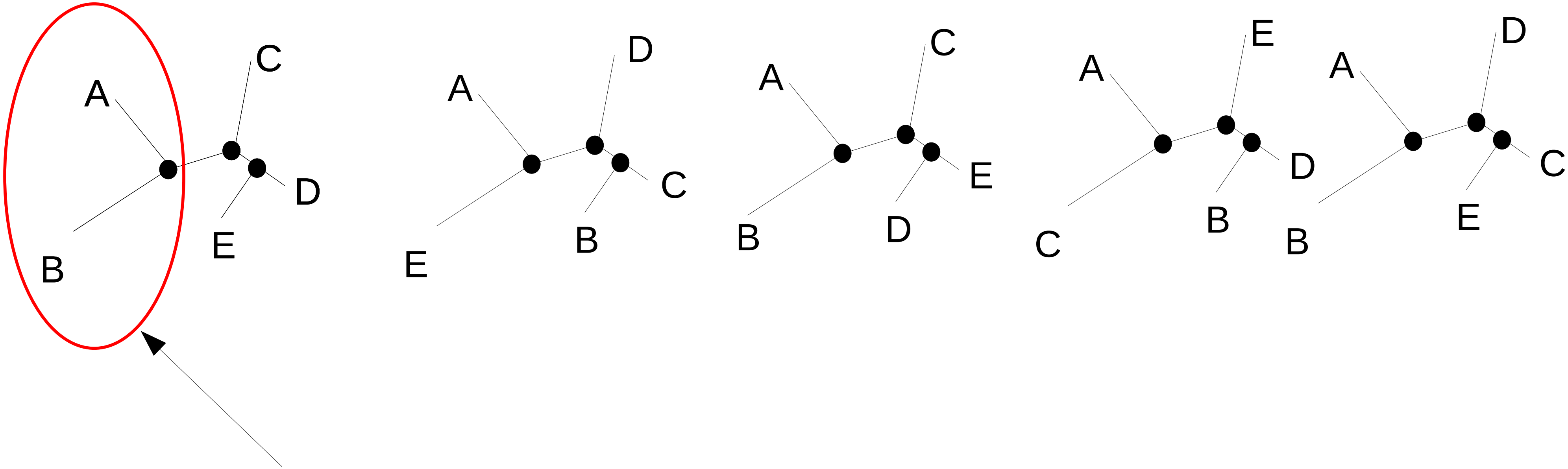
Bayesian Statistics and Markov Chain Monte Carlo Methods

Alexandros Stamatakis

Questions

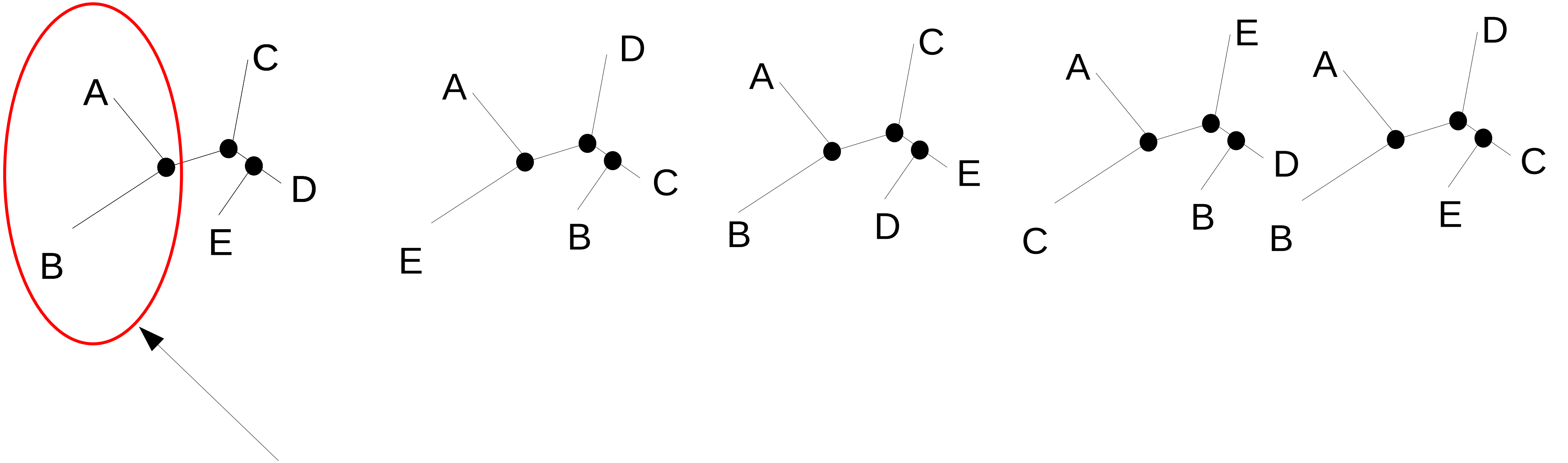
- How is it the posterior distribution of trees used to find the best phylogenetic tree?
- → it is not used for this, the key concept of Bayesian statistics is that we are using distributions (or samples thereof) for everything (not point estimates) and then use **summary statistics** to investigate the parameters we are interested in
- So we can use for instance a consensus tree building algorithm to construct a tree (an average essentially) from the posterior tree set

