

# Molekulare Bioinformatik

Wintersemester 2011/2012

Martinetz

Institut für Neuro- und Bioinformatik  
Universität zu Luebeck

20.12.2011





# Shannon Entropie als Maß für Unkenntnis

$$H(X) = H[P(X)] = - \sum_{i=1}^N P(X = x_i) \log P(X = x_i)$$

bezeichnet man als die (Shannon)-Entropie der Zufallsgröße  $X$  bzw. des Set von Wahrscheinlichkeiten  $P(X = x_i)$ .

Es gilt  $H(X) \geq 0$ , und es gilt  $H(X) = 0$  für den Fall  $P(x_{i^*}) = 1$ .

$H(X)$  (Unkenntnis) wird maximal, falls jeder Zustand als gleichwahrscheinlich angenommen werden muss, also  $P(X = x_i) = 1/N$  mit  $N$  Anzahl der Zustände.

Mit Anzahl  $N$  möglicher Zustände steigt maximale Unkenntnis  $H(1/N)$  monoton an.



# Maximale Shannon Entropie

Welche  $P(X = x_i) = P_i$  maximieren  $H(X) = -\sum_{i=1}^N P_i \log P_i$   
 unter der Nebenbedingung  $\sum_{i=1}^N P_i = 1$ ?

Lagrange: Finde  $P_i$ , die

$$L = -\sum_{i=1}^N P_i \log P_i + \beta \left( 1 - \sum_{i=1}^N P_i \right)$$

maximieren, mit  $\beta$  als Lagrange-Parameter.



# Shannon Entropie

Die partielle Ableitung nach einem  $P_j$  liefert

$$\frac{\partial L}{\partial P_j} = -\log P_j - 1 - \beta. \quad (1)$$

und wird Null für  $P_j = e^{-1-\beta}$ .

Alle  $P_i$  haben gleichen Wert. Daher  $P_i = 1/N$ , da Summe 1.

Wie erwünscht wird Shannon-Entropie maximal, wenn alle Zustände gleich wahrscheinlich.



# Shannon Entropie für unabhängige Subsysteme

Besteht das System aus zwei Subsystemen, deren Zustände  $X, X'$  voneinander unabhängig sind, so ist die Wahrscheinlichkeit, dass System im Zustand  $x_i, x'_{i'}$  zu finden,  $P(X = x_i, X' = x'_{i'}) = P_i P_{i'}$ .

Unsere Unkenntnis ist dann

$$\begin{aligned} H(X, X') &= - \sum_{i=1}^N \sum_{i'=1}^{N'} P_i P_{i'} \log(P_i P_{i'}) \\ &= - \sum_{i=1}^N \sum_{i'=1}^{N'} P_i P_{i'} \log P_i - \sum_{i=1}^N \sum_{i'=1}^{N'} P_i P_{i'} \log P_{i'} \\ &= - \sum_{i=1}^N P_i \log P_i - \sum_{i'=1}^{N'} P_{i'} \log P_{i'} \\ &= H(X) + H(X') \end{aligned}$$



## Beispiel: Informationsgehalt eines Buches

Annahme: 2000 Zeichen pro Seite und 200 Seiten. 30 verschiedene Zeichen (mit Umlauten).

Ohne Vorwissen müssen die Zeichen als unabhängig angenommen werden. Dann ist die Unsicherheit (Entropie) die Summe über Unsicherheit an jeder Zeichenposition.

Unsicherheit:

$$H(X) = -400.000 \sum_{i=1}^{30} \frac{1}{30} \log_2 \frac{1}{30} = 2 \times 10^6 \text{ bit} = 250 \text{ kByte}.$$

Wenn wir das Buch schwarz auf weiss gezeigt bekommen, sinkt Unsicherheit auf Null. Der Informationsgewinn ist 250 kByte.



## Beispiel: Informationsgehalt eines Genoms

Annahme:  $3 \times 10^9$  bps. Jede der vier Basen gleich wahrscheinlich.

$$H(X) = -3 \times 10^9 \sum_{i=1}^4 \frac{1}{4} \log_2 \frac{1}{4} = 6 \times 10^9 \text{ bit} \approx 1 \text{ GigaByte}.$$

Informationsgewinn durch Sequenzierung: 1 *GigaByte*.

