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**

In Vorlesung 4 behandeln wir einige ausgewählte Themen, die sich mit genomischen Sequenzen beschäftigen.

Zunächst einmal geht es um die Frage, wo eigentlich die Gene sind.

Dann geht es darum, in Promoterbereichen bzw. Enhancerbereichen mögliche Bindestellen von Transkriptionsfaktoren zu identifizieren.

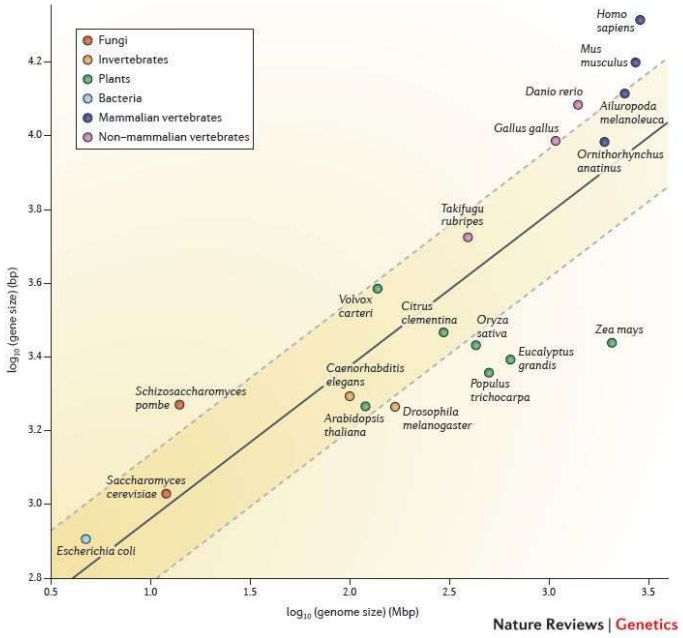
Zu Repeatsequenzen gibt es nur 2 Folien.

Zum Schluss geht es kurz um die Prozessierung von Next Generation Sequencing-Daten auf ein Referenzgenom und um die Annotation von SNP-Positionen.

Wonach suchen wir: Länge von Genen

Wie lang sind Gene im Mittel?

Generell enthalten längere Genome längere Gene.



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

Zunächst beschäftigen wir uns mit der Identifikation von Genen. Die Abbildung zeigt auf der x-Achse die Länge (in Millionen Basenpaaren) für verschiedene Organismen.

Bakterien (links unten) haben, wie erwartet, die kürzesten Genome. Der Mensch (links oben) gehört zu den Organismen mit den längsten Genomen.

Auf der y-Achse ist die mittlere Länge der Gene der einzelnen Organismen aufgetragen.

Interessanterweise haben Organismen mit sehr langen Genomen auch sehr lange Gene. Dies sind man daran, dass alle Organismen entlang einer Diagonale aufgereiht sind.

Z.B. ist das Genom von *Arabidopsis thaliana* nur 135 Mb lang, das menschliche Genom aber 3.2 Gb. Das ist ein Faktor von 23.

Arabidopsis hat aber über 27.000 Protein-kodierende Gene (https://plants.ensembl.org/Arabidopsis_thaliana/Info/Annotation/), der Mensch aber nur etwa 20.000 (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6100717/).

Worin suchen wir: offene Leserahmen / Leseraster

offenes Leseraster, ***open reading frames***, abgekürzt ORF:

Als offene Leserahmen bezeichnet man längere DNA-Abschnitte, die ausschließlich aus aminosäurecodierenden Tripletts (Basentriplett) bestehen und nicht durch Stop-Codonen unterbrochen sind.

Offene Leseraster können für Proteine codierende Regionen darstellen, müssen jedoch nicht immer codierende Funktionen haben.

Die **einfachste** Methode, DNA Sequenzen zu finden, die für Proteine kodieren, ist nach **offenen Leserahmen** zu suchen.

https://www.spektrum.de/lexikon/biologie/offenes-leseraster/47378

Auf https://application.wiley-vch.de/HOME/bioinformatik/sequ/sequ_ueb/orf_ueb/orfs.htm steht dazu:

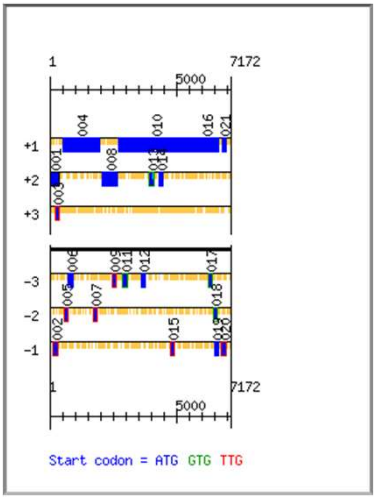
„Ein *ORF* ist ein Stück DNA, welches von einem Start- und einem Stoppcodon flankiert wird und eine, ganzzahlig durch 3 teilbare Anzahl von Basen (die Codonen englisch *codons*) umfasst. Die Menge der *ORFs* ist eine Obermenge der *Gene*; dies sind diejenigen *ORFs*, die tatsächlich von der Zelle in Proteine übersetzt werden. Da jeder der sechs möglichen Leserahmen codieren kann, überlappen sich *ORFs* häufig. Es ist die hohe Kunst der Genidentifikation, aus der Menge der *ORFs* genau die Menge der Gene herauszufiltern."

Identifikation von Genen

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.

Die Abbildung zeigt ein DNA-Fragment, bestehend aus 7172 Nucleotiden. Darin sind 20 offene Leserahmen enthalten, die länger als 150 Nucleotide sind.

Diese ORFs sind in den sechs Leserahmen (hier mit +1,+2,+3, -1, -2, -3 markiert) eingetragen.



https://application.wiley-vch.de/HOME/bioinformatik/sequ/sequ_ueb/orf_ueb/orfs.htm

Dieses Beispiel stammt von der angegebenen Webseite des Wiley-Verlags. Dort können Sie interaktiv auf die blau markierten Leserahmen draufklicken und erhalten dann die jeweilige Aminosäuresequenz angezeigt, für die dieser Leserahmen kodieren würde.

Identifikation von Genen

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 **Translationsstart** fungiert, bis zu dem **nächsten Stopcodon** in demselben offenen Leserahmen, gewöhnlich eine gute Vorhersage für die Protein-kodierende Region.

Ohne Kommentare.

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.

„Extrinische“ Methoden ziehen Information „von außen“ hinzu, also z.B. das bestätigte Wissen, dass eine ähnliche Genomregion in einem anderen Organismus für ein Gen kodiert.

