Algorithmische Bioinformatik

Gene Finding und Markov-Modelle

Ulf Leser

Wissensmanagement in der Bioinformatik

Ziel dieser Vorlesung

- Einblick in statistische Verfahren
- Statistisches Patternmatching verstehen (eine Variante)
- Biologischer Hintergrund: Struktur von Genen

Inhalt der Vorlesung

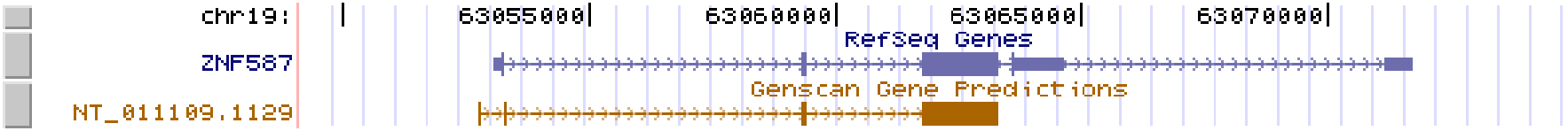
- Gene Finding
- Struktur von Genen
- CpG Inseln und Markov Modelle

Gene Finding

- Wichtigster Bestandteil eines Genoms sind seine Gene
  - Unsere Definition: Teil eines Chromosoms, der in ein Protein übersetzt wird
- Wie kann man Gene finden?
  - Experimentell: mRNA sequenzieren – im Genom suchen
    - Findet Gene nur teilweise
    - Findet nur schwer Splicevarianten
    - Findet nur Gene, die in den untersuchten Proben exprimiert werden
  - Homologie: Ähnliche Sequenzen in evolutionär entfernten Spezies
    - Generiert nur Hypothese, keinen Beweis (z.B. Pseudo Genes)
    - Findet auch nicht-kodierende, aber konservierte Bereiche
    - Findet gerade speziesspezifische Gene nicht

Gene Prediction

- Kann man Gene vorhersagen?
  - Ist an der Sequenz eines Gens irgendwas besonderes?
  - Kann man die Unterschiede aus bekannten Genen lernen?
  - Kann man das Gelernte zur Vorhersage neuer Gene benutzen?
- Gene Prediction
  - Aktuelle Verfahren benutzen alle verfügbaren Informationen
    - Basenzusammensetzung, Bindungsstellen, Splicesignale, Phylogenie, ...
    - GRAIL, GeneWise, Gene-ID, GeneScan, ...
  - Vorhergesagte Gene werden oft als „putative“ in die aktuellen Genomannotationen übernommen



Inhalt der Vorlesung

- Gene Finding
- Struktur von Genen
- CpG Inseln und Markov Modelle

Prokaryoten versus Eukaryoten

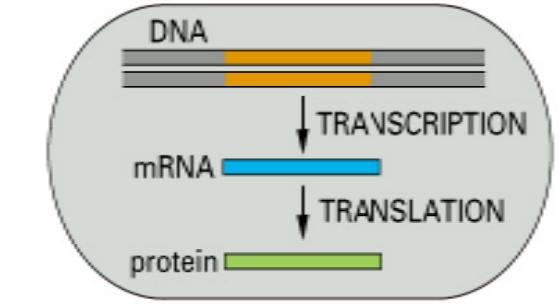


Figure 6–21 part 2 of 2. Molecular Biology of the Cell, 4th Edition.

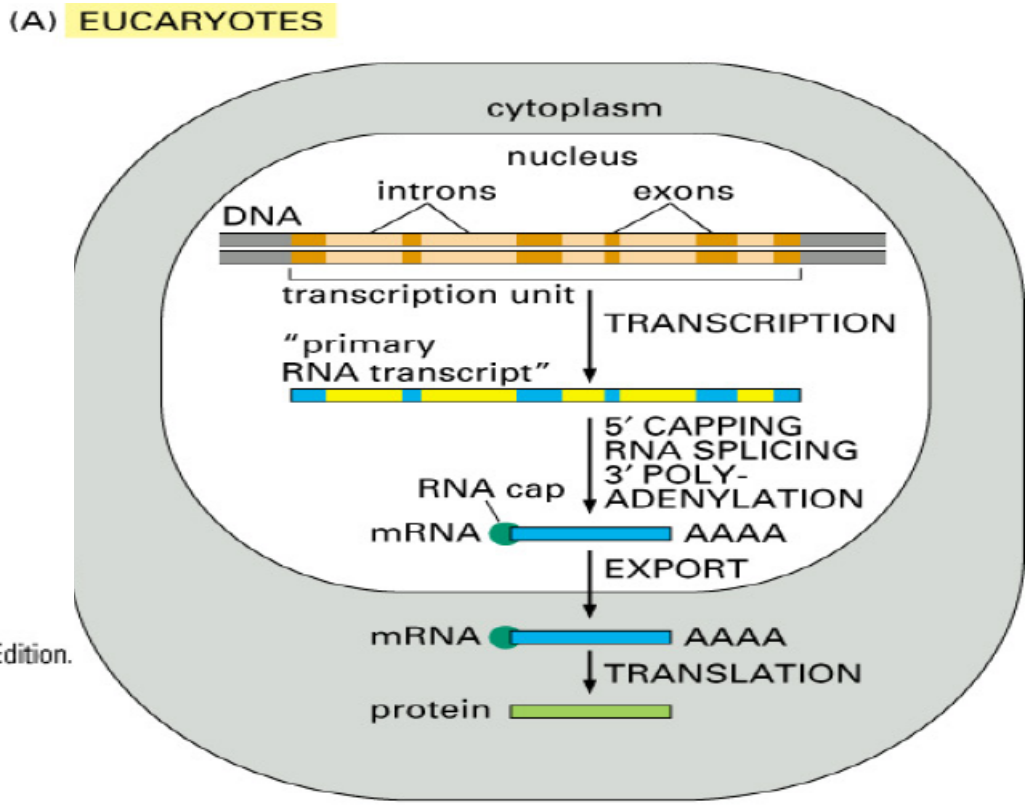
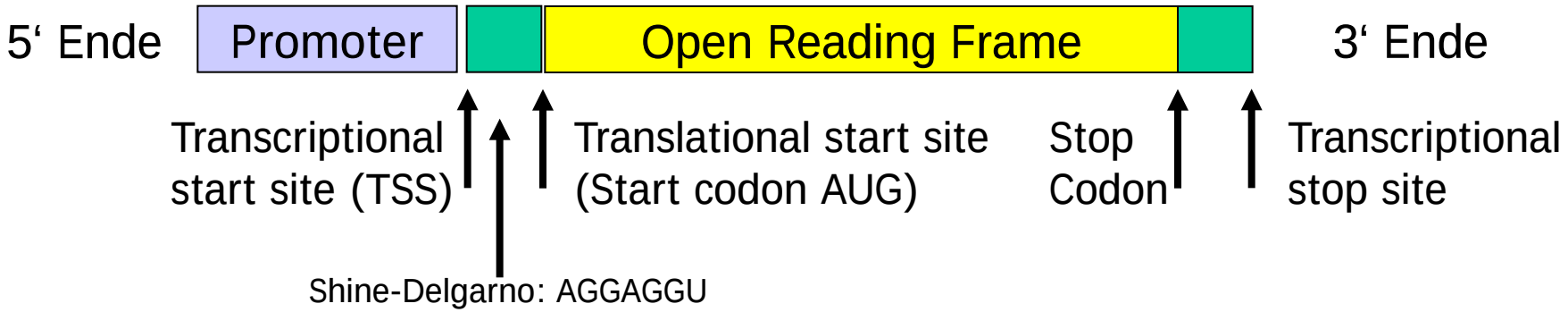


Figure 6–21 part 1 of 2. Molecular Biology of the Cell, 4th Edition.

