Genregulation und Erkennungssequenzen

Peter N. Robinson

Institut für medizinische Genetik
Charité Universitätsmedizin Berlin

21. Dezember 2015

Überblick: Heute

- Einführung in transkriptionelle Regulation
- Consensussequenzen
- Motive: Transkriptionsfaktorbindungsmotive
- Motive in Sequenzen finden
- Neue Motive entdecken

Outline

Die Kontrolle der Genexpression

- Die Expression menschlicher Gene auf verschiedenen Ebenen gesteuert
- Eines der wichtigsten Ziele in der Bioinformatik ist das Verständnis der Netzwerke, welche die Genexpression steuern
- Regulation erfolgt auf mehreren Ebenen
  - Transkription
  - Posttranskriptional
  - Epigenetische Mechanismen (beruhen nicht auf Veränderung der Gensequenz)
- Heute: Kurze Einführung in transkriptionelle Regulation

cis und trans

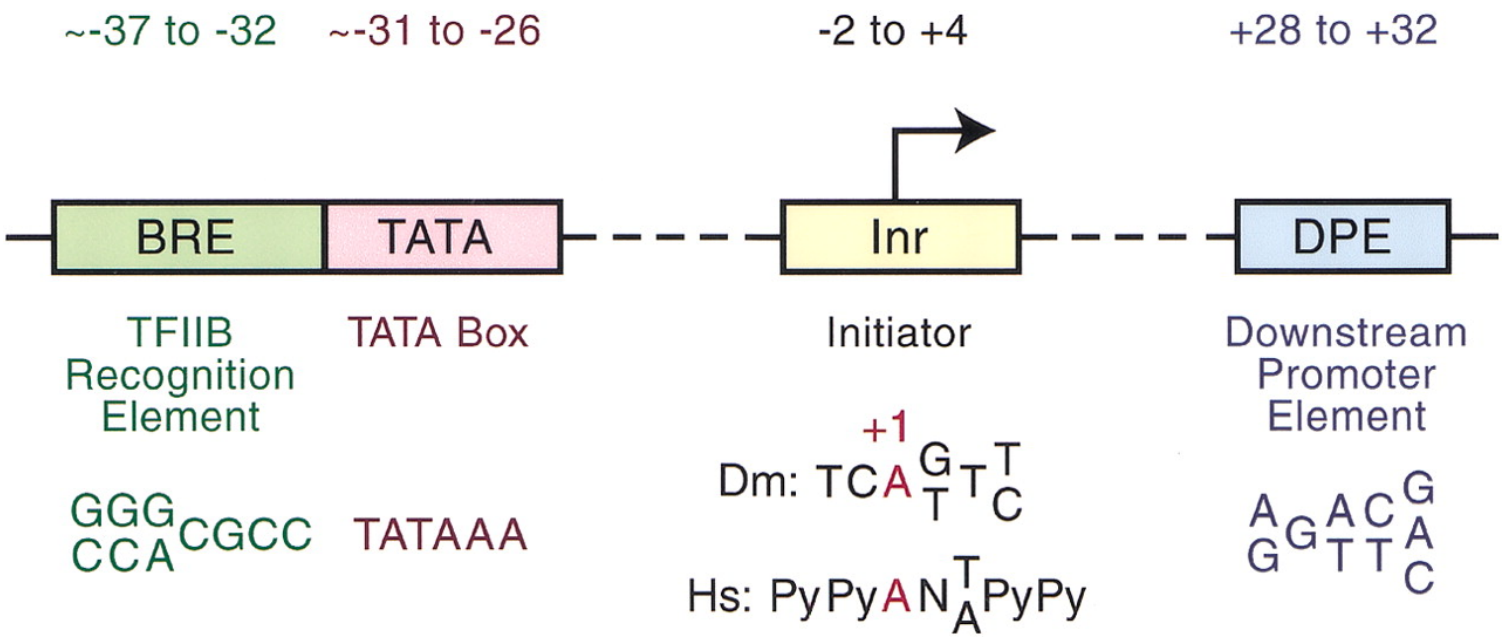
- Gallia cisalpina: *Gallien diesseits der Alpen*
- Gallia transalpina: *Gallien jenseits der Alpen*

Wikipedia commons

cis und trans (2)

- Genregulation auf Ebene der Transkription erfolgt über die Bindung von Proteinen an regulatorische DNA-Sequenzen
- trans-Faktoren: Die regulatorischen Proteine (Transkriptionsfaktoren) werden von entfernt liegenden Genen kodiert und gelangen erst nach ihrer Synthese im Zytoplasma und ggf. Aktivierung wieder in den Zellkern an ihren Wirkungsort, deshalb bezeichnet man sie als trans-aktiv
- cis-Sequenzen: Die regulatorische DNA-Sequenzen, an welche die trans-Faktoren binden, liegen i.d.R. in der Nachbarschaft des regulierten Gens und werden deshalb als cis-aktiv bezeichnet

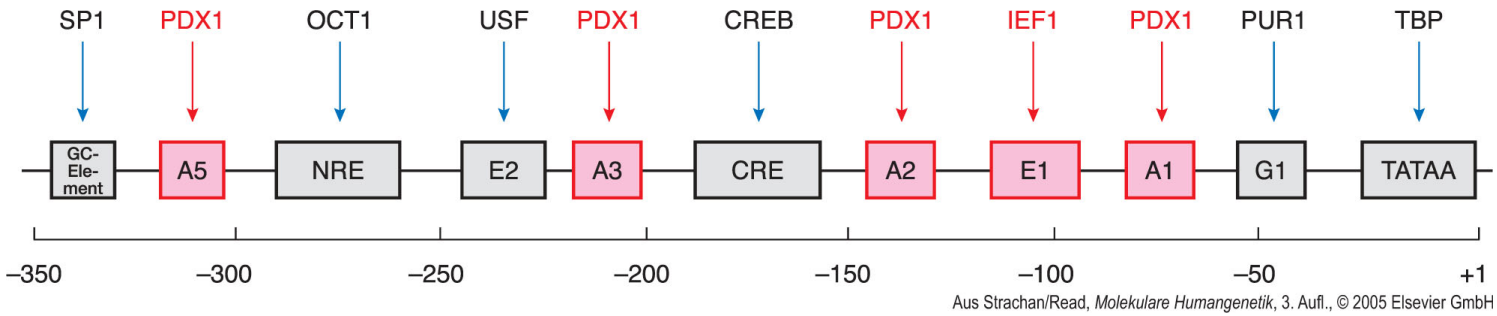
Kernpromotoren



Butler JE, Kadonaga JT (2002) The RNA polymerase II core promoter: a key component in the regulation of gene expression. *Genes Dev* **16**:2583-92.

- Der Core-Promoter bindet an mehr oder weniger ubiquitäre Transkriptionsfaktoren und ermöglicht so eine basale Transkriptionsrate, die durch Aktivität des weiter upstream gelegenen Promoters bzw. von Enhancern oder Silencern modifiziert werden kann
- Elemente eines Core-Promoters: TATA-Box, INR, DPE, BRE

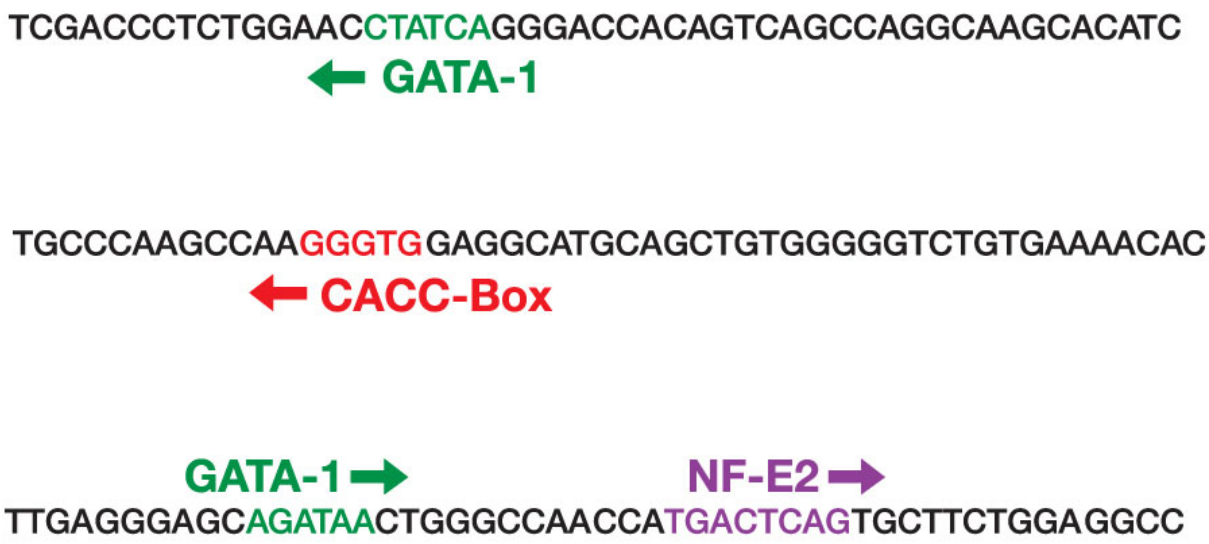
Insulinpromoter



Aus Strachan/Read, *Molekulare Humangenetik*, 3. Aufl., © 2005 Elsevier GmbH

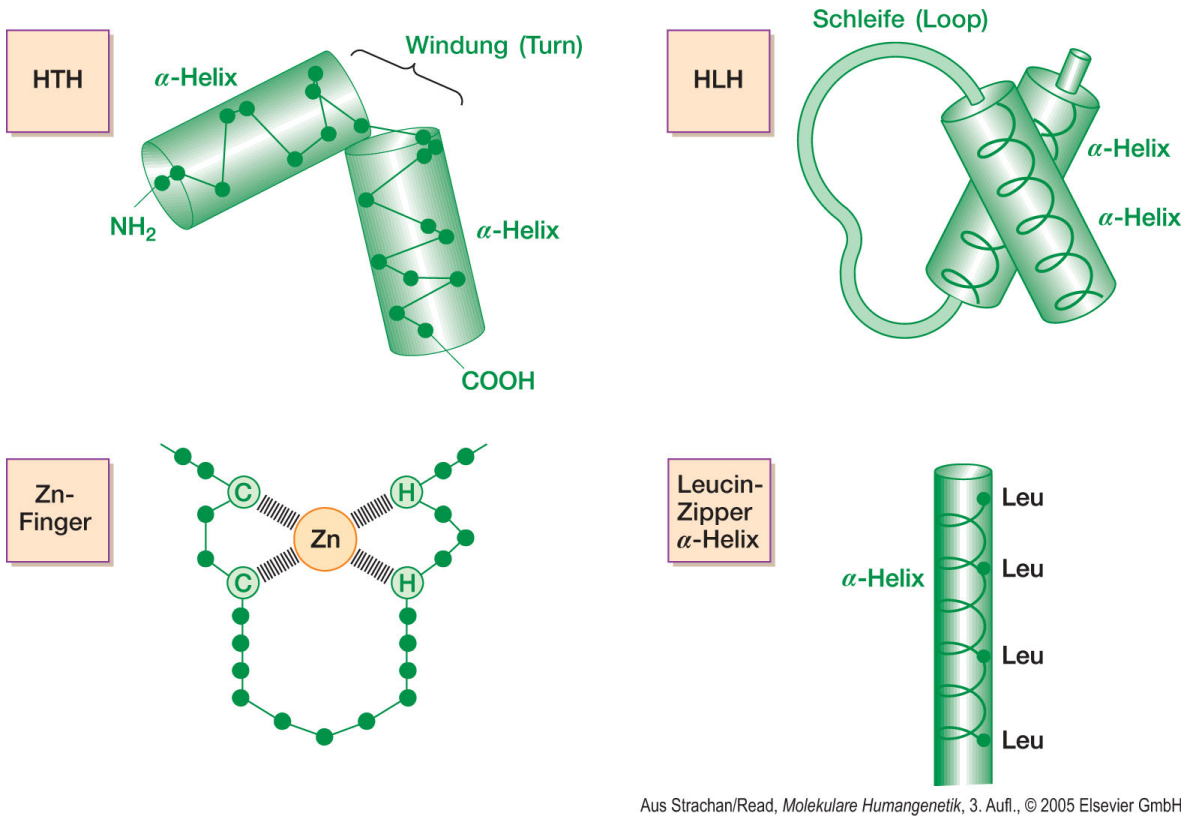
- Ubiquitäre oder allgemein exprimierte Transkriptionsfaktoren: schwarz
- für β-Zellen des Pankreas spezifische Faktoren: rot
- CRE, cAMP-Response-Element; NRE, negativ regulatorisches Element.
- PDX1 bindet an vier Sequenzmotive mit der Struktur C(C/T)TAATG, die im Insulinpromotor vorkommen (A1, A2, A3, A5)

Promoter des α-Globingens



- Die regulatorische HS-40-Sequenz des α-Globingens enthält viele Erkennungselemente für erythroidspezifische Transkriptionsfaktoren.

