

# 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

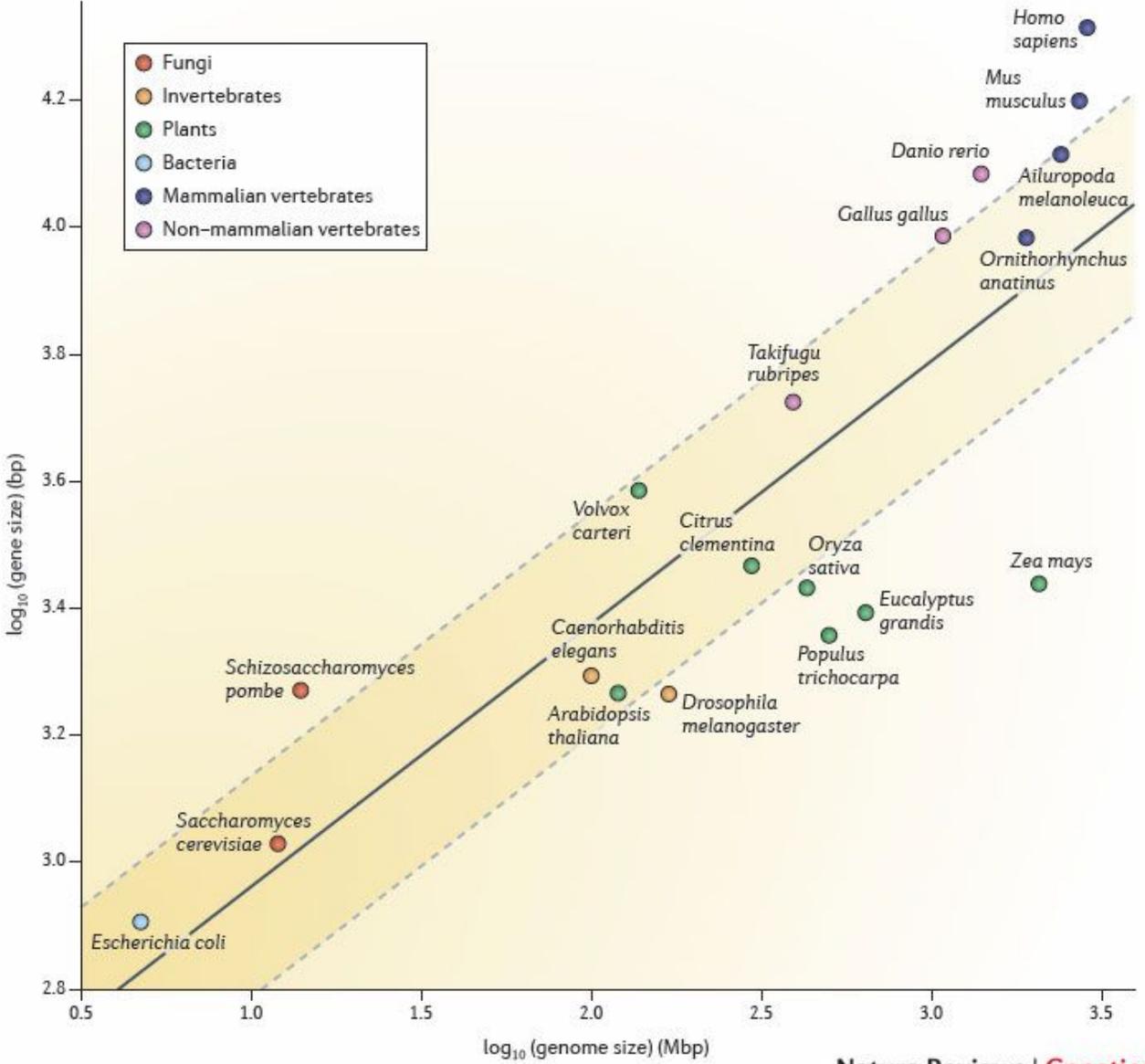
- **Mapping** von **NGS-Daten** auf **Referenzgenom**

- **Alignment zweier Genom-Sequenzen**

  - Suffix Bäume

# Länge von Genen

Generell haben längere Genome auch längere Gene.



Yandell & Ence, *Nature Reviews Genetics* 13, 329–342 (2012)

# 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. kodierende und nicht-kodierende Abschnitte bzw. Intron, Promoter, Exon.*

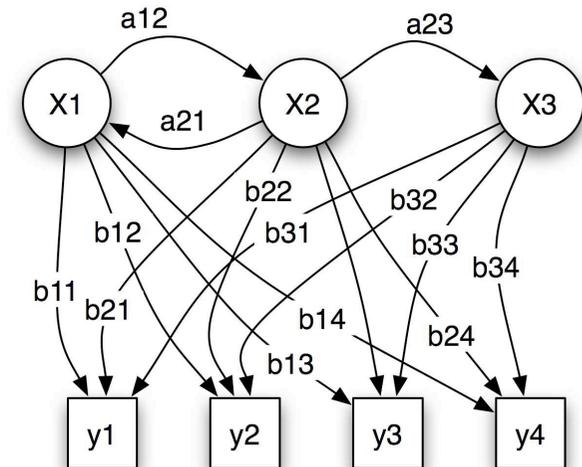
Zwischen den Zuständen X1 und X2 und zurück und von X2 nach X3 sind hier Übergänge erlaubt.

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)

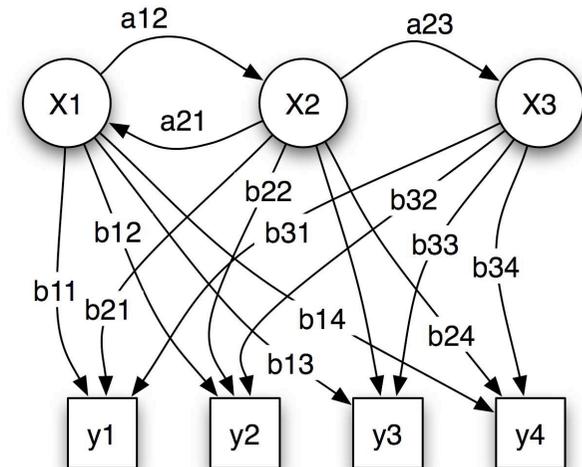
Die Topologie des Graphen gibt an, zwischen welchen Zuständen Übergänge erlaubt sind. Diese gibt man bei der Spezifikation des HMM vor.

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  $a_{ij}$  und  $b_{ij}$  müssen in der Trainingsphase des HMM hergeleitet werden.

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



# Generkennung von prokaryotischen Genen mit Glimmer3

Table 2. Glimmer3 prediction accuracy when trained on confirmed genes

Genome			Glimmer3 Predictions				versus Glimmer2.13			
Organism	GC%	# Genes	3' Matches		5' & 3' Matches	Extra	3' Match	5' & 3'	Extra	
<i>A.fulgidus</i>	49	1165	1162	99.7%	841	72.2%	1308	-2	-67	-59
<i>B.anthraxis</i>	35	3132	3119	99.6%	2717	86.7%	2345	+6	+726	-77
<i>B.subtilis</i>	44	1576	1559	98.9%	1379	87.5%	2886	+11	+413	-539
<i>C.tepidum</i>	57	1292	1284	99.4%	867	67.1%	778	+2	-33	-190
<i>C.perfringens</i>	29	1504	1501	99.8%	1360	90.4%	1177	-1	+244	-28
<i>E.coli</i>	51	3603	3525	97.8%	3014	83.7%	942	+16	+693	-632
<i>G.sulfurreducens</i>	61	2351	2320	98.7%	1883	80.1%	1107	+15	+541	-380
<i>H.pylori</i>	39	915	908	99.2%	785	85.8%	774	+1	+46	-94
<i>P.fluorescens</i>	63	4535	4484	98.9%	3412	75.2%	1896	+14	+731	-704
<i>R.solanacearum</i>	67	2512	2468	98.2%	1922	76.5%	1091	+72	+646	-326
<i>S.epidermidis</i>	32	1650	1646	99.8%	1496	90.7%	767	+3	+338	-66
<i>T.pallidum</i>	53	575	569	99.0%	397	69.0%	568	+3	+55	-296
<i>U.parvum</i>	26	327	325	99.4%	292	89.3%	297	0	+19	-17
Averages:				99.1%		81.1%		+11	+335	-262

Glimmer2 und Glimmer3 verwenden Varianten von Markov-Modellen.

Sie sind sehr erfolgreich (> 99%) bei der **Identifizierung** von prokaryotischen **Genen**. Allerdings ist die akkurate Erkennung des **Genstarts** schwieriger (81.1%).

Delcher et al. Nucl. Acids. Res. 27, 4636 (1999)

# Generkennung mit Hidden Markov Modellen

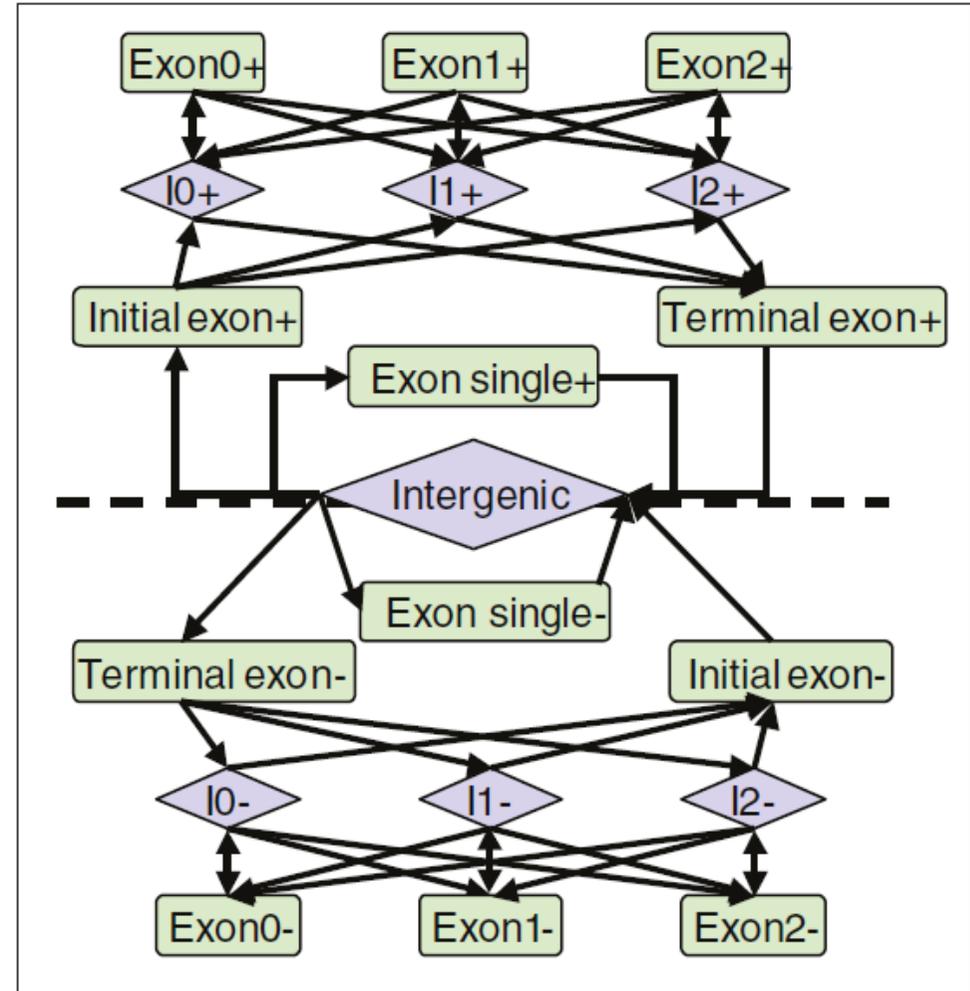
Bei der Generkennung für eukaryotische Gene möchte man bestimmen, wo in einem Genom Exons (E) und Introns (I) sind.

Der Output sind die bekannten exprimierten Sequenzen.

Für eine Eingabesequenz 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.

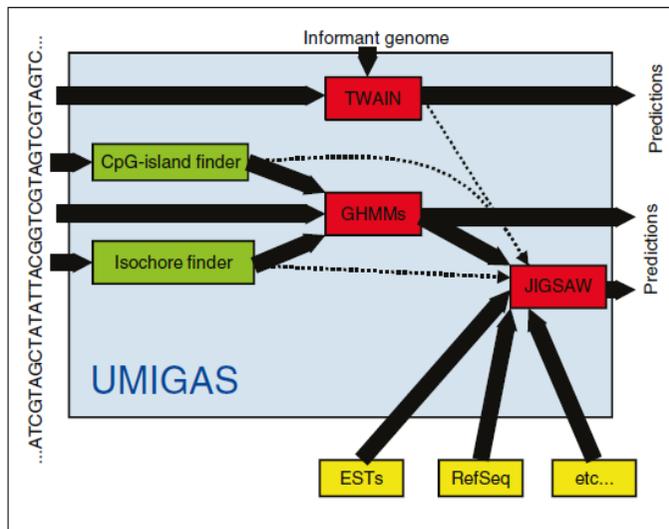
Allen et al. Genome Biol. 7, S9 (2006)



**Figure 4**

State-transition diagram of the GHMM for GlimmerHMM. The dashed line in the middle separates the positive strand and negative strand portions of the model. Each state in the GHMM is implemented as a separate submodel, such as a weight array matrix or an IMM (interpolated Markov models).