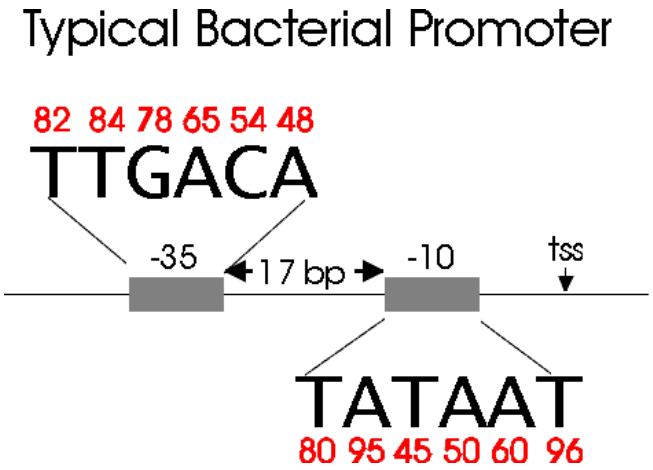
Quelle: William Stafford Noble

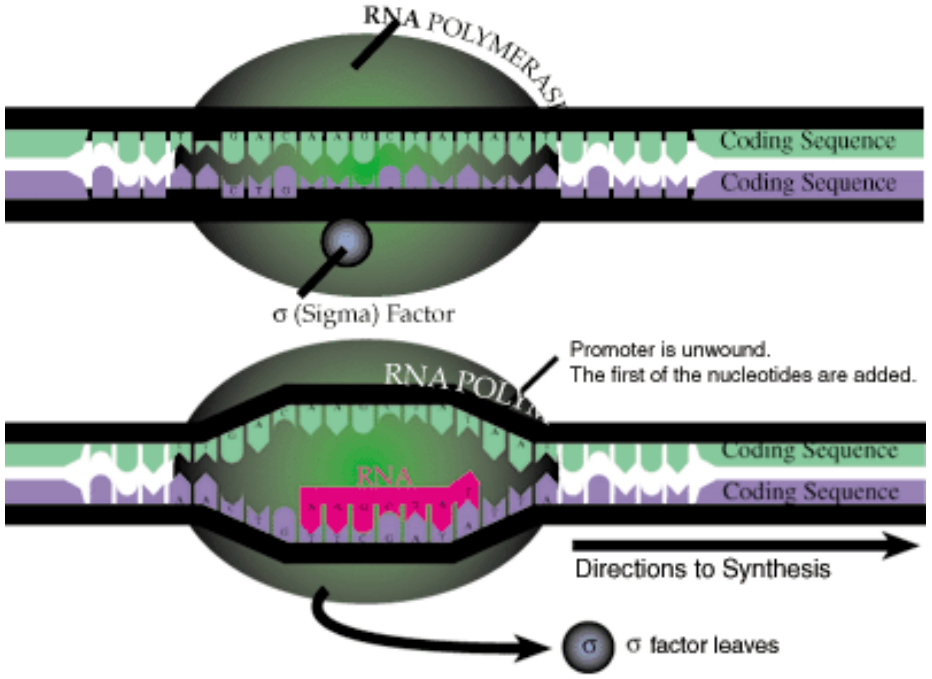
Gene in Prokaryoten

- Haben eine vergleichsweise einfache Struktur
  - Relativ feste Start- und Stopcodons
  - Open Reading Frame (ORF): Sequenz zwischen Start- und Stopcodon von >100 Basen Länge; Länge durch 3 teilbar
  - Sequenzsignale für Anfang und Ende der Transkription (TSS)
  - Promoterregion: Konservierte Motive im Abstand von -35 bzw. -10 Basen von der Transcriptional Start Site (TSS)



Promoter Region und RNA Polymerase





Quelle: Blackwell Pub., 11th hour

- RNA Polymerase: Komplex aus verschiedenen Proteinen
- Sigma-Faktoren erkennen unterschiedliche DNA-Motive
  - Produktion der Sigma-Faktoren hängt von Umwelt ab und regelt z.T. die Reaktion der Zelle auf Umwelteinflüsse
- Polymerase bindet erst, wenn Sigma-Faktor gebunden

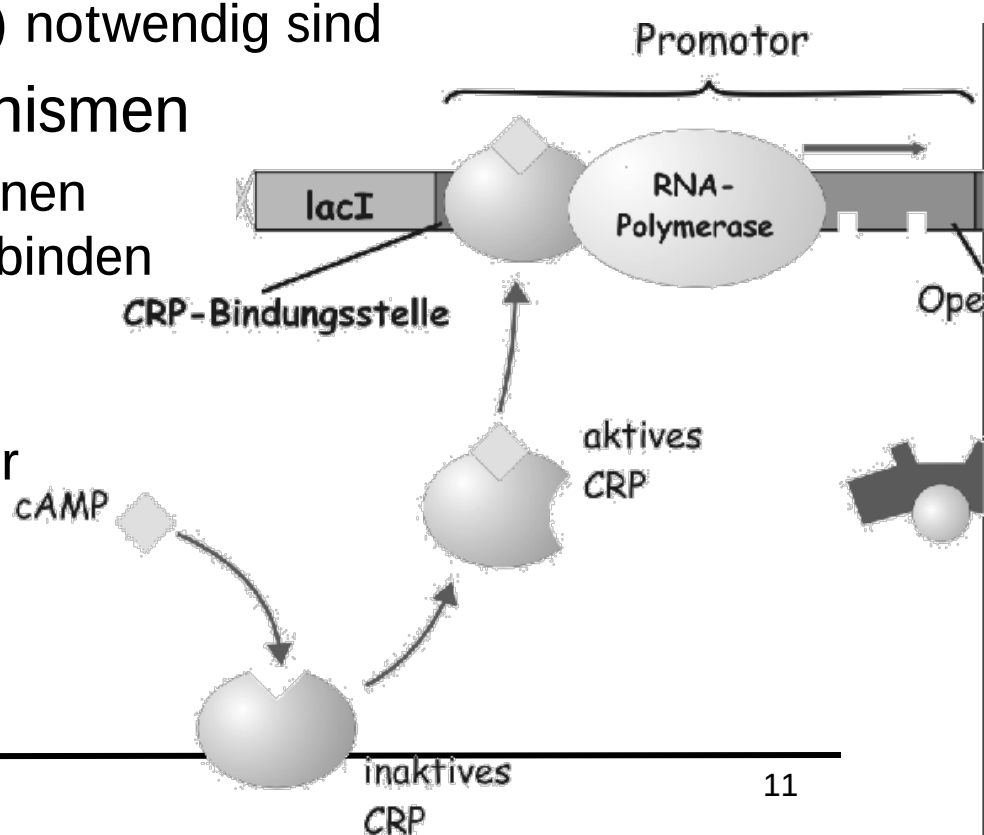
Sigma-Faktoren

Faktor	Erkennungs-sequenz -35	Erkennungs-sequenz -10	Bedingungen
σ^{70}	**TTGACA**	**TTGACA**	Normal (~70% aller Gene)
σ^{32}	**CTTGAA**	**CTTGAAA**	Hitzestress
σ^{54}	**CTGGCAC**	**CTGGCAC**	Stickstoffmangel
σ^{28}	**TAAA**	**CTAAA**	...
...	...		...

