen Stickersystemen erzeugte Sprache gibt, die nicht kontextfrei ist. Die von von diesen Systemen erzeugten Sprachen werden als ASL bezeichnet, als Allgemeine Stickersprachen (arbitrary sticker language). Die Ergebnisse des 3. Abschnittes siedeln die ASL in der Chomsky-Hierarchie an.

2 Stickersysteme

2.1 Einführung in die Stickersysteme

Zunächst beginnen wir mit dem Konzept der Stickersysteme. Wie bereits erwähnt bestehen DNS-Stränge aus einer Doppelhelix. Würde man diese Helix "aufwickeln", so hätte man zwei übereinanderliegende Nukleotidketten, die aus A, T, G und C bestehen. Dabei ist jedes obenliegenden Nukleotid komplementär zu dem Darunterliegenden. A ist dabei komplementär zu T und G ist komplementär zu C. Diese Komplementarität wird auch als Watson-Crick-Komplementarität bezeichnet. Diese Doppelstränge besitzen nun zwei "klebrige" Sticky-Ends, an denen weiter DNS-Stränge durch Ligation und Abkühlung angehängt werden können. Um diesen Vorgang zu modellieren betrachtet man ein Alphabet V und eine symmetrische Relation $\rho \subseteq V \times V$. Dabei wird bei den Stickersystemen von der in Natur insofern abstrahiert, daß Stickersysteme über einem Alphabet arbeiten, was mehr als die Terminale A, G, C und T enthält. Die Relation ρ wird Komplementaritätsrelation genannt. Da ein DNA - Molekülen aus Doppelsträngen bestehen, betrachten wir die übereinanderliegenden Elemente (x_1, x_2) als Teil eines Doppelstrings aus der Menge $V^* \times V^*$. Die Elemente werden geschrieben als $\begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$. In diesem Fall ist jedoch nichts über die Koplementarität der beiden Elemente x_1 und x_2 ausgesagt. Sind die beiden Elementen x_1 und x_2 komplementär, so schreiben wir $\begin{bmatrix} x_1 \\ x_2 \end{bmatrix}_\rho$ wobei ρ die Komplementaritätsrelation angibt. Wir schreiben ausserdem

$$\begin{bmatrix} V \\ V \end{bmatrix}_\rho = \{ \begin{bmatrix} a \\ b \end{bmatrix} \mid a; b \in V, (a,b) \in \rho, WK_\rho(V) = \begin{bmatrix} V \\ V \end{bmatrix}_\rho^*$$

$WK_\rho(V)$ wird Watson Crick Domäne genannt über einem Alphabet V bezüglich der Komplemenatritätsrelation ρ. Die Elemente $\begin{bmatrix} a_1 \\ b_1 \end{bmatrix} \begin{bmatrix} a_2 \\ b_2 \end{bmatrix} \dots \begin{bmatrix} a_n \\ b_n \end{bmatrix} \in WK_\rho(V)$werden auch als $\begin{bmatrix} w_1 \\ w_2 \end{bmatrix}$ geschrieben, wobei $w_1 = a_1 a_2 \dots a_n$ ist beziehungsweise $w_2 = b_1 b_2 \dots b_n$. Solche Elemente $\begin{bmatrix} w_1 \\ w_2 \end{bmatrix} \in WK_\rho(V)$werden vollständige Doppelstränge genannt, w_1 und w_2 bezeichnet man als Einzelstränge. Beachte den Unterschied zwischen $\begin{pmatrix} x \\ y \end{pmatrix}$ und $\begin{bmatrix} x \\ y \end{bmatrix}$: Bei $\begin{bmatrix} x \\ y \end{bmatrix}$ handelt es sich um wohlgeformte DNA Moleküle, bei $\begin{pmatrix} x \\ y \end{pmatrix}$ nicht. Betrachtet man nun wohlgeformte Doppelstrangsequenzen so impliziert dies, daß

- die beiden Einzelstränge w_1 und w_2 die gleiche Länge haben
- abhängig von der Relation ρ komplementär sind

