1 Schnelle Einführung in bedingte Wahrscheinlichkeit

1.1 Kurzfassung

Bedingte Wahrscheinlichkeit: $\Pr(X|Y)$, d.h. die Wahrscheinlichkeit von X, gegeben Y.

1.2 Beispiel

Demografie Deutschlands:

Alter	Männliche Personen (Millionen)	Weibliche Personen (Millionen)
≤ 14	5,9	5,6
15-64	27,8	26,8
≥ 65	6,8	9,5
Summe	40,5	41,9
Gesamtsumme	82,4	

(Quelle: `en.wikipedia.org/wiki/Demographics_of_Germany`)

Wahrscheinlichkeiten unter Personen in Deutschland:

- $\Pr(\text{weiblich}) = 41{,}9/82{,}4 \approx 0{,}51$ (d.h. 51%)
  - In Worten: die Wahrscheinlichkeit, dass eine zufällige Person in Deutschland weiblich ist.
- $\Pr(\geq 65) = (6{,}8 + 9{,}5)/82{,}4 \approx 0{,}20$
- $\Pr(\text{weiblich}|\geq 65) = 9{,}5/(6{,}8 + 9{,}5) \approx 0{,}58$
  - In Worten: die Wahrscheinlichkeit, dass eine zufällige Person in Deutschland weiblich ist, gegeben dass diese Person 65 oder älter ist.
- $\Pr(\geq 65|\text{weiblich}) = 9{,}5/(5{,}6 + 26{,}8 + 9{,}5) = 9{,}5/41{,}9 \approx 0{,}23$
- $\Pr(\text{female}|\leq 14) = 5{,}6/(5{,}9 + 5{,}6) \approx 0{,}49$
- $\Pr(\text{weiblich}|\geq 65) > \Pr(\text{weiblich})$, weil Frauen im Durchschnitt länger als Männer leben.

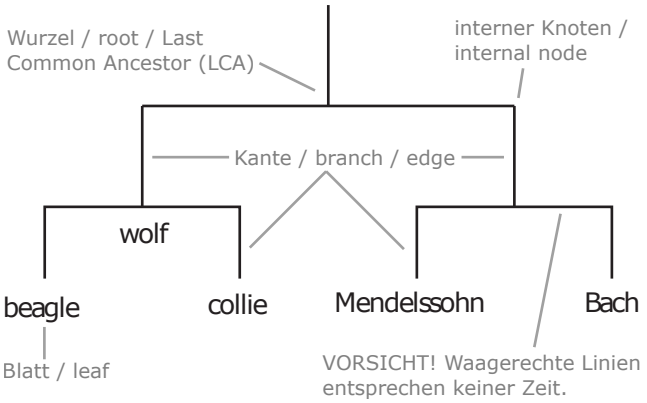
1.3 Ein Krimi mit bedingten Wahrscheinlichkeiten

Wenn Sie Lust haben, lesen Sie den Wikipedia-Artikel über Sally Clark: `https://en.wikipedia.org/wiki/Sally_Clark`. (Es gibt keine deutsche Version.) Sally Clark wurde wegen eines Beweises verurteilt, der statistisch gesehen fragwürdig war. Unter der Unterübertitel "Statistical evidence" im Artikel werden die verschiedenen statischen Fehler erläutert.

2 Übersicht über Phylogenie-Vorlesung

- Was ist Phylogenie / Warum machen wir das?
- Phylogenetische Bäume
- Wie schätzen wir Bäume ein?
  - Distanzmatrizen
  - UPGMA-Algorithmus
- Bootstrapping / statistische Signifikanz
- Projektchen über Riboswitches
- Molekulare-Uhr-Hypothese
- Wahrscheinlichkeits-basierte Phylogenie
  - Jukes-Cantor-Modell
  - Bessere Distanz-Matrix
  - Wahrscheinlichkeit eines phylogenetischen Baumes
  - Wo ist die Wurzel?

3 Phylogenetischer Baum



4 Newick-Format-Beispiel

```
((Bach:0.5,Beagle:0.5):0.2,Schlange:0.7);
```

5 Matrix im UPGMA-Beispiel

	a	b	c	d	e
a		17	21	31	23
b			30	34	21
c				28	39
d					43

6 Distanzbasierte Methoden

Erstellung einer Distanzmatrix (z.B. mit normalisierter Hamming-Distanz):

- Eingabe: ein multiples Alignment
- Ausgabe: eine Distanzmatrix

Einschätzung eines phylogenetischen Baumes (z.B. mit dem UPGMA-Algorithmus):

- Eingabe: eine Distanzmatrix
- Ausgabe: ein phylogenetischer Baum

7 Definition der Distanz in UPGMA

$$D(X,Y) = \frac{1}{|X| \cdot |Y|} \sum_{i \in X} \sum_{j \in Y} D(i,j)$$

gegeben die Mengen X und Y, die Sequenzen enthalten. $|X|$ heißt die Anzahl von Sequenzen in der Menge X.