Eine „intrinsische“ Methode wurde an den Eigenschaften von bekannten Genen (in diesem Organismus oder in verwandten Organismen) trainiert. Für die Vorhersage von neuen Genen werden keine zusätzlichen externen Informationen verwendet.

Hidden Markov Modell (HMM)

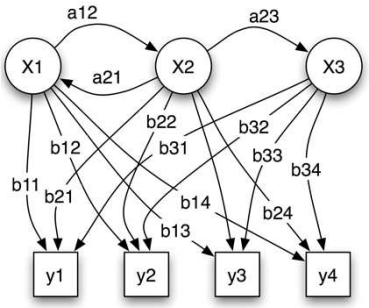
Ein Hidden Markov Modell ist ein stochastisches Modell.
Man repräsentiert es üblicherweise durch einen Graph,
der die verschiedenen Zustände verbindet.

Im Modell rechts gibt es 3 **„verborgene“ Zustände**:
X1, X2, X3.

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) Zustände der
Ausgabe.



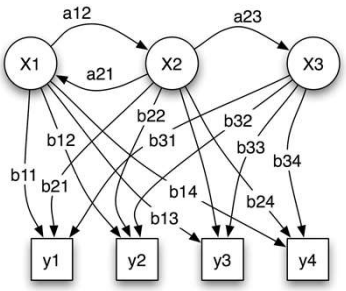
Hidden Markov Modelle (HMM) sind stochastische Modelle, die auf Markov-Ketten beruhen. Eine Markov-Kette ist ein spezieller stochastischer Prozess, bei dem zu jedem Zeitpunkt die Wahrscheinlichkeiten aller zukünftigen Zustände nur vom momentanen Zustand abhängen (= Markov-Eigenschaft).

Man stellt ein HMM üblicherweise durch einen gerichteten mathematischen Graph dar (wie in der Abbildung gezeigt). Ein solcher Graph besteht aus Knoten (als Kreise gezeichnet, diese bedeuten Zustände) und gerichteten Kanten (Pfeile, diese bedeuten Übergänge zwischen 2 Zuständen).

Ein Markov-Modell ordnet jedem Zustand eine Ausgabe zu (gezeichnet als Quadrate), die ausschließlich vom aktuellen Zustand (bzw. Zustandsübergang) abhängig ist

Hidden Markov Modell (HMM)

Im Falle einer Genvorhersage entsprechen die **„verborgenen“ Zustände** X1, X2, X3 den funktionellen *Bereichen der DNA, z.B. kodierende und nicht-kodierende Abschnitte bzw. Intron, Promoter, Exon.*



Dies ist das, was wir gerne vom Modell als Vorhersage erhalten möchten.

y1 bis y4, die sichtbaren Zustände der Ausgabe, sind *im Fall der Gen-Vorhersage die beobachteten Sequenzen.*

Bei der Genvorhersage auf DNA-Sequenzen hat man 4 sichtbare Zustände der Ausgabe für jede der beobachteten Nukleotidbasen.

Ein Hidden Markov Model ist ein Markov-Modell, bei dem nur die Sequenz der Ausgaben beobachtbar ist und die Sequenz der Zustände verborgen („hidden“) bleibt. Daher rührt der Name „hidden“ Markov-Modell.

Auf dieser Folie ist dargestellt, welches bei der Gen-Vorhersage die verborgenen Zustände sind und welches die Ausgaben sind.

Trainieren eines Hidden Markov Modells (HMM)

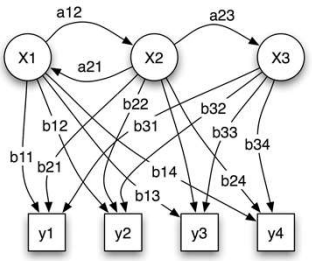
Die Topologie des Graphen gibt an, zwischen welchen Zuständen Übergänge erlaubt sind. Diese gibt man bei der Spezifikation des HMM vor.

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.

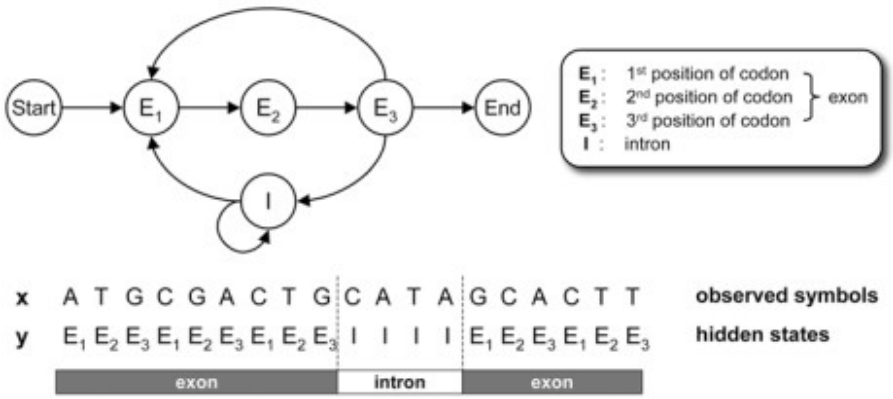
Ein HMM besteht also aus der Topologie und den trainierten Wahrscheinlichkeiten.



Die „Kunst“ beim Erstellen eines HMMs ist es, eine Topologie (d.h. die möglichen Zustände und die Ausgaben) vorzugeben, die (a) möglichst einfach ist, aber (b) die relevanten biologische Zustände möglichst gut erfasst.

Nachdem man die Topologie festgelegt hat, muss man die Übergangswahrscheinlichkeiten a und b anhand eines Goldstandard-Datensatzes von bekannten Genen „trainieren“. Dazu gibt es etablierte Algorithmen wie den forward-Algorithmus und den backward-Algorithmus und den Viterbi-Algorithmus.

HMM für eukaryotische Gene



x	A	T	G	C	G	A	C	T	G	C	A	T	A	G	C	A	C	T	T	observed symbols	
y	E_1	E_2	E_3	E_1	E_2	E_3	E_1	E_2	E_3	I	I	I	I	E_1	E_2	E_3	E_1	E_2	E_3	hidden states	

exon | intron | exon

Ein einfaches HMM zur Modellierung eukaryotischer Gene.