- Verschiedene σ-Faktoren binden an versch. Sequenzmotive
  - E.Coli hat 7 Faktoren; andere Spezies haben mehr/weniger
- Motive müssen nicht perfekt erhalten sein
  - Dargestellt sind Consensus-Sequenzen
  - Je größer die Abweichung, desto geringer die Expression des regulierten Gens

Regeln und Abweichungen

- Nicht alle Gene haben eigene Promoter
  - Operons: Gruppen von Genen, deren Expression durch einen gemeinsamen Promoter reguliert wird (nur in Prokaryoten)
  - Z.B. Gruppen von Genen, die zur Bewältigung einer Aufgabe (Hitzestress, Zellteilung, etc.) notwendig sind
- Weitere Regulationsmechanismen
  - Unterdrückung: Proteine können zwischen Promotor und TSS binden und Bindung der Polymerase verhindern
  - Aktivierung: Bindung weiterer Proteine in der Nähe des Promoters kann Effizienz der Expression erhöhen



Open Reading Frames (ORFs)

- Prokaryotische Gene haben keine Introns
- Nahezu alle DNA ist kodierend
- Open Reading Frame
  - Bereich auf dem Chromosom, der kodierend sein könnte
  - Sollte länger als 60 Codons sein (trifft für ~98% aller Gene zu)
  - Start-Codon AUG (meistens)
    - AUG kodiert auch für Methionin – kein eindeutiges Signal
  - Stop-Codons UAA, UAG, UGA
- Kann man relativ leicht und schnell finden

Gene Prediction in Prokaryoten

- Verfügbare Evidenzen
  - ORFs
  - Konservierte Promotor-Sequenzen
  - In einem ORF ist die dritte Base jedes Codons häufiger gleich als statisch erwartet
    - Grund: Spezies favorisieren spezifische Codons für Aminosäuren, bei denen es mehrere Möglichkeiten gibt
  - Transcriptional Stop Site, Shine-Delgardo-Sequenz, ...
- Wenn man diese Eigenschaften (fast) alle gefunden hat, hat man mit hoher Wahrscheinlichkeit ein Gen
  - Wahrscheinlichkeit eines Falsch-Positiven Hits für ein beliebiges ORF der Länge 60 Codons
    - 60-mal kein Stop-Codon sehen: (61/64)^60 ~ 4%

Eukaryoten

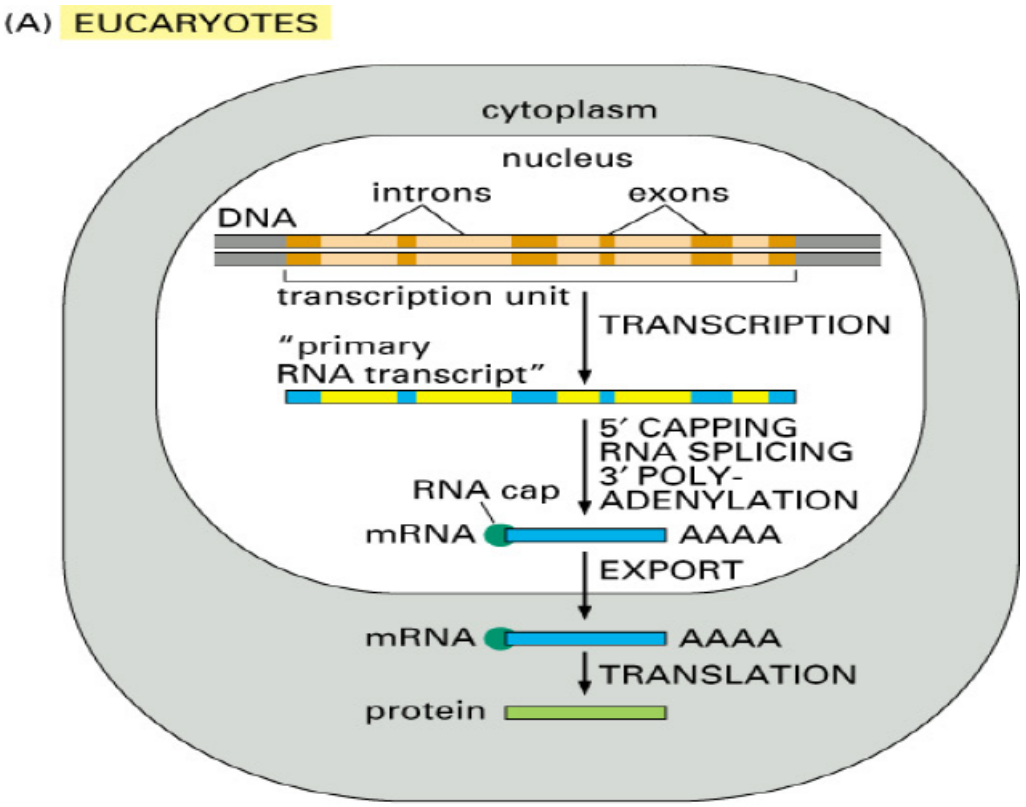
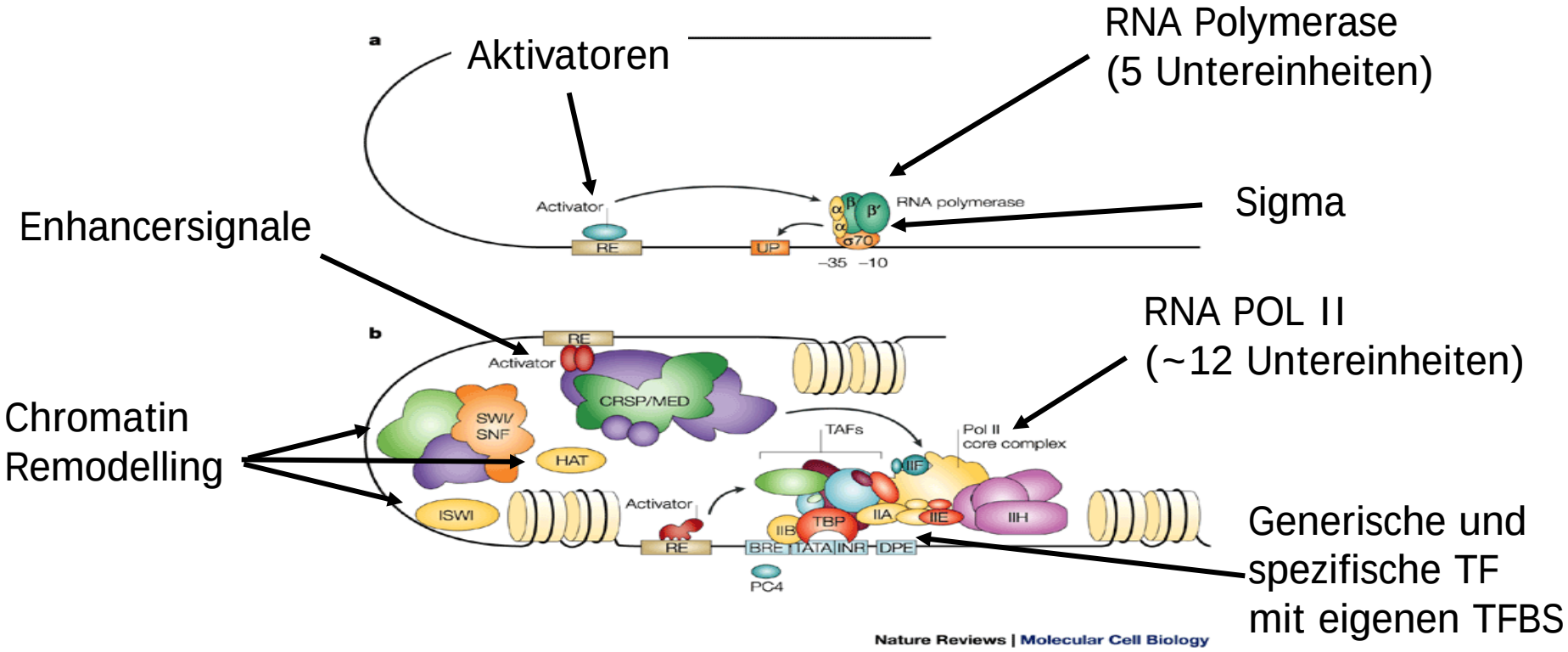


Figure 6–21 part 1 of 2. Molecular Biology of the Cell, 4th Edition.