Strukturmotive



Aus Strachan/Read, *Molekulare Humangenetik*, 3. Aufl., © 2005 Elsevier GmbH

- Strukturmotive, die häufig in Transkriptionsfaktoren vorkommen
  - HTH, Helix-Turn-Helix; HLH: Helix-Loop-Helix

Bindung an DNA

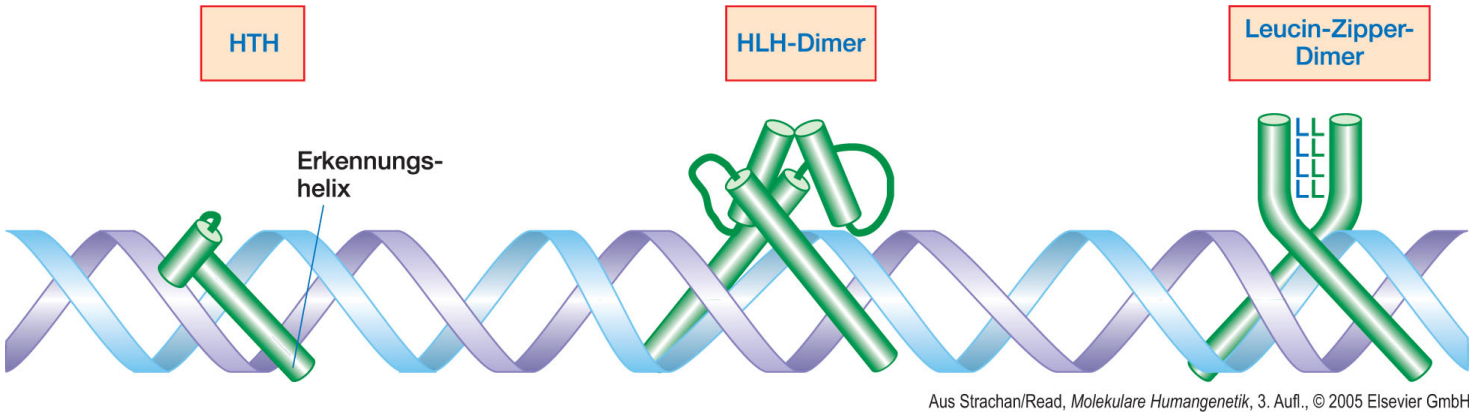
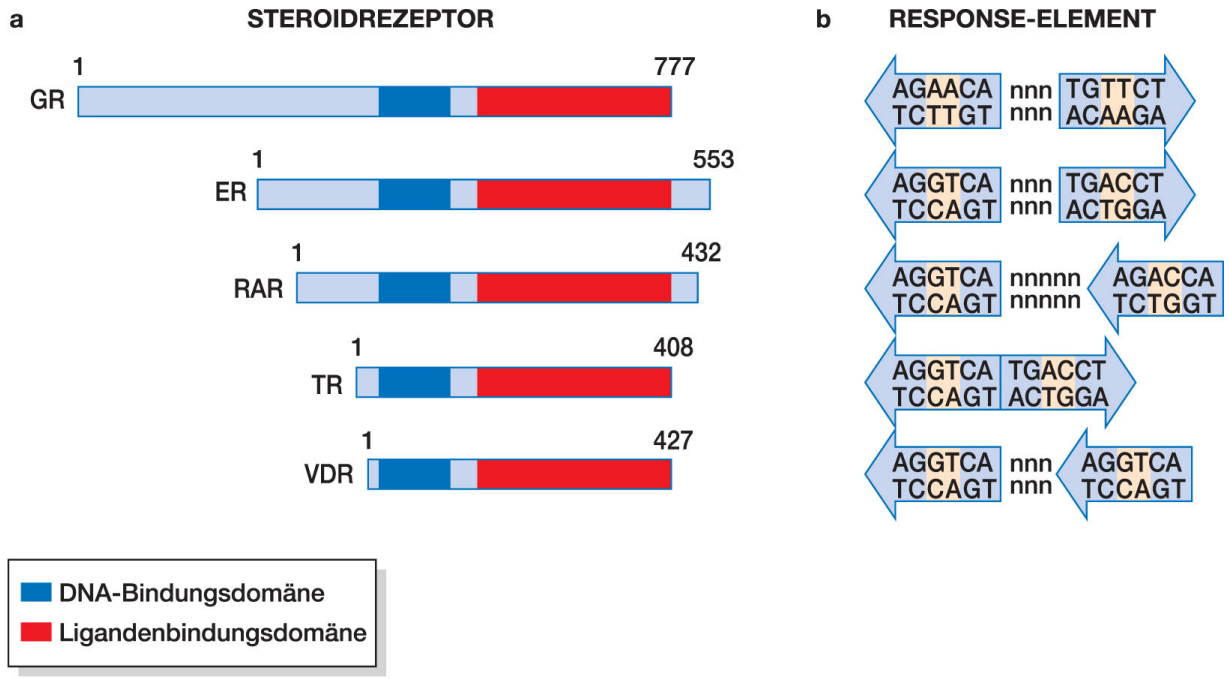


Abb. 10.9 Die Bindung von konservierten Strukturmotiven in Transkriptionsfaktoren an die Doppelhelix. HLH-Heterodimere und Leucin-Zipper-Heterodimere ermöglichen noch eine weitere Regulationsebene (siehe Text).

- Die Bindung von konservierten Strukturmotiven in Transkriptionsfaktoren an die Doppelhelix.

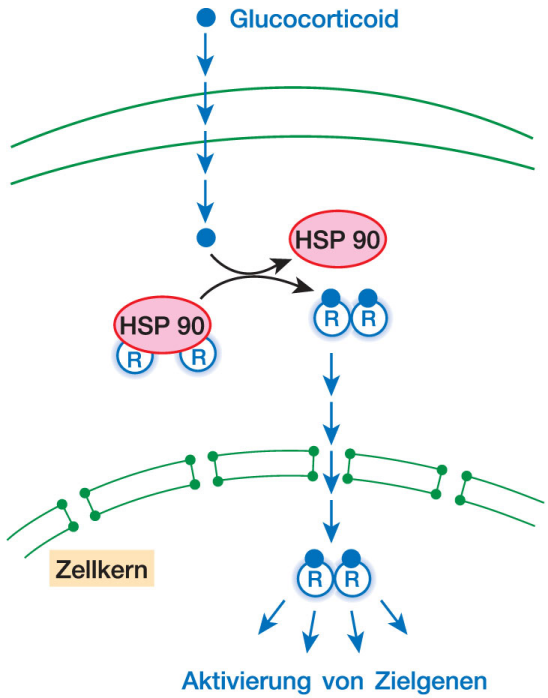
Steroidrezeptoren



Aus Strachan/Read, *Molekulare Humangenetik*, 3. Aufl., © 2005 Elsevier GmbH

- ER, Östrogenrezeptor; GR, Glucocorticoidrezeptor; PR, Progesteronrezeptor; RAR, Retinsäurerezeptor; TR, Thyroxinrezeptor; VDR, Vitamin D-Rezeptor.
- Kleine hydrophobe Hormone diffundieren durch die Plasmamembran, binden an und aktivieren Kellkernhormonrezeptoren, welche in den Kern wandern und an spezifische DNA-Response-Elemente binden, und somit die (50–100) Zielgene aktivieren

Steroidrezeptoren (2)



- Der Glucocorticoidrezeptor wird normalerweise durch die Bindung des Inhibitorproteins Hsp90 inaktiviert. Die Bindung von Glucocorticoiden an den Rezeptor setzt Hsp90 frei, der Rezeptor dimerisiert und aktiviert dann spezifische Gene, die im Promotor ein Glucocorticoid-Response-Element enthalten.

Outline

Konsensussequenz

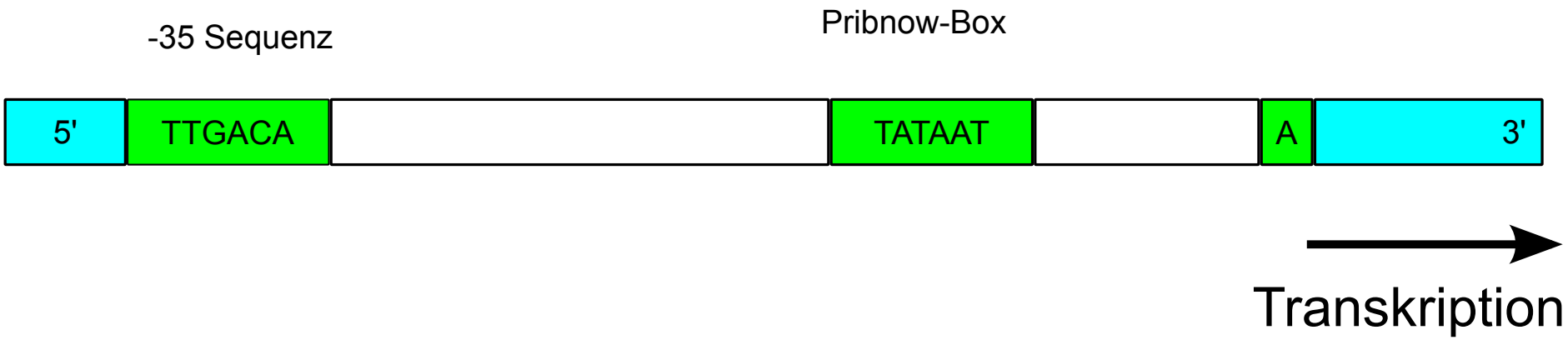
```
Ba-g   (s)   TGGGTTA> TATAGGGGGTAGT
Ba-g   (g)   TGGG-TA> TATAGAGGGACAT

Tca          TGGTTTT> TATACGTGAAAGC
Dpu    (g)   ATTGATA> TATACGAGAGAAT
Dpu    (s)   TTTCCTA> TATAGGGAAGAGG
Afr    (g)   GACCTTA> TATAAGGGGGTCA
Afr    (s)   TCTTGTA> TATAGAGAAGTCA

Consensus         TA> TATANGRRRR
```

- Diejenige Sequenz von Nukleotiden oder Aminosäuren bezeichnet, welche in der Summe am wenigsten von einer gegebenen Menge von entsprechenden Mustersequenzen abweicht.
- Beispiel `A[CT]N{A}YR`
  - `A`: A koomt immer an der angegebenen Stelle vor
  - `[CT]` entweder C oder T
  - `N`: beliebig
  - `{A}` beliebige Base außer A
  - `Y`: beliebige Pyrimidinbase
  - `R`: beliebige Purinbase

Pribnow box



- Prokaryotische Promotoren weisen häufig eine -35 upstream Sequenz sowie eine -10 upstream Sequenz (auch “Pribnow-Box”) auf
- Die Grafik zeigt eine Consensussequenz für die Pribnow-Box, bei der jeweils die häufigste Base dargestellt wird

Pribnow box

- Es fällt jedoch auf, dass die Positionen 3–5 eine nur schwache Vorliebe für bestimmte Basen haben, nur 49%–59% aller Pribnow-Boxen weisen die angegebene Base auf

T	A	T	A	A	T
82%	89%	52%	59%	49%	89%

- Von 322 Pribnow-Sequenzen in *e. coli* weisen nur 15 die eigentliche ”Konsensussequenz“ auf (4.7%)

Djordjevic M (2011) Redefining Escherichia coli σ(70) promoter elements: -15 motif as a complement of the -10 motif.

J Bacteriol **193**:6305–6314

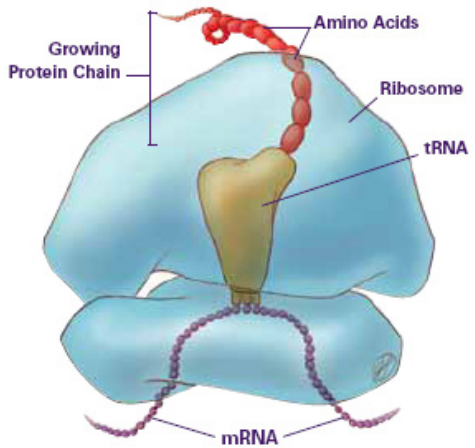
Pribnow box

- Diese Darstellung der Pribnow-Box-Sequenz vermittelt viel mehr Informationen über die tatsächlichen Sequenzen
- Heute wollen wir untersuchen, welche Probleme Consensussequenzen haben und wie PSSMs, Sequenzlogos usw funktionieren.