Diese Eigenschaften gelten von Natur aus für DNA-Doppelstränge. Für das Modell der Stickersysteme benötigen wir jedoch auch nicht vollständige DNA-Moleküle. Solche Moleküle besitzen links, rechts oder an beiden Seiten Überhänge. Die Menge der unvollständigen Moleküle bezeichnen wir als Menge

$$W_\rho(V) = L_\rho(V) \cup R_\rho(V) \cup LR_\rho(V),$$

wobei die einzelnen Teilmengen definiert sind als

$$L_\rho(V) = (\left(\begin{array}{c} \lambda \\ V^* \end{array} \right) \cup \left(\begin{array}{c} V^* \\ \lambda \end{array} \right)) \left[\begin{array}{c} V \\ V \end{array} \right]_\rho^*,$$

$$R_\rho(V) = \left[\begin{array}{c} V \\ V \end{array} \right]_\rho^* (\left(\begin{array}{c} \lambda \\ V^* \end{array} \right) \cup \left(\begin{array}{c} V^* \\ \lambda \end{array} \right)),$$

$$LR_\rho(V) = (\left(\begin{array}{c} \lambda \\ V^* \end{array} \right) \cup \left(\begin{array}{c} V^* \\ \lambda \end{array} \right)) \left[\begin{array}{c} V \\ V \end{array} \right]_\rho^+ (\left(\begin{array}{c} \lambda \\ V^* \end{array} \right) \cup \left(\begin{array}{c} V^* \\ \lambda \end{array} \right)).$$

$L_\rho(V)$ bezeichnet dabei die DNA-Moleküle mit linkem Überhang, $R_\rho(V)$ die Moleküle mit rechtem Überhang und $LR_\rho(V)$ die Moleküle mit sowohl rechtem als auch linkem Überhang. Abbildung 1 zeigt alle Möglichkeiten von unvollständigen DNA-Molekülen.

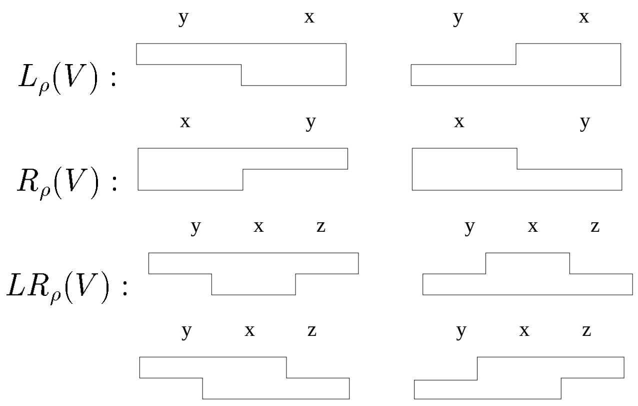


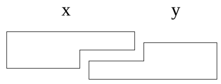
Abbildung 1: Mögliche Formen von unvollständigen DNA-Molekülen

Beachte, daß die Überhänge auch leer sein können. Daraus ergibt sich, daß $WK_\rho(V) \subset W_\rho(V)$ ist. Die einzelnen Elemente von $W_\rho(V)$ werden als Dominos bezeichnet. Um nun verschiedene Dominos zusammenzukleben, beginnt man mit einem sogenanten "guten" Anfang. Als einen guten Anfang bezeichnet man ein Domino, an dem mindestens an einer Stelle die Bedingung $\left[\begin{array}{c} a \\ b \end{array} \right]$ gilt, wobei a und b $\neq \lambda$ sind. Eine weitere Voraussetzung bei einem guten Anfang ist, daß, falls mehrere Doppelsequenzen aus $WK_\rho(V)$ sind, diese Sequenzen aufeinanderfolgen. Einen solchen guten Anfang kann man ganz allgemein schreiben als

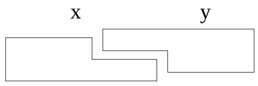
$$x = x_1 x_2 x_3,$$

wobei $x_2 \in WK_\rho(V) - \left[\begin{array}{c} \lambda \\ \lambda \end{array} \right]\}$ und $x_1, x_3 \in \left(\begin{array}{c} V^* \\ \lambda \end{array} \right) \cup \left(\begin{array}{c} \lambda \\ V^* \end{array} \right)$. Das "kleben" oder "sticking" ist definiert als $\mu(x, y)$, wobei wir ein DNA-Molekül an die rechte Seite des bisherigen Dominos anhängen. Dabei werden ingesamt 8 verschiedene Fälle des stickings unterschieden:

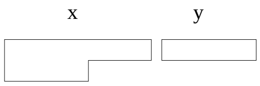
1.) $x_3 = \begin{pmatrix} u \\ \lambda \end{pmatrix}, y = \begin{pmatrix} \lambda \\ v \end{pmatrix} y' \mid u, v \in V^*$ so daß $\begin{bmatrix} u \\ v \end{bmatrix} \in WK_\rho(V)$ und $y' \in L_\rho(V)$, dann ist $\mu(x,y) = x_1 x_2 \begin{bmatrix} u \\ v \end{bmatrix} y'$.



