autor instrukcji: Małgorzata Kotulska

Bioinformatyka i Biologia Obliczeniowa
Laboratorium 8

1. Wyznacz drzewo UPGMA dla 5 sekwencji na podstawie poniższej macierzy odległości:

	X1	X2	X3	X4	X5
X1	0	13	9	6	13
X2	13	0	4	15	16
X3	9	4	0	11	12
X4	6	15	11	0	15
X5	13	16	12	15	0

2. Na podstawie dopasowania wielosekwencyjnego kilku wybranych gatunków, korzystając z fragmentów (lub całej ich długości) dopasowywanych sekwencji:
   a. zredukuj MSA i wyznacz macierz odległości miedzy sekwencjami, korzystając z wyników modelu Jukesa- Cantora
   b. wyznacz drzewo filogenetyczne UPGMA korzystając z obliczonych odległości
   c. porównaj swoje drzewo do drzew filogenetycznych generowanych automatycznie przez różne narzędzia do MSA.

Model Jukesa i **Cantora:**

$$p_{ij}(t) = \begin{cases} \frac{1}{4} + \frac{3}{4} \cdot e^{-4\alpha t} & \text{dla } i = j \\ \frac{1}{4} - \frac{1}{4} \cdot e^{-4\alpha t} & \text{dla } i \neq j \end{cases}$$

Gdzie α to tempo podstawień jednego spośród *nukleotydów* w sekwencji na dowolny inny.
Odsetek pozycji *D*, na których można *zaobserwować* różnice między dwoma sekwencjami, w zależności od chwili czasu *t*, wynosi:

$$D(t) = 3 \cdot p_{i,j}(2t) = \frac{3}{4} - \frac{3}{4} e^{-8\alpha t}$$

Natomiast ***średnia liczba wszystkich podstawień***, które rzeczywiście nastąpiły (nie wszystkie są widoczne po dłuższym czasie) na pozycję w sekwencji wynosi *d* :

$$\mathrm{d} = 2t \cdot 3\alpha = 6\,\alpha\, t$$

$$\mathrm{d} = \lim_{t \to 0}(D) = 6\alpha\, t$$

Można więc oszacować rzeczywistą ***liczbę wszystkich podstawień*** jako:

$$d = -\left(\frac{3}{4} \ln\left(1 - \frac{4}{3} D\right)\right)$$

Jaka zależność będzie obowiązywała dla sekwencji aminokwasów?

3. Do analizy rodzin białek wykorzystaj serwisy:

np. *Pfam, PRINTS, Prosite, CATH-Gene3D*