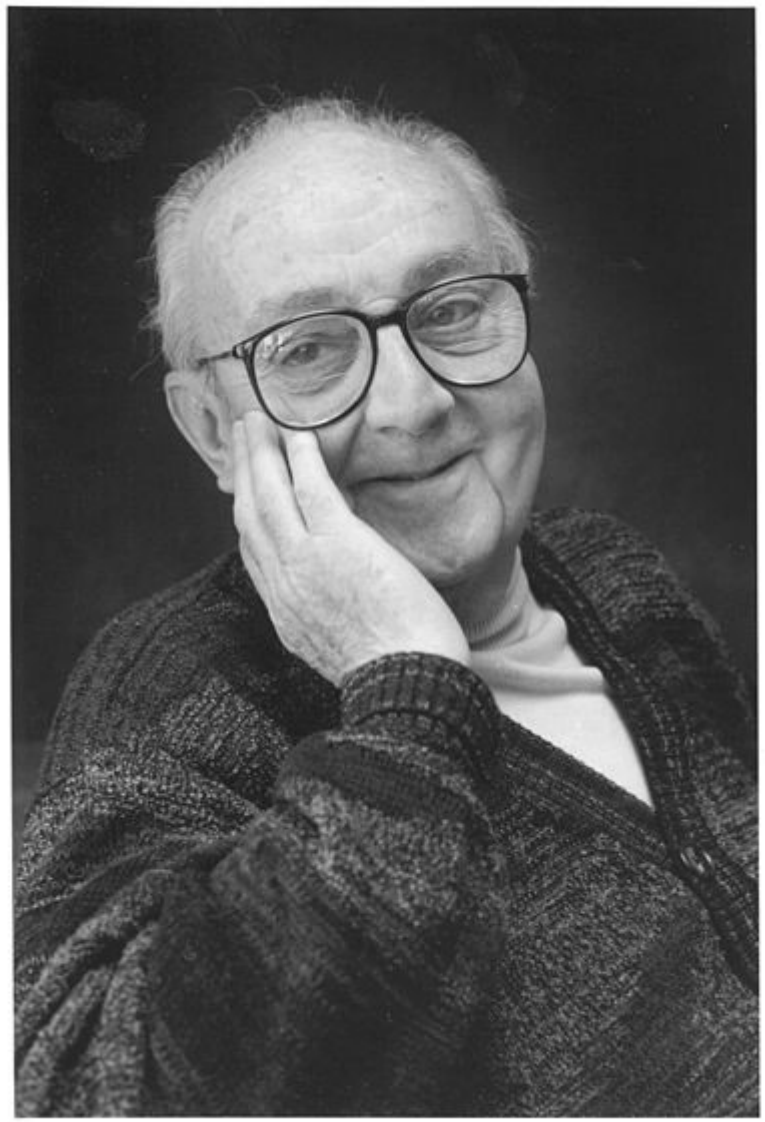
Model based phylogenetics

Pradipta Ray,

BIOL 6385 / BMEN 6389,

The University of Texas at Dallas

(some material based on content by PR in Eric Xing's 10-810 Carnegie Mellon class)



Wikipedia

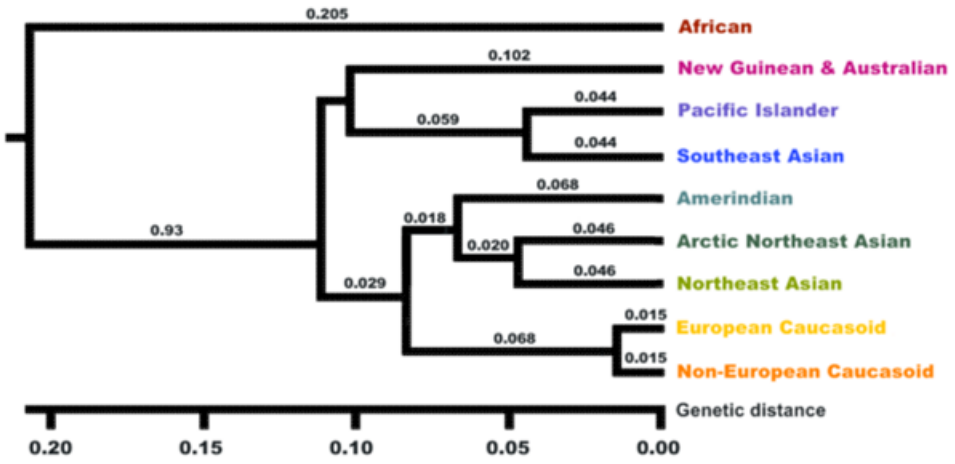
All models are wrong, but some are useful.

- George Box, 1979

2 schools of phylogeny reconstruction

- Distance based methods

Linkage tree for 9 population clusters showing genetic distances (F_{ST})
(Cavalli-Sforza *et al.*, 1994:80)



F_{ST} distance matrix for the 9 clusters shown above
(x10,000 with standard errors obtained by bootstrap analysis)

	AFR	NEC	EUC	NEA	ANE	AME	SEA	PAI	NGA
African	0.0								
Non-European Caucasian	1340.0 ± 301	0.0							
European Caucasian	1655.6 ± 416	154.7 ± 29	0.0						
Northeast Asian	1979.1 ± 452	640.4 ± 134	938.2 ± 217	0.0					
Arctic North-east Asian	2008.5 ± 387	708.2 ± 160	746.7 ± 210	459.7 ± 98	0.0				
Amerindian	2261.4 ± 434	955.5 ± 204	1038.2 ± 276	746.5 ± 183	577.4 ± 89	0.0			
Southeast Asian	2206.3 ± 529	939.6 ± 262	1240.4 ± 339	630.5 ± 299	1039.4 ± 326	1341.7 ± 418	0.0		
Pacific Islander	2505.4 ± 648	953.7 ± 230	1344.7 ± 354	723.8 ± 262	1181.2 ± 331	1740.7 ± 544	436.7 ± 87	0.0	
New Guinean and Australian	2472.0 ± 536	1179.1 ± 189	1345.7 ± 231	734.4 ± 118	1012.5 ± 257	1457.9 ± 283	1237.9 ± 277	808.7 ± 264	0.0

metric

Distance based methods

- Most distance metrics don't fit a tree : giving rise to inconsistent trees (data and trees don't agree)

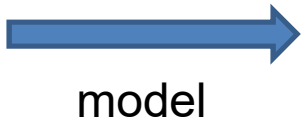
- Difficult to rationally choose one tree over another (is one tree better than another ? is one hypothesis better than another ?)

- Difficult to predict ancestral states (what are the patterns of evolutionary change ?)

2 schools of phylogeny reconstruction

- Character based methods

A B C D E
A X C D E
A X Y D E



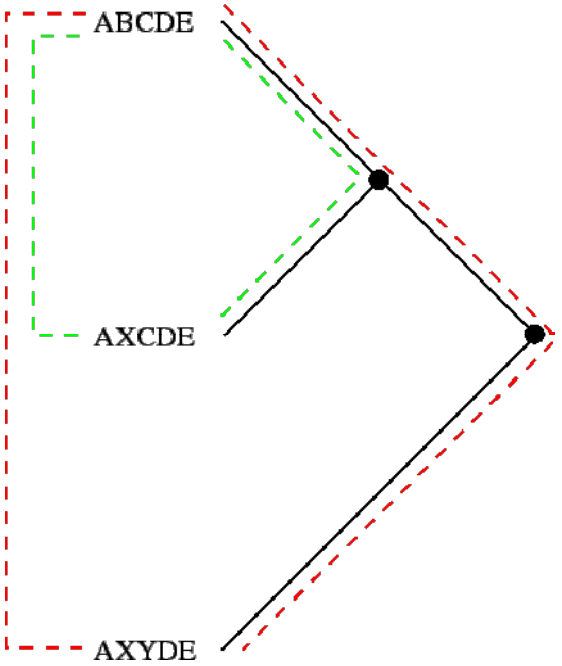


Figure: A Phylogenetic Tree

molgen.mpg.de

UT DALLAS
The University of Texas at Dallas

Character based methods

- Explicitly model how the characters change
  - easier to predict ancestral states
  - model can be used to score candidate trees
  - no trees are wrong, they can be better or worse than others in light of the data and model (score)
  - a hierarchical clustering will still be generated : the inter cluster “distance”s are implicit
- Remember our stochastic processes for evolution ?
  - Model $P (X_t = i \mid X_0 = j)$

Can we model non genomic data ?

- Simply use a discrete character set to model phenotype(s) (eg. meristic features : no of vertebra in spinal column : 1, 2, 3, 4, …)

- Define stochastic process with same number of states (eg. X_t can take value of no of vertebra)

Continuous data

- Continuous time, continuous value Markov process
  - Wiener process / brownian motion

- Usually not done in practice, a better idea is to "discretize" the continuous quantity into a number of bins
  - Model evolution over bin index

Model based evolution

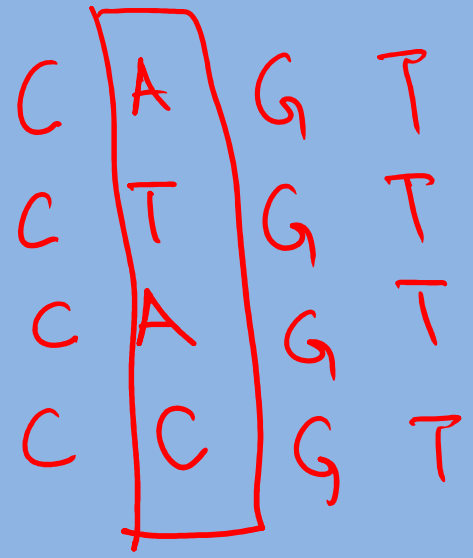
- Given the data
  - Generate each possible tree
  - Score each tree with the model
  - Pick the tree whose “score” is the “best”
    - Score for probabilistic models = Likelihood = P (data | model)
    - Best score for probabilistic models = Highest likelihood

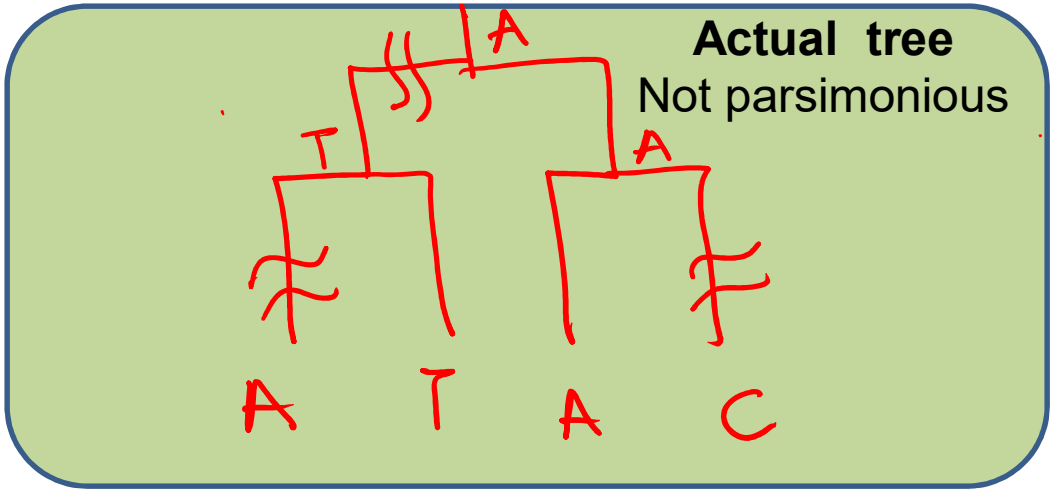
Parsimony

- Fewest substitutions to explain alignment data
  - Is min substitutions equivalent to Occam's razor / minimal assumptions - unclear
- Build a tree where branch length = no of substitutions (how can it be > 1 ?)
- Minimize sum of branch lengths
  - No selection (could be modelled)
  - Not all data may be parsimonious
  - May be many equally parsimonious trees
  - Ambiguous ancestral sequences for same tree
  - Some (many!) character alignments not used

Data

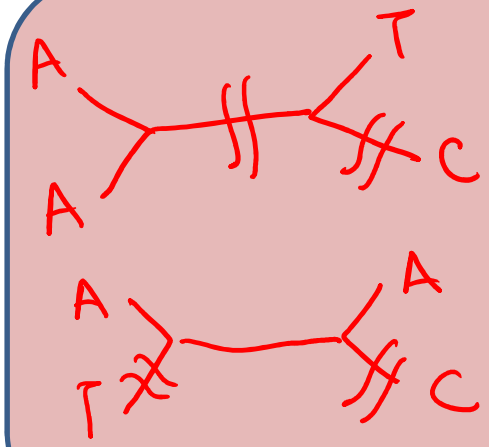
Only 1 column is informative





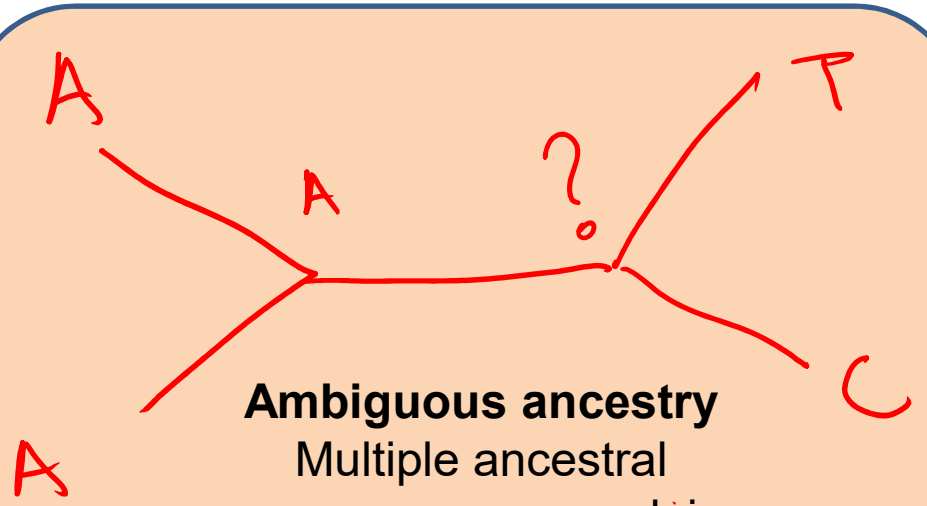
Actual tree

Not parsimonious



Ambiguous topology

Multiple most parsimonious trees



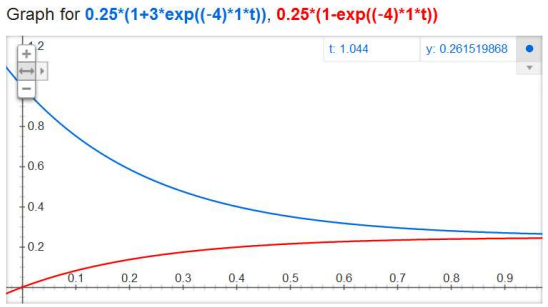
Ambiguous ancestry

Multiple ancestral sequences can explain data

Some facts about parsimony

- Not a non parametric method (doesn't have a clear data dependent parametric structure)
- May be inconsistent (may converge to wrong tree even with unlimited training data)
- Likelihood models (eg. ones where no observed change is preferred to change always) may be sufficient (but not necessary) to approximate parsimony

  sometimes (the likelihood model does more)
- Assumptions (difficult to make explicit)
  - minimum evolution
  - independence across sites (weighted variant)
  - agnostic to nature of change (weighted variant)



J C probabilities as a fn of time

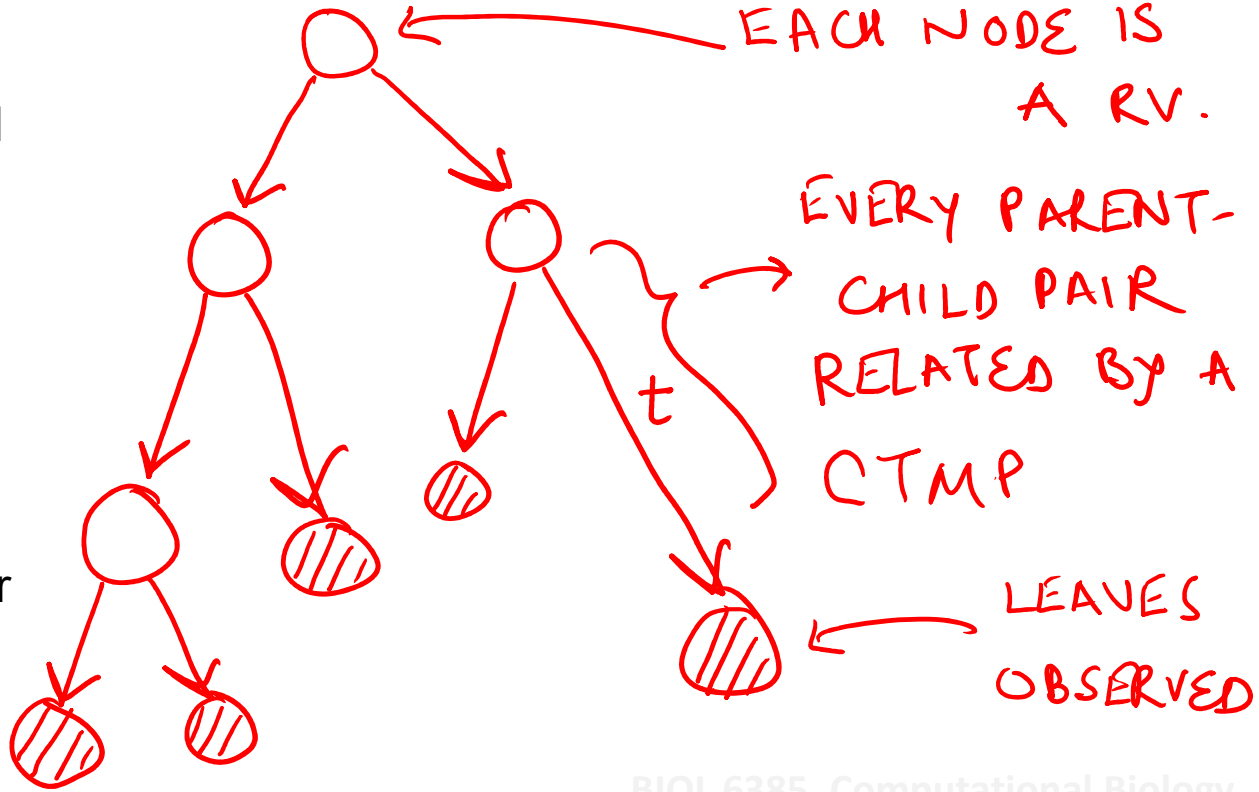
Maximum likelihood framework

- What are our random variables ? Which are observed ? How are they related ?

Typically, parameters of the CTMPs are assumed to be global

Each fork = 2 independent CTMPs

Lineage specific modelling is restricted to different time intervals for the stoch processes (branch lengths)



The model

- Topology : which rv gives rise to which rv
- Branch lengths : time intervals for which the stochastic processes run
- CTMP parameters : the evolutionary matrix – could be specific for each branch, could be universal for the tree, or somewhere in between

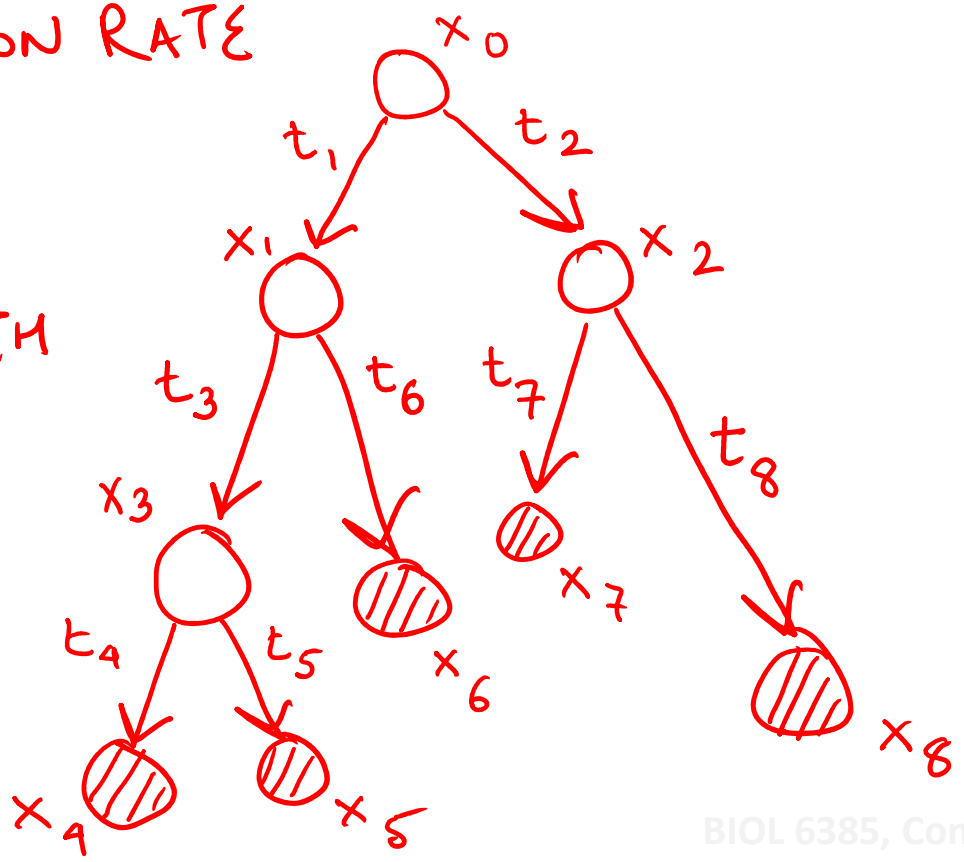
Meaning of the branch length

Length of t ⟶ NOT real time

FIX MUTATION RATE AT 1

⇓

BRANCH LENGTH = EXPECTED NO OF SUBST / SITE



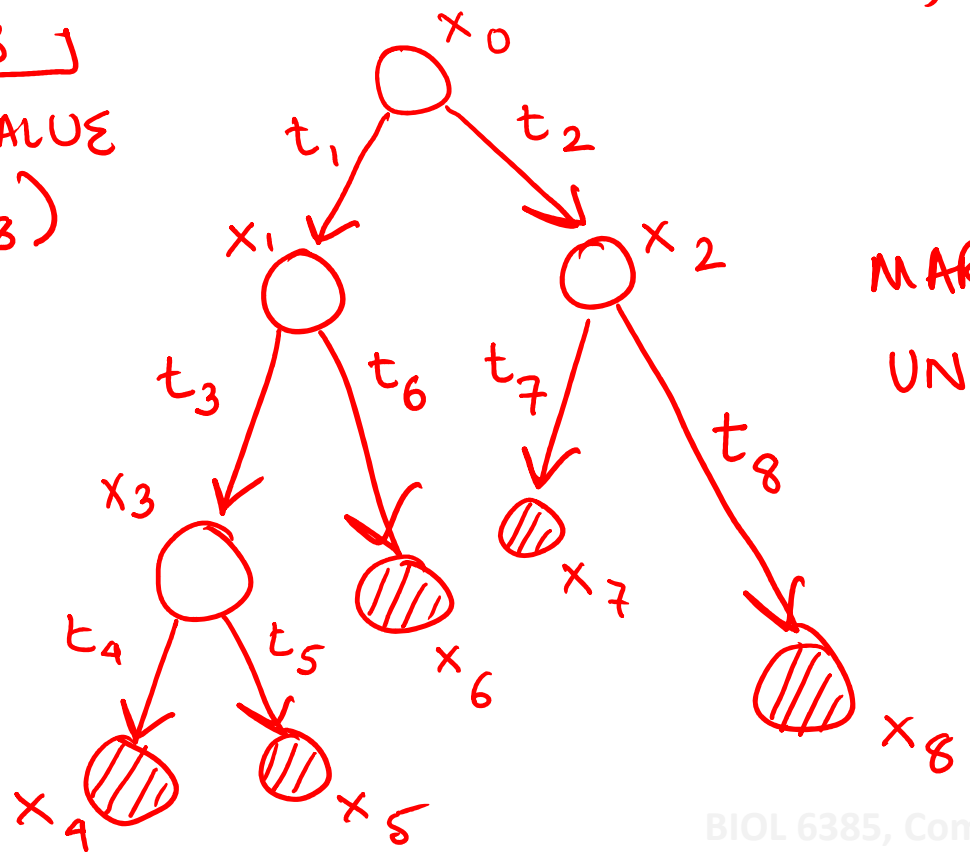
Likelihood of a single site

$$P(x_4, x_5, x_6, x_7, x_8 \mid \theta, T)$$

$$= \sum_{x_0} \sum_{x_1} \sum_{x_2} \sum_{x_3} P(x_0, x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8 \mid \theta, T)$$

CYCLE THRU EACH VALUE

$T = (t_1, t_2, \ldots, t_8)$



MARGINALIZE UNOBSERVED R.V.'S

Likelihood of a single site

$$P(x_4, x_5, x_6, x_7, x_8 \mid \theta, T)$$

$$= \sum_{x_0} \sum_{x_1} \sum_{x_2} \sum_{x_3} P(x_8 \mid x_2, \theta, t_8) P(x_7 \mid x_2, \theta, t_7) P(x_6 \mid x_1, \theta, t_6)$$

$$\times P(x_5 \mid x_3, \theta, t_5)$$

$$\times P(x_4 \mid x_3, \theta, t_4)$$

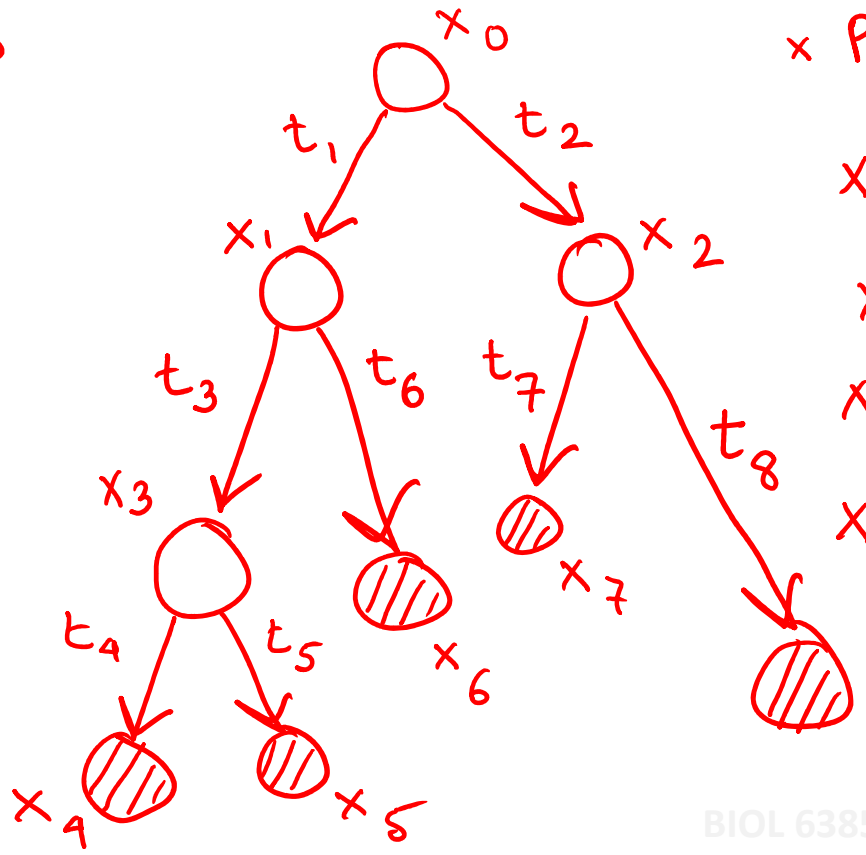
$$\times P(x_3 \mid x_1, \theta, t_3)$$

$$\times P(x_2 \mid x_0, \theta, t_2)$$

$$\times P(x_1 \mid x_0, \theta, t_1)$$

$$\times P(x_0 \mid \theta)$$

WHY?



