V4 – Analyse von Genomsequenzen

- **Gene identifizieren**

  Intrinsische und Extrinsische Verfahren:

  Homologie bzw. Hidden Markov Modelle

- **Transkriptionsfaktorbindestellen** identifizieren

  Position Specific Scoring Matrices (PSSM)

- Ganz kurz: finde **Repeat-Sequenzen**

  Suche nach bekannten Repeat-Motiven

- **Mapping** von **NGS-Daten** auf **Referenzgenom**

- **Alignment zweier Genom-Sequenzen**

  Suffix Bäume

Identifikation von Genen

Die **einfachste** Methode, DNA Sequenzen zu finden, die für Proteine kodieren, ist nach **offenen Leserahmen** (**open reading frames** oder ORFs) zu suchen.

In jeder Sequenz gibt es 6 mögliche offene Leserahmen:
3 ORFs starten an den Positionen 1, 2, und 3 und gehen in die 5‘ 3‘ Richtung,
3 ORFs starten an den Positionen 1, 2, und 3 und gehen in die 5‘ 3‘ Richtung des komplementären Strangs.

In prokaryotischen Genomen werden Protein-kodierende DNA-Sequenzen gewöhnlich in mRNA transkribiert und die mRNA wird ohne wesentliche Änderungen direkt in einen Aminosäurestrang übersetzt.

Daher ist der längste ORF von dem ersten verfügbaren Met codon (**AUG**) auf der mRNA, das als **Codon** für den **Transkriptionsstart** fungiert, bis zu dem **nächsten Stopcodon** in demselben offenen Leserahmen,
gewöhnlich eine gute Vorhersage für die Protein-kodierende Region.

Vorhersage von Genen in Genomsequenzen

Etwa die Hälfte aller Gene kann durch Homologie zu anderen bekannten Genen oder Proteinen gefunden werden („**extrinsische Methode**“).

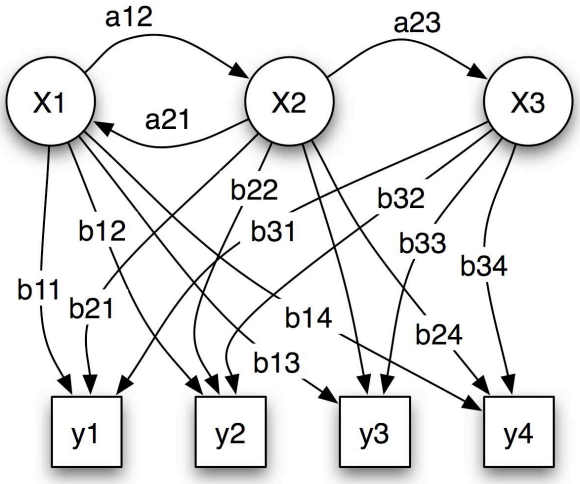
Dieser Anteil wächst stetig, da die Anzahl an sequenzierten Genomen und bekannten cDNA/EST Sequenzen kontinuierlich wächst.

Um die übrige Hälfte an Genen zu finden, muss man Vorhersage-Methoden einsetzen („**intrinsische Methoden**“),
die an einem Goldstandard-Datensatz mit bekannten Genen trainiert wurden.

Hidden Markov Modell (HMM)

Ein Hidden Markov Modell ist ein Graph, der verschiedene Zustände verbindet.

Im Modell rechts gibt es 3 **„verborgene“ Zustände**: X1, X2, X3.
In unserem Fall sind dies Bereiche der DNA,
z.B. Intron, Promoter, Exon.



Zwischen den Zuständen X1 und X2 und zurück und von X2 nach X3 sind hier Übergänge erlaubt.

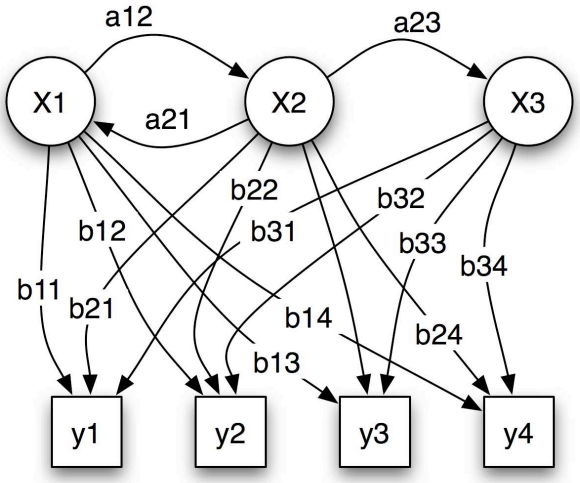
Die Übergangswahrscheinlichkeiten hierfür sind a12, a21 und 23.

y1 bis y4 sind die möglichen (sichtbaren) Output-Zustände.
Im Fall der Gen-Vorhersage also die Beobachtung, ob die entsprechenden DNA-Abschnitte als mRNA-Sequenzen exprimiert werden oder nicht.

Die Output-Zustände werden aus den verborgenen Zuständen mit den Wahrscheinlichkeiten b11 bis b34 erzeugt.

Hidden Markov Modell (HMM)

Die Topologie des Graphen gibt an, zwischen welchen Zuständen Übergänge erlaubt sind.
Diese gibt man bei der Spezifikation des HMM vor.



Jeder Übergang hängt nur von den beiden Zuständen *i* und *j* ab, zwischen denen der Übergang stattfindet, nicht von früheren Zuständen.

(Diese Eigenschaft gilt allgemein für Markov-Modelle)

Die Übergangswahrscheinlichkeiten aij und bij müssen in der Trainingsphase des HMM hergeleitet werden.

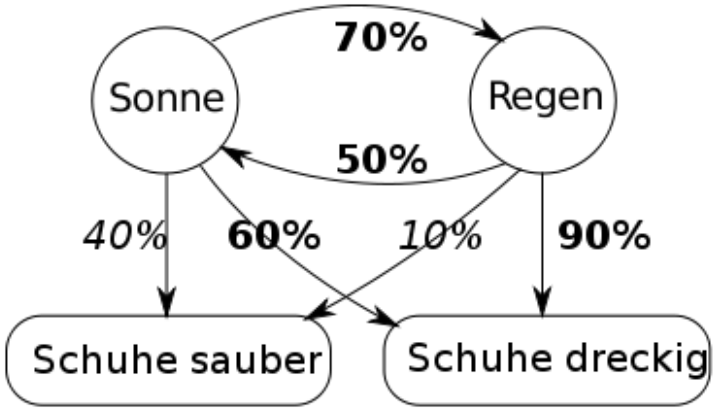
Ein HMM besteht also aus der Topologie und den trainierten Wahrscheinlichkeiten.

Wettervorhersage mit Hidden Markov Modell

Ein Gefangener im Kerkerverlies möchte das aktuelle Wetter herausfinden.

Er weiß, dass auf einen sonnigen Tag zu 70 % ein Regentag folgt und dass auf einen Regentag zu 50 % ein Sonnentag folgt.

Weiß er zusätzlich, dass die Schuhe der Wärter bei Regen zu 90 % dreckig, bei sonnigem Wetter aber nur zu 60 % dreckig sind, so kann er durch Beobachtung der Wärterschuhe Rückschlüsse über das Wetter ziehen.



Verborgener Zustand

Beobachtung

www.wikipedia.de

Generkennung mit Hidden Markov Modellen

Bei der Generkennung möchte man bestimmen, wo in einem Genom Exons (E) und Introns (I) sind.

Der Output sind die bekannten exprimierten Sequenzen.

Aus dieser soll jedem Basenpaar der günstigste verborgene Zustand (E/I) zugeordnet werden.

Bei Markov-Modelle hängt der Zustand des *i*-ten Buchstaben nur von seinem direkten Vorgänger, dem (*i* -1)- ten Buchstaben ab.

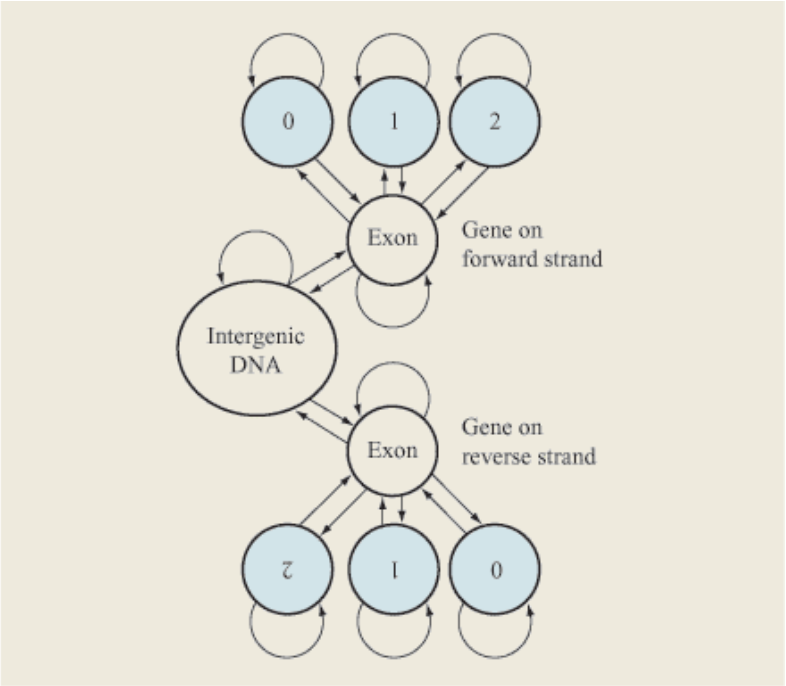


Figure 2

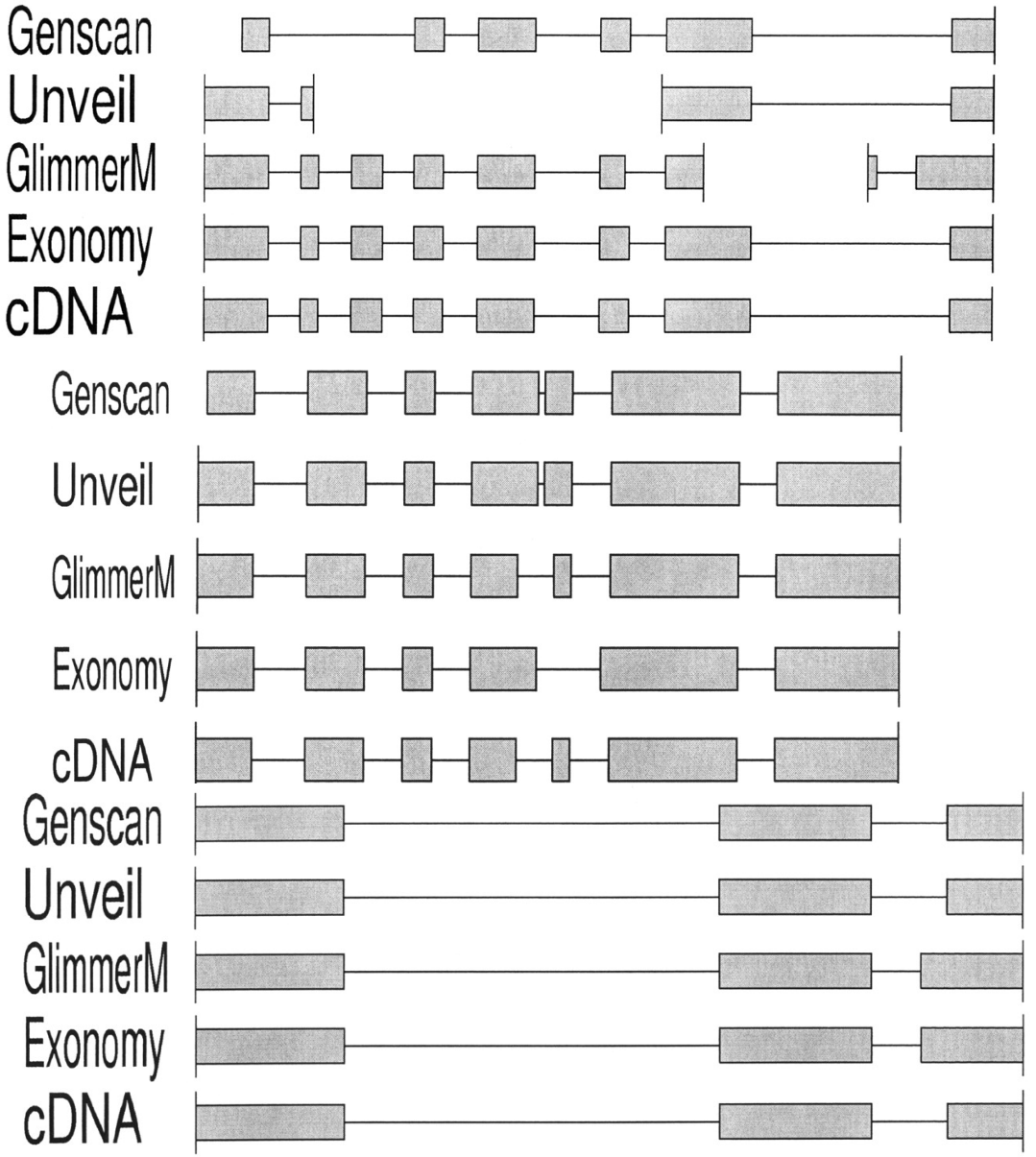
Abbreviated gene HMM model. The HMM is split into two symmetrical parts: genes on the forward or reverse strand of the DNA sequence (DNA sequence can be read in two directions). Each gene model contains a central exon state which has an emission of nucleotides tuned to recognize protein coding regions. Interrupting the exons are introns; three intron states are used, since there are three relative positions at which an intron can interrupt a coding triplet of DNA bases. These introns are distinguished by their "phase" — 0, 1, or 2.

Vergleich von Genvorhersage-Methoden

Ein Beispiel, in dem Exonomy die Gene richtig erkennt.

Ein Beispiel, in dem GlimmerM die Gene richtig erkennt.

Ein Beispiel, in dem Unveil die Gene richtig erkennt (auch Genscan).



Majoros et al. Nucl. Acids. Res. 31, 3601 (2003)

Promotervorhersage in *E.coli*

Um *E.coli* Promoter zu analysieren kann man eine Menge von Promoter-sequenzen bzgl. der Position alignieren, die den bekannten **Transkriptionsstart** markiert und in den Sequenzen nach konservierten Regionen suchen.

→ *E.coli* Promotoren enthalten 3 konservierte Sequenzmerkmale

- eine etwa 6bp lange Region mit dem Konsensusmotif **TATAAT** bei Position **-10**

- eine etwa 6bp lange Region mit dem Konsensusmotif **TTGACA** bei Position **-35**

- die **Distanz** zwischen den beiden Regionen von etwa 17bp ist relativ konstant

Machbarkeit der Motivsuche mit dem Computer?

Transkriptionsfaktorbindestellen (TFBS) mit einem Computerprogramm zu identifizieren ist schwierig, da diese aus kurzen, entarteten Sequenzen bestehen, die häufig ebenfalls durch Zufall auftreten.

→ Das Problem lässt sich daher schwer eingrenzen

• die Länge des gesuchten Motivs vorher nicht bekannt

• das Motiv braucht zwischen verschiedenen Promotern nicht stark konserviert sein.