Outline

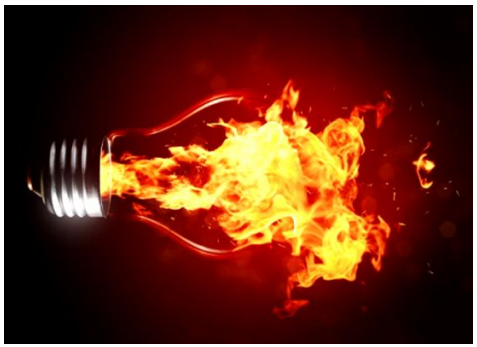
Ribosomen-Bindungssequenz

- Wir werden einige Themen in der Informationstheorie anhand von Ribosomen-Bindungssequenzen untersuchen
- Bakterielle mRNAs enthalten eine kurze Ribosomen-Bindungssequenz, die sich wenige Nukleotide vor dem AUG-Startkodon befindet
- Diese Sequenz bildet mit der RNA der kleinen ribosomalen Einheit Basenpaarungen aus und bringt das AUG-Startkodon in die richtige Position für die Initiation der Translation



Ribosomen-Bindungssequenz

- Die Ribosomen müssen die Ribosomen-Bindungssequenzen mit einer hohen Effizienz finden, um einen Selektionsnachteil zu vermeiden
- Finden die Ribosomen die korrekten Bindungssequenzen nicht, werden die entsprechenden Proteine nicht hergestellt
- Binden sie unspezifisch, werden zufällige Sequenzen übersetzt, was Energie verschwendet



Ribosomen-Bindungssequenz

- E. coli hat etwa **2600** Gene
- All diese Gene weisen eine Ribosomen-Bindungssequenz vor dem Translationsstart auf
- Alle mRNAs zusammen haben ca. 4.7×10^6 Nukleotide
- Das Ribosom muss also jeweils eine von 2600 unter 4.7×10^6 Sequenzen finden!

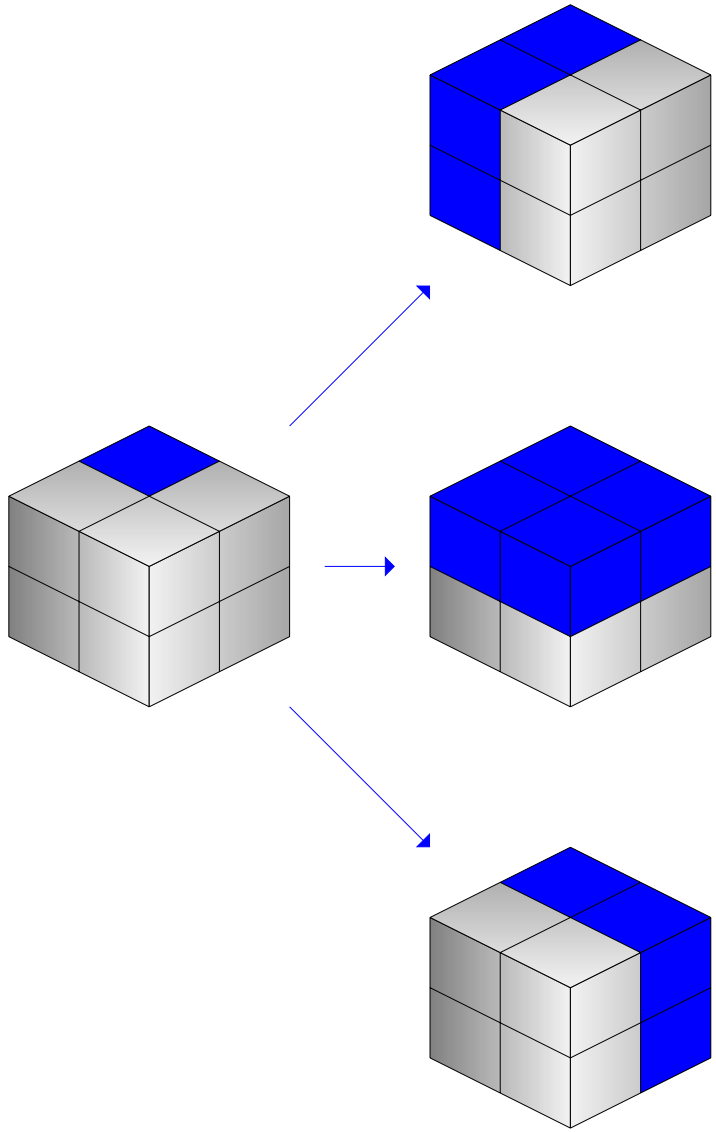

Hunts Needle in a Haystack

Ribosomen-Bindungssequenz

- Gemäß der Informationstheorie muss das Ribosom $\log_2 \frac{4.7 \times 10^6}{2600} \sim 11$ “Entscheidungen” treffen, um eine solche Sequenz zu finden...
- Wie funktioniert das?

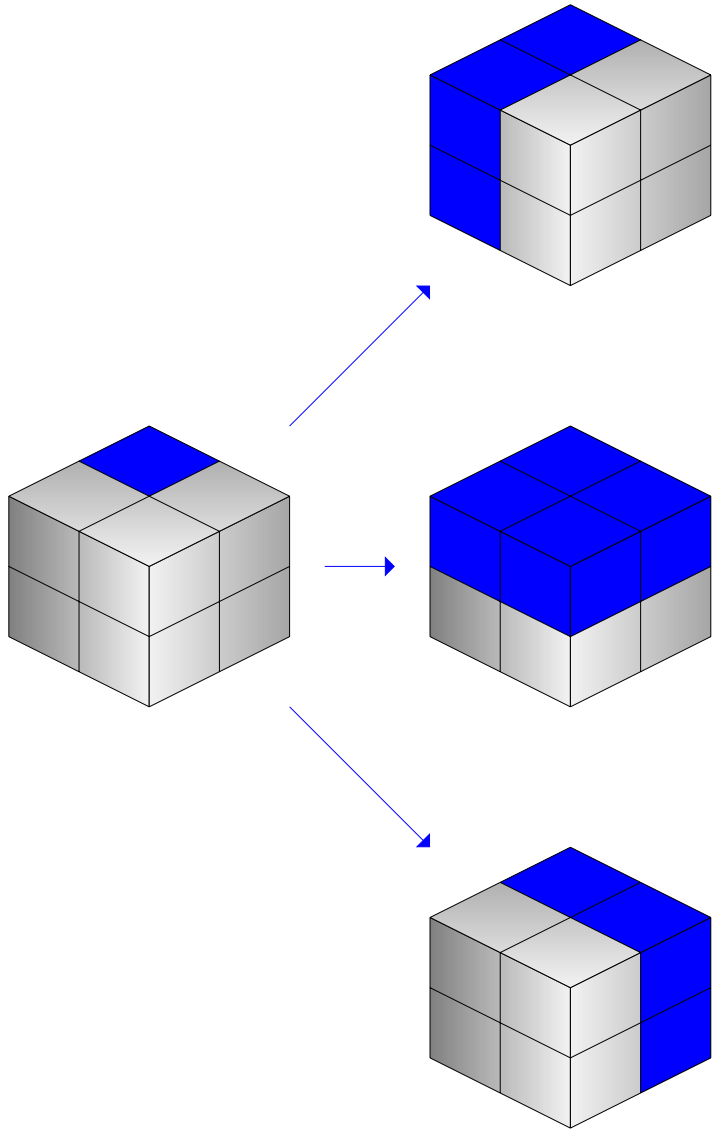
Schlau suchen

- Ich habe 8 Schachteln als Würfeln gestapelt.
- In eine Schachtel habe ich einen Schoko-Riegel versteckt
- Du kannst Ja/Nein fragen stellen, um die Schachtel zu identifizieren
- Was ist die geringste Anzahl Fragen, um zu garantieren, dass du den Schoko-Riegel kriegst?



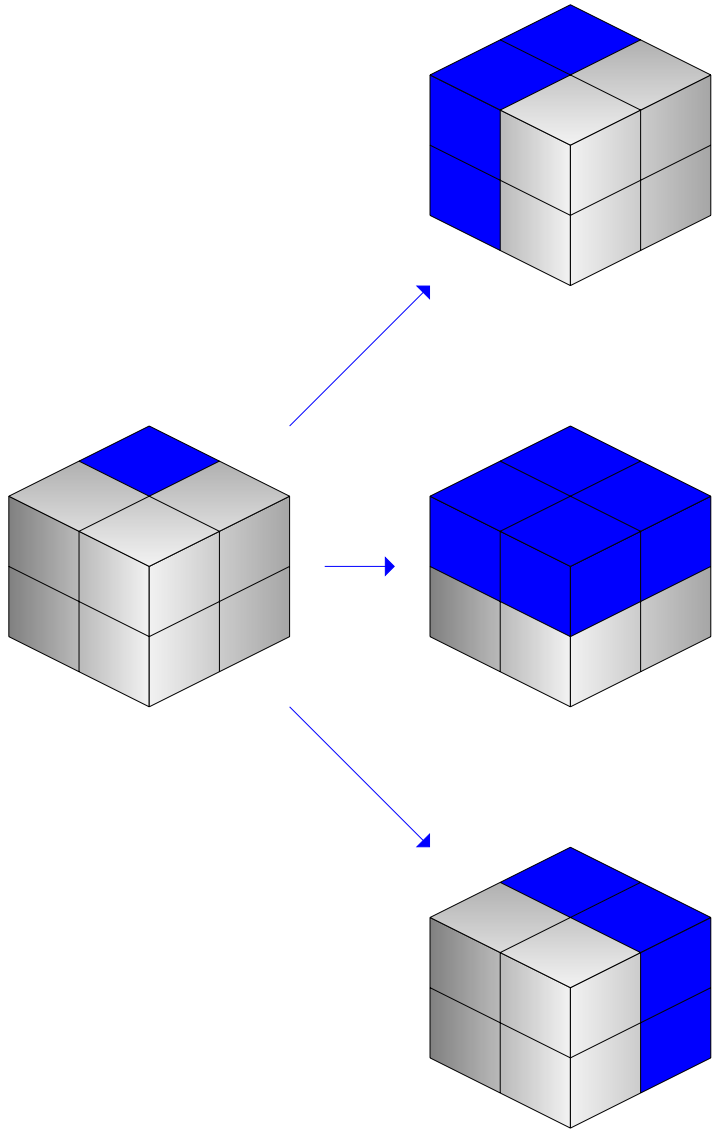
Dumme Frage ...

- Ist der Schokoriegel in Schachtel 1? in Schachtel 2? 3? 4? 5? 6? 7? 8?
- Mit etwas Pech musst du 8mal fragen, im Schnitt 4mal

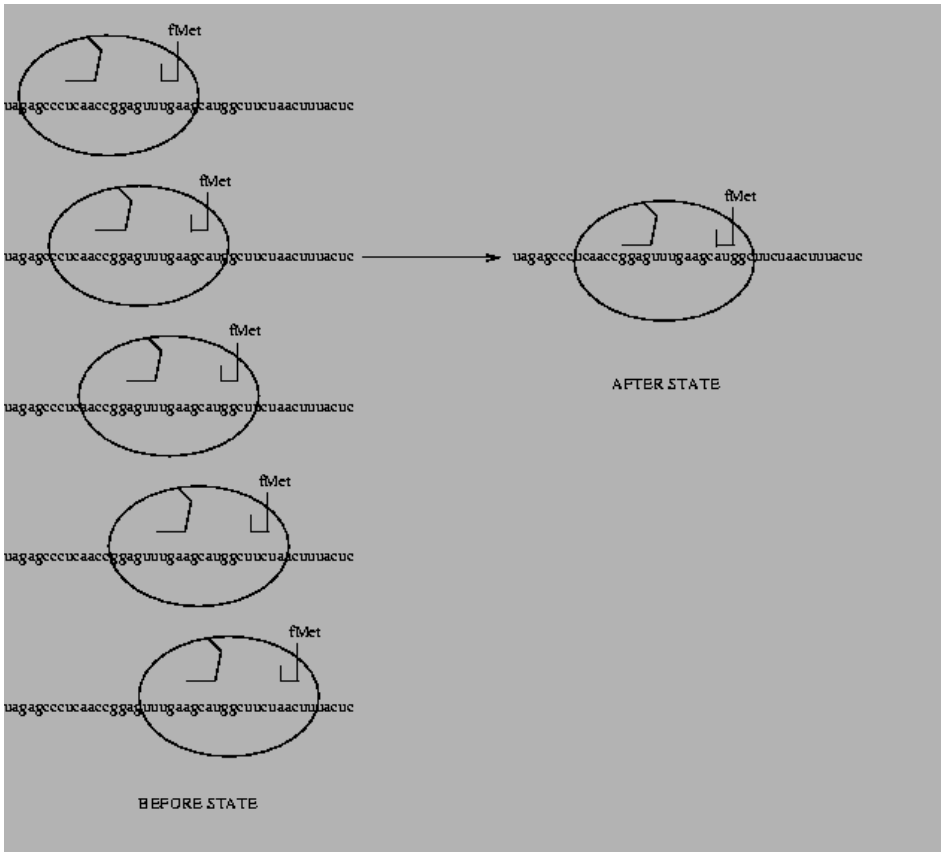


Schlaue Frage ...