Back to Phylogenetics



What's the posterior probability of bipartition *AB|CDE*?

Back to Phylogenetics



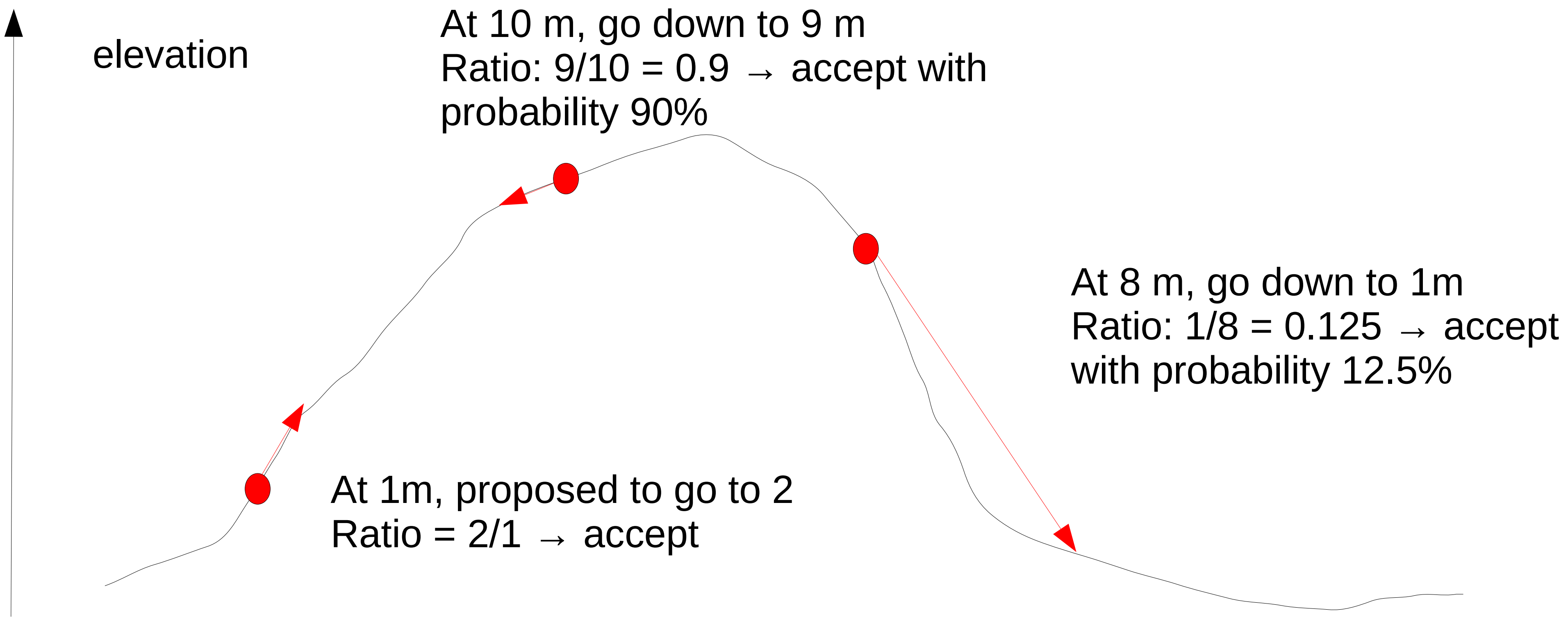
What's the posterior probability of bipartition *AB|CDE*?
We just count from the sample generated by MCMC, here it's *3/5 → 0.6*
This approximates the true proportion (posterior probability) of bipartition *AB|CDE*
if we have run the chain long enough and **if** it has converged

Questions

- On slide 48 it is said that steep downwards steps are almost never accepted. (i.e. not never) but how can it be that R>=1 at a downward step?
- I am not sure I understand the question correctly, but for R >= 1 the move will always be accepted and in our simple example with the robot this only happens if it goes higher up.
- If R = 0.25 (4m → 1m) for example, we draw a random number *r* uniformly at random between 0 and 1, if r > 0.25 we reject the move, if r <= 0.25 we accept the move