Yoon, Curr. Genomics 10, 402 (2009)

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
A.fulgidus	49	1165	1162	99.7%	841	72.2%	1308	−2	−67	−59
B.anthracis	35	3132	3119	99.6%	2717	86.7%	2345	+6	+726	−77
B.subtilis	44	1576	1559	98.9%	1379	87.5%	2886	+11	+413	−539
C.tepidum	57	1292	1284	99.4%	867	67.1%	778	+2	−33	−190
C.perfringens	29	1504	1501	99.8%	1360	90.4%	1177	−1	+244	−28
E.coli	51	3603	3525	97.8%	3014	83.7%	942	+16	+693	−632
G.sulfurreducens	61	2351	2320	98.7%	1883	80.1%	1107	+15	+541	−380
H.pylori	39	915	908	99.2%	785	85.8%	774	+1	+46	−94
P.fluorescens	63	4535	4484	98.9%	3412	75.2%	1896	+14	+731	−704
R.solanacearum	67	2512	2468	98.2%	1922	76.5%	1091	+72	+646	−326
S.epidermidis	32	1650	1646	99.8%	1496	90.7%	767	+3	+338	−66
T.pallidum	53	575	569	99.0%	397	69.0%	568	+3	+55	−296
U.parvum	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)

Dies sind Ergebnisse mit den bekannten Glimmer2 und Glimmer3-Programmen. Gen-Vorhersagen für prokaryotische Genome funktionieren im Allgemeinen sehr gut.

Generkennung mit Hidden Markov Modellen

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

Die Ausgabe sind die bekannten exprimierten Sequenzen.

Für eine Eingabesequenz soll jedem Basenpaar der günstigste verborgene Zustand (E/I) zugeordnet werden.

Bei **Markov-Modellen** 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)

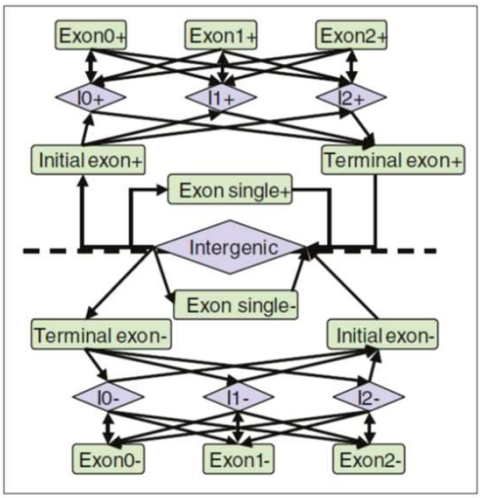
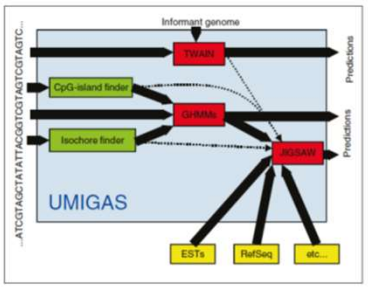


Figure 4
State-transition diagram of the GHMM for GlimmerHMM. The dashed line in the middle separates the positive strand and negative strand portions of the model. Each state in the GHMM is implemented as a separate submodel, such as a weight array matrix or an IMM (interpolated Markov models).

Für eukaryotische Genome verwendet man oft HMM-Topologien, wie die in der Abbildung gezeigte Topologie.

Der Algorithmus „marschiert“ dann quasi durch das Genom, von Base zu Base, und bestimmt jedes Mal, ob diese Base noch in einer intergenischen Region liegt, oder ob hier ein Exon beginnen sollte.

Generkennung von menschlichen Genen mit JIGSAW



Durch Hinzunahme zusätzlicher Information konnten etwa ¾ der menschlichen Gene präzise vorhergesagt werden.
Nur 3% der Gene wurden überhaupt nicht gefunden (rot umkreist).

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)

Gezeigt wird hier die Genauigkeit der Gen-Vorhersage für menschliche Gene mit dem Tool JIGSAW (im Jahr 2006).

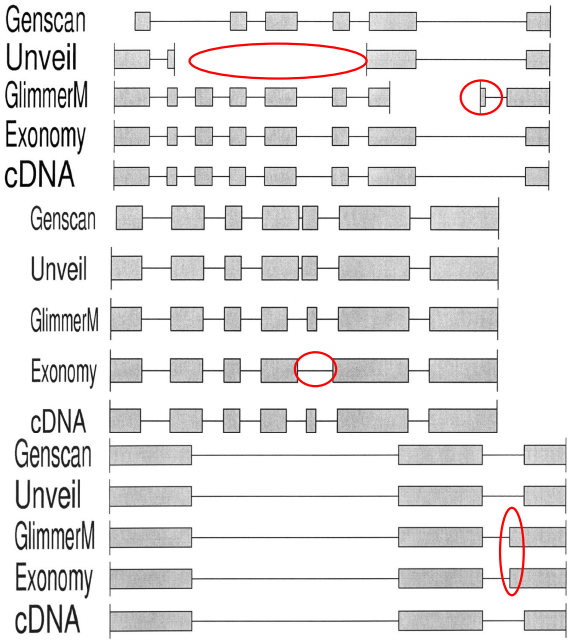
Als Zusatzinformation wird ein Tool verwendet, mit dem man CpG-Inseln identifizieren kann.

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).

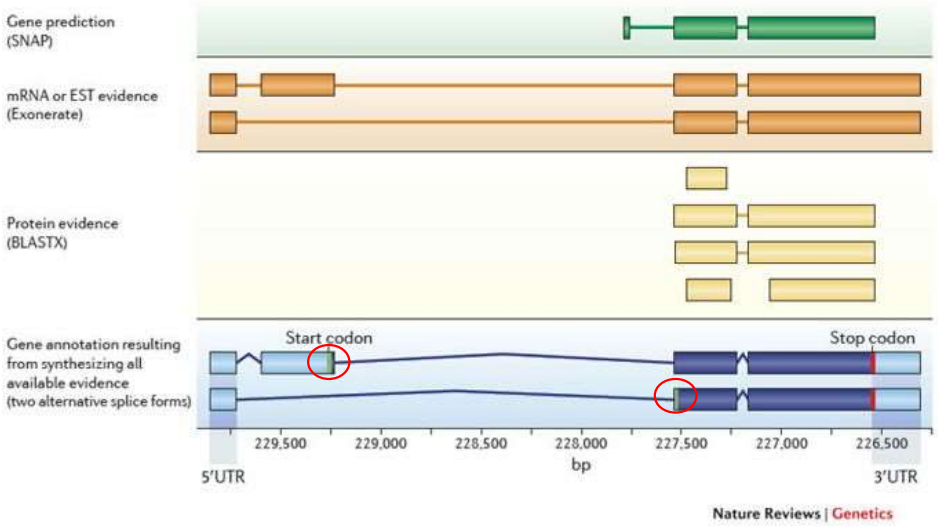


Majoros et al. Nucl. Acids. Res. 31, 3601 (2003)

Dieses Beispiel zeigt, dass verschiedene Tools durchaus leicht unterschiedliche Ergebnisse liefern können.