Likelihood of a single site

$$P(x_4, x_5, x_6, x_7, x_8 \mid \theta, T)$$

$$= \sum_{x_0} \sum_{x_1} \sum_{x_2} \sum_{x_3} P(x_8 \mid x_2, \theta, t_8) P(x_7 \mid x_2, \theta, t_7) P(x_6 \mid x_1, \theta, t_6)$$

$$\times P(x_5 \mid x_3, \theta, t_5)$$

$$\times P(x_4 \mid x_3, \theta, t_4)$$

$$\times P(x_3 \mid x_1, \theta, t_3)$$

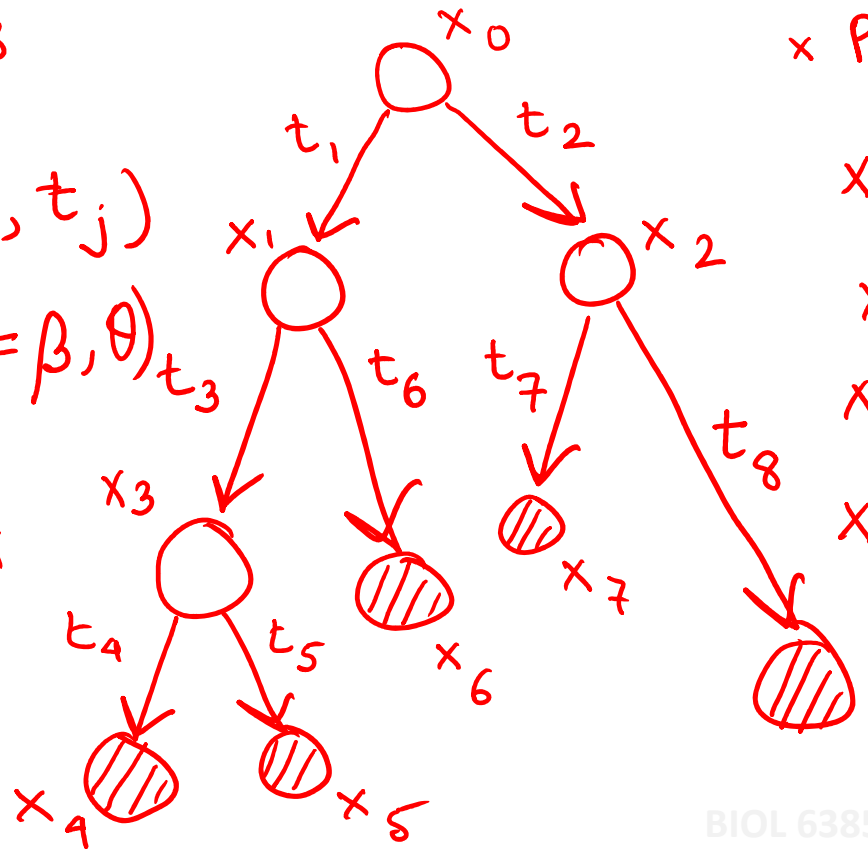
$$\times P(x_2 \mid x_0, \theta, t_2)$$

$$\times P(x_1 \mid x_0, \theta, t_1)$$

$$\times P(x_0 \mid \theta)$$

$$P(x_j \mid x_i, \theta, t_j) \equiv P(Y_t = \alpha \mid Y_0 = \beta, \theta)_{t_3}$$

↑ CALCULATE EACH TERM USING A CTMP



Quick reminder

eg. Jukes Cantor, 1969

$$Q = \begin{pmatrix} * & \frac{\mu}{4} & \frac{\mu}{4} & \frac{\mu}{4} \\ \frac{\mu}{4} & * & \frac{\mu}{4} & \frac{\mu}{4} \\ \frac{\mu}{4} & \frac{\mu}{4} & * & \frac{\mu}{4} \\ \frac{\mu}{4} & \frac{\mu}{4} & \frac{\mu}{4} & * \end{pmatrix}$$

$$P = \begin{pmatrix} \frac{1}{4} + \frac{3}{4}e^{-t\mu} & \frac{1}{4} - \frac{1}{4}e^{-t\mu} & \frac{1}{4} - \frac{1}{4}e^{-t\mu} & \frac{1}{4} - \frac{1}{4}e^{-t\mu} \\ \frac{1}{4} - \frac{1}{4}e^{-t\mu} & \frac{1}{4} + \frac{3}{4}e^{-t\mu} & \frac{1}{4} - \frac{1}{4}e^{-t\mu} & \frac{1}{4} - \frac{1}{4}e^{-t\mu} \\ \frac{1}{4} - \frac{1}{4}e^{-t\mu} & \frac{1}{4} - \frac{1}{4}e^{-t\mu} & \frac{1}{4} + \frac{3}{4}e^{-t\mu} & \frac{1}{4} - \frac{1}{4}e^{-t\mu} \\ \frac{1}{4} - \frac{1}{4}e^{-t\mu} & \frac{1}{4} - \frac{1}{4}e^{-t\mu} & \frac{1}{4} - \frac{1}{4}e^{-t\mu} & \frac{1}{4} + \frac{3}{4}e^{-t\mu} \end{pmatrix}$$

Likelihood of a single site

$$P(x_4, x_5, x_6, x_7, x_8 \mid \theta, T)$$

$$= \sum_{x_0} \sum_{x_1} \sum_{x_2} \sum_{x_3} P(x_8 \mid x_2, \theta, t_8) P(x_7 \mid x_2, \theta, t_7) P(x_6 \mid x_1, \theta, t_6)$$

$$\times P(x_5 \mid x_3, \theta, t_5)$$

$$\times P(x_4 \mid x_3, \theta, t_4)$$

$$\times P(x_3 \mid x_1, \theta, t_3)$$

$$\times P(x_2 \mid x_0, \theta, t_2)$$

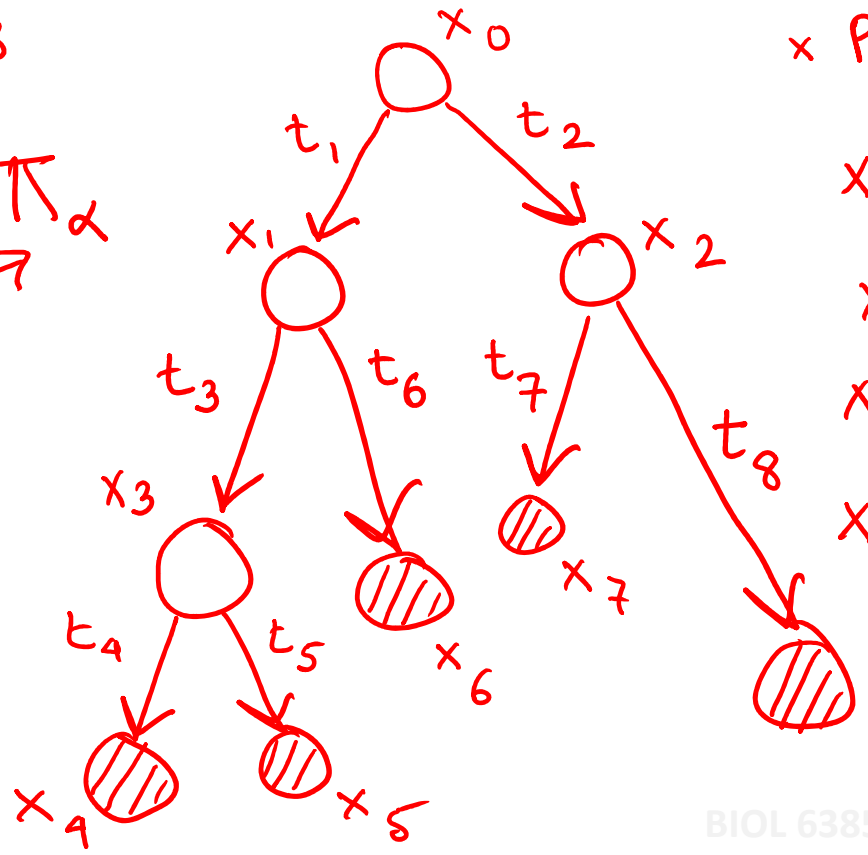
$$\times P(x_1 \mid x_0, \theta, t_1)$$

$$\times P(x_0 \mid \theta)$$

$P(x_0 = \alpha) = \pi_\alpha$

STATIONARY / EQUILIBRIUM DISTR.

WHY?



Likelihood of a single site

$$P(x_4, x_5, x_6, x_7, x_8 \mid \theta, T)$$

$$= \sum_{x_0} \sum_{x_1} \sum_{x_2} \sum_{x_3} P(x_8 \mid x_2, \theta, t_8) P(x_7 \mid x_2, \theta, t_7) P(x_6 \mid x_1, \theta, t_6)$$

$$\times P(x_5 \mid x_3, \theta, t_5)$$

$$\times P(x_4 \mid x_3, \theta, t_4)$$

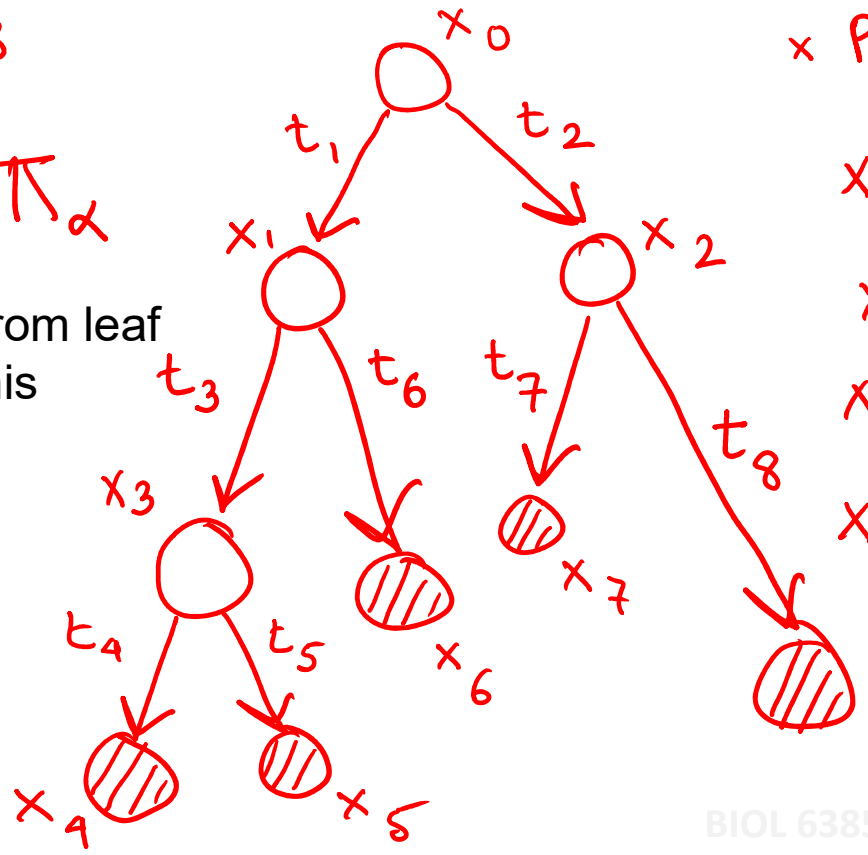
$$\times P(x_3 \mid x_1, \theta, t_3)$$

$$\times P(x_2 \mid x_0, \theta, t_2)$$

$$\times P(x_1 \mid x_0, \theta, t_1)$$

$$\times P(x_0 \mid \theta)$$

$$P(x_0 = \alpha) = \pi_\alpha$$

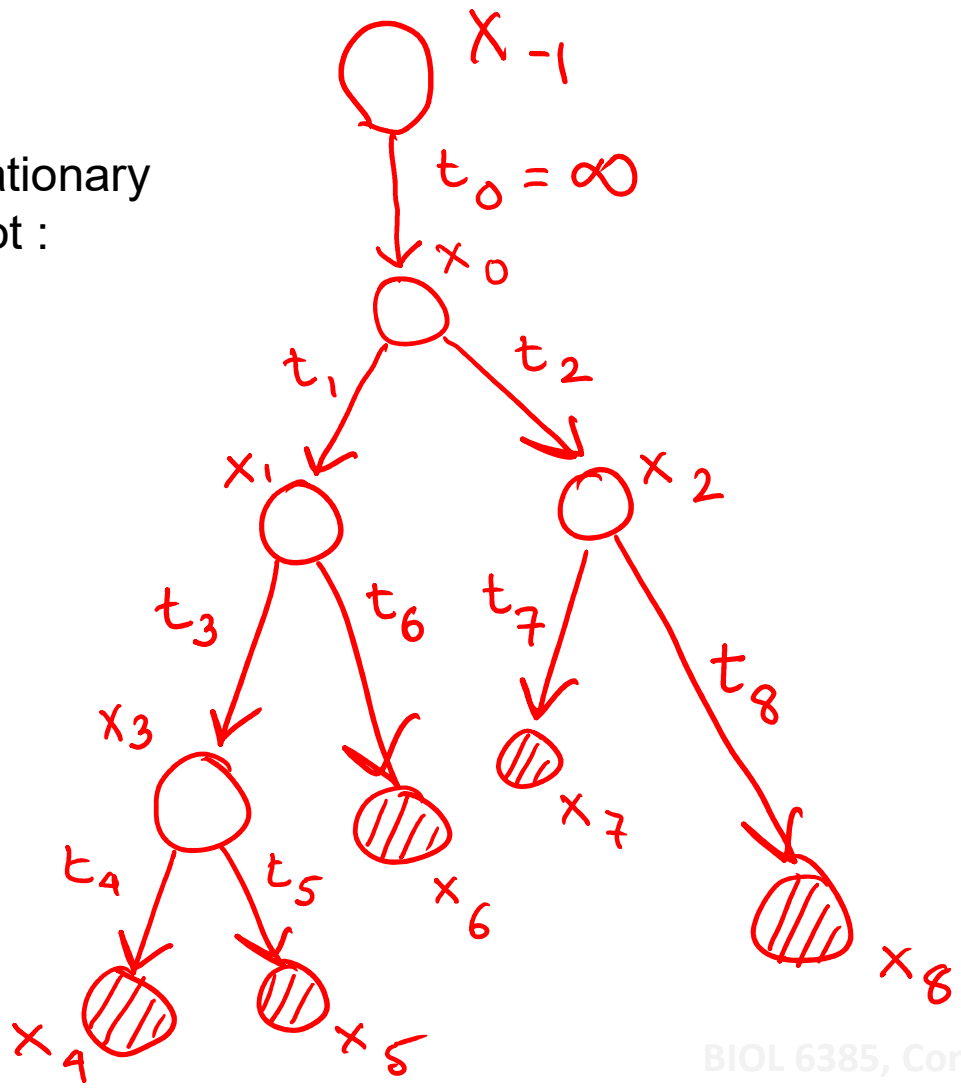


Is the shortest distance from leaf to root long enough for this assumption ?

No !

Likelihood of a single site

Motivation for stationary distribution at root :



Likelihood of a single site

$$P(x_4, x_5, x_6, x_7, x_8 \mid \theta, T)$$

$$= \sum_{x_0} \sum_{x_1} \sum_{x_2} \sum_{x_3} P(x_8 \mid x_2, \theta, t_8) P(x_7 \mid x_2, \theta, t_7) P(x_6 \mid x_1, \theta, t_6)$$

$$\times P(x_5 \mid x_3, \theta, t_5)$$

$$\times P(x_4 \mid x_3, \theta, t_4)$$

$$\times P(x_3 \mid x_1, \theta, t_3)$$

$$\times P(x_2 \mid x_0, \theta, t_2)$$

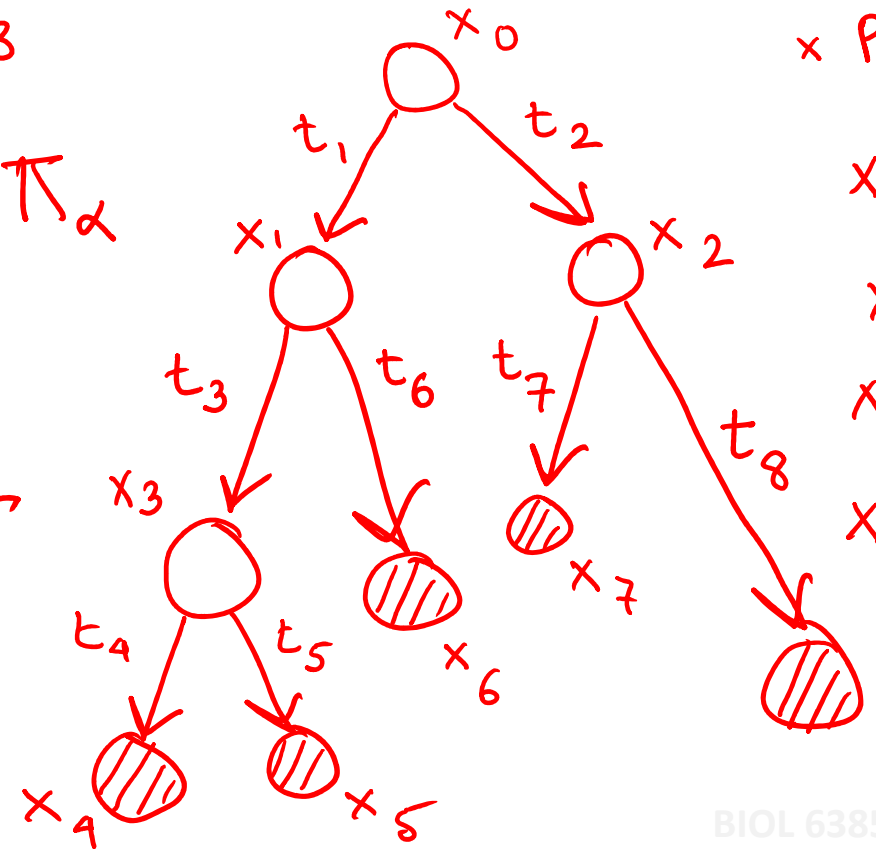
$$\times P(x_1 \mid x_0, \theta, t_1)$$

$$\times P(x_0 \mid \theta)$$

$$P(x_0 = \alpha) = \pi_\alpha$$

IN REALITY, AN APPROXIMATION

IN A BAYESIAN SETTING, USED AS PRIOR



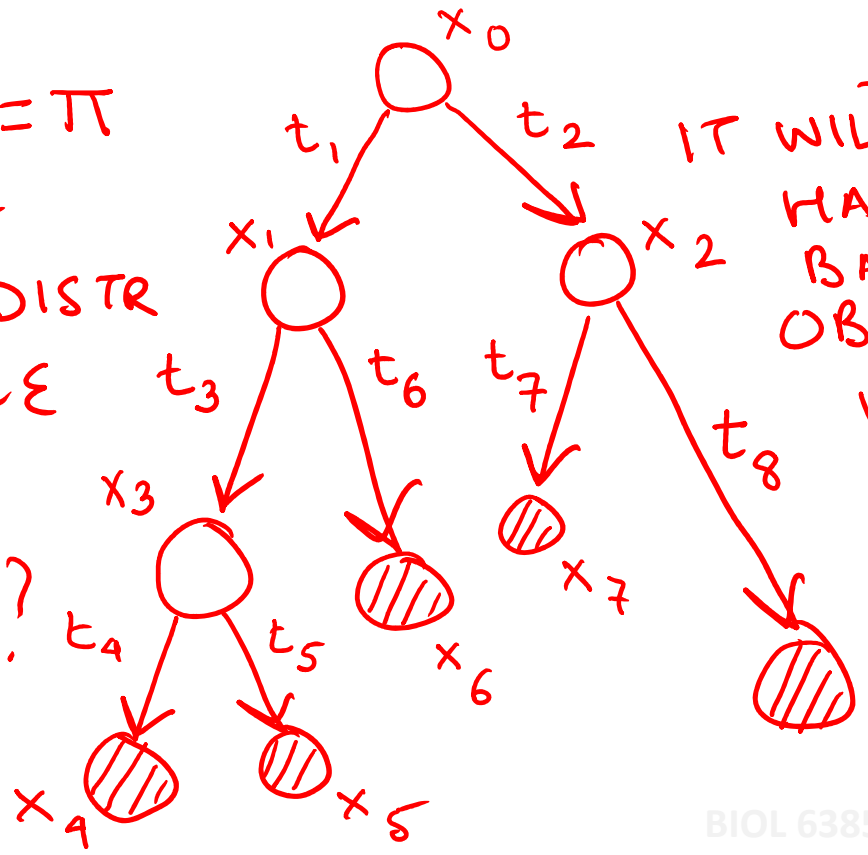
Wait, wont the distr get stuck ?

CONSIDER THE STOCHASTIC PROCESS AS A FUNCTION THAT INCREMENTALLY MODIFIES $P_t(x)$

i.e. $SP(P_t(x)) \rightarrow P_{t+h}(x)$

π is a fixed point: $SP(\pi) = \pi$

SO WONT THE NUCLEOTIDE DISTR ALONG THE TREE BE π EVERYWHERE ?



ANS

IT WILL, IF WE DONT HAVE OBSERVATIONS

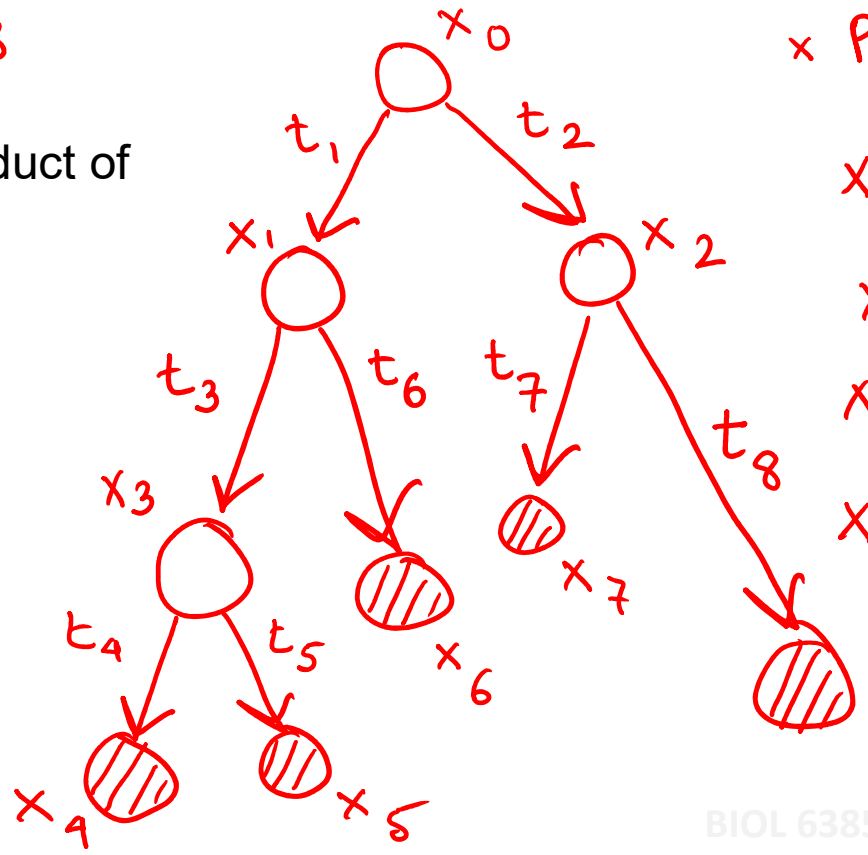
BAYESIAN: OBSERVATIONS WILL MODIFY OUR PRIOR (π) INTO A POSTERIOR, BASED ON MODEL

Brute force computation

$$P(x_4, x_5, x_6, x_7, x_8 \mid \theta, T)$$

$$= \sum_{x_0} \sum_{x_1} \sum_{x_2} \sum_{x_3} P(x_8 \mid x_2, \theta, t_8) P(x_7 \mid x_2, \theta, t_7) P(x_6 \mid x_1, \theta, t_6)$$

$$\times P(x_5 \mid x_3, \theta, t_5) \times P(x_4 \mid x_3, \theta, t_4) \times P(x_3 \mid x_1, \theta, t_3) \times P(x_2 \mid x_0, \theta, t_2) \times P(x_1 \mid x_0, \theta, t_1) \times P(x_0 \mid \theta)$$



Sum of products ➔ Product of sums

Horner's Rule
(Horner, 1830)
(Zhu Shijie, 1303)