• die Sequenzen, mit denen man nach dem Motiv sucht, brauchen nicht notwendigerweise dem gesamten Promoter entsprechen

Suche nach gemeinsamen Sequenzmotiven

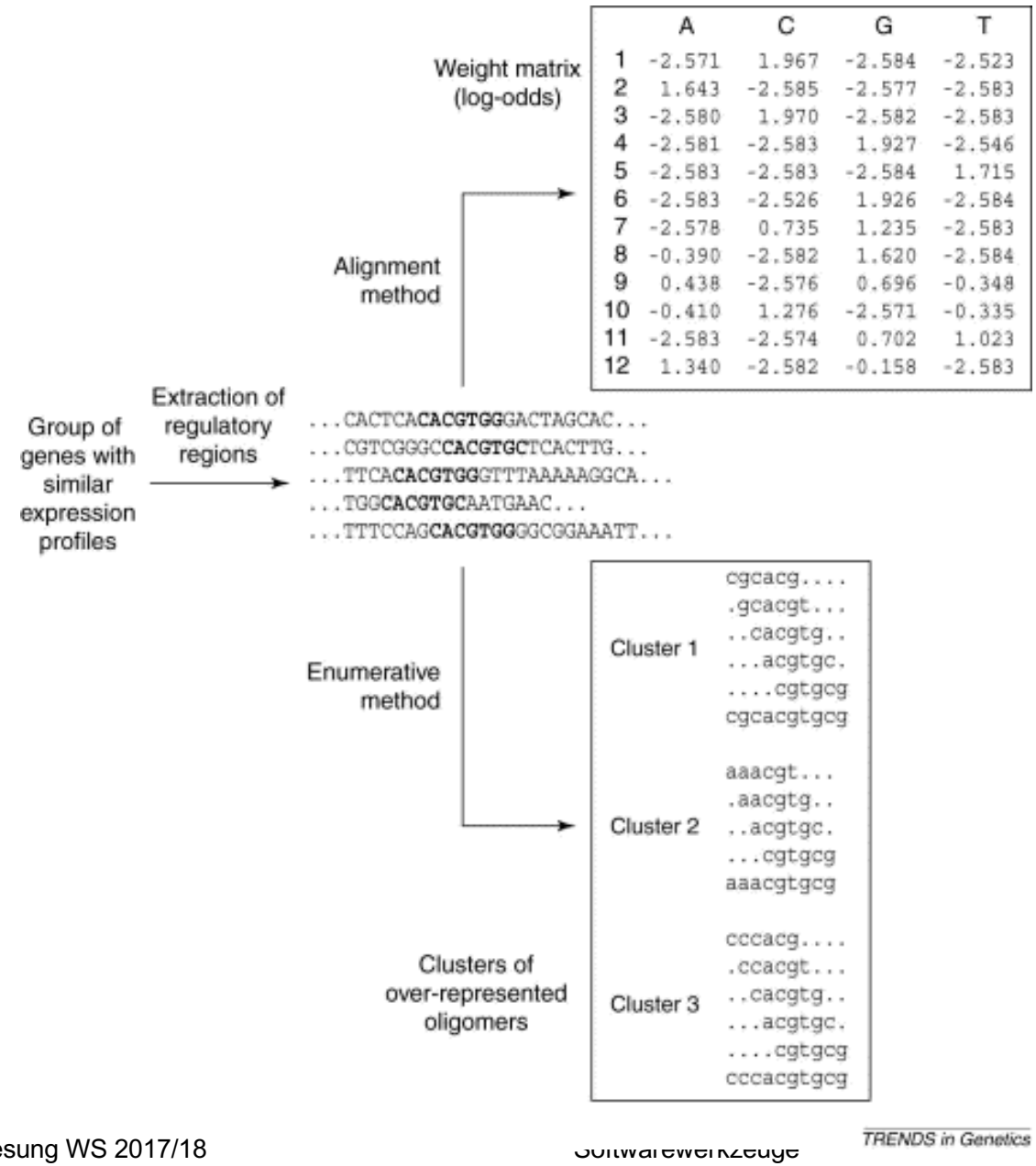
Wird seit der Verfügbarkeit von Microarray Gen-Expressionsdaten eingesetzt.

Durch Clustern erhält man Gruppen von Genen mit ähnlichen Expressionsprofilen (z.B. solche, die zur selben Zeit im Zellzyklus aktiviert sind)

Hypothese, dass dieses Profil, zumindest teilweise, durch eine ähnliche Struktur der für die transkriptionelle Regulation verantwortlichen cis-regulatorischen Regionen verursacht wird.

→ Suche nach gemeinsamen Motiven in upstream Region des TSS dieser Gene (z.B. -100 bp für Prokaryoten bzw. -2000 bp für Eukaryoten).

Motif-Identifizierung



	A	C	G	T
1	-2.571	1.967	-2.584	-2.523
2	1.643	-2.585	-2.577	-2.583
3	-2.580	1.970	-2.582	-2.583
4	-2.581	-2.583	1.927	-2.546
5	-2.583	-2.583	-2.584	1.715
6	-2.583	-2.526	1.926	-2.584
7	-2.578	0.735	1.235	-2.583
8	-0.390	-2.582	1.620	-2.584
9	0.438	-2.576	0.696	-0.348
10	-0.410	1.276	-2.571	-0.335
11	-2.583	-2.574	0.702	1.023
12	1.340	-2.582	-0.158	-2.583

Cluster 1:
```
cgcacg....
.gcacgt...
..cacgtg..
...acgtgc.
....cgtgcg
cgcacgtgcg
```

Cluster 2:
```
aaacgt...
.aacgtg..
..acgtgc.
...cgtgcg
aaacgtgcg
```

Cluster 3:
```
cccacg....
.ccacgt...
..cacgtg..
...acgtgc.
....cgtgcg
cccacgtgcg
```

Ohler, Niemann
Trends Gen 17, 2 (2001)

Positions-spezifische Gewichtsmatrix

Populäres Verfahren wenn es eine Liste von Genen gibt, die ein TF-Bindungs-motiv gemeinsam haben. Bedingung: gute MSAs müssen vorhanden sein.

Alignment-Matrix: wie häufig treten die verschiedenen Buchstaben an jeder Position im Alignment auf?

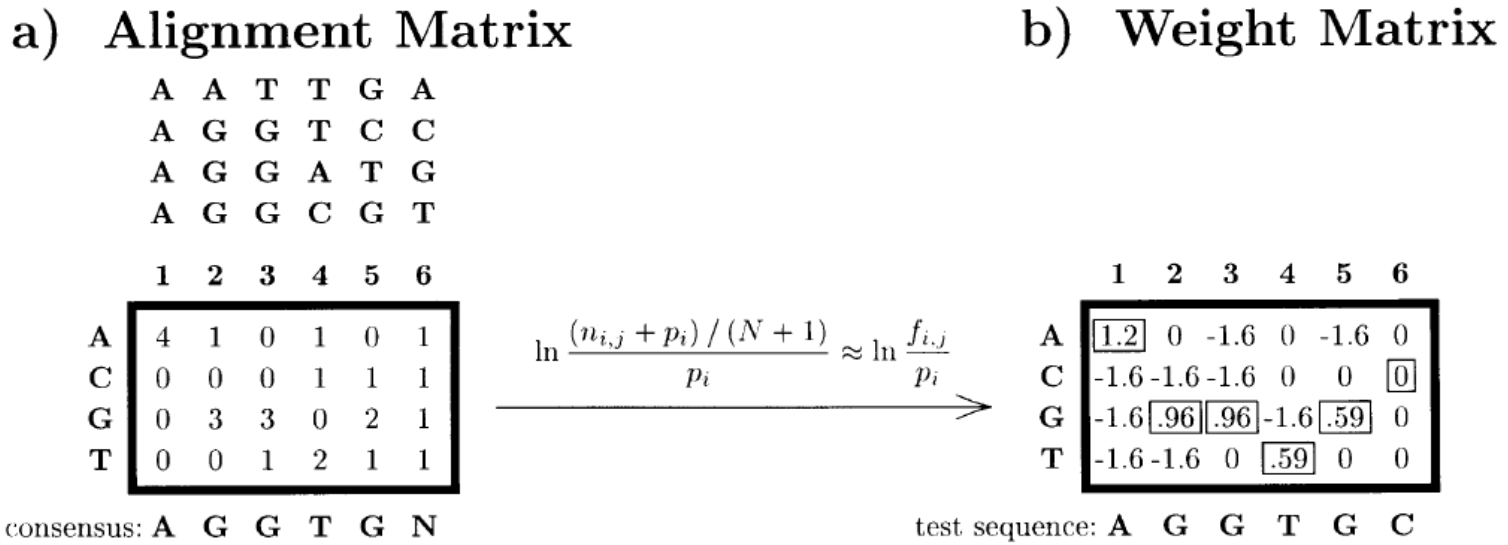


Fig. 1. Examples of the simple matrix model for summarizing a DNA alignment. (**a**) An alignment matrix describing the alignment of the four 6-mers on top. The matrix contains the number of times, $n_{i,j}$, that letter i is observed at position j of this alignment. Below the matrix is the consensus sequence corresponding to the alignment (N indicates that there is no nucleotide preference). (**b**) A weight matrix derived from the alignment in (a). The formula used for transforming the alignment matrix to a weight matrix is shown above the arrow. In this formula, N is the total number of sequences (four in this example), p_i is the *a priori* probability of letter i (0.25 for all the bases in this example) and $f_{i,j} = n_{i,j}/N$ is the frequency of letter i at position j. The numbers enclosed in blocks are summed to give the overall score of the test sequence. The overall score is 4.3, which is also the maximum possible score with this weight matrix.

Hertz, Stormo (1999) Bioinformatics 15, 563

Positions-spezifische Gewichtsmatrix

Beispiele für Matrizen, die von YRSA verwendet werden:

```
A  [11  0  0 10  0  2 11  3  0  4 ]
C  [ 0  0 11  0  6  3  0  8  0  0 ]
G  [ 0  0  0  1  0  0  0  0 11  7 ]
T  [ 0 11  0  0  5  6  0  0  0  0 ]
ID: MYO001
NAME: ABF1
SOURCE: Church Lab: ABF1.mot
LINK: http://atlas.med.harvard.edu/motifs/ABF1.mot
---

A  [4 5 5 0 4 0 9 0 0 0 6 1 4 0 0 ]
C  [0 0 0 0 0 9 0 8 9 8 0 1 1 1 3 ]
G  [0 2 1 0 4 0 0 0 0 0 1 4 1 0 0 ]
T  [5 2 3 9 1 0 0 1 0 1 2 3 3 8 6 ]
ID: MYO002
NAME: AFT1
SOURCE: Church Lab: AFT1.mot
LINK: http://atlas.med.harvard.edu/motifs/AFT1.mot
---

A  [ 7  9  0  2  1  0 10  0  0  0  0 ]
C  [ 2  0  0  1  0 10  0 10  0  0  0 ]
G  [ 0  0  2  4  1  0  0  0 10  0 10 ]
T  [ 1  1  8  3  8  0  0  0  0 10  0 ]
ID: MYO003
NAME: CBF1
SOURCE: Church Lab: CBF1.mot
LINK: http://atlas.med.harvard.edu/motifs/CBF1.mot
---
```

http://forkhead.cgb.ki.se/YRSA/matrixlist.html

Datenbank für eukaryotische Transkriptionsfaktoren: TRANSFAC

BIOBase / TU Braunschweig / GBF

Relationelle Datenbank

6 Dateien:

FACTOR Wechselwirkung von TFs

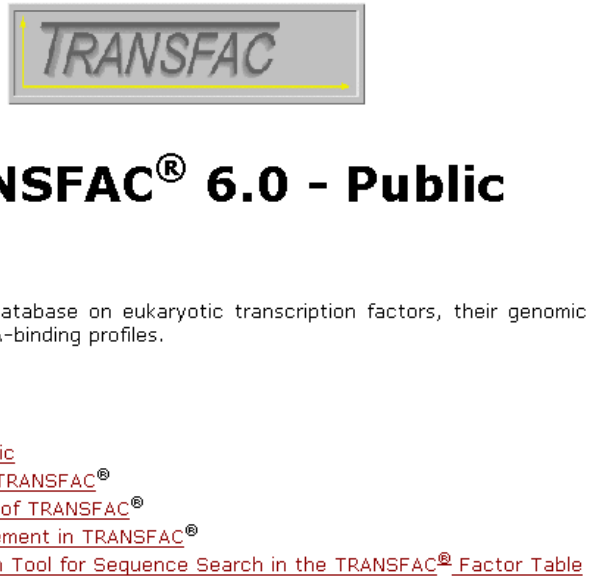
SITE ihre DNA-Bindungsstelle

GENE durch welche sie diese Zielgene regulieren

CELL wo kommt Faktor in Zelle vor?

MATRIX TF Nukleotid-Gewichtungsmatrix

CLASS Klassifizierungsschema der TFs



Wingender et al. (1998) J Mol Biol 284,241

Datenbank für eukaryotische Transkriptionsfaktoren: TRANSFAC

BIOBase / TU Braunschweig / GBF

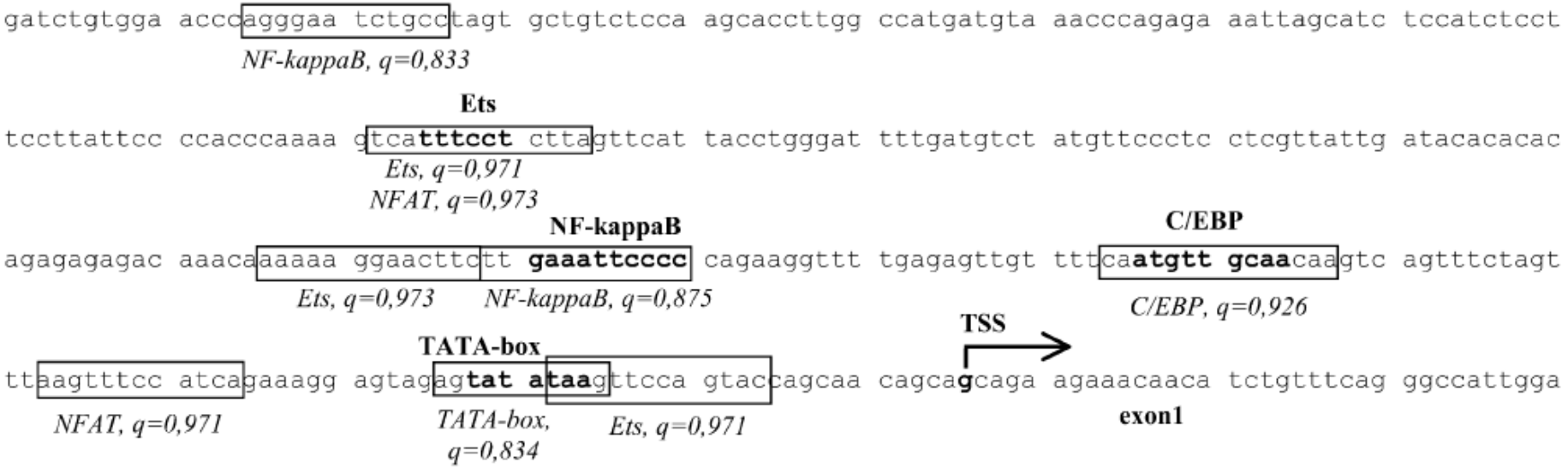


Figure 1. Application of specific profiles provided by the Match™ program. Potential binding sites found by Match™ are boxed, name of the transcription factor and score for the match are given under the sequence. (**A**) The immune-specific profile (with modified cut-offs) is applied to find potential binding sites within the promoter sequence of the human IL-12 p40 gene (EMBL accession no. AY008847, positions 2101 to 2460). Known binding sites for transcription factors are shown in bold, the name of the transcription factor is given above the sequence. The transcription start site (TSS) is indicated by an arrow.

Matys et al. (2003) Nucl Acid Res 31,374

Identifizierung von Repeats: RepeatMasker

RepeatMasker: durchsucht DNA Sequenzen auf
- eingefügte Abschnitte, die **bekannten Repeat-Motiven** entsprechen
(dazu wird eine lange Tabelle mit bekannten Motiven verwendet)
und
- auf **Regionen geringer Komplexität** (z.B. lange Abschnitt AAAAAAAA).

Output:
- detaillierte Liste, wo die Repeats in der Sequenz auftauchen und
- eine modifizierte Version der Input-Sequenz, in der die Repeats „**maskiert**“ sind, z.B. durch N‘s ersetzt sind.

