Biochemisch inspirierte Berechnungen

Stringoperationen

Peter Leupold

Vorlesung Wintersemester 2009/2010

Duplication

Wir gehen aus von einem Spezialfall des Splicing, nämlich dem Splicing eines DNA-Stranges mit sich selbst, d.h. mit einem zweiten Strang derselben Basensequenz.

Erfolgt der Schnitt des Restriktionsenzzyms in beiden Strängen an derselben Stelle, dann hat die ganze Operation keinen sichtbaren Effekt.

Bei Schnitten an zwei verschiedenen Stellen entsteht hingegen ein Strang, der die Sequenz zwischen den Schnittstellen zweimal enthält sowie ein zweiter Strang aus dem diese Sequenz gelöscht wird.

Duplication

Derartige Verdopplungen in Genomen sind ein durchaus häufiges Phänomen und somit ein Mechanismus der Genmutation.

Die Rückverfolgung der Duplikationsgeschichte eines Chromosoms kann also Hinweise auf die Verwandtschaftsgeschichte verschiedener Populationen derselben Spezies liefern.

Andererseits sind Wiederholungen in Zeichenketten eines der ältesten Themen dessen, was heute Theoretische Informatik ist.

Axel Thue

Vor etwa einhundert Jahren veröffentlichte der Nowegische Mathematiker Axel Thue (Semi-Thue-Systeme) mehrere Arbeiten über Wiederholungen in Zeichenreihen:

1906: Über unendliche Zeichenreihen.

1912: Über die gegenseitige Lage gleicher Teile verschiedener Zeichenreihen.

1914: Probleme über die Veränderung von Zeichenreihen nach gegebenen Regeln.

Alle in: Norske Videnskabers Selskabs Skrifter Mat.-Nat. Kl. (Kristiania)

Vermeidbarkeit (Avoidability)

Die "moderne" Terminologie für die Probleme, die Axel Thue behandelte:

Eine Wort vermeidet ein Muster u, wenn es keinen Faktor der Form $\phi(u)$ für einen nichtlöschenden Morphismus ϕ enthält.

- Ein Wort heisst *quadratfrei*, wenn es das Muster vv vermeidet.
- Ein Wort heisst *kubenfrei*, wenn es das Muster v^3 vermeidet.
- Ein Wort heisst n^+*-frei*, wenn es evtl. eine Wiederholung n-ten Grades enthält, aber keine Wiederholung m-ten Grades für irgendein $m > n$ (beide rationale Zahlen).

Vermeidbarkeit (Avoidability)

Vermeidbarkeit hängt in der Regel auch von der Länge der Wörter ab. Trivialerweise lassen sich Muster der Länge ℓ in Wörtern, die kürzer sind als ℓ, vermeiden.

Besonders interessant ist es also, zu wissen, ob sich ein Muster in Wörtern beliebiger Länge vermeiden lässt.

Dies ist gleichbedeutend damit, dass es unendlich viele Wörter gibt, die das Muster vermeiden.

Nach dem Zornschen Lemma bedeutet dies wiederum, dass es ein unendliches Wort gibt, das das Muster vermeidet.

Vermeidbarkeit (Avoidability)

Unendliche Wörter sind in unserem Zusammenhang rechtsunendlich.

Das heisst, das Wort beginnt links mit einem ersten Buchstaben, setzt sich nach rechts fort und hat keinen letzten Buchstaben:

$$ababcabbcaabac\ldots$$

Die Menge aller unendlichen Wörter über einem Alphabet Σ wird mit Σ^{ω} bezeichnet.

Thues Ergebnisse

Satz

Über einem Alphabet von zwei Buchstaben gibt es kein unendliches quadratfreies Wort.

- λ
- a
- b
- ab
- ba
- aba
- bab

Thues Ergebnisse

Satz

Über einem Alphabet von drei oder mehr Buchstaben gibt es unendliche quadratfreie Wörter.

Satz

Über einem Alphabet von zwei Buchstaben gibt es unendliche 2^+-freie Wörter.

Das Thue-Morse Wort ist definiert durch $\mu(a) := ab$ und $\mu(b) := ba$.

Weiterhin der Morphismus:

$$\delta(c) := a,\ \delta(b) := ab,\ \delta(a) := abb.$$

Dann ist $\mu^\omega(a)$ 2^+-frei, und $\delta^{-1}(\mu^\omega(a))$ ist quadratfrei (δ^{-1} ist im allgemeinen nicht wohldefiniert).

