

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 β 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: 3×10^9 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 15×10^6 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.

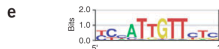
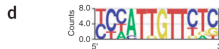
Wieviel Information tragen diese Sequenzabschnitte?

Sequence Motifs und Sequence Logos

a HEM13 CCCATTGTTCTC
 HEM13 TTTCTGGTTCTC
 HEM13 TCAATTGTTTAG
 ANB1 CTCATTGTTGTC
 ANB1 TCCATTGTTCTC
 ANB1 CCTATTGTTCTC
 ANB1 TCCATTGTTTCGT
 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

A	0027000000010
C	464100000505
G	000001800112
T	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

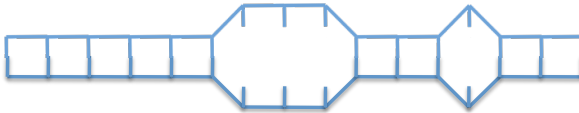
Der mikroskopische Zustand eines Gases ist durch Ort und Impuls jedes der $N \approx 10^{23}$ Gasteilchen beschrieben.

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.

Beispiel Doppelhelix



Uns interessiert, an wievielen Stellen die DNA-Helix aufgetrennt ist.

Dies hängt statistisch ab von der Umgebungstemperatur.

Bei niedriger Temperatur T treten nur wenige aufgetrennte Basenpaare auf, bei hohen Temperaturen kann sich die Helix vollständig auftrennen.

Wie können wir zu Aussagen über den mikroskopischen Zustand der Helix kommen?



Beispiel Doppelhelix

Kodierung des mikroskopischen Zustandes: $s_i \in \{0, 1\}$ beschreibt, ob das i -te Basenpaar zusammen ($s_i = 0$) oder aufgetrennt ist ($s_i = 1$).

Der Zustand der Helix ist dann beschrieben durch den Vektor $\vec{s} = (s_1, s_2, \dots, s_N)$, mit N als die Gesamtzahl der Basenpaare.

Die entscheidende Größe eines physikalischen Systems ist dessen Energie E .

Für die Auftrennung eines Basenpaares ist die Energie ε notwendig.

Sind n Paare aufgetrennt, so besitzt das System die Energie

$$E = n \cdot \varepsilon.$$



Energie des Systems gegeben

Frage: Wenn wir nur den aktuellen Energiegehalt E des Systems kennen (System sei isoliert und damit $E = \text{const.}$), welche Aussagen können wir über \vec{s} treffen?

Dann sind

$$n = \frac{E}{\varepsilon}$$

Basenpaare aufgetrennt.

Aber welche mit welcher Wahrscheinlichkeit?

Was können wir über die Wahrscheinlichkeit $P(\vec{s})$ sagen, die Helix im Zustand \vec{s} anzutreffen?



Maximierung der Unkenntnis

Wir wissen nur, daß

$$E = \sum_{i=1}^N \varepsilon \cdot s_i = \text{const.}$$

Ansatz: Da wir sonst nichts über den Zustand des Systems wissen, setzen wir dasjenige $P(\vec{s})$ an, welches unsere Unkenntnis über das System unter der Randbedingung

$$\sum_{i=1}^N \varepsilon \cdot s_i = E = \text{const.}$$

maximiert. Wir nehmen also maximale Entropie an.



Maximierung der Entropie

Unsere Unkenntnis ist gegeben durch die Entropie

$$H = - \sum_{\text{alle } \vec{s}} P(\vec{s}) \log P(\vec{s})$$

Wir suchen also diejenige Verteilungsfunktion $P(\vec{s})$, die H maximiert und für die $\sum_{\text{alle } \vec{s}} P(\vec{s}) = 1$ (Normierung), gleichzeitig aber auch $\sum_{i=1}^N \varepsilon \cdot s_i = E$ gilt.

$P(\vec{s})$ ist also überall Null, wo $\sum_{i=1}^N \varepsilon s_i = E$ nicht gilt.