# Generkennung von menschlichen Genen mit JIGSAW



Durch Hinzunahme zusätzlicher Information konnten etwa  $\frac{3}{4}$  der menschlichen Gene präzise vorhergesagt werden. Nur 3% der Gene wurden überhaupt nicht gefunden.

Table 2

## Results of applying JIGSAW with all available evidence

	Gene Sn	Gene Sp	Exon Sn	Exon Sp	Nuc Sn	Nuc Sp	Missed Genes	Missed Exons	Inserted Exons
JIGSAW-mRNA	48%	60%	76%	93%	84%	97%	17%	11%	4%
JIGSAW-All-EST	52%	52%	77%	88%	91%	91%	6%	10%	8%
JIGSAW-KnownGene	71%	74%	76%	95%	87%	96%	7%	12%	3%
JIGSAW-All	74%	70%	80%	92%	93%	94%	3%	7%	6%
KnownGene	77%	73%*	78%	82%	89%	94%	13%	10%	4%
JIGSAW-EGASP	73%	66%	81%	89%	95%	92%	4%	6%	8%

\* KnownGene predicts multiple transcripts per gene locus with transcript specificity of 47%. The percentage of test genes and exons that do not overlap a prediction are listed in the Missed Genes and Missed Exons columns, respectively. The rightmost column shows the percentage of predicted exons inserted into true introns. See text for details. Nuc, nucleotide; Sn, sensitivity; Sp, specificity.

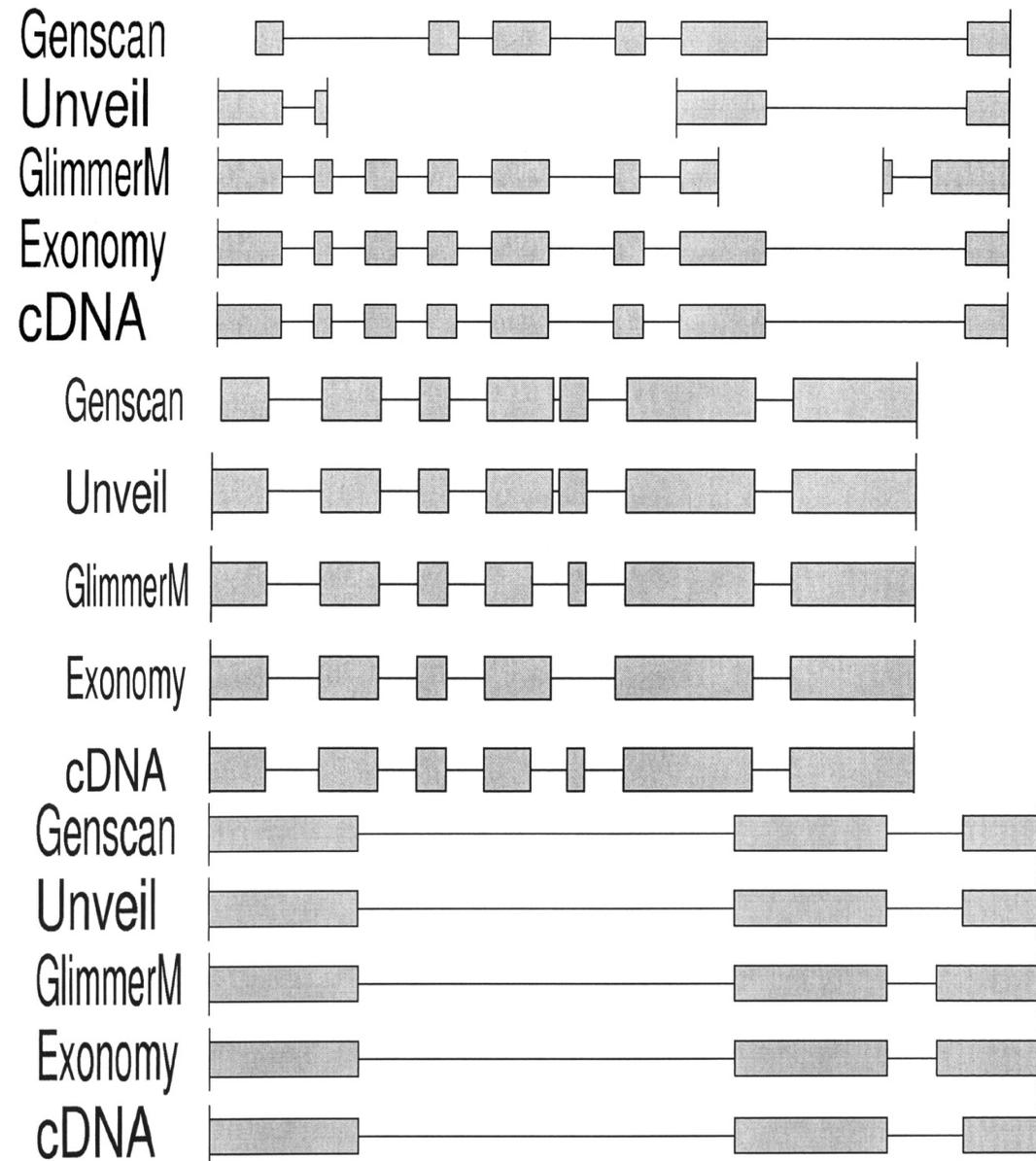
Allen et al. Genome Biol. 7, S9 (2006)

# Vergleich von Genvorhersage-Methoden

Ein Beispiel, in dem Exonomy die Gene richtig erkennt.

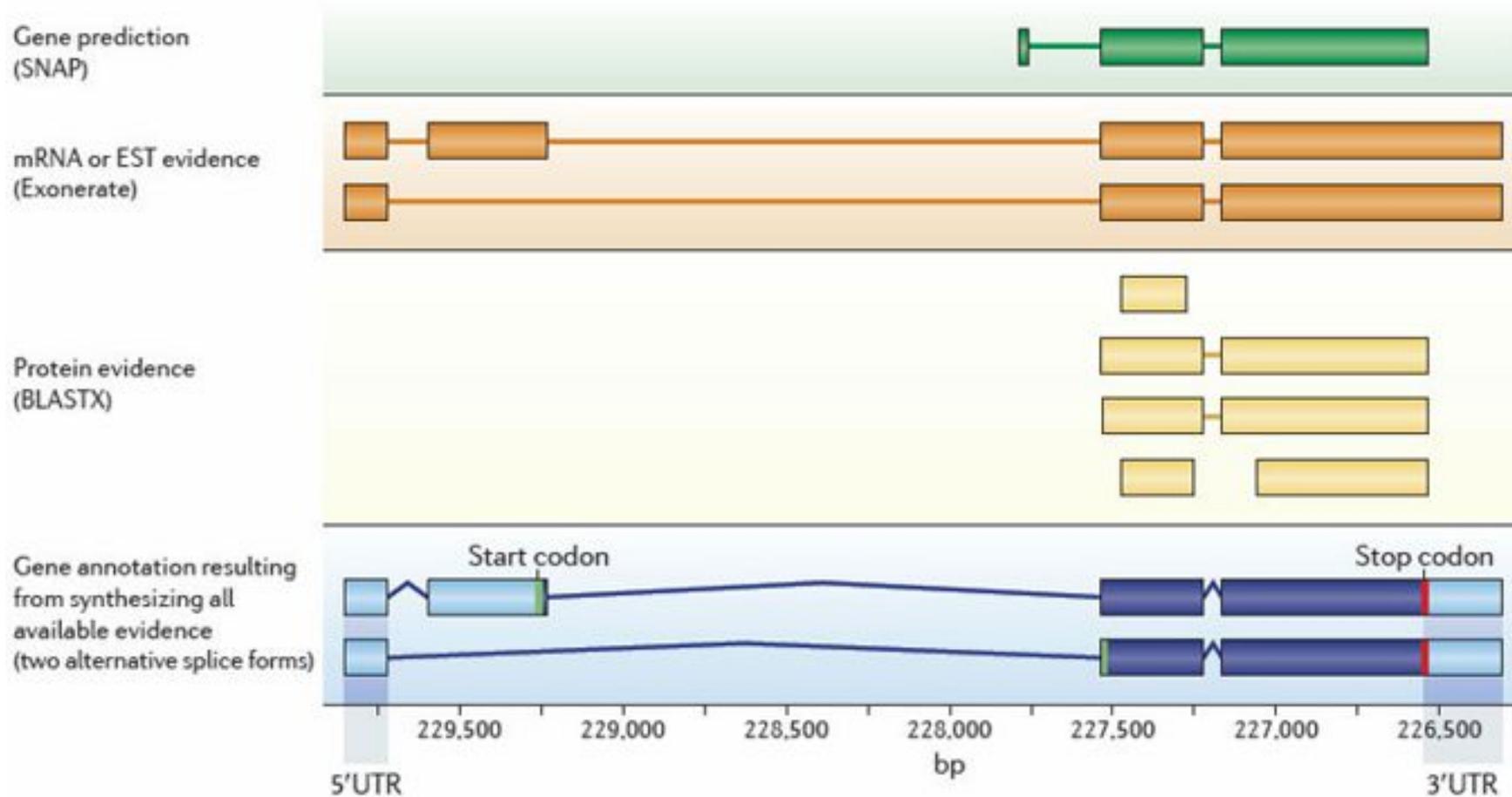
Ein Beispiel, in dem GlimmerM die Gene richtig erkennt.

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



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

# Vorhersage von Genen

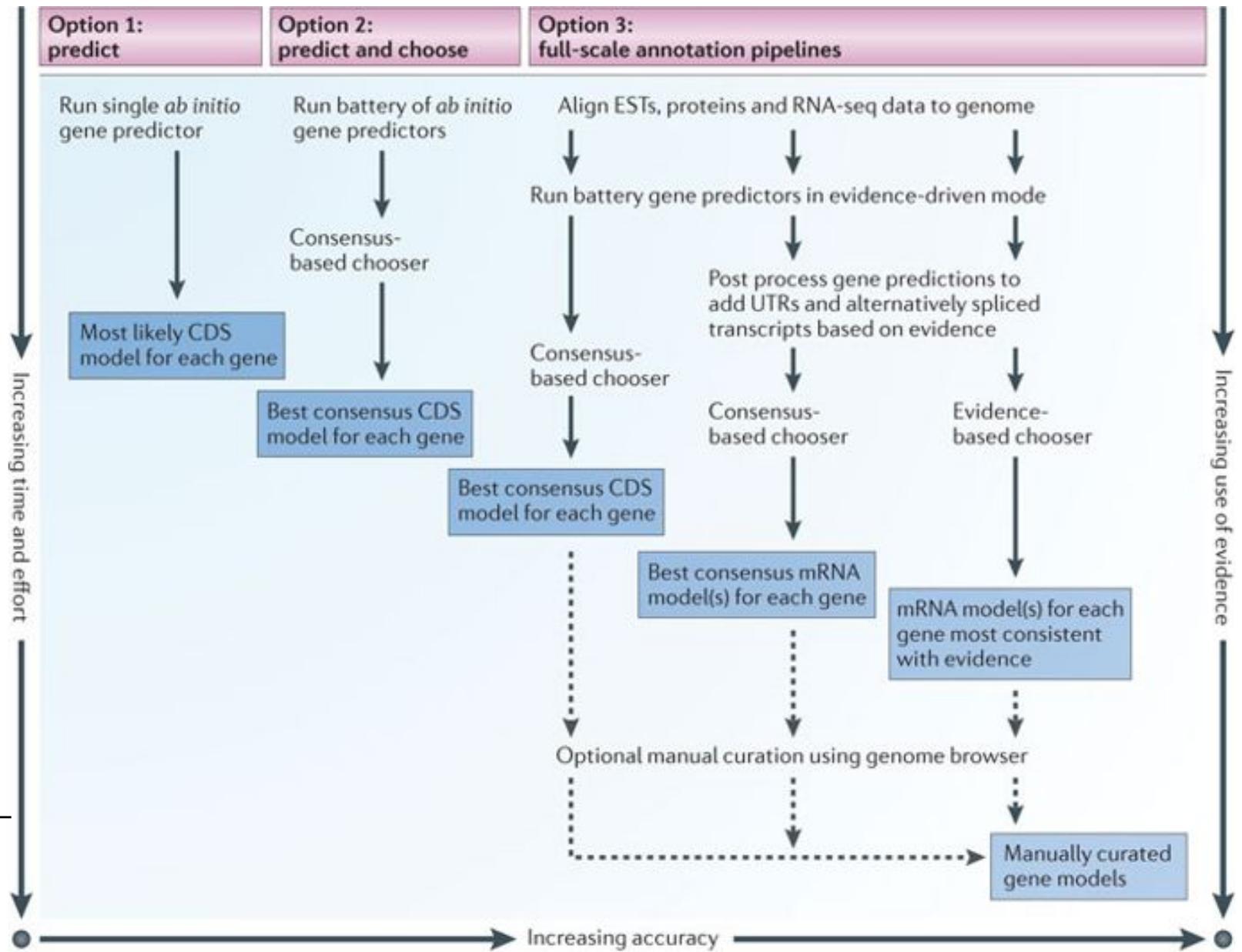


Nature Reviews | Genetics

Yandell & Ence, *Nature Reviews Genetics* 13, 329–342 (2012)

Tools zur Vorhersage von Gene ermitteln nur die kodierenden Abschnitte!

