

# Sequenzvergleiche mit Suffixbäumen

Jürgen Kleffe  
Institut für Molekularbiologie und Biochemie  
Bereich Molekularbiologie UND Bioinformatik  
Arnimallee 22, 14195 Berlin

## Zusammenfassung

Angesichts des exponentiellen Wachstums der Sequenzdatenbanken gewinnen effektive Verfahren zum Sequenzvergleich immer größere Bedeutung. Wir geben deshalb eine kurze Einführung in die Anwendung von Suffixbäumen. Diese Technik erlaubt Sequenzvergleiche in mit der Sequenzlänge proportional anwachsender Rechenzeit und ist damit unverzichtbar für den Umgang mit diesen Daten.

# Motivation

Taglich gibt es neue Nachrichten vom Fortgang der Genomforschung. Die im Februar 2001 veroffentlichte Sequenz des Humangenoms umfat ungefahr 94% der gesamten chromosomalen DNA. Es existieren zwei alternative Entwurfe.

"The golden path":

<http://genome.ucsc.edu>

"The Human Genome Draft Sequence":

[http:// www.ncbi.nlm.nih.gov/genome/guide/human/](http://www.ncbi.nlm.nih.gov/genome/guide/human/)

Diese Sequenzen bilden jedoch nur einen kleinen Teil der insgesamt verfugbaren Genomdaten, die taglich um mehrere Millionen Basenpaare wachsen. Jeden Tag werden uber 100 neue Proteinsequenzen und wenigstens eine neue Proteinstruktur in die Datenbanken aufgenommen.

Verbinden Sie sich mit der GenBank, um den aktuellen Stand zu erfahren!

Enthalt sie schon 16 Milliarden Basenpaare?

<http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=nucleotide>

Die europaischen Datenbanken erreichen sie unter

<http://SRS.EBI.ac.uk>

Die Auswertung der gesammelten Daten bleibt jedoch deutlich zuruck. Schwierige Probleme bereiten schon einfache Vergleiche so groer Mengen von Sequenzen. Die Suche nach moglichst langen gemeinsamen Teilstucken zweier Genome erfordert einen enormen technischen Aufwand.

## Auf einfache Weise geht es nicht

Um lange gemeinsame Teilsequenzen zweier Genome zu finden, mussen beginnend in zwei beliebigen Positionen beider Sequenzen Buchstabe fur Buchstabe verglichen werden, bis eine Differenz (mismatch) auftritt. Abbildung 1 zeigt das Prinzip eines solchen primitiven Sequenzvergleichs.

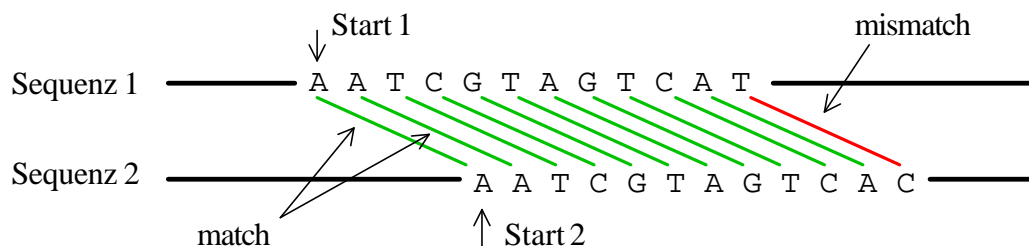


Abbildung 1: Vergleich zweier Sequenzen 1 und 2 beginnend in den Startpunkten Start1 und Start 2. Nach 12 Buchstabenvergleichen wird eine gemeinsame Teilsequenz der Lange 11 gefunden.

Wenn ein solcher Vergleich im Mittel nur eine Mikrosekunde dauert, dann sind fur zwei Sequenzen mit  $10^7$  Basenpaaren  $10^{14}$  solche Vergleiche durchzufuhren, die  $10^8$  Sekunden, d.h.

mehr als 3 Jahre dauern würden. Die Ursache ist der quadratisch mit der Sequenzlänge wachsende Rechenaufwand eines solchen naiven Verfahrens.

### **Mathematik und Informatik müssen helfen**

Die Genomdatenbanken blieben für die wissenschaftliche Forschung wertlos, wenn nicht Mathematiker und Informatiker raffiniertere Methoden zum Sequenzvergleich entwickelt hätten. Sie beruhen meist auf Suffixbäumen oder Suffixfeldern und benötigen für den Sequenzvergleich nur wenige Sekunden. Wie viele Sekunden es genau sind, das hängt vom Algorithmus ab und ist Gegenstand des Wettbewerbs der Erfinder neuer Algorithmen, die mit der Genomforschung zu sehr gefragten Leuten wurden. Ihre Programme entsprechen den Waschanlagen und Sieben der alten Goldsucher in den Rocky Mountains. Das neue Gold heißt "Genetische Information" und dient der Medikamentenentwicklung und Biotechnologie. Es muß schnell und gründlich gewaschen werden. Kein Stäubchen darf entkommen.

**(Links zu Suffixfeldern und Suffixbäumen)**

# Suffixbäume

Suffixbäume haben die Technik zum Vergleich von Texten revolutioniert. Sie erlauben die Lösung vieler Suchaufgaben in sehr kurzer Zeit, weil der notwendige Aufwand nur linear mit den Längen der untersuchten Texte wächst. Wir geben hier Schritt für Schritt einen Einblick in diese Technik. Viel zitiert werden die Publikationen von McCreight (1976), Ukkonen (1985), Landau & Vishkin (1988) und Gusfield (1997)

## Was ist ein Suffix?

Der Suffix Nummer  $i$  einer Sequenz  $S$  ist die Teilsequenz, die mit dem  $i$ -ten Nukleotid  $S[i]$  von  $S$  beginnt und bis an das Ende von  $S$  reicht. Eine Sequenz hat genau so viele Suffixe wie Nukleotide. Abbildung 2 zeigt die Liste der Suffixe der Sequenz  $S = \text{aaattttttt}$ .

Sequenz:	aaattttttt
Suffixnummern:	0123456789
-----	-----
Suffix 0:	aaattttttt
Suffix 1:	aatttttttt
.....	.....
.....	.....
Suffix 8:	tt
Suffix 9:	t

Abbildung 2: Liste der Suffixe von  $S = \text{aaattttttt}$ . Suffix 0 ist die Sequenz selbst. Die Suffixnummer definiert den Anfang des Suffixes in der Sequenz.

## Was ist ein Suffixbaum?

Alle Suffixe einer Sequenz  $S$  können in eine Baumstruktur eingetragen werden, die Suffixbaum heißt. Abbildung 3 zeigt den Suffixbaum der Sequenz  $S = aaattttttt$ .

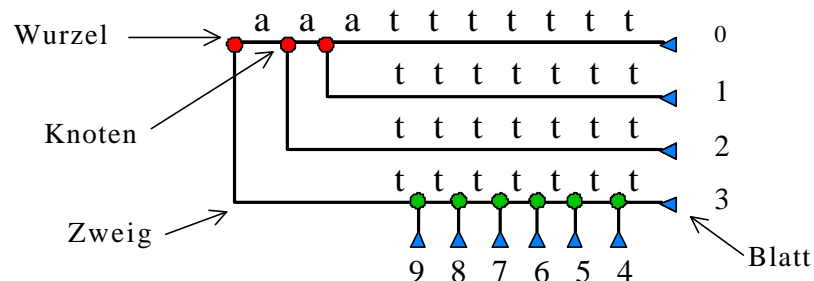


Abbildung 3: Suffixbaum der Sequenz  $S=aaattttttt$ . Rote Kreise kennzeichnen explizite Knoten; grüne Kreise implizite Knoten. Blätter sind durch ein Dreieck gekennzeichnet

Wenn wir von der Wurzel den Weg zu einem beliebigen Blatt nehmen und die auf den passierten Zweigen stehenden Nukleotide zu einer Teilsequenz zusammen fügen, erhalten wir den Suffix von  $S$ , dessen Nummer am Zielblatt steht. Prüfen Sie das nach.

Der Suffixbaum hat für jeden Suffix genau ein Blatt. Die Verzweigungen des Baumes heißen Knoten. Sie werden explizite Knoten genannt, wenn wenigstens zwei der den Knoten in Richtung der Blätter verlassenden Zweige Nukleotide tragen (rot). Implizite Knoten (grün) sind keine echten Verzweigungen. Sie markieren lediglich das Ende eines Suffix auf einem Zweig.