Für die Sequenzvergleiche wird eine effiziente Implementation des Smith-Waterman-Gotoh Algorithmus verwendet.

http://www.gene-regulation.com

Zusammenfassung

Es gibt große Datenbanken (z.B. TRANSFAC) mit Informationen über Promoterstellen. Diese Informationen sind experimentell überprüft.

Microarray-Daten erlauben es, nach gemeinsamen Motiven von ko-regulierten Genen zu suchen.

Auch möglich: gemeinsame Annotation in der Gene Ontology etc.

TF-Bindungsmotive sind oft überrepräsentiert in der 1000 bp-Region upstream. Die klare Funktion dieser Bindungsmotive ist oft unbekannt.

Allgemein gilt:
- relativ wenige TFs regulieren eine große Anzahl an Genen
- es gibt globale und lokale TFs
- Gene werden üblicherweise durch mehr als einen TF reguliert

http://www.gene-regulation.com

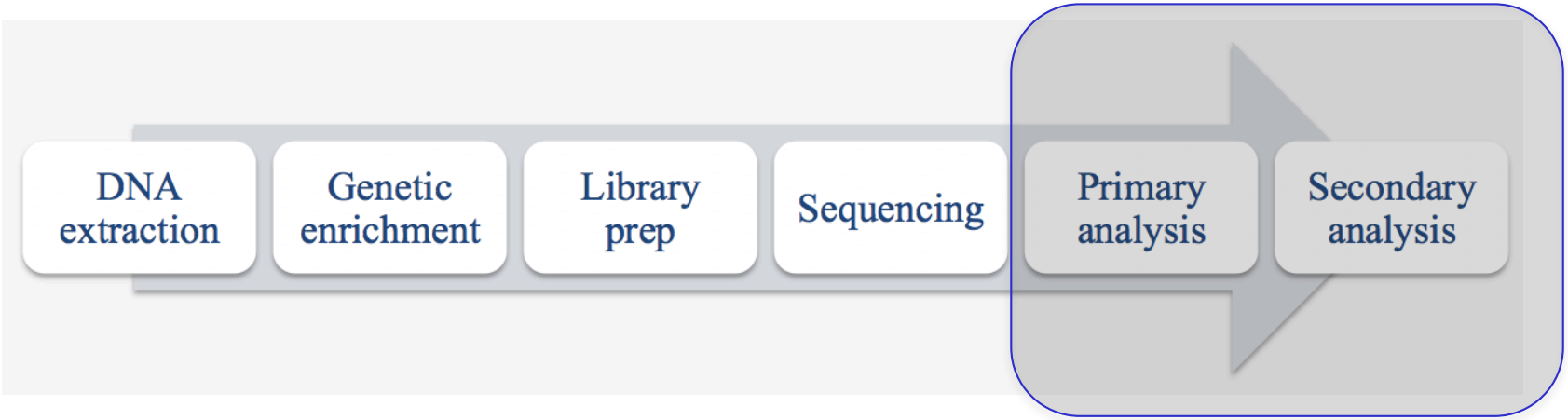
Prozessierung von NGS-Daten

- Ganzgenomsequenzierung = Whole Genome Sequencing (WGS)
- Anwendung von WGS für mikrobielle Isolate
- Qualitätskontrolle der Sequenzierungs-reads
- Alignment
- SNP calling
- Genomvisualisierung
- Genomassemblierung

Hier wird dies Thema nur grob vorgestellt,
NGS-Prozessierung wird genauer in Vorlesungen von
Prof. Keller, Prof. Marschall und Dr. Schulz behandelt.

Danksagung für Folien: Mohamed Hamed

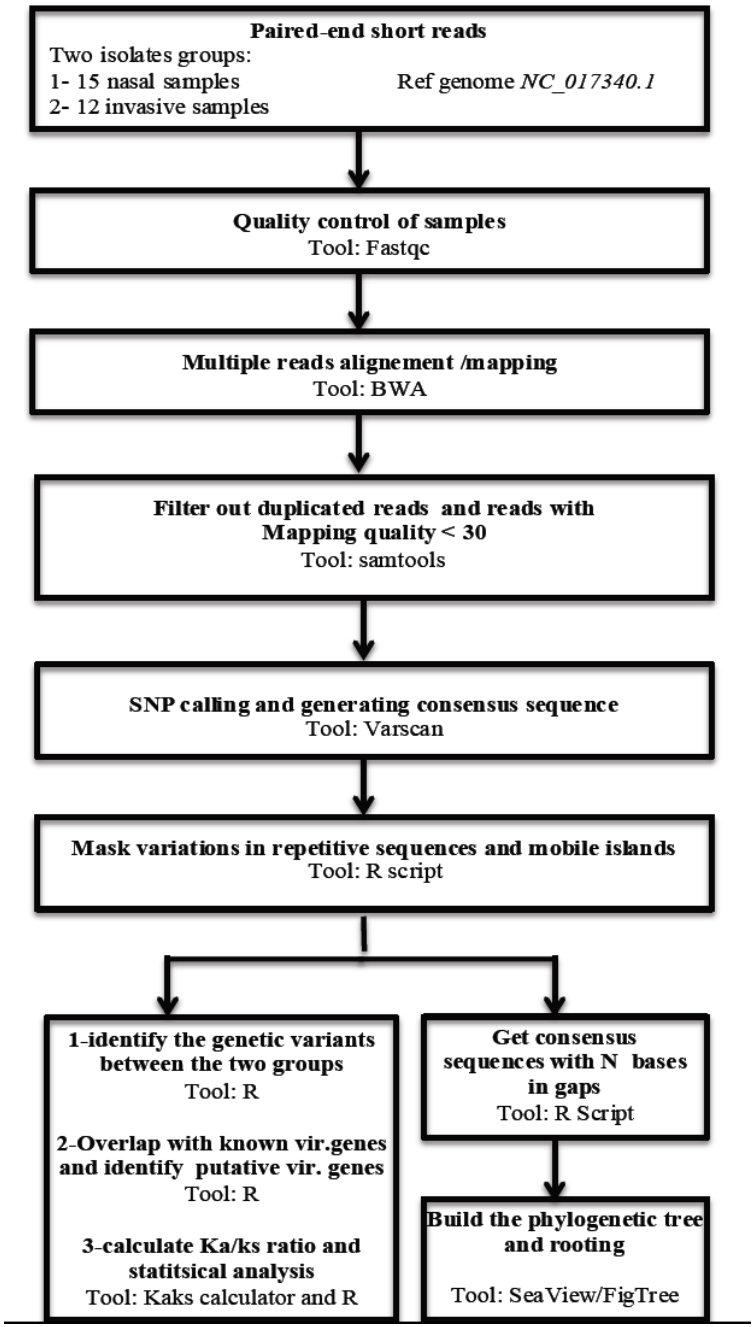
NGS Pipeline im Überblick



1. Extraktion der DNA aus biologischer Probe
2. Genetic enrichment: Manchmal soll nur eine kleine Region des Genoms sequenziert werden (einzelne Gene bzw. nur die Exons bei Sequenzierung von eukaryot. Genomen). Die Extraktion dieser Regionen aus dem Genome nennt man Anreicherung (enrichment).
3. Vorbereitung der Bibliothek (Library prep): Für viele Sequenziermaschinen muss die DNA für die Sequenzierung vorbereitet werden.
4. Die eigentliche Sequenzierung
5. Rohanalyse (primary analysis): Alignment / Assemblierung, SNP calling
6. Eigentliche Analyse (secondary analysis): Identifizierung von kausalen SNPs variants, phänotypische Charaktisierung (z.B. Virulenzfaktoren)

Wir konzentrieren uns auf die Schritte 5 und 6

WGS Pipeline für bakterielle Phylotypisierung



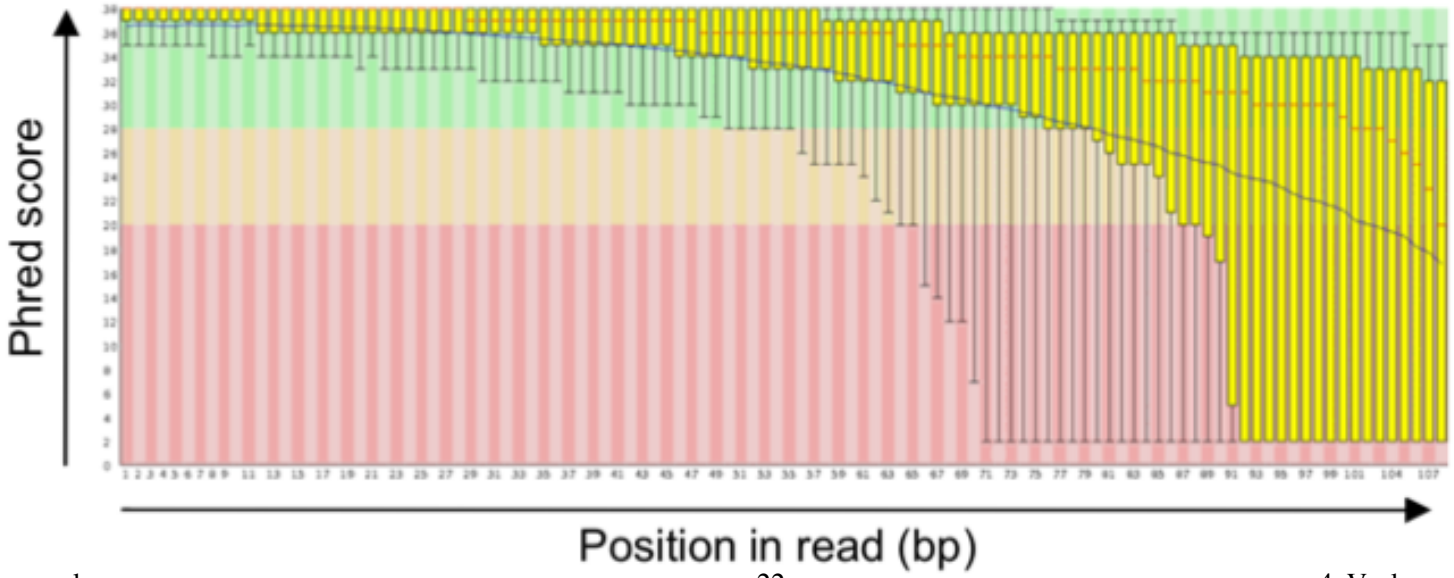
Quality (Phred) score

Phred Score (Q):

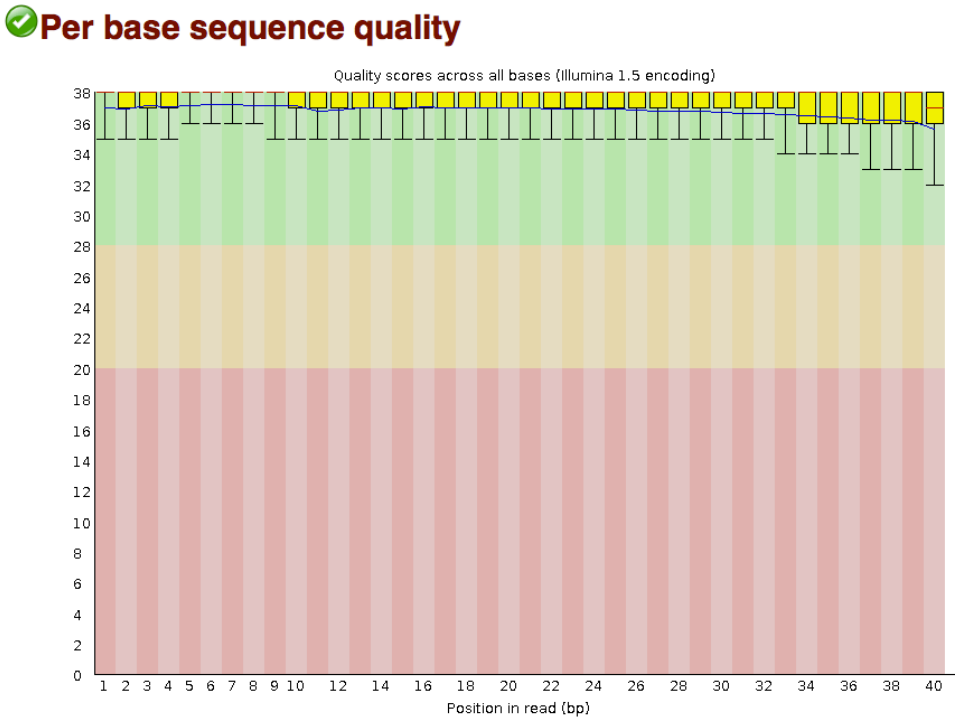
$Q = -10 \times \log_{10} P$

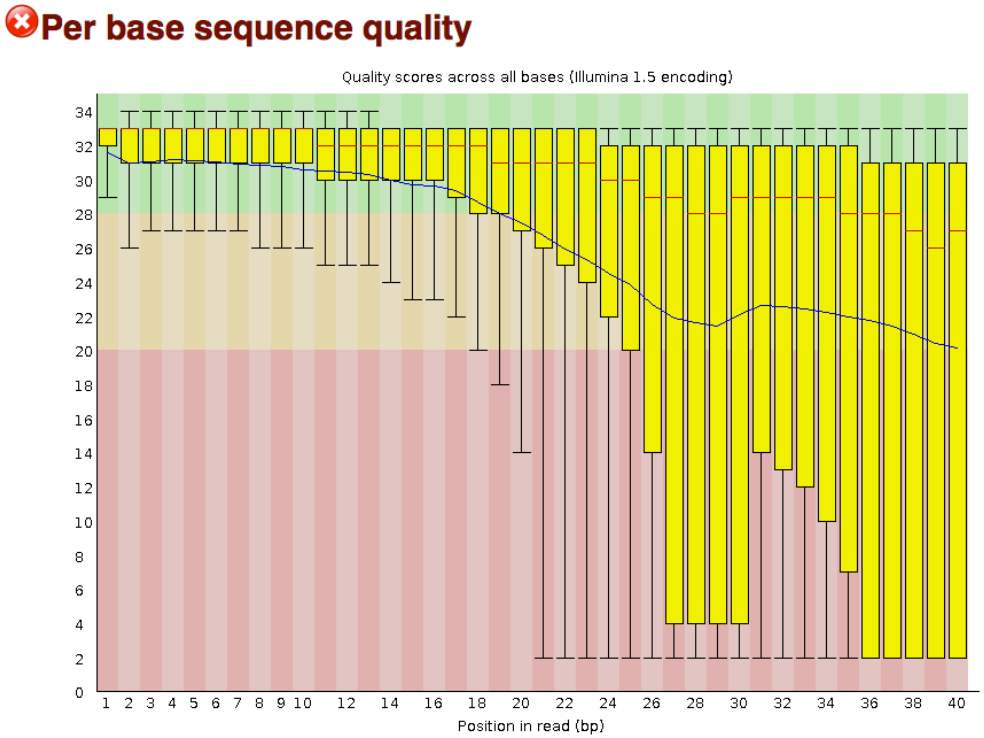
P ist eine Abschätzung für den Fehler des Base-calling aus den Rohdaten der Sequenzierung. D.h. ein Fehler von 0.1% (10^{-3}) ergibt Q = 30.

Base Qualitäts-scores nehmen üblicherweise am Ende der reads ab
Deshalb werden die reads vor dem Alignment-Schritts „getrimmt“, d.h. gekürzt.