8 Schnellerer UPGMA-Algorithmus: Mathe

Wir wollen die Vereinigung der Mengen Y_1 und Y_2. Wir nehmen weiter an: $Y = Y_1 \cup Y_2$ und $Y_1 \cap Y_2 = \emptyset$. D.h. jeder Bestandteil von Y ist entweder in Y_1 oder Y_2 enthalten, aber nicht in beiden.

Mit dem folgenden Trick können wir $D(X, Y)$ schneller berechnen:

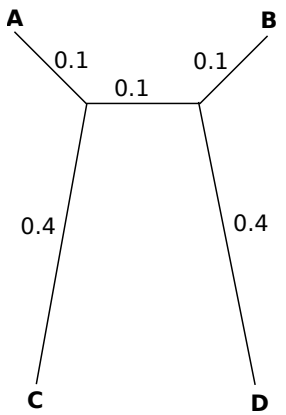
$$
\begin{aligned}
D(X,Y) =& \frac{1}{|X| \cdot |Y|} \sum_{i \in X} \sum_{j \in Y} D(i,j) && \text{(einfach die Definition von } D(X,Y)\text{)} \\
=& \frac{1}{|X| \cdot |Y|} \left(\left(\sum_{i \in X} \sum_{j \in Y_1} D(i,j) \right) + \left(\sum_{i \in X} \sum_{j \in Y_2} D(i,j) \right) \right) && \text{(unterteilen Elementen von } Y \text{ in } Y_1 \text{ und } Y_2\text{)} \\
=& \frac{1}{|X| \cdot |Y|} \left(|X| \cdot |Y_1| \cdot \overbrace{\left(\frac{1}{|X| \cdot |Y_1|} \sum_{i \in X} \sum_{j \in Y_1} D(i,j) \right)}^{\text{das müssen wir schon berechnet haben—es ist } D(X,Y_1)} \right. \\
& \left. + |X| \cdot |Y_2| \cdot \underbrace{\left(\frac{1}{|X| \cdot |Y_2|} \sum_{i \in X} \sum_{j \in Y_2} D(i,j) \right)}_{\text{das müssen wir schon berechnet haben—es ist } D(X,Y_2)} \right) \\
=& \frac{1}{|X| \cdot |Y|} \left(|X| \cdot |Y_1| \cdot D(X,Y_1) + |X| \cdot |Y_2| \cdot D(X,Y_2) \right) \\
=& \frac{(|Y_1| \cdot D(X,Y_1) + |Y_2| \cdot D(X,Y_2))}{|Y|} && (|X| \text{ hebt sich auf})
\end{aligned}
$$

9 Bootstrapping

- Typische Frage: wie sicher ist es, dass Bach und Beagle enger verwandt als Schlangen sind?
- Verfahren: durch zufällige Alignments eine Signifikanz einschätzen.

10 Molekulare-Uhr-Hypothese

Dieser Baum passt zur Molekulare-Uhr-Hypothese nicht. Der UPGMA-Algorithm kann diesen Baum nicht ausgeben.



11 Annahmen nach Jukes-Cantor

- Keine Selektion
- Nukleotide unabhängig
- A,C,G,T: Mutationsrate sind gleich
- A,C,G,T: durchschnittlich so häufig wie einander
- Zeitumkehrinvarianz / time reversible

12 Jukes-Cantor-Distanz

$$
\begin{aligned}
p &= 1 - \left(1/4 + 3/4 \left(1 - \frac{4}{3}\alpha \right)^t \right) \\
p &= 3/4 - 3/4 \left(1 - \frac{4}{3}\alpha \right)^t \\
\frac{4}{3}p &= 1 - \left(1 - \frac{4}{3}\alpha \right)^t \\
1 - \frac{4}{3}p &= \left(1 - \frac{4}{3}\alpha \right)^t \\
\log\left(1 - \frac{4}{3}p \right) &= t \log\left(1 - \frac{4}{3}\alpha \right) \\
t &= \log\left(1 - \frac{4}{3}p \right) / \log\left(1 - \frac{4}{3}\alpha \right)
\end{aligned}
$$