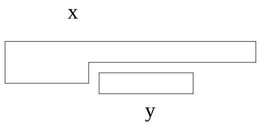
2.) $x_3 = \begin{pmatrix} \lambda \\ v \end{pmatrix}, y = \begin{pmatrix} u \\ \lambda \end{pmatrix} y' \mid u, v \in V^*$ so daß $\begin{bmatrix} u \\ v \end{bmatrix} \in WK_\rho(V)$ und $y' \in L_\rho(V)$, dann ist $\mu(x,y) = x_1 x_2 \begin{bmatrix} u \\ v \end{bmatrix} y'$.



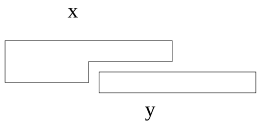
3.) $x_3 = \begin{pmatrix} u_1 \\ \lambda \end{pmatrix}, y = \begin{pmatrix} u_2 \\ \lambda \end{pmatrix} \mid u_1, u_2 \in V^*$ dann ist $\mu(x,y) = x_1 x_2 \begin{pmatrix} u_1 u_2 \\ \lambda \end{pmatrix}$.



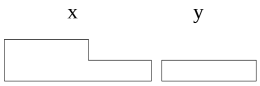
4.) $x_3 = \begin{pmatrix} u_1 u_2 \\ \lambda \end{pmatrix}, y = \begin{pmatrix} \lambda \\ v \end{pmatrix} \mid u_1, u_2, v \in V^*$ so daß $\begin{bmatrix} u_1 \\ v \end{bmatrix} \in WK_\rho(V)$ dann ist $\mu(x,y) = x_1 x_2 \begin{bmatrix} u_1 \\ v \end{bmatrix} \begin{pmatrix} u_2 \\ \lambda \end{pmatrix}$.



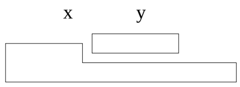
5.) $x_3 = \begin{pmatrix} u \\ \lambda \end{pmatrix}, y = \begin{pmatrix} \lambda \\ v_1 v_2 \end{pmatrix} \mid u, v_1, v_2 \in V^*$ so daß $\begin{bmatrix} u \\ v_1 \end{bmatrix} \in WK_\rho(V)$, dann ist $\mu(x,y) = x_1 x_2 \begin{bmatrix} u \\ v_1 \end{bmatrix} \begin{pmatrix} \lambda \\ v_2 \end{pmatrix}$.



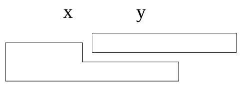
6.) $x_3 = \begin{pmatrix} \lambda \\ v_1 \end{pmatrix}, y = \begin{pmatrix} \lambda \\ v_2 \end{pmatrix} \mid v_1, v_2 \in V^*$, dann ist $\mu(x,y) = x_1 x_2 \begin{pmatrix} \lambda \\ v_1 v_2 \end{pmatrix}$.



7.) $x_3 = \begin{pmatrix} \lambda \\ v_1 v_2 \end{pmatrix}, y = \begin{pmatrix} u \\ \lambda \end{pmatrix} \mid u, v_1, v_2 \in V^*$ so daß $\begin{bmatrix} u \\ v_1 \end{bmatrix} \in WK_\rho(V)$, dann ist $\mu(x,y) = x_1 x_2 \begin{bmatrix} u \\ v_1 \end{bmatrix} \begin{pmatrix} \lambda \\ v_2 \end{pmatrix}$.



8.) $x_3 = \begin{pmatrix} \lambda \\ v \end{pmatrix}, y = \begin{pmatrix} u_1 u_2 \\ \lambda \end{pmatrix} \mid v, u_1, u_2 \in V^*$ so daß $\begin{bmatrix} u_1 \\ v \end{bmatrix} \in WK_\rho(V)$ dann ist $\mu(x,y) = x_1 x_2 \begin{bmatrix} u_1 \\ v \end{bmatrix} \begin{pmatrix} u_2 \\ \lambda \end{pmatrix}$.



In der gleichen Weise wird die Funktion $\mu(y,x)$ definiert, wobei ein neues Molekül von links an das vorhergehende Dominostück angehangen wird. Das Ankleben an das rechte Ende wird nicht vom Ankleben an das linke Ende unterschieden, da in beiden Fällen von einem guten Anfangsdomino ausgegangen wird und nach der Operation wiederum ein gutes Anfangsdomino vorliegt. Wichtig ist nur die Reihenfolge der Sequenzen an den sticky ends. Erinnern wir uns, daß wir untersuchen wollen, welche Zeichenketten wir mit Stickersystemen erzeugen können und versuchen diese in die Chomsky-Hierarchie einzuordnen. Bekanntlich beruhen die Chomsky-Grammatiken auf Termersetzung. Die erzeugende Operation für Stickersysteme ist das Ankleben von DNA-Molekülen an einen, wie weiter oben beschriebenen, "guten Anfang".