AVOID 4^N SUMS FOR N+1 LEAVES

Horner's rule

- Push terms as far to the left as possible

$$\sum_x \sum_y \sum_z f_1(x,y,z) f_2(x,y) f_3(y,z) \times f_4(z,x) f_5(x) f_6(y) f_7(z)$$

$$= \sum_x f_5(x) \sum_y f_6(y) f_2(x,y) \underbrace{\sum_z f_8(x,y,z)}$$

$$= \sum_x f_5(x) \underbrace{\sum_y f_6(y) f_2(x,y) f_9(x,y)}$$

$$= \sum_x f_5(x) f_{10}(x) = \text{constant}$$

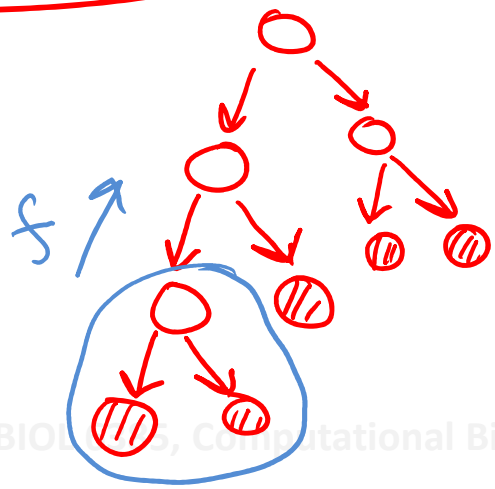
Felsenstein's pruning algo

$$P(x_4, x_5, x_6, x_7, x_8 \mid \theta, T)$$

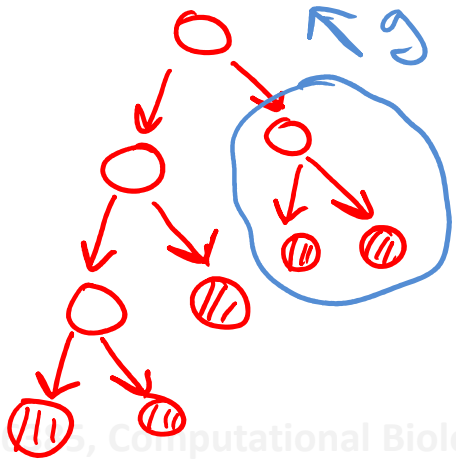
$$= \sum_{x_0} \sum_{x_1} \sum_{x_2} \sum_{x_3} P(x_8 \mid x_2, \theta, t_8) P(x_7 \mid x_2, \theta, t_7) P(x_6 \mid x_1, \theta, t_6)$$
$$\times P(x_5 \mid x_3, \theta, t_5)$$
$$\times P(x_4 \mid x_3, \theta, t_4)$$
$$\times P(x_3 \mid x_1, \theta, t_3)$$
$$\times P(x_2 \mid x_0, \theta, t_2)$$
$$\times P(x_1 \mid x_0, \theta, t_1)$$
$$\times P(x_0 \mid \theta)$$

$$= \sum_{x_0} P(x_0 \mid \theta)$$
$$\times \sum_{x_1} P(x_1 \mid x_0, \theta, t_1) P(x_6 \mid x_1, \theta, t_6)$$
$$\times \sum_{x_2} P(x_2 \mid x_0, \theta, t_2) P(x_7 \mid x_2, \theta, t_7) P(x_8 \mid x_2, \theta, t_8)$$
$$\times \sum_{x_3} P(x_3 \mid x_1, \theta, t_3) P(x_5 \mid x_3, \theta, t_5) \cdot P(x_4 \mid x_3, \theta, t_4)$$

$$P(D|M) = \sum_{x_0} P(x_0|\theta) \times \sum_{x_1} P(x_1|x_0,\theta,t_1) P(x_6|x_1,\theta,t_6) \times \sum_{x_2} P(x_2|x_0,\theta,t_2) P(x_7|x_2,\theta,t_7) P(x_8|x_2,\theta,t_8) \times \underbrace{\sum_{x_3} P(x_3|x_1,\theta,t_3) P(x_5|x_3,\theta,t_5) \cdot P(x_4|x_3,\theta,t_4)}_{f(x_1)}$$



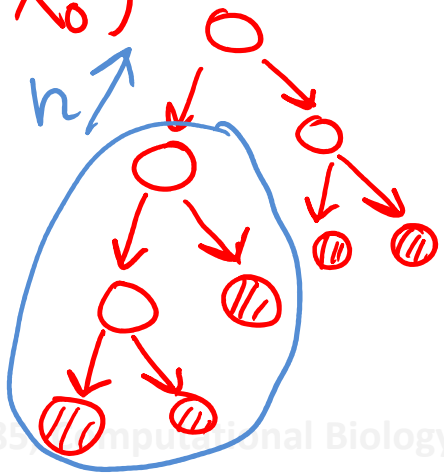
$$P(D|M) = \sum_{x_0} P(x_0|\theta) \times \sum_{x_1} P(x_1|x_0,\theta,t_1) P(x_6|x_1,\theta,t_6) \times f(x_1) \times \underbrace{\sum_{x_2} P(x_2|x_0,\theta,t_2) P(x_7|x_2,\theta,t_7) P(x_8|x_2,\theta,t_8)}_{g(x_0)}$$



$$P(D|M) = \sum_{x_0} P(x_0|\theta) \times g(x_0)$$

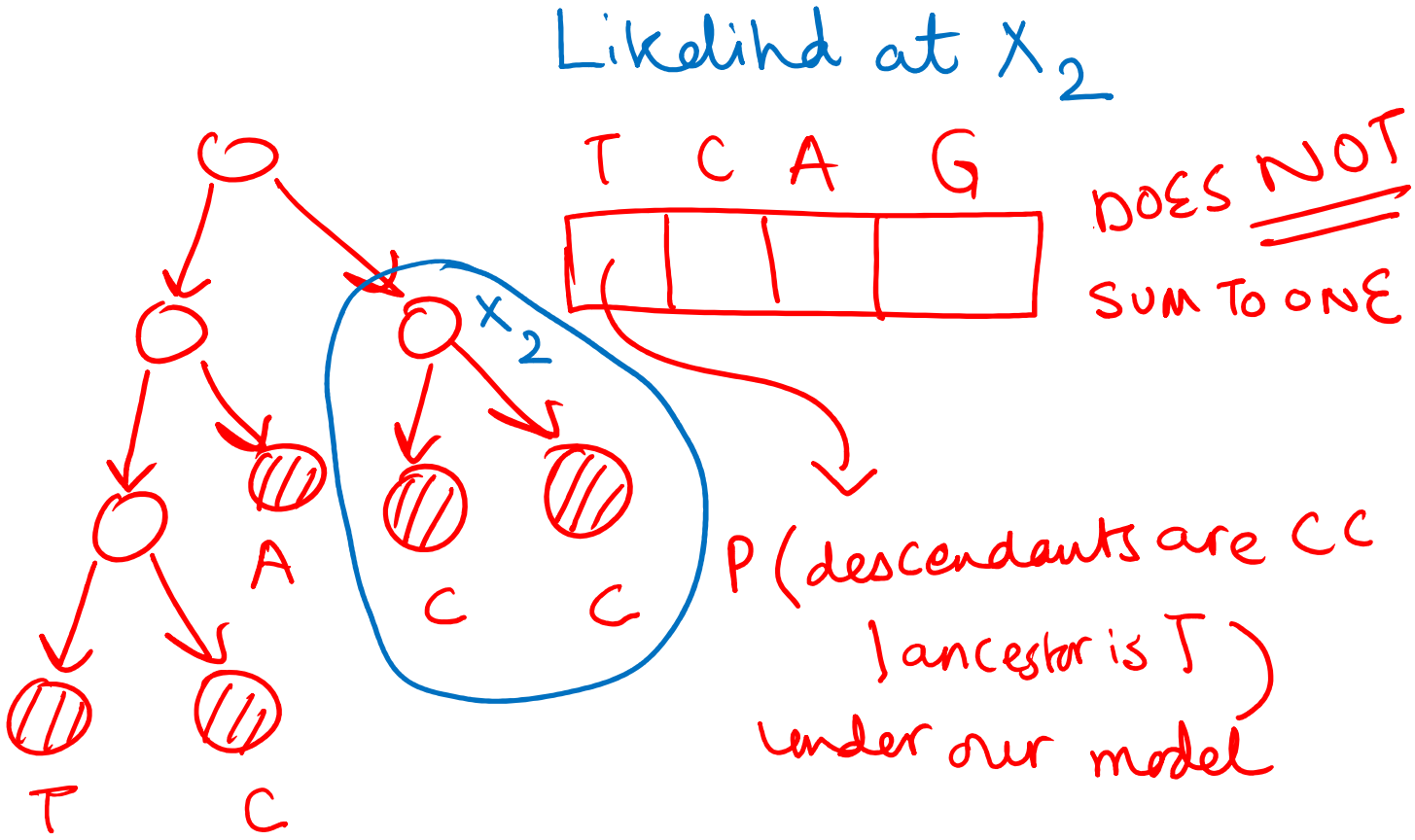
$$\times \underbrace{\sum_{x_1} P(x_1|x_0,\theta,t_1) P(x_6|x_1,\theta,t_6) \times f(x_1)}_{h(x_0)}$$

$$= \sum_{x_0} P(x_0|\theta)\, g(x_0)\, h(x_0)$$

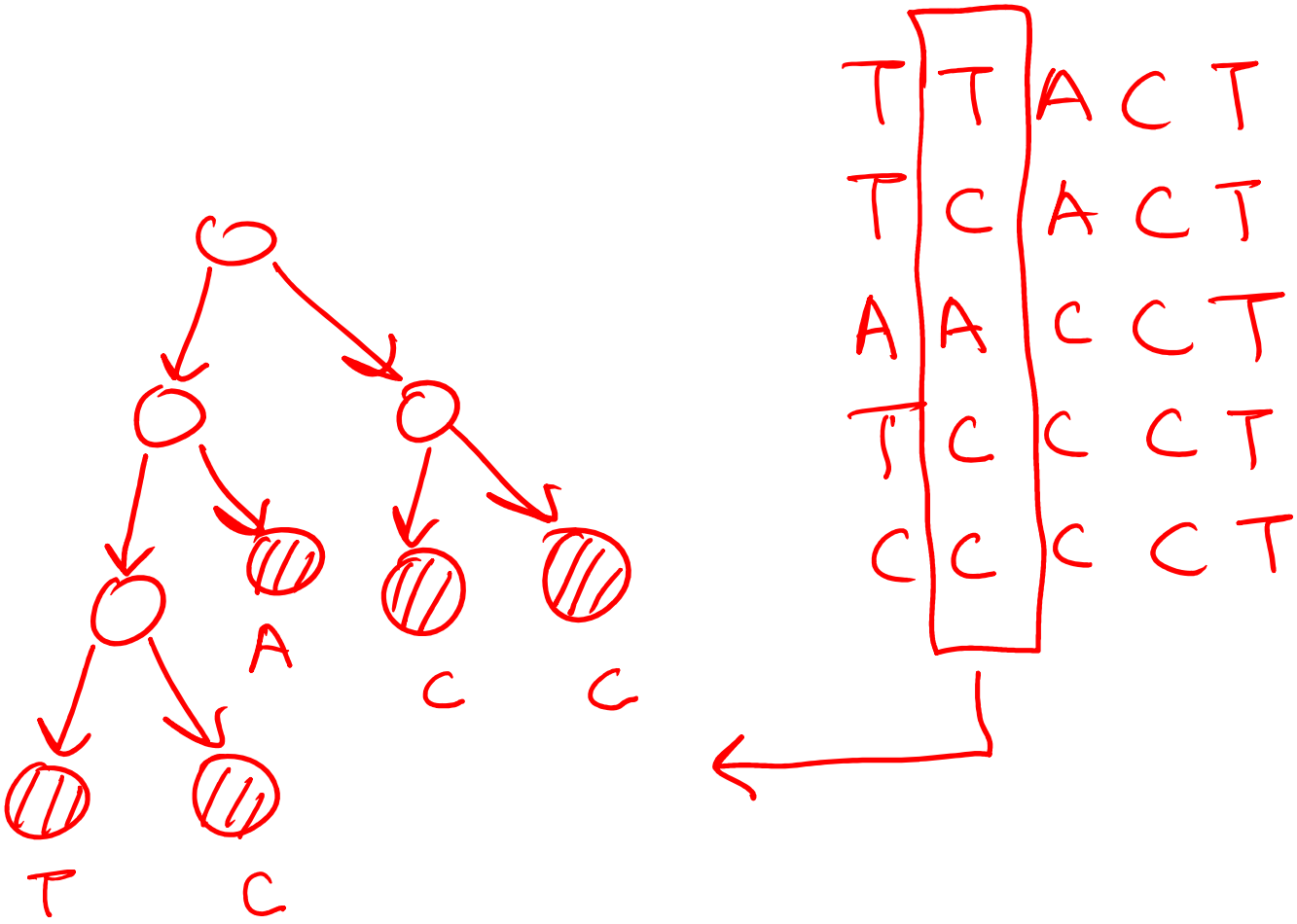


Handtracing the pruning algo

Ancestral likelihoods come free !

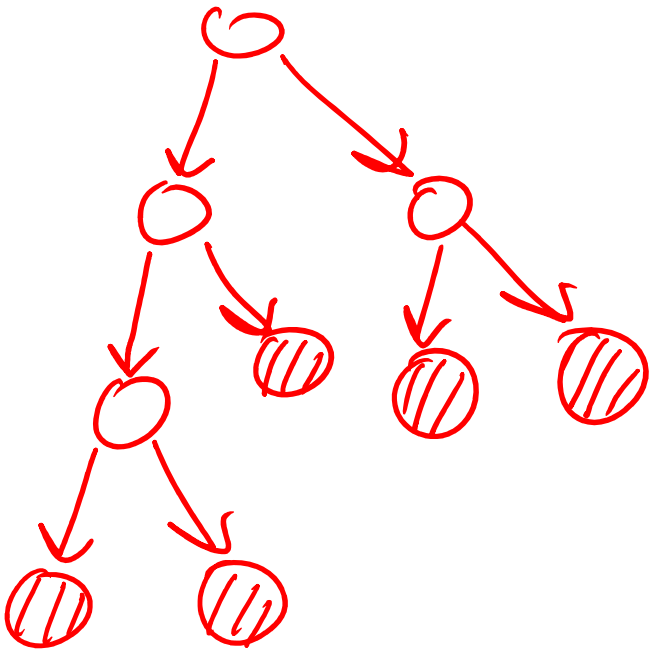


A recursive formulation



A recursive formulation

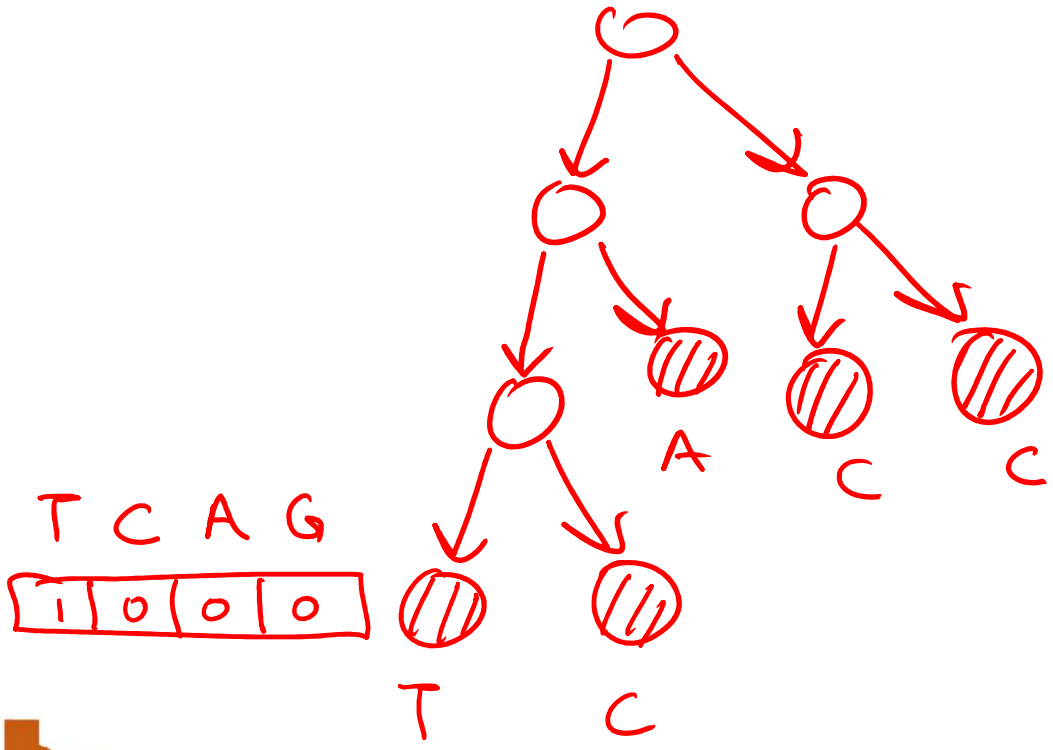
- At the observed leafs :



$$L_i(\alpha) = 1 \quad \text{if } x_i = \alpha$$

$$= 0 \quad \text{if } x_i \neq \alpha$$

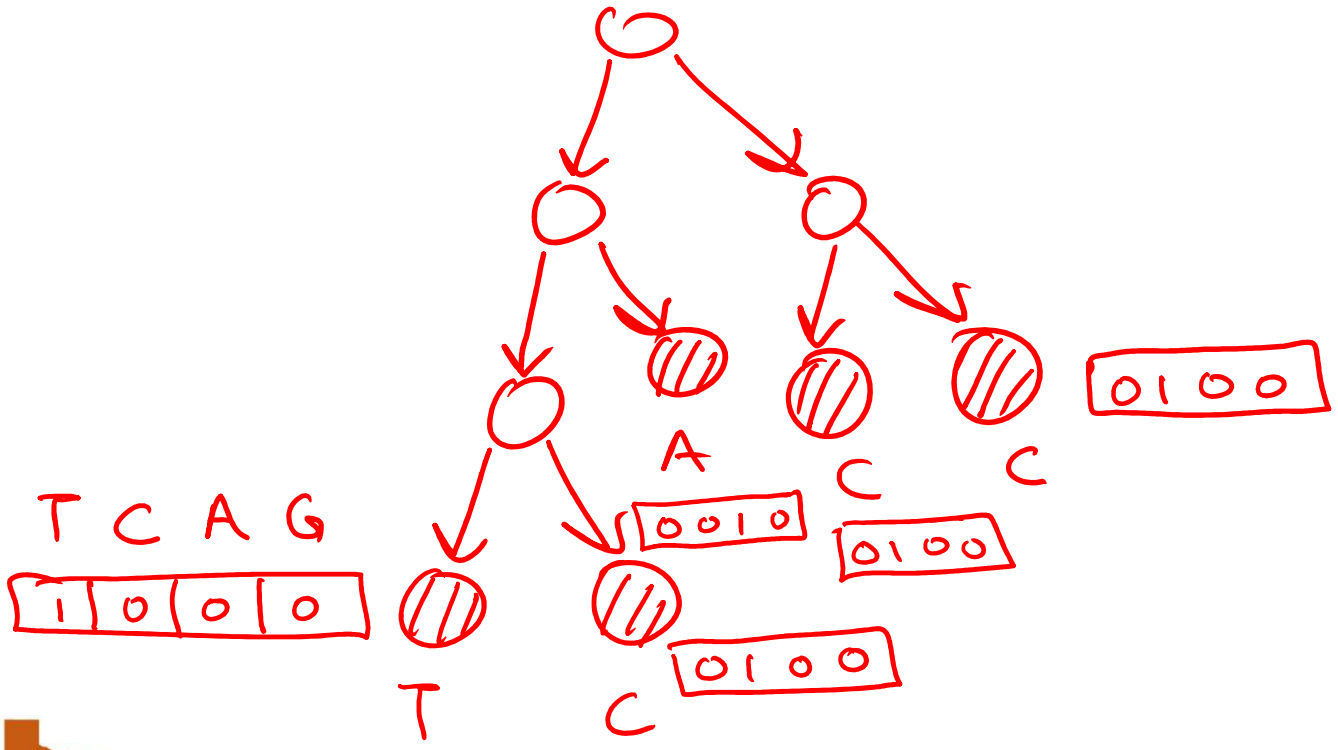
A recursive formulation



$$L_i(\alpha) = 1 \quad \text{if } x_i = \alpha$$

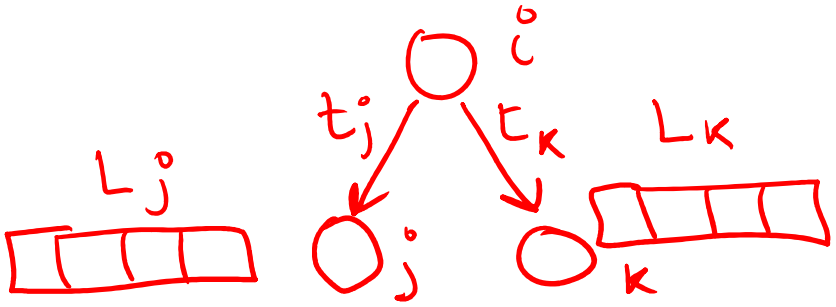
$$= 0 \quad \text{if } x_i \neq \alpha$$

A recursive formulation



A recursive formulation

- For the interior nodes :

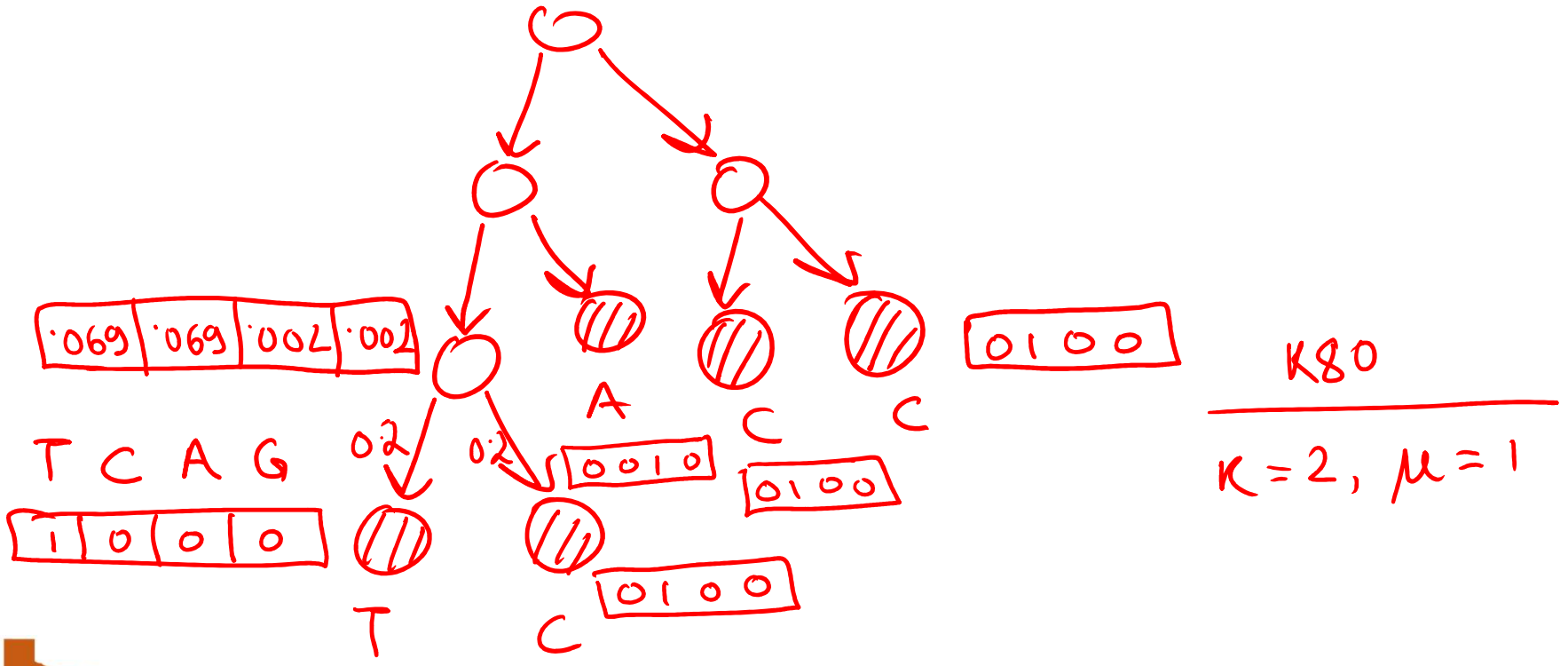


$$L_i(\alpha) = \left[\sum_{\beta} P(\alpha \to \beta \mid t_j, \theta) L_j(\beta)\right] \times \left[\sum_{\gamma} P(\alpha \to \gamma \mid t_k, \theta) L_k(\gamma)\right]$$