Der Goldstandard ist hier die experimentell bestimmte cDNA-Sequenz, die aus den transkribierten mRNAs zurückübersetzt wurde.

Vorhersage von Genen



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

Tools zur Vorhersage von Gene ermitteln nur die kodierenden Abschnitte, keine Proteinisoformen (Transkripte)!

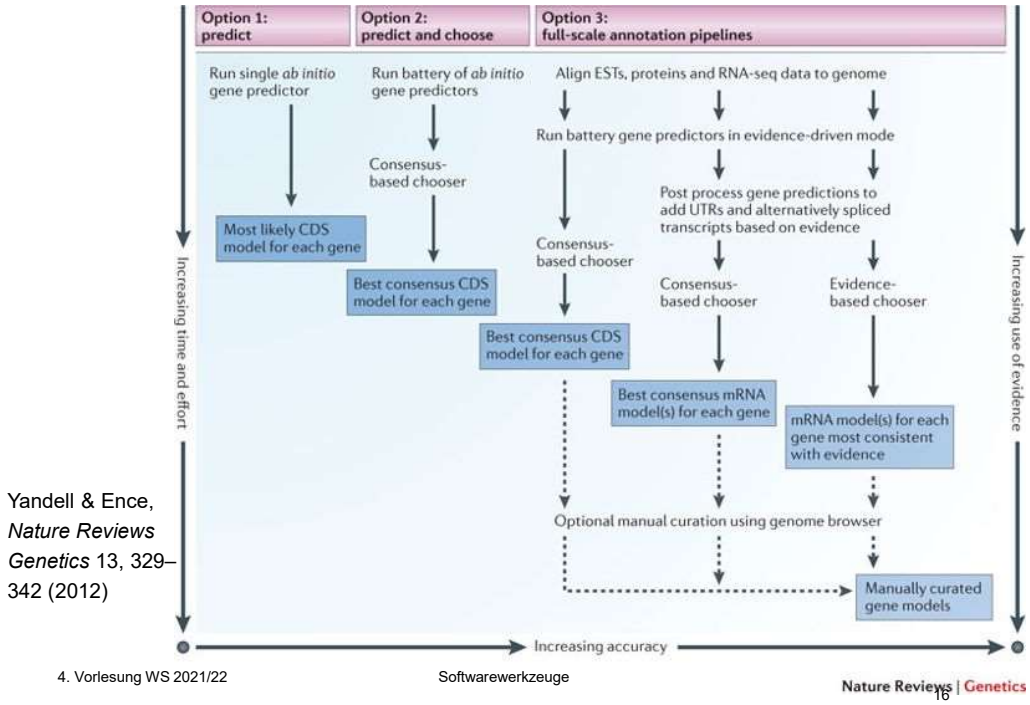
Die oberste Zeile enthält das Ergebnis eines Genvorhersage-Programms.

Die tatsächliche mRNA-Sequenz ist vorne und hinten noch um die 5‘-UTR und 3‘-UTR Regionen verlängert.

Durch Blast-Suche könnte man feststellen, dass die 3 grünen Boxen als Proteine in anderen Organismen vorliegen.

In der unteren Zeile sind die 2 tatsächlich vorkommenden Spleißvarianten gezeigt. Die grün markierten Bereiche markieren 2 mögliche Startcodonen für die Translation.

Annotation von Genomen



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

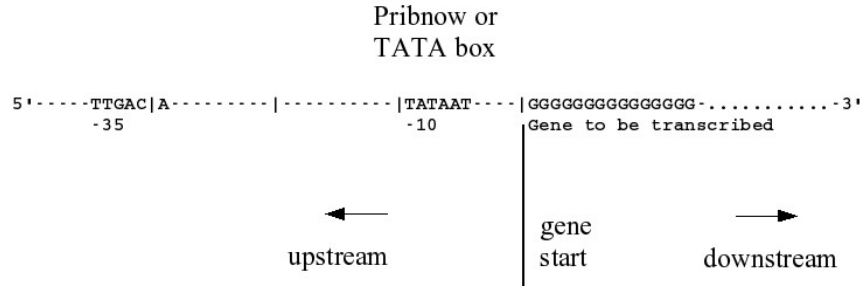
In dem Artikel von Yandell und Ence sind etliche Software-Tools aufgelistet, mit denen man die einzelnen Schritte durchführen kann.

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



Promoterabschnitte in Bakterien wie *E.coli* sind stets nach einem festen Schema aufgebaut. Bei der Position -10 vor dem Transkriptionsstart liegt die TATA-Sequenz, wo das TATA-binding protein bindet. Bei Position -35 liegt ein weiteres stark konserviertes TTGACA-Motif.

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

Die Suche nach möglichen Transkriptionsfaktor-Bindestellen ist zwar technisch nicht schwierig (dies werden wir gleich sehen), allerdings etwas schwammig greifbar.

Suche nach gemeinsamen Sequenzmotiven

Wird seit der Verfügbarkeit von Microarray Gen-Expressionsdaten eingesetzt.

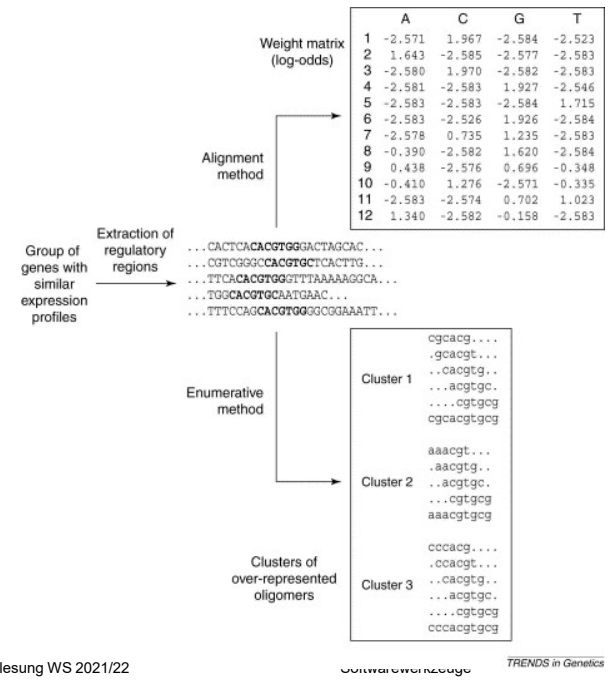
Durch Clustern erhält man Gruppen von Genen mit ähnlichen Expressionsprofilen (z.B. solche, die zur selben Zeit im Zellzyklus aktiviert sind)

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).