2.2 Klassifizierung von Allgemeinen Stickersystemen

Im vorhergehenden Abschnitt haben wir eine kurze Einführung zu Stickersystemen kennengelernt. Im weiteren werden nun Allgemeine Stickersysteme vorgestellt. Ein Stickersystem ist ein Konstrukt

$$\gamma = (V, \rho, A, D).$$

Dabei ist V ein Alphabet, $\rho \subseteq V \times V$ eine symmetrische Relation, $A \subseteq LR\rho(V)$ (im weiteren auch Axiom genannt) und $D \subseteq W_\rho(V) \times W_\rho(V)$. Die Relation ρ bezeichnet dabei die Komplementaritätsfunktion. Eine Berechnung mit Stickersystemen geht nun aus von einem Domino $x_1 \in A$ und endet in einem vollständigen Doppelstrang $x_k \in WK_\rho(V)$. Daraus folgt eine Berechnungssequenz $\sigma : x_1 \Rightarrow x_2 \Rightarrow \ldots \Rightarrow x_k$. Wir abstrahiern hierbei ein wenig von den realen DNA-Molekülen, indem wir sagen wir erzeugen bei der Berechnung eine Sequenz von Zeichenketten, einen String. Bei der Berechnung kann man nun verschiedene Restriktionen zulassen, zum Beispiel kann man für die Menge D nur Paare (u,v) zulassen, in denen entweder $u = \lambda$ oder $v = \lambda$ ist. Die durch diese Restriktion erzeugte Sprache heißt Einseitiges Stickersystem. Wir werden an dieser Stelle jedoch nur den allgemeinsten Fall betrachten, die Allgemeinen Stickersysteme. Einzige Vorraussetzung für die erzeugte Sprache ist, daß x_r ein vollständiges Molekül ist. Wir bezeichnen die erzeugte Sprache ASL(n), für allgemeine oder willkürliche Stickersprache. Das n steht hierbei für "nicht eingeschränkt". Einschränkungen der Berechnungssequenz σ wären zum

Beispiel, daß σ primitiv ist, oder die Überhänge der Moleküle $x_i \mid 1 \leq i \leq k$ in ihrer Länge beschränkt sind. Primitiv bedeutet wiederum, daß keine der Zwischenberechnungen von σ einen vollständigen Doppelstrang ergibt, sondern x_r das erste und einzige vollständige Molekül ergibt. Überhangsbeschränkt durch einen Wert d heißt, daß kein Überhang einer Zwischenberechnung $x_i \mid 1 \leq i \leq k$ grösser als d ist.

3 Die Mächtigkeit von Allgemeinen Stickersystemen

Wir werden nun zeigen, daß die Sprachen, die von Allgemeinen Stickersystemen erzeugt werden können in der Klasse der kontextsensitiven Sprachen enthalten sind. Weiterhin zeigen wir, daß es Sprachen gibt, die in ASL enthalten sind, jedoch nicht kontextfrei sind.

3.1 $ASL(n) \subseteq CS$

Daß die Allgemeinen Stickersprachen eine Teilmenge der kontextsensitiven Sprachen sind, sieht man daran, daß man bei Stickersystemen (das gilt sowohl für ASL als auch für andere Stickersysteme) keine Löschoperationen hat. Man kann daher einen linear beschränkten Automaten bauen, der die Stickersprachen erkennt. Aus der Automatentheorie [2] wissen wir, daß ein LBA genau die kontextsensitiven Sprachen erkennt. Dauraus folgt in einer straightforward Konstruktion, daß $ASL(n) \subseteq CS$.