Shannon Entropie

Aufgabe: Suche $P(\vec{s})$ welches

$$H = - \sum_{\text{alle } \vec{s}} P(\vec{s}) \log P(\vec{s})$$

maximiert, unter den Nebenbedingungen

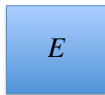
$$\sum_{\text{alle } \vec{s}} P(\vec{s}) = 1 \quad \text{und} \quad \sum_{i=1}^N \varepsilon \cdot s_i = E$$

Mit Lagrange-Multiplikatoren erhalten wir $P(\vec{s}) = \text{const.}$. Also gilt

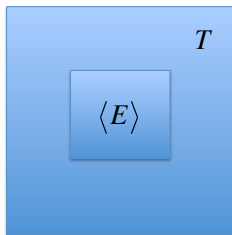
$$P(\vec{s}) = 0 \quad \text{falls} \quad \sum_{i=1}^N \varepsilon \cdot s_i \neq E$$

$$= \left(\frac{N!}{n!(N-n)!} \right)^{-1} \quad \text{falls} \quad \sum_{i=1}^N \varepsilon \cdot s_i = E.$$

System im Wärmebad



Bislang war das System isoliert. Dann E konstant. Die Zustände, die das System einnimmt, nennt man mikrokanonische Gesamtheit.



Für uns realistischer ist System im Wärmebad mit Temperatur T . Der Energieinhalt des Systems fluktuiert dann um einen Mittelwert $\langle E \rangle$. Die Zustände, die das System einnimmt, nennt man kanonische Gesamtheit.



System im Wärmebad

Die Energie des Gesamtsystems (System plus Wärmebad) ist konstant, der Energieinhalt des Systems schwankt um einen Mittelwert $\langle E \rangle$.

Auch hier stellen wir die Frage, mit welcher Wahrscheinlichkeit $P(\vec{s})$ das System den Zustand \vec{s} annimmt.

Nebenbedingung ist diesmal, dass $\langle E \rangle$ konstant ist.

Auch hier bestimmen wir $P(\vec{s})$ wieder durch Maximierung unserer Unkenntnis, also der Entropie.



Maximale Entropie im Wärmebad

Aufgabe: Wir suchen dasjenige $P(\vec{s})$, welches

$$H = - \sum_{\text{alle } \vec{s}} P(\vec{s}) \log P(\vec{s})$$

maximiert, diesmal unter den Nebenbedingungen

$$\sum_{\text{alle } \vec{s}} P(\vec{s}) = 1 \quad \text{und} \quad \sum_{\text{alle } \vec{s}} P(\vec{s}) E(\vec{s}) = \langle E \rangle$$

$E(\vec{s})$ bezeichnet die Energie des Systems im Zustand \vec{s} .



Maximale Entropie im Wärmebad

Ansatz mit Lagrangemultiplikatoren:

$$L = - \sum_{\text{alle } \vec{s}} P(\vec{s}) \ln P(\vec{s}) + \alpha \left(1 - \sum_{\text{alle } \vec{s}} P(\vec{s}) \right) + \beta \left(\langle E \rangle - \sum_{\text{alle } \vec{s}} P(\vec{s}) E(\vec{s}) \right)$$

Es gibt (bei der DNA) 2^N verschiedene \vec{s} . Wir nehmen jetzt ein \vec{s}^* heraus und leiten L nach $P(\vec{s}^*)$ ab:

$$\frac{\partial L}{\partial P(\vec{s}^*)} = -\ln P(\vec{s}^*) - 1 - \alpha - \beta E(\vec{s}^*)$$

Dies wird Null für

$$P(\vec{s}^*) = \underbrace{e^{-1-\alpha}}_{\text{Normierungskonstante } Z^{-1}} \cdot e^{-\beta E(\vec{s}^*)}$$



Maximale Entropie im Wärmebad

Z wird so gewählt, dass die Summe über alle Wahrscheinlichkeiten 1 ergibt:

$$\sum_{\text{alle } \vec{s}} P(\vec{s}) = Z^{-1} \sum_{\text{alle } \vec{s}} e^{-\beta E(\vec{s})} = 1$$

also

$$Z = \sum_{\text{alle } \vec{s}} e^{-\beta E(\vec{s})}$$

Z wird auch als Zustandssumme bezeichnet.

Für die Wahrscheinlichkeiten $P(\vec{s})$ erhalten wir dann die berühmte **Boltzmann-Verteilung**

$$P(\vec{s}) = \frac{e^{-\beta E(\vec{s})}}{Z}$$