Quelle: William Stafford Noble

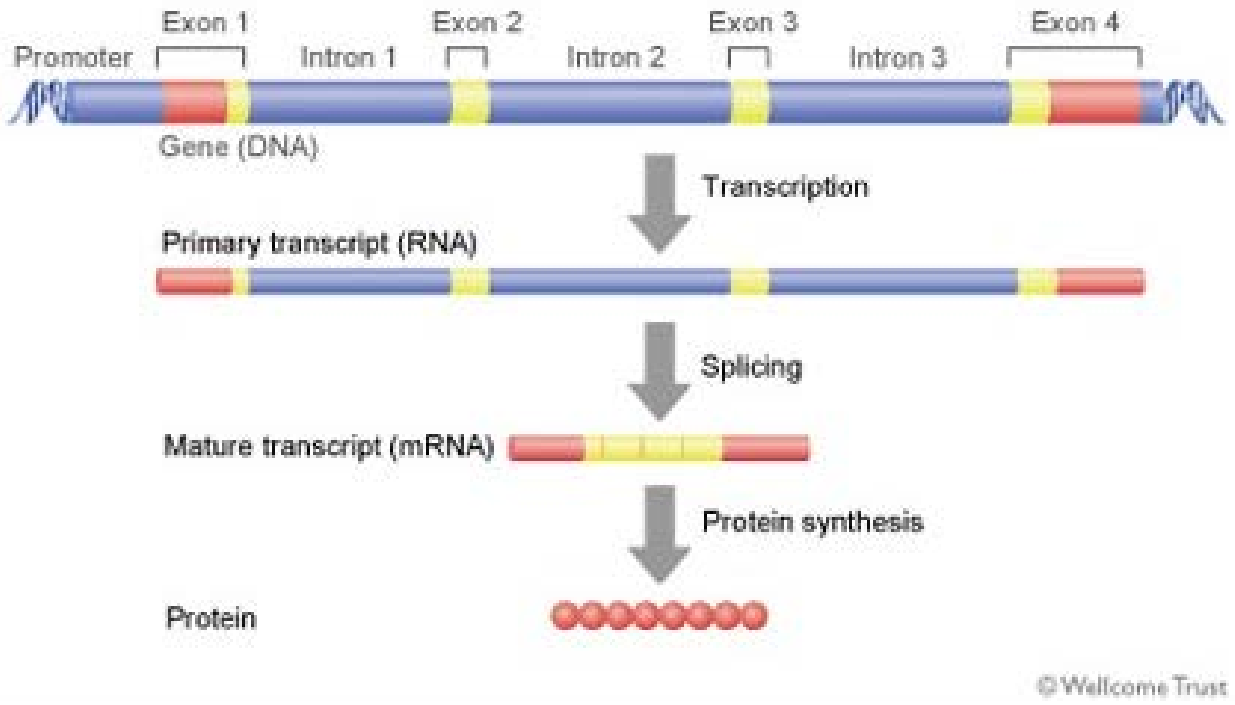
- Introns: variable Zahl/Länge
  - können >MB lang sein
- Differentielles Splicing
- 3 RNA-Polymerasen
- Promoterregionen können MB'n entfernt sein
- Polymerase bindet nur bei Vorhandensein mehreren Transcription Factors (TF)
  - Mensch: ~2000 TF
  - Expression benötigt im Schnitt ~5 gebundene TFs
- Viel nicht-kodierende DNA
- ...

Polymerase Initiation Complex



- Warum so komplex? Unterschiedliche Expressionsmuster
  - Viele Gewebetypen mit spezifischen Aufgaben
  - Entwicklungsprozess jedes Individuums mit verschiedenen Stadien

Grobe Genstruktur bei Eukaryoten

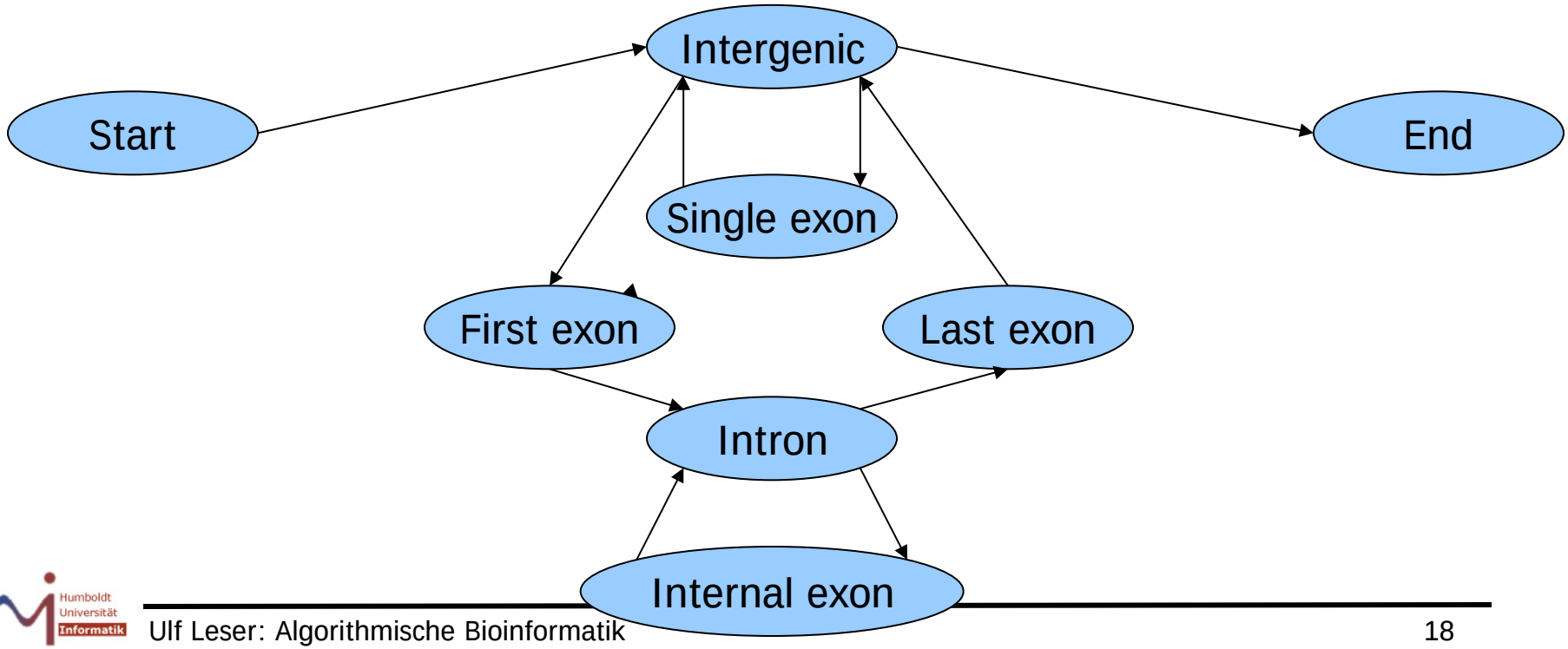


Modellierung: Module

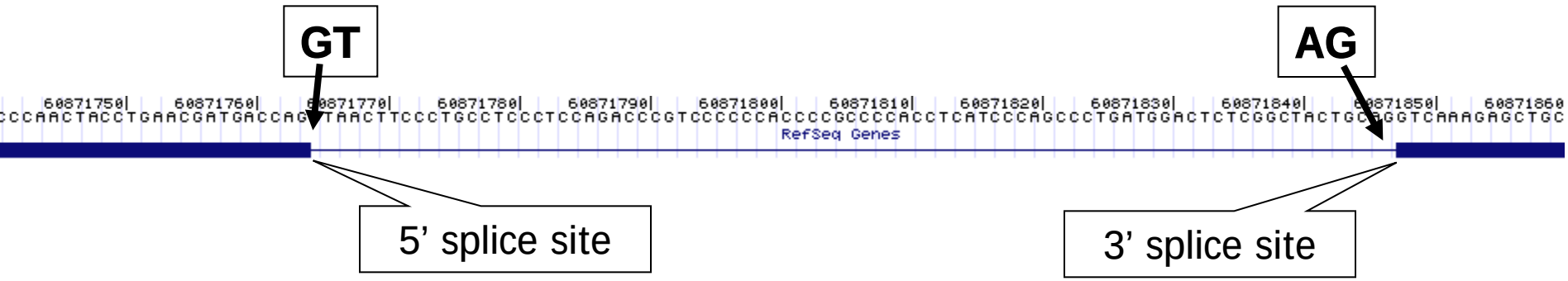
- Exons, Introns, ... nennen wir Module eines Gens
  - Signale: Feste Länge (kurz) und „relativ“ feste Sequenz
    - Splicestellen, Start- und Stop-Codons, TFBS
  - Blöcke: Keine feste Länge, variable Sequenz
    - Exons, Introns, UTRs, Promoterregionen
- Wie kann man ein Gen samt seiner Modulstruktur finden?
  - Module haben meistens keine feste Grenzen
  - Verschiedene Arten von Modulen haben best. Eigenschaften
    - Länge von Coding Regions durch 3 teilbar
    - Exons sind meistens kürzer als Introns, Introns können sehr lang sein
    - Start- und Stop-Codons (nicht feste, aber stark präferiert)
    - Splicestellen sind 99% konserviert (GT, AG)
    - Exons und Introns haben unterschiedliche Basenzusammensetzung
    - ...