3.2 $ASL(n) - CF \neq \emptyset$

Wir werden nun zeigen, daß es mindestens eine Sprache gibt, die mit Allgemeinen Stickersystemen erzeugt werden kann, die nicht kontextfrei ist. Dazu genügt es eine solche Sprache zu konstruieren.Man betrachte folgendes Stickersystem:

$$\gamma = (V, \rho, A, D),$$

$$V = \{a, b, c\},$$

$$\rho = \{(a,a), (b,b), (c,c)\},$$

$$A = \{\begin{bmatrix} a \\ a \end{bmatrix}\},$$

$$D = \{(\begin{pmatrix} b \\ \lambda \end{pmatrix}, \begin{pmatrix} b \\ \lambda \end{pmatrix}), (\begin{pmatrix} c \\ \lambda \end{pmatrix}, \begin{pmatrix} \lambda \\ \lambda \end{pmatrix}), (\begin{pmatrix} \lambda \\ b \end{pmatrix}, \begin{pmatrix} \lambda \\ \lambda \end{pmatrix}), (\begin{pmatrix} \lambda \\ c \end{pmatrix}, \begin{pmatrix} \lambda \\ b \end{pmatrix})\}$$

Es gilt

$$WK_\rho(V) = \{\begin{bmatrix} a \\ a \end{bmatrix}, \begin{bmatrix} b \\ b \end{bmatrix}, \begin{bmatrix} c \\ c \end{bmatrix}\}^*.$$

Weiterhin sind wegen der Bauart der Dominos nur Worte in der Sprache enthalten die folgende Form haben :

$$\{\begin{bmatrix} b \\ b \end{bmatrix}, \begin{bmatrix} c \\ c \end{bmatrix}\}^* \begin{bmatrix} a \\ a \end{bmatrix} \begin{bmatrix} b \\ b \end{bmatrix}^m$$

Wegen der Dominos ($\begin{pmatrix} b \\ \lambda \end{pmatrix}$,$\begin{pmatrix} b \\ \lambda \end{pmatrix}$) gilt das vor $\begin{bmatrix} a \\ a \end{bmatrix}$ genau soviele $\begin{pmatrix} b \\ b \end{pmatrix}$ stehen wie danach, also genau m Stück. Die Dominos ($\begin{pmatrix} \lambda \\ c \end{pmatrix}$,$\begin{pmatrix} \lambda \\ b \end{pmatrix}$ garantieren, daß links von $\begin{bmatrix} a \\ a \end{bmatrix}$ genau so viele $\begin{bmatrix} c \\ c \end{bmatrix}$ stehen wie $\begin{bmatrix} b \\ b \end{bmatrix}$ rechts davon, also ebenfalls m Stück. Zusammengenommen ergibt sich folgende Sprache:

$$ASL(n)_\gamma = \{c^m b^m a b^m\}.$$

Diese Sprache ist bekanntlich nicht kontextfrei [2]. Wir haben damit also gezeigt, daß es eine Allgemeine Stickersprache gibt, die nicht kontextfrei ist, was zu beweisen war.

4 Zusammenfassung der Ergebnisse

In diesem Artikel wurden Stickersysteme, speziell die Allgemeinen Stickersysteme eingeführt. Sie modellieren den physikalisch-chemischen Prozess von Ligation und Aushärtung, die eingesetzt werden, um DNA-Doppelstränge miteinander zu verketten. Es kam dabei heraus, daß sich, betrachtet man die Nukleotide als Zeichenketten, Sprachen erzeugen lassen. Abstrahierend vom natürlichen Modell können auch andere Zeichen als die Nukleotid-Terminale A, T, C, und G verwendet werden. Wie haben die von der allgemeinsten Form der Stickersysteme erzeugten Sprache ASL(n) betrachtet und diese in die Chomsky-Hierarchie der formalen Sprachen eingeordnet. Wir haben folgendes festgestellt:

- $ASL(n) \subseteq CS$
- $ASL(n) - CF \neq \emptyset$

Was noch an Untersuchungen bleibt, ist die Mächtigkeit von anderen Stickersystemen zu untersuchen, wie zum Beispiel den Einseitigen Stickersystemen oder zu untersuchen, wie sich eine Beschränkung der ASL(n) auf die Mächtigkeit de Sprache auswirkt. Man könnte ASL(b) untersuchen, wobei im Gegensatz zu ASL(n) die Einschränkung gilt, daß die Maximallänge der Überhänge bei der Erzeugung beschränkt ist (bounded). Interessant wäre auch die Umsetzung in die Praxis um Sprachen in einem Reagenzglas zu erzeugen, anstatt sie von einer Maschine berechnen zu lassen. Bleibt, die Entwicklung von DNA-Computing und die Auswirkungen auf die Informationstechnik zu beobachten.

Literatur

[1] G. Paun, G. Rozenberg und A. Salomaa. *DNA Computing - New Computing Paradigms.* Springer-Verlag, Berlin Heidelberg, 1998

[2] John E. Hopcroft und Jeffrey D. Ullman. *Introduction to automata theory, languages and computation.* Addison-Wesley, 1994