$\theta \rightarrow$ CTMP

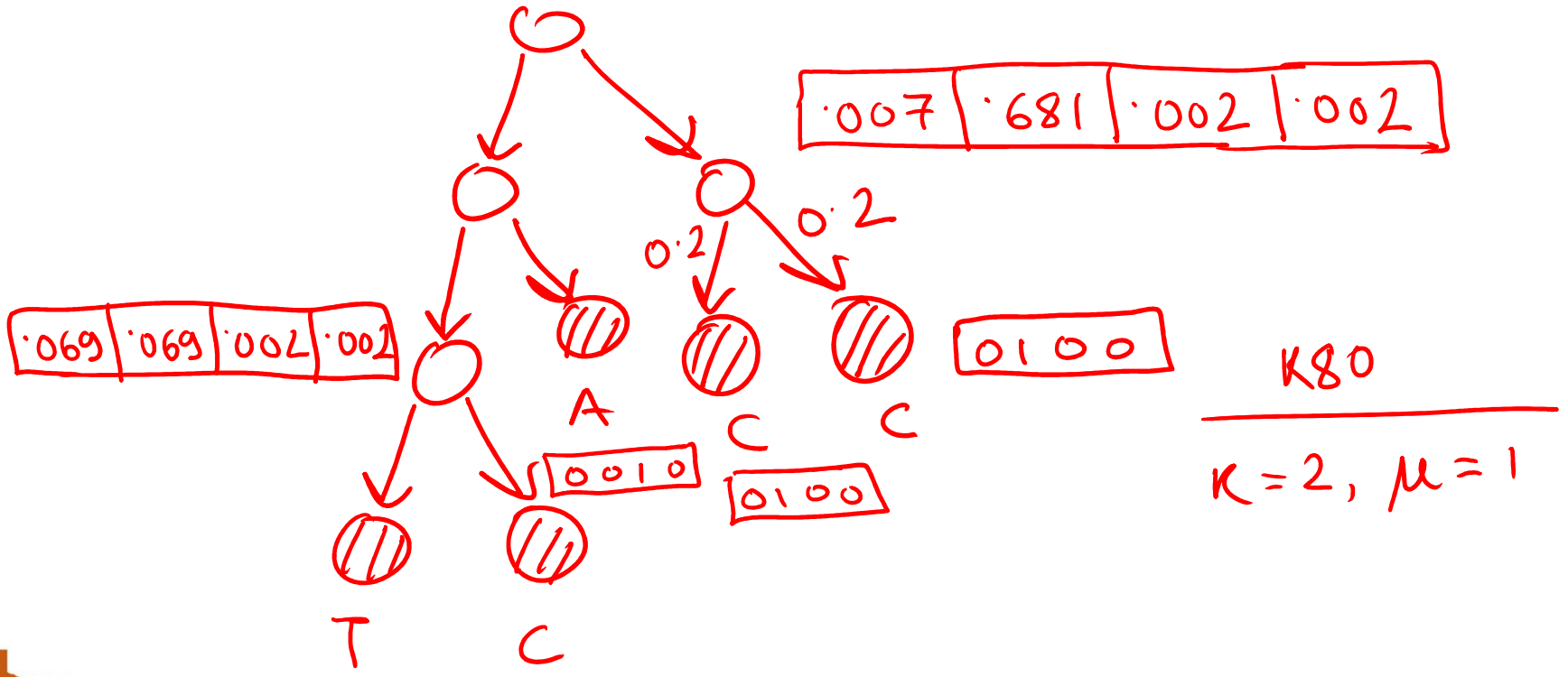
A recursive formulation

- Applied at X3



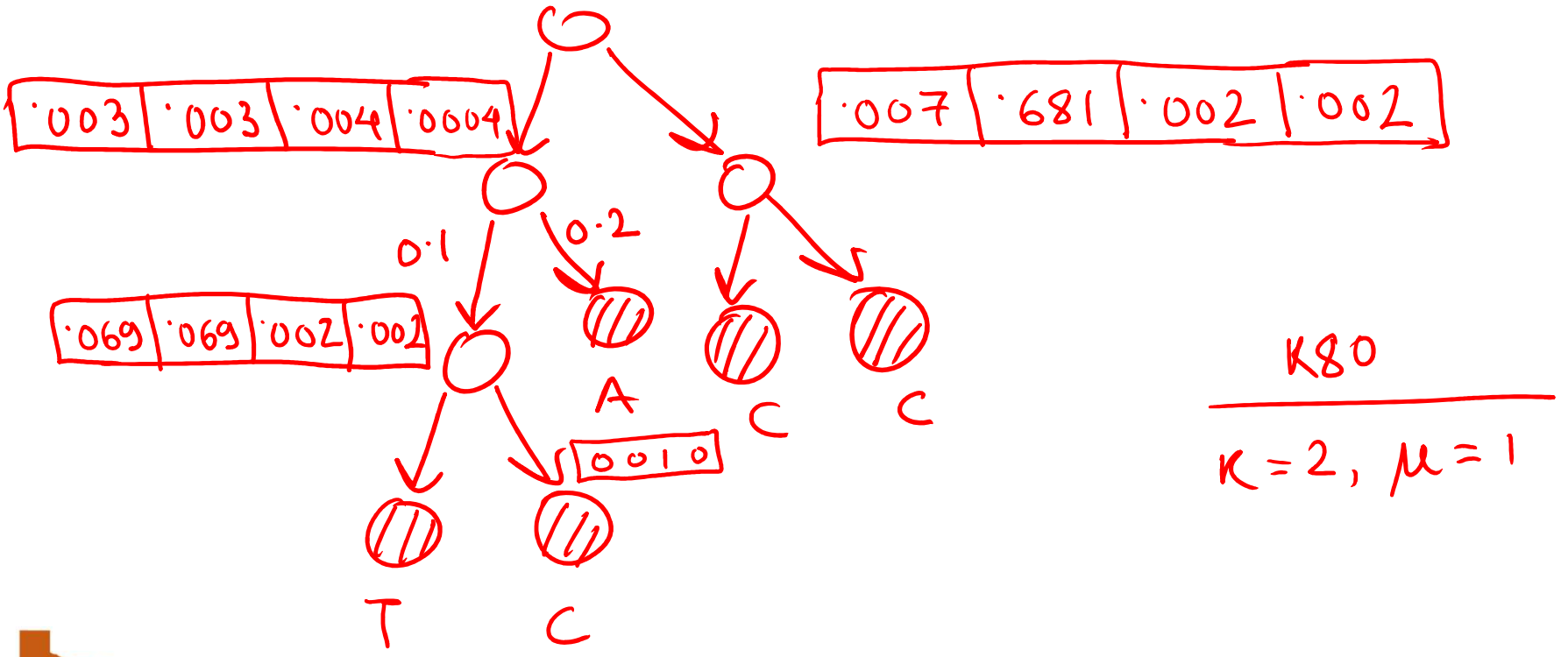
A recursive formulation

- Applied at X2



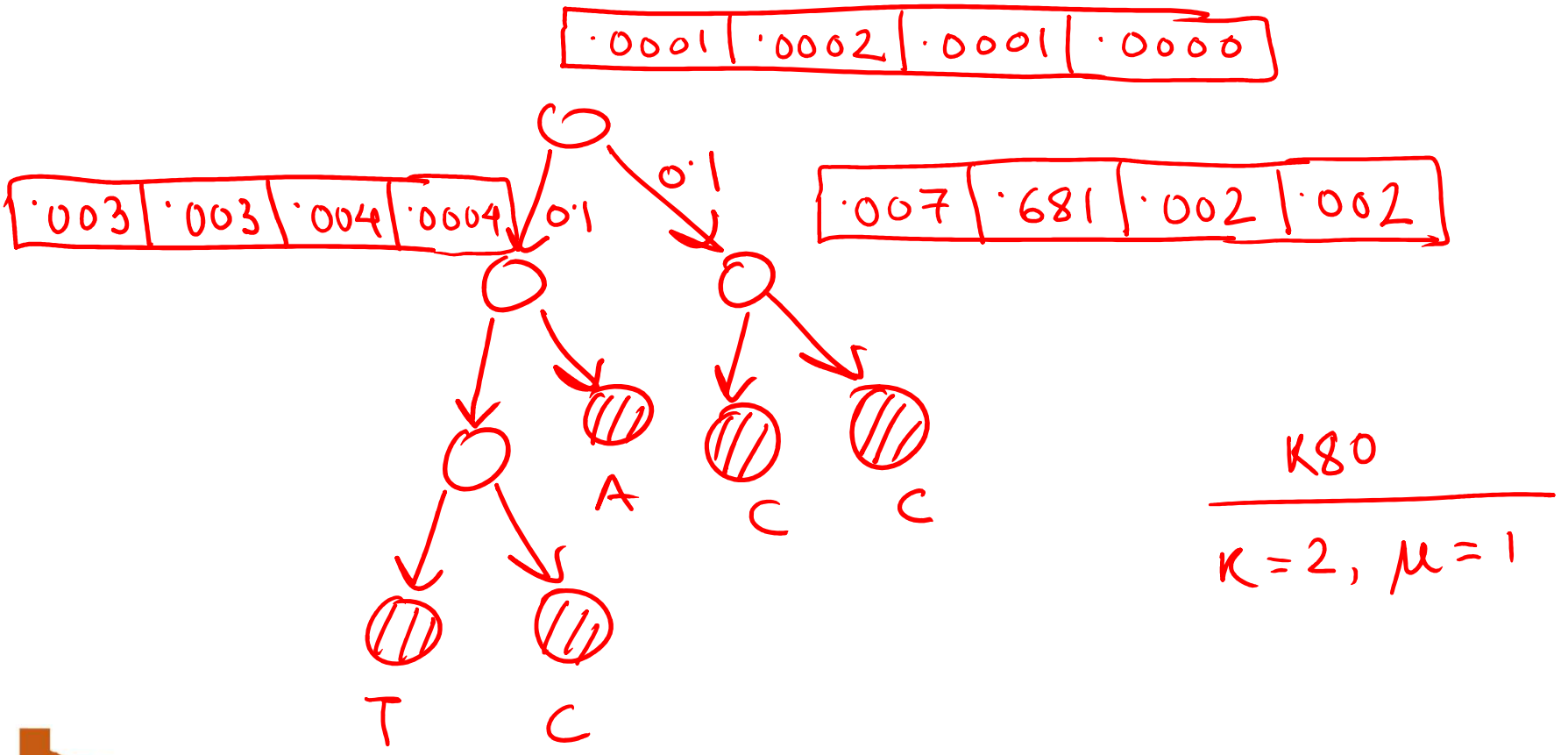
A recursive formulation

- Applied at X1



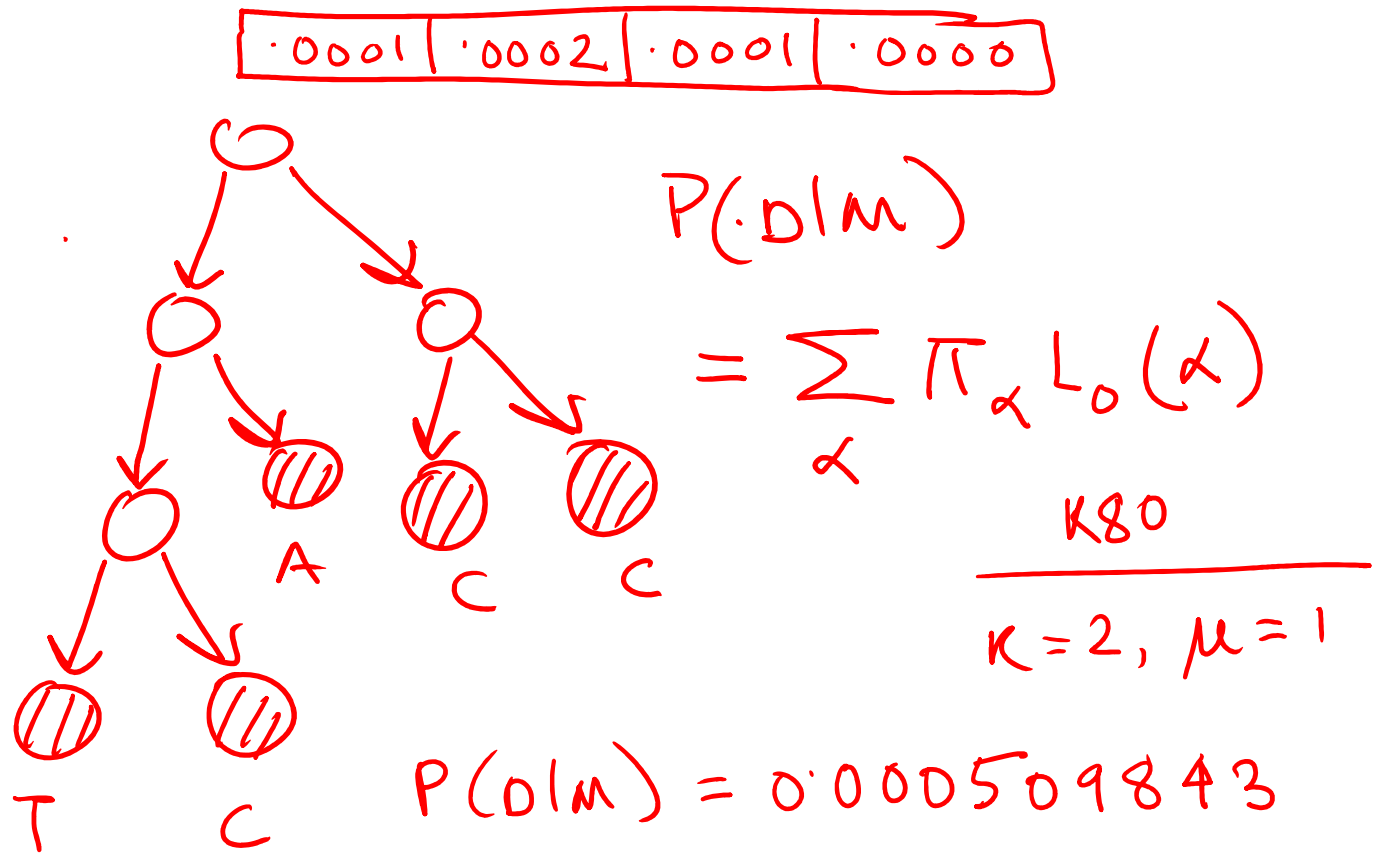
A recursive formulation

- Applied at X0 : the root. Where's P (D | M) ?



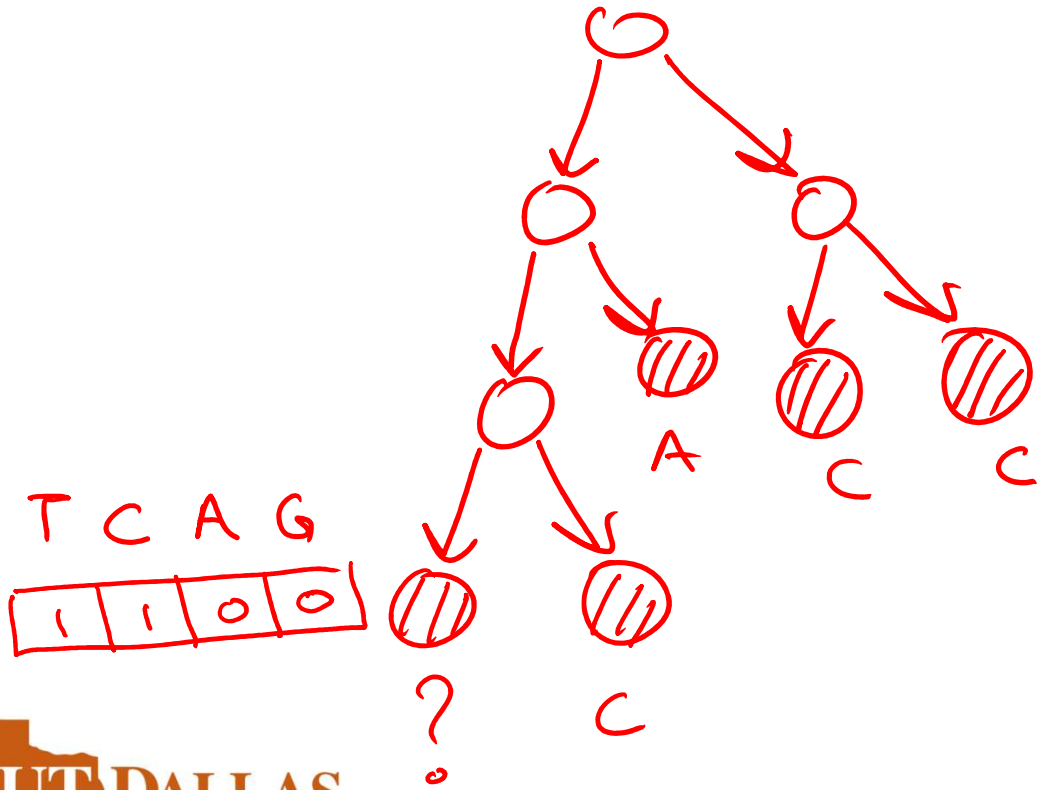
A recursive formulation

- At the root, factor in the stationary distr :



Modelling ambiguity

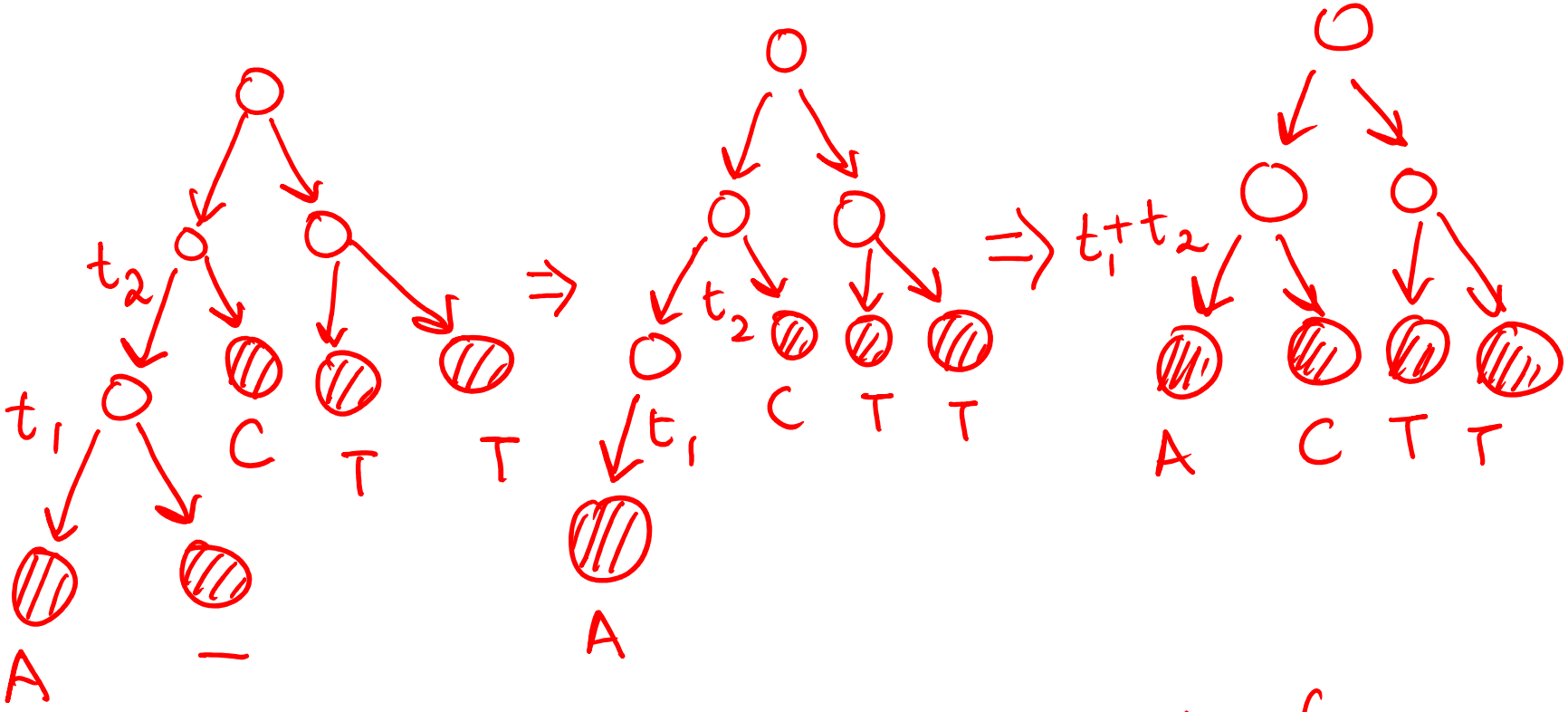
- Easy to model ambiguity at leaf



Handling gaps

- 3 unsatisfactory ways :
  - Throw away gapped columns (underestimate mutation rate; lose lot of data)
  - Treat the "-" as a fifth character / state in the stochastic process
    - Same framework used to model nucleotide change and indel creation (estimation is hard)
  - Treat gaps as hidden variables and marginalize
    - Commonly used (underestimates mutation rate)

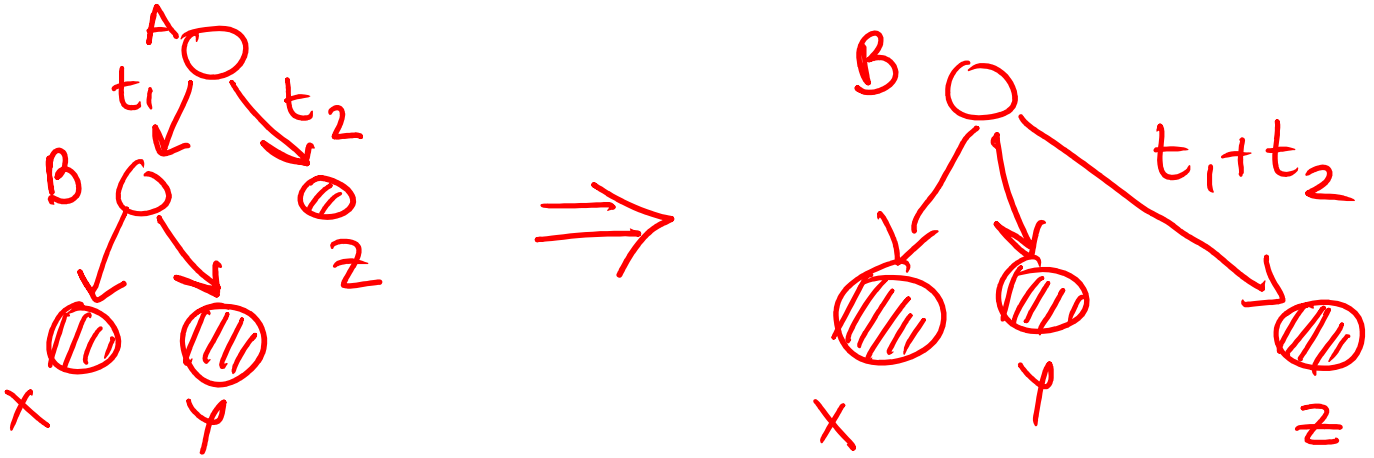
What happens when you marginalize a leaf ?



Or, put  at leaf

Predicting ancestral sequence

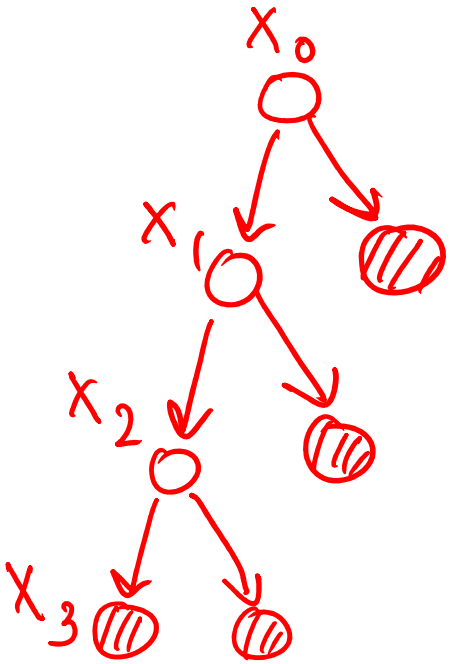
- Didn't we just do that ?
  - Only using data under that node, what if we want to use all the leaf nodes ?



  - Use a ternary tree !
  - Then pick nucleotide corr to max likelihood

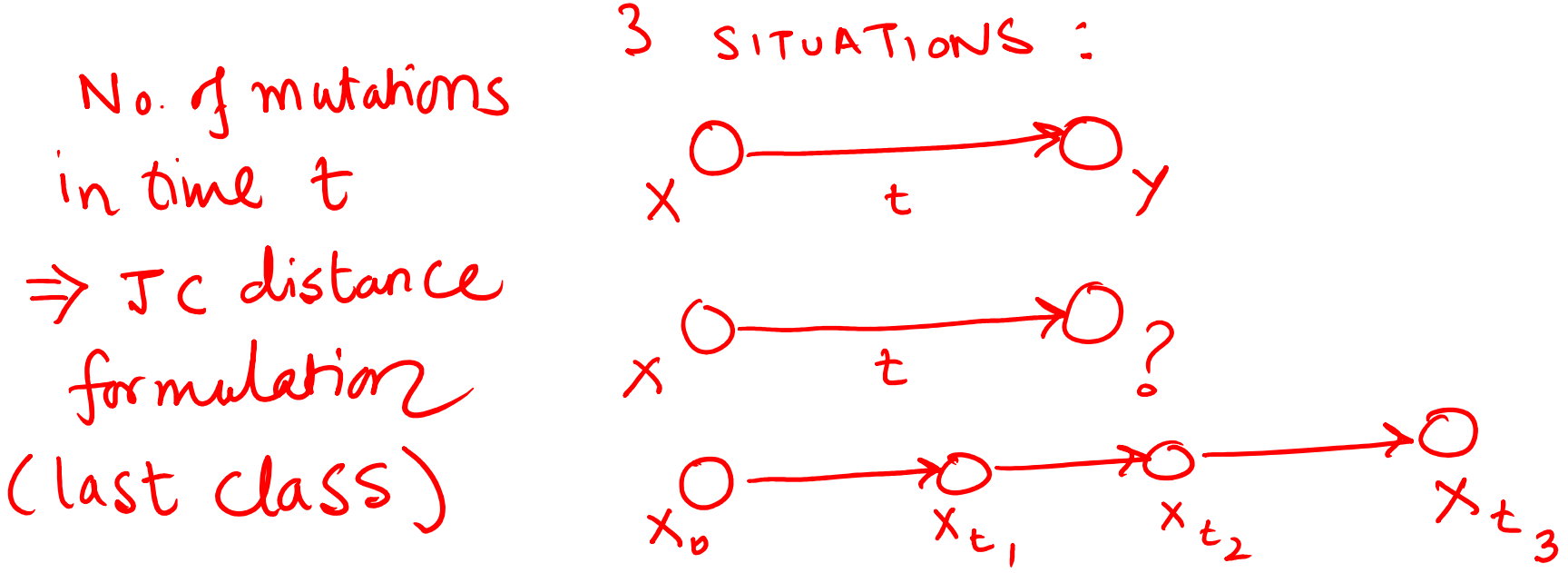
Joints and marginals

- Posterior decoding and Viterbi may give different results, remember ?



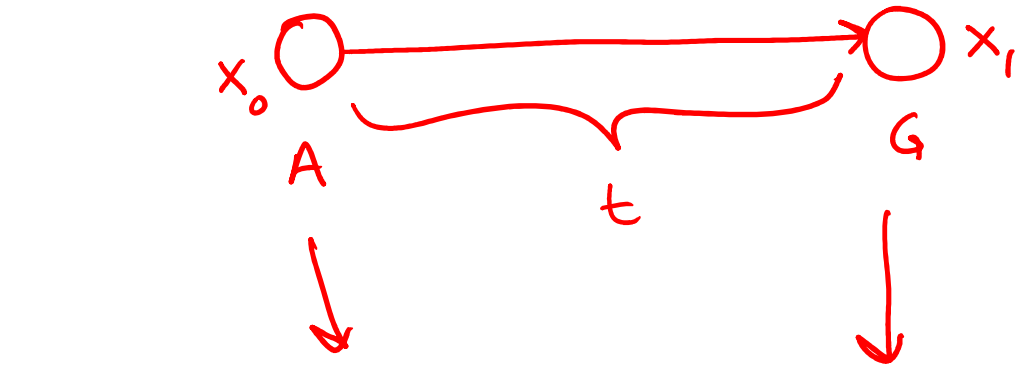
The most likely joint of (X_0, X_1, X_2) may be diff. from most likely X_0, X_1, & X_2 marginally

ML trajectories & no of mutations



- ML trajectories are sometimes well defined :
  - http://books.nips.cc/papers/files/nips22/NIPS2009_0822.pdf

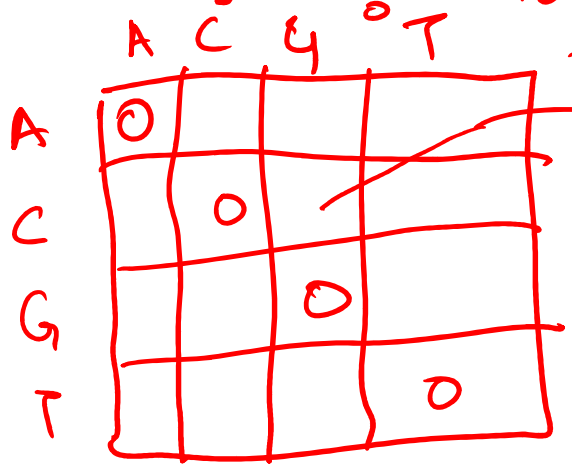
A simpler problem



WHAT IS A MOST LIKELY TO CHANGE TO?

WHAT IS G MOST LIKELY TO HAVE CHANGED FROM?

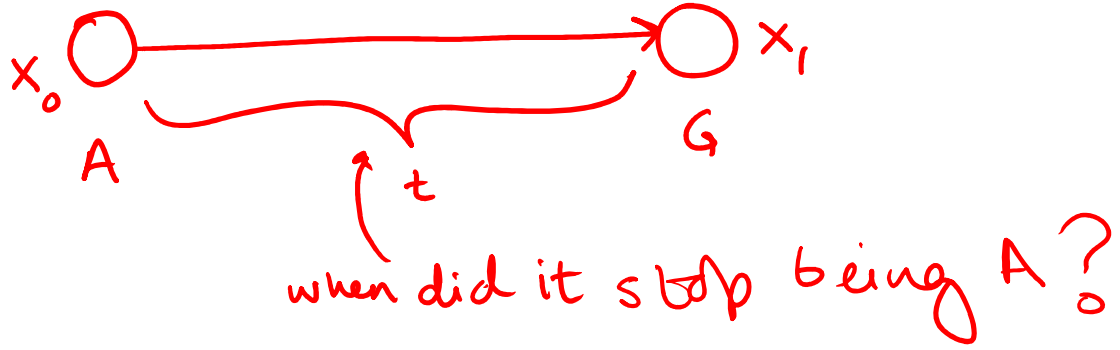
JUMP CHAIN



$$\frac{q_{ij}}{-q_{ii}}$$

WHAT IS THE ML PATH FOR 2 HOP MODEL fr. A TO G?