Morses Wiederentdeckung

Thues Veröffentlichungen fanden keine weite Verbreitung. So wurden seine Ergebnisse im Laufe des letzten Jahrhunderts mehrmals neu entdeckt.

Vielleicht die interessanteste darunter ist die von Morse im Zusammenhang mit unendlichen Schachpartien.

Mit einer kleinen Regeländerung würde die Existenz quadratfreier Wörter unendlich lange Schachpartien ermöglichen.

Der Satz von Myhill-Nerode

Die *(rechts-)syntaktische Kongruenz* einer Sprache L ist

$$u \sim_L v :\leftrightarrow \forall w \in \Sigma^*(uw \in L \leftrightarrow vw \in L).$$

Dies ist eine Äquivalenzrelation.

Satz (Kleene / Myhill / Nerode)

Eine Sprache L ist regulär, genau dann wenn $\sim_L$ von endlichem Index ist.

Der Index einer Äquivalenzrelation ist die Zahl ihrer Äquivalenzklassen.

Duplication

Wir verwenden für die Verdopplung von Faktoren (Teilwörtern) folgende Relation:

$$u \heartsuit v :\Leftrightarrow \exists z[z \in \Sigma^+ \wedge u = u_1 z u_2 \wedge v = u_1 z z u_2]$$

$\heartsuit^*$ ist die reflexive, transitive Hülle der Relation $\heartsuit$.

Teilweise betrachten wir $\heartsuit$ auch als Textersetzungssystem $\{z \to z^2 : z \in \Sigma^+\}$.

Längenbeschränkte Duplication

Da sich die allgemeine Duplikationsrelation als sehr kompliziert bzw. schwer zu beschreiben herausgestellt hat, betrachten wir auch folgende beiden längenbeschränkten Varianten:

$$u \heartsuit^{\leq n} v :\Leftrightarrow \exists w [u = u_1 w u_2 \wedge v = u_1 w w u_2 \wedge |w| \leq n]$$

und

$$u \heartsuit^{= n} v :\Leftrightarrow \exists w [u = u_1 w u_2 \wedge v = u_1 w w u_2 \wedge |w| = n].$$

Wir sprechen von n-beschränkter bzw. von n-Duplication.

Duplication

Definition

Die von einem Wort w erzeugte Duplication Sprache ist

$$w^{\heartsuit} := \{u : w \heartsuit^* u\}.$$

Die Sprachen $w^{\heartsuit \leq n}$ und $w^{\heartsuit = n}$ sind analog definiert.

Auf eine ganze Sprache wenden wir die Relationen in der offensichtlichen Weise an, also zum Beispiel

$$L^{\heartsuit} := \bigcup_{w \in L} w^{\heartsuit}.$$

Beschränkte Duplication

Satz

Die Klasse der regulären Sprachen ist unter 2-beschränkter Duplikation abgeschlossen.

Beweis an der Tafel...

Satz

Die Klasse der regulären Sprachen ist für $n \geq 4$ nicht unter n-beschränkter Duplikation abgeschlossen.

Beweis an der Tafel...

Duplication über zweibuchstabigem Alphabet

Satz

Die Klasse der regulären Sprachen über zweibuchstabigem Alphabet ist unter Duplikation abgeschlossen.

Sei nun $\rightarrow$ die Ableitungsrelation des Textersetzungssytems $R = \{a \rightarrow aa, b \rightarrow bb, ab \rightarrow abab, ba \rightarrow baba\}$, das die Sprache $w^{\heartsuit \leq 2}$ generiert.

Lemma

Für jedes Wort $u \in \{a, b\}^*$ gelten $ab \xrightarrow{*} abubab$, $ab \xrightarrow{*} abuaab$, und $ab \xrightarrow{*} abuab$.

Satz

Über zweibuchstabigem Alphabet gilt $w^{\heartsuit \leq n} = w^{\heartsuit \leq 2}$ und somit $w^{\heartsuit} = w^{\heartsuit \leq 2}$ für alle Wörter w und alle $n \geq 2$.

Duplication über dreibuchstabigem Alphabet

Satz

Die Klasse der regulären Sprachen über dreibuchstabigem Alphabet ist nicht unter Duplikation abgeschlossen.

Dies lässt sich ähnlich zeigen wie die Tatsache, dass die regulären Sprachen für $n \geq 4$ nicht unter n-beschränkter Duplikation abgeschlossen sind.

Dies lässt die Frage offen, ob der Duplikationsabschluss kontextfrei ist oder nicht — kontextsensitivität ist offensichtlich gegeben.