## Was sollen Suffixbäume?

Ein Suffixbaum ist hervorragend geeignet, um einen Text in einer Sequenz zu finden. Zum Beispiel wollen wir den Text  $T = att$  in der Sequenz  $S = aaattttttt$  suchen. Wir beginnen in Abbildung 3 an der Wurzel des Suffixbaums und wählen den Zweig, der mit dem Buchstaben  $a$  beginnt, gelangen zum nächsten Knoten, wählen dort den Zweig, der mit dem Buchstaben  $t$  beginnt und finden schließlich  $T = att$  auf diesem Zweig. Die Suffixnummer am Ende des Zweigs sagt uns, daß  $T$  in Position 2 von  $S$  auftritt. *Die Anzahl der notwendigen Buchstabenvergleiche war 3, also gleich der Anzahl der Buchstaben in  $T$ .* Im Fall des Texts  $T = ttt$  wählen wir von der Wurzel den Zweig, der mit dem Buchstaben  $t$  beginnt und finden  $T$  am impliziten Knoten mit der Suffixnummer 7. Alle Suffixnummern von unterhalb dieses Knotens liegenden weiteren Knoten bezeichnen alternative Positionen von  $T$  in  $S$ . Wenn wir den Suffixbaum eines vollständigen Chromosoms aufgebaut haben, können wir für jede Nukleotidsequenz der Länge  $n$  in nur  $n$  Schritten ermitteln, ob und wo sie überall im Chromosom enthalten ist.

## Wie baut man einen Suffixbaum?

Wir schildern zunächst ein einfach zu verstehendes aber zeitaufwendiges Verfahren. Es sei  $S=aaatttttt$  die Sequenz, von der wir den Suffixbaum erstellen wollen.

Der Suffix 0 ist die Sequenz selbst. Wir tragen ihn entlang des ersten Zweiges ein.

a a a t t t t t t t 0

Suffix 1 beginnt  $aat\dots$  und sein Eintrag, ausgehend von der Wurzel liefert den zweiten Zweig und den ersten Knoten.

a a a t t t t t t t 0  
 ↗ |  
 t t t t t t t t 1

Völlig analog tragen wir Suffix 2 und Suffix 3 ein.

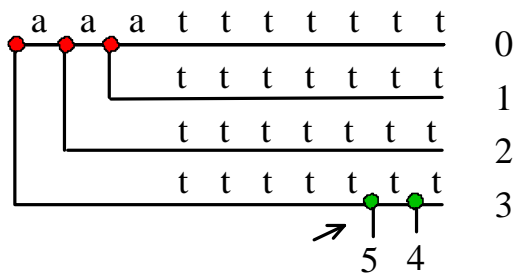
a a a t t t t t t t 0  
 ↗ |  
 t t t t t t t t 1  
 |  
 t t t t t t t t 2

↓ a a a t t t t t t t 0  
 | | |  
 t t t t t t t t 1  
 | | |  
 t t t t t t t t 2  
 | | | |  
 t t t t t t t t 3

Suffix 4 beginnt wie Suffix 3 mit  $ttt\dots$ . Sein Eintrag in den Suffixbaum endet inmitten eines Zweiges. Wir müssen deshalb einen impliziten Knoten hinzufügen, um das Suffixende zu bezeichnen.

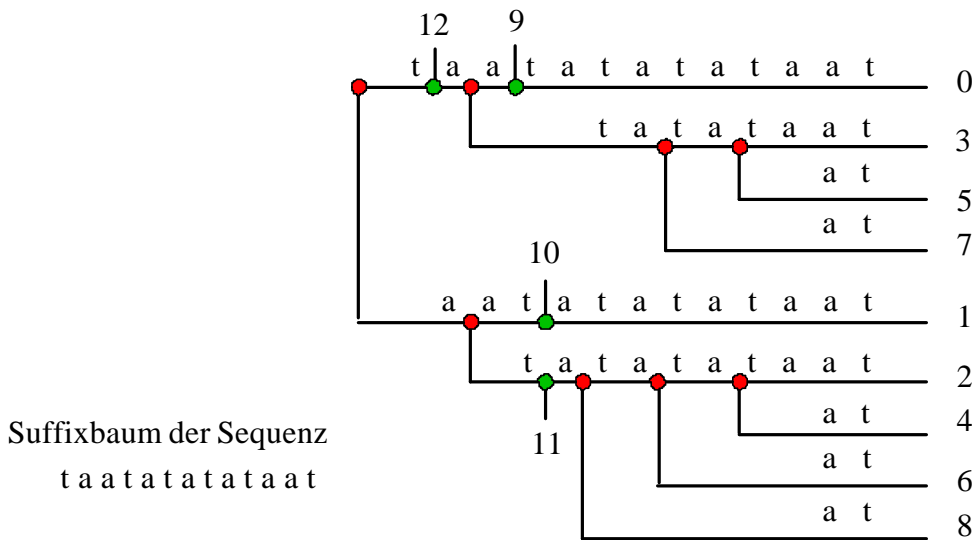
a a a t t t t t t t 0  
 | | |  
 t t t t t t t t 1  
 | | |  
 t t t t t t t t 2  
 | | | |  
 t t t t t t t t 3  
 ↗ |  
 4

Suffix 5 endet im Suffixbaum einen Buchstaben vor dem Ende von Suffix 4.



Es ist klar, wie fortgefahen werden muß, bis alle Suffixe in den Baum eingetragen sind. Erstellen Sie selbst den Suffixbaum der Sequenz "taataataataat".

Lösung: (soll erst nach Anklicken erscheinen)



### Wie groß ist ein Suffixbaum?

Wird der Suffixbaum einer langen Sequenz nicht riesig groß? Dauert es nicht ewig, alle Knoten eines solchen Baumes abzusuchen? Erstellen Sie noch einmal einen Suffixbaum. Achten Sie jetzt darauf, wie viele neue Knoten beim Hinzufügen jedes Suffixes entstehen. Sie werden beobachten, daß beim Eintragen jedes Suffixes höchstens ein neuer Knoten (explizit oder implizit) entsteht. Damit ist die Anzahl der Knoten in einem Suffixbaum nicht größer als die Sequenzlänge. Das ist eine kleine Zahl. Die Anzahl der expliziten Knoten kann noch einmal deutlich kleiner sein. Computer haben deshalb kein Problem, alle Knoten eines Suffixbaums nacheinander aufzusuchen.

### Wie lange dauert der Aufbau eines Suffixbaums?

Profis bauen Suffixbäume sehr viel schneller als zuvor beschrieben und brauchen auch nur wenig Platz, um sie abzuspeichern. Speicherplatz und Rechenzeit wachsen proportional, d.h. linear, mit der Länge der betrachteten Sequenz. Linear wachsender Speicherplatz ist leicht realisiert, wenn jeder Zweig des Baumes nicht mit der ihm zugeordneten Teilsequenz, sondern

nur mit dessen Anfangs- und Endposition in der betrachteten Sequenz beschriftet wird. Der Aufbau solcher Suffixbäume in linear wachsender Rechenzeit erfordert jedoch eine Reihe von raffinierteren Tricks. Wollen Sie wissen, wie das geht?

# Gesetzmäßigkeiten beim Aufbau von Suffixbäumen

Suffixbäume werden erstellt, indem man alle Suffixe der Reihe nach in einen Baum einträgt. Die Geschicklichkeit, mit der das geschieht, unterscheidet zwischen naiven und raffinierten Algorithmen. Die letzteren nutzen eine Reihe von Gesetzmäßigkeiten, um alle Suffixe in linear mit der Sequenzlänge wachsender Zeit einzutragen.