The road to simulation



WAIT TIME / HOLD TIME

$\exp(-q_{ii})$

← MEMORY LESS

$$\text{MEAN} = \frac{1}{-q_{ii}}$$

Speeding up calculations

- Minimize floating point operations (precision arithmetic)

- Pre calculate likelihoods on branches
- Pre calculate likelihoods on some patterns on subtrees
- Optimize ordering of sums

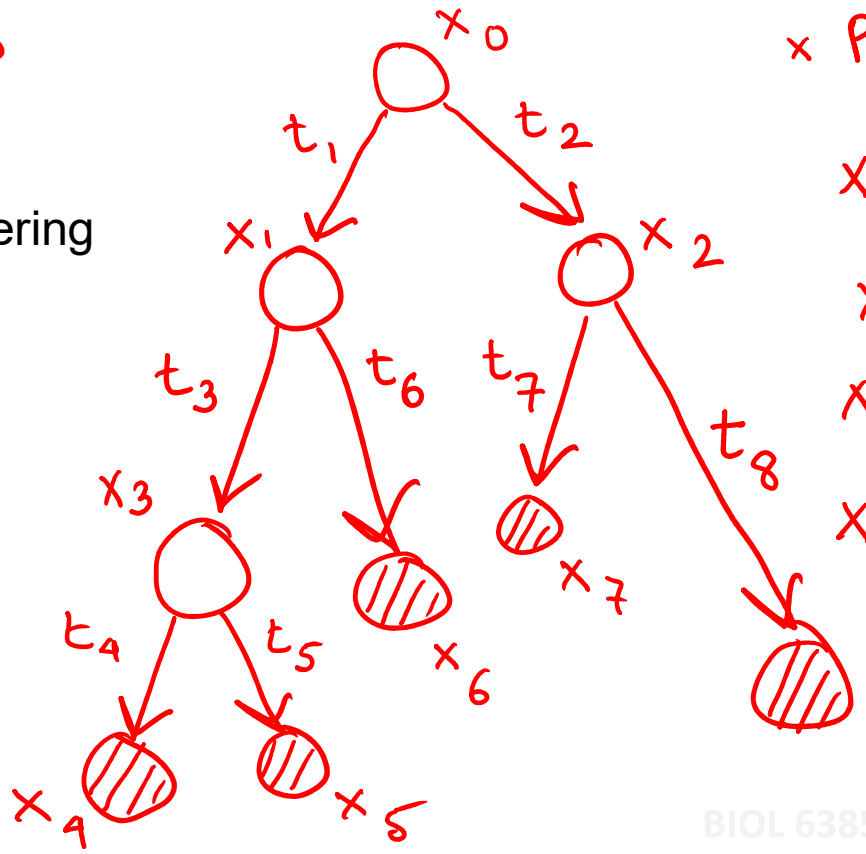
Speeding up calculations

$$P(x_4, x_5, x_6, x_7, x_8 \mid \theta, T)$$

$$= \sum_{x_0} \sum_{x_1} \sum_{x_2} \sum_{x_3} P(x_8 \mid x_2, \theta, t_8) P(x_7 \mid x_2, \theta, t_7) P(x_6 \mid x_1, \theta, t_6)$$

$$\times P(x_5 \mid x_3, \theta, t_5) \times P(x_4 \mid x_3, \theta, t_4) \times P(x_3 \mid x_1, \theta, t_3) \times P(x_2 \mid x_0, \theta, t_2) \times P(x_1 \mid x_0, \theta, t_1) \times P(x_0 \mid \theta)$$

What is the optimal ordering of the sums ?



Likelihood of full alignment

- Likelihood of a multiple sites
  - Simplest model : independent sites, single model

$$P(A \mid M) = \prod_i P(A_i \mid M)$$

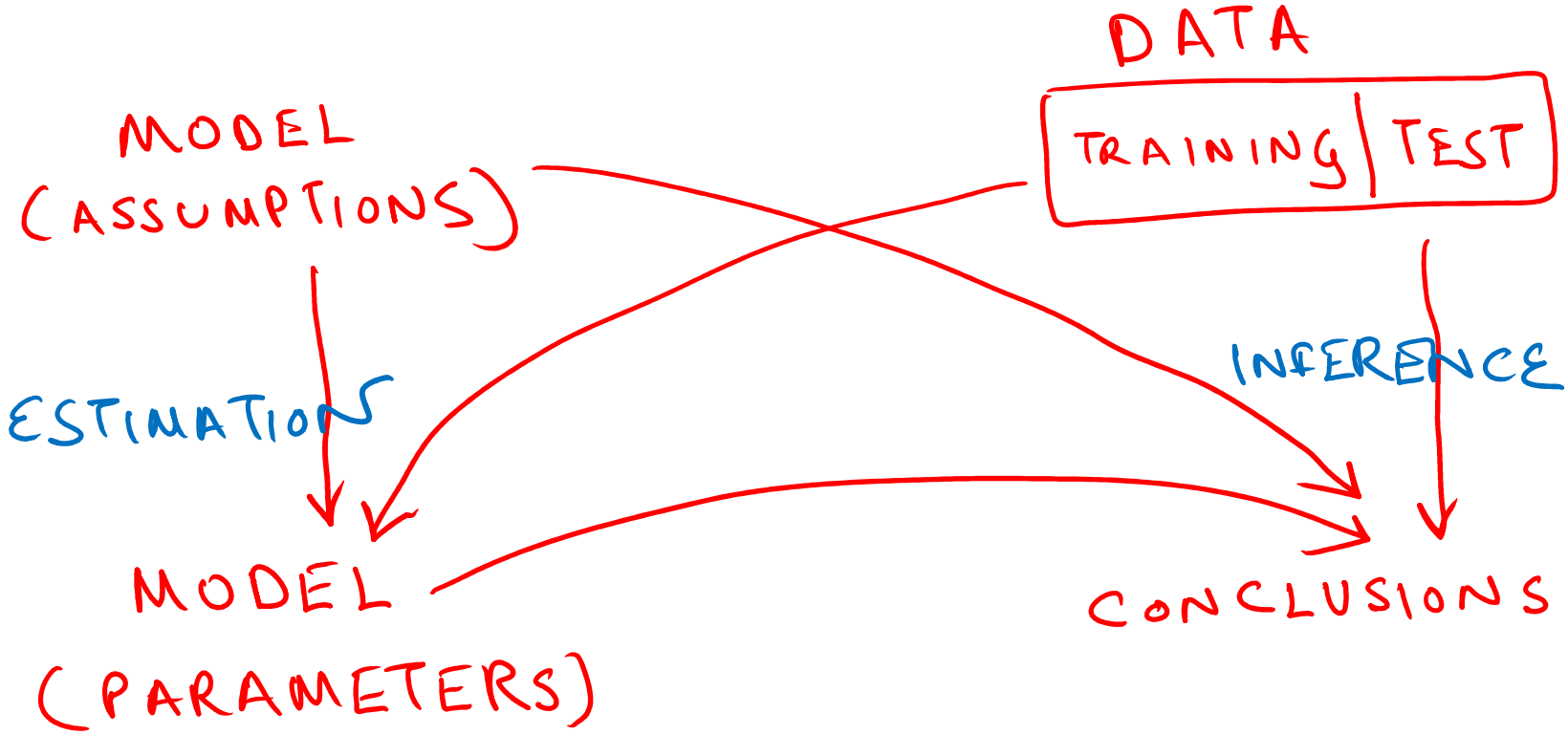
↳ PRUNING ALGO. ON ONE ALIGNED SITE AT A TIME

  - Independent sites, diff known models

$$= \prod_i P(A_i \mid M_i)$$

WHAT IF M_i'S ARE UNKNOWN? COMING SOON...

Estimation & inference



Estimation

- Given the data
  - Generate each possible tree
  - Score each tree with the model
  - Pick the tree whose “score” is the “best”
    - Score for probabilistic models = Likelihood = P (data | model)
    - Best score for probabilistic models = Highest likelihood

For maximum likelihood estimation

- Given the data
  - Generate each possible tree
  - Find P (data | model) :likelihd fn: for each tree
  - Pick the tree with highest likelihd fn
  - Typically find $\frac{\partial}{\partial\theta}(L(\theta)) = \frac{\partial}{\partial\theta}(P(D|\theta)) = 0$

    for max likelihood parameters

One snag

Topology space is discrete

How to take the derivatives ?

Instead …

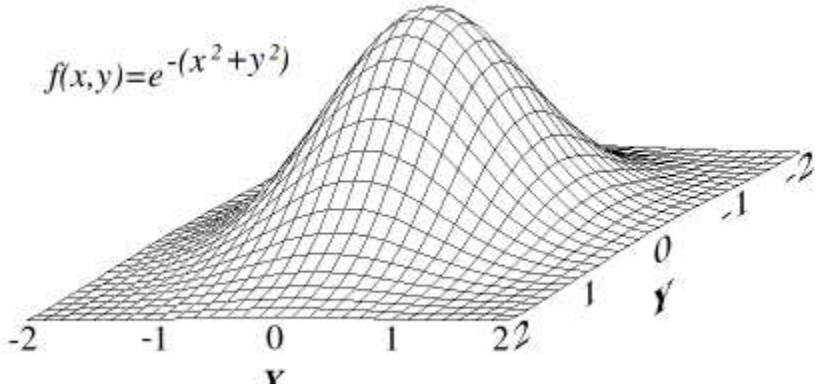
- Let us assume for now that the topology is known, and we want to optimize the CTMP parameters and branch lengths

- Multivariate optimization
  - should we change them one at a time ? Or all at once ?

- Either way, we'll need to try many parameter values : reason why closed form CTMP prob are reqd

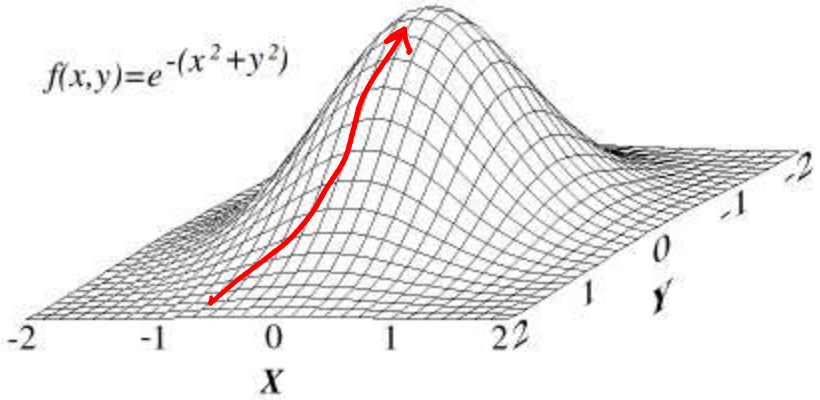
Line search and grid search

- Univariate optimization
  - sample along the single dimension at regular intervals to pick highest / lowest scoring point
- Multivariate optimization
  - sample along all dimensions at regular intervals to pick highest / lowest scoring point
    - Curse of dimensionality



Smarter way : gradient ascent

- Start in one position
  - move in direction of steepest upward likelihood gradient



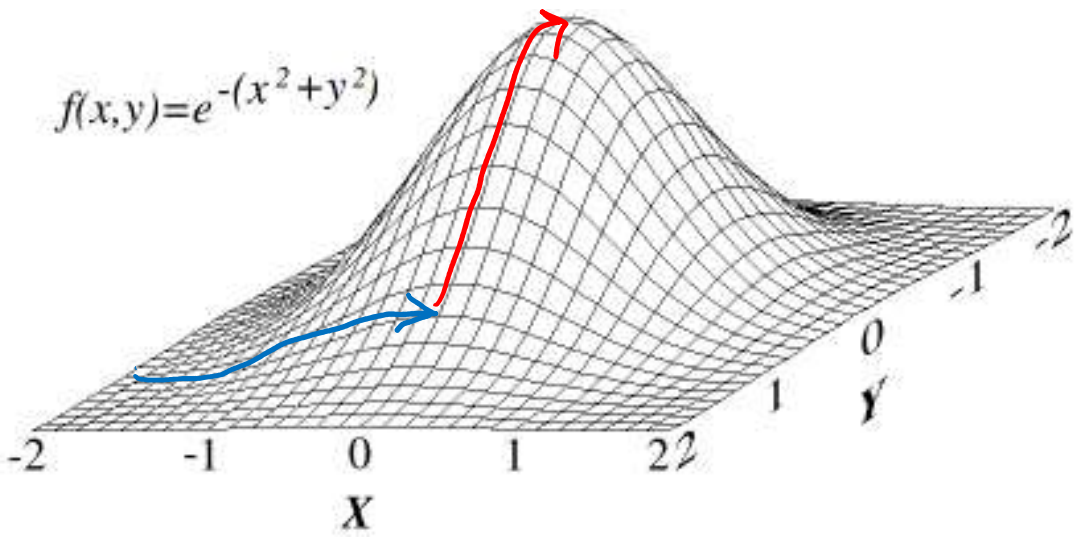
Absolute astronomy

Optimization: vast mathematical body of work

Basic underpinning : to reach the top of the hill quickly, climb the slope in the steepest direction always (or almost always)

Gradient ascent

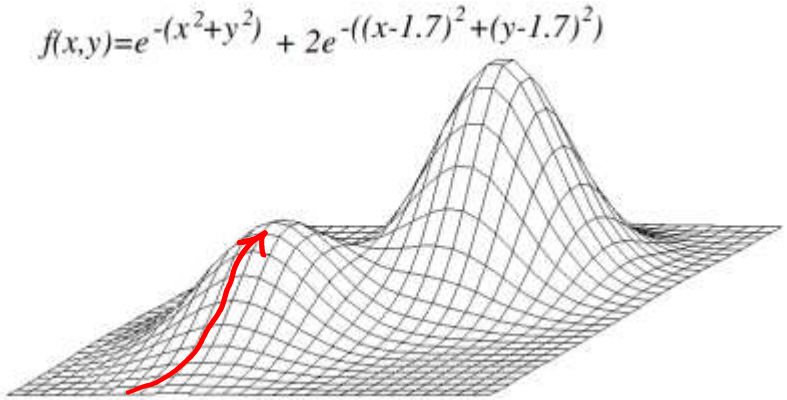
- Optimizing one variable at a time



Absolute astronomy

The curse of gradient ascent

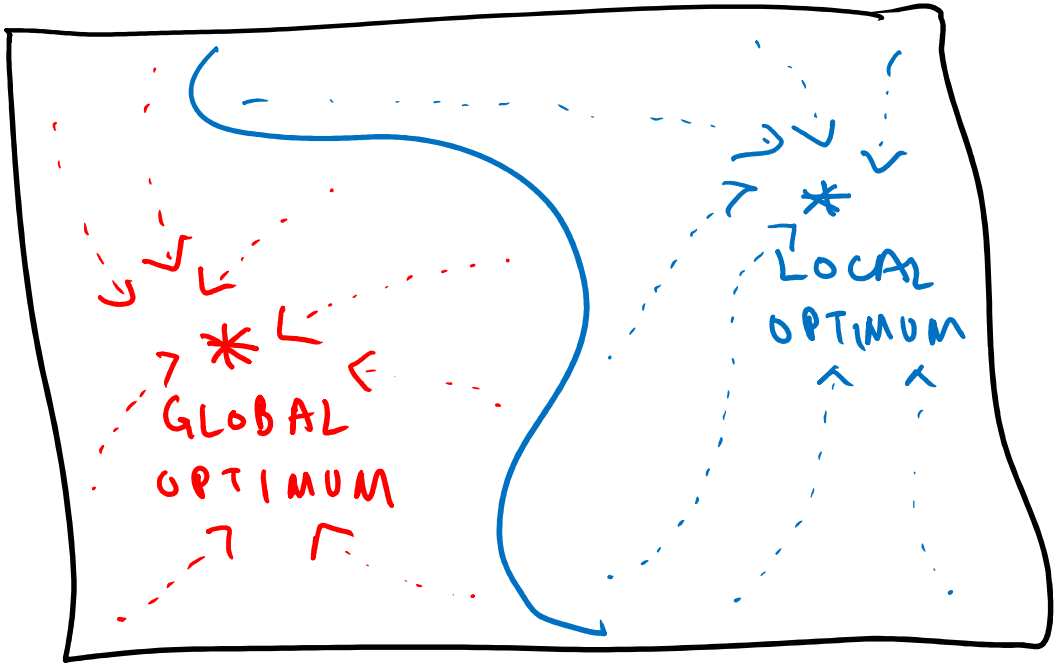
- Local maxima



Absolute astronomy

Orbits and attractors

- Search procedure = dynamical system



Why not …

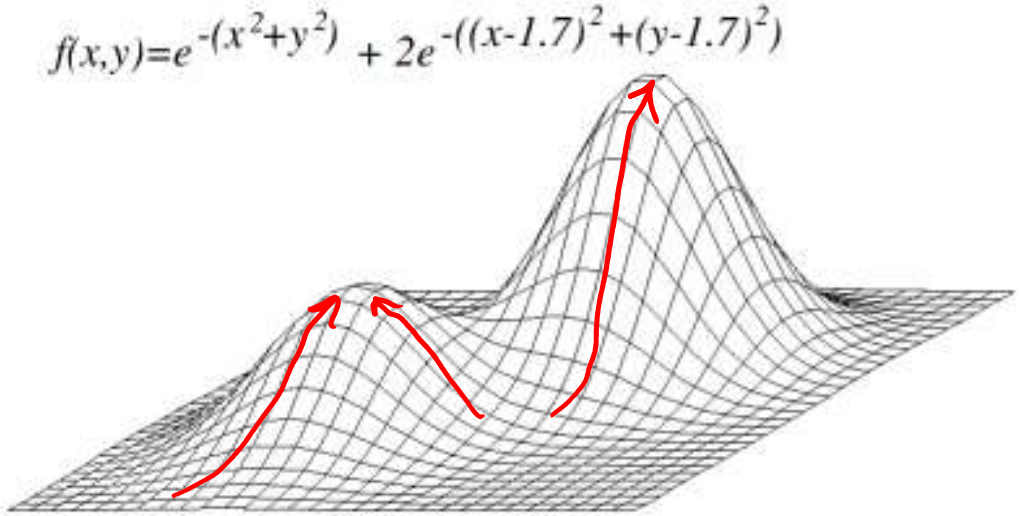
- Why not figure out the boundaries of the attractor initially and dispense with the iterations ?

Why not …

- Why not figure out the boundaries of the attractor initially and dispense with the iterations ?
  - Because the likelihood landscape will change with the data ! And so will the attractors, and their boundaries !

Avoiding local maxima

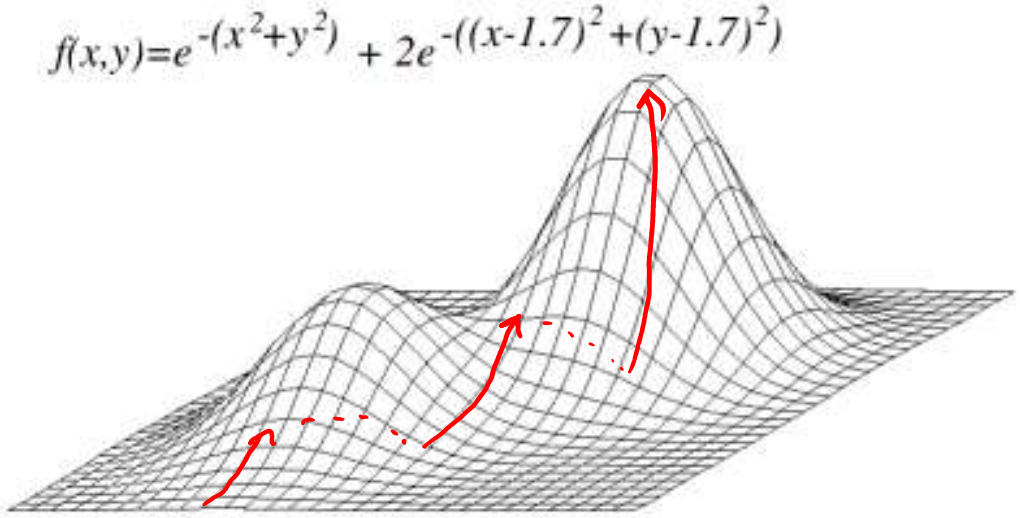
- Random restarts : run repeatedly with different initial guesses : sooner or later sample all attractors



Avoiding local maxima

- Simulated annealing : Occasionally break the rules, and jump around in the space randomly : do this less often as search progresses and score improves

Jumps should be large enough to switch attractors , or frequent enough (in the beginning) to incrementally move from one attractor to another



Simulated annealing + random restarts possible

What are our dimensions ?

- Branch lengths for each branch
- Stochastic process parameters

- No meaningful way to reduce dimensionality easily

Learning topology

Problem of structure learning

difficult and intractable problem

- discrete space
- difficult to parameterize

Learning topology

- Given the data
  - Generate each possible tree topology
  - Optimize branch lengths and CTMP parameters such that P(data|model) for that topology is the highest
  - Pick the topology – parameter combination s.t. likelihood is highest

Can we really generate all topologies?

- No, but formal search strategies
  - Heuristic, but in general the longer we search the better the chance we find the global optimum
  - Better heuristics = less time to generate results of some determined quality level (model score)
  - Better heuristics = better quality results (model score) after searching for a fixed amount of time
  - Trade - off

Our search space

- Is it the space of all trees ?
  - Yes, but we have tools to optimize branch lengths, and evolutionary parameters efficiently
- Is it the space of all tree topologies ?
  - Yes, we heuristically sample tree topologies, since discrete topology – space is hard to optimize numerically

Random sampling

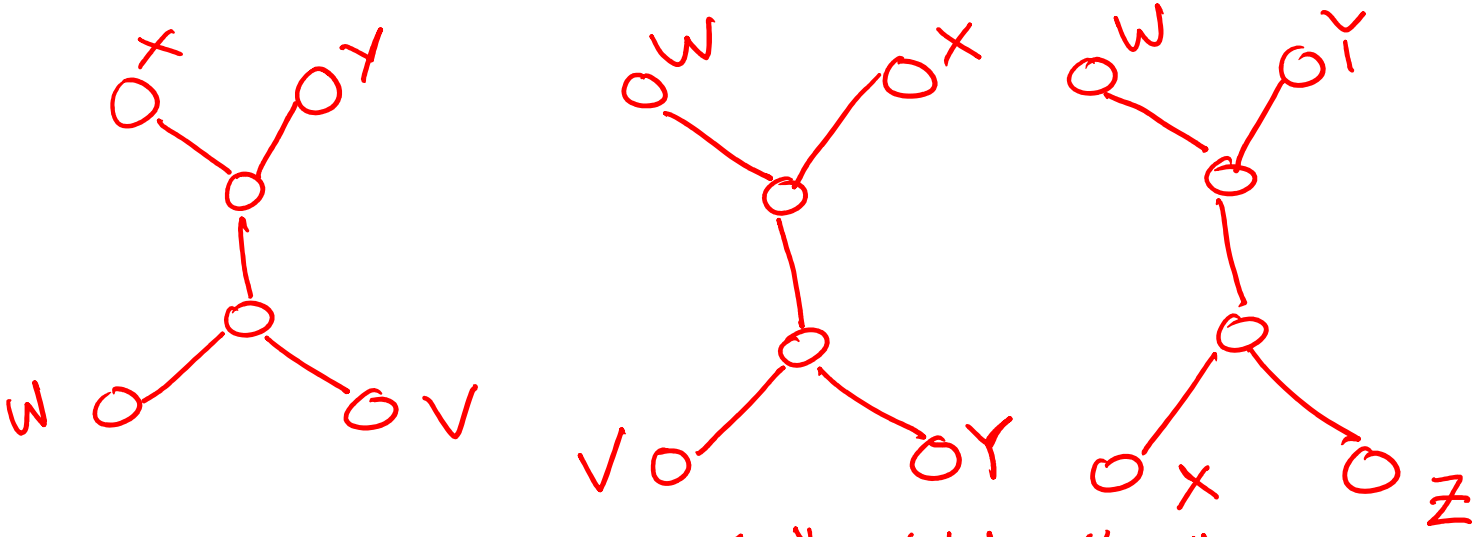
- Keep generating new topologies randomly and score them: remember the best score

- Doomed to repeat mistakes + most work is thrown away
  - goal should be to improve upon previously found high scoring topologies by incrementally changing the topology : discipline of search

Hill climbing

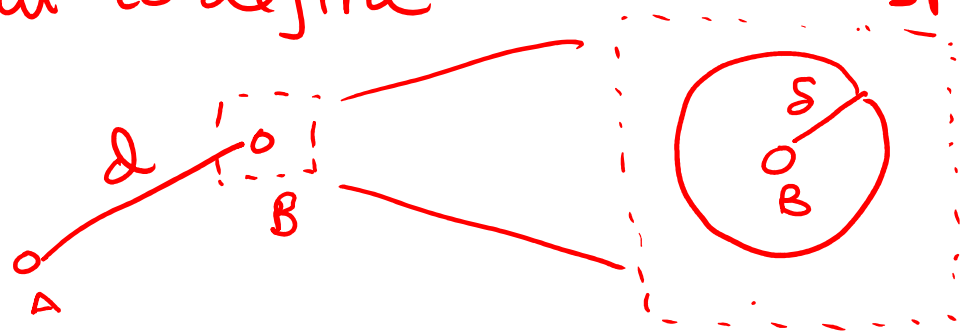
- Similar to gradient ascent : for discrete spaces
- At any point in topology-space, check the max likelihood score of all neighbors and move in direction of biggest increase of likelihood

Phylogeny topology - space



Concept of "neighbor" & "distance" (metric space) difficult to define

consider 2D - euclidean space :



δ is small

NEIGHBOR

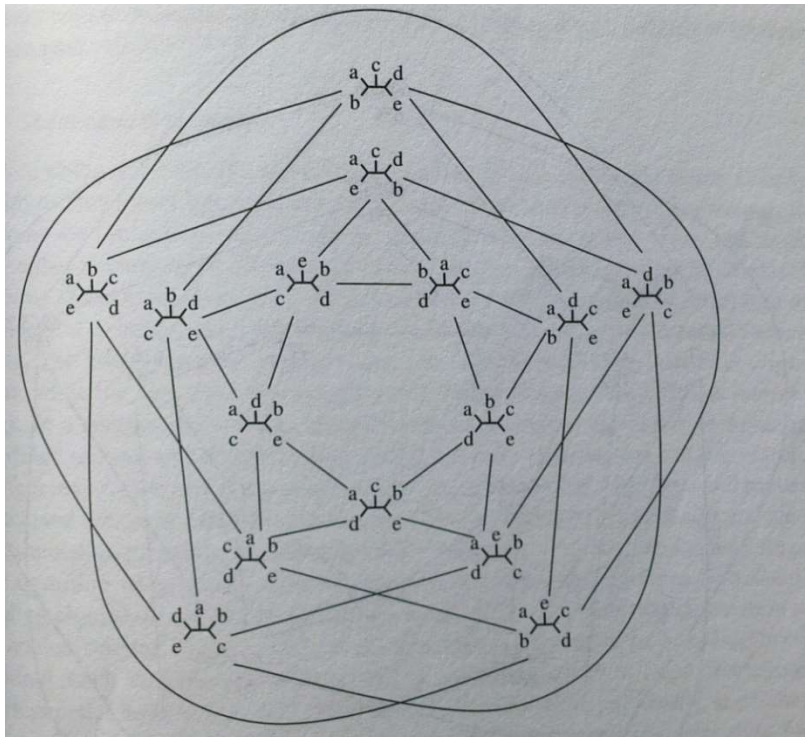
DISTANCE $\overline{AB} = \sqrt{(y_2-y_1)^2+(x_2-x_1)^2}$

Wait, didn't we already study distances in phylogeny ?