- Ist der Schokoriegel auf der linken Seite? Unten? Vorne?
- Die Antwort auf diese Fragen halbiert die Möglichkeiten, vermittelt somit einen Bit Information
- Insgesamt müssen wir $\log_2 8 = 3$ Fragen stellen

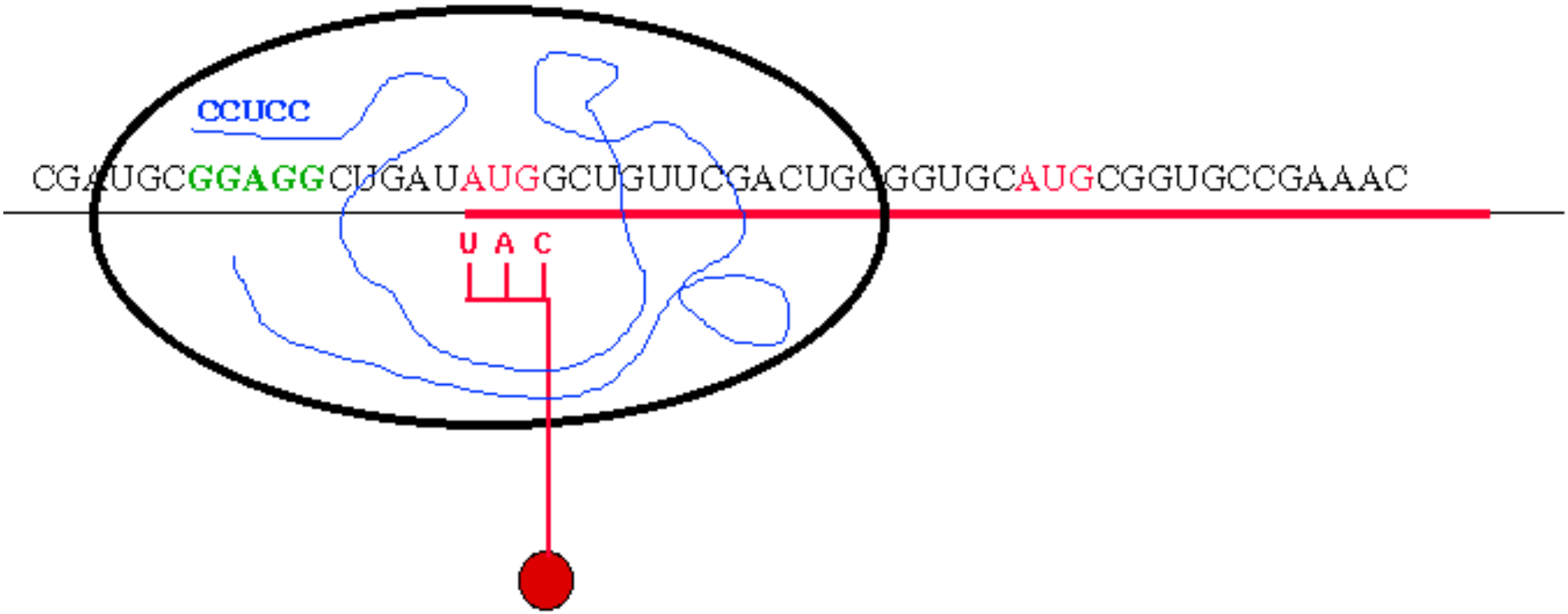


Abtasten ...



- Ein Ribosom tastet die Sequenzen in einem zufälligen Prozess von Brownscher Bewegung ab
- Findet es eine korrekte Sequenz, bindet es sie und initiiert die Translation

Abtasten ...



- Kleine ribosomale Einheit mit Initiator Met-tRNA bzw. Proteinfaktoren, welche die Ribosomen-Bindungssequenz und das AUG-Startkodon erkennen

Abtasten ...

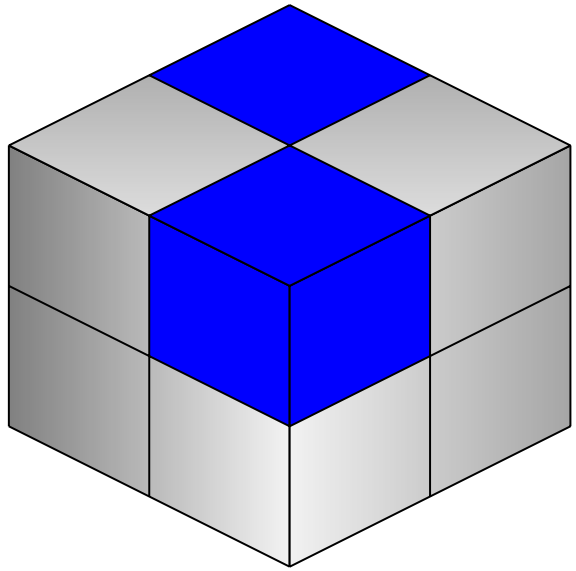
- Für den Schokoriegel brauchten wir $\log_2 8 = 3$ Fragen
- Hätten wir den Schokoriegel unter 16 Schachteln versteckt, wären $\log_2 16 = 4$ Fragen erforderlich
- Im Allgemeinen gilt für G Schachteln

$$B = \log_2 G \tag{1}$$

- d.h., $G = 2^B$, und wir brauchen B Bits für G Schachteln

Mehrere Riegel ...

- Falls 2 Schachteln einen Schokoriegel haben, kannst du einen Bit ignorieren, da du ihn nicht brauchst um *einen* Riegel zu finden
- Falls 4 Schachteln einen Schokoriegel haben, kannst du zwei Bits ignorieren...
- im Allgemeinen kann man für γ Schachteln mit Riegeln $\log_2 \gamma$ Bits ignorieren
- d.h., wir brauchen lediglich $\log_2 G - \log_2 \gamma$ Bits, um den Riegel zu finden



$R_{frequency}$

Wir definieren $\log_2 G - \log_2 \gamma$ als $R_{frequency}$

$$R_{frequency} = \log_2 G - \log_2 \gamma \tag{2}$$

Wir können $R_{frequency}$ auch wie folgt schreiben

$$R_{frequency} = \log_2 \frac{G}{\gamma} \tag{3}$$

- $\frac{G}{\gamma}$ stellt die Wahrscheinlichkeit dar, womit wir im Rahmen einer zufälligen Suche erwarten können, den gewünschten Riegel (bzw. eine Bindungssequenz für das Ribosom) zu finden

$R_{frequency}$

- Für die Ribosomen-Bindungssequenzen gilt $G = 4.7 \times 10^6$ und $\gamma = 2600$, so dass $R_{frequency} = 10.8$ Bits pro Bindungssequenz
- $R_{frequency} = 10.8$ ergab sich aus der Anzahl von Ribosomen-Bindungssequenzen und der Größe des Transkriptoms und stellt eine Vorhersage dar, wieviele Bits Information eine Bindungssequenz aufweisen muss.

Outline

Informationsgehalt

- Gegeben sei eine Zufallsvariable X, die einen von endlich vielen spezifischen Werten annehmen kann, $\{x_1, x_2, \ldots, x_n\}$
- Die entsprechenden Wahrscheinlichkeiten sind $\{p_1, p_2, \ldots, p_n\}$
- Die Wahrscheinlichkeit von x_i beträgt p_i und $\sum_{i=1}^{n} p_i = 1$
- Notation: $p(X = x)$ oder einfach $p(x)$ bezeichnet die Wahrscheinlichkeit, dass unsere Zufallsvariable X den Wert x annimmt
- Die verschiedenen Werte, die X annehmen kann, werden wir als die Ausgänge (Realisierungen) von X bezeichnen

Informationsgehalt

Shannon definierte den Informationsgehalt eines Ausgangs x als:

$$h(x) = \log_2 \frac{1}{p(x)} = -\log_2 p(x) \tag{4}$$

Entropie

Shannon definierte die Entropie der Zufallsvariablen X als den durchschnittlichen Informationsgehalt über alle möglichen Realisierungen von X:

$$H(X) = \sum_{i=1}^{n} p(x_i) \log_2 \frac{1}{p(x_i)} \tag{5}$$

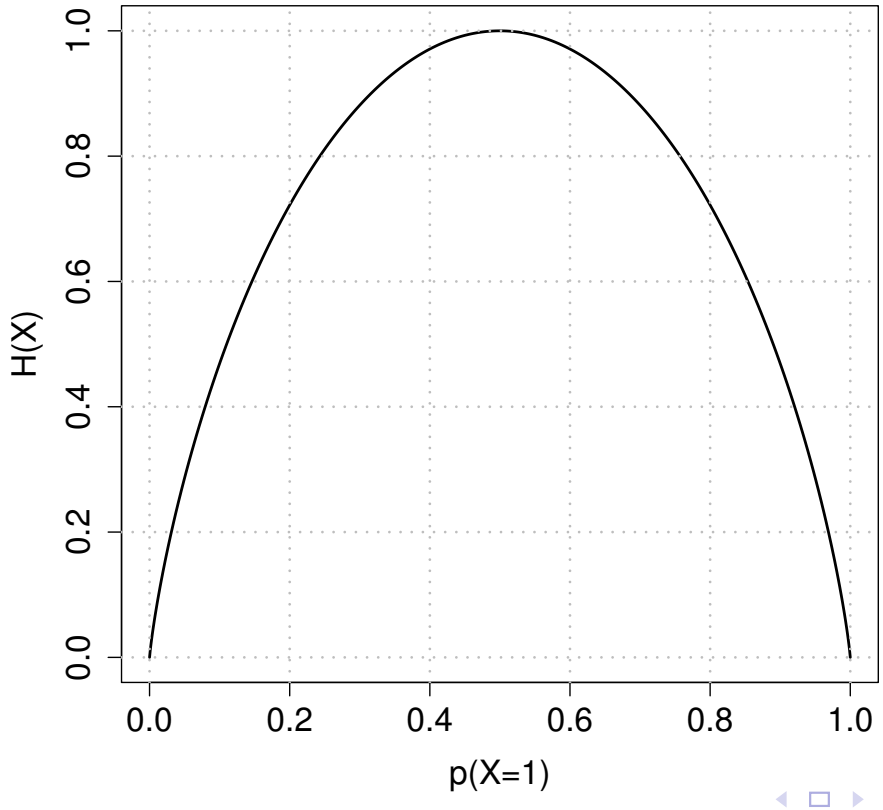
Entropie

- Entropie und Informationsgehalt werden in der Einheit “Bit” angegeben
- Die Entropie einer Zufallsvariablen X lässt sich als die “Ungewissheit” über den Ausgang von X interpretieren
- Falls eine Zufallsvariable X nur einen Ausgang hat (d.h., keine Ungewissheit), dann ist die Entropie von X Null

$$H(X) = p(x') \log_2 \frac{1}{p(x')} = 1 \times -1 \times \log_2 1 = 0$$

Entropie

- Es lässt sich einfach zeigen, dass die Entropie von X dann maximal ist, wenn maximale Ungewissheit über deren Ausgang herrscht – d.h., kein Ausgang ist wahrscheinlicher als ein anderer, d.h., X folgt einer uniformen Verteilung



Und was hat dies mit Bindungssequenzen zu tun?

	-3	-2	-1	+1	+2	+3	+4	+5	+6
A	36	56	11	0	0	60	69	7	14
C	36	13	3	0	0	3	9	6	14
G	18	14	78	100	0	35	12	82	19
T	11	17	8	0	100	3	11	5	53

Prozente für die vier Nukleotide an der 5' Spleißsequenz

Shapiro MB, Senapathy P (1987) RNA splice junctions of different classes of eukaryotes: sequences statistics and functional implications in gene expression.

Nucl Acids Res **15**:7155–7174

Und was hat dies mit Bindungssequenzen zu tun?