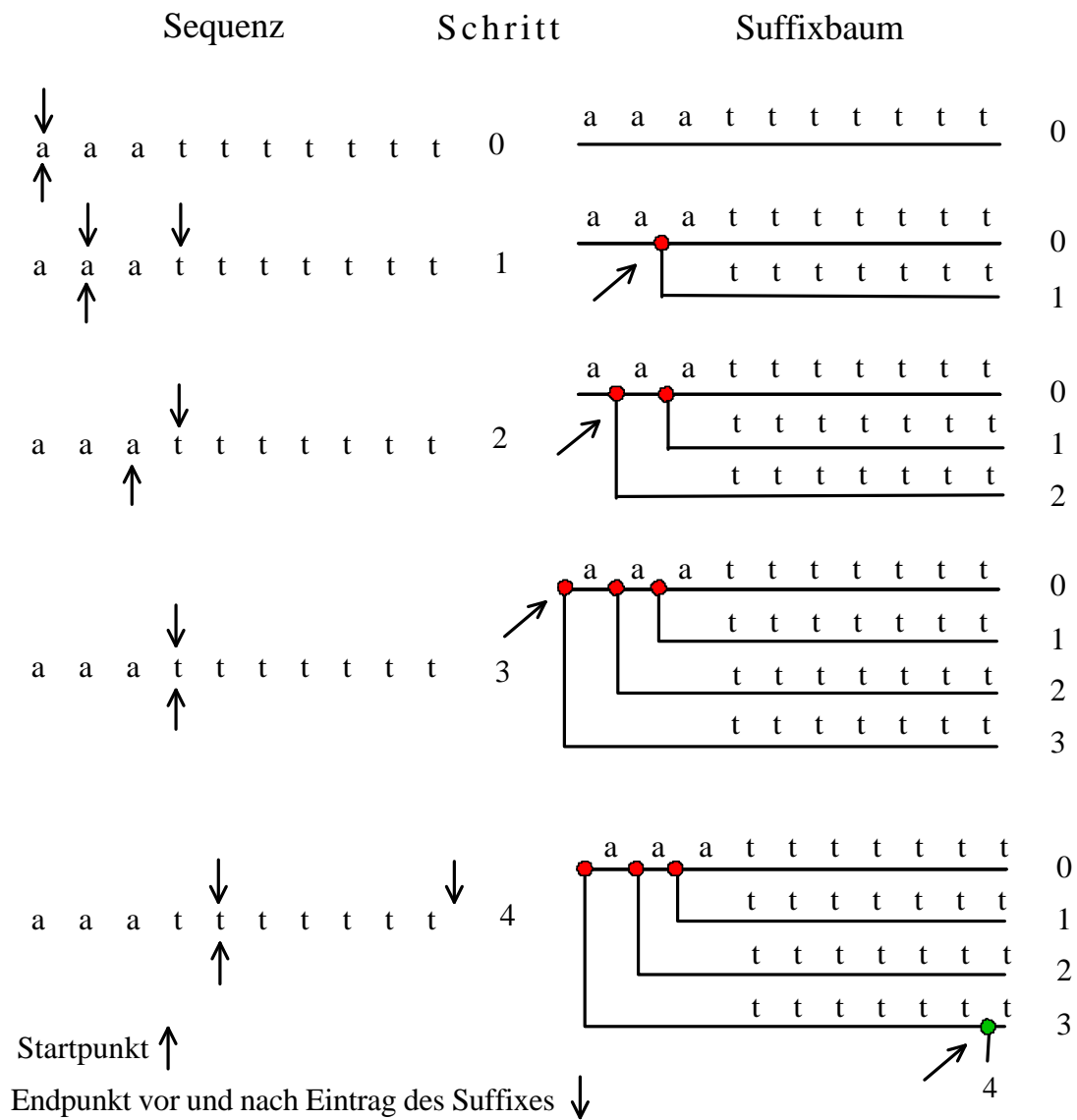


Abb. 4: Das Fortschreiten von Start- und Endpunkt beim Aufbau des Suffixbaumes in Abbildung 3. Der Pfeil zeigt auf den jeweils aktiven Knoten.

Wir nehmen an, daß bereits ein Suffixbaum existiert, der mindestens den Suffix 0 enthält. Zum Eintrag des nächsten Suffixes in den bestehenden Baum beginnen wir an der Wurzel und verfolgen Buchstabe für Buchstabe den neuen Suffix bis ein Konflikt auftritt. Das ist der Fall, wenn an einem Knoten kein Zweig mit dem passenden Buchstaben fortsetzt oder wenn inmitten eines Zweiges eine Differenz auftritt. In beiden Fällen müssen wir an dieser Stelle einen neuen

Zweig hinzufügen und ihn mit dem Rest des betrachteten Suffixes beschreiben. Das ist die zuvor beschriebene naive Methode. Sie ist der Ausgangspunkt unserer Überlegungen.

Zur Durchführung verwenden wir zwei Zeiger in die Sequenz. Den einen nennen wir Startpunkt. Er zeigt auf das erste Zeichen des einzutragenden Suffixes. Der andere heißt Endpunkt. Er zeigt auf das erste Zeichen des Suffixes, mit dem der zuletzt neu hinzugefügte Zweig beschriftet wurde. Der Knoten, von dem dieser Zweig ausgeht, heißt aktiver Knoten. Abbildung 4 zeigt, wie sich Startpunkt, Endpunkt und aktiver Knoten während der ersten 5 Schritte beim Bau des Suffixbaums für die Sequenz  $S = \text{aaattttttt}$  ändern. Der Endpunkt wandert im vierten Schritt hinter das Sequenzende, weil Suffix 4 bereits vollständig im Baum enthalten ist.

Start- und Endpunkt bezeichnen immer den Abschnitt des zuletzt eingetragenen Suffixes, der bereits im Baum enthalten war. Das ist der Teil des Suffixes, der mit dem bestehenden Baum verglichen werden mußte. Er kann allgemein sehr lang werden. Bei gut gemischten, z.B. stochastischen Sequenzen, ist dem aber eine Grenze gesetzt. Enthält eine Sequenz keine zwei identischen Teilsequenzen der Länge 20, so können Start- und Endpunkt nicht weiter als 20 Positionen auseinander rücken. Der Eintrag jedes Suffixes verlangt dann höchstens 20 Buchstabenvergleiche. Damit ist solch ein naiver Algorithmus oft schon recht schnell. Die Laufzeit steigt aber stark an, wenn die Sequenz nur aus einem einzigen Buchstaben besteht oder lange derartige Teilsequenzen enthält. Genomische Sequenzen enthalten oft hohe Anzahlen sich wiederholender Sequenzwörter. Sie führen ebenfalls zu langen Laufzeiten dieses naiven Verfahrens. Solchen Problemen kann zum Teil mit einer einfachen Beobachtung begegnet werden.

**Satz 1:        Nach jedem Schritt enthält der Suffixbaum alle Teilsequenzen,  
                 die vor dem Endpunkt beginnen und bis an den Endpunkt reichen.**

Erinnern wir uns. Zu jedem Zeitpunkt markieren Start- und Endpunkt den Abschnitt des zuletzt eingetragenen Suffixes, der bereits im Baum enthalten war und den wir mit  $A$  bezeichnen und Suffixanfang nennen. Deshalb ist  $A$  auch Anfang eines bereits früher eingetragenen Suffixes. Das folgt aus der Methode, mit der wir neue Suffixe eintragen und den Endpunkt festlegen. Die betrachtete Sequenz enthält also eine Kopie von  $A$ , die vor dem Startpunkt beginnt. Das Gleiche gilt auch für jeden Suffix von  $A$ , der deshalb auch schon früher einmal in den Suffixbaum eingetragen wurde. Das ist ganz klar, wenn die Kopie des Suffixes von  $A$  vor dem Startpunkt beginnt. Anderenfalls ist die Argumentation ist etwas aufwendiger, wie in Abbildung 5 dargestellt. Der zuletzt eingetragene Suffixanfang ist dort  $A = \text{taaataaat}$ . Ein Suffix  $\text{aaat}$  von  $A$  (grün), dessen Abbild in der Kopie von  $A$  nicht vor dem Startpunkt beginnt (rot), muß auch in einer früheren Position von  $A$  existieren (blau), dessen Abbild in der Kopie von  $A$  vor dem Startpunkt beginnt und damit schon früher in den Baum eingetragen wurde.

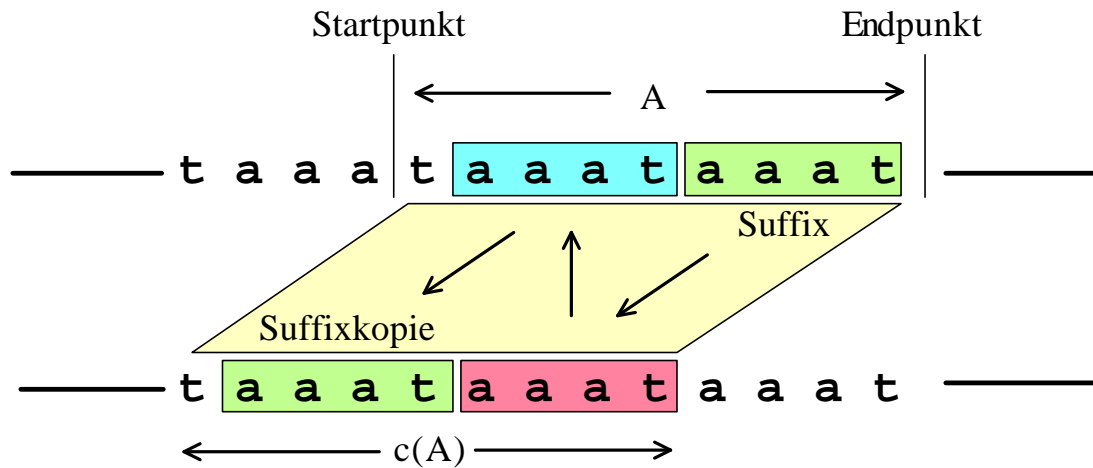


Abbildung 5: Der zuletzt eingetragene Suffixanfang A besitzt eine früher in der Sequenz liegende Kopie  $c(A)$ , die mit A überlappt. Beginnt die Kopie eines Suffixes von A (rot) hinter dem Startpunkt, so gibt es weitere Kopien dieses Suffixes, von denen wenigstens eine vor dem Startpunkt beginnt.