Qualitätskontrolle - FASTQC





Die Sequenzqualität pro Base ist im linken Beispiel durchgehend sehr hoch (d.h. Q > 30),

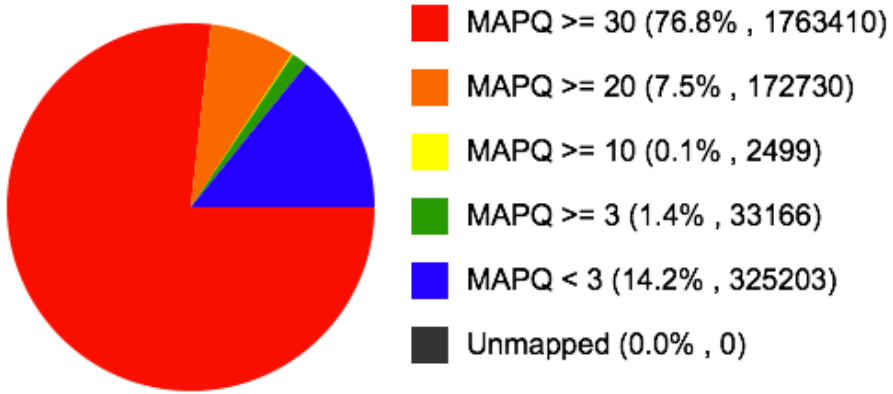
im rechten Beispiel etwa ab Position 20 aber unzuverlässig.

Qualitätskontrolle im Alignment-Schritt

Auch bei der Alignierung mit dem Referenz-Genom muss bewertet werden, ob den Reads zweifelsfreie Positionen zugeordnet werden können.

SAMStat: monitoring biases in next generation sequencing data

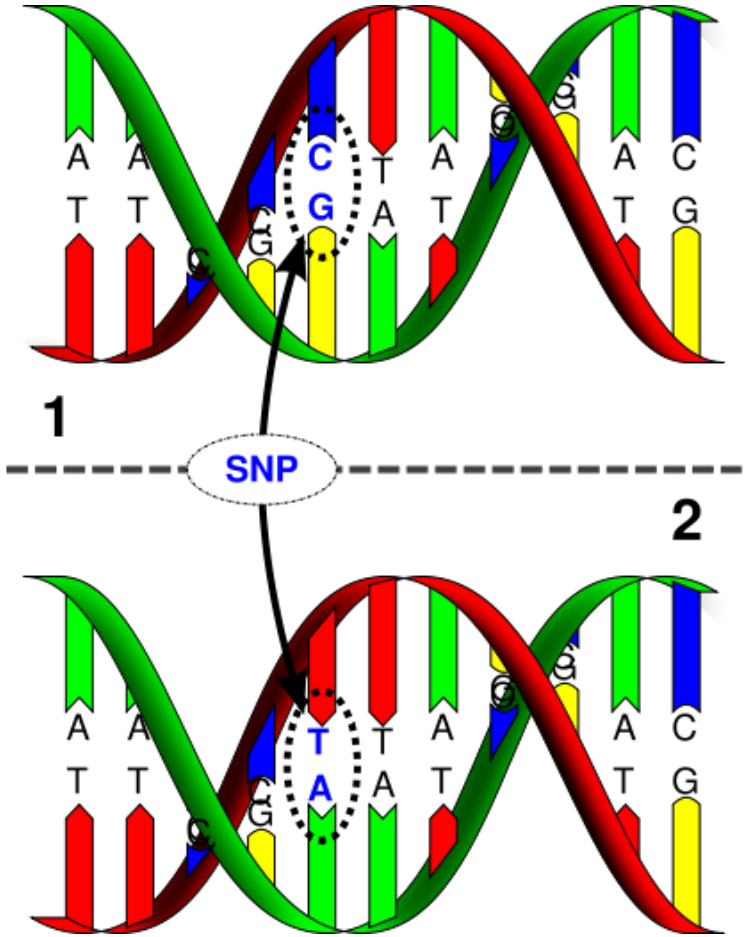
Timo Lassmann*, Yoshihide Hayashizaki and Carsten O. Daub



Verteilung der Mapping Qualitätsscores

- Alle Reads werden entfernt, deren Mapping-Qualität geringer als 30 ist, d.h. die Fehlerwahrscheinlichkeit, dass der read auf eine andere Region gemappt wird, ist 0.1% und höher.
- Entfernung von duplizierten reads, da diese die Qualität des SNP-Calling beeinflussen.

Biologie von SNP-Mutationen



http://www.science.marshall.edu/murraye/341/Images/416px-Dna-SNP_svg.png

Verschiedene menschliche Genome unterscheiden sich etwa an jeder 1000-ten Base.

Die meisten Variationen sind Unterschiede einzelner Basen.

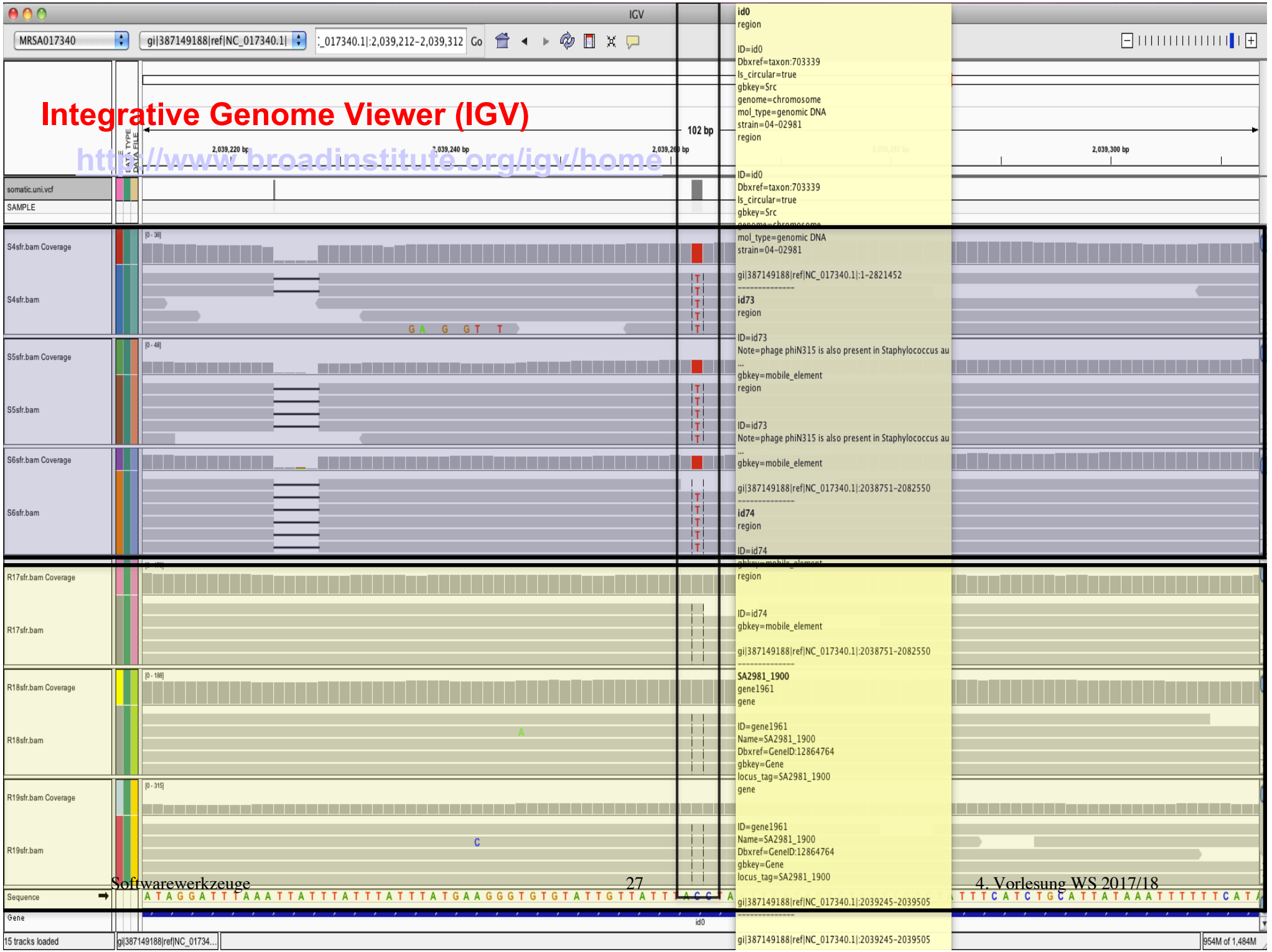
Polymorphismus: vererbter Unterschied

Somatische Mutation: erworbener Unterschied

Mögliche Gründe für Abweichungen in Alignments

- Ein wahrer SNP
- Experimenteller Fehler
  - Fehler bei Präparierung der Bibliothek oder bei der PCR
  - Base calling Fehler während Analyse von Rohdaten
- Fehler beim Alignment oder beim Mapping-Schritt
- Fehler in der Sequenz des Referenzgenoms

- Gebräuchliche Software Tools:
  - Samtools/bcftools
  - Gatk
  - Varscan
  - Snv-mix
- Die Ausgabe des Alignments ist im VCF Format (Variant Call Format)


IGV
MRSA017340
gi|387149188|ref|NC_017340.1|
:_017340.1|:2,039,212-2,039,312
Go
Integrative Genome Viewer (IGV)
http://www.broadinstitute.org/igv/home
102 bp
2,039,220 bp
2,039,240 bp
2,039,260 bp
2,039,300 bp
somatic.uni.vcf
SAMPLE
S4sfr.bam Coverage
S4sfr.bam
S5sfr.bam Coverage
S5sfr.bam
S6sfr.bam Coverage
S6sfr.bam
R17sfr.bam Coverage
R17sfr.bam
R18sfr.bam Coverage
R18sfr.bam
R19sfr.bam Coverage
R19sfr.bam
Sequence
Gene
id0
region
ID=id0
Dbxref=taxon:703339
Is_circular=true
gbkey=Src
genome=chromosome
mol_type=genomic DNA
strain=04-02981
region
ID=id0
Dbxref=taxon:703339
Is_circular=true
gbkey=Src
mol_type=genomic DNA
strain=04-02981
gi|387149188|ref|NC_017340.1|:1-2821452
id73
region
ID=id73
Note=phage phiN315 is also present in Staphylococcus au
gbkey=mobile_element
region
ID=id73
Note=phage phiN315 is also present in Staphylococcus au
gbkey=mobile_element
gi|387149188|ref|NC_017340.1|:2038751-2082550
id74
region
ID=id74
region
ID=id74
gbkey=mobile_element
gi|387149188|ref|NC_017340.1|:2038751-2082550
SA2981_1900
gene1961
gene
ID=gene1961
Name=SA2981_1900
Dbxref=GeneID:12864764
gbkey=Gene
locus_tag=SA2981_1900
gene
ID=gene1961
Name=SA2981_1900
Dbxref=GeneID:12864764
gbkey=Gene
locus_tag=SA2981_1900
gi|387149188|ref|NC_017340.1|:2039245-2039505
gi|387149188|ref|NC_017340.1|:2039245-2039505
15 tracks loaded
954M of 1,484M

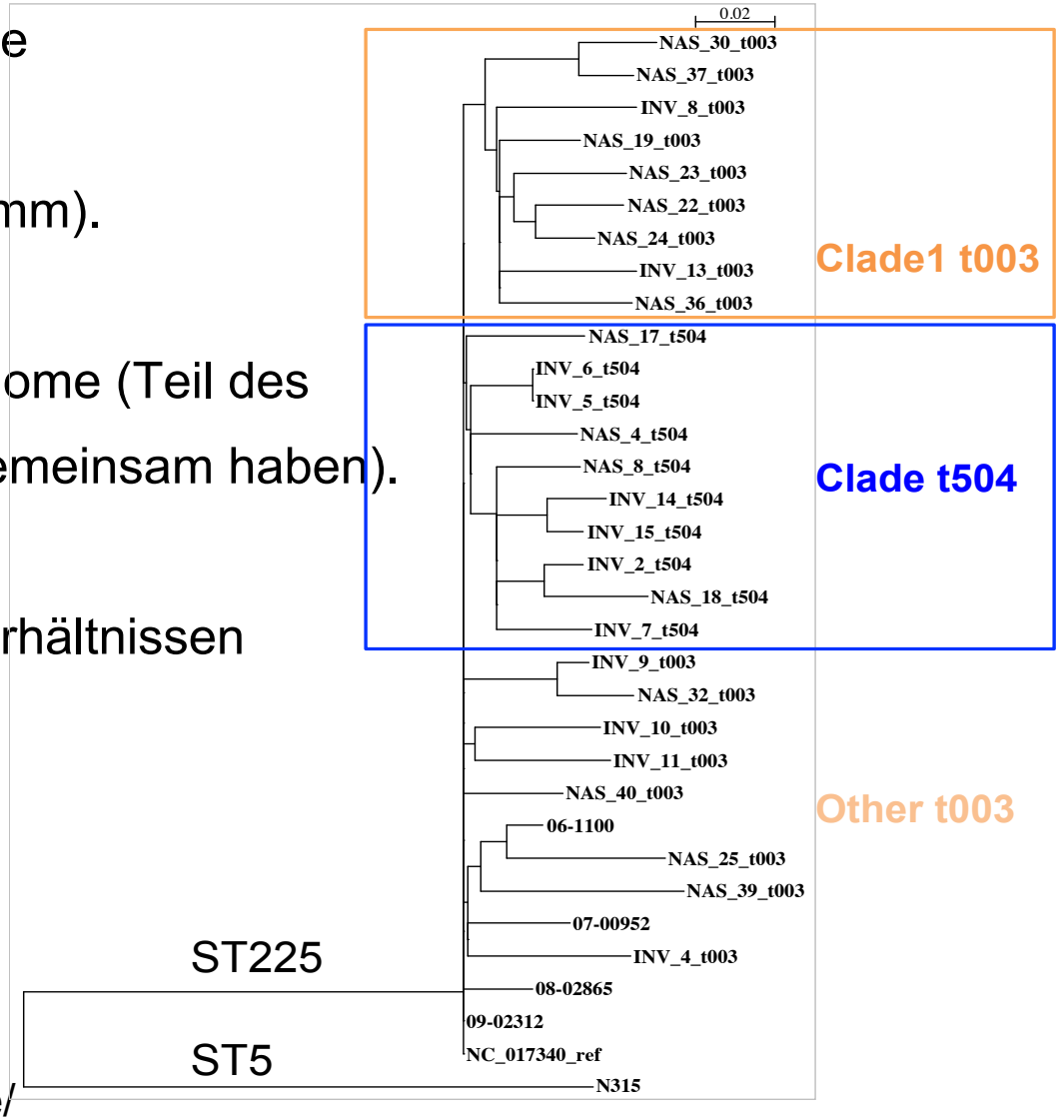
Phylogenetischer Baum aus core-genome SNPs

Input: WGS-Sequenzen für verschiedene *Staphylococcus aureus* Stämme (nas: nasaler Stamm; inv: invasiver Stamm).

Schritt 1: identifiziere SNPs im core-genome (Teil des *S. aureus*-Genoms, das alle Stämme gemeinsam haben).

Schritt 2: konstruiere Verwandschaftsverhältnissen zwischen den Stämmen.

Ausgabe: phylogenetischer Baum



Hamed et al. (2015)
Infection, Genetics and Evolution

- Tools
  - FigTree http://tree.bio.ed.ac.uk/software/
  - SeaView http://pbil.univ-lyon1.fr/software/seaview3.html

Whole Genome Alignment (WGA)

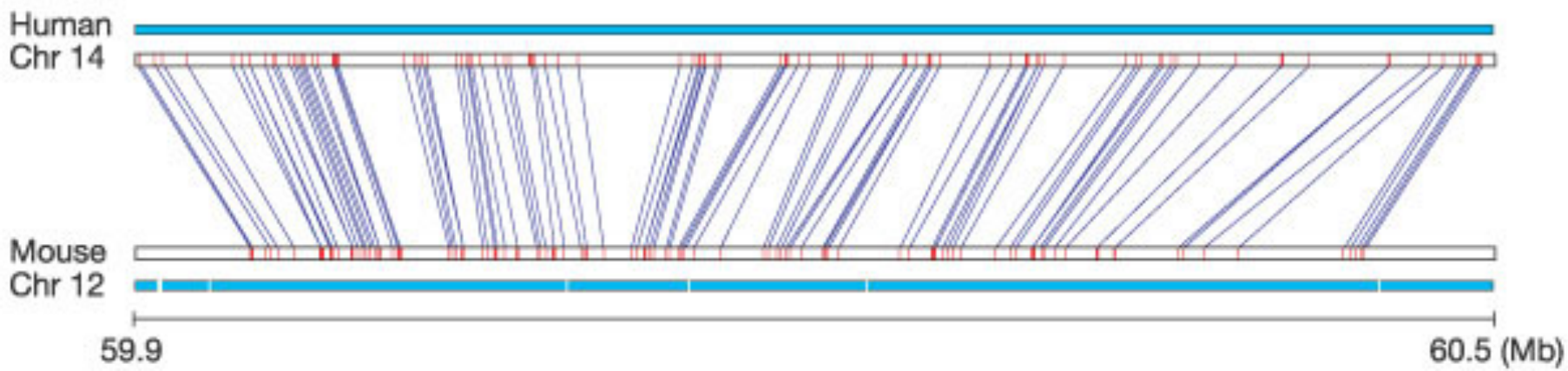
Wenn die genomische DNA-Sequenz eng verwandter Organismen verfügbar wird, kann man ein Alignment von zwei Genomen konstruieren.