# Annotation von Genomen



Yandell & Ence,  
*Nature Reviews  
 Genetics* 13, 329–  
 342 (2012)

## Promotervorhersage in *E.coli*

Um *E.coli* Promoter zu analysieren kann man eine Menge von Promotersequenzen 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

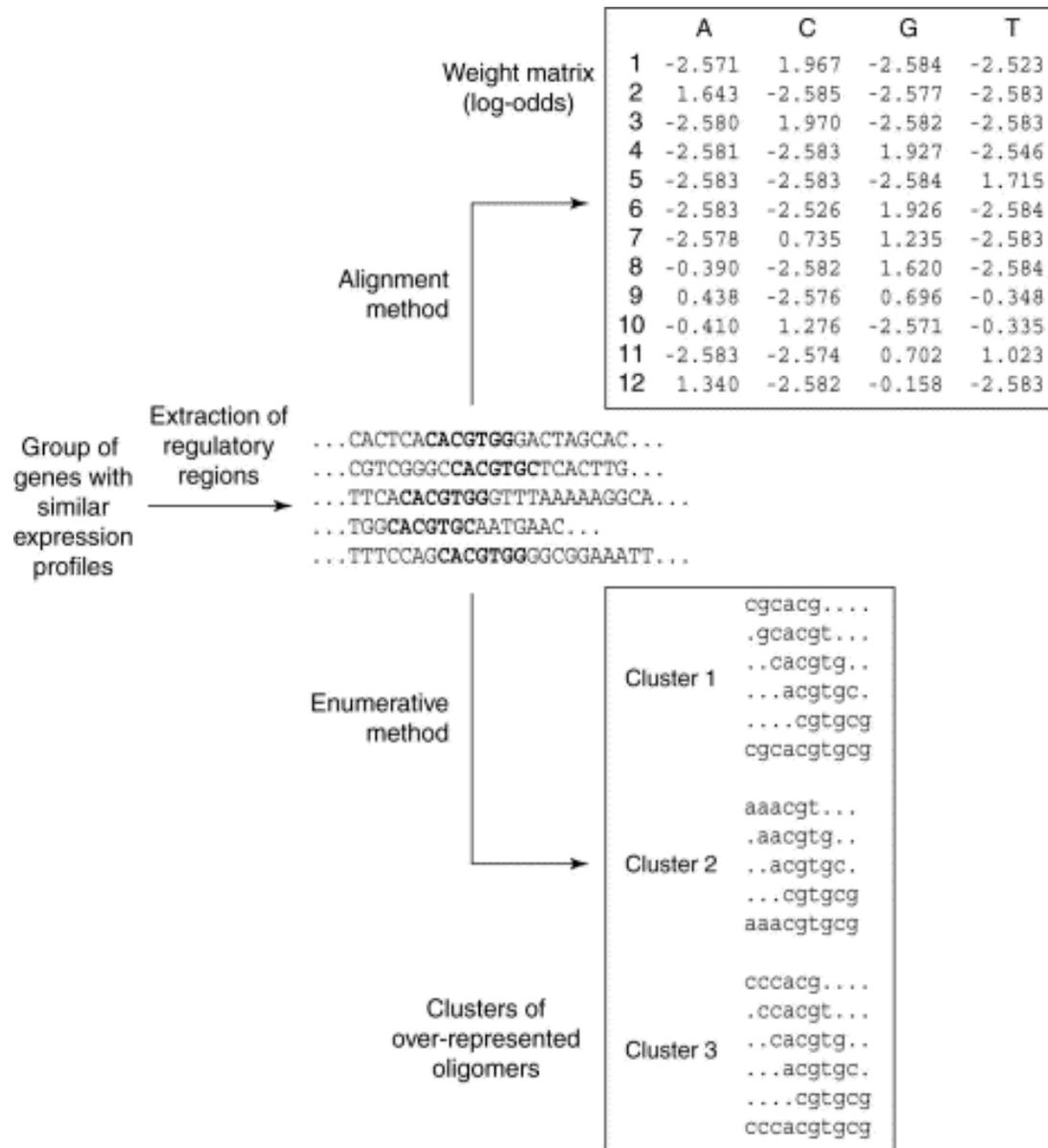
Wird seit der Verfügbarkeit von Microarray Gen-Expressionsdaten eingesetzt.

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



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?

a) Alignment Matrix

	A	A	T	T	G	A
	A	G	G	T	C	C
	A	G	G	A	T	G
	A	G	G	C	G	T
	1	2	3	4	5	6
A	4	1	0	1	0	1
C	0	0	0	1	1	1
G	0	3	3	0	2	1
T	0	0	1	2	1	1

consensus: A G G T G N

$$\ln \frac{(n_{i,j} + p_i) / (N + 1)}{p_i} \approx \ln \frac{f_{i,j}}{p_i}$$

b) Weight Matrix

	1	2	3	4	5	6
A	1.2	0	-1.6	0	-1.6	0
C	-1.6	-1.6	-1.6	0	0	0
G	-1.6	.96	.96	-1.6	.59	0
T	-1.6	-1.6	0	.59	0	0

test sequence: A G G T G C

**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.

# 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: MY0001
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: MY0002
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: MY0003
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

## TRANSFAC<sup>®</sup> 6.0 - Public

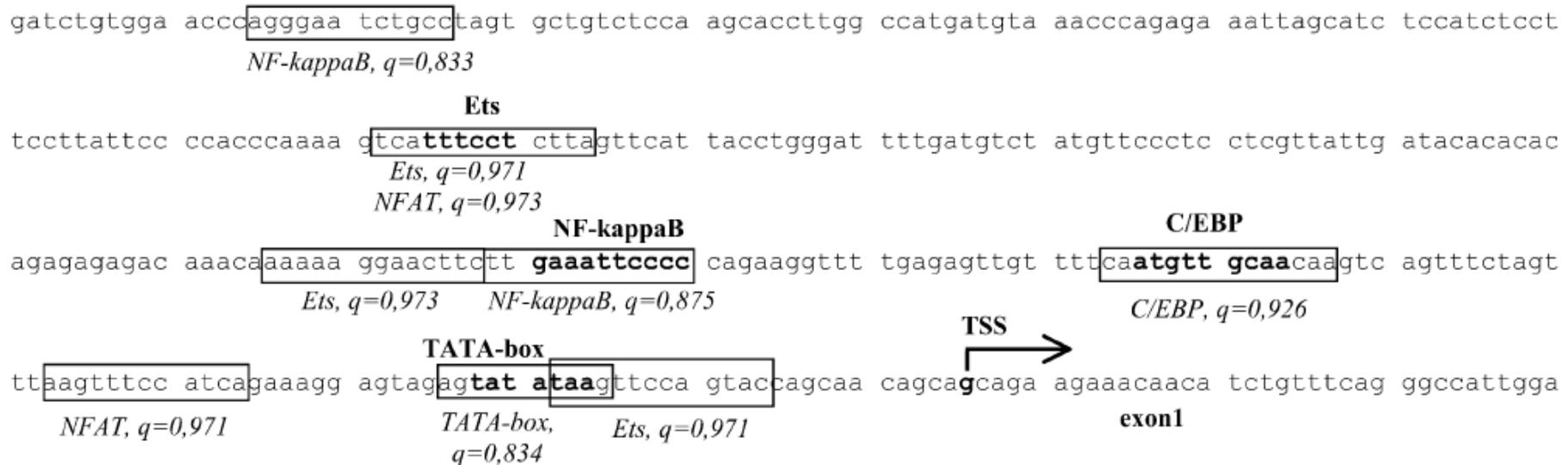
TRANSFAC<sup>®</sup> is the database on eukaryotic transcription factors, their genomic binding sites and DNA-binding profiles.

- [Search](#)
- [Classification](#)
- [Documentation](#)
- [Fungal Metabolic](#)
- [Pax factors in TRANSFAC<sup>®</sup>](#)
- [The green site of TRANSFAC<sup>®</sup>](#)
- [Quality Management in TRANSFAC<sup>®</sup>](#)
- [Tfblast: Search Tool for Sequence Search in the TRANSFAC<sup>®</sup> Factor Table](#)

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.

Wegen der relativen kurzen Länge der TFBS-Motive und den wenigen invarianten Nukleotiden darin gibt es manche Motive Millionen mal im menschlichen Genom. Nur etwa 1 aus 500 ist tatsächlich davon mit einem TF besetzt!

Z.B. enthält das Maus-Genom ~8 Millionen Treffer für das Bindemotiv des GATA-binding factor 1. Aber nur ~15,000 DNA Segmente hatten den TF in roten Blutzellen wirklich gebunden (Hardison and Taylor, 2012).

# 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 AAAAAAAAAA).

## 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>

# Prozessierung von NGS-Daten

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

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

Danksagung für Folien: Mohamed Hamed

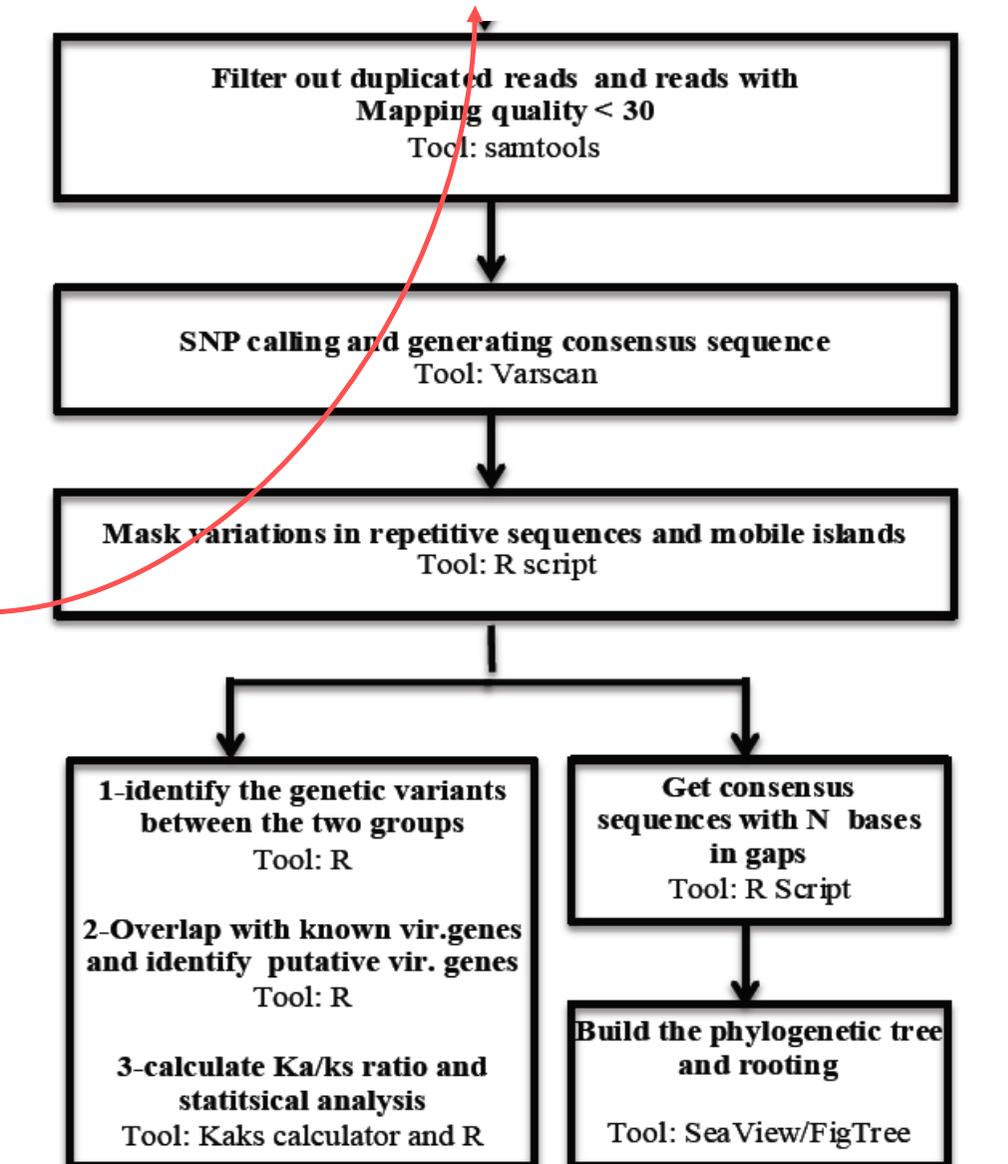
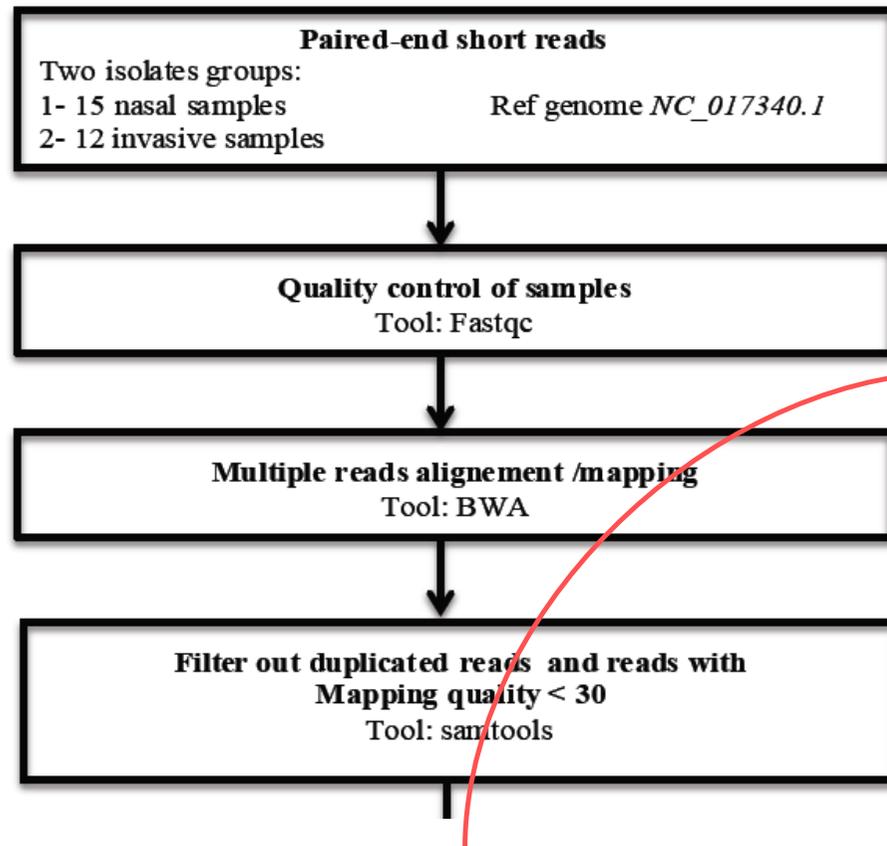
# NGS Pipeline im Überblick