Man sucht meist gemeinsame Motive in den Promotersequenzen von Genen, die sich in Experimenten als „ko-exprimiert“ herausgestellt haben. Sie sind also in den bestimmten Bedingungen entweder gemeinsam angeschaltet oder gemeinsam ausgeschaltet. Dies deutet auf einen gemeinsamen Regulationsmechanismus durch einen (oder mehrere) Transkriptionsfaktor hin.

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)

Für die Identifizierung von Sequenzmotiven kann man entweder eine Gewichtsmatrix (oben rechts) oder eine abzählende Methode (unten) verwenden.

Bei der abzählenden (enumerativen) Methode sucht man nach Clustern von exakt überlappenden k-meren. Mit dieser Methode werden wir uns hier nicht weiter beschäftigen.

Auf den nächsten Folien wird die obere Methode (Gewichtsmatrix) detailliert erklärt.

Positions-spezifische Gewichtsmatrix (PSSM)

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?

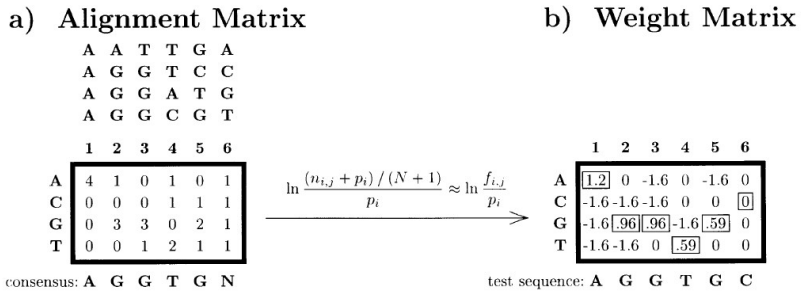


Fig. 1. Examples of the simple matrix model for summarizing a DNA alignment. (**a**) An alignment matrix describing the alignment of the four 6-mers on top. The matrix contains the number of times, $n_{i,j}$, that letter i is observed at position j of this alignment. Below the matrix is the consensus 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

Im Englischen benützt man sowohl die Ausdrücke position-specific scoring matrix (PSSM) als auch position-specific weight matrix (PSWM). Sie bedeuten im Wesentlichen dasselbe. Man verwendet eine **Matrix** mit den Dimensionen (Anzahl an Positionen) x (Anzahl möglicher Buchstaben). Im den hier gezeigten Beispiel geht es um DNA-Sequenzen. Es gibt also die vier Buchstaben A, C, G und T. Es werden sechs aufeinanderfolgende Positionen in der Sequenz betrachtet.

Wir zählen ab, wie häufig jeder Buchstabe an jeder Position in einer Menge von alignierten Sequenzen vorkommt (hier: 4 Sequenzen) und tragen dies in die „Alignment Matrix“ ein. Daraus könnte man nun die darunter angegebene „Consensus Sequenz“ bilden, die an jeder Position die häufigste Base enthält. Je nachdem wie hoch die Variabilität ist, bedeutet die Bildung einer Consensus-Sequenz einen hohen Informationsverlust.

Die linke Alignment-Matrix ist wesentlich aussagekräftiger als die Consensus-Sequenz. Man verwendet meist die logarithmierte Version davon, nämliche die rechte „Gewichtsmatrix“.

Die Formel gibt an, wie die Einträge der rechten Matrix aus denen der linken Matrix berechnet werden. Wie bei den BLOSUM-Austauschmatrizen gilt einfach wieder: log (observed / expected). Im Nenner steht die erwartete a priori Häufigkeit jeder Base. Hier wurde der Einfachheit halber eine Gleichverteilung angenommen. Deshalb hat jede Base die Häufigkeit p = 0.25. N ist die Anzahl an Sequenzen (hier: vier), n_{ij} ist die Häufigkeit der Base i an Position j in den betrachteten Sequenzen. Z.B. kommt A an der ersten Position 4 mal vor. In der rechten Matrix enthält man daraus ln ((4+0.25) / 5) / 0.25) = ln (3.4) = 1.223

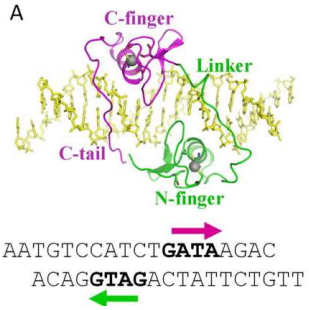
Man addiert zu der Häufigkeit an jeder Position p_i hinzu, um Probleme zu vermeiden, falls n_{ij} gleich Null ist. Der Logarithmus von Null ist nämlich nicht definiert. Außerdem kann man argumentieren, dass man ja nur eine bestimmte Anzahl an Sequenzen vorliegen hat. Falls nun noch mehrere hinzukämen, dann wäre die Häufigkeit der Base i im Mittel gleich p_i. Da man die mittlere Häufigkeit im Zähler addiert hat, teilt man dann durch N+1, nicht durch N.

Positions-spezifische Gewichtsmatrix

Frequenzmatrix: die Felder enthalten die Anzahl an Sequenzen, die Base x in Spalte y enthalten.

Beispiel aus JASPAR-Datenbank für *homo sapiens*:

GATA1 (11 Positionen) ist ein Zink-Finger



A [	22209	17328	3953	1314	49692	67550	2206	2567	7397	26545	22656	]
C [	12209	14489	62419	595	710	1292	1238	65937	3025	11186	15261	]
G [	13955	11088	2712	652	856	988	618	1471	765	14358	9325	]
T [	23455	28923	2744	69267	20570	1998	67766	1853	60641	19739	24586	]
Position	1	2	3	4	5	6	7	8	9	10	11	

http://jaspar.genereg.net/
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3978094/

Um die numerische Effizienz bei der Durchsuchung von Genomen zu verbessern, verwendet man ganzzahlige Gewichtsmatrizen. Hier ist die Matrix für den bekannten Transkriptionsfaktor GATA1 gezeigt.

Die Abbildung zeigt eine Kristallstruktur, wie die DNA-Bindedomäne des verwandten GATA3-Faktors an eine 20-mer palindromische GATA-Bindestelle bindet.

Der GATA-Faktor bindet in der großen Furche der DNA (major groove). Eine volle DNA-Umdrehung hat eine Periodizität von 10 bp. Dies erklärt, warum Bindemotive typischerweise zwischen 5 und 10 nt lang sind.

Sequenzlogos repräsentieren Bindemotive

Ein **Logo** stellt jede Spalte des Alignments als einen Stapel Buchstaben dar.