Informationsgewinn durch Hinweis, dass menschliches Genom?

Menschliches Genom hat  $15 \times 10^6$  SNPs. Dann Unsicherheit

$$H(X) = -15 \times 10^6 \sum_{i=1}^2 \frac{1}{2} \log_2 \frac{1}{2} = 15 \times 10^6 \text{ bit} \approx 2 \text{ MegaByte}$$

Informationsgehalt des Hinweis:  $(6 \times 10^9 - 15 \times 10^6) \text{ bit}$ .



# Sequence Motifs und Sequence Logos

Sequence motifs: kurze, wiederkehrende Sequenzen auf der DNA mit eventueller biologischer Funktion.

Häufig sequenzspezifische Bindungsstellen, z.B. für Proteine wie Nukleasen oder Transkriptionsfaktoren.

Wichtig für Analyse genetischer regulatorischer Netzwerke.

Ein bestimmter Transkriptionsfaktor bindet an eine bestimmte Menge von Sequenzabschnitten.

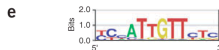
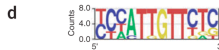
Wieviel Information tragen diese Sequenzabschnitte?

# Sequence Motifs und Sequence Logos

**a** HEM13 CCCATTGTTCTC  
 HEM13 TTTCTGGTTCTC  
 HEM13 TCAATTGTTTAG  
 ANB1 CTCATTGTTGTC  
 ANB1 TCCATTGTTCTC  
 ANB1 CCTATTGTTCTC  
 ANB1 TCCATTGTTTCG  
 ROX1 CCAATTGTTTTG

**b** YCHATTGTTCTC

**c** A 002700000010  
 C 464100000505  
 G 000001800112  
 T 422087088261



Menge der Sequenzen, an denen der Rox1 Transkriptionsfaktor auf dem *Saccharomyces cerevisiae* Genome bindet.

Häufigkeit jeder Base an entsprechender Position.

Sequence Logos



# Informationsgehalt von Sequence Motifs

Annahme: in den binding motifs kommt jede Base an einer Position mit einer Wahrscheinlichkeit vor, die ihrer Häufigkeit in den binding motifs entspricht.

Annahme: jede Position ist unabhängig von den anderen.

Dann ist die Information im binding motif gleich der Summe der Information über die Positionen  $i$ .



# Informationsgehalt von Sequence Motifs

Ohne Vorwissen:  $H_i = - \sum_{b=A,C,G,T} \frac{1}{4} \log_2 \frac{1}{4} = 2 \text{ bit}$

Mit Wissen, dass Sequenz ein binding motif ist:

$$H_i = - \sum_{b=A,C,G,T} f_b^i \log_2 f_b^i$$

$f_b^i$  ist Frequenz, mit der Base  $b$  an Position  $i$  auftaucht.

Die Information an Position  $i$  ist dann

$$I_i = 2 + \sum_{b=A,C,G,T} f_b^i \log_2 f_b^i$$

**c**

<b>A</b>	0027000000010
<b>C</b>	464100000505
<b>G</b>	000001800112
<b>T</b>	422087088261

**d****e**



# Statistische Physik

Moleküle sind ständig in Bewegung und verändern Ort und Form (mikroskopischer Zustand).

Energieinhalt  $E$  spielt eine Rolle.

Temperatur  $T$  spielt eine Rolle.

Auch Druck  $p$  spielt häufig eine Rolle.

Die Parameter  $E, T, p$  sind makroskopische Parameter.

Statistische Physik schafft Verbindung zwischen mikroskopischen Zuständen und makroskopischen Parametern.



# Beispiel Gas

Der mikroskopische Zustand eines Gases ist durch Ort und Impuls jedes der  $N \approx 10^{23}$  Gasteilchen beschrieben.

Die makroskopischen Meßgrößen sind der Druck  $p$  und die Temperatur  $T$ .

Was kann man bei Messung von  $p$  und  $T$  über den mikroskopischen Zustand des Gases sagen?

Die Konzepte sind sehr universell und auf beliebige Systeme mit vielen Freiheitsgraden übertragbar.