Durch Anwendung von Satz 1 können wir sehr effektiv den Suffixbaum einer beliebig langen Sequenz erstellen, die nur aus Wiederholungen eines einzigen Buchstabens besteht.

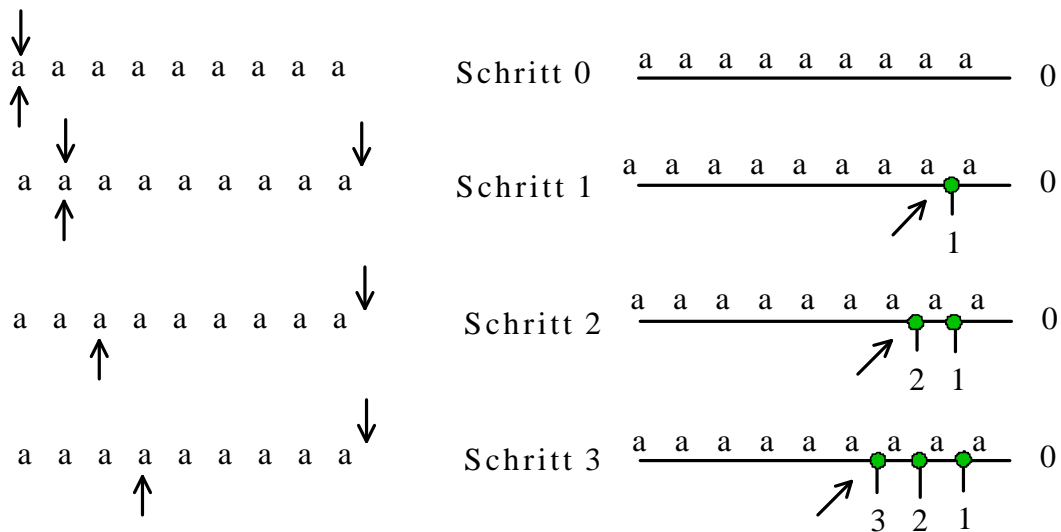


Abbildung 6: Aufbau des Suffixbaums der Sequenz  $S = \text{aaaaaaaaaaaa}$ .

Im Schritt 0 tragen wir den Suffix 0 ein. Im folgenden Schritt stellen wir fest, daß Suffix 1 bereits im Baum enthalten ist. Der Endpunkt wandert hinter das Sequenzende. Nun wissen wir, daß alle folgenden Suffixe bereits im Baum enthalten sind. Da der für den Eintrag in Frage kommende Zweig jeweils länger als der einzutragende Suffix ist, kann das Ende des Suffixes auf dem Zweig berechnet werden. Deshalb erfolgt jeder Eintrag in konstanter Zeit und der Aufwand zum Bau des Suffixbaums ist proportional zur Sequenzlänge.

Ebenso schnell kann der in Abbildung 7 gezeigte Suffixbaum erstellt werden.

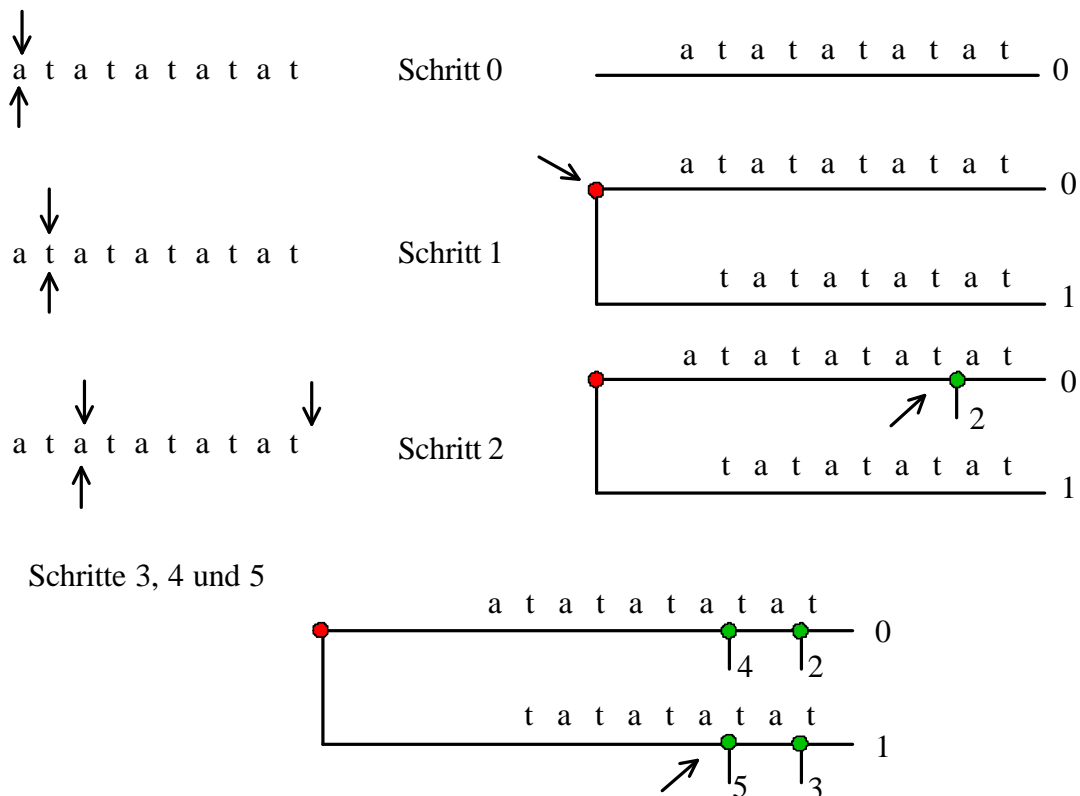


Abbildung 7: Aufbau des Suffixbaums für die Sequenz  $S=atataatataat$

Nachdem Schritt 1 nur einen Buchstabenvergleich erforderte, ist in Schritt 2 der Suffix 2 bereits im Baum enthalten. Zum Eintrag aller folgenden Suffixe vergleichen wir lediglich den ersten Buchstaben, um den richtigen Zweig zu wählen, und berechnen dann das Ende des Suffixes auf dem Zweig. Die Gesamtanzahl der erforderlichen Buchstabenvergleiche ist gleich der Sequenzlänge.

Die Anwendung von Satz 1 führt ganz allgemein zu einer Beschleunigung des Aufbaus von Suffixbäumen. Um die Teilsequenz vom Startpunkt bis an den Endpunkt im Suffixbaum zu finden, können wir von Zweig zu Zweig springen, ohne alle Buchstaben einzeln entlang der Zweige vergleichen zu müssen. Beginnt ein Zweig mit dem richtigen Buchstaben, so setzt er auch bis zum nächsten Knoten oder den Endpunkt richtig fort. Das folgt, weil beginnend an der Wurzel, jede Teilsequenz nur genau einmal im Suffixbaum enthalten ist. Erst ab dem Endpunkt muß Buchstabe für Buchstabe verglichen und der Endpunkt verschoben werden bis eine

Differenz auftritt und das Eintragen des neuen Suffixes beendet wird. *Da der Endpunkt immer nur vorwärts schreitet ist die damit verbundene Anzahl von Buchstabenvergleichen gleich der Sequenzlänge.* Der zeitkritische Teil des Algorithmus ist deshalb das Aufsuchen der Teilsequenz zwischen Start- und Endpunkt im bereits vorhandenen Baum. Dabei springen wir von Knoten zu Knoten, deren Anzahl den Aufwand zum Eintragen eines neuen Suffixes bestimmt. Für Sequenzen, die sich aus wiederholenden langen Worten zusammensetzen, führt dieser Trick zu einer erheblichen Leistungssteigerung. Mit der Methode der Suffixzeiger kann jedoch auch noch das Absuchen der Knoten von der Wurzel des Baumes her weitgehend vermieden werden  
(Link Suffixzeiger).

# Suffixzeiger

Eine raffinierte Methode zum Eintragen der Suffixe benutzt Suffixzeiger. Betrachten wir die Abbildung 8. Sie zeigt in Teil i) einen einzutragenden Suffixanfang und in Teil ii) die entsprechende Situation im betreffenden Ausschnitt des Suffixbaums.

Der Suffix beginnend am Startpunkt wurde bis zum Endpunkt in den Baum eingetragen. Am Ende des Suffixanfangs  $A = \text{taaataaat}$  wurde ein neuer mit dem Buchstaben  $t$  beginnender Zweig eingerichtet, der den Suffix repräsentiert. Dieser Zweig wurde einem bereits vorhandenen Knoten  $X$  hinzugefügt oder es wurde ein neuer Knoten  $X$  zur Einrichtung dieses Zweigs geschaffen.