Die Höhe jedes Buchstabens ist proportional zur **beobachteten Frequenz** der entsprechenden Aminosäure oder Nukleotids.

Die Gesamthöhe jeden Stapels ist proportional zur **Sequenzkonservierung** and dieser Position.

Sequenzkonservierung wird als Unterschied zwischen der maximal möglichen Entropie oder der Entropie der beobachteten Verteilung der Symbole definiert:

$$R_{seq} = S_{max} - S_{obs} = \log_2 N - \left(- \sum_{n=1}^{N} p_n \log_2 p_n \right)$$

p_n : beobachtete Häufigkeit von Symbol *n* an einer bestimmten Sequenzposition

N : Anzahl an verschiedenen Symbolen (DNA/RNA: 4, Protein: 20).

Crooks et al. Genome Research 14:1188–1190 (2004)

Die Bindepräferenz eines Transkriptionsfaktors lässt sich sehr gut als sogenanntes Sequenzlogo visualisieren.

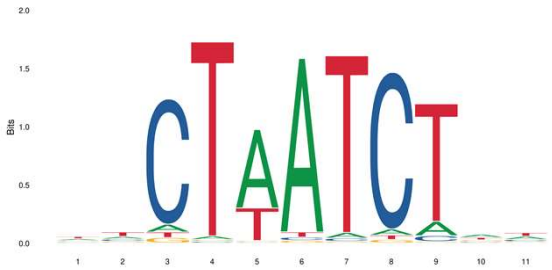
Wenn an einer Position alle 4 Nukleotide gleich häufig auftreten würden (es also gar keine Präferenz gibt), dann hat diese Position die „Entropie“ log2 N.

Zumeist gibt es aber eine deutliche Präferenz für eine oder zwei Nukleotide. Dann ist die Entropie dieser Position geringer. In der Formel wird dies durch die Shannon-Entropie (Summenformel) ausgedrückt.

Wenn die tatsächliche Entropie sehr klein ist, zieht man einen sehr kleinen Term von log2N ab, und erhält einen großen Buchstaben.

Positions-spezifische Gewichtsmatrix

Sequenzlogo für GATA1



A [	22209	17328	3953	1314	**49692**	**67550**	2206	2567	7397	26545	22656	]
C [	12209	14489	**62419**	595	710	1292	1238	**65937**	3025	11186	15261	]
G [	13955	11088	2712	652	856	988	618	1471	765	14358	9325	]
T [	23455	28923	2744	**69267**	**20570**	1998	**67766**	1853	**60641**	19739	24586	]

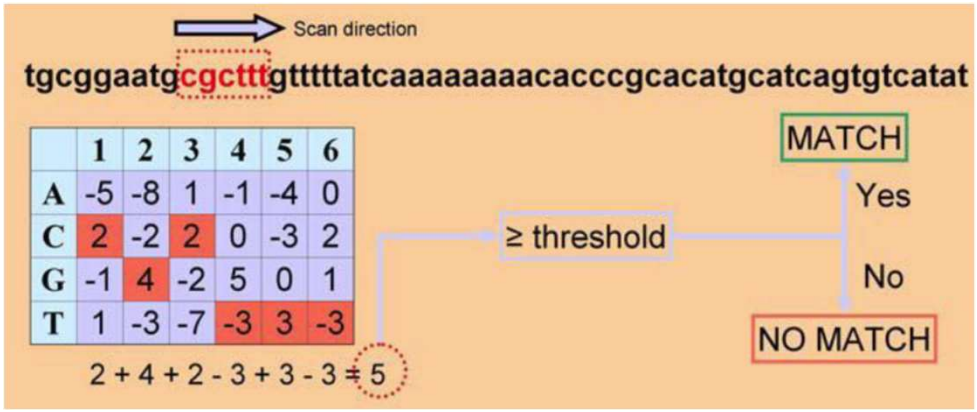
http://jaspar.genereg.net/

Dies ist das Sequenzlogo für den GATA1-Transkriptionsfaktor, das mit der vorgestellten Formel berechnet wurde.

Die Positionen 3, 4, 6-9 haben die höchste Präferenz für ein bestimmtes Nukleotid.

Dies sieht man auch in der zugehörigen Gewichtsmatrix. Dort sind die Maxima jeder Spalte fett markiert.

PWM-Motive finden: z.B. PWMScan



Der p-Wert einer PWM-Bewertung *x* wird als die Wahrscheinlichkeit definiert, dass eine zufällige *k*-mer Sequenz der Länge der PWM eine Bewertung ≥ *x* hat (für die Basenzusammensetzung des Genoms).

Ambrosini et al. Bioinformatics 34, 2483–2484 (2018)
https://www.cs.cmu.edu/~02710/Lectures/Motifs2015.pdf

Um nun ein Bindemotiv zu finden, rutscht der Algorithmus gewissermaßen über die Sequenz und bewertet an jeder Position, wie gut die n nächsten Nukleotide zu der Gewichtsmatrix passen.

In diesem Beispiel enthält die oben gezeigte Sequenz die roten Nukleotide CGCTTT. Der Algorithmus zählt die entsprechenden Häufigkeiten der Gewichtsmatrix zusammen und entscheidet, ob diese Bewertung oberhalb eines Schrankenwerts liegt oder nicht. Falls dies der Fall ist, dann wird dies vom Programm als mögliche Bindestelle für diesen Transkriptionsfaktor ausgegeben.

Repeats in genomischen Sequenzen

Viele Genome enthalten hoch repititive DNA-Abschnitte.
Man teilt diese Sequenzen in fünf Kategorien auf:

Simple Repeats - Duplikation mehrerer DNA Basen (typisch 1-5bp) wie A, CA, CGG etc.

Tandem Repeats - oft in den Centromeren und Telomeren von Chromosomen. Dies sind Duplikate von komplexen 100-200 bp langen Sequenzen.

Segmental Duplications - Große Blöcke von 10-300 kB Länge, die an eine andere Stelle des Genoms kopiert wurden.

Interspersed Repeats: Processed Pseudogenes, Retrotranscripts, SINES - Non-functional copies of *DNA Transposons*, *Retrovirus Retrotransposons*, *Non-Retrovirus Retrotransposons (LINES)*

Ungefähr 50% des menschlichen Genoms wird derzeit als **repetitiv** angesehen.
http://www.repeatmasker.org/faq.html

Das menschliche Genom (sowie viele andere Genome) enthält viele repetitive Sequenzen, die etliche Male in (nahezu) unveränderter Form im Genom auftauchen.