- Over taxa, yes
- Over trees, no

What does tree space look like ?

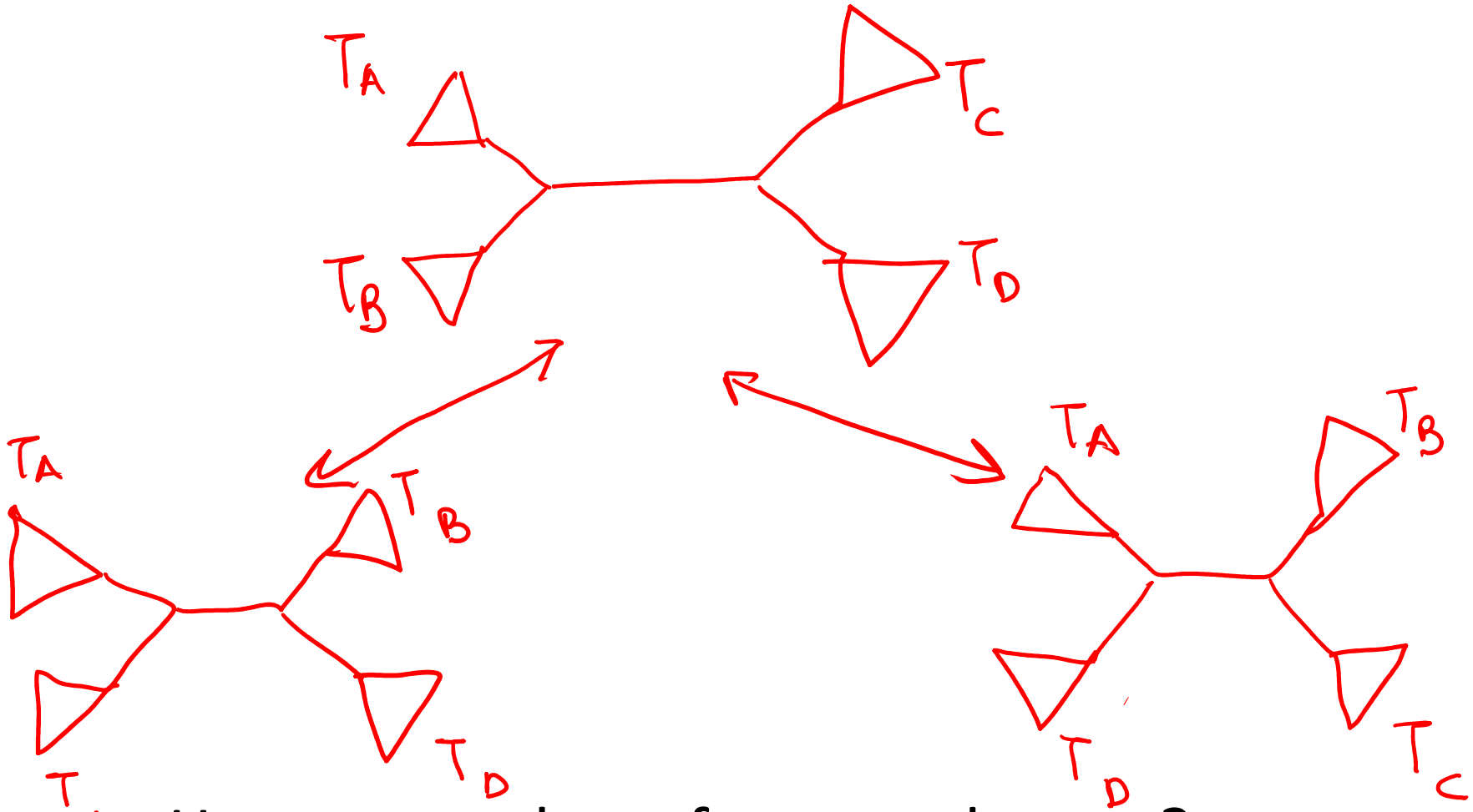
- Edges represent neighbors (wait, who are neighbors?)



Z Yang

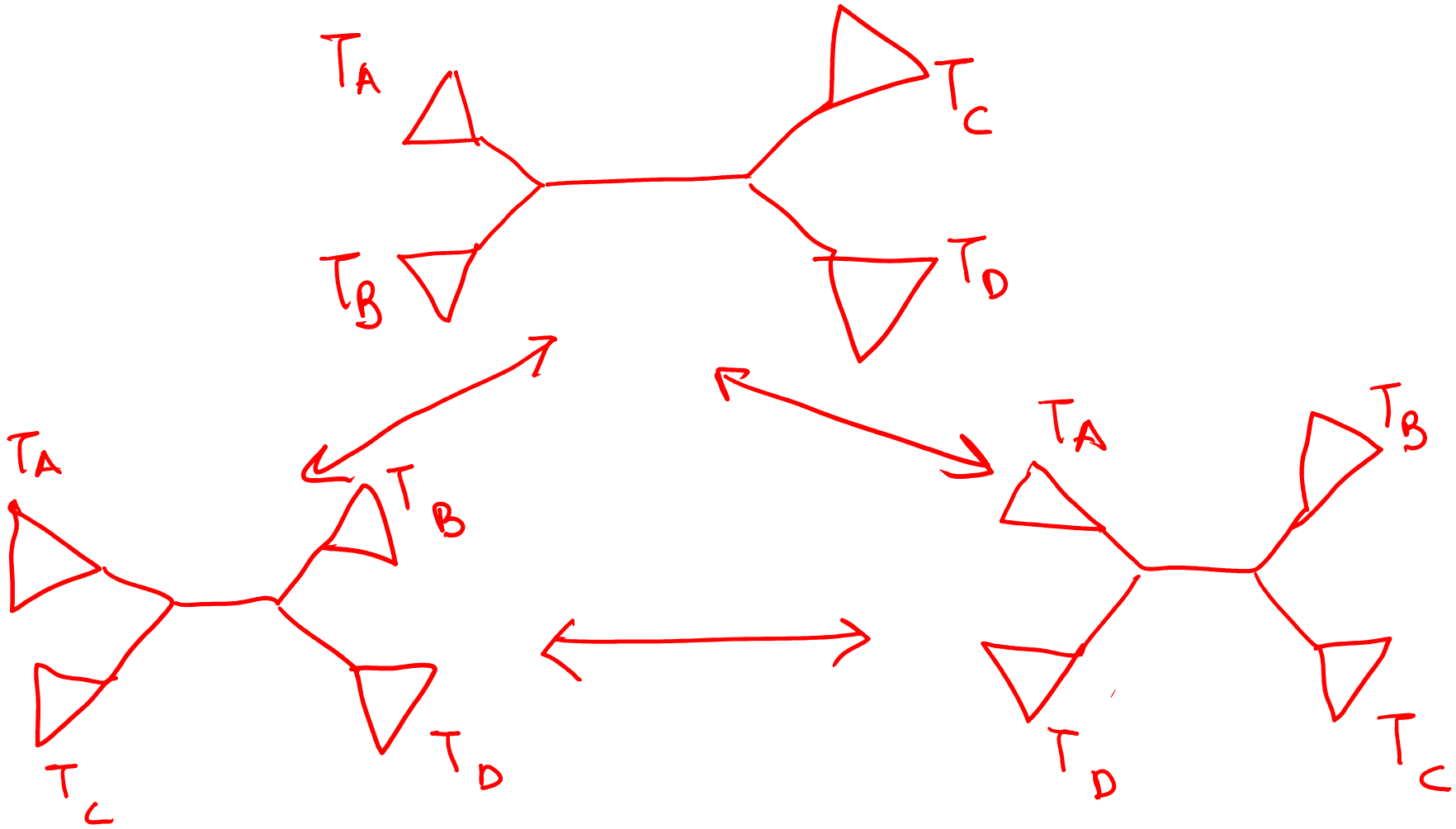
- x is a neighbor of y = y can be built by changing x in a small way

Nearest neighbor interchange

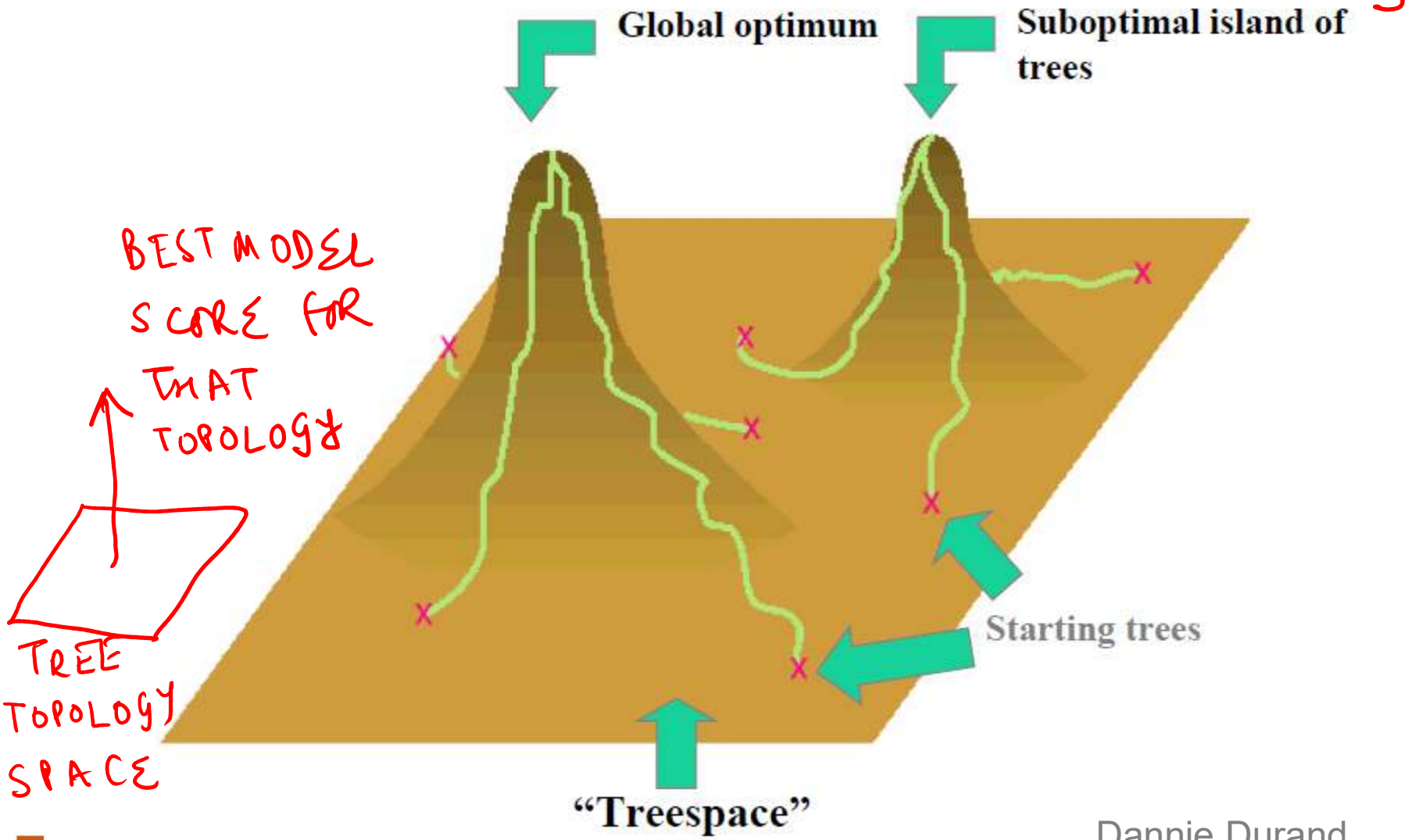


- How to use these for rooted trees ?
  - Use the unrooted equivalent and convert back

In fact …



Heuristic search: HILL CLIMBING



Local and global optima in topology - space

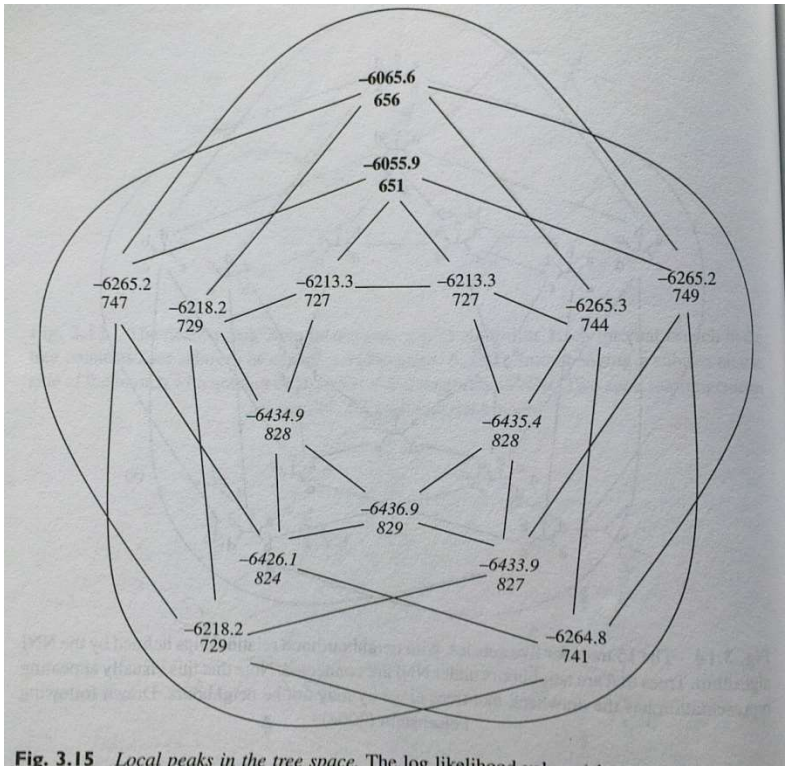


Fig. 3.15 *Local peaks in the tree space.* The log-likelihood

Z Yang

Traditional AI search

… techniques (like admissible heuristics and A*) doesn't work

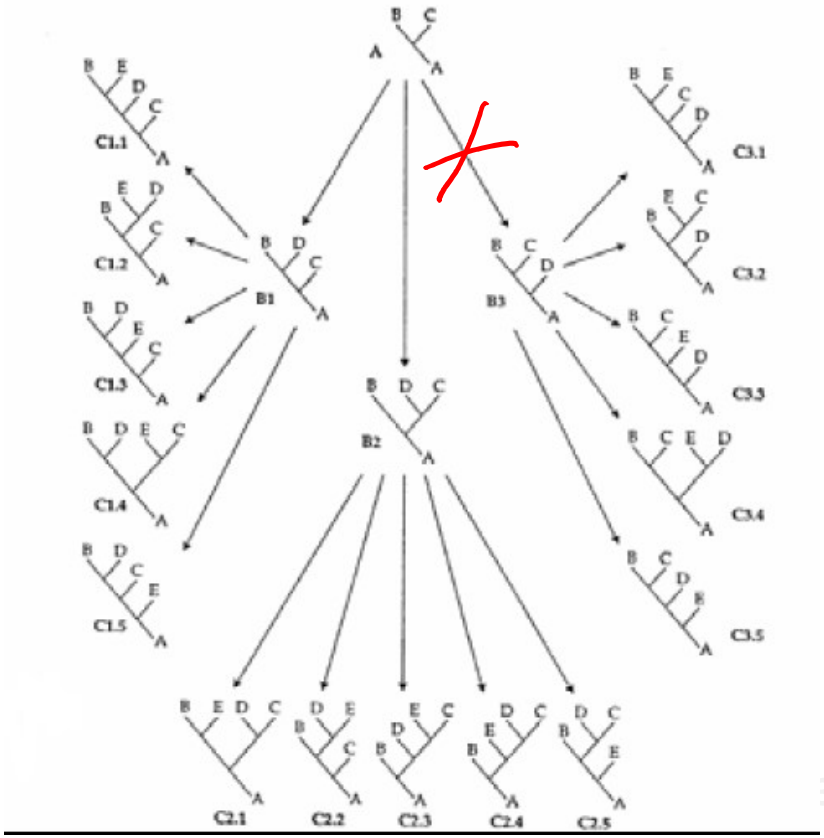
Difficult to estimate how far away we are from our goal (best model score)

Branch and bound

- Rule out ("prune") or de prioritize some parts of the search space : similar to MSA ?

Wait, this isnt our tree space !

This is okay, since each point in our original space is reachable in this space !



In practice …

- Instead of a formal search, we often use :
  - known tree topologies
    - based on trees constructed using distance based methods

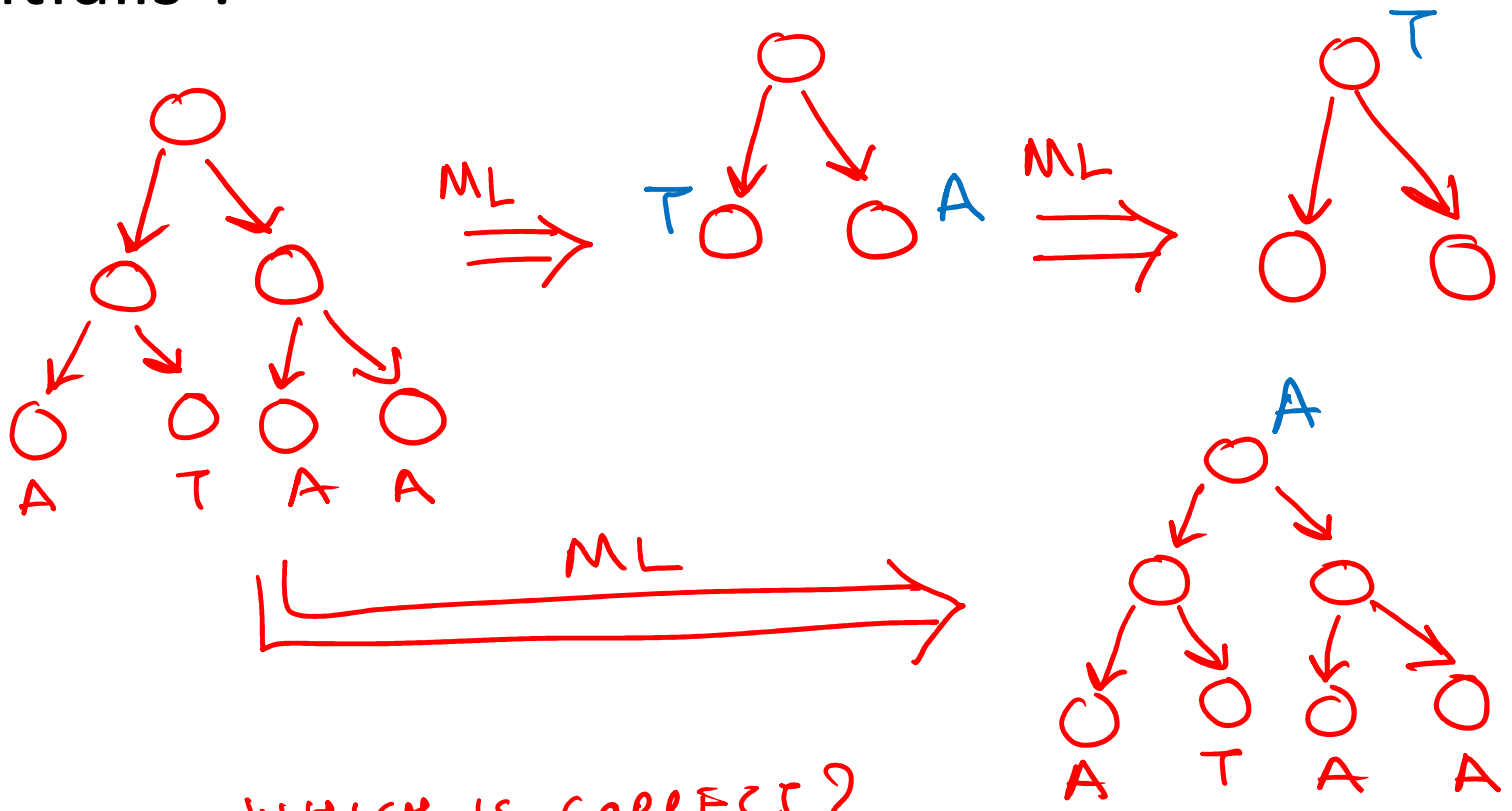
- Consider this as a very strong prior over the topologies

Search : take home

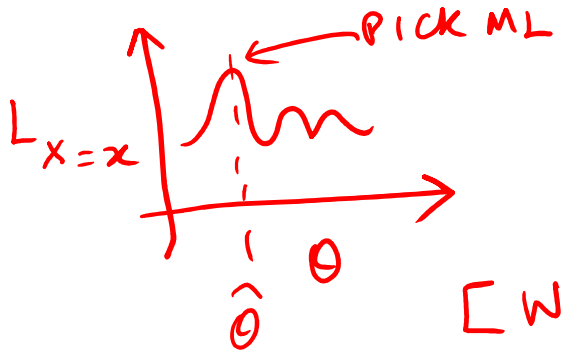
- Search longer = better results
- Better heuristics & better strategy = better results
- Usually no guarantees = may get stuck in local optimum
- All points in search space should be potentially reachable by our strategy

Using inferred results

- In the next step of an analysis : what are the pitfalls ?



Ball of uncertainty



WHAT IS THE CHANCE THAT WE PICKED THE RIGHT θ?

[WHOLE THEORY: PAC-LEARNING]

- EVERYTIME WE PICK A SINGLE θ, WE "FORGET" SOME DATA
- ERRORS ACCUMULATE
- BETTER IDEA: TREAT θ AS A R.V.
  - FIND ITS DISTRIBUTION

Bayesians and frequentists

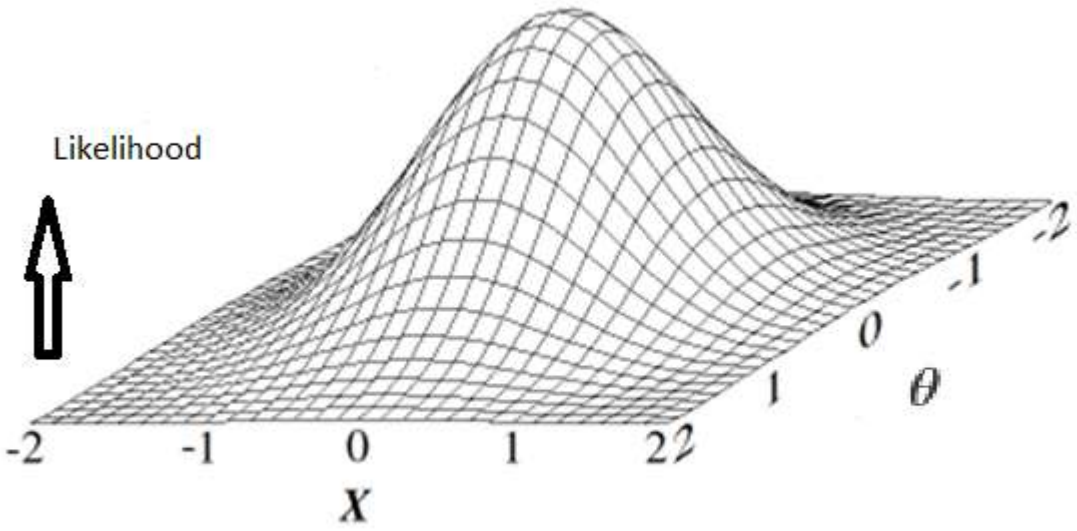
- Frequentists
  - parameters are unknown constants : find the constant that maximizes the likelihood

- Bayesians
  - parameters are themselves rv s : find the distribution of the parameters

Likelihoods

$P(D \mid M) = P(x \mid \theta) = L(\theta)$

$= P(x; \theta)$



SUMMING ALONG x GIVES 1

Likelihoods

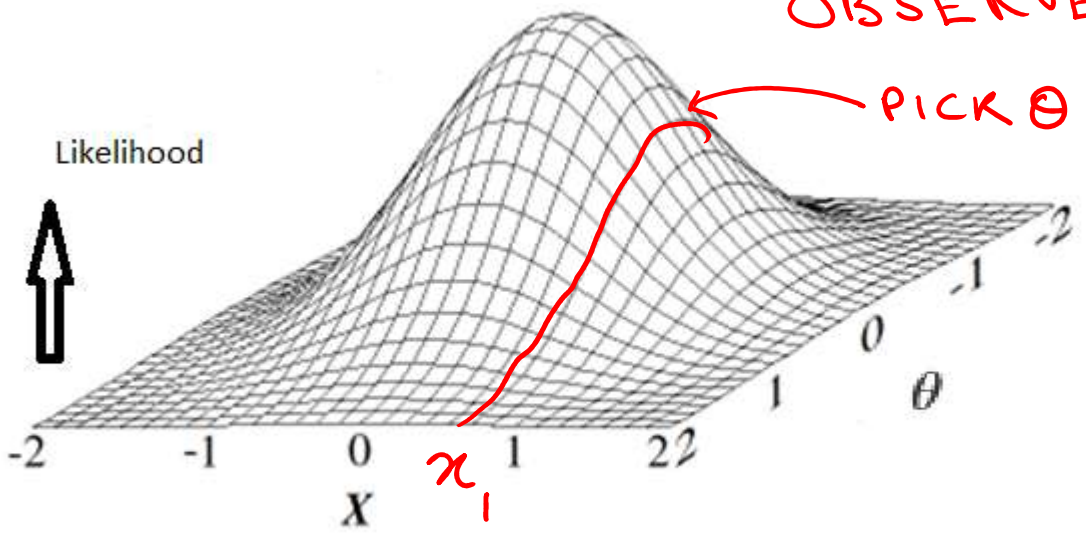
$P(D \mid M) = P(X \mid \theta) = L(\theta)$