- Die Angabe, dass Position *i* von einer Bindungssequenz immer ein bestimmtes Nukleotid ist, erfordert 2 Bit Information (wir wählen eins von vier Dingen, z.B. U aus A,C,G,U
- Falls Position *i* zur Hälfte A und zur Hälfte G ist, entspricht dies eine Selektion von 2 aus 4 Nukleotiden, also 1 Bit
- Bevor das Ribosom bindet, “sieht” es 4 Basen und hat sozusagen eine Ungewissheit von $\log_2 4 = 2$ Bits
- Nach der Bindungsreaktion verringert sich die Ungewissheit: kann an der Position *i* nur ein Nukleotid gebunden werden, beträgt die Restungewissheit $\log_2 1 = 0$ Bit. Falls das Ribosom an der Stelle eines von zwei Nukleotiden zulässt, beträgt die Restungewissheit $\log_2 2 = 1$ Bits

Und was hat dies mit Bindungssequenzen zu tun?

- Im Allgemeinen beträgt die Entropie nach Bindung an Position ℓ

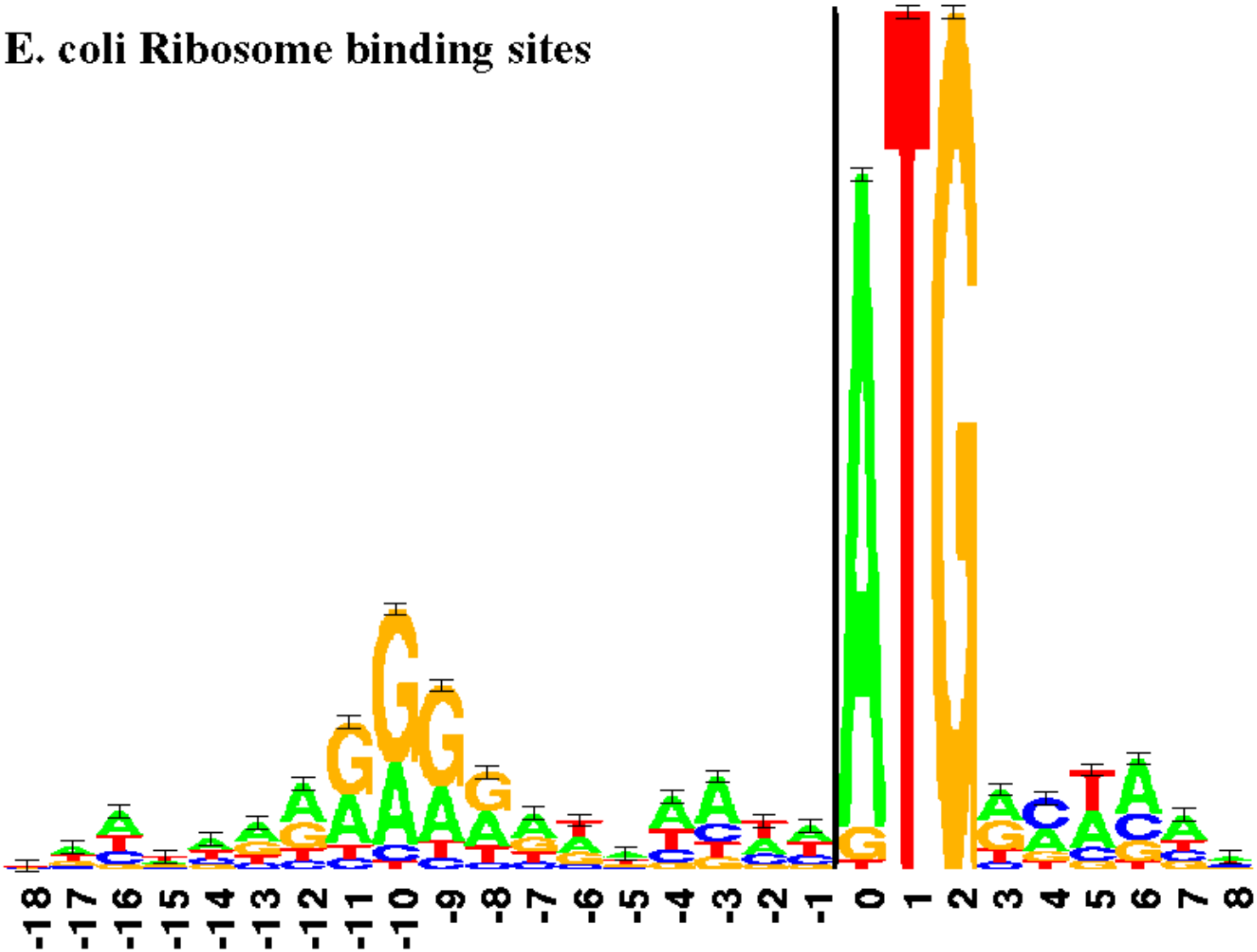
$$H(\ell) = -\sum_{b \in \{a,c,g,t\}} f(b,\ell) \log f(b,\ell) \tag{6}$$

- Wir können hiermit die Reduktion an Ungewissheit an jeder Position der Bindungsmatrix berechnen als

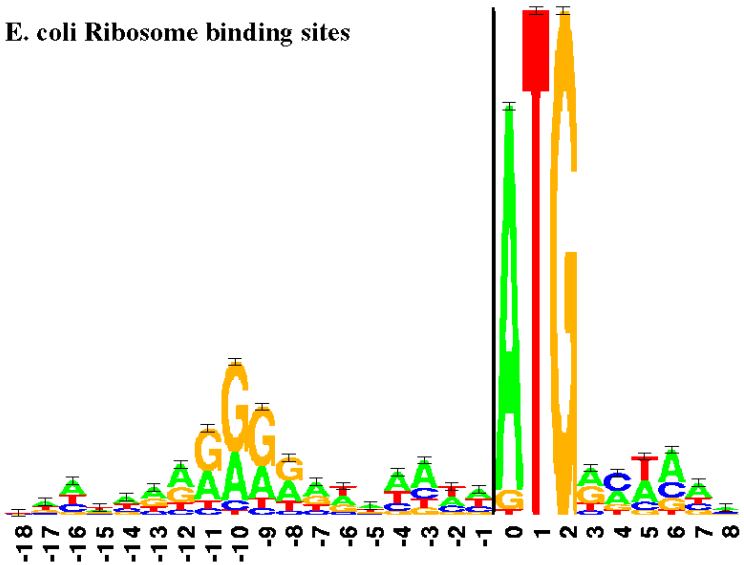
$$R_{sequence}(\ell) = 2 - H(\ell) \tag{7}$$

- (Der Einfachkeit halber nehmen wir eine uniforme Hintergrund-Verteilung der vier Nukleotide im Genom an. Genauer ist $R_{sequence}(\ell) = H(n) - H(\ell)$, wo $H(n)$ die Entropie der Hintergrund-Verteilung darstellt).

Sequenzlogo-E-coli Ribosomen-Bindungssequenz



Sequenzlogo-E-coli Ribosomen-Bindungssequenz



- Die Höhe an Position ℓ entspricht $R_{sequence}(\ell) = 2 - H(\ell)$
- Die relative Höhe der Buchstaben entspricht der Häufigkeit der entsprechenden Nukleotide an der Position

$R_{sequence}$

- Unter der Annahme, dass die verschiedenen Position unabhängig sind, können wir nun die Information der Bindungssequenz als Summe der Information an den einzelnen Positionen berechnen

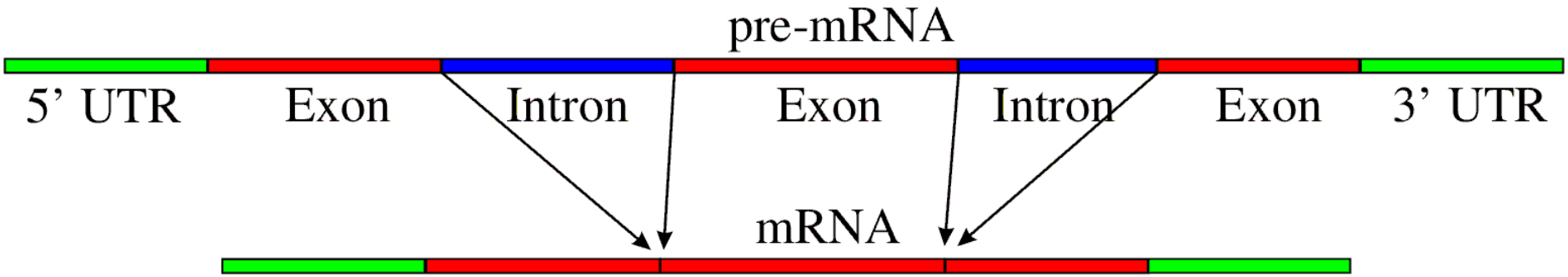
$$R_{sequence} = \sum_i R_{sequence}(i) \tag{8}$$

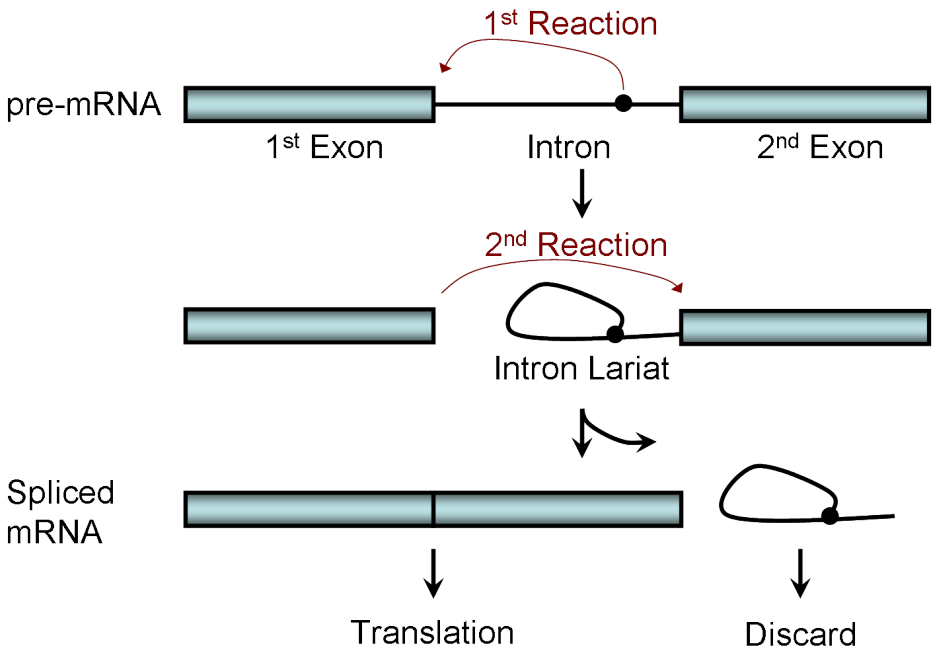
- Für die Ribosomen-Bindungssequenz erhalten wir $R_{sequence} = 11.0$: Fast denselben Wert wie für $R_{frequency}$!
- Intuitiv: es gibt gerade genug Information in den Ribosomen-Bindungssequenzen ($R_{sequence}$), damit sie innerhalb des Transkriptoms der Zelle gefunden werden können ($R_{frequency}$).
- Ähnliche Beobachtungen können für viele andere (aber nicht alle!) DNA- oder RNA-Bindungsproteine gemacht werden

Schneider TD (1994) Sequence logos, machine/channel capacity, Maxwell's demon, and molecular computers: a review of the theory of molecular machines. *Nanotechnology* **5**:1–18.

