

# Statistical Alignment

Jotun Hein & Rune Lyngsø

Remember that under the TKF91 model of sequence evolution we have

$$p_0^N(t) = E(t) \quad (1)$$

$$p_n^N(t) = N(t)B(t)^{n-1} \quad \text{for } n > 0 \quad (2)$$

$$p_n^H(t) = H(t)B(t)^{n-1} \quad (3)$$

$$p_n^I(t) = I(t)B(t)^n \quad (4)$$

where  $p_i^N(t)$  is the probability that a nucleotide is deleted but leaves  $i$  descendants after time  $t$ ,  $p_i^H(t)$  is the probability that a nucleotide survives with  $i - 1$  further descendants after time  $t$ , and  $p_i^I(t)$  is the probability that the immortal link has  $i$  descendants after time  $t$  (i.e. that  $i$  nucleotides have been inserted at the beginning of the sequence). The shorthands used for convenience in these expressions are given by

$$\beta(t) = \frac{1 - e^{(\lambda - \mu)t}}{\mu - \lambda e^{(\lambda - \mu)t}}, \quad E(t) = \mu\beta(t), \quad N(t) = (1 - e^{-\mu t} - \mu\beta(t))(1 - \lambda\beta(t)),$$

$$I(t) = 1 - \lambda\beta(t), \quad B(t) = \lambda\beta(t), \quad H(t) = e^{-\mu t}(1 - \lambda\beta(t))$$

- a. Assume the TKF91 model of sequence evolution with nucleotide substitution described by the Jukes-Cantor single parameter model with rate  $\alpha$ . Let parameters be  $\alpha t = 0.2$ ,  $\mu t = 0.1$ , and  $\lambda t = 0.09$ . What is the likelihood of observing homologous sequences  $s1 = \text{AG}$  and  $s2 = \text{G}$ ?
- b. What is the probability of the most probable alignment of these two sequences?
- c. What is the most probable alignment?
- d. What is the probability of observing  $s1$  and  $s2$  as non-homologous sequences, *i.e.* assuming they are not descendants from the same ancestral sequence?
- e. How does things change, if at all, if you let  $s1 = \text{G}$  and  $s2 = \text{AG}$ , *i.e.* swap the ancestor-descendant relationship?