$= P(X; \theta)$

FREQUENTIST

OBSERVES $X = x_1$

PICK θ WHICH MAXIMIZES $L_{X=x_1}(\theta)$

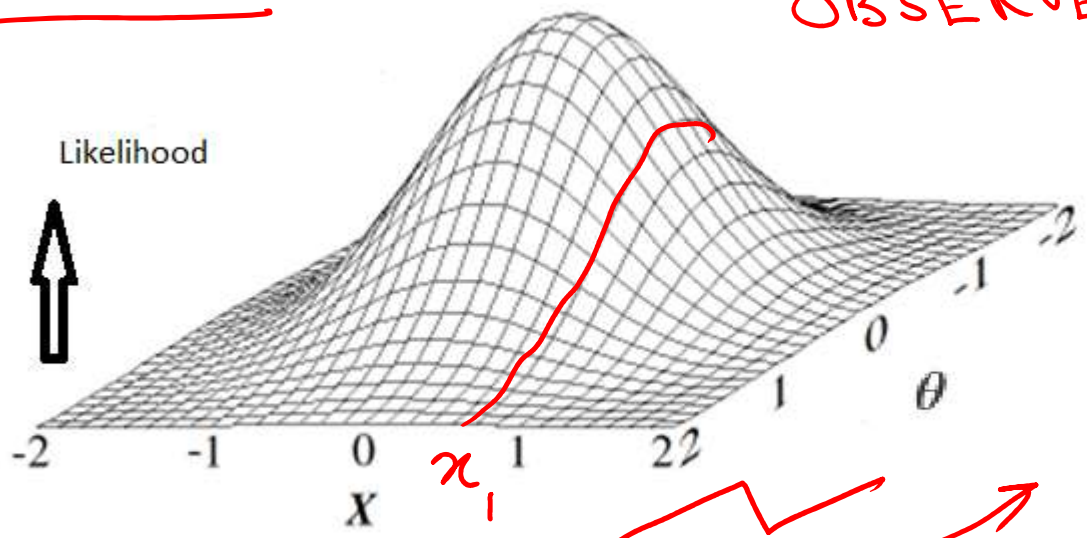


Likelihoods

$$P(X,\theta) = P(X|\theta) \cdot P(\theta)$$

BAYESIAN

OBSERVES $X = x_1$



If FORCED, PICK θ WHICH MAXIMIZE POSTERIOR

POSTERIOR

PRIOR PROB

POSTERIOR PROB

θ

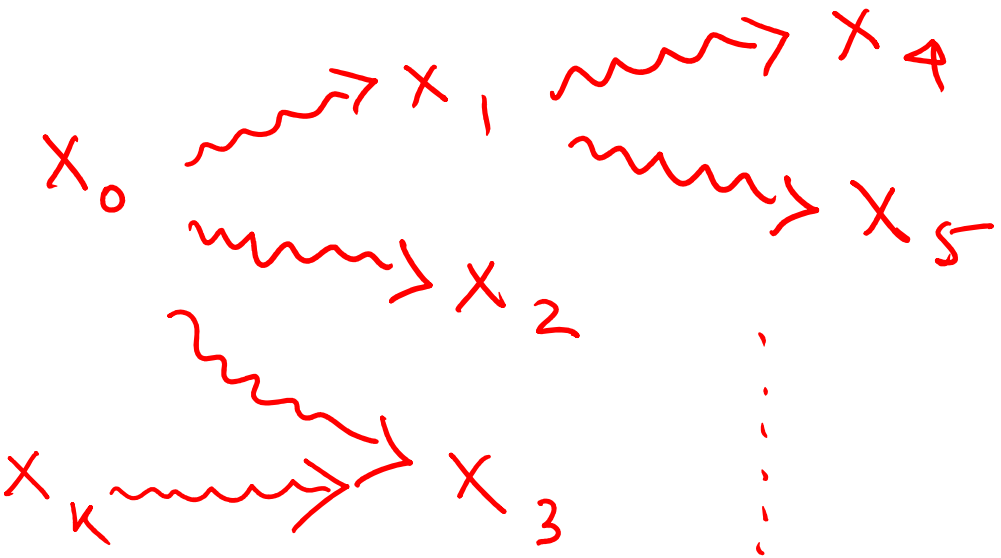
θ

UNNORMALIZED → POST(θ) = $L_{X=x_1}(\theta)$ PRIOR(θ)

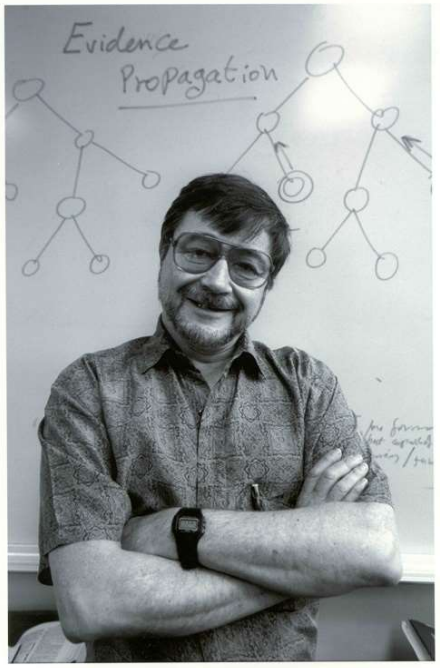
Bayesian networks

CAN BE USED FOR ML & BAYESIAN ANALYSIS

- Think of a sequential generative process : r. v. X_0 gives rise to X_1, X_2, … ; maybe in collaboration with other Xi s
- Each of these in turn give rise to more RVs



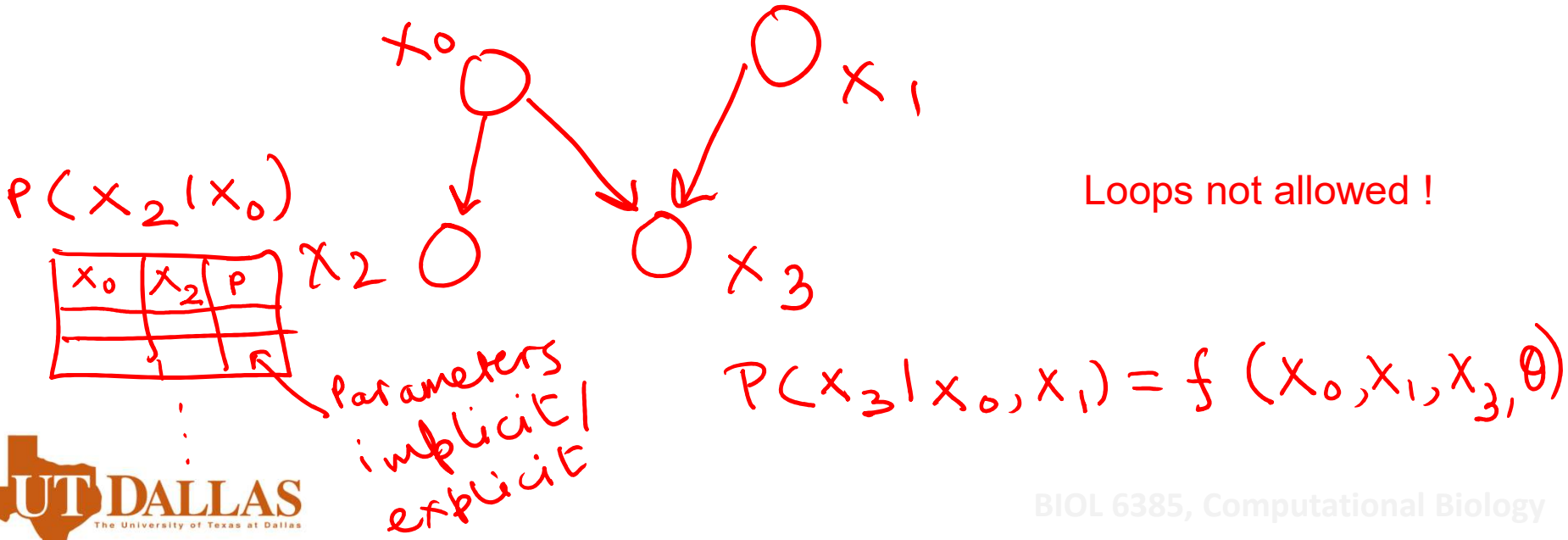
wikipedia



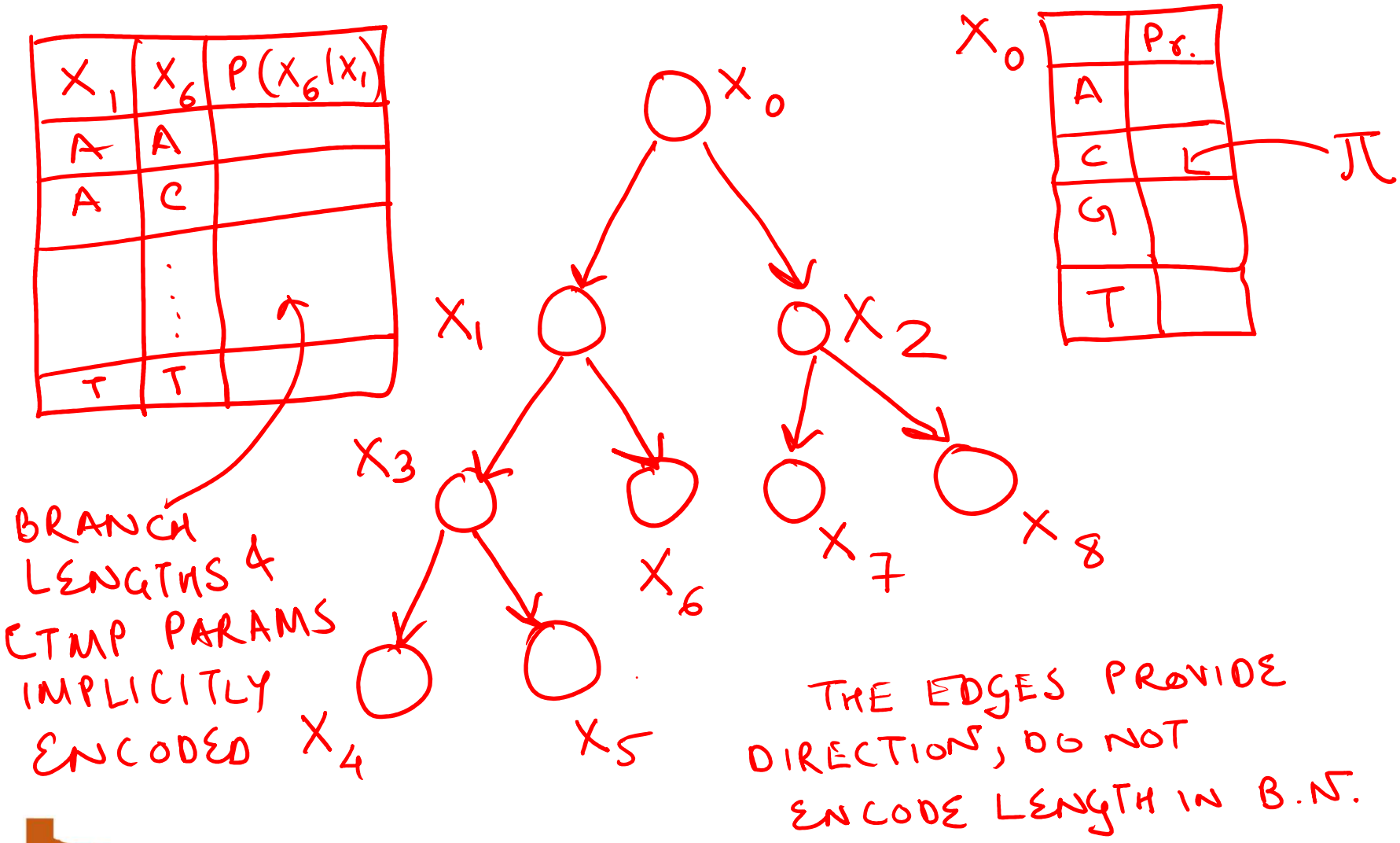
Judea Pearl

Bayesian networks

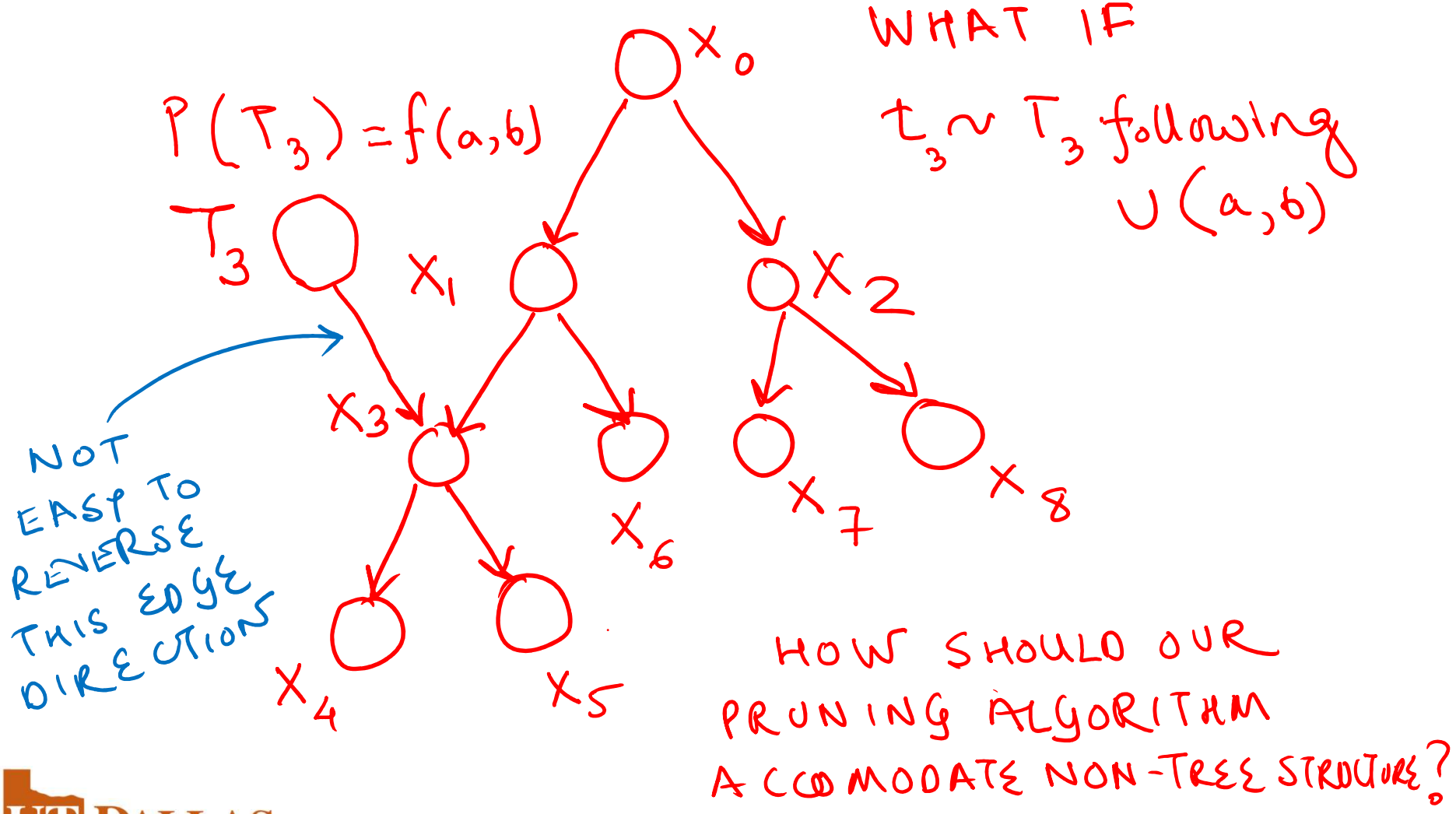
- The process of RVs "giving rise" to another RV can be captured by local conditional distributions (shown by tables – discrete support, or function – continuous support)



Representing our phylogeny as BN



Bayesian trees : priors on parameters



Comparing methods

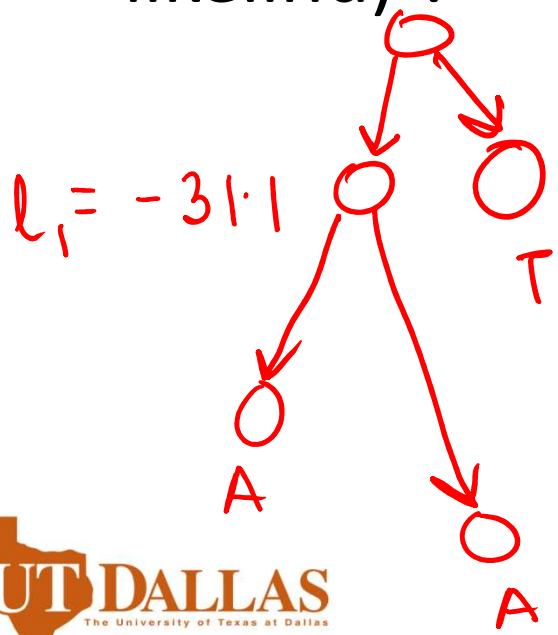
- Identifiability
  - Are parameters confounded given the observed data ?
- Consistency
  - Does it converge to the right tree as data set increases ?
- Efficiency
  - Is the variance low ? For unbiased estimators, bounded by Cramer Rao bound
- Robustness
  - Does performance degrade smoothly when model assumptions are violated ?

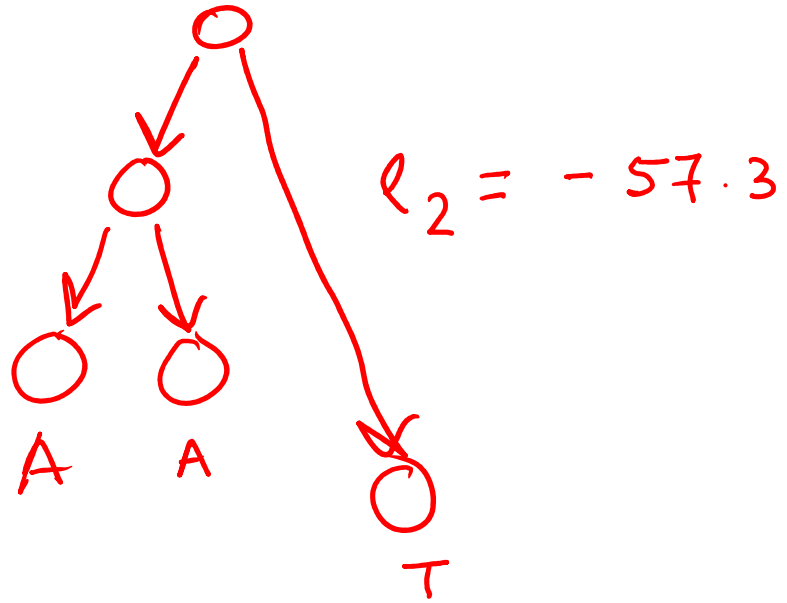
Comparing trees

- How different are two trees ?
- Which one is better in the light of the data ?
  - Were the two reconstructed using the same model (assumptions) ?
    - Comparing trees based on the same model
    - Comparing trees across models

Which tree is better ?

- Under the same model : same likelihood fn
- Which one has higher likelihood (or log likelihd) ?





Which tree is better ?

- Likelihd ratio test : 2 models
  - Ho : null hypothesis
  - H1 : alternative hypothesis
  - typically Ho is a special case of H1

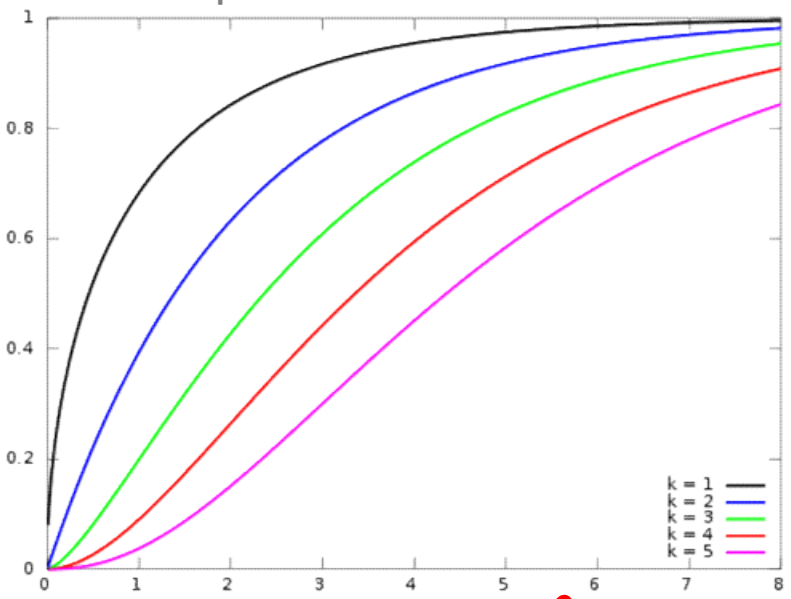
e.g. JC 69 is a special case of K80 with $K = 1$

$H_0 \rightarrow p_0$ parameters,
$H_1 \rightarrow p_1$ parameters
} → find log likelihd of ML parameters

$$2(ll_1 - ll_0) \overset{\text{approx.}}{\sim} \chi^2_{p_1 - p_0}$$

Which tree is better ?

wikipedia



χ^2

eg

$\ell_0 = -6262.01 \leftarrow$ JC69

$\ell_1 = -6113.86 \leftarrow$ K80

↑ MORE PARAMS, HIGHER LIKELIHD

What is $p_1 - p_0$?

$2(\ell_1 - \ell_0) = 296.3 \gg \chi^2_{1,1\%} = 6.63$

H_1 accepted

Which tree is better ?

- What if models are non nested ?

- Aikake Information Criterion : whose AIC is better = - 2 log(L) + 2 p

p → no of params

- Bayesian Information Criterion : whose BIC is better = - 2 log(L) + log(n) p

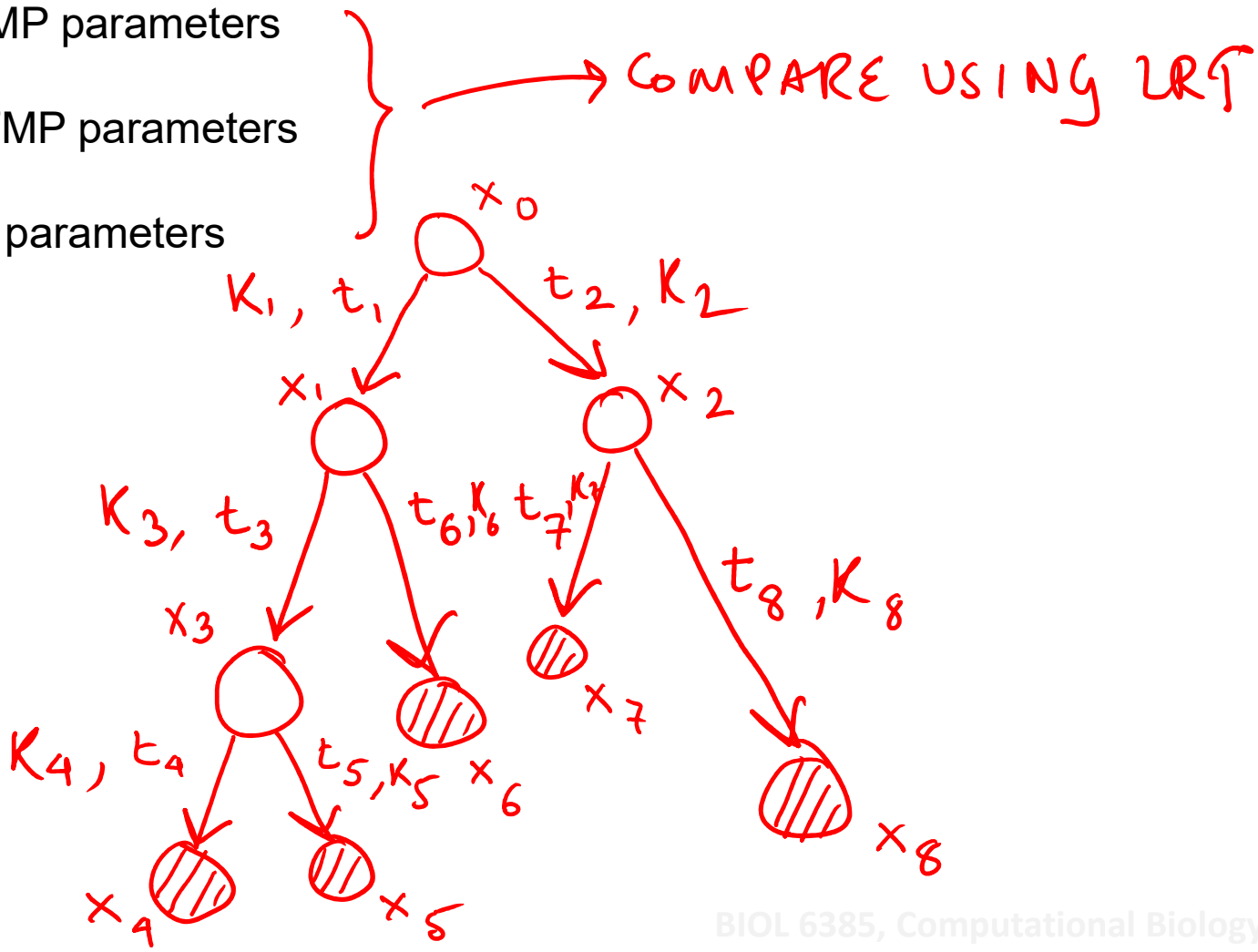
n → data set size

Testing differential selection

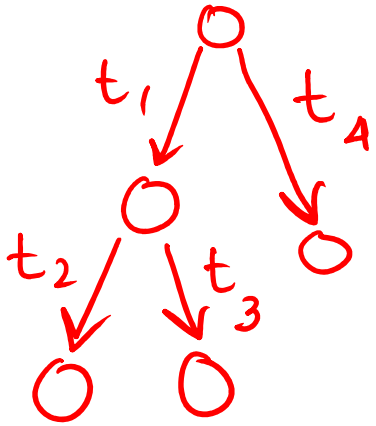
Branch specific CTMP parameters

Lineage specific CTMP parameters

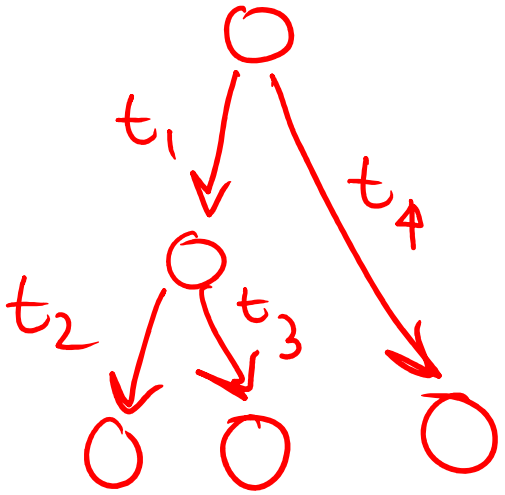
Single set of CTMP parameters



Testing the molecular clock



AND



FIND ML VALUES & CORR. LOG LIKELIHD

w/ constraints:

$t_2 = t_3$, $t_1 + t_2 = t_4$

[USE LAGRANGE MULTIPLIER]

FIND ML VALUES & LOG LIKELIHD

COMPARE W/ LRT

Can we really test for the molecular clock ?

- We test that the root is equidistant from all the leaves
  - A weaker assertion than that of the molecular clock. Why ?