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 Charakterisierung (z.B. Virulenzfaktoren)

Wir konzentrieren uns auf die Schritte 5 und 6

# WGS Pipeline für bakterielle Phylotypisierung



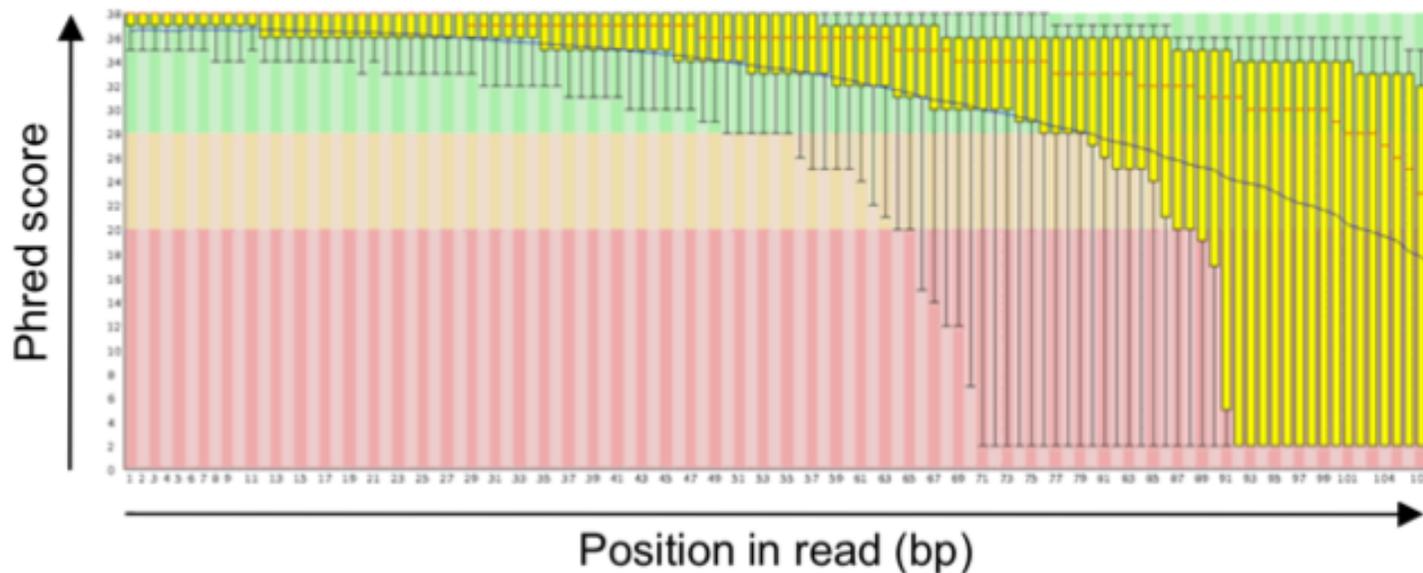
# Quality (Phred) score

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-Schritt „getrimmt“, d.h. gekürzt.

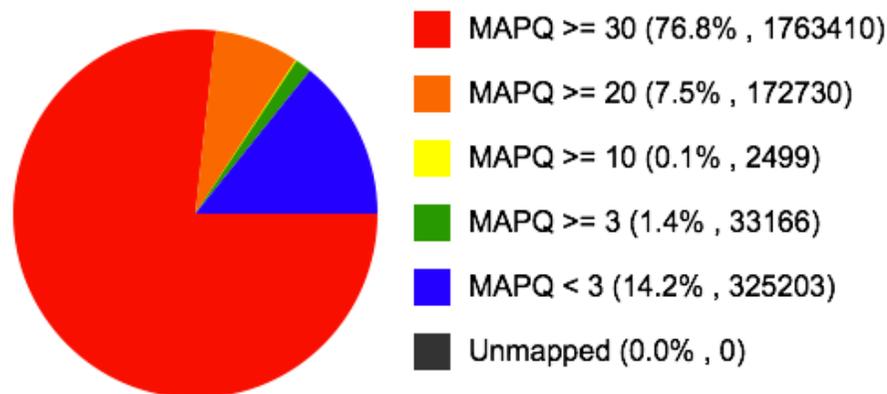


# Qualitätskontrolle im Alignment-Schritt

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**

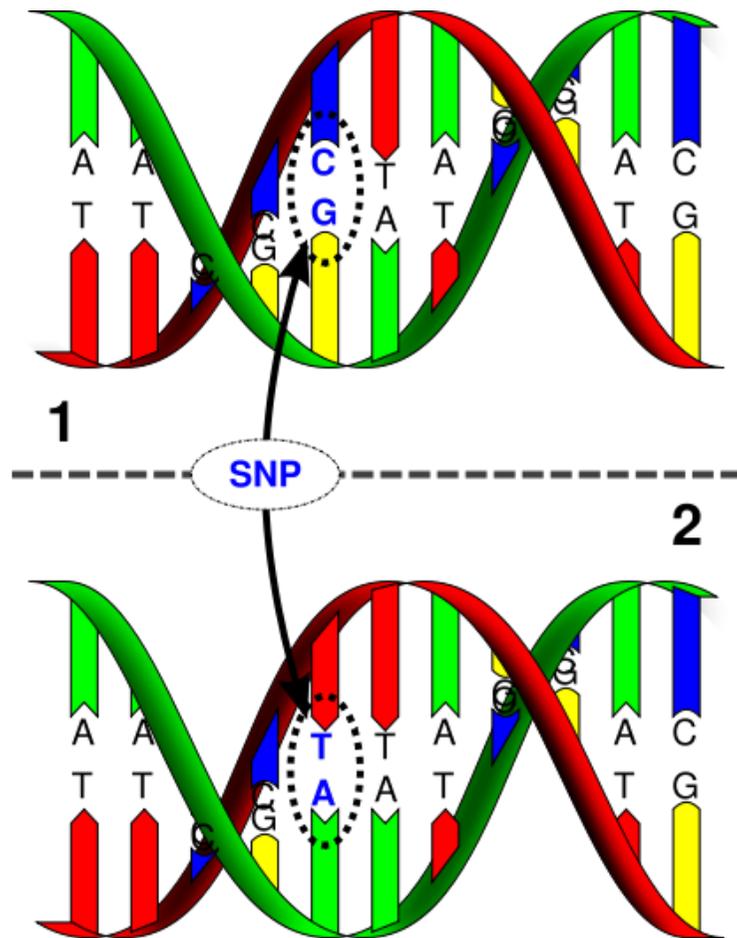
Timo Lassmann\*, Yoshihide Hayashizaki and Carsten O. Daub



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



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

[http://www.science.marshall.edu/murraye/341/Images/416px-Dna-SNP\\_svg.png](http://www.science.marshall.edu/murraye/341/Images/416px-Dna-SNP_svg.png)

# 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>

15 tracks loaded | gi|387149188|ref|NC\_017340.1 |

954M of 1,484M

<p>somatic.unl.vcf</p> <p>SAMPLE</p>		102 bp	<p>id0 region</p> <p>ID=id0 Dbxref=taxon:703339 Is_circular=true gbkey=Src genome=chromosome mol_type=genomic DNA strain=04-02981 region</p>
<p>S4sfr.bam Coverage</p> <p>S4sfr.bam</p>	<p>[0 - 38]</p> <p style="text-align: center;">G A G G T T</p>	T T T T	<p>ID=id0 Dbxref=taxon:703339 Is_circular=true gbkey=Src genome=chromosome mol_type=genomic DNA strain=04-02981</p> <p>gi 387149188 ref NC_017340.1 :1-2821452</p> <p>id73 region</p> <p>ID=id73 Note=phage phiN315 is also present in Staphylococcus au ... gbkey=mobile_element region</p>
<p>S5sfr.bam Coverage</p> <p>S5sfr.bam</p>	<p>[0 - 48]</p>	T T T T	<p>ID=id73 Note=phage phiN315 is also present in Staphylococcus au ... gbkey=mobile_element region</p>
<p>S6sfr.bam Coverage</p> <p>S6sfr.bam</p>	<p>[0 - 48]</p>	T T T T	<p>gbkey=mobile_element</p> <p>gi 387149188 ref NC_017340.1 :2038751-2082550</p> <p>id74 region</p> <p>ID=id74 gbkey=mobile_element region</p>
<p>R17sfr.bam Coverage</p> <p>R17sfr.bam</p>	<p>[0 - 180]</p>	T T T T	<p>region</p> <p>ID=id74 gbkey=mobile_element region</p> <p>gi 387149188 ref NC_017340.1 :2038751-2082550</p>
<p>R18sfr.bam Coverage</p> <p>R18sfr.bam</p>	<p>[0 - 180]</p> <p style="text-align: center;">A</p>	T T T T	<p>SA2981_1900 gene1961 gene</p> <p>ID=gene1961 Name=SA2981_1900 Dbxref=GeneID:12864764 gbkey=Gene locus_tag=SA2981_1900 gene</p>
<p>R19sfr.bam Coverage</p> <p>R19sfr.bam</p>	<p>[0 - 315]</p> <p style="text-align: center;">C</p>	T T T T	<p>ID=gene1961 Name=SA2981_1900 Dbxref=GeneID:12864764 gbkey=Gene locus_tag=SA2981_1900 gene</p>
<p>Sequence</p> <p>Gene</p>	<p>ATAGGATTAAATTATTTATTTATTTATGAAGGGTGTGATTGTTATTACCA</p>	29	<p>gi 387149188 ref NC_017340.1 :2039245-2039505</p>

