

Maximum Likelihood Problem (IBO level – conceptual)

Maximum Likelihood Problem (IBO level – conceptual)

You have aligned four DNA sequences of 100 bp. You run a maximum likelihood analysis with three different substitution models: JC69 (equal rates, equal base frequencies), HKY85 (different transition/transversion rates, unequal base frequencies), and GTR+G (general time reversible with gamma-distributed rate heterogeneity). The log-likelihood scores are:

- JC69: -2450.3
- HKY85: -2401.7
- GTR+G: -2395.2

Tasks:

1. Which model fits the data best? Justify your answer. (Consider the number of free parameters if needed.)
2. You perform a likelihood ratio test (LRT) between HKY85 and GTR+G. The difference in likelihood is 6.5, and the degrees of freedom = 3 (GTR adds three extra parameters). Is the improvement significant? (Critical chi-square value for $df=3$ at $\alpha=0.05$ is 7.815.)
3. What does the gamma shape parameter (α) estimate? If $\alpha < 1$, what does that indicate about rate variation among sites?
4. A common mistake is overfitting – choosing a complex model when a simpler model is sufficient. Which model would you recommend and why?
5. In IBO practicals, you may be asked to interpret a likelihood tree. What does the bootstrap support value (e.g., 95%) mean?