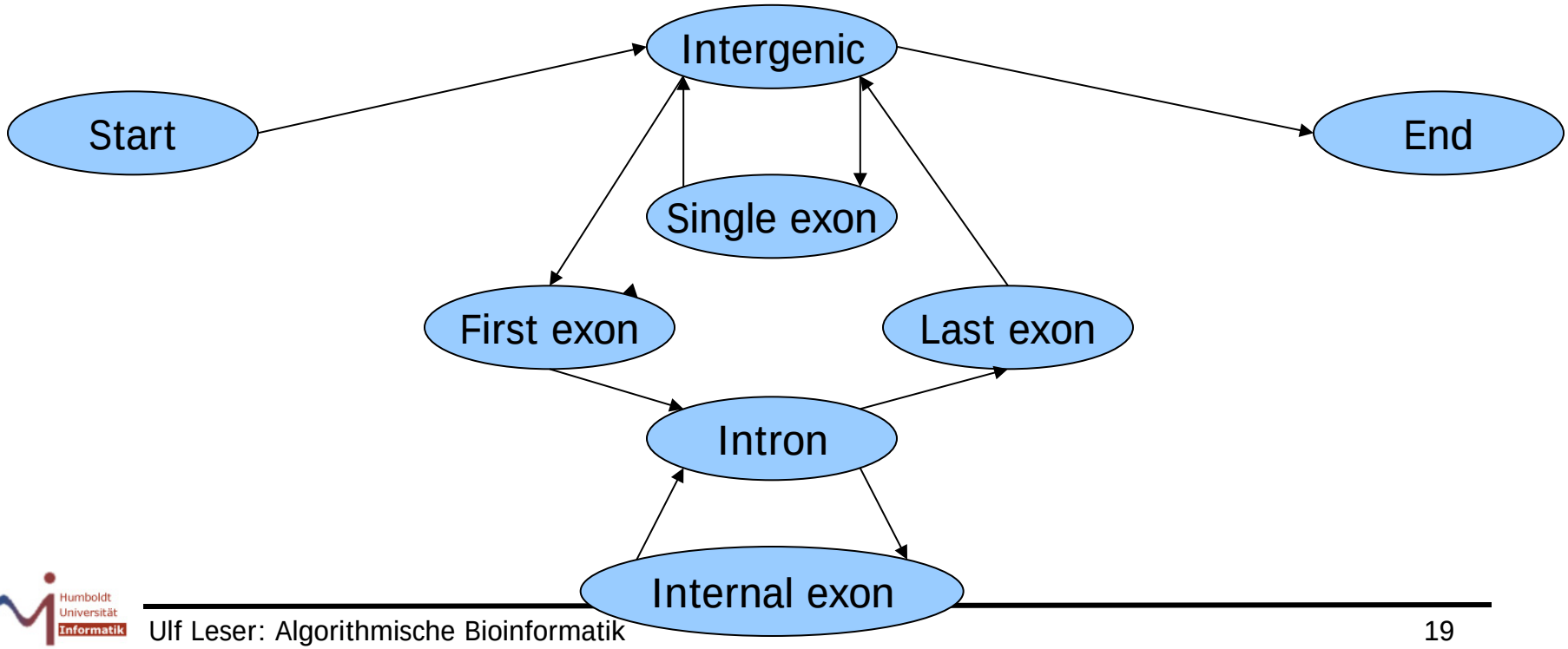
Einfaches Zustandsmodell

- Stellen wir uns vor, jede Base hat einen Zustand
  - Die Modulart, zu der sie gehört
- Folgende Übergänge sind erlaubt
  - Übergänge von Zustand Z zu sich selbst nicht enthalten