Outline

RNA-Spleißen

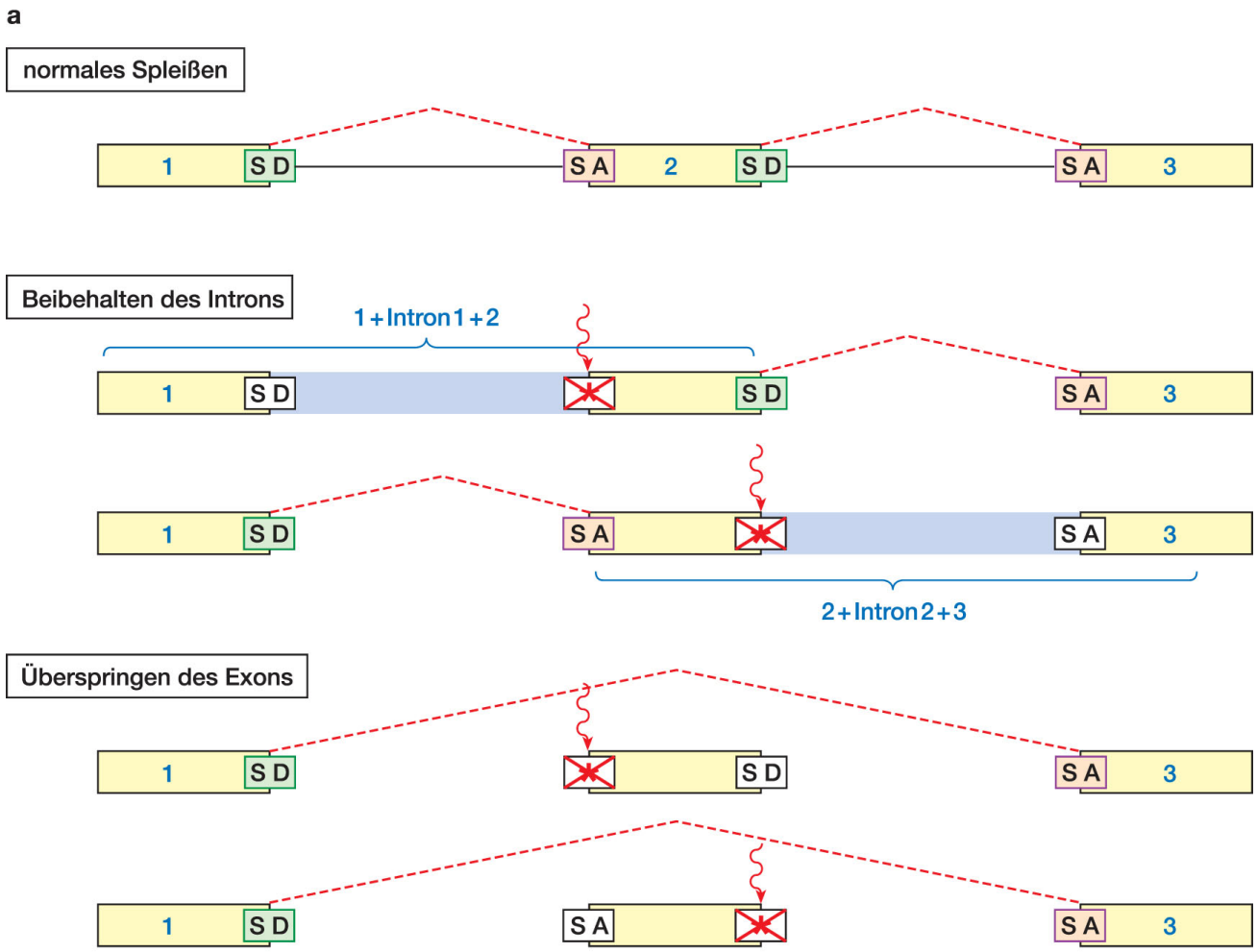




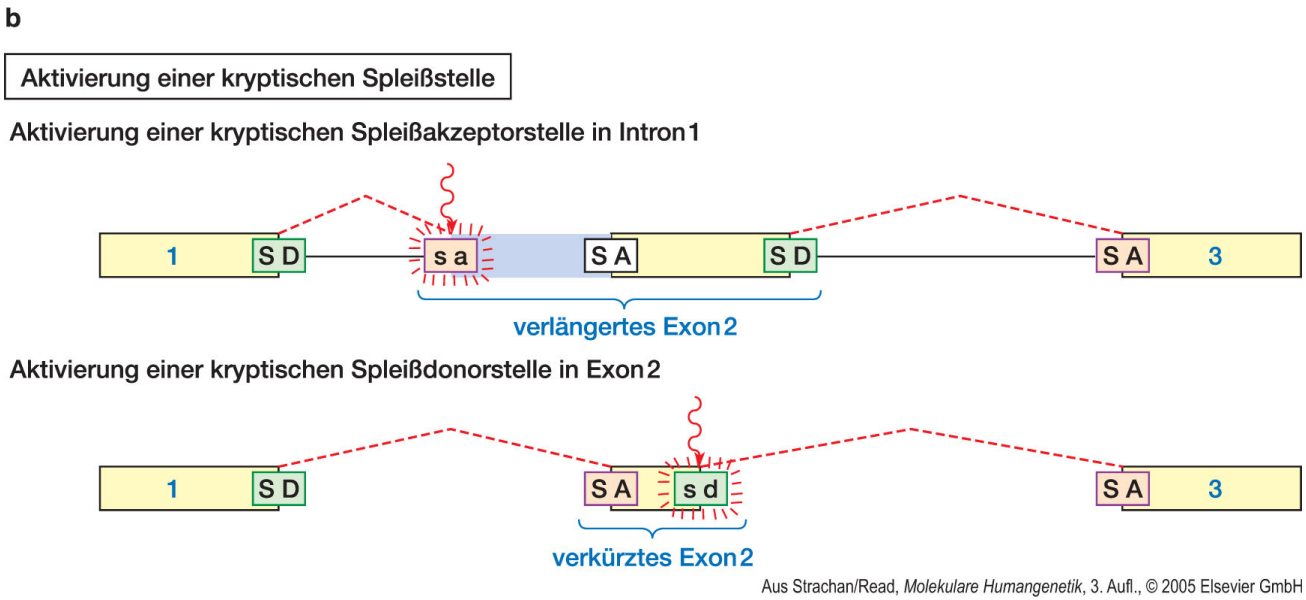
Wikipedia

- GT-AG-Regel

Exon-Skipping vs. Intronretention

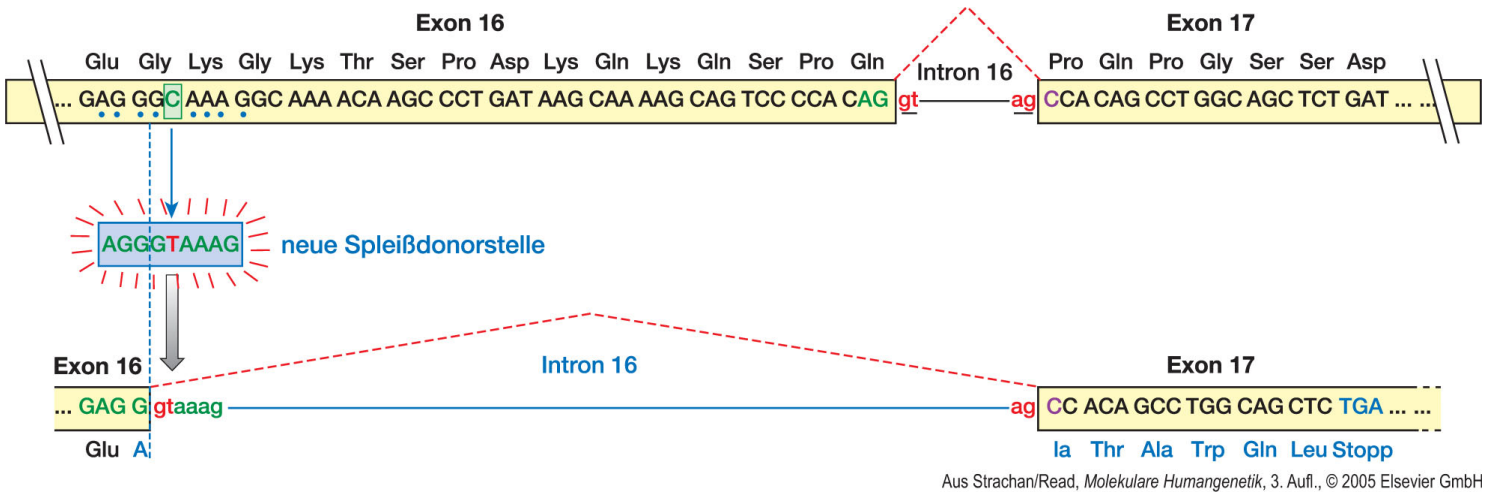


Aktivierung einer kryptischen Spleißstelle



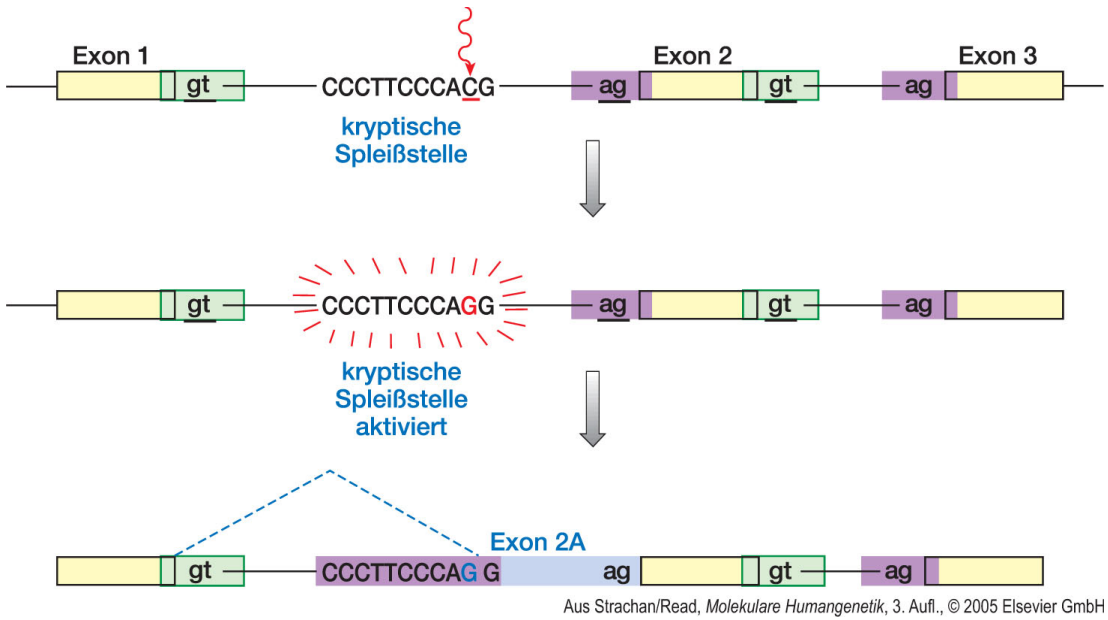
Aus Strachan/Read, *Molekulare Humangenetik*, 3. Aufl., © 2005 Elsevier GmbH

Beispiel: Aktivierung einer kryptischen Spleißstelle



- ”Stille” Mutation: GGC = Gly, GGC = Gly
- Durch die Mutation wird jedoch eine kryptische Spleißstelle aktiviert, die Spleißreaktion verläuft fehlerhaft
- Eine der bei der Gliedergürtel-Muskeldystrophie Typ 2A nachgewiesenen Mutationen

Beispiel: Aktivierung einer kryptischen Spleißstelle



Aus Strachan/Read, *Molekulare Humangenetik*, 3. Aufl., © 2005 Elsevier GmbH

- Aktivierung einer kryptischen Spleißstelle innerhalb eines Introns

R_i: Information einzelner Sequenzen

- Die individuelle Information R_i (in Bit) einer bestimmten Spleißsequenz i kann als das Skalarprodukt zwischen der Sequenz und der Gewichtsmatrix berechnet werden
- R_i entspricht der freien Energie der Bindung
- $R_{sequence}$: Durchschnittliche für das Spleißen erforderliche Information
- $R_{sequence} = 7.92 \pm 0.09$ Bit für 10 bp lange Spleißsequenzen
- Starke Spleißsequenzen haben $R_i >> R_{sequence}$, schwache Sequenzen $R_i < R_{sequence}$
- Idee: Mutationen der Spleißsequenz reduzieren R_i

R_i: Information einzelner Sequenzen

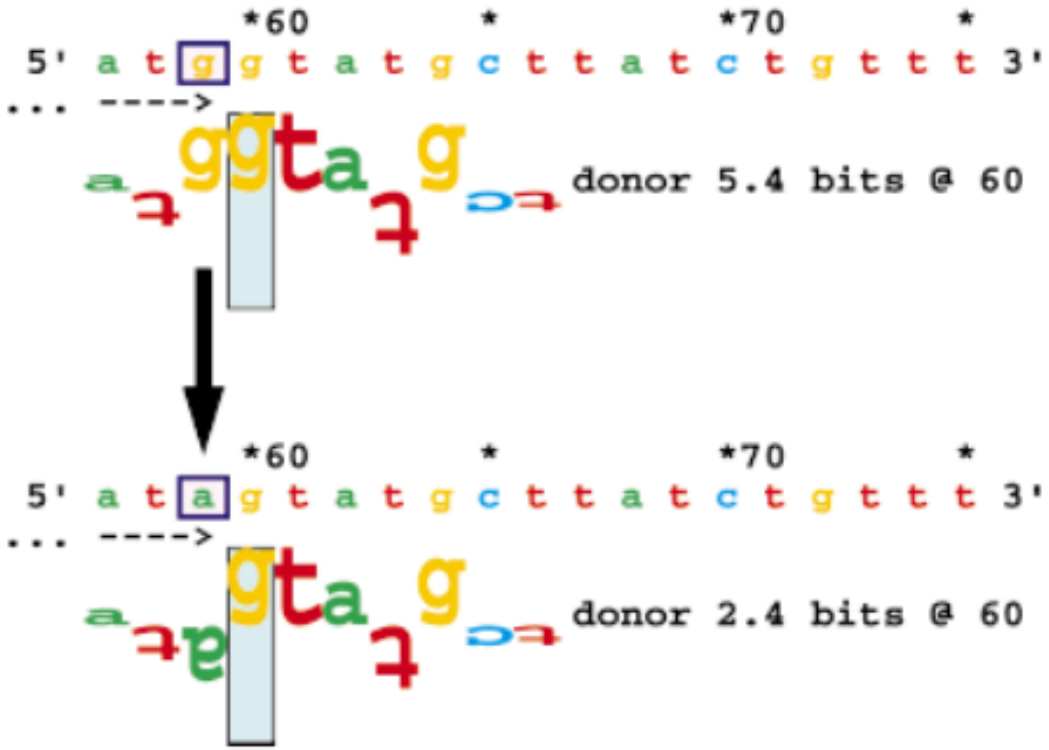
- Die Gewichtsmatrix für die Spleißsequenz berechnet sich als $R_{iw} = 2 - (-\log_2 f(b, \ell))$, wo $f(b, \ell)$ die Frequenz von Base b an Position ℓ angibt.