Globale Genom-Alignments machen nur für eng verwandte Organismen Sinn.

Im anderen Fall muss man zuerst die genomischen Rearrangements betrachten.

Dann kann man die **systenischen Regionen** (Regionen, in denen Gen-Reihenfolge des nächsten gemeinsamen Vorfahrens in beiden Spezies konserviert blieb) betrachten und **lokale Genom-Alignments** dieser Regionen produzieren.

Konservierung von Syntenie zwischen Mensch und Maus

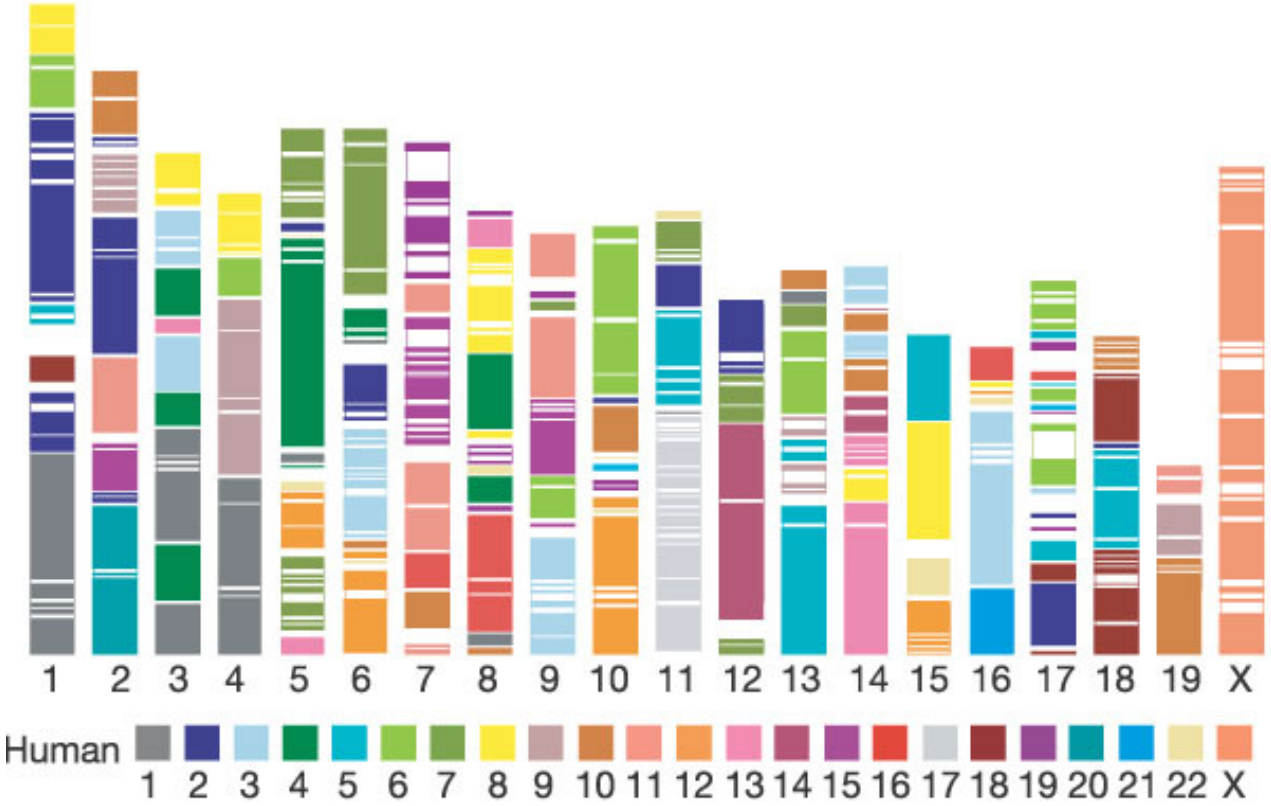


Ein typisches 510-kb Segment des Maus-Chromosoms 12, das mit einem 600-kb Stück des menschlichen Chromosom 14 verwandt ist.
Blaue Linien: reziprok eindeutige Treffer in beiden Genomen.
Rote Markierungen kennzeichnen die Länge der passenden Regionen.

Die Abstände zwischen diesen „Landmarks“ sind im Maus-Genom kleiner als im Mensch, was mit der 14% kürzeren Gesamtlänge des Genoms übereinstimmt.

The mouse genome. *Nature* **420**, 520 - 562

Entsprechung syntenischer Regionen



342 Segmente und 217 Blöcke >300 kb mit konservierter Syntenie im Mensch sind im Maus-Genom markiert.

Jede Farbe entspricht einem bestimmten menschlichen Chromosom.

The mouse genome. *Nature* **420**, 520 - 562

Sensitivität

Im globalen Mensch:Maus Alignment sind mehr als eine Millionen Regionen stärker als 70% konserviert (auf 100-bp Level)
– diese Regionen decken > 200 Million bp ab.

Nur 62% von ihnen werden von (lokalen) BLAT-Treffern abgedeckt.

Dies bedeutet, daß man 38% der konservierten Abschnitte nur durch das globale Alignment finden kann!

Idee: lokales Alignment soll als **Anker-Verfahren** für anschliessendes globales Alignment dienen.
Dadurch hofft man, viele zusätzliche konservierte Regionen ausserhalb der Anker-Regionen zu finden.

Couronne, ..., Dubchak, Genome Res. 13, 73 (2003)

Ankerbasierte Methoden für WGA

Diese Methoden versuchen, sich entsprechende Teile der Buchstabenfolgen der betrachteten Sequenzen zu finden, die wahrscheinlich zu einem globalen Alignment gehören werden.
(Diese teilweisen Treffer können durch lokale Alignments gefunden werden).
Sie bilden „Anker“ in den beiden zu alignierenden Sequenzen.

In diesen Methoden werden zuerst die Ankerpunkte aligniert und dann die Lücken dazwischen geschlossen.

MUMmer ist eine sehr erfolgreiche Implementation dieser Strategie für das Alignment zweier genomischer Sequenzen.

Was ist MUMmer?

- A.L. Delcher *et al*. 1999, 2002 Nucleic Acids Res.
- http://www.tigr.org/tigr-scripts/CMR2/webmum/mumplot

- nimm an, dass zwei Sequenzen eng verwandt sind (sehr ähnlich)
- MUMmer kann zwei bakterielle Genome in weniger als 1 Minute alignieren
- nutzt **Suffix-Bäume** um Maximal Unique Matches zu finden
- Definition eines Maximal Unique Matches (MUM):
  - Eine Subsequenz, die in beiden Sequenzen genau einmal ohne Abweichungen vorkommt und in keine Richtung verlängert werden kann.
- Grundidee: ein MUM ausreichender Länge wird sicher Teil eines globalen Alignments sein.

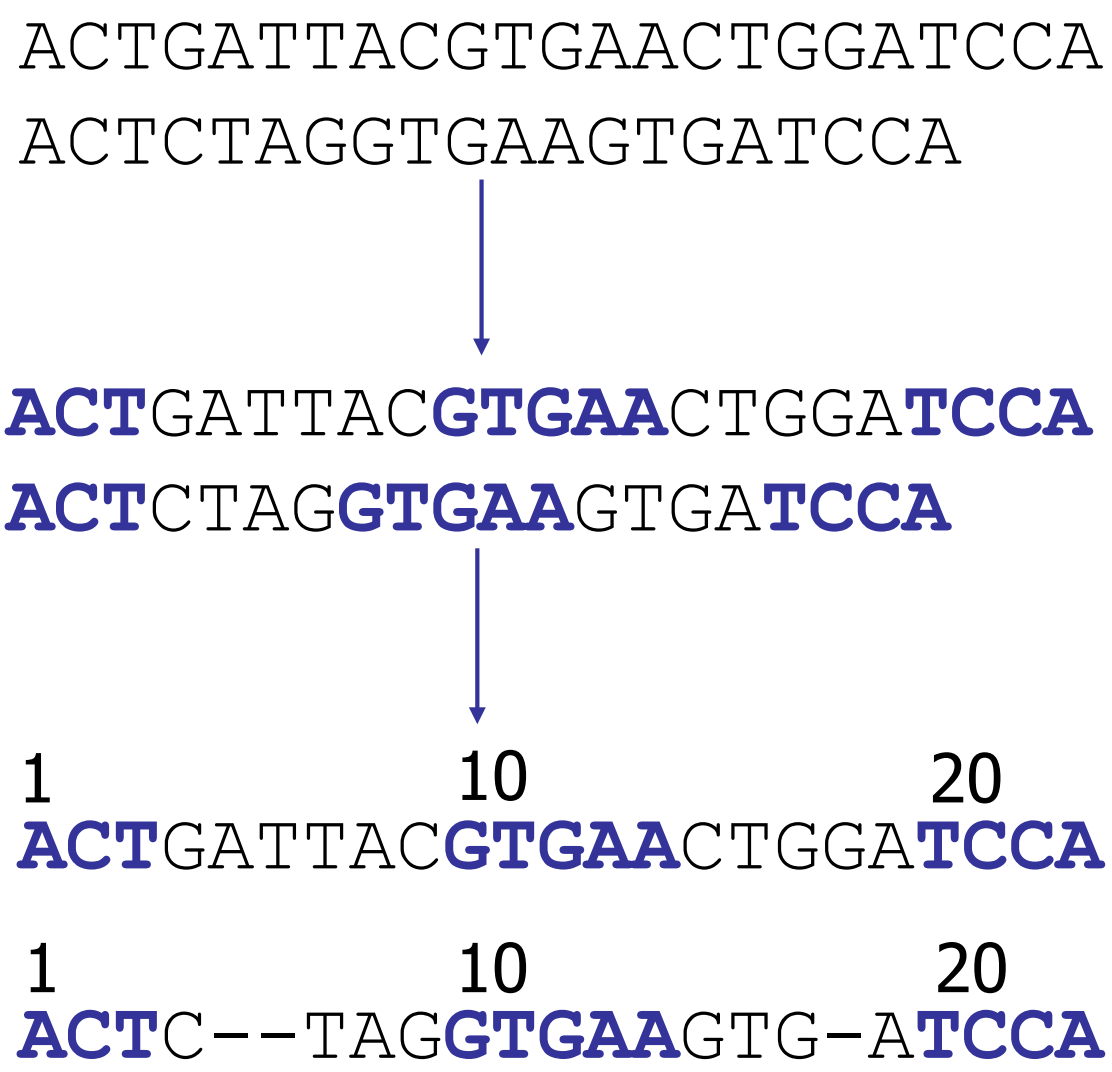
Genome A: `tcgatcGACGATCGCGGCCGTAGATCGAATAACGAGAGAGCATAAcgactta`
Genome B: `gcattaGACGATCGCGGCCGTAGATCGAATAACGAGAGAGCATAAtccagag`

A maximal unique matching subsequence (MUM) of 39 nt (shown in uppercase) shared by Genome *A* and Genome *B*. Any extension of the MUM will result in a mismatch. By definition, an MUM does not occur anywhere else in either genome.

Delcher et al. Nucleic Acids Res 27, 2369 (1999)

MUMmer: wichtige Schritte

- Erkenne MUMs (Länge wird vom Benutzer festgelegt)



Definition von MUMmers

- Für zwei Strings S1 und S2 und einen Parameter *l*
- Der Substring *u* ist eine MUM Sequenz wenn gilt:
  - $|u| > l$
  - *u* kommt genau einmal in S1 und genau einmal in S2 (Eindeutigkeit) vor
  - Für jeden Buchstaben *a* kommt weder *ua* noch *au* sowohl in S1 als auch in S2 vor (Maximalität)

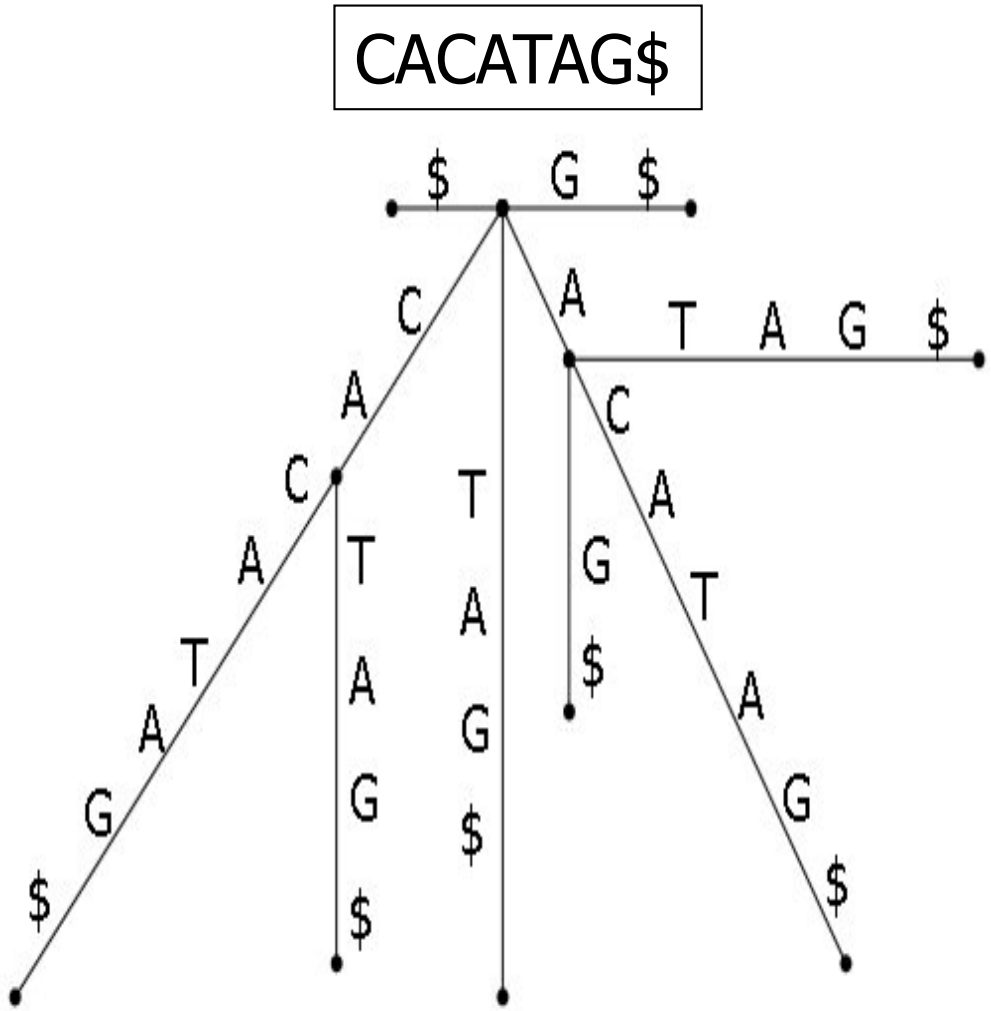
Wie findet man MUMs?