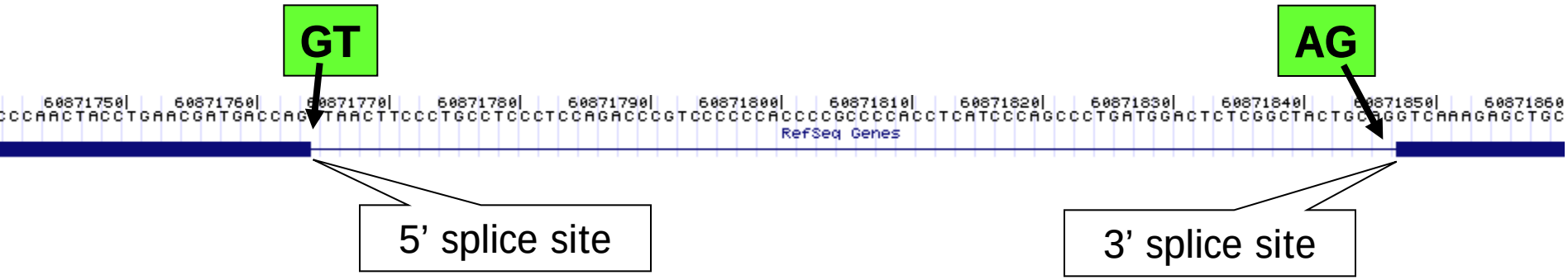


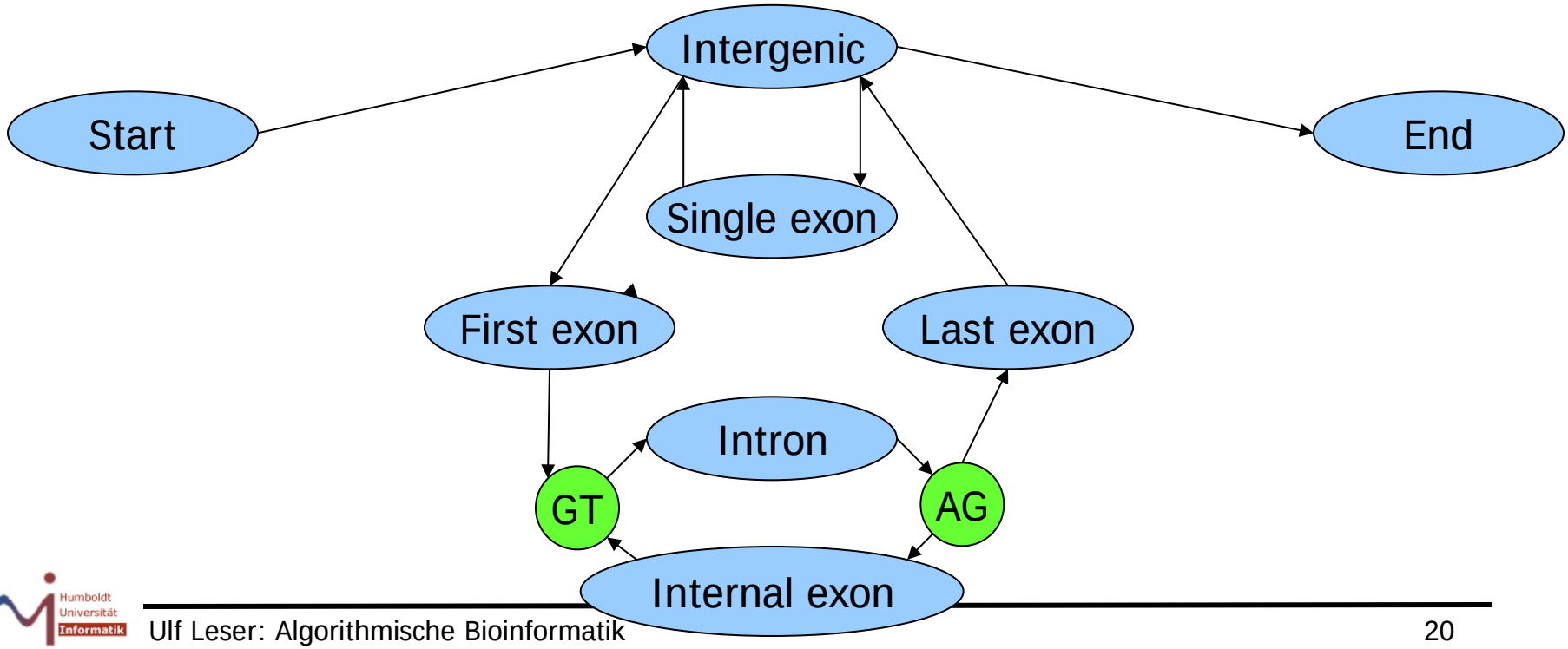
Exon-Intron-Grenzen



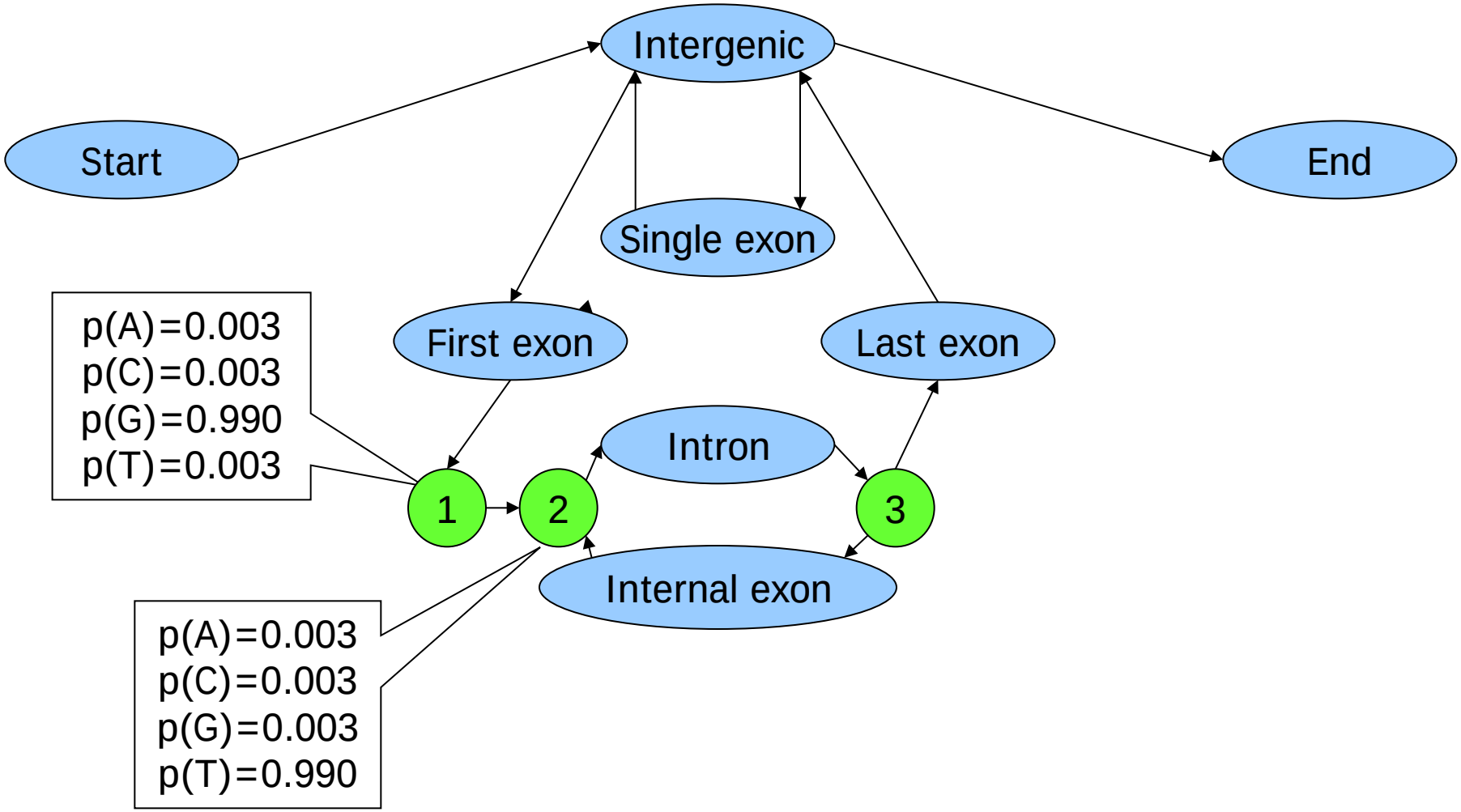


Signale für Exons/Introns

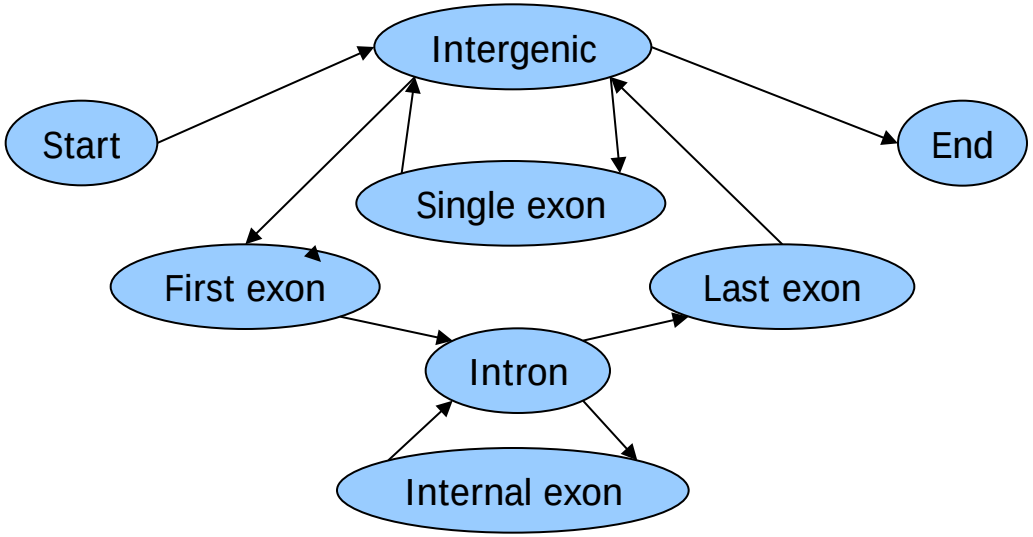




Wahrscheinlichkeiten

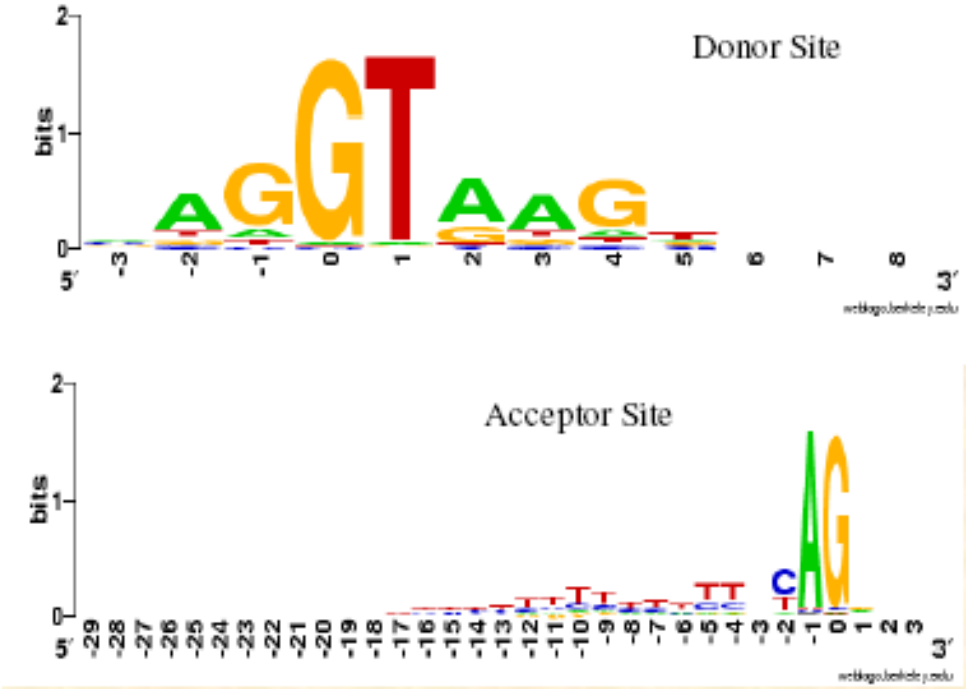


Probabilistische Automaten



- Module sind Zustände des Modells
- Zustände emittieren Basen
- Emissionen erfolgen nur mit einer bestimmten Wahrscheinlichkeit
- Pfeile sind Zustandsübergänge
- Auch Übergänge haben eine bestimmte Wsk
- Das ist ein Hidden Markov Model (HMM)