Can we really test for the molecular clock ?

- We test that the root is equidistant from all the leaves
  - A weaker assertion than that of the molecular clock. Why ?
  - Mutation rates could be different in precisely the same or compensatory ways along each lineage, and this would still hold
    - We can only assert total amount of mutation from start of clock is same in all lineages

Calibration with real time

- If molecular clock hypothesis holds : branch length = expected no of substitutions should be linear to real time

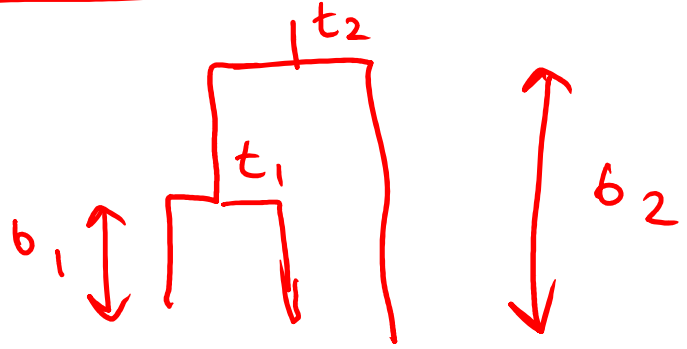
$b_1 = t_1 r \Rightarrow b_2 = t_2 \cdot r$

(r: KNOWN)

$= t_2 \cdot \frac{b_1}{t_1}$

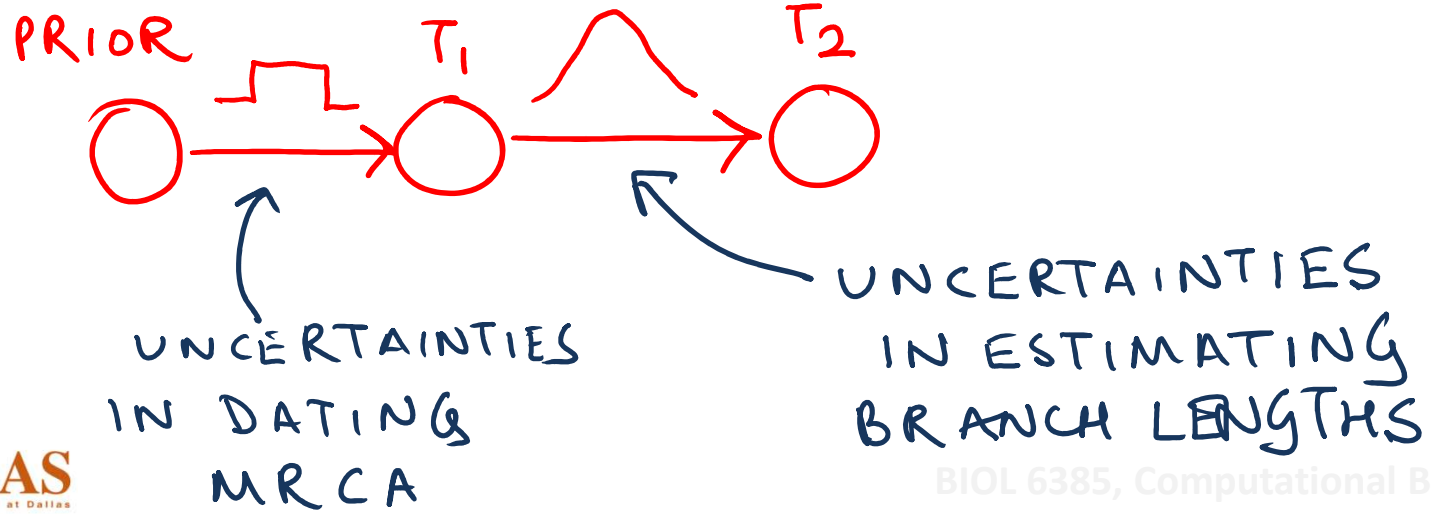
$t_2 = \frac{b_2 \cdot t_1}{b_1}$

3 SPECIES, INFINITE DATA

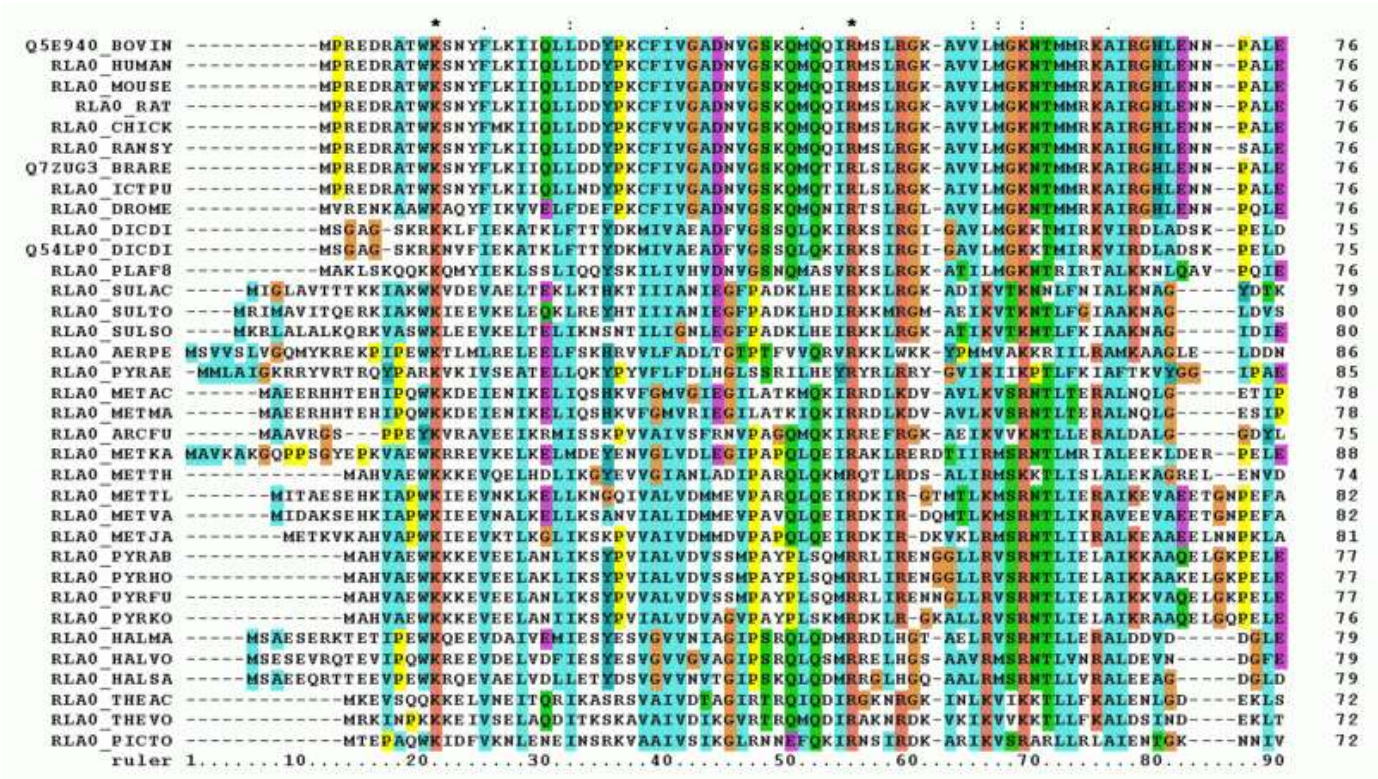


Uncertainty in calibration

- Inaccurate branch lengths
- Molecular dating of fossils come with error bars
- Common ancestor or not – determined by character data : how to determine how far it is from MRCA ?



Accounting for multiple models

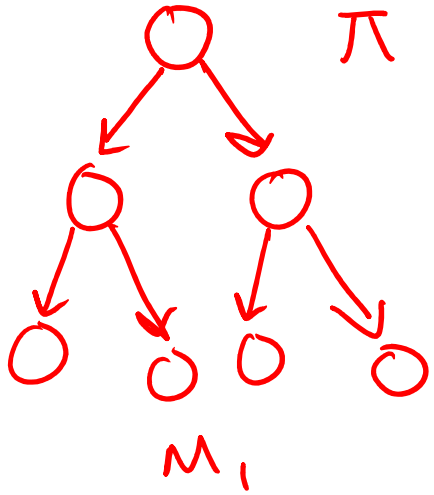


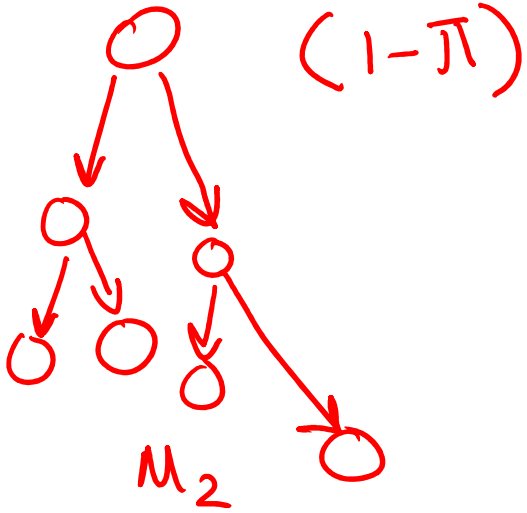
wikipedia

- Are all the characters generated by the same evolutionary model ?

Modelling variation : horizontally

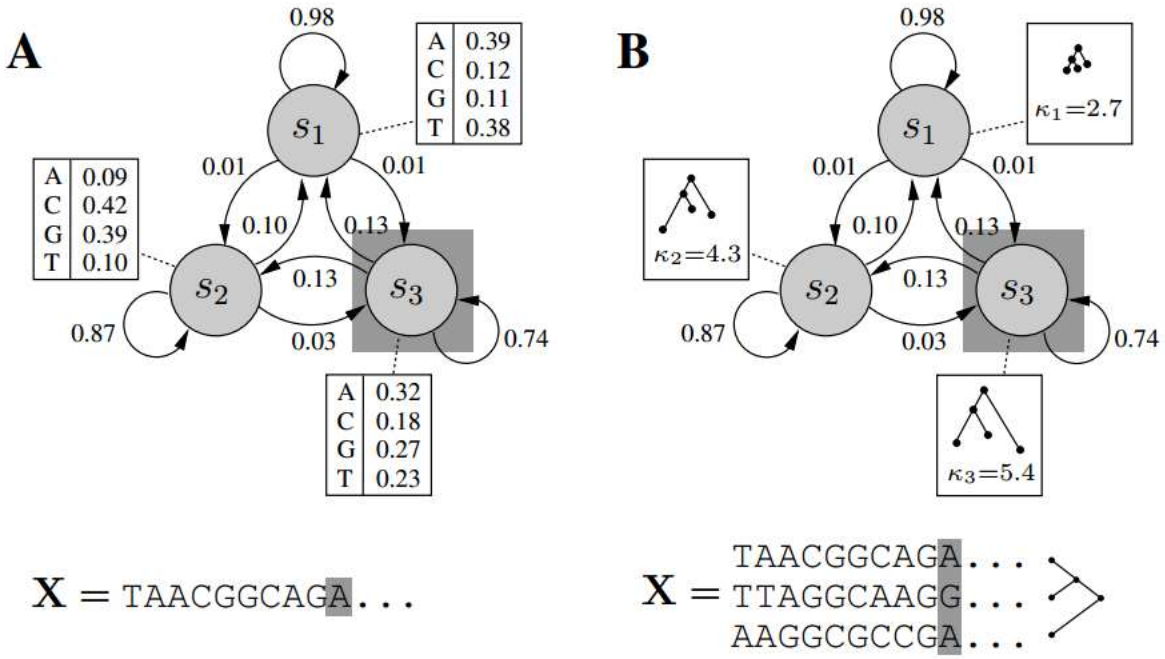
- Along the genome, is there spatial correlation in evolutionary parameters (say, mutation rates) ?
  - If yes, markov model
  - If no, mixture model





Modelling variation : horizontally

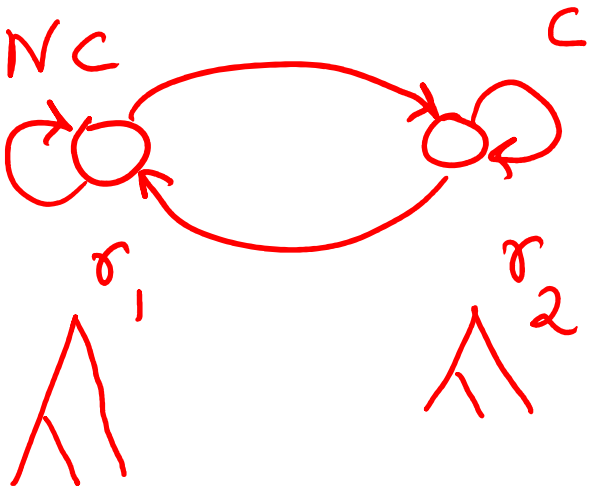
- Phylogenetic HMMs



Siepel & Haussler, JCB, 2004

Modelling spacers, repeats, conserved regions and other evolutionary events

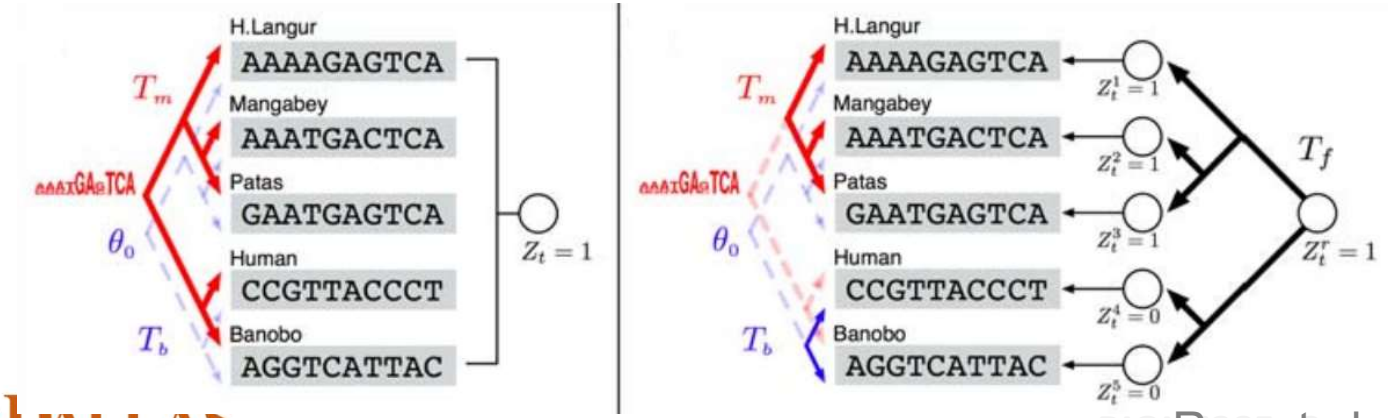
- PhastCons : 2 state phylogenetic HMM modelling evolutionary rates for conserved and non conserved sites



$\sigma_2 < \sigma_1$

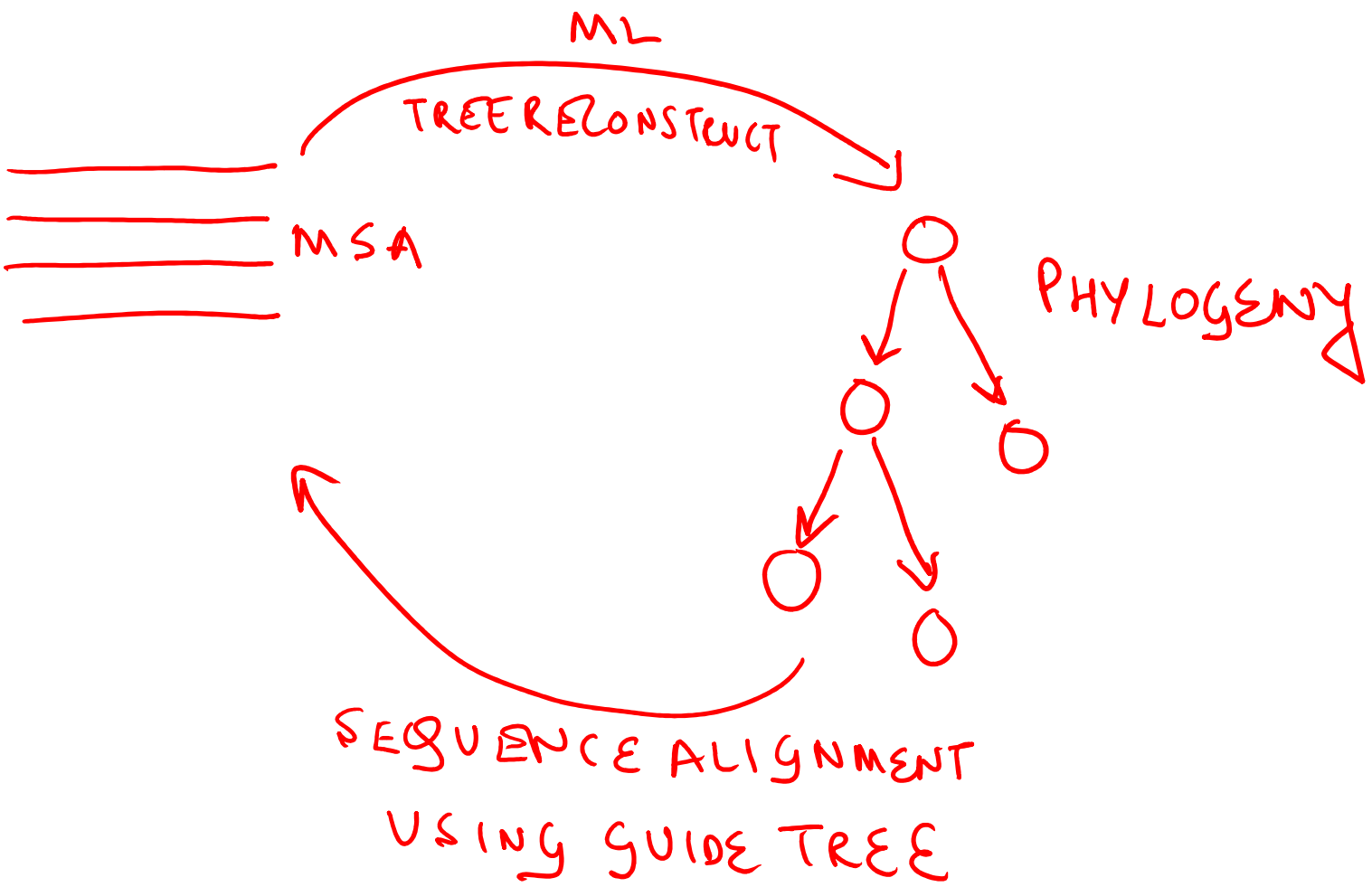
Modelling variation : vertically

- In the phylogeny, is there correlation in evolutionary parameters inside subtrees ?
  - If yes, model a mixture of phylogenies : mixture components drawn from another phylogeny
  - If no, model a mixture of phylogenies : mixture components drawn independently



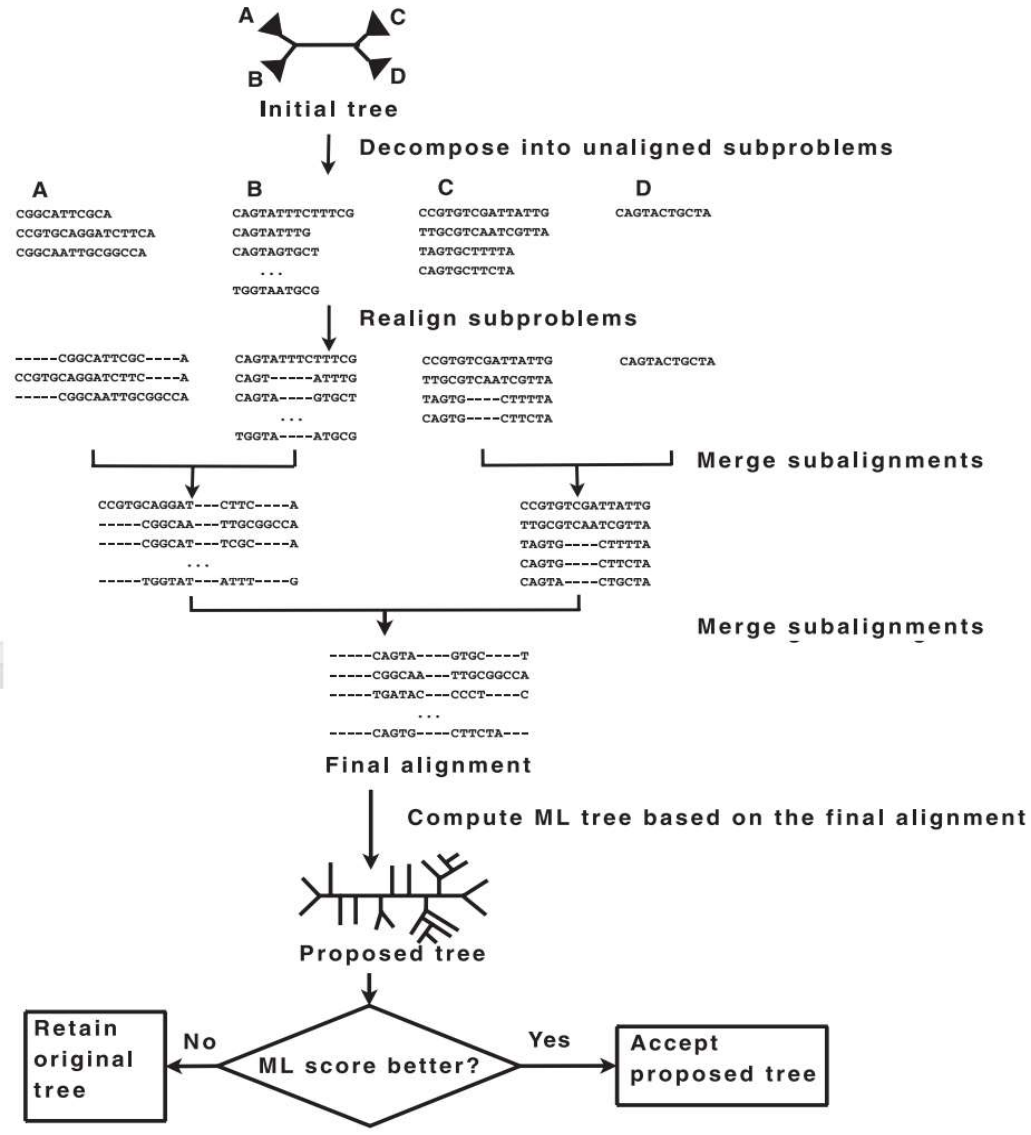
Ray et al, PloS CB, 2008

MSA – phylogeny co-construction



Practical tree building

Warnow Lab, Science, 2009



UT DALLAS
The University of Texas at Dallas

Controversies and the NFL theorem

- If there is no “wrong” model, merely better or worse models in the light of a data set
  - how do we falsify an evolutionary hypothesis ? Isnt that a cornerstone of the scientific process ?

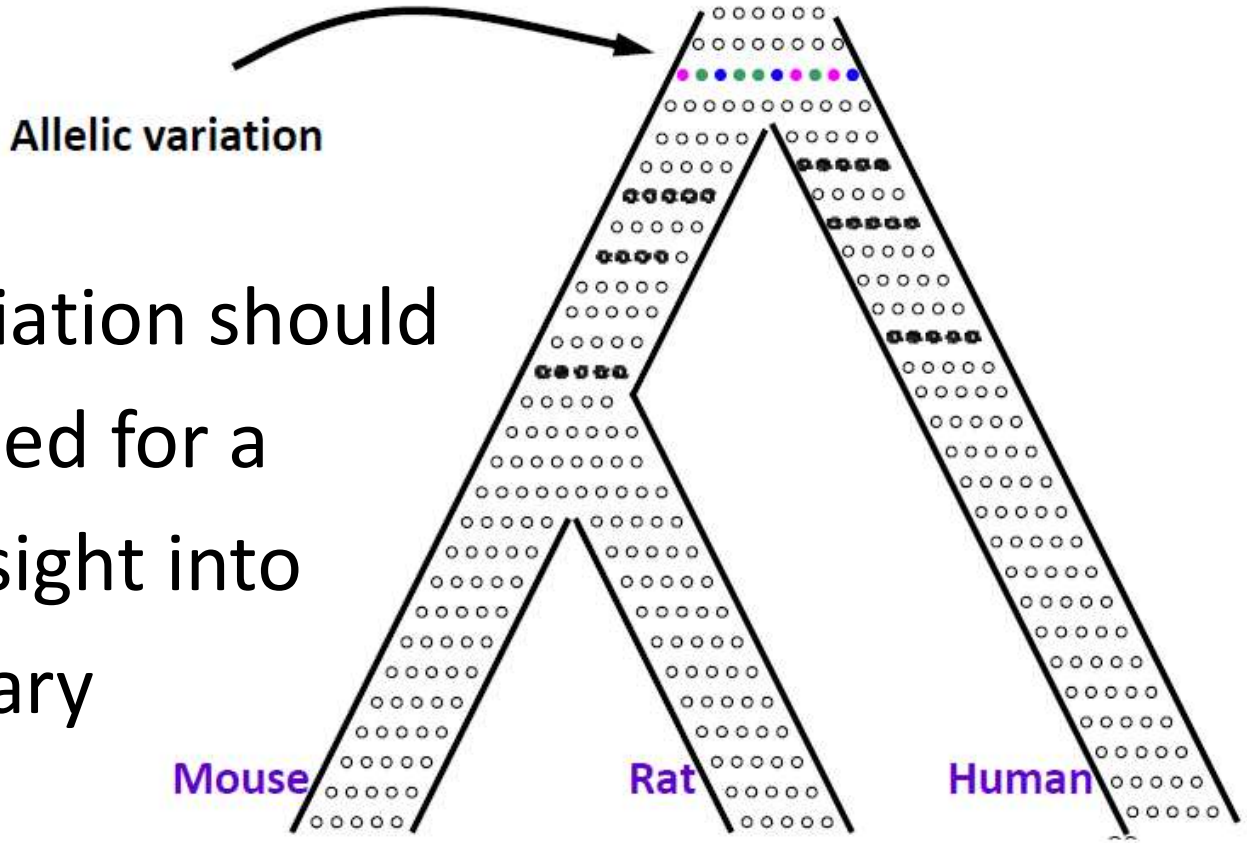
- Typically, in the light of some data, we say one theory is better than the other
  - higher score : better fit
  - explains more data : more general
  - typically traded off

Controversies and the NFL theorem

- No free lunch theorems : complicated mathematical theorems
  - in the case of statistical learning, one key tenet
    - If you don't assume anything about a data set, the only thing you can learn about the data set is the set itself
      - Assumptions about the data = learning bias
- Take home message : never cherry pick examples
  - there is a vast repository of evolutionary data : cherrypicking can always bolster any model

The real phylogenetic tree

- It’s a jungle out there !



Allelic variation should be modelled for a clearer insight into evolutionary dynamics

Acknowledgements

- Eric Xing
- Dannie Durand