Identifizierung von Repeats: RepeatMasker

Programm **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).

Ausgabe:
- 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.repeatmasker.org/

Das Programm RepeatMasker verwendet eine lange Tabelle mit den Sequenzen bekannter Repeats.

Prozessierung von NGS-Daten

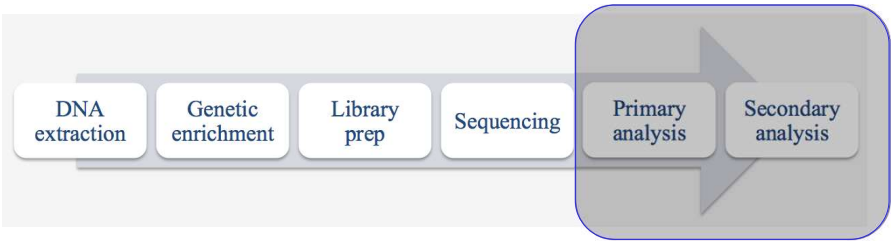
- Ganzgenomsequenzierung = Whole Genome Sequencing (WGS)
- Anwendung von WGS für mikrobielle Isolate
- Qualitätskontrolle der Sequenzierungs-reads
- Alignment
- SNP calling
- Genomvisualisierung
- Genomassemblierung

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

Danksagung für Folien: Mohamed Hamed

Wir behandeln nun kurz auch die Prozessierung von NGS-Daten und speziell die Identifizierung von SNPs.

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 SNP Varianten, phänotypische Charaktisierung (z.B. Virulenzfaktoren)

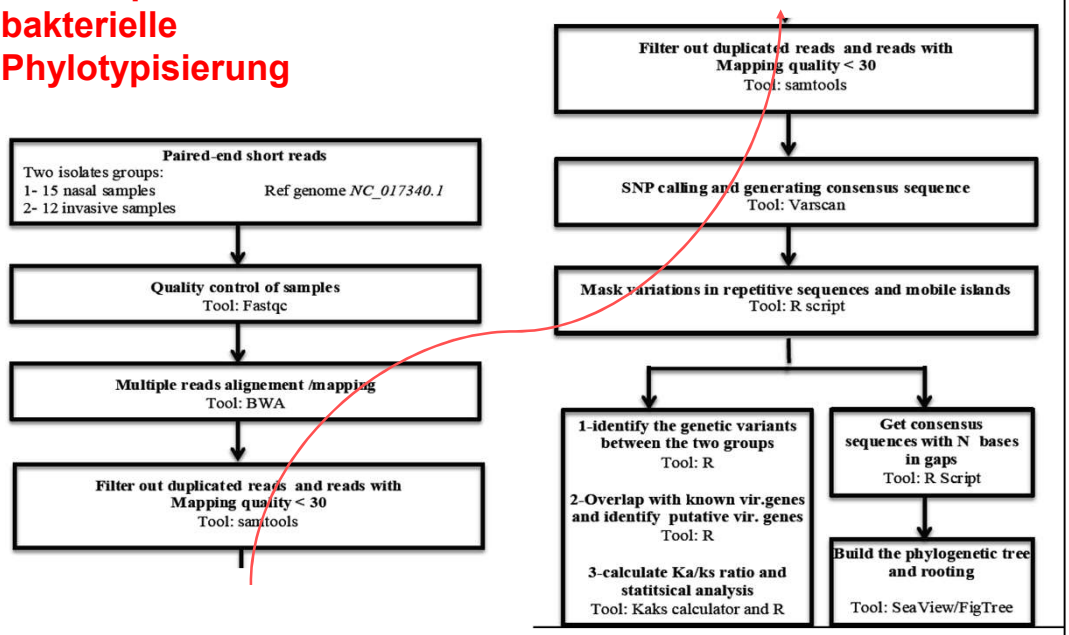
Wir konzentrieren uns auf die Schritte 5 und 6

@Prof.keller NGS lect, 2014

Ein NGS-Sequenzierungsprojekt besteht aus vielen experimentellen Schritten (oben links) und der Auswertung am Computer.

Wir werden uns in dieser Vorlesung nur mit den Schritten 5 und 6 beschäftigen.

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.

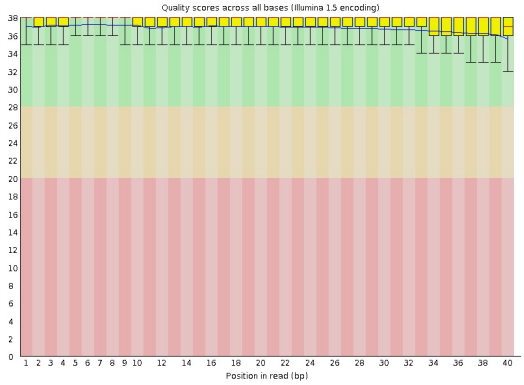
Base Qualitäts-scores nehmen üblicherweise am Ende der reads ab.

Deshalb werden die reads vor dem Alignment-Schritts „getrimmt“, d.h. gekürzt, z.B. mit dem Programm cutadapt.

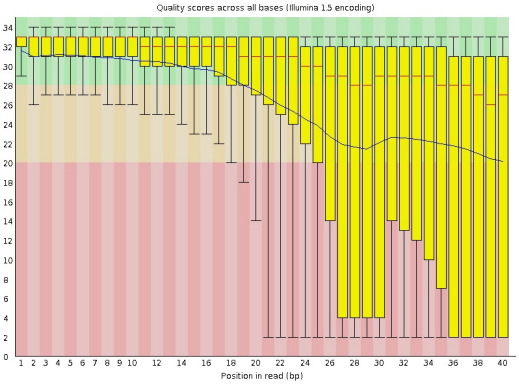
Quality (Phred) score

Das Programm FastQC
http://www.bioinformatics.babraham.ac.uk/projects/fastqc/
Ist ein Quasi-Standard.

„Gute“ Illumina-Daten



„Schlechte“ Illumina-Daten



http://www.bioinformatics.babraham.ac.uk/projects/fastqc/bad_sequence_fastqc.html

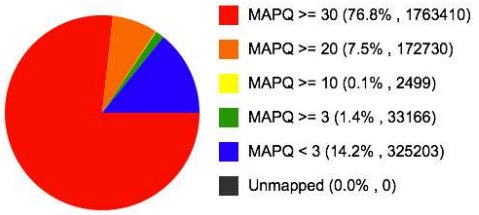
Die Sequenzierqualität von Reads nimmt üblicherweise zum Ende des Reads hin ab.