Nach Satz 1 tritt der gleiche Suffixanfang  $A$  bereits früher in der Sequenz auf (Kopie des Suffixanfangs). Er wurde aber von einem anderen Sequenzbuchstaben gefolgt, da sonst der Endpunkt weiter rechts liegen würde. In der Abbildung ist es der Buchstabe  $a$ .

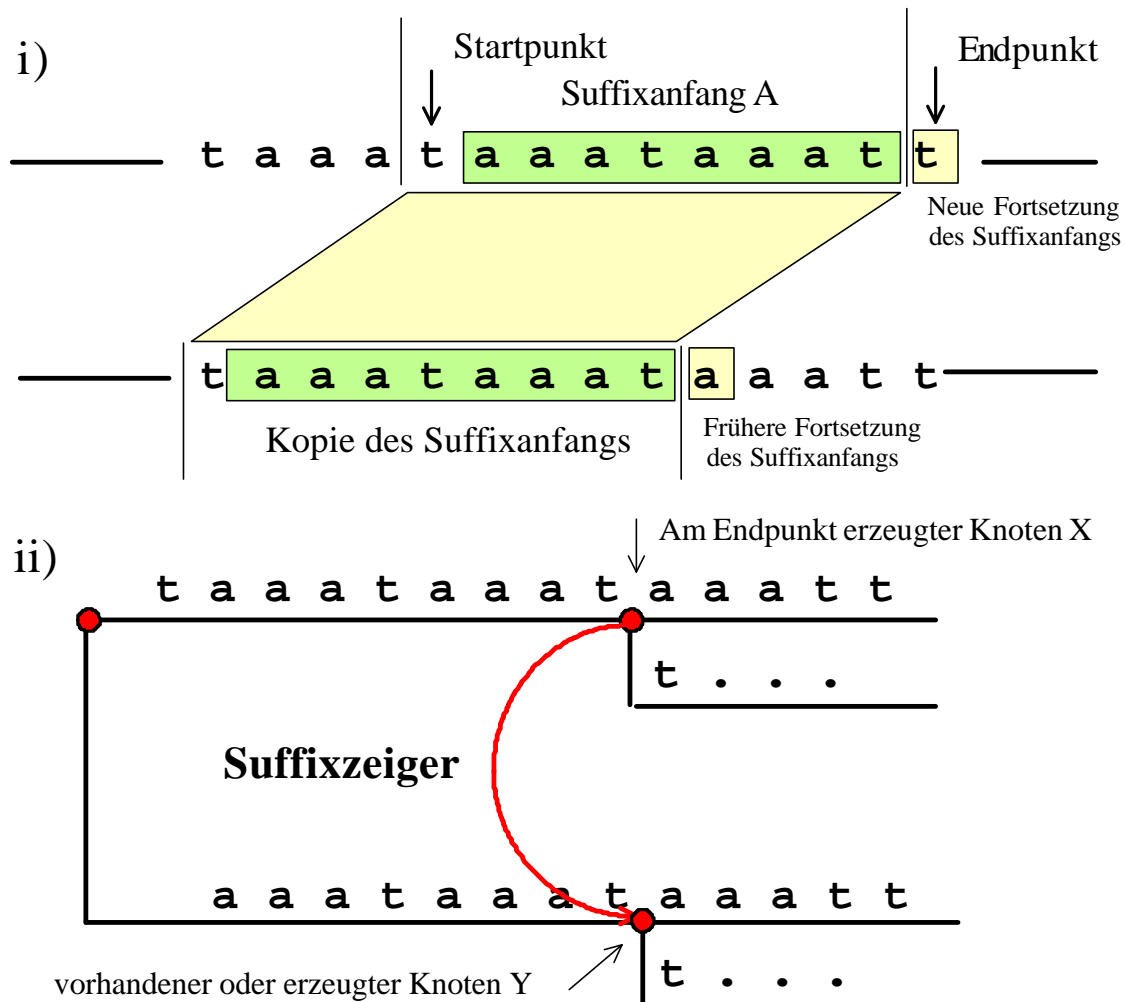


Abbildung 8: Die Existenz oder Einrichtung des Knotens  $X$  beim Eintragen der Teilsequenz zwischen Start- und Endpunkt führt im nächsten Schritt notwendigerweise zum Knoten  $Y$  (siehe Text).

Wir wollen nun den Anfang des nächsten Suffix (grüner Kasten) in den Baum eintragen. Auch dieser trat bereits früher in der Sequenz auf. Er ist deshalb im Baum enthalten und wird auch von dem Buchstaben a gefolgt. Beim Eintrag des neuen Suffixanfangs (grüner Kasten) gelangen wir deshalb am Endpunkt an eine Stelle, von der aus bereits mit dem Buchstaben a fortgesetzt werden kann. Entweder ist a der nächste Buchstabe auf einem Zweig, oder es existiert bereits ein Knoten Y, von dem wenigstens ein mit a beginnender Zweig ausgeht. Wir müssen aber an dieser Stelle mit dem Buchstaben t fortfahren, d.h. den Knoten Y einrichten, falls er noch nicht vorhanden ist.

Der Knoten Y steht in einer wichtigen Beziehung zum zuvor betrachteten Knoten X. An ihm endet der erste Suffix von A, d.h. von der Sequenz, die am Knoten X endet. Genau diesen wollen wir im folgenden Schritt im Suffixbaum aufsuchen. Deshalb merken wir uns diese Beziehung durch Einrichtung eines Zeigers am Knoten X, der auf den Knoten Y gerichtet ist und Suffixzeiger genannt wird. Wenn wir das konsequent tun, folgt das wichtige Resultat.

**Satz 2:** **Jeder Knoten, mit eventueller Ausnahme des aktiven Knoten, besitzt einen Suffixzeiger.**

Wird ein schon vorhandener Knoten zum aktiven Knoten, so besitzt er bereits einen Suffixzeiger. Wird der aktive Knoten neu erstellt, so besitzt er noch keinen Suffixzeiger, bekommt ihn aber im nächsten Schritt.

Suffixzeiger erlauben das Aufsuchen von Suffixanfängen in kürzerer Zeit. Diese Methode ist in Abbildung 9 schematisch dargestellt. Anstatt einen neuen Suffixanfang vom der Wurzel des Suffixbaums her zu suchen, beginnen wir am zuletzt besuchten, dem aktiven Knoten X. War dieser bereits vorhanden, dann besitzt er einen Suffixzeiger und wir folgen diesem (grün) direkt zum Endpunkt des nächsten Suffixanfangs im Knoten Y. Anderenfalls merken wir uns die Teilsequenz aa, mit der der Zweig zum Knoten X beschriftet ist, springen an seinen Anfang, den Knoten Z1, dann mit Hilfe des Suffixzeigers (rot) zum Knoten Z2 und verfolgen von dort die gemerkte Teilsequenz aa im bestehenden Suffixbaum. So kürzen wir beim Eintragen des nächsten Suffixes den unter Umständen langen Weg von der Wurzel des Baumes ab.



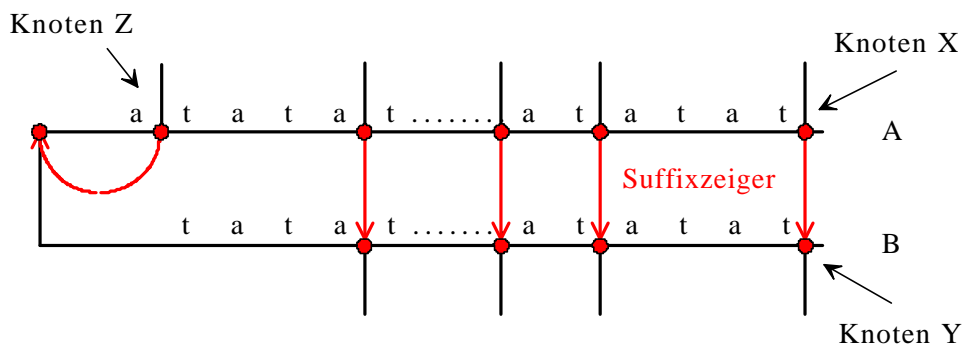


Abbildung 10: Schematische Darstellung des im Satz 4 beschriebenen Zusammenhangs zwischen den aktuellen Knotentiefen von durch Suffixzeiger verbundenen Knoten.