	-3	-2	-1	+1	+2	+3	+4	+5	+6
A	0.36	0.56	0.11	0.0	0.0	0.60	0.69	0.7	0.14
C	0.36	0.13	0.3	0.0	0.0	0.3	0.9	0.6	0.14
G	0.18	0.14	0.78	1.0	0.0	0.35	0.12	0.82	0.19
T	0.11	0.17	0.8	0.0	1.0	0.3	0.11	0.5	0.53

- $s(b, \ell, j)$: 1 für Base b an Position ℓ, sonst 0

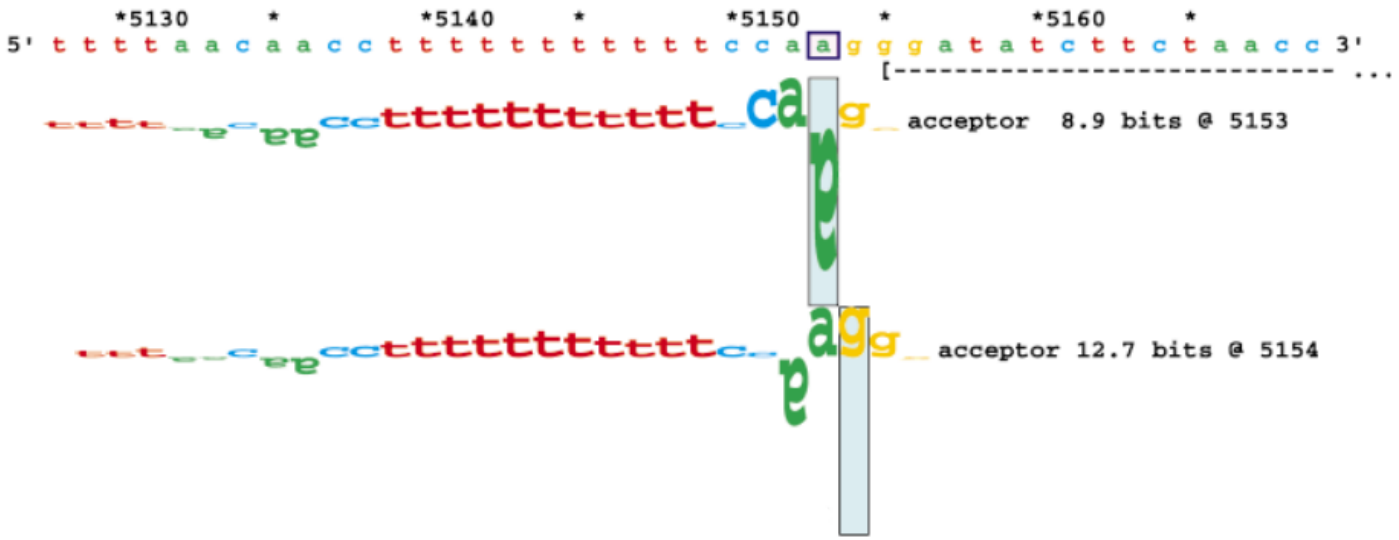
$$R_i(j) = \sum_{\ell} \sum_{b \in \{a,c,g,t\}} s(b, \ell, j) R_{iw}(b, \ell) \tag{9}$$

R_i und Mutationen



- G>A Mutation an Position +1 der Donorsequenz von Exon 6 des *COL1A2*-Gens
- Reduktion von R_i von 5.4 auf 2.4, Spleißdefekt experimentell nachgewiesen

R_i und Mutationen



- A>G Mutation an Position -2 der Acceptorsequenz von Intron 3 des *IDS*-Gens
- Reduktion des R_i-Wertes der normalen Acceptorsequenz bei gleichzeitiger Erzeugung einer neuen Akzeptorsequenz ein Nukleotid stromaufwärts

R_i und Mutationen

- Durchschnittliche Reduktion von R_i bei Spleißmutationen, welche zur Beeinträchtigung des Spleißens führten: $\Delta\overline{R_i} = -7.67 \pm 3.95$ Bit (Donor) bzw. $\Delta\overline{R_i} = -5.97 \pm 3.50$ Bit (Akzeptor)
- Kann verwendet werden, um die Auswirkung einer uncharakterisierten Sequenzvariante vorherzusagen

Rogan PK, Faux BM, Schneider TD (1998) Information analysis of human splice site mutations. *Hum Mutat* **12**:153–171.

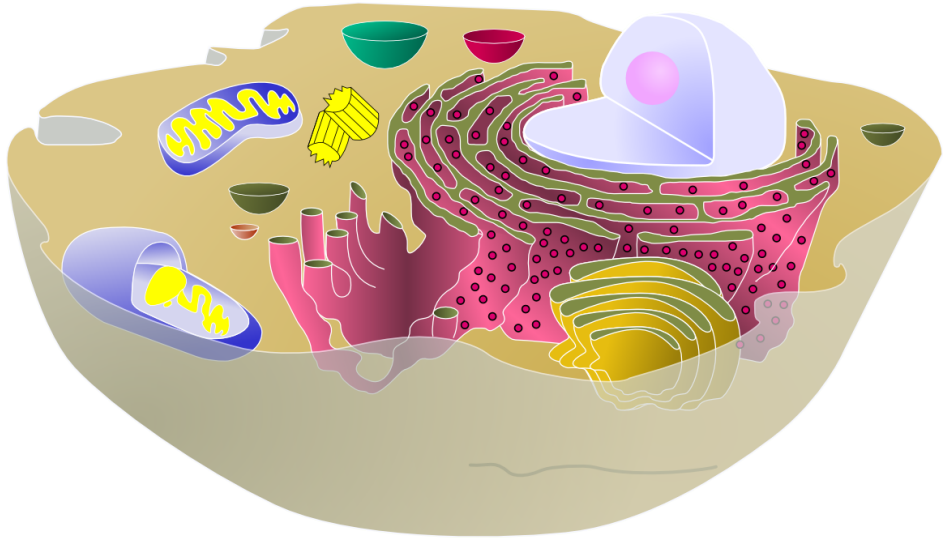
Outline

Neue Motive entdecken

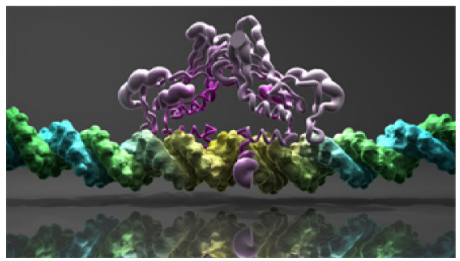
- Man kann mit verschiedenen Methoden DNA-Segmenten identifizieren, welche Bindungssequenzen für einen bestimmten Transkriptionsfaktor enthalten, ohne dass man die genaue Lokalisation der Bindungssequenzen innerhalb der Segmente kennt

ChIPSEQ: Method (1)

- DNA-bindende Proteine werden mit Formaldeyd an die DNA querverlinkt
- Das Chromatin wird isoliert und fragmentiert



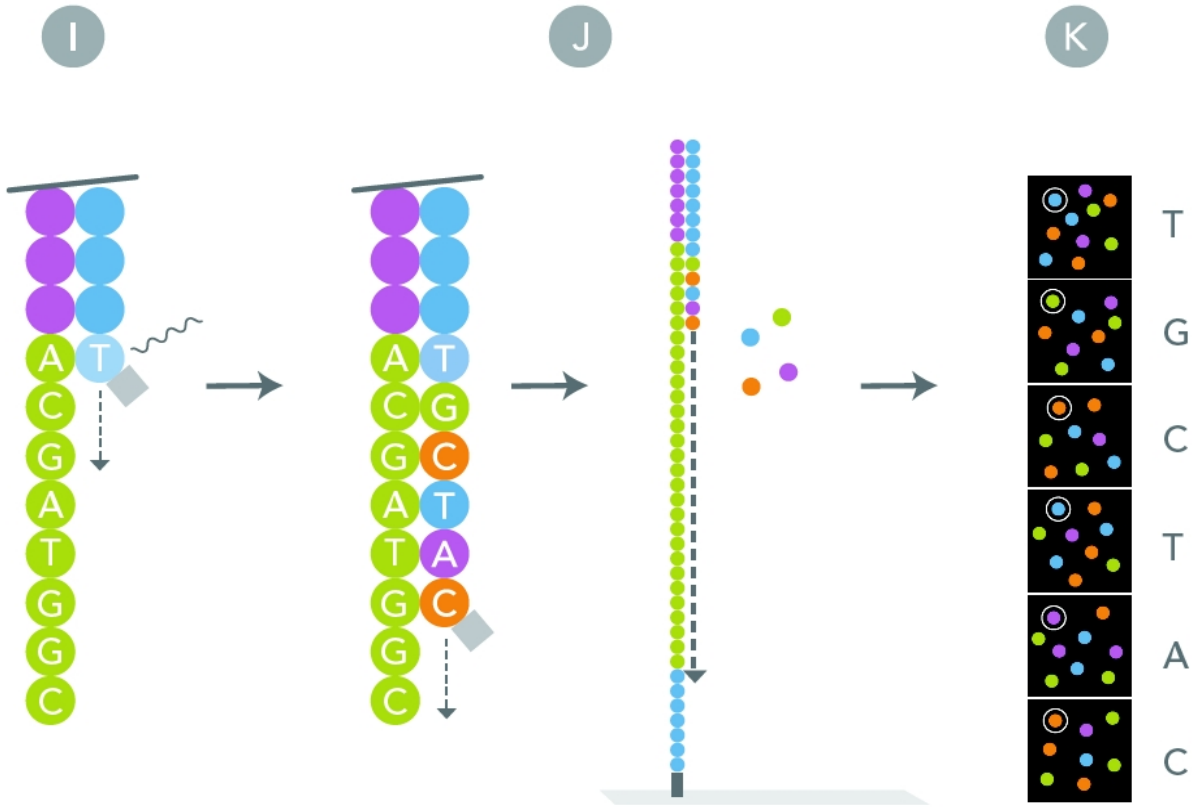
Crosslink proteins to DNA and lyse cells



ChIPSEQ: Method (2)

- Mittels eines spezifischen Antikörpers wird das DNA-bindende Protein und die daran verlinkten DNA-Fragmente angereichert (Immunpräzipitation)
- DNA wird fregesetzt und die Proteine verdaut

