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

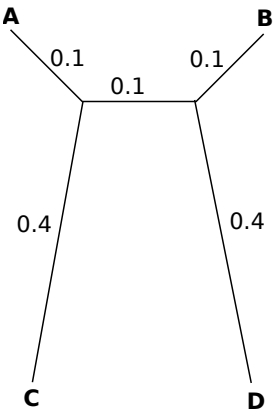
- Was ist Phylogenie / Warum machen wir das?
- Phylogenetische Bäume
- Wie schätzen wir Bäume ein?
  - Distanzmatrizen
  - UPGMA-Algorithmus
- Bootstrapping / statistische Signifikanz
- Molekulare-Uhr-Hypothese
- Wahrscheinlichkeits-basierte Phylogenie
  - Jukes-Cantor-Modell
  - Bessere Distanz-Matrix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

3 Bootstrapping

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

4 Molekulare-Uhr-Hypothese

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



5 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
- (Nur für Vorlesung, zur Vereinfachung)
  - Molekulare-Uhr-Hypothese
  - Distanz = Zeit

6 Jukes-Cantor-Distanz

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

Typische Form (p = sichtbare Mutationsrate, t = Zeit bzw. Distanz):

$$t = -\frac{3}{4}\log\left(1 - \frac{4}{3}p\right)$$

7 Modelle

- Jukes-Cantor
- Kimura, 1981
- GTR (Generalized Time-Reversible): am häufigsten verwendet (heutzutage)

8 Baum mittels Wahrscheinlichkeiten berechnen

Strategie: Maximum-Likelihood-Methode

Die Likelihood ist Pr(Alignment|Baum, Modell). Wir wollen diese Likelihood maximieren.

Annahmen zur Vereinfachung (nur für die Vorlesung):

- alle Sequenzen haben nur ein Nukleotid
- das Nukleotid jedes Knotens (auch internen Knotens) ist bekannt
- Struktur des Baumes ist bekannt
- Kantenlängen sind bekannt
- Baum = Wurzel mit 2 Kindern

Methode zur Verwurzelung (engl. rooting)

- Ignorieren. Ist oft nicht wichtig.
- Außengruppe (engl. outgroup) benutzen.
- "Zentrum" des Baumes (nicht verlässlich).

9 Felsenstein-Algorithmus