In Abbildung 10 haben wir einen Teil eines Suffixbaums dargestellt. Es ist X der Knoten am Ende einer Teilsequenz A, der einen Suffixzeiger zum Knoten Y besitzt. Wir nehmen an, daß alle dem Knoten X vorangehenden Knoten bereits Suffixzeiger besitzen. Diese zeigen notwendigerweise auf verschiedene Knoten auf dem Weg von der Wurzel des Baumes zum Ende der Teilsequenz B, dem ersten Suffix von A. Nur der Suffixzeiger eines unmittelbar auf die Wurzel folgenden Knotens Z zeigt möglicherweise auf diese. Deshalb gilt:

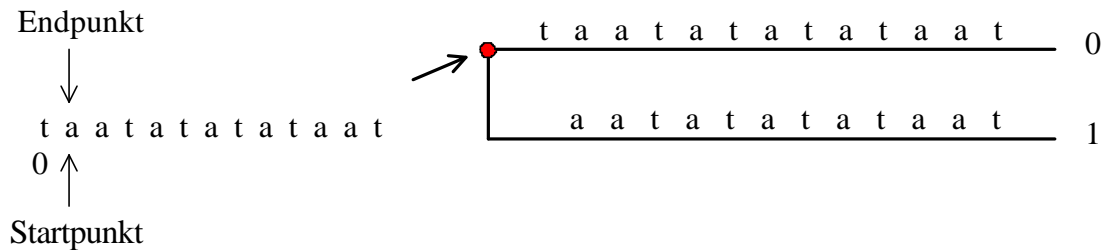
**Satz 4:** **Zeigt der Suffixzeiger eines Knotens X auf den Knoten Y und besitzen alle Elternknoten von X einen Suffixzeiger, so ist die Knotentiefe von X höchstens eins größer als die aktuelle Knotentiefe von Y.**

Mit diesem Resultat können wir eine einfache Bilanz für die aktuelle Tiefe des aktiven Knotens X in Abbildung 9 ziehen. Besitzt er einen Suffixzeiger, so besitzen nach Satz 2 alle Knoten des Baumes diese Eigenschaft und nach Satz 4 ist die aktuelle Knotentiefe von Y höchstens um eins kleiner als die von X. Besitzt X keinen Suffixzeiger, so erfüllt Knoten Z1 die Bedingung von Satz 4. Die aktuelle Knotentiefe vermindert sich um höchstens zwei beim Übergang von X nach Z2, steigt dann aber um mindestens eins beim Übergang von Z2 nach Y. In beiden Fällen ist die aktuelle Knotentiefe von Y, einem Elternknoten des neuen aktiven Knoten, höchstens eins kleiner als die vom alten aktiven Knoten X. Sie steigt aber zusätzlich mit jedem Zwischenknoten Z3, der auf dem Weg von Z2 nach Y liegt. Wären es in jedem Schritt des Verfahrens mehr als 2 zusätzliche Knoten Z3, so würde die aktuelle Tiefe des neuen aktiven Knotens Y sich in jedem Schritt um mindestens 2 vergrößern und schnell die Sequenzlänge überschreiten. Das ist unmöglich. Um dieses Ereignis auszuschließen, darf die Gesamtanzahl der in allen Verfahrensschritten zusätzlich besuchten Knoten Z3 nicht das Dreifache der Sequenzlänge überschreiten. Das ist nur eine sehr grobe obere Schranke. Die tatsächliche Anzahl der besuchten Knoten Z3 ist meist viel kleiner. Aber selbst im ungünstigsten Fall ist der durch die Knoten Z3 insgesamt verursachte Rechenaufwand proportional zur Sequenzlänge.

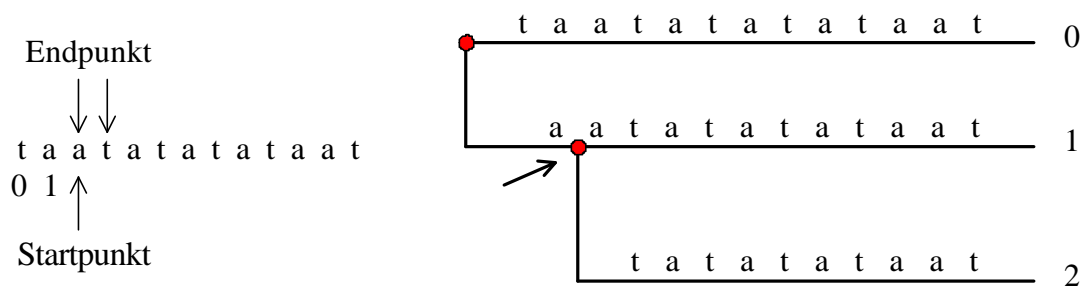
## So erstellen wir den Suffixbaum ganz schnell

Wir betrachten als Beispiel die Sequenz  $S=taatatataat$  und erstellen den Suffixbaum unter Verwendung von Suffixzeigern. Für jeden Schritt gibt es eine Abbildung. Sie erlaubt die Überprüfung aller Behauptungen in den Sätzen 1 bis 4. Vergewissern Sie sich. Nur so erfahren Sie, ob der Inhalt der Sätze richtig verstanden wurde. In den Abbildungen zeigen die beiden mit Endpunkt beschrifteten Pfeile auf den Endpunkt vor und nach dem Eintrag eines Suffixes.

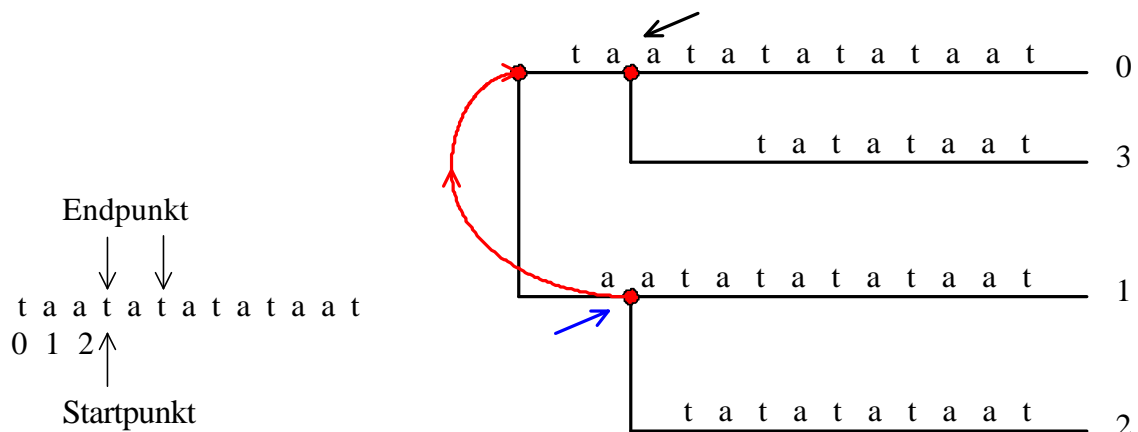
Wir nehmen an, daß Suffix 0 bereits in den Baum eingetragen ist.



Zum Eintrag von Suffix 1 beginnen wir an der Wurzel und müssen bereits dort einen neuen Zweig einrichten. Der aktive Knoten (schwarzer Pfeil) ist die Wurzel.

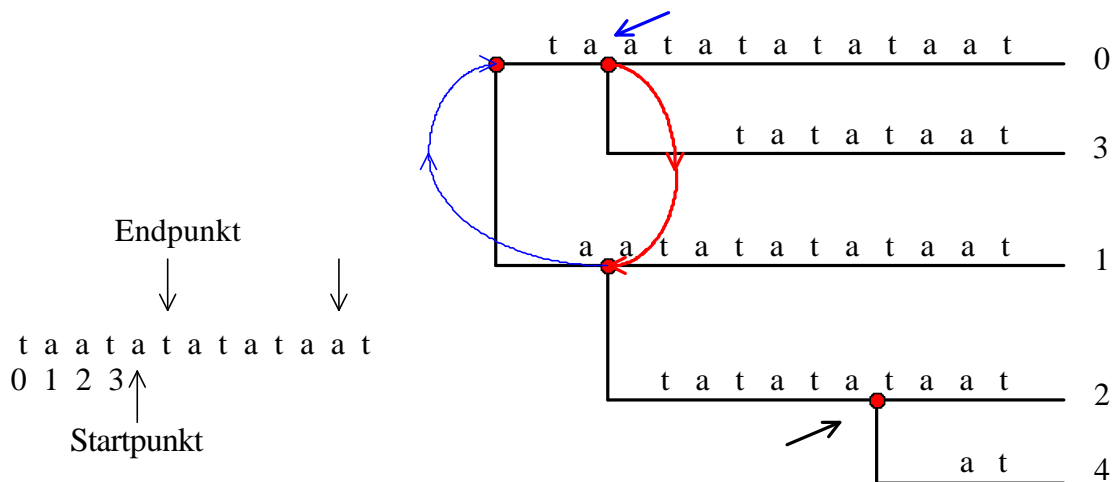


Zum Eintrag von Suffix 2 beginnen wir deshalb an der Wurzel, finden den Buchstaben a bereits vor und errichten den neuen aktiven Knoten (schwarzer Pfeil). Da der Buchstabe a bereits im Baum enthalten war, bewegt sich der Endpunkt einen Platz vorwärts. Um die Aussage von Satz 2 zu rechtfertigen, stellen wir uns vor, daß die Wurzel einen Suffixzeiger auf sich selbst besitzt.