Solexa: Sequenzierung



I Erste Base verlängern, ablesen, Block entfernen

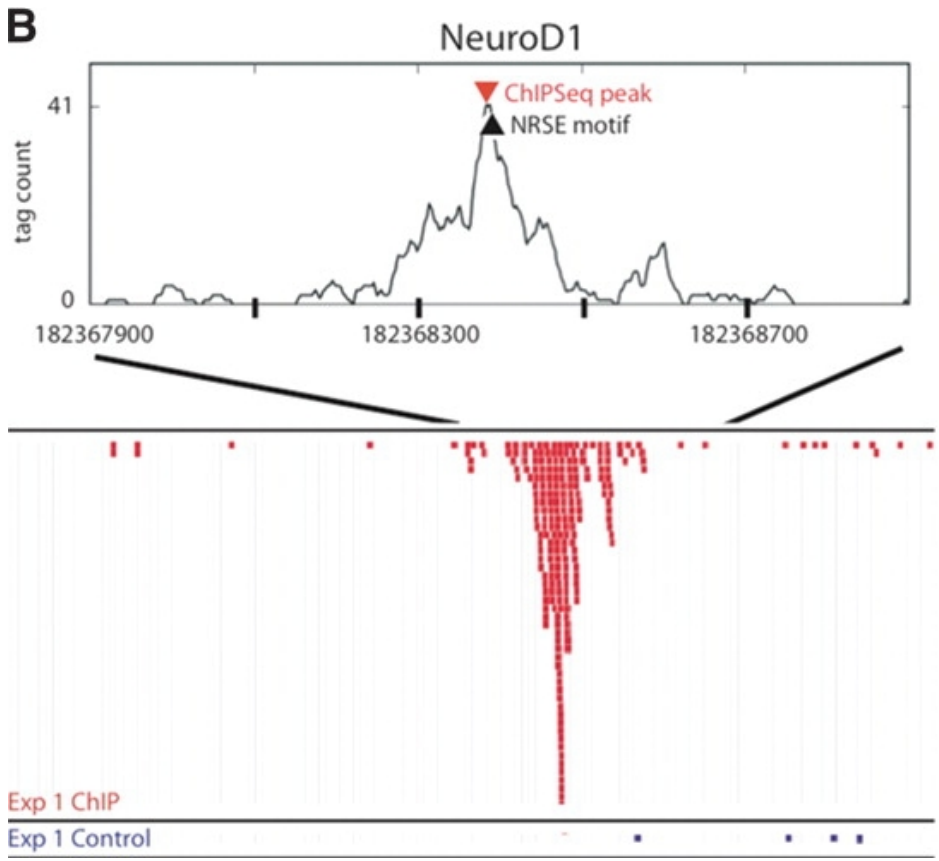
J Wiederholen!

K Basen bestimmen (“calls”)

! The fun begins

Identifying clusters of reads

- Cluster von Sequenz-Reads zeigen die (wahrsscheinlichen) Transkriptionsfaktorbindungssequenzen (TFBS) an
- Ein typischer Versuch ergibt Hunderte oder Tausende solcher Cluster
- Ziel: Das Bindungsmotiv des Untersuchten Transkriptionsfaktor identifizieren



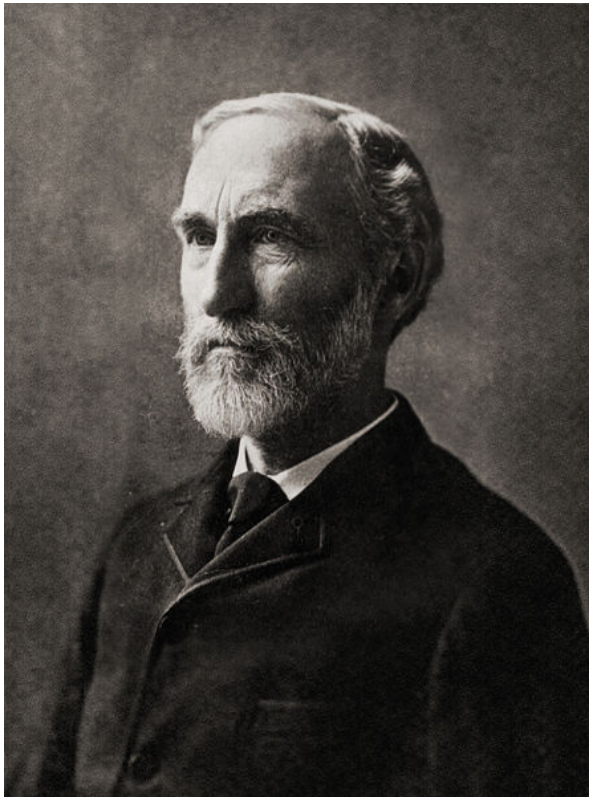
Entdeckung von TFBS

- Zahlreiche Algorithmen, ein ziemlich aktives Forschungsgebiet in der Bioinformatik
- Wir wollen heute die Methode “Gibb’s Sampling” unter die Lupe nehmen

Lawrence CE, Altschul SF, Boguski MS, Liu JS, Neuwald AF, Wootton JC (1993) Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment. *Science* **262**:208-14.

Gibb's Sampling

- Wir suchen kurze Sequenzmotive ohne Lücken (gaps)
- Ein Muster wird als PSSM (position specific scoring/weight matrix) beschrieben
- Die Daten: N Sequenzen $S_1, S_2, \ldots, S_n$
- Angabe über die Länge L des gesuchten Motivs



Josiah Willard Gibbs (1839–1903)

Gibb's Sampling

Schritt 1

- Von den n Sequenzen wird eine Sequenz z per Zufall ausgesucht
- Berechne die PFM[1] $q_{i,j}$ und die Hintergrundfrequenzen p_j aus den aktuellen Positionen a_k in allen Sequenzen außer z (a_k stellt also die Startposition des Motivs in Sequenz k dar)
- Für die i^{te} Position des Motivs haben wir $n-1$ beobachtete Nukleotide; wir bezeichnen die Anzahl von Nukleotid j an Position i als $c_{i,j}$

$$q_{i,j} = \frac{c_{i,j}}{n-1} \tag{10}$$

[1] PFM = *Position Frequency Matrix*.

Gibb's Sampling

Schritt 1: Pseudocounts

- Basenzahlen bei 10 beobachteten Sequenzen
- Sind wir wirklich sicher, dass an der zweiten Position grundsätzlich kein A erscheinen darf?

Pos	1	2	3	4	5
A	7	0	0	5	0
C	0	2	2	1	8
G	2	0	0	1	1
T	1	8	8	3	1

Gibb's Sampling

Schritt 1: Pseudocounts

- Um Artefakte, die sich aus der kleinen Stichprobe ergeben, zu vermeiden, fügen wir +1 zu jeder Zelle hinzu (Pseudocount)

Pos	1	2	3	4	5
A	8	1	1	6	1
C	1	3	3	2	9
G	3	1	1	2	2
T	2	9	9	4	2

- Dies ergibt die Positionsfrequenzmatrix (PFM):

Pos	1	2	3	4	5
A	0.57	0.07	0.07	0.43	0.07
C	0.07	0.21	0.21	0.14	0.64
G	0.21	0.07	0.07	0.14	0.14
T	0.14	0.64	0.64	0.29	0.14

Gibb's Sampling

Schritt 1: Pseudocounts

- Das heißt, wir berechnen die positionsspezifische Basenfrequenz nicht als

$$q_{i,j} = \frac{c_{i,j}}{N-1} \tag{11}$$

- Sondern (wo $B = \sum_i b_i$, die Summe der Pseudocounts, im Falle von Nukleotiden 4)

$$q_{i,j} = \frac{c_{i,j} + b_i}{N-1+B} \tag{12}$$

- Die Hintergrundfrequenzen p_i werden über die nicht-PFM-Positionen analog berechnet

Gibb's Sampling

Schritt 2: Sampling

- Mit der aktuellen PFM wird die Wahrscheinlichkeit des Motifs an jeder möglichen Position von der herausgelassenen Sequenz z berechnet $(1, 2, \ldots, M - L + 1$ für Sequenzlänge M und Motiflänge $L)$: P_x
- Ähnlich wird die Wahrscheinlichkeit berechnet, dass jedes Segment $(1, 2, \ldots, M - L + 1)$ durch die Hintergrundsverteilung generiert wurde: Q_x
- So können wir jedem Segment ein Gewicht $A_x = \log \frac{P_X}{Q_x}$ zuordnen
- Ein Segment wird gemäß den A_x per Zufall ausgewählt und wird das a_z für Sequenz z für die nächste Iteration

Gibb's Sampling

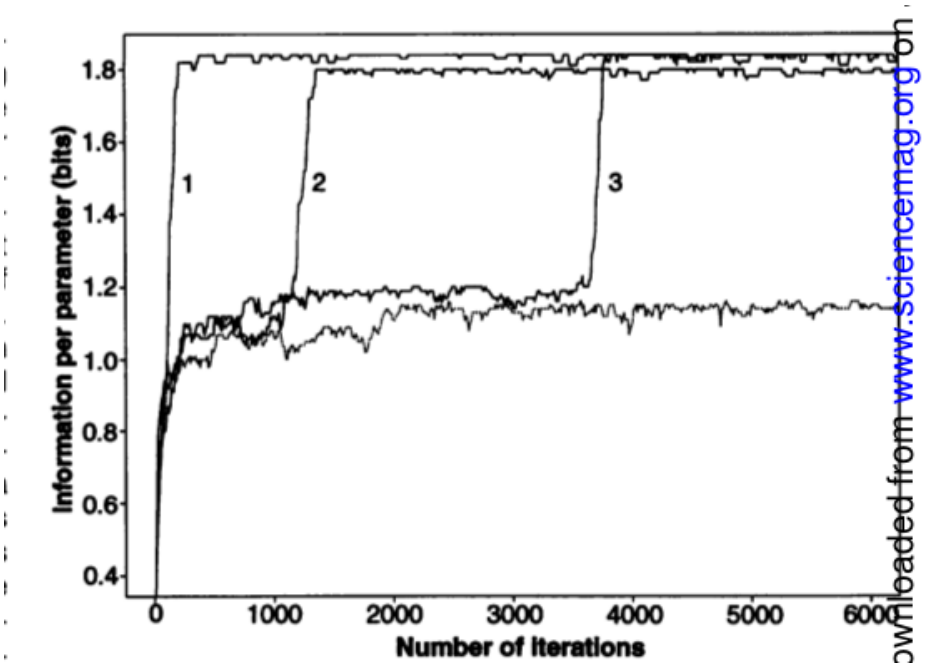
Schritt 2: Sampling

- Die Iteration wird fortgesetzt mit immer neuen herausgelassenen Sequenzen:
  1. berechne $q_{i,j}$ mit $n-1$ Sequenzen
  2. bestimme a_z für die ausgelassene Sequenz z
- Je akkurater die Werte $q_{i,j}$ für die PFM sind, desto besser können die a_k bestimmt werden, was wiederum zu noch besseren Werten $q_{i,j}$ führt.
- Der Algorithmus konvergiert nicht, aber der Gesamtscore der PFM[2] verbessert sich typischerweise von Iteration zu Iteration bis ein Plateau errreicht wird

[2] der Score kann als $\sum R_i$ berechnet werden

Gibb's Sampling

Konvergenzverhalten



- Drei "runs" zeigen unterschiedliche Konvergenzverhalten (Der Algorithmus ist nicht deterministisch)

Lawrence CE, Altschul SF, Boguski MS, Liu JS, Neuwald AF, Wootton JC (1993) Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment. *Science* **262**:208-14.

Gibb's Sampling

Input: A set of DNA sequences $S_1, S_2, \ldots, S_N$ and the motif width W
Output: The starting position a_k of the motif in each sequence S_k;
A PSSM $\mathcal{Q} = [q_{i,j}]$ for the putative motif model

begin
Initialization:
Randomly select a position a_k for the motif in each sequence S_k
Estimate the background base frequencies p_j, for j from 1 to J, to obtain $\mathcal{P}$
Repeat until convergence:
Predictive update step:
Randomly select a sequence S_z from the input sequences
Take the set of putative binding sites $\{S_k[a_k, a_k + W - 1] \mid 1 \leq k \leq N, k \neq z\}$
Estimate the PSSM $\mathcal{Q}$ from $\{S_k[a_k, a_k + W - 1] \mid 1 \leq k \leq N, k \neq z\}$
Sampling step:
Estimate $P(S_z[n, n + W - 1] \mid \mathcal{Q})$ for every position n in sequence S_z
Estimate $P(S_z[n, n + W - 1] \mid \mathcal{P})$ for every position n in sequence S_z
Randomly select a new position a_z in S_z according to $L(S_z[n, n + W - 1] \mid \mathcal{P}, \mathcal{Q})$
end

zum Schluss

- Email: peter.robinson@charite.de

weiterführende Literatur

- Schneider TD (1994) Sequence logos, machine/channel capacity, Maxwell's demon, and molecular computers: a review of the theory of molecular machines. *Nanotechnology* **5**:1–18.
- Lawrence CE, Altschul SF, Boguski MS, Liu JS, Neuwald AF, Wootton JC (1993) Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment. *Science* **262**:208-14.
- Stormo GD (2010) Motif discovery using expectation maximization and Gibbs' sampling. *Methods Mol Biol* **674**:85-95.