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

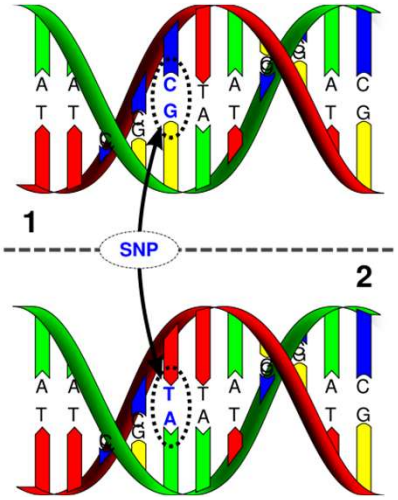
Timo Lassmann*, Yoshihide Hayashizaki and Carsten O. Daub



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

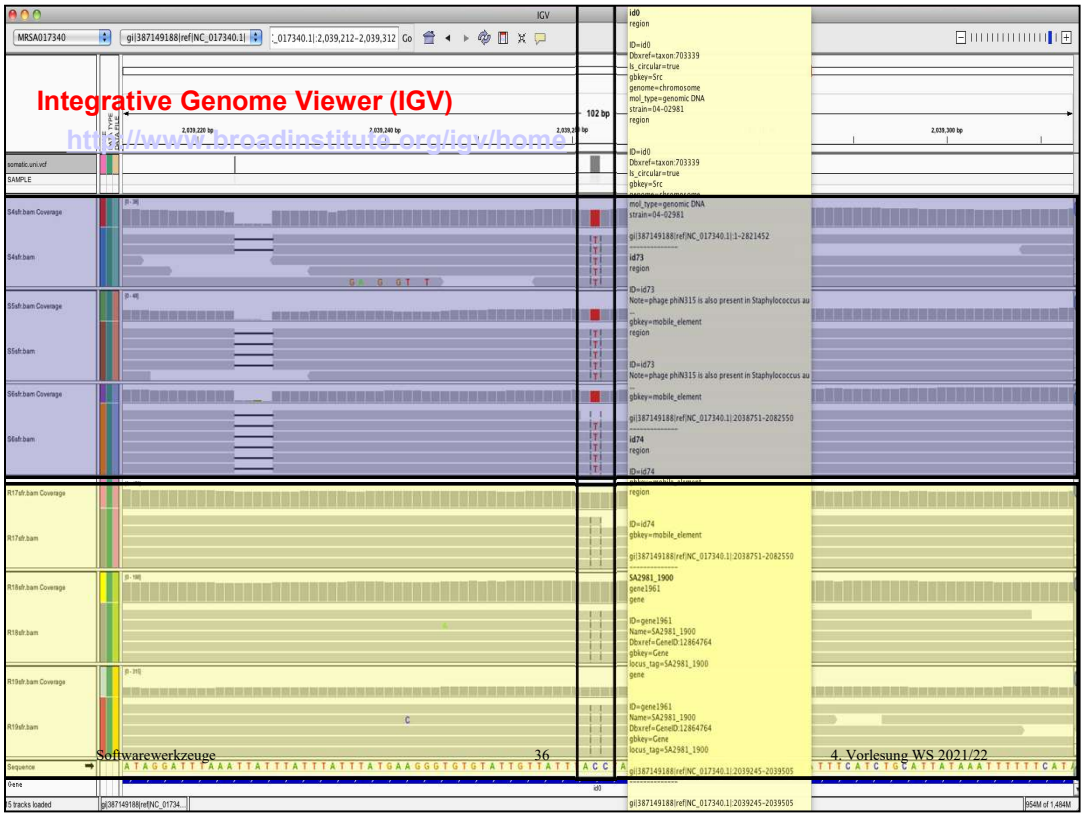
Mögliche Gründe für Abweichungen in Alignments

- 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)

SNPs können Unterschiede zwischen gesunden und kranken Zellen bewirken.

SNPs können aber auch einfache die Unterschiede zwischen verschiedenen Individuen hervorrufen.



Auf dieser Folie sind oben lila unterlegt die Genomsequenzen von 3 sensitiven Stämmen gezeigt (S4 bis S6) und darunter gelb unterlegt die Sequenzen von drei resistenten Stämmen (R17 bis R19).

Jede einzelne Bahn enthält einzelne Sequenzreads eines Stamms. Ganz unten ist die Referenzsequenz gezeigt.

In der schwarz umrahmten Position enthalten alle Reads der sensitiven Stämme ein SNP (C->T), aber keiner der resistenten Stämme.

Solche Positionen sind Kandidaten für Mutationen, die Gene virulent werden lassen können. Alle solchen Fälle werden abgespeichert und einzeln analysiert.

Phylogenetischer Baum aus core-genome SNPs

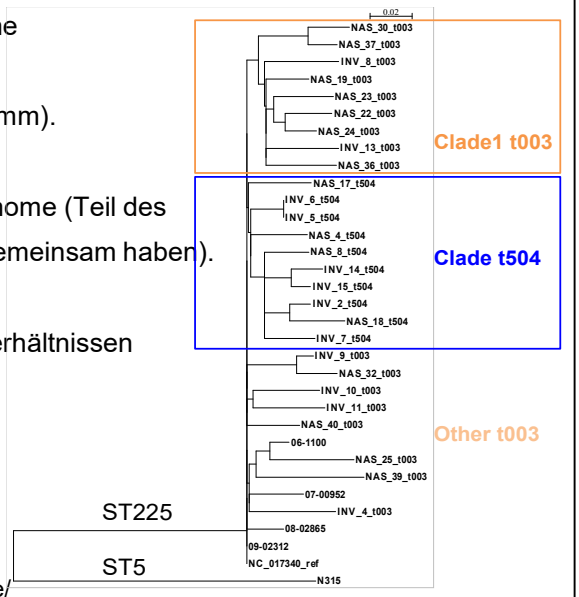
Input: WGS-Sequenzen für verschiedene *Staphylococcus aureus* Stämme (nas: nasaler Stamm; inv: invasiver Stamm).

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

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



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

Hier gezeigt ist ein phylogenetischer Stammbaum für die Ganzgenomsequenzen einiger *Staphylococcus aureus* Stämme.

Man verwendet nur die Gene, die alle Stämme gemeinsam besitzen („core genome“), und in diesen Genen nur die Positionen, an denen in einem Stamm SNPs gegenüber dem Referenzgenom auftreten.

Der Grund hierfür ist, dass in den anderen Positionen ja alle Stämme dasselbe Nukleotid enthalten.

Zwischen jeweils zwei Stämmen bestimmt man dann ein Distanzmaß proportional zur Anzahl an SNP-Positionen, in denen sie sich unterscheiden.