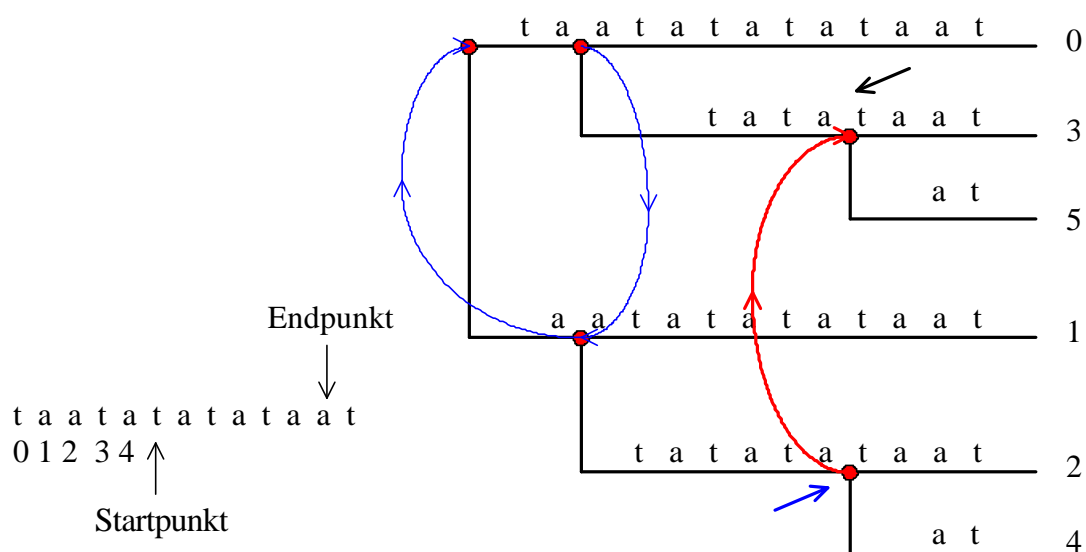


Immer wenn Endpunkt und Startpunkt übereinstimmen, müssen wir den Eintrag des neuen Suffixes an der Wurzel beginnen und einen Suffixzeiger vom alten aktiven Knoten (blauer Pfeil) zur Wurzel eintragen (rot). Wir finden beim Eintrag von Suffix 3 die Buchstaben t und a bereits

im Baum und richten den neuen aktiven Knoten (schwarzer Pfeil) ein. Der Endpunkt rückt zwei Positionen vor.

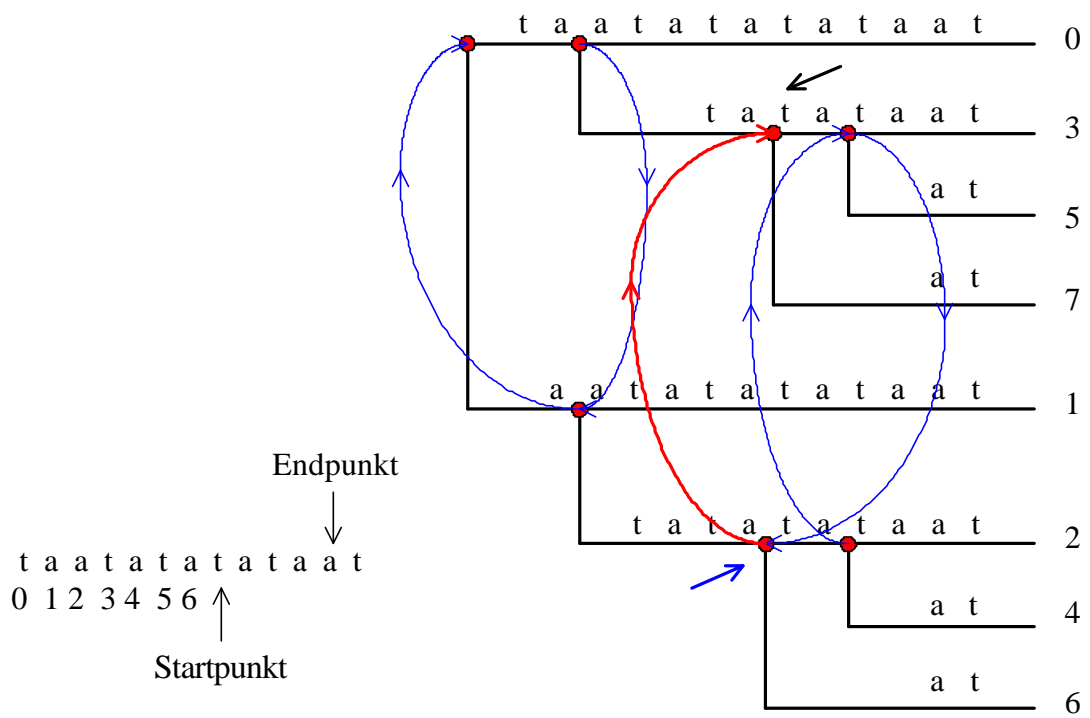
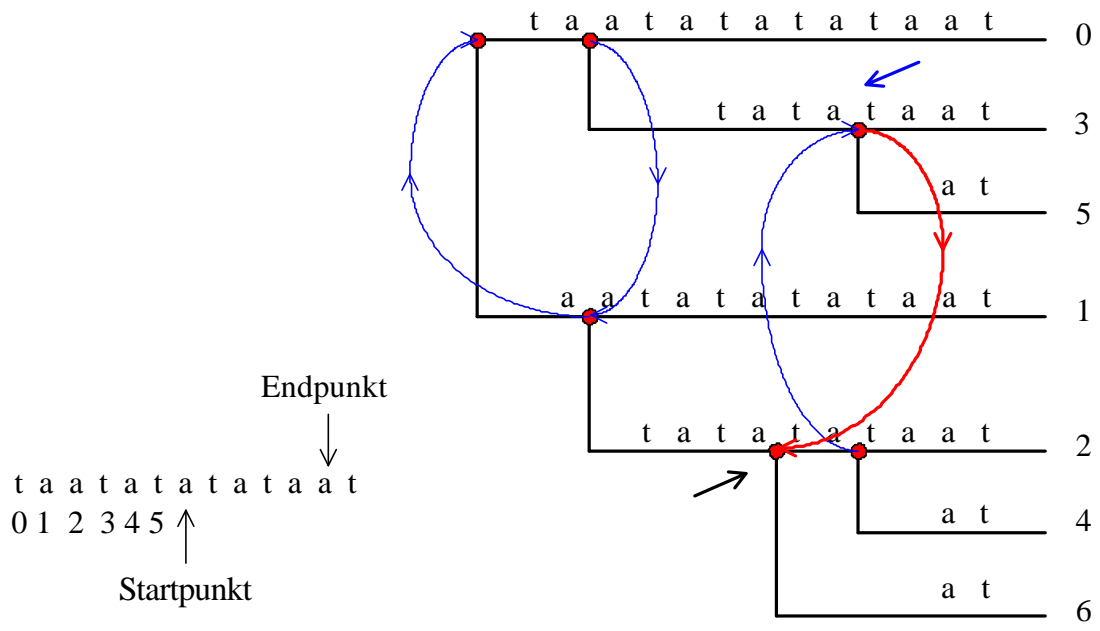