Softwarewerkzeuge

4. Vorlesung WS 2017/18

# 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 Verwandtschaftsverhältnissen zwischen den Stämmen.

Ausgabe: phylogenetischer Baum

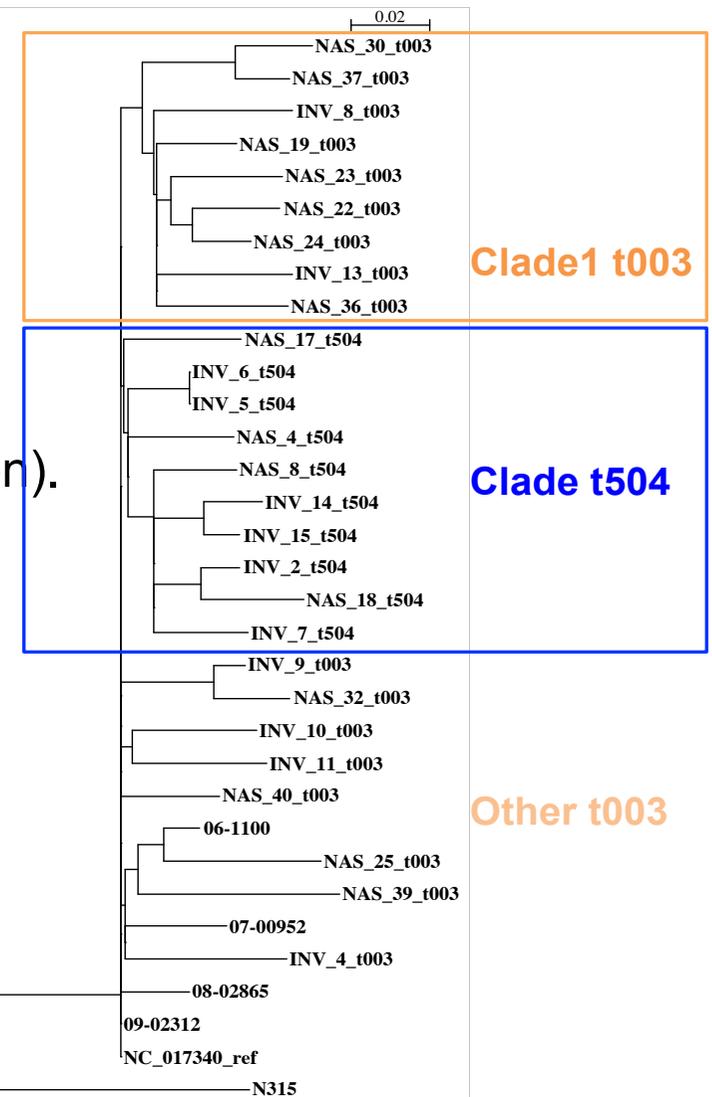
## •Tools

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

CC5

ST225

ST5



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

# Whole Genome Alignment (WGA)

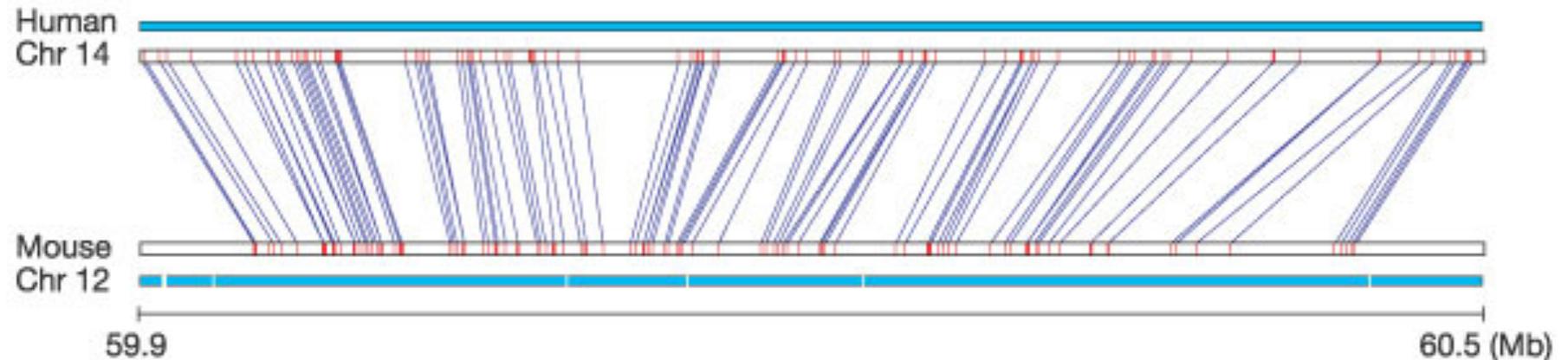
Wenn die genomische DNA-Sequenz eng verwandter Organismen verfügbar wird, kann man ein Alignment von zwei Genomen konstruieren.

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

Im anderen Fall muss man zuerst die genomischen Rearrangements betrachten.

Dann kann man die **systemischen 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



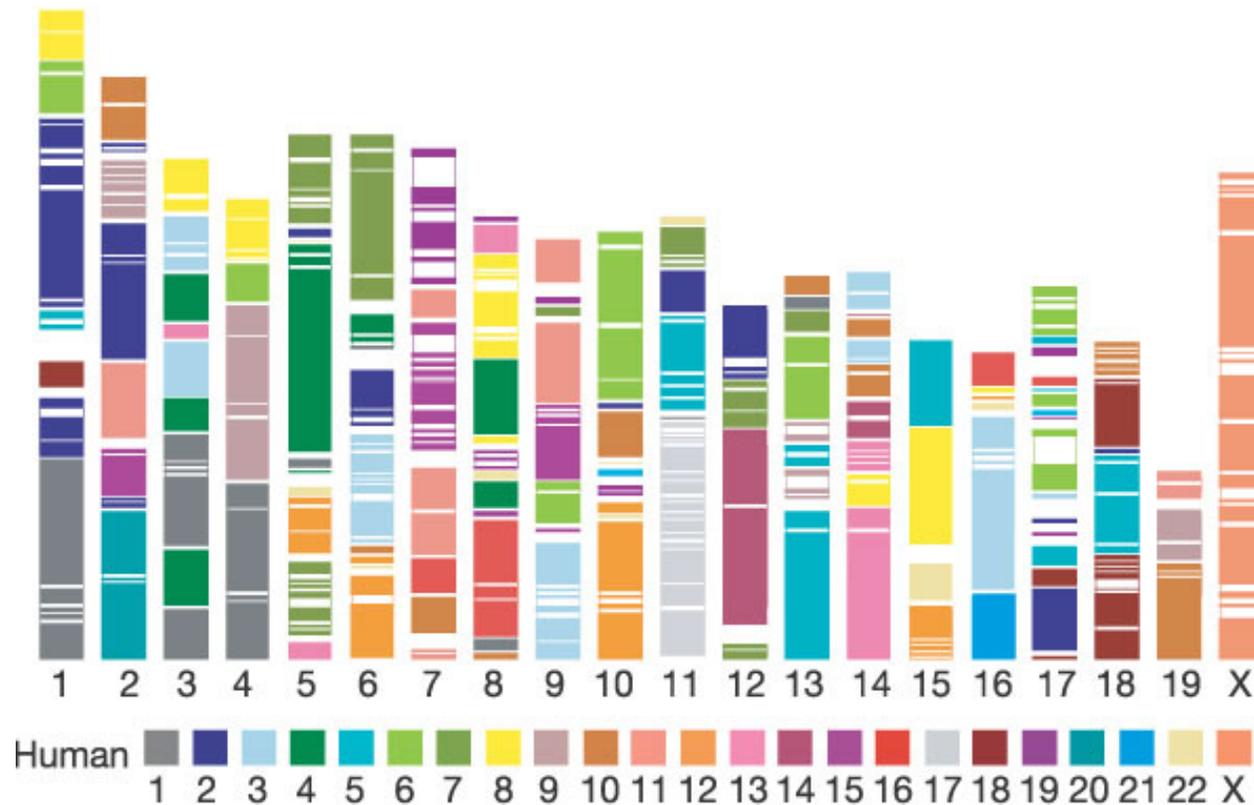
Ein typisches 510-kb Segment des Maus-Chromosoms 12, das mit einem 600-kb Stück des menschlichen Chromosom 14 verwandt ist.

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.

# 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.

# 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: tcgatcGACGATCGCGCCGTAGATCGAATAACGAGAGAGCATAAcgactta

Genome B: gcattaGACGATCGCGCCGTAGATCGAATAACGAGAGAGCATAAtccagag

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)

ACTGATTACGTGAACTGGATCCA

ACTCTAGGTGAAGTGATCCA



**ACT**GATTAC**GTGAA**CTGGAT**TCCA**

**ACT**CTAG**GTGAA**GTGAT**TCCA**



1                      10                      20  
**ACT**GATTAC**GTGAA**CTGGAT**TCCA**

1                      10                      20  
**ACT**C--TAG**GTGAA**GTG-A**TCCA**

## 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?

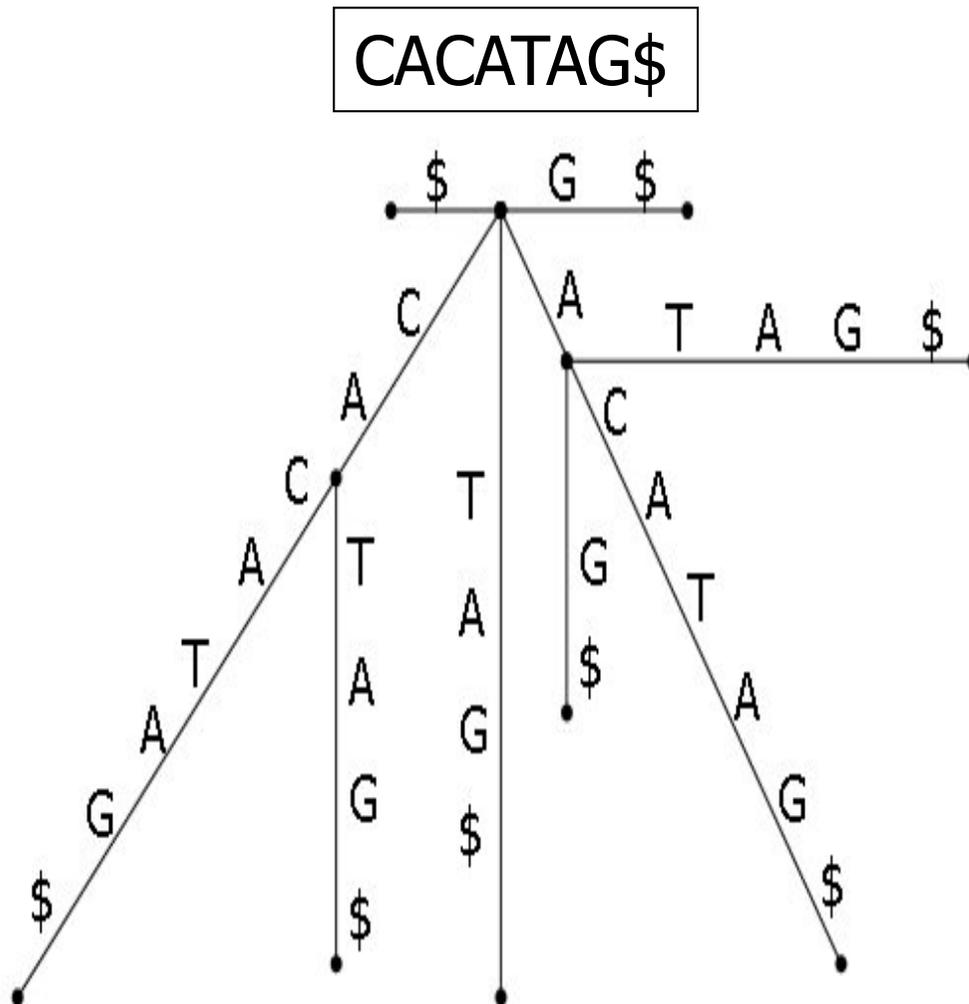
- Naiver Ansatz
  - Vergleiche alle Teilsequenzen von A mit allen Teilsequenzen von B.  
Dies dauert  $O(n^n)$
- verwende Suffix-Bäume als Datenstruktur
  - ein naiver Ansatz, einen Suffix-Baum zu konstruieren hat eine quadratische Komplexität in der Rechenzeit und dem Speicherplatz
  - durch cleverere 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  $N$  Suffixes.
- 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

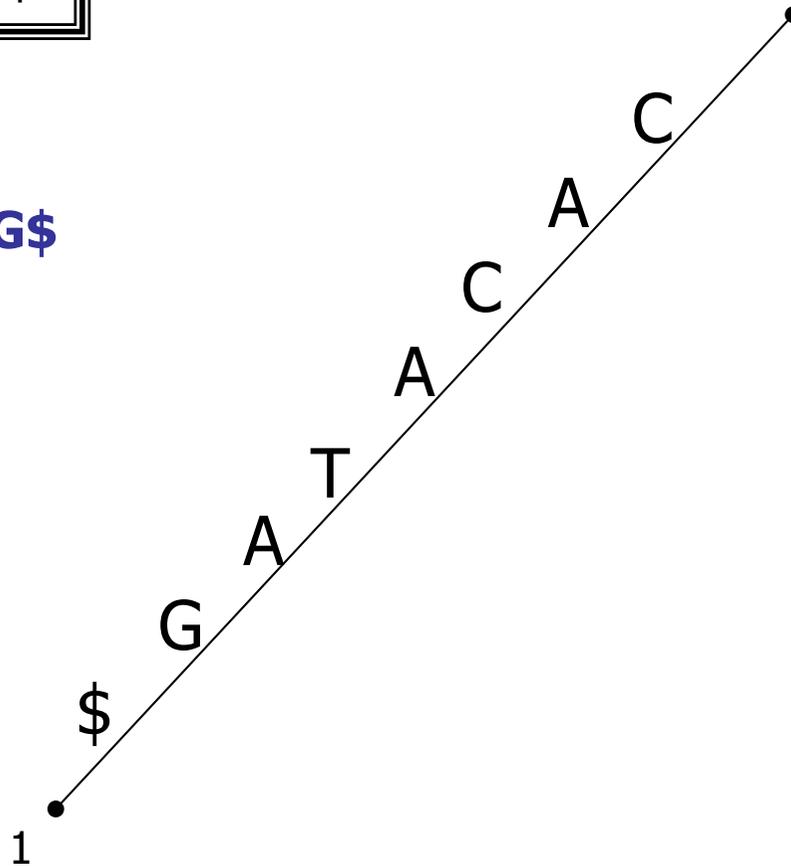


# Konstruktion eines Suffix-Baums

CACATAG\$

Suffixes:

**1. CACATAG\$**

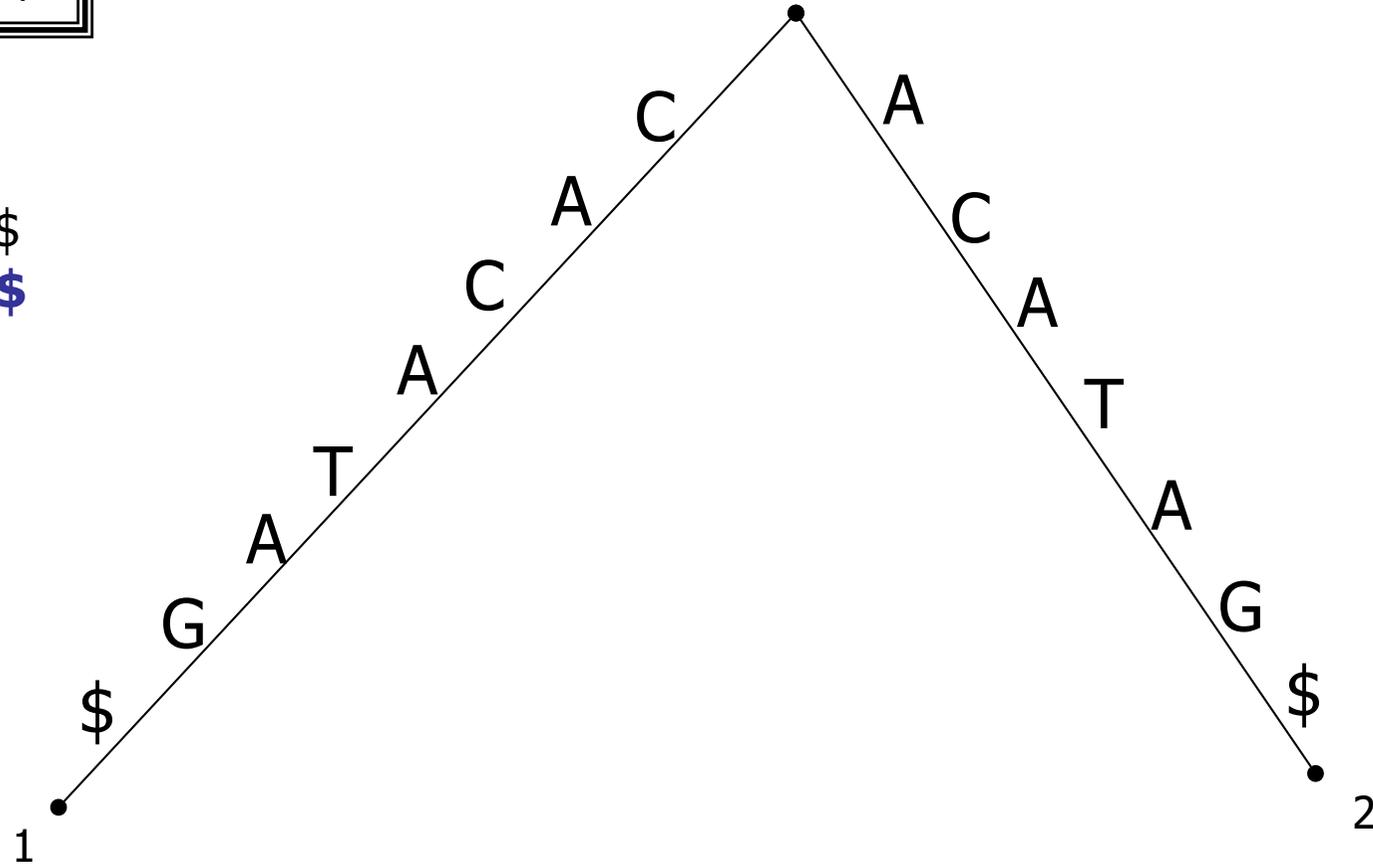


# Konstruktion eines Suffix-Baums

CACATAG\$

Suffixes:

1. CACATAG\$
2. ACATAG\$

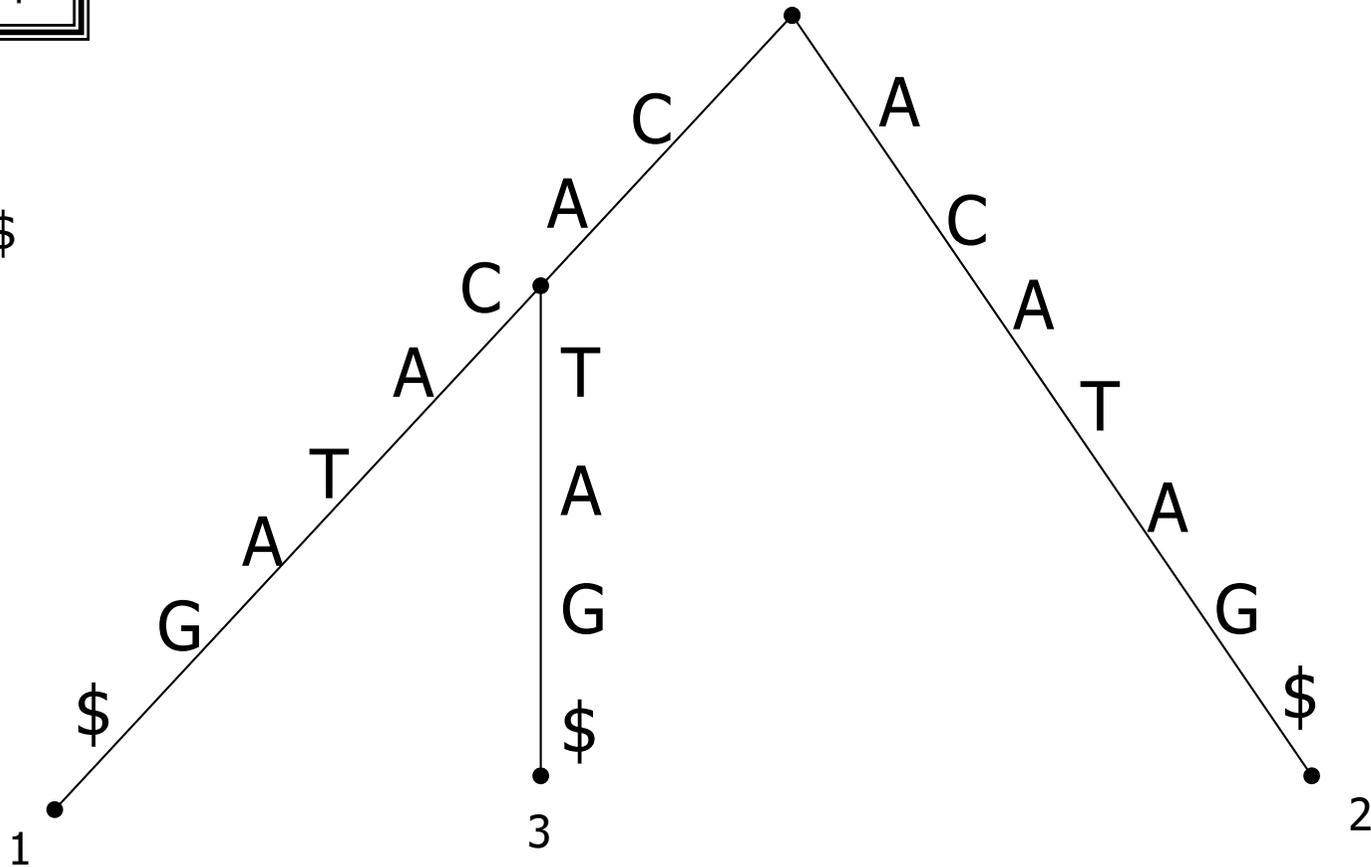


# Konstruktion eines Suffix-Baums

CACATAG\$

Suffixes:

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

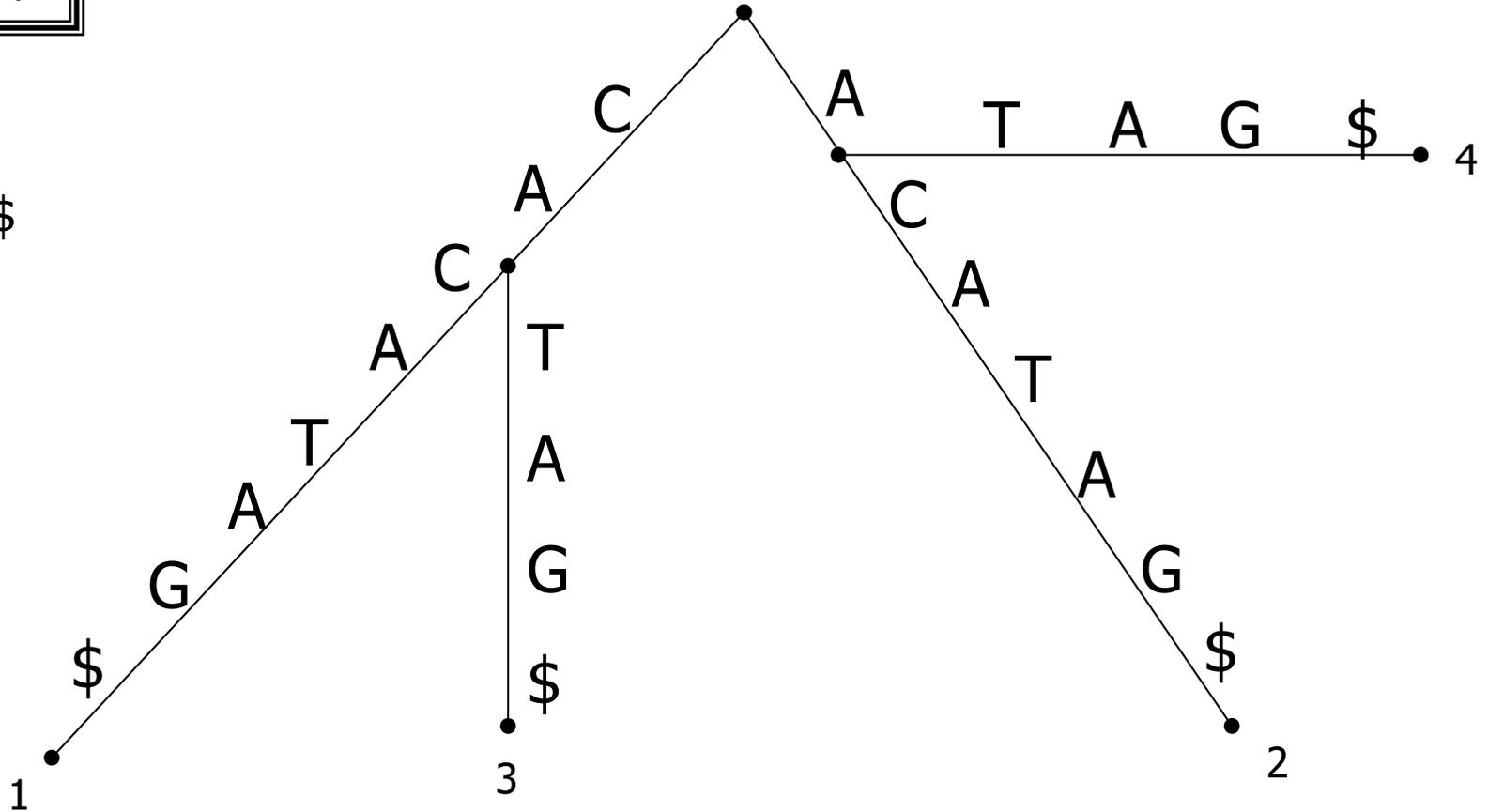


# Konstruktion eines Suffix-Baums

CACATAG\$

Suffixes:

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

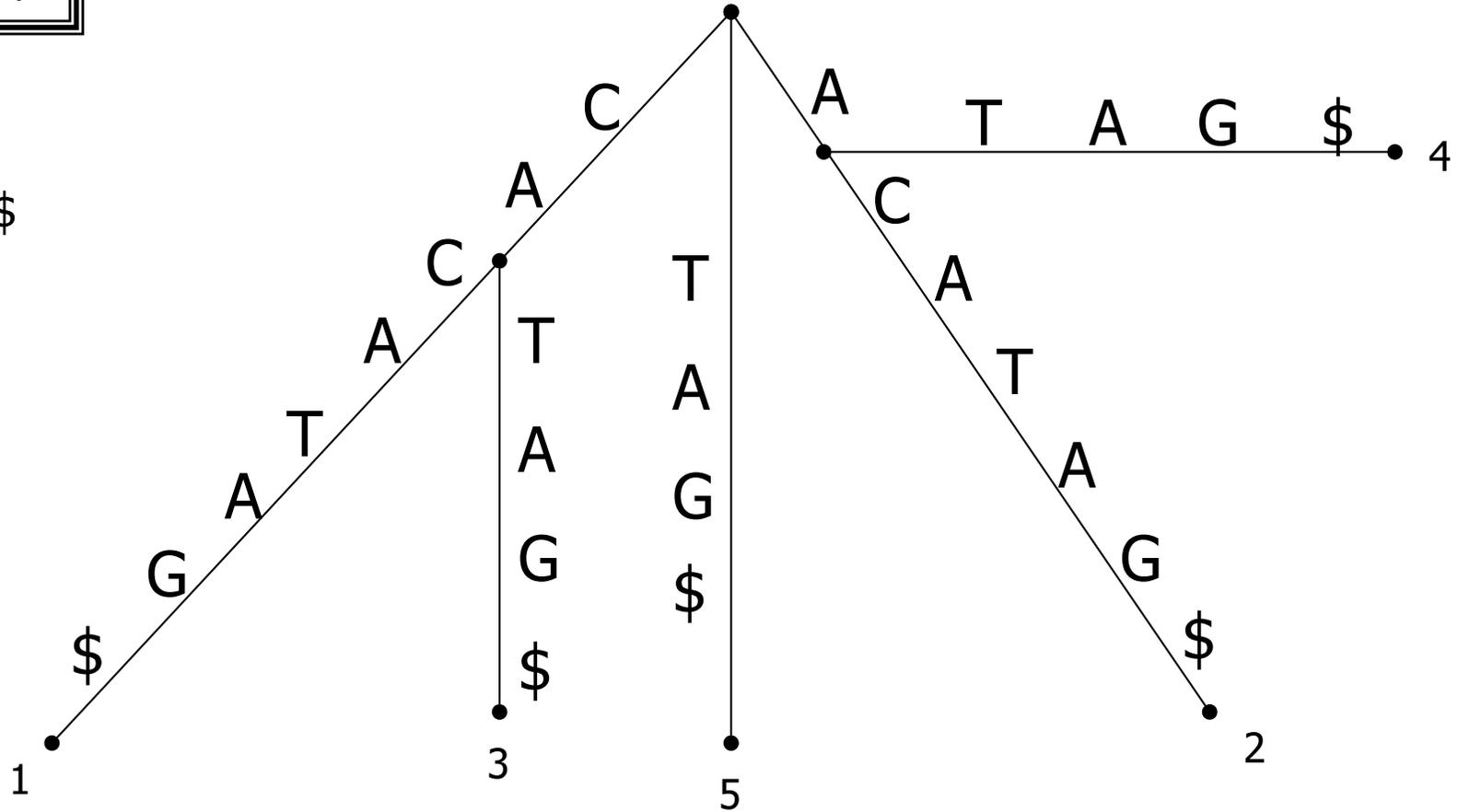


# Konstruktion eines Suffix-Baums

CACATAG\$

Suffixes:

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



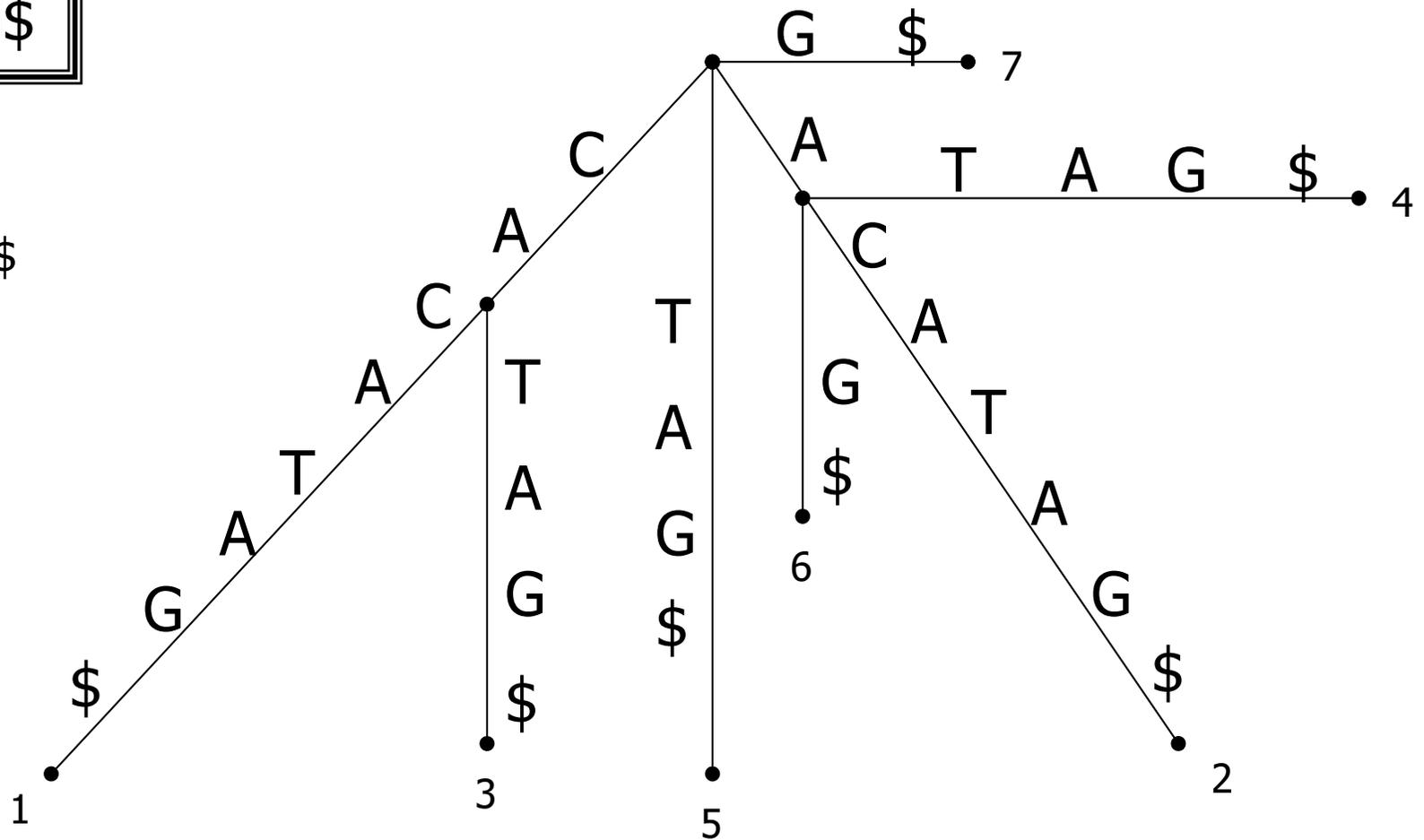


# Konstruktion eines Suffix-Baums

CACATAG\$

Suffixes:

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

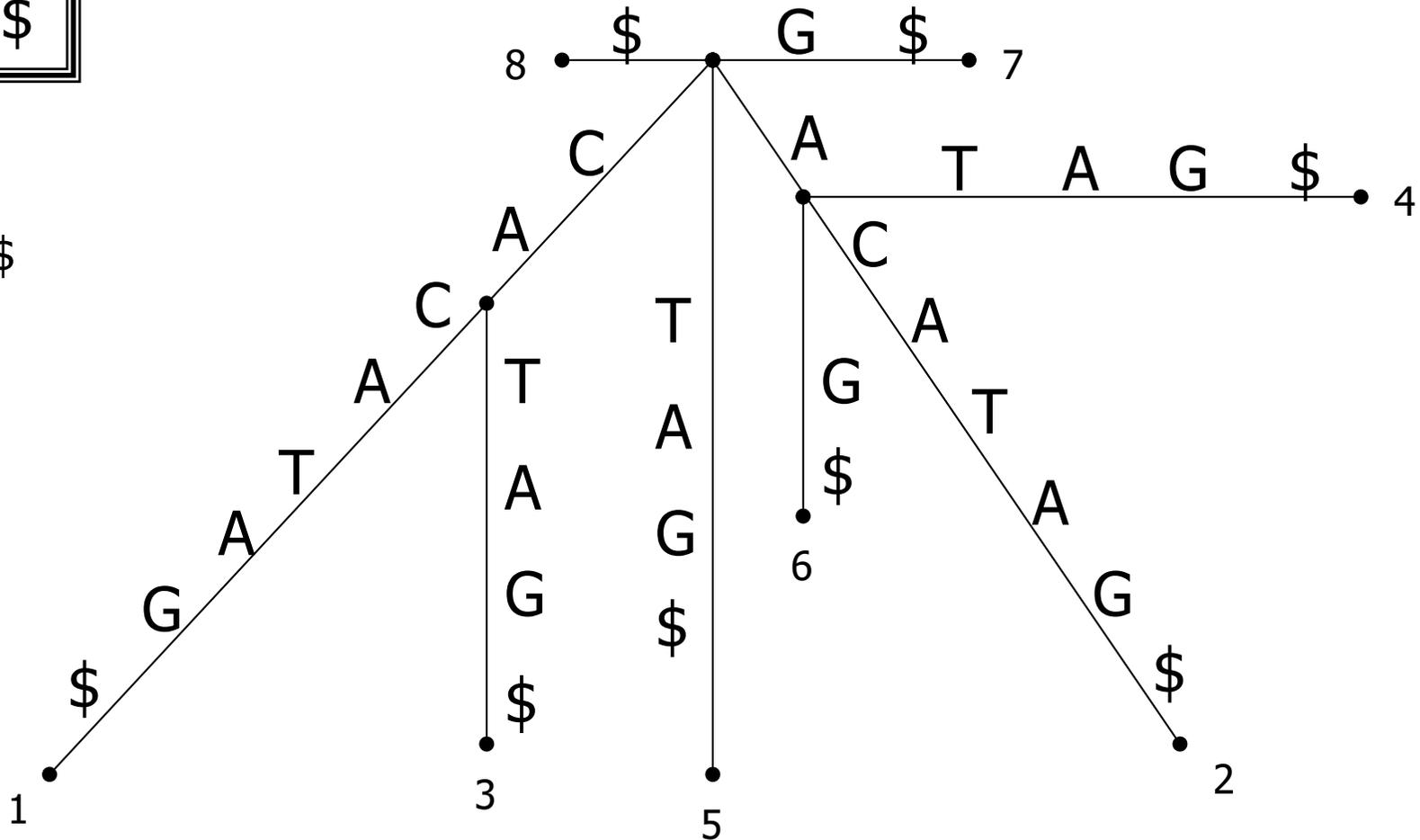


# Konstruktion eines Suffix-Baums

CACATAG\$

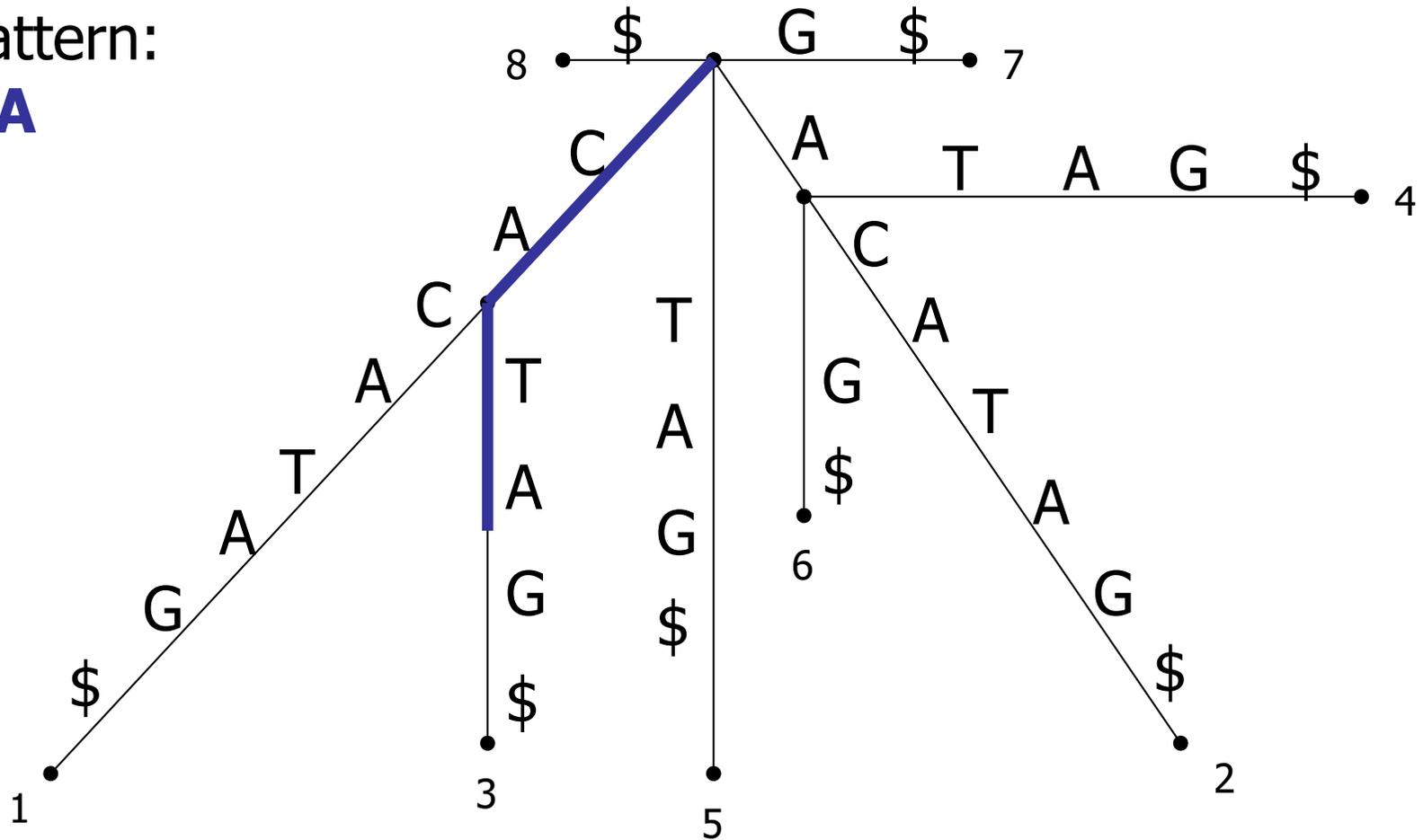
Suffixes:

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



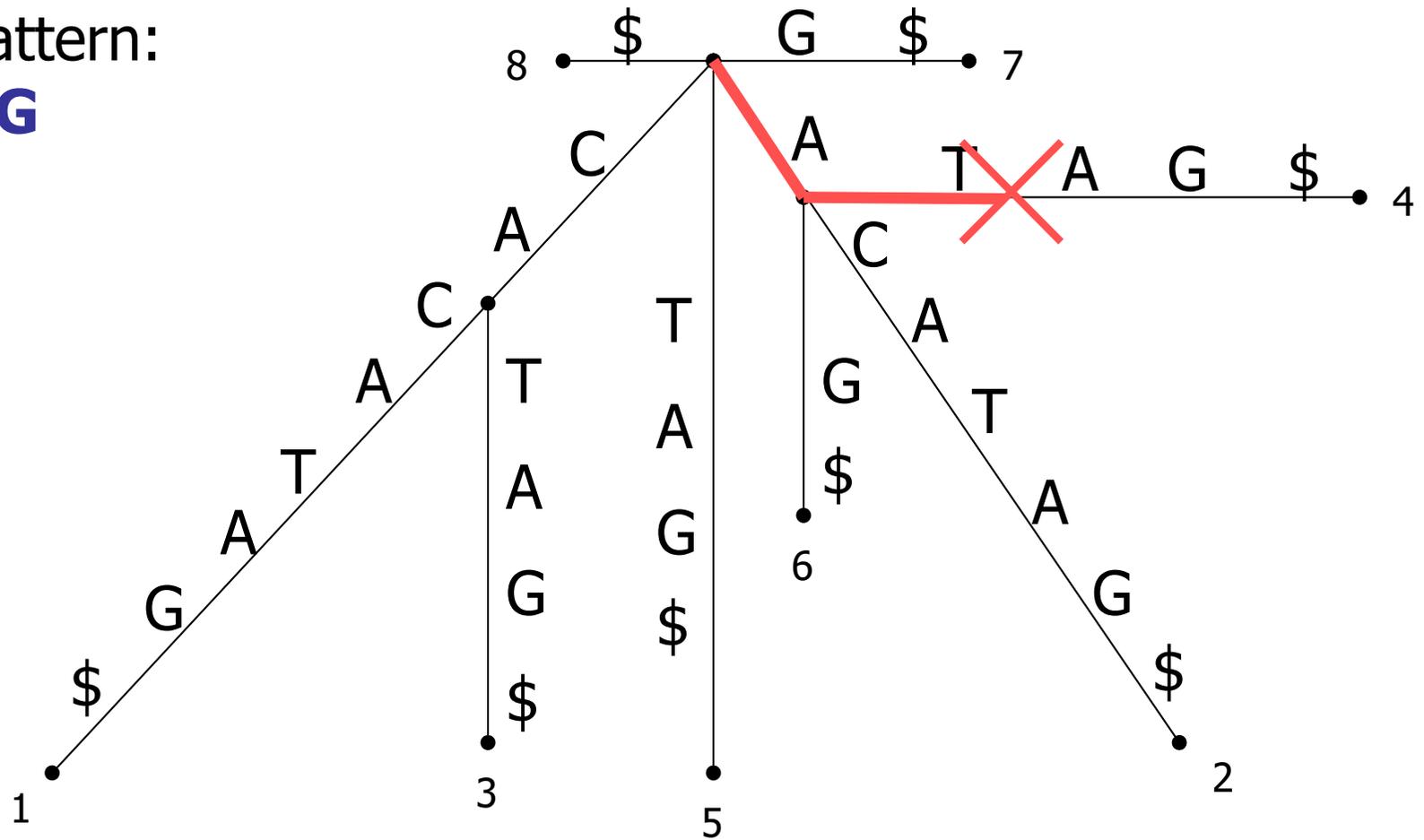
# Suchen in einem Suffix-Baum

Search Pattern:  
**CATA**



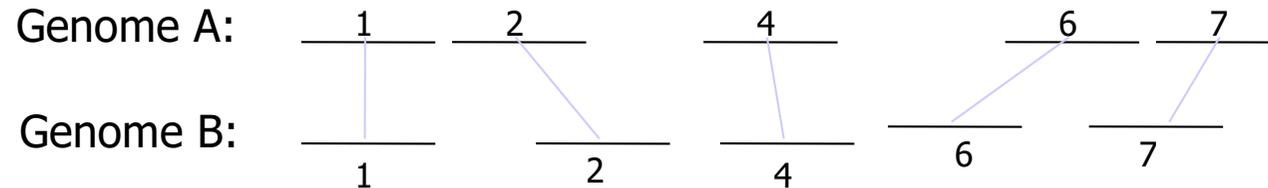
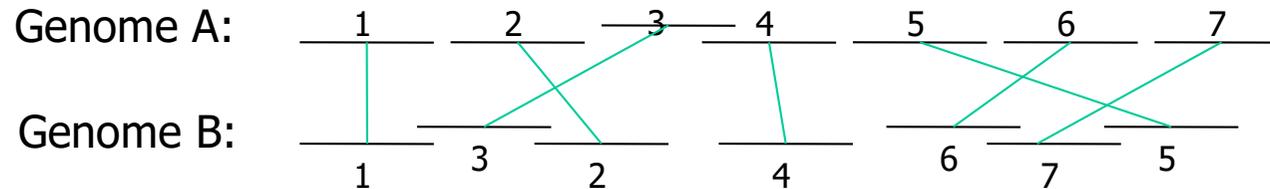
# Suchen in einem Suffix-Baum

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

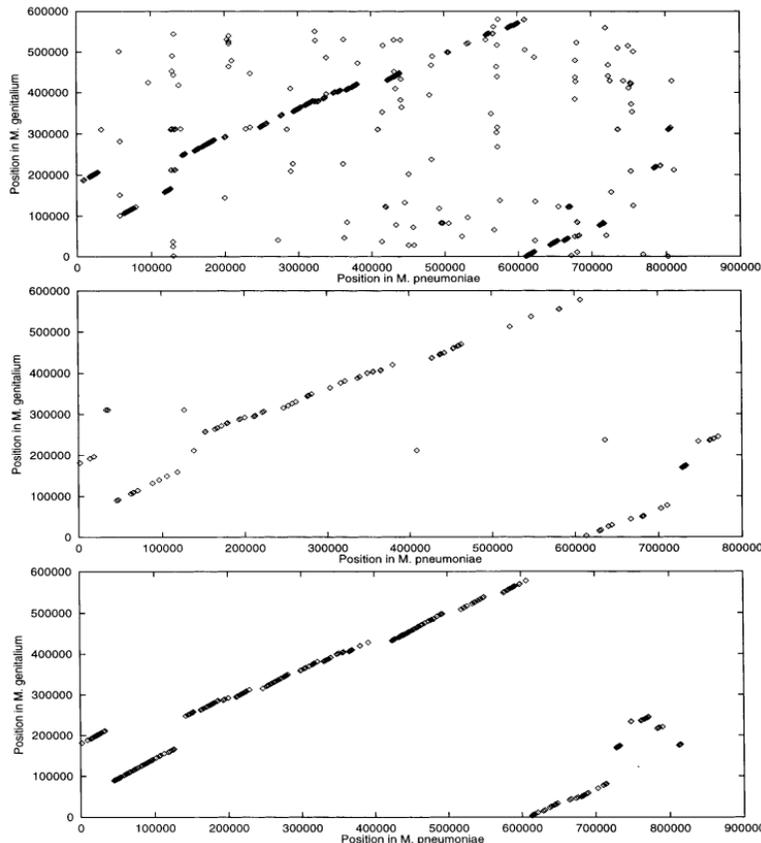
Unten: Alignment mit MUMs. 5 Translokationen.

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.



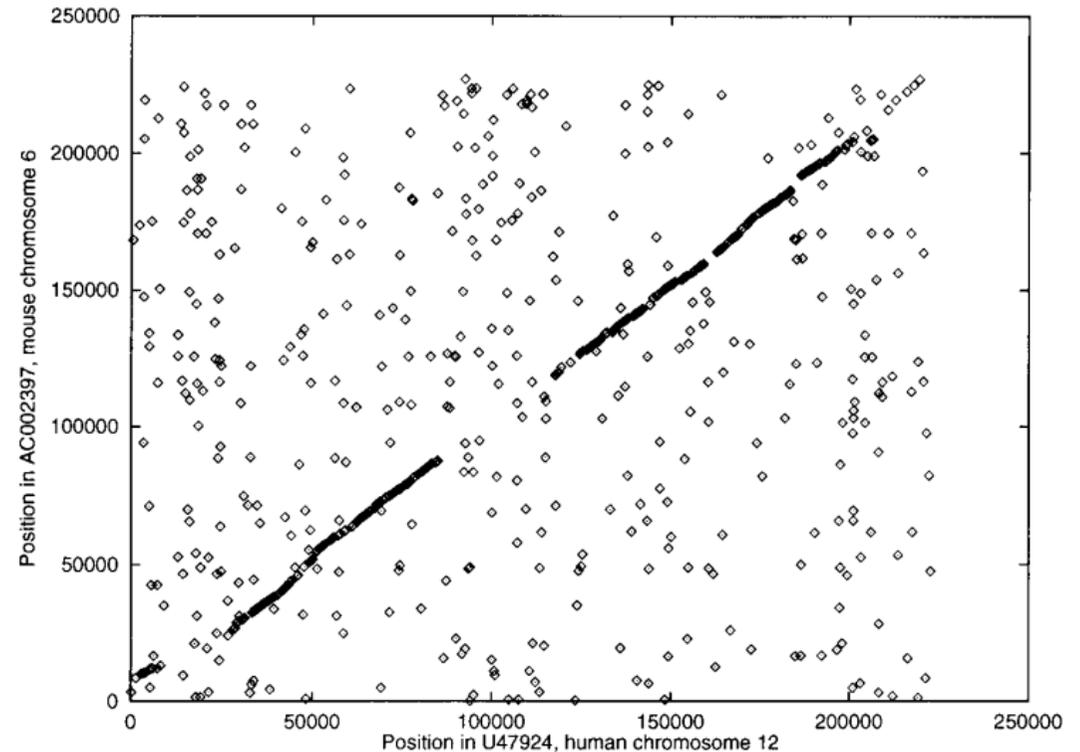
Delcher et al. Nucleic Acids Res 27, 2369 (1999)

## 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

- Gene in prokaryotischen Genomen können sehr zuverlässig mit Bioinformatik-Tools gefunden werden, den genauen Genstart aber nur in etwa 80%.
- Für eukaryotische Gene ist die Heranziehung von Zusatzinformation vorteilhaft.
- TFBS-Motive identifizieren mögliche (!) Bindestellen von Transkriptionsfaktoren.
- Die Anwendung der Suffix-Bäume war ein Durchbruch für die Alignierung ganzer Genome.