- Naiver Ansatz
  - Vergleiche alle Teilsequenzen von A mit allen Teilsequenzen von B.
    Dies dauert $O(n^n)$

- verwende Suffix-Bäume als Datenstruktur
  - ein naiver Ansatz, einen Suffix-Baum zu konstruieren hat
    eine quadratische Komplexität in der Rechenzeit und dem Speicherplatz

  - durch klevere Benutzung von Pointern gibt es lineare Algorithmen in Rechenzeit und Speicherplatz wie den Algorithmus von McCreight

Suffix-Bäume



Suffix-Bäume sind seit über 30 Jahren wohl etabliert.

Einige ihrer Eigenschaften:

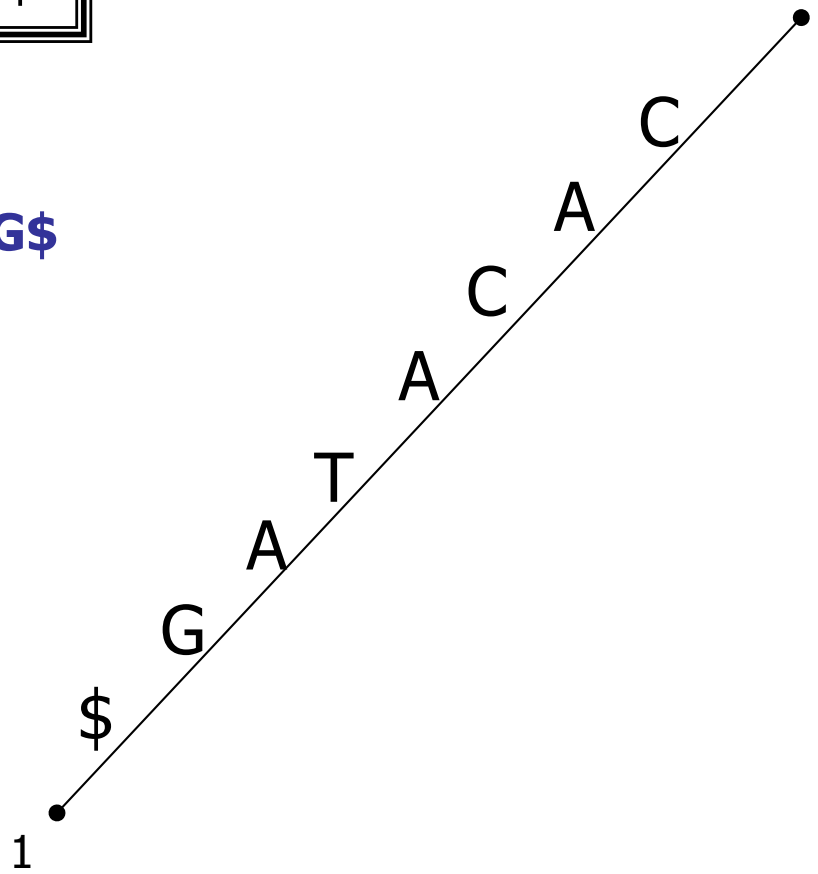
- ein “Suffix” beginnt an jeder Position *I* der Sequenz und reicht bis zu ihrem Ende.
- Eine Sequenz der Länge *N* hat *N* Suffices.
- Es gibt *N* Blätter.
- Jeder interne Knoten hat mindest zwei Kinder.
- 2 Kanten aus dem selben Knoten können nicht mit dem selben Buchstaben beginnen.
- Am Ende wird $ angefügt

Konstruktion eines Suffix-Baums

CACATAG$

Suffixes:

1. CACATAG$

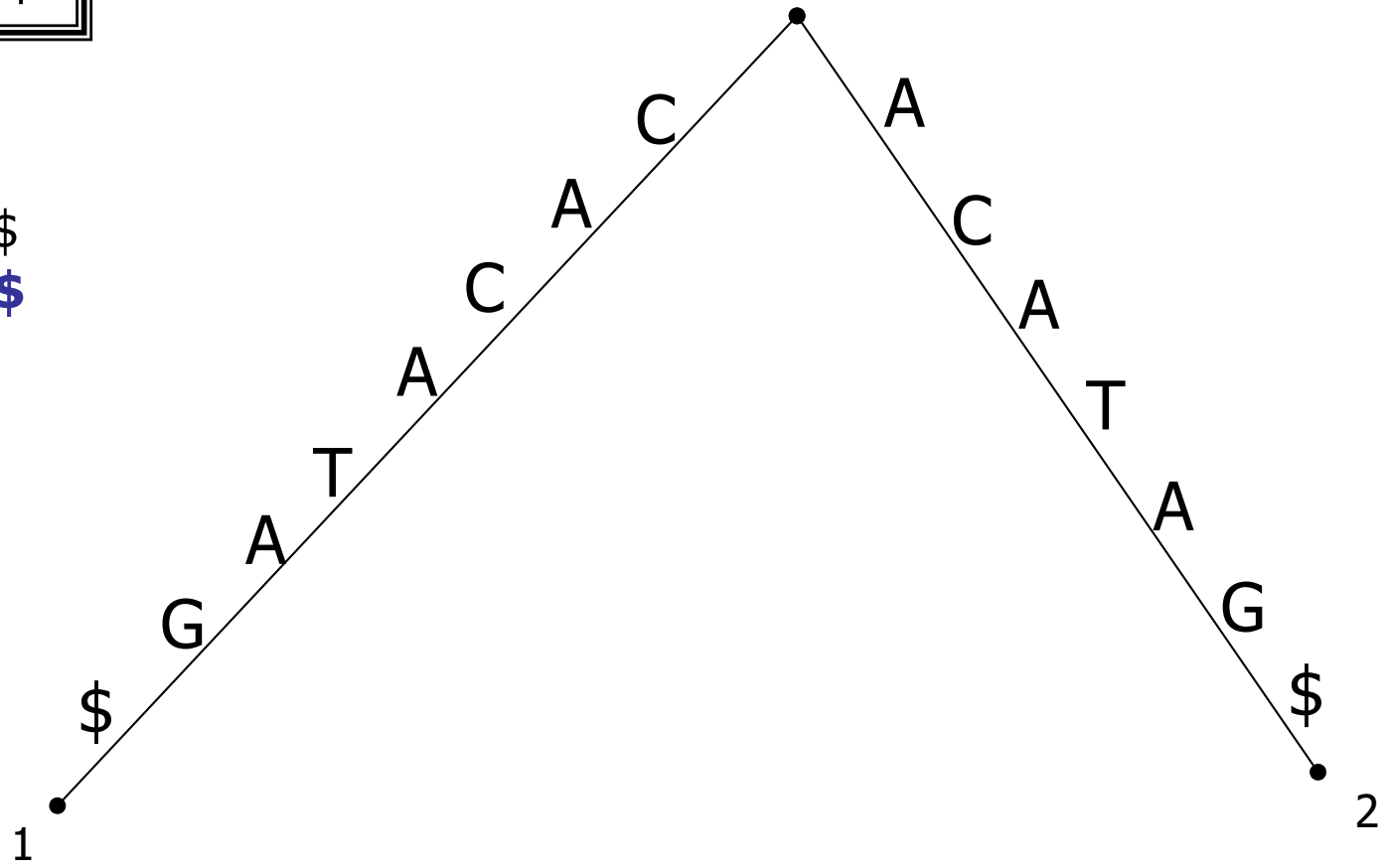


Konstruktion eines Suffix-Baums

CACATAG$

Suffixes:

1. CACATAG$
2. **ACATAG$**

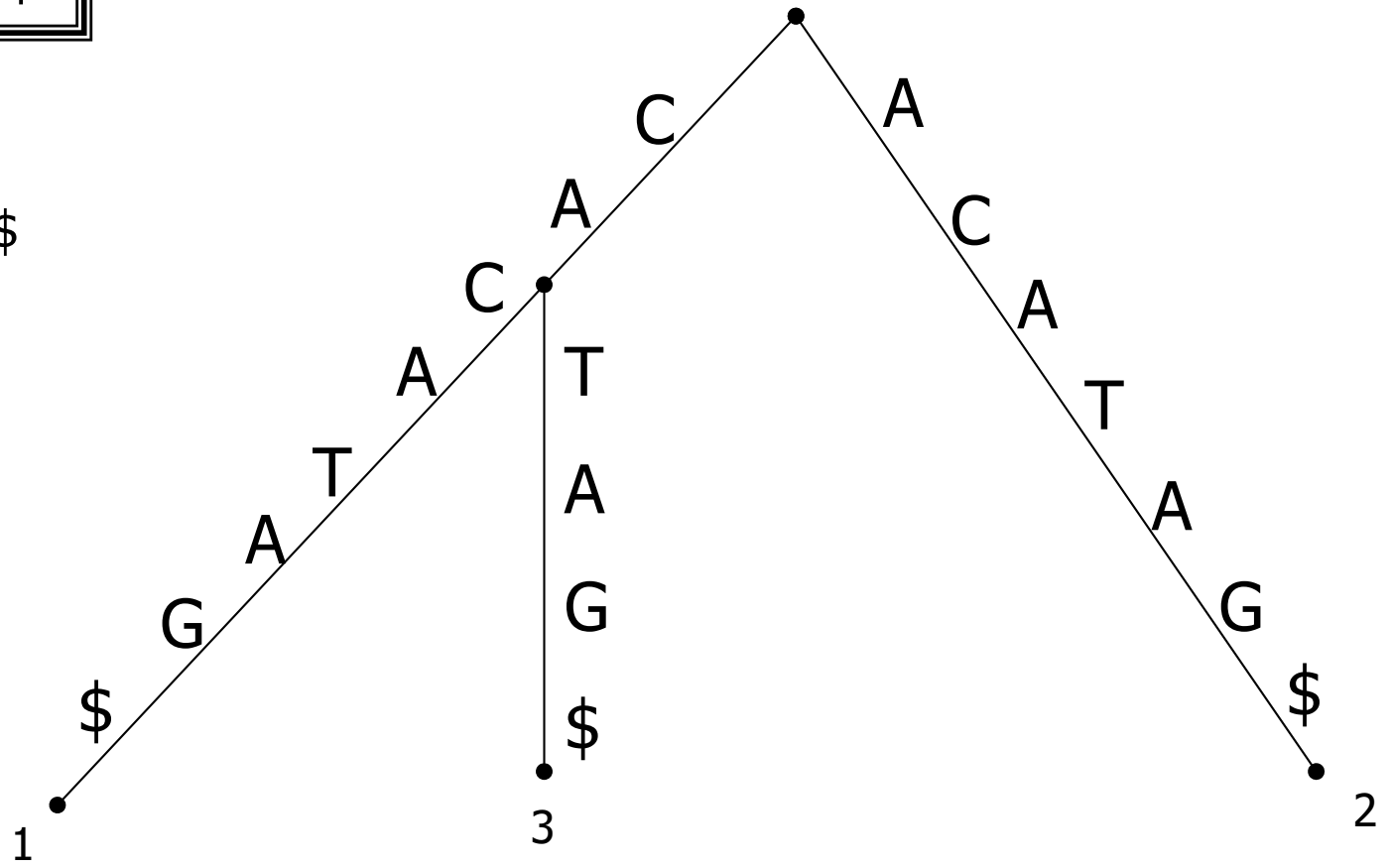


Konstruktion eines Suffix-Baums

CACATAG$

Suffixes:

1. CACATAG$
2. ACATAG$
3. **CATAG$**

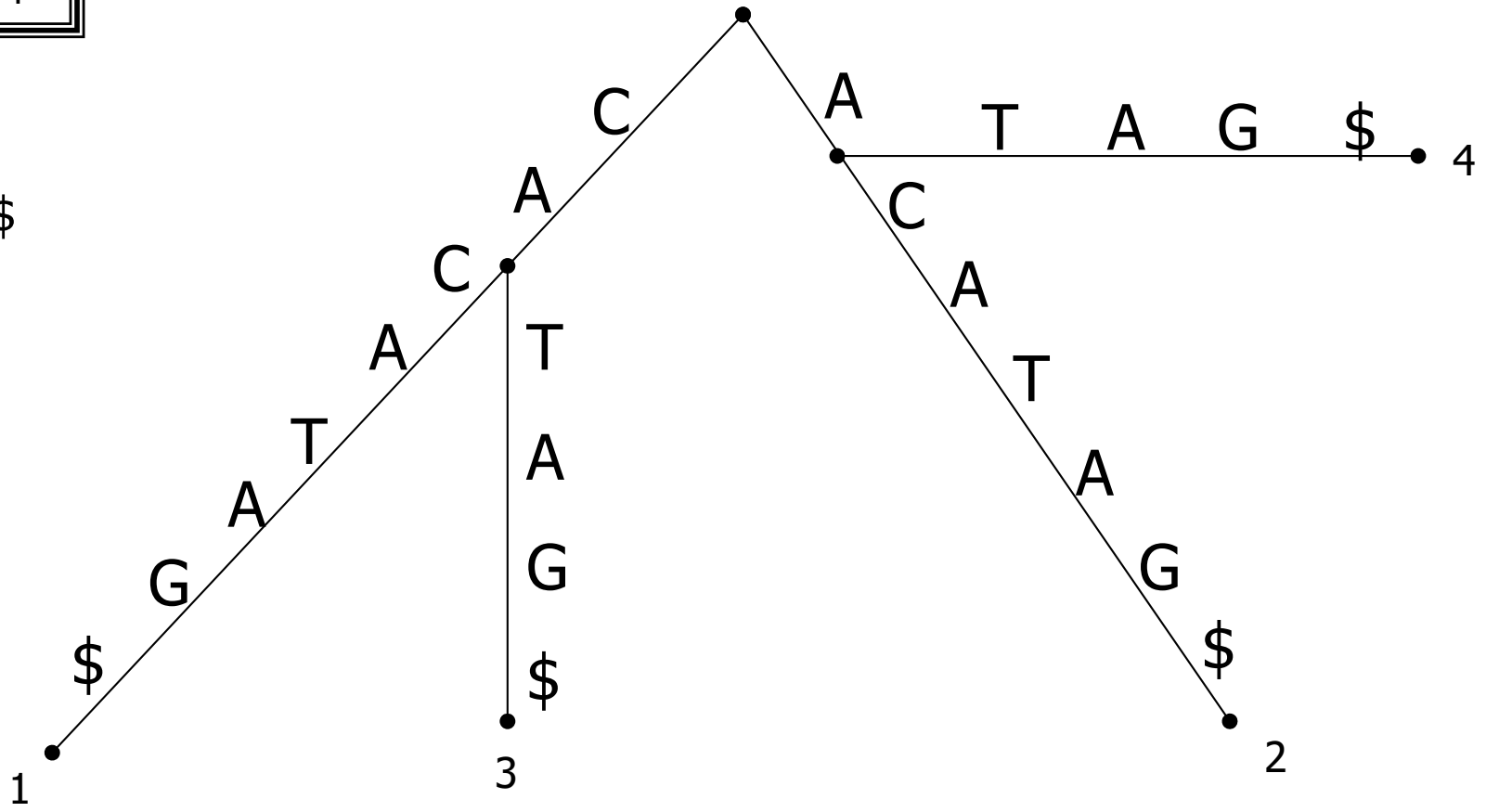


Konstruktion eines Suffix-Baums

CACATAG$

Suffixes:

1. CACATAG$
2. ACATAG$
3. CATAG$
4. **ATAG$**

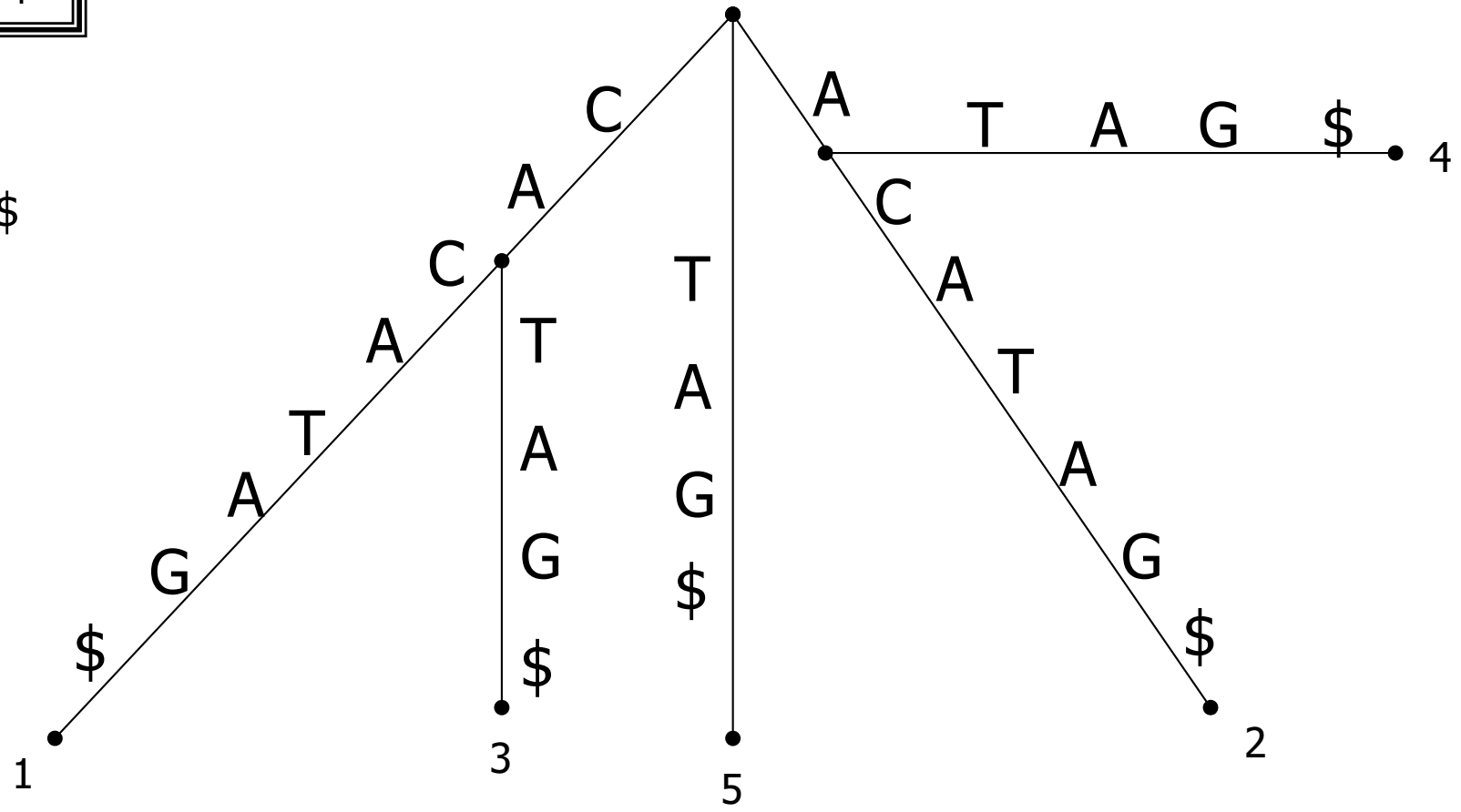


Konstruktion eines Suffix-Baums

CACATAG$

Suffixes:

1. CACATAG$
2. ACATAG$
3. CATAG$
4. ATAG$
5. **TAG$**

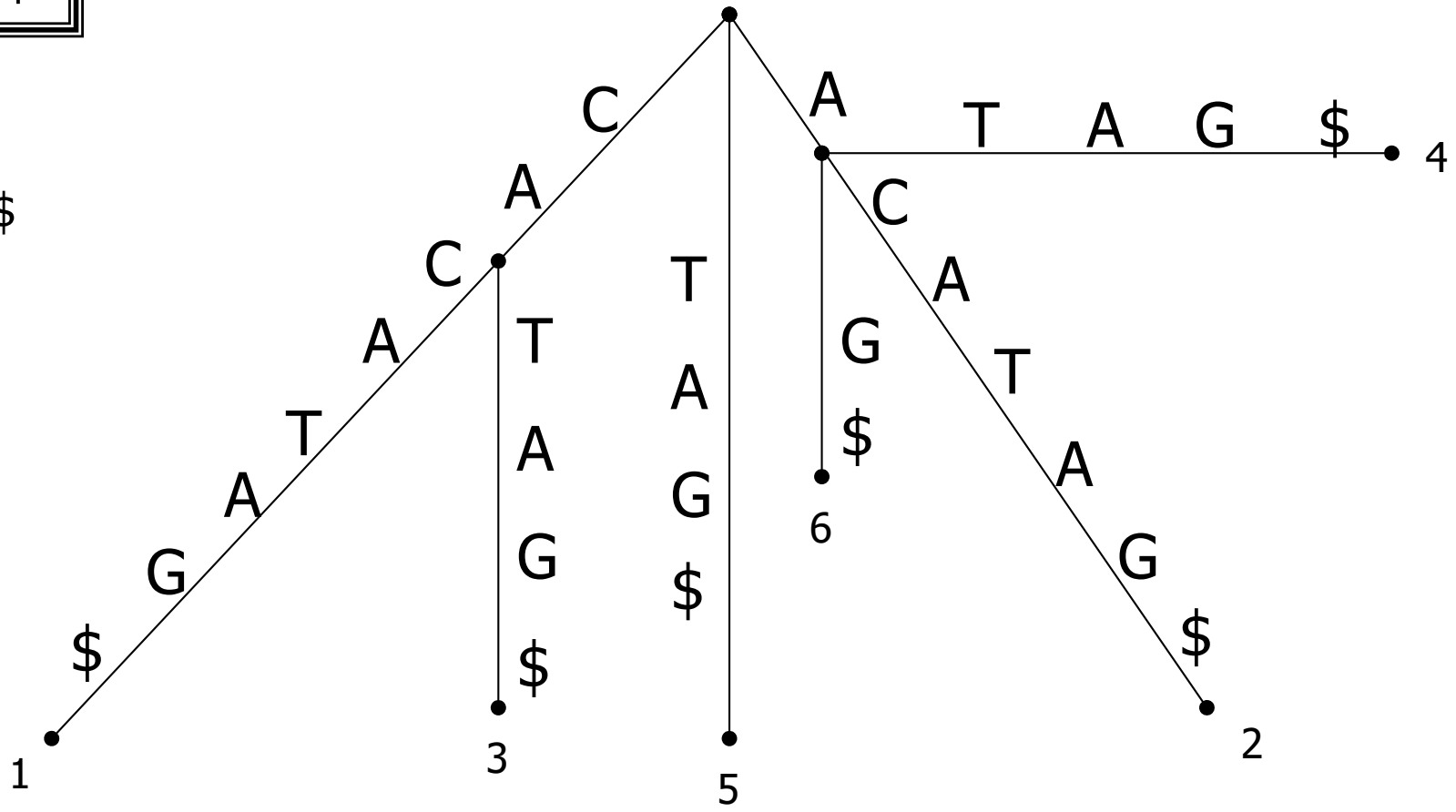


Konstruktion eines Suffix-Baums

CACATAG$

Suffixes:

1. CACATAG$
2. ACATAG$
3. CATAG$
4. ATAG$
5. TAG$
6. **AG$**

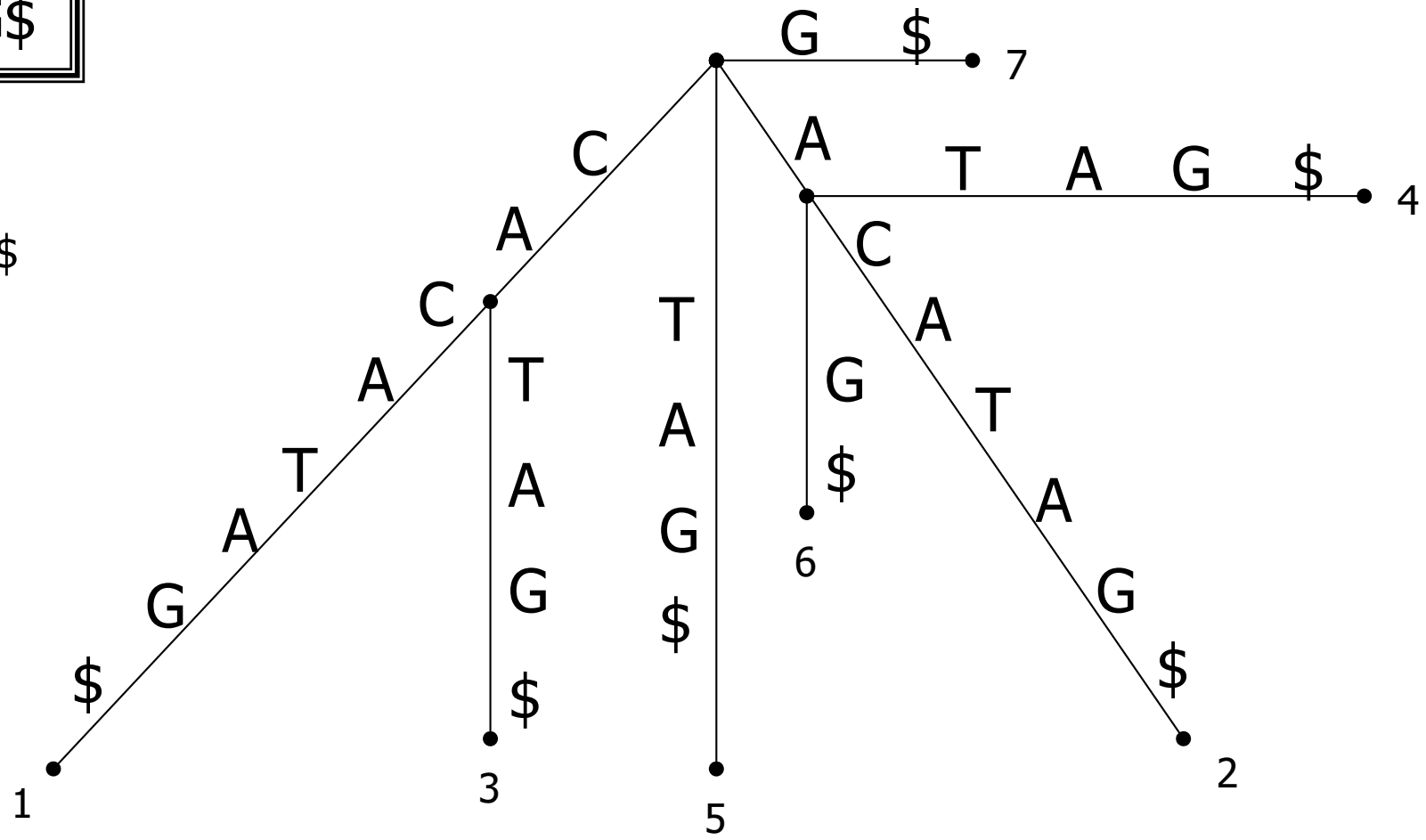


Konstruktion eines Suffix-Baums

CACATAG$

Suffixes:

1. CACATAG$
2. ACATAG$
3. CATAG$
4. ATAG$
5. TAG$
6. AG$
7. **G$**

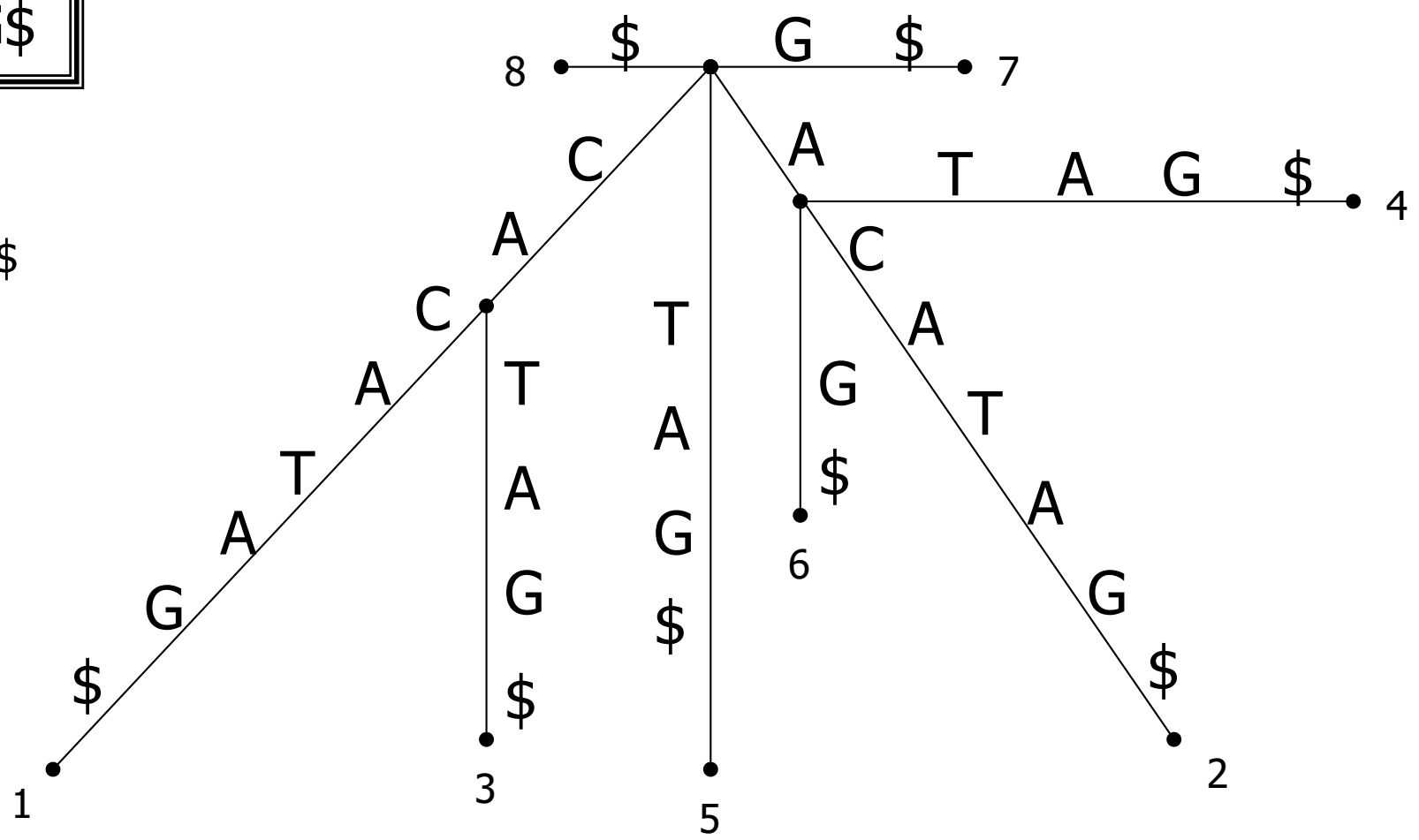


Konstruktion eines Suffix-Baums

CACATAG$

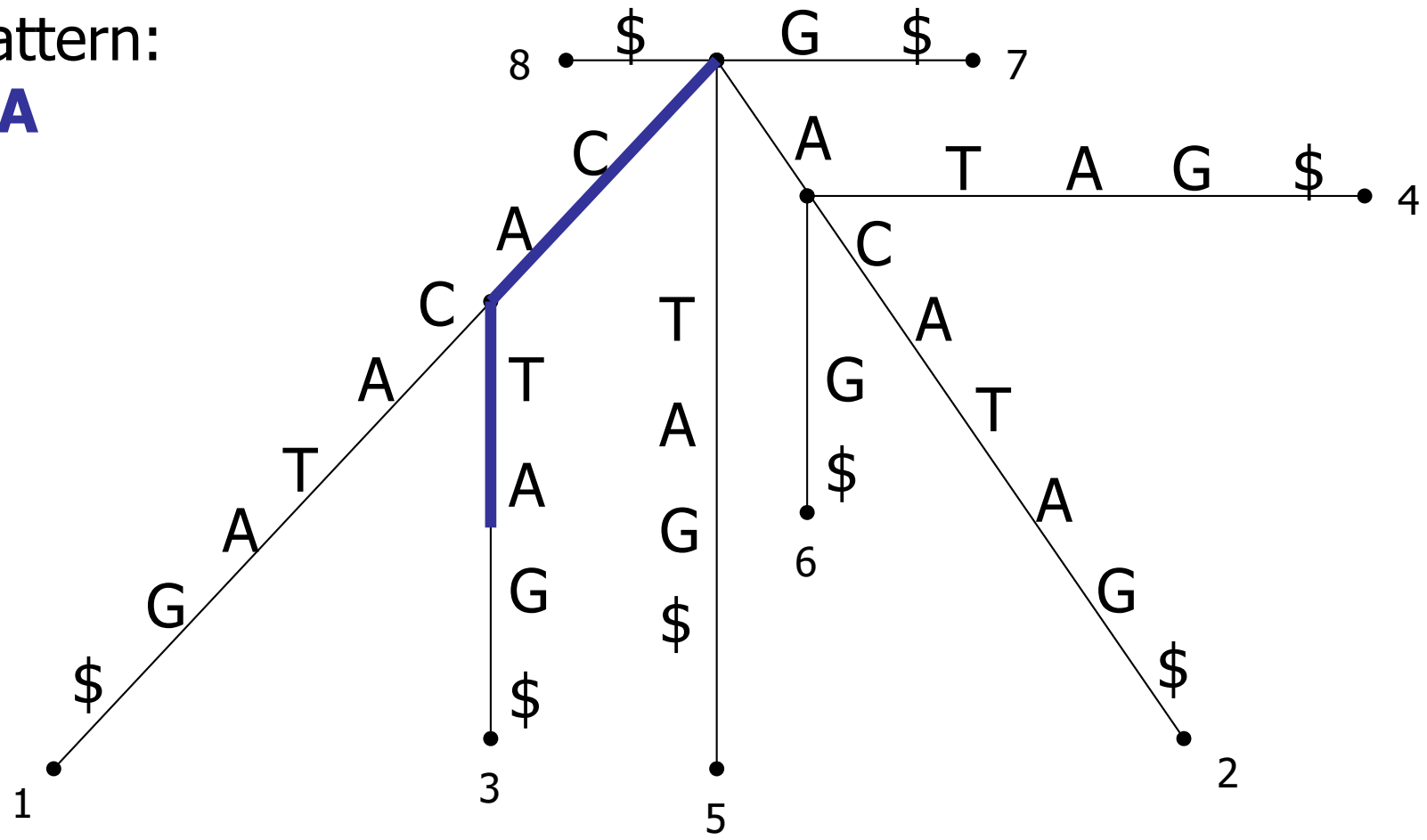
Suffixes:

1. CACATAG$
2. ACATAG$
3. CATAG$
4. ATAG$
5. TAG$
6. AG$
7. G$
8. $



Suchen in einem Suffix-Baum

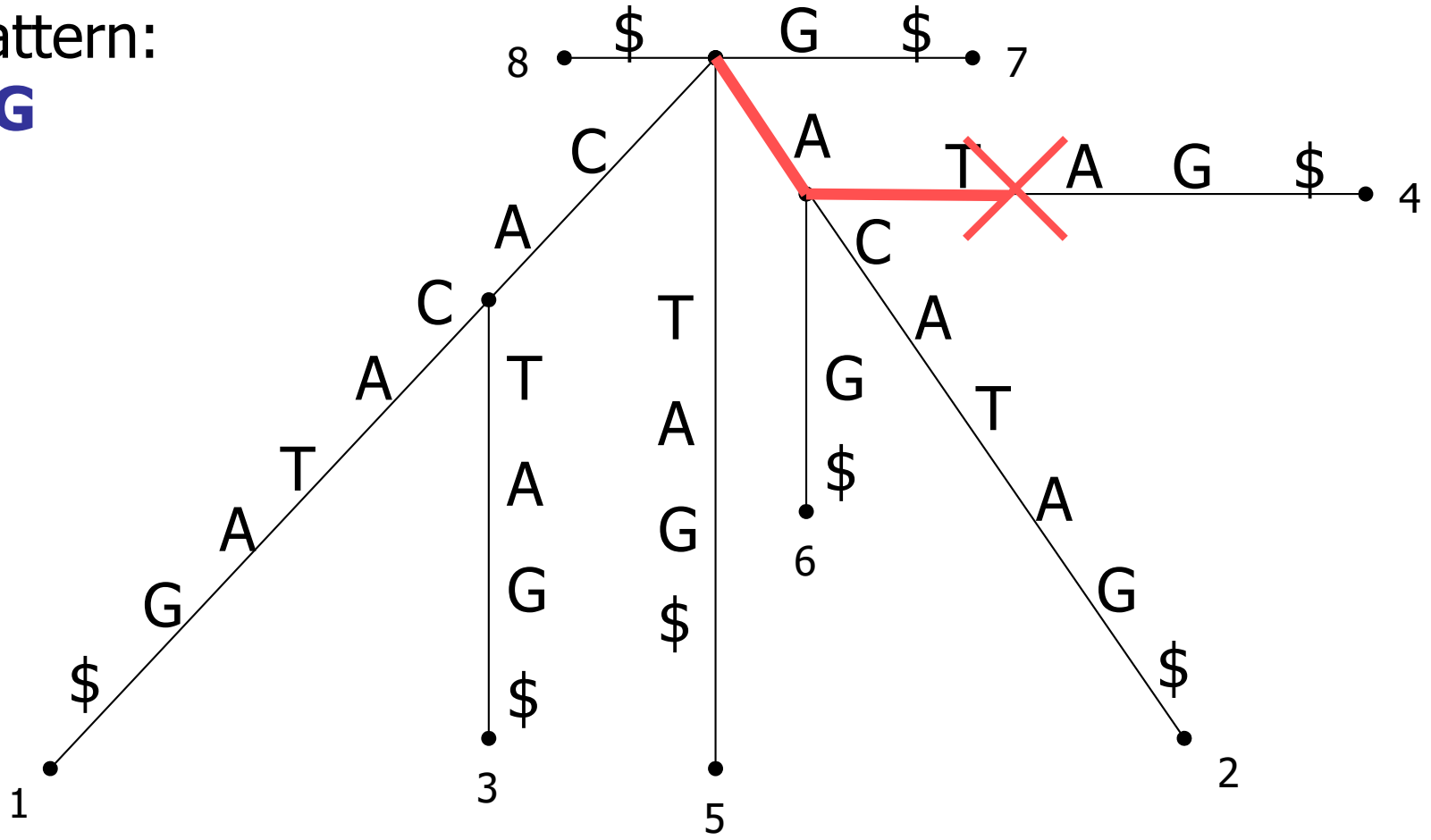
Search Pattern:
CATA



Suchen in einem Suffix-Baum

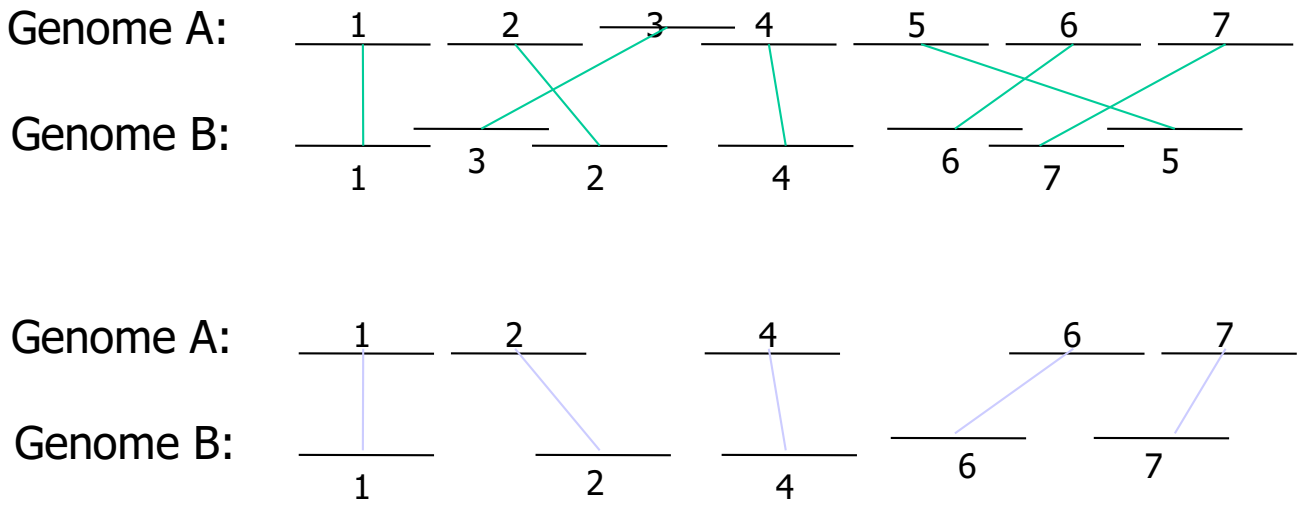
Search Pattern:

ATCG



Sortieren der MUMs

- MUMs werden nach ihren Positionen in Genom A sortiert

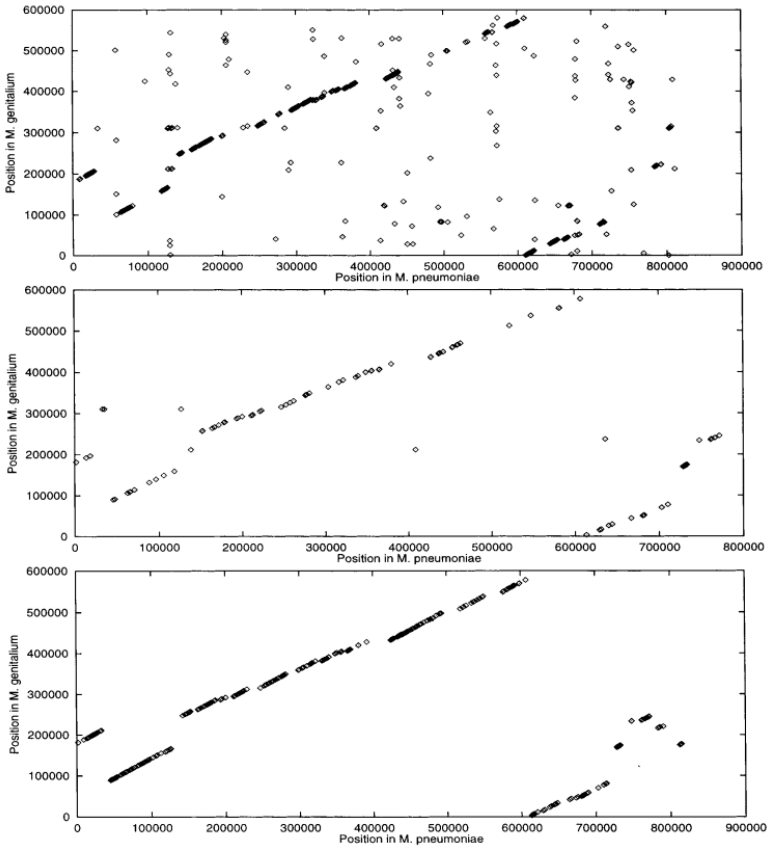


Jeder MUM ist nur mit seiner Nummer gekennzeichnet, ohne Berücksichtigung seiner Länge.

Das obere Alignment zeigt alle MUMs.
Die Verschiebung von MUM 5 in Genom B zeigt eine Transposition an.
Die Verschiebung von MUM 3 könnte ein Zufallstreffer oder Teil einer inexakten Repeat-Sequenz sein.

Unteres Alignment: suche in beiden Genomen die längste gemeinsam ansteigende Folge an Subsequenzen

Beispiel: Alignment zweier Mikroorganismen



Das Genom von *M.genitalium* ist nur etwa 2/3 so lang wie das von *M.pneumoniae*.

Obere Abbildung: FASTA-Alignment von *M.genitalium* und *M.pneumoniae.*

Mitte: Alignment mit 25mers

Unten: Alignment mit MUMs. 5 Translokationen.

Ein Punkt bedeutet jeweils einen Treffer zwischen den Genomen.
FASTA-Plot: ähnliche Gene
25-mer-Plot: 25-Basen-Sequenz, die in beiden Sequenzen genau einmal vorkommt.
MUM-Plot: MUM-Treffer.

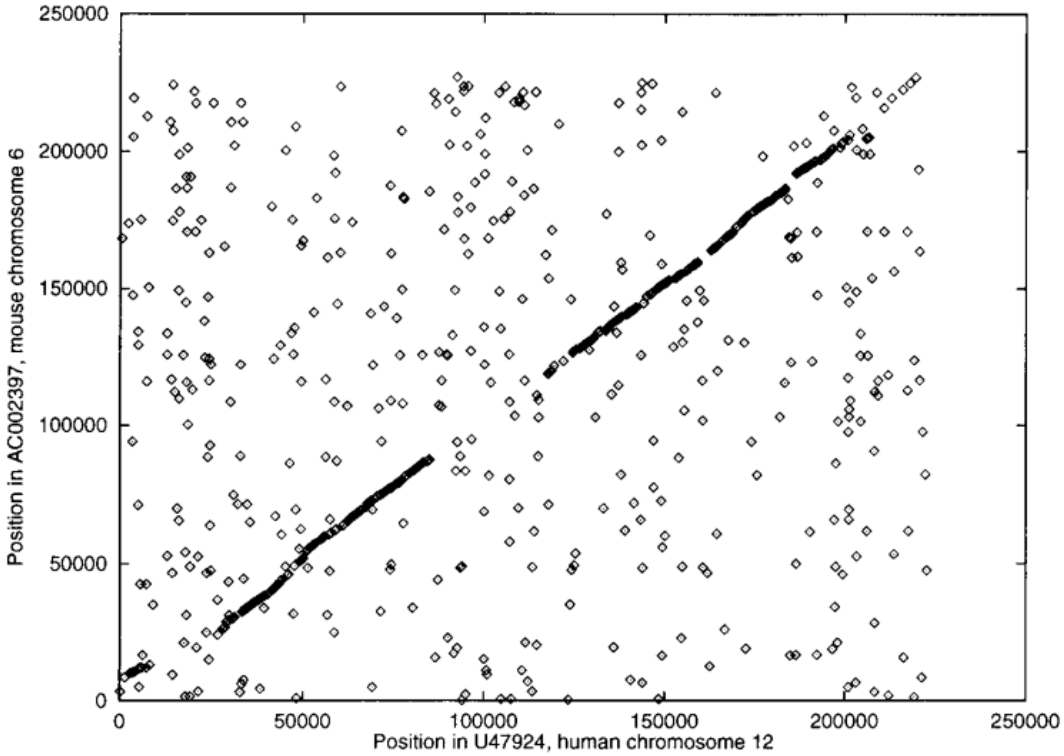
Delcher et al. Nucleic Acids Res 27, 2369 (1999)

Beispiel: Alignment Mensch:Maus

Alignment von weiter entfernt liegenden Spezies:
Mensch gegen Maus.

Hier: Alignment einer 222 930 bp Teilsequenz auf dem menschlichen Chromosom 12, accession no. U47924, gegen eine 227 538 bp lange Teilsequenz des Mauschromosoms 6.

Jeder Punkt des Plots entspricht einem MUM von [ge]15 bp.



Delcher et al. Nucleic Acids Res 27, 2369 (1999)

Zusammenfassung

- Die Anwendung der Suffix-Bäume war ein Durchbruch für die Alignierung ganzer Genome

- MUMmer 2 besitzt zusätzliche Verbesserung für die Rechenzeit und den Speicherplatz
  - die Verwendung von Suffix-Arrays anstatt von Suffix-Bäumen gibt eine verbesserte Datenstruktur ($\rightarrow$ Stefan Kurtz, Hamburg)
  - es wird nun möglich, mehr als zwei Genome zu alignieren (implementiert in MGA)