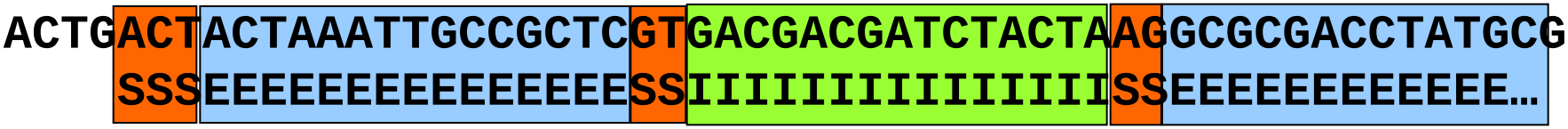
Echte Splicestellen



- Auch Basen links/rechts vom Signal sind konserviert
- Kann man als weitere Zustände in das Modell aufnehmen

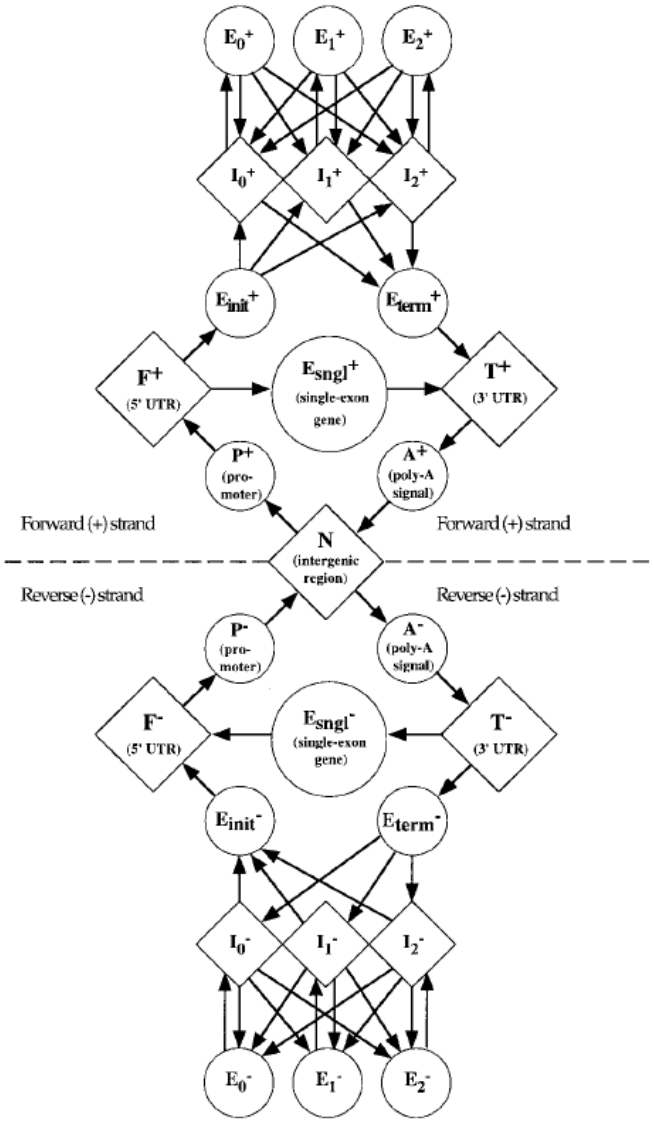
Probleme (informell)

- Einer gegebenen DNA-Sequenz kann man erst mal nicht ansehen, aus welcher Zustandssequenz sie am wahrscheinlichsten generiert wird
  - Alle emittieren A,C,G,T, nur mit (geringfügig) unterschiedlicher Wsk
- Problem 1: Gegeben eine Sequenz und ein Modell: Finde die Modulgrenzen (also die Zustandsübergänge)



- Problem 2: Gegeben viele Gene: Finde die Übergangs- und Emissionswahrscheinlichkeiten des Modells
  - Und womöglich das Modell selber

Beispiel: GeneScan



- Burge, C. and Karlin, S. (1997). "Prediction of complete gene structures in human genomic DNA." *J Mol Biol* 268(1): 78-94.
- Modell mit 27 Zuständen
- Erkennungsgenauigkeit (1997)
  - ~90% für Basen (in Gen oder nicht)
  - ~80% für: In Exon oder nicht
  - ~43% für komplete Genstruktur
- Trainingsdaten: ~400 humane Gene

Inhalt der Vorlesung

- Gene Finding
- Struktur von Genen
- CpG Inseln und Markov Modelle

CpG Inseln

- Mit “CpG” bezeichnet man das Nukleotidpaar CG
  - CpG: Hintereinander auf einem Strang, nicht die Paarung C-G
  - „p“: Phosphodiesterbrücke zwischen C und G
- CpG's sind statistisch überraschend selten im humanen (und anderen eukaryotischen) Genom
  - Das C in CpG kann methyliert werden
  - Dadurch höhere Mutabilität
- Aber: Im Bereich ab ca. 1500 Basen vor einem Gen ist die Dichte an CpG „normal“
  - Erklärung: Methylierung erhöht die Histon-Bindung der DNA
  - Dadurch wird die Expression erschwert
  - Zusätzliches Regulationsprinzip
  - Wird eng mit gewebespezifischen Expressionsmustern assoziiert

CpG Inseln

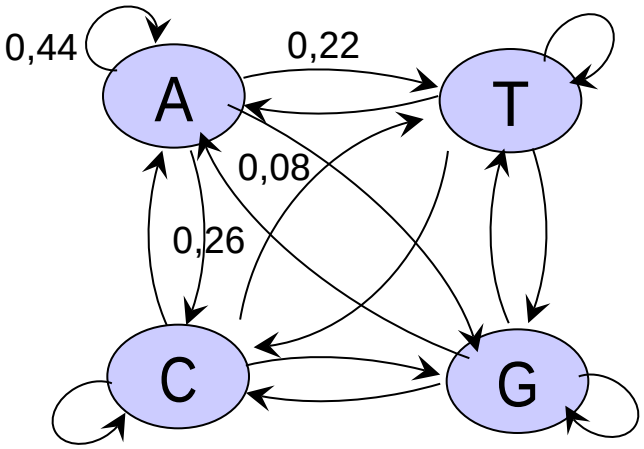
- CpG-Inseln
  - Sequenzabschnitte, in denen mehr CpG als erwartet (bezogen auf absolute Häufigkeit im Genom) vorkommen
  - Die meisten CpG Inseln liegen vor Genen
  - Die meisten Gene liegen hinter einer CpG Insel
- Wie kann man für einen Sequenzabschnitt entscheiden, ob er eine CpG Insel ist?
  - Wir wissen, dass bestimmte Dinukleotide häufiger sind als sonst
    - Nach C kommt häufiger ein G
  - Erster Versuch: Markov-Modelle

Markov-Modell (oder Markov-Kette)