The Robot Metaphor

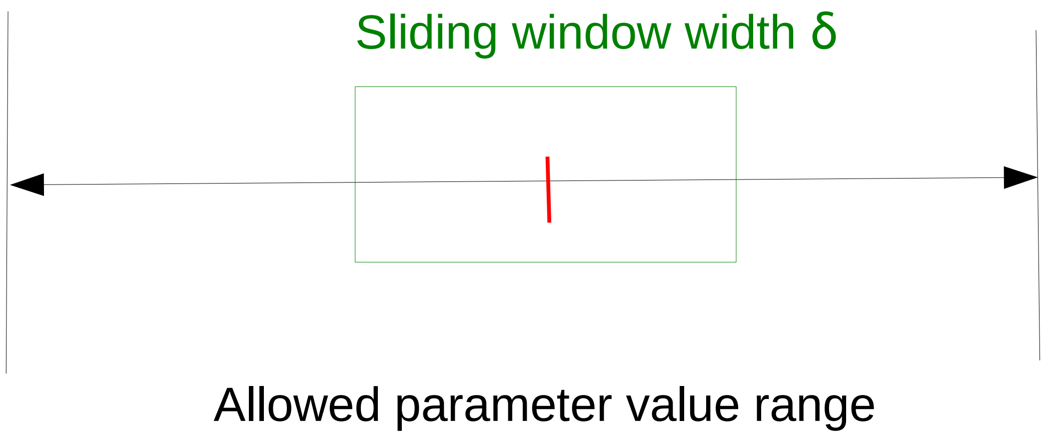
- Drop a robot onto an unknown planet to explore its landscape



Questions

- On slide 88 it is said, that delta can be auto-tuned. Does this mean, that delta is also in the parameter-vector θ?
- No, this is not a parameter of the *model* whose posterior probability we attempt to approximate but a parameter of the *specific proposal mechanism*
- The way this is done in practice is that the acceptance rate is monitored and calculated for say 10,000 successive proposals/ generations and subsequently the parameter is adapted

Sliding Window Proposal



Notes:
1. The hastings ratio of this move is *1*
2. The edge cases can be handled by back-projection
3. The window size δ can be tuned itself (auto-tuning) to obtain an acceptance rate of ≈ ¼
4. This proposal can be used, e.g., for the *α*-shape parameter of the *Γ* function in rate heterogeneity models

Questions

- If we change the substitution model during rjMCMC, how do we know the rates? (i.e. if we start from Jukes-Cantor, there are all rates 1/4, but if we then apply a split move to 11112, how do we determine the new entries of the rate matrix?)

Stolen from: *Bayesian Phylogenetic Model Selection Using Reversible Jump Markov Chain Monte Carlo*, 2004. n_i is the group size (number of rates in the group)

A split move works by picking one of the $N(M)$ groups with at least two substitution types at random. The group to be split will be designated $\mathbf{k}$. The substitution types in that group are then divided randomly, with the constraint that there is at least one item in each group. This results in two groups, $\mathbf{k}'_i$ and $\mathbf{k}'_j$. For example, if the current model, M, is 112334, then only two groups are available that could be potentially split: the groups with substitutions *AC* and *AG* or with substitution types *CG* and *CT*. One of these two groups would be chosen at random (with each having an equal probability of being chosen), and the substitution types in the group randomly divided into two new groups. In this case, if $\mathbf{k} = (AC, AG)$, then the resulting groups would be $\mathbf{k}'_i = (AC)$ and $\mathbf{k}'_j = (AG)$. To find the new rates for the two new groups, r'_i and r'_j, we first generate a uniform random variable on the interval $(-n_i r, n_j r)$, denoted u, where r is the rate of the group to be split. The new rates are then

$$r'_i = r + \frac{u}{n_i}$$

and

$$r'_j = r - \frac{u}{n_j}.$$

Problem #1
Model Proposals

- Split move

  Chose a set of substitution rates with > *1* member at random

  111222 (two-parameter model)

  and split it randomly into two rates

  111223 (three-parameter model)

- Merge move

  Chose two substitution rate sets at random

  111223

  and merge them into one substitution rate set

  111222

Problem #1
Model Proposals

- Split move

  Chose a set of substitution rates with > *1* member at random

  111222 (two-parameter model)

  and split it randomly in

  111223 (three-param

- Merge move

  Chose two substit...tion rate sets at random

  111223

  and merge them into one substitution rate set

  111222

Clear to everyone what the respective rate matrix looks like?