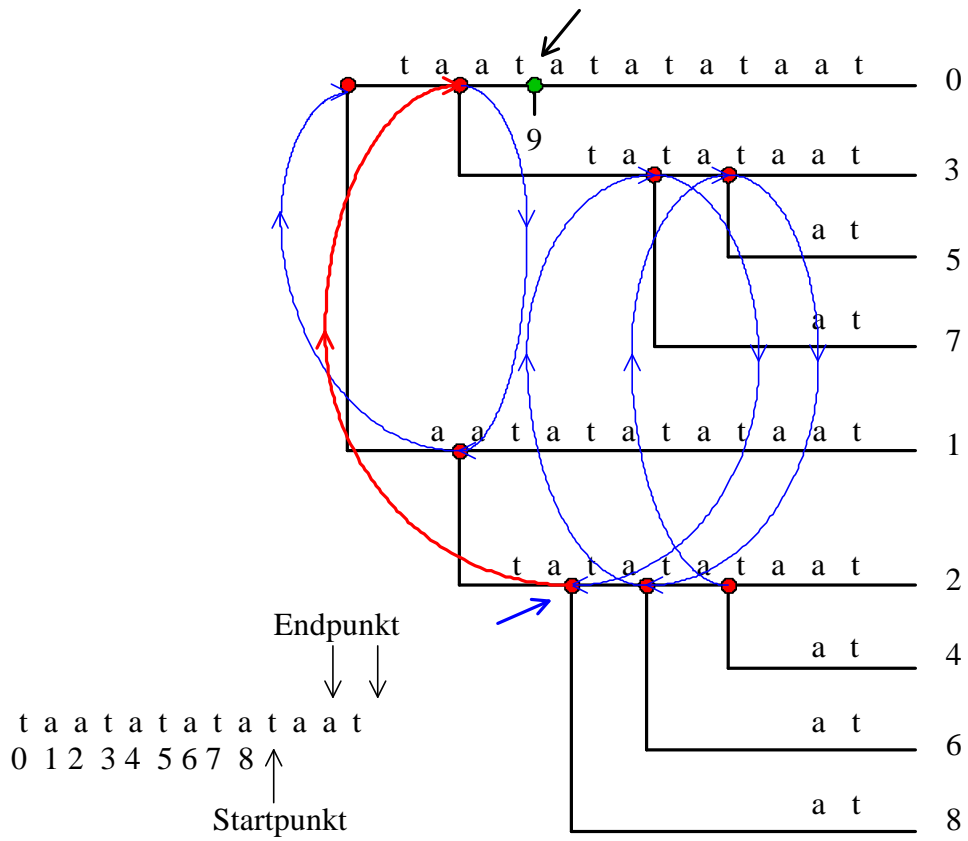
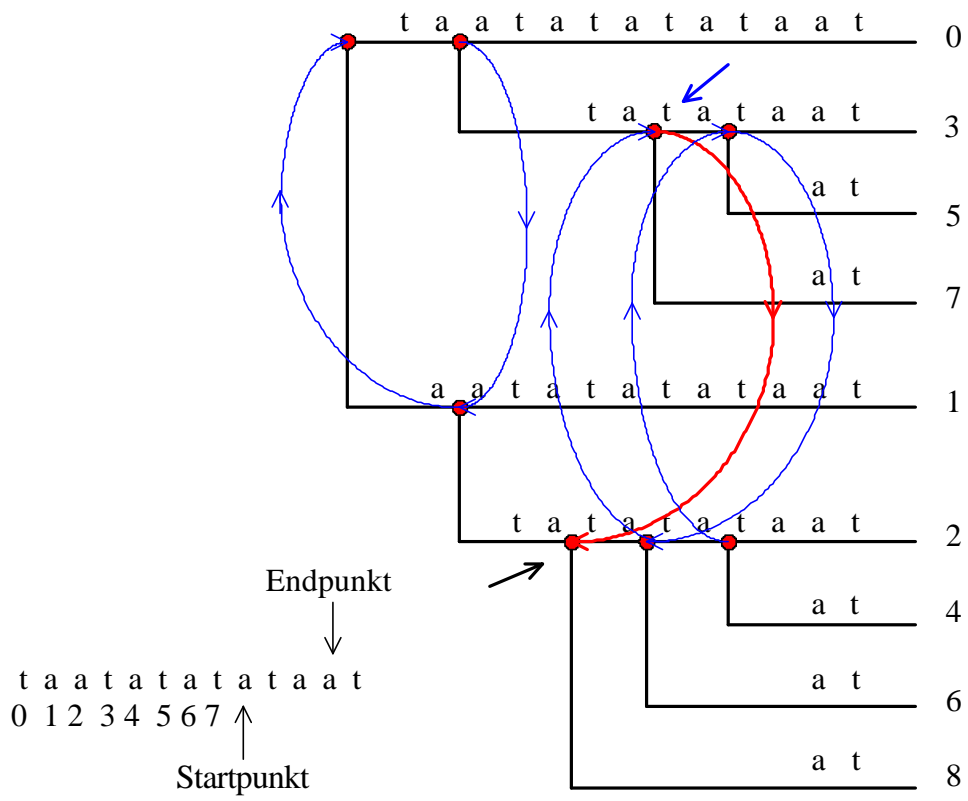
Zum Eintrag von Suffix 4 beginnen wir am alten aktiven Knoten (blauer Pfeil). Er besitzt noch keinen Suffixzeiger. Wir müssen deshalb zum Elternknoten zurück gehen. Immer wenn dieser die Wurzel ist, merken wir uns nur den ersten Suffix der Teilsequenz auf dem Zweig von der Wurzel zum alten aktiven Knoten. In unserem Fall ist es der Buchstabe a. Sein Eintrag in den Baum endet an einem schon vorhandenen Knoten. Wir tragen den Suffixzeiger (roter Pfeil) ein. Dann folgen wir dem Baum auf dem Zweig zur Suffixnummer 2. Wir finden einen großen Teil von Suffix 4 dort vor, errichten den neuen aktiven Knoten (schwarzer Pfeil) und einen neuen Zweig, der Suffix 4 repräsentiert. Der Endpunkt ist weit an das Sequenzende gerückt.



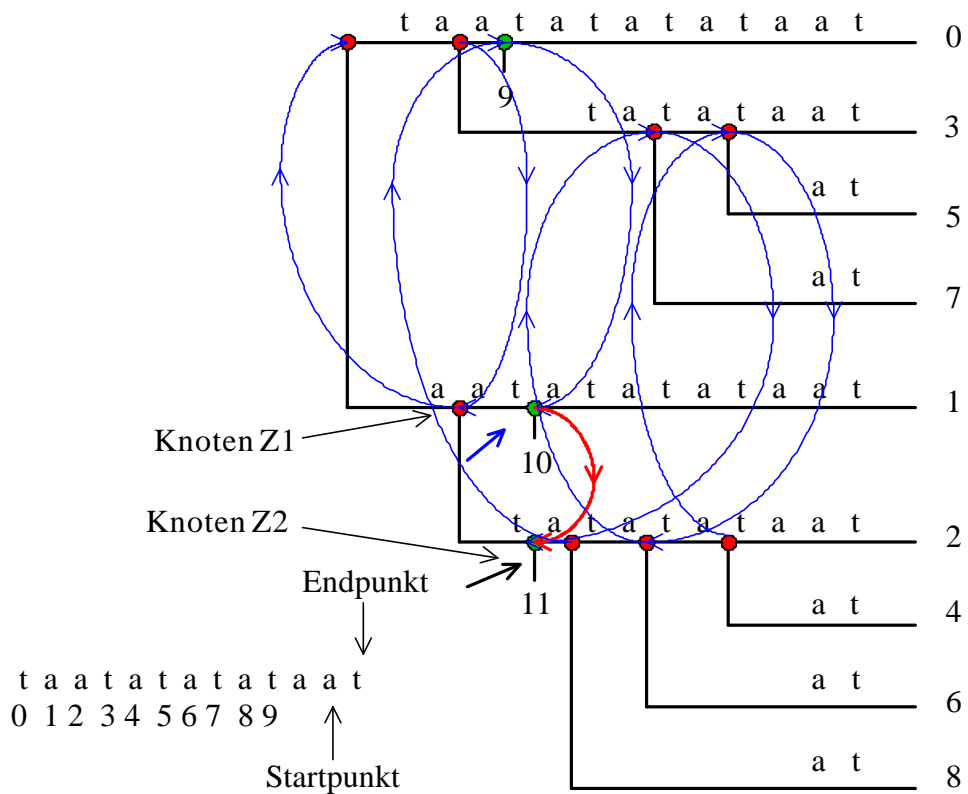
Zum Eintrag von Suffix 5 merken wir uns die Teilsequenz *tatata* auf dem Zweig zum zuletzt aktiven Knoten (blauer Pfeil), nutzen den Suffixzeiger vom Elternknoten zur Wurzel und tragen von dort beginnend die gemarkte Teilsequenz *tatata* in den Baum ein. Sie endet am neuen aktiven Knoten (schwarzer Pfeil), den wir einrichten. Es entsteht der rote Suffixzeiger.

Das Eintragen der Suffixe 6, 7 und 8 erfolgt völlig analog. Versuchen Sie es selbst. Der Startpunkt rückt in jedem Schritt näher an den Endpunkt. Wir zeigen die entsprechenden Abbildungen unkommentiert.









Zum Eintragen von Suffix 11 gehen wir vom aktiven Knoten (blauer Pfeil) zum Knoten Z1, folgen dem Suffixzeiger zur Wurzel und tragen von dort a t ein. Wir erhalten den impliziten Knoten Z2, der zum aktiven Knoten wird und erstellen den neuen Suffixzeiger (rot).