- Definition
  Gegeben ein Alphabet Σ. Ein Markov-Modell erster Ordnung ist ein sequentieller stochastischer Prozess (Zustandsfolge) über $|\Sigma|$ Zuständen $s_1, \ldots, s_n$ mit
  - *Jeder Zustand s_i emittiert genau ein Zeichen aus Σ*
  - *Keine zwei Zustände emittieren das selbe Zeichen*
  - *Für eine Folge $z_1,z_2,\ldots$ von Zuständen gilt:*
    $$p(z_t=s_t|z_{t-1}=s_{t-1}, z_{t-2}=s_{t-2},\ldots, z_1=s_1) = p(z_t=s_t|z_{t-1}=s_{t-1})$$
  - *Die $a_{0,i} = p(z_1=s_i)$ heißen Startwahrscheinlichkeiten*
  - *Die $a_{i,j} = p(z_t=s_j|z_{t-1}=s_i)$ heißen Übergangswahrscheinlichkeiten*
  - *Außerdem:* $\sum_i a_{i,j} = 1$
- Bemerkung
  - Wsk eines Zustands hängt nur vom Vorgängerzustand ab

Visualisierung

- Jeder Zustand einer Markov-Kette emittiert ein eindeutiges Zeichen des Alphabets
  - Daher können wir Zustände und Zeichen verschmelzen
    - Bei HMM geht das nicht, daher trennen wir jetzt schon in der Definition
- Markov-Modell als Zustandsgraph
  - Knoten sind die Zeichen des Alphabets (Zustände)
  - Kanten sind mit Übergangswahrscheinlichkeiten beschriftet



Hier sind alle Zustände mit allen verbunden; das muss nicht so sein (a_{ij}=0)

Bemerkung

- Eins zweidimensionale quadratische Matrix M heisst (spalten-)stochastisch, wenn folgendes gilt
  - $\forall i,j: 0 \leq M[i,j] \leq 1$
  - $\forall j: \Sigma\, M[i,j] = 1$

- Die Übergangsmatrix $M = [a_{i,j}]$ eines Markov-Modells ist eine stochastische Matrix

Wahrscheinlichkeit einer Zustandsfolge

- Gegeben ein Markov-Modell M mit Übergangswsk a und eine Sequenz S von Zeichen aus Σ
- Wir lassen den stochastischen Prozess laufen; M wird eine Sequenz S erzeugen
- Wie groß ist die Wsk, dass M genau S erzeugt?

$$p(S \mid M) = p(z_1 = S[1]) * \prod_{i=2..n} p(z_i = S[i] \mid z_{i-1} = S[i-1])$$

$$= a_{0,S[1]} * \prod_{i=2..n} a_{S[i-1],S[i]} = a_{0,1} * \prod_{i=2..n} a_{i-1,i}$$

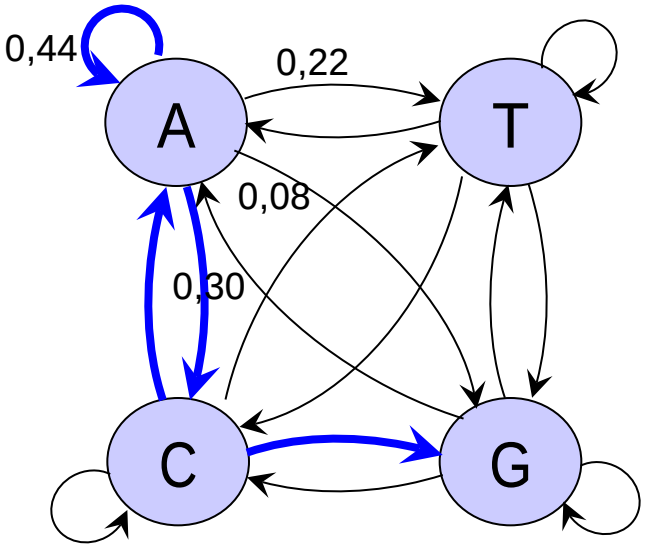
  - Deterministisch: Da Zustände eindeutige Zeichen emittieren, kann jedes S nur durch genau eine Zustandsfolge erzeugt werden

Vereinfachung

- Startzustände machen die Formeln hässlich
- Vereinfachung
  - Einführung eines expliziten neuen Startzustands s_0
  - Jede Zustandsfolge beginnt mit $z_0=s_0$
  - Seine Wahrscheinlichkeit ist fix 1 und er emittiert kein Zeichen des Alphabets
  - Damit

$$p(S \mid M) = a_{0,1} * \prod_{i=2..n} a_{i-1,i} = \prod_{i=1..n} a_{i-1,i}$$

Beispiel



$$P(CAACG|M) = p(z_1=C|z_0)* p(z_2=A|z_1=C) * p(z_3=A|z_2=A) * p(z_4=C|z_3=A) * p(z_5=G|z_4=C)$$

$$= a_{0C} * a_{CA} * a_{AA} * a_{AC} * a_{CG}$$

CpG Inseln revisited

- Wie unterscheiden sich CpG Inseln von anderen Sequenzen?
- Durch ihre Übergangswahrscheinlichkeiten

M+	A	C	G	T
A	.180	.274	.426	.120
C	.171	.368	.274	.188
G	.161	.339	.375	.125
T	.079	.355	.384	.182

M-	A	C	G	T
A	.300	.205	.285	.210
C	.233	.298	.078	.302
G	.248	.246	.298	.208
T	.177	.239	.292	.292

Quelle: [DEK+03]; 48 humane CpG Islands, 60.000 Basen

CpG Inseln erkennen

- Erster Versuch: Wir bilden zwei Markov-Modelle
  - Modell M+ für die Übergangshäufigkeiten in CpG Inseln
  - Modell M- für die Übergangshäufigkeiten in normaler Sequenz
  - Berechnung des Log-Odds-Score

$$s = \log\left(\frac{p(S \mid M+) * p(M+)}{p(S \mid M-) * p(M-)}\right) = \sum_{i=1}^{n} \frac{\log(a^{+}_{i-1,i})}{\log(a^{-}_{i-1,i})} + \log\left(\frac{p(M+)}{p(M-)}\right)$$

- s>0: Die Sequenz ist wahrscheinlich eine CpG Insel
  - Je größer s, desto wahrscheinlicher
- s<0: Die Sequenz ist wahrscheinlich keine CpG Insel

CpG Inseln finden

- Aber: Die Frage: „Ist Sequenz S eine CpG Insel?“ ist nicht wirklich relevant
- Wichtiger: „Wo in S sind CpG Inseln?“
- Problem: Die Markov-Kette kann überall in S beginnen
- Lösung 1: Sliding Window (sei |S|=n)
  - Wir schieben ein Fenster der Größe w über S
  - Für jede Position bestimmten wir den Score s mit M+ und M-
  - Laufzeit: O(n) – Wie?
  - Problem: Welches w?
    - CpG Inseln haben keine fixen Längen
- Besser wäre ein längenunabhängiger Mechanismus

Geschichte

- Andrej Andrejewitsch Markov (1856-1922)
  - Russischer Mathematiker
  - Entwickelte Markov-Ketten-Modelle für Anwendungen in der Sprache
  - Statistische Analyse der Buchstabenfolgen in Novellen
  - Markov, A. A. (1913). "Beispiel statistischer Untersuchungen des Textes ‚Eugen Onegin', das den Zusammenhang von Ereignissen in einer Kette veranschaulicht (Original in Russisch)." *Bulletin de l'Academie Imperiale des Sciences de St.-Petersbourg*: 153-162.

Markov-Modelle höherer Ordnung

- Markov-Modelle höherer Ordnung
  - Markov-Modell der Ordnung k : Wsk von Zustand z_i hängt von den k Vorgängern ab
  - Nicht ausdrucksstärker als Markov-Ketten der Ordnung 1
  - Jedes Markov-Modell der Ordnung k mit n Zuständen kann in ein Markov-Modell der Ordnung 1 mit n^k Zuständen überführt werden
    - Beispiel : Zustände eines Markov-Modells der Ordnung 3 für DNA-Sequenzen: AAA, AAC, AAT, AAG, ACA, ...
- Kann aber intuitivere Modellierung ermöglichen

Selbsttest

- Was ist ein Sigma-Faktor?
- Erklären Sie die Struktur eines eukaryotischen Gens
- Was ist der Unterschied zwischen der Transcriptional Start Site und der Translational Start Site?
- Was ist ein Markov Modell? Was ist ein MM dritter Ordnung?