Phylogenetic trees IV
Maximum Likelihood

Gerhard Jäger

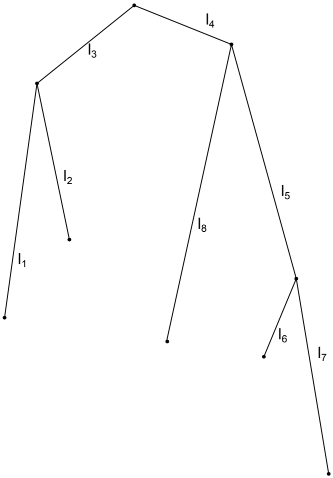
Words, Bones, Genes, Tools

February 28, 2018

Theory

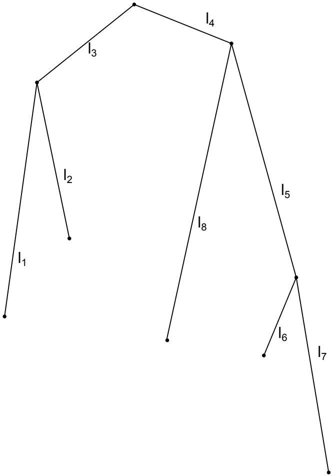
Recap: Continuous time Markov model

$$\begin{aligned} P(t) &= \begin{pmatrix} s + re^{-t} & r - re^{-t} \\ s - se^{-t} & r + se^{-t} \end{pmatrix} \\ \pi &= (s, r) \end{aligned}$$



Likelihood of a tree

background reading: Ewens and Grant (2005), 15.7



- simplifying assumption: evolution at different branches is independent
- suppose we know probability distributions v_t and v_b over states at top and bottom of branch l_k
- $\mathcal{L}(l_k) = v_t^T P(l_k) v_b$

Likelihood of a tree

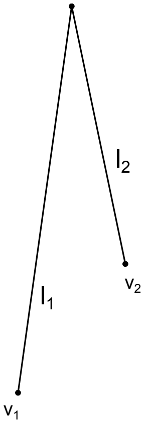
- likelihoods of states $(0,1)$ at root are

$$v_1^T P(l_1) v_2^T P(l_2)$$

- log-likelihoods

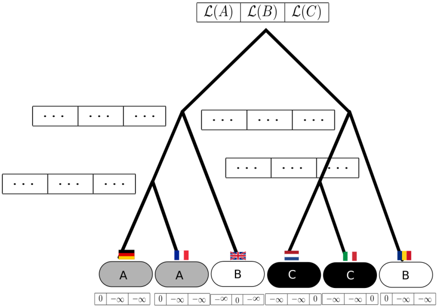
$$\log(v_1^T P(l_1)) + \log(v_2^T P(l_2))$$

- log-likelihood of larger tree: recursively apply this method from tips to root



Likelihood of a tree

$$\mathcal{L}(\textit{mother})_i = \prod_{d\in \textit{daughters}} \sum_{1\le j\le n} (P(t)_{i,j}\mathcal{L}(d)_j),$$


$\mathcal{L}(A)$ | $\mathcal{L}(B)$ | $\mathcal{L}(C)$
... | ... | ...
... | ... | ...
... | ... | ...
... | ... | ...
A
A
B
C
C
B
0 | $-\infty$ | $-\infty$
0 | $-\infty$ | $-\infty$
$-\infty$ | 0 | $-\infty$
$-\infty$ | $-\infty$ | 0
$-\infty$ | $-\infty$ | 0
0 | $-\infty$ | $-\infty$

(Log-)Likelihood of a tree

- this is essentially identical to Sankoff algorithm for parsimony:
  - $\text{weight}(i, j) = \log P(l_k)_{ij}$
  - weight matrix depends on branch length → needs to be recomputed for each branch
- overall likelihood for entire tree depends on probability distribution on root
- if we assume that root node is in equilibrium:

$$\mathcal{L}(\text{tree}) = (s, r)^T \mathcal{L}(\text{root})$$

- does not depend on location of the root (→ time reversibility)
- this is for one character — likelhood for all data is product of likelihoods for each character

(Log-)Likelihood of a tree

- likelihood of tree depends on
  - branch lengths
  - rates for each character
- likelihood for tree *topology:*

$$\mathcal{L}(\text{topology}) = \max_{l_k:\ k \text{ is a branch}} \mathcal{L}(\text{tree}|\vec{l_k})$$

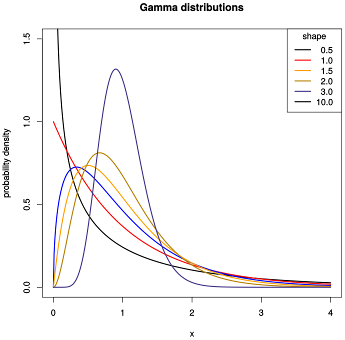
(Log-)Likelihood of a tree

- Where do we get the rates from?
- different options, increasing order of complexity
  1. $s = r = 0.5$ for all characters
  2. $r =$ empirical relative frequency of state 1 in the data (identical for all characters)
  3. a certain proportion p_{inv} (value to be estimated) of characters are *invariant*
  4. rates are *gamma distributed*

Gamma-distributed rates

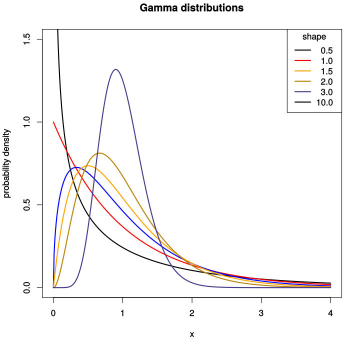
- we want allow rates to vary, but not too much
- common method (no real justification except for mathematical convenience)
  - equilibrium distribution is identical for all characters
  - rate matrix is multiplied with coefficient λ_i for character i
  - λ_i is random variable drawn from a *Gamma distribution*

$$\mathcal{L}(r_i = x) = \frac{\beta^\beta x^{(\beta-1)} e^{-\beta x}}{\Gamma(\beta)}$$



Gamma-distributed rates

- overall likelihood of tree topology: integrate over all λ_i, weighted by Gamma likelihood
- computationally impractical
- in practice: split Gamma distribution into n discrete bins (usually $n = 4$) and approximate integration via Hidden Markov Model



Modeling decisions to make

aspect of model	**possible choices**	**number of parameters to estimate**
branch lengths	unconstrained	$2n-3$ (n is number of taxa)
	ultrametric	$n-1$
equilibrium probabilities	uniform	0
	empirical	1
	ML estimate	1
rate variation	none	0
	Gamma distributed	1
invariant characters	none	0
	p_{inv}	1

This could be continued — you can build in rate variation across branches, you can fit the number of Gamma categories . . .

Model selection

- tradeoff
  - rich models are better at detecting patterns in the data, but are prone to over-fitting
  - parsimoneous models less vulnerable to overfitting but may miss important information
- standard issue in statistical inference
- one possible heuristics: **Akaike Information Criterion** (AIC)

$$\text{AIC} = -2 \times \log \text{likelihood} + 2 \times \text{number of free parameters}$$

- the model minimizing AIC is to be preferred

Example: Model selection for cognacy data/ UPGMA tree

model no.	branch lengths	eq. probs.	rate variation	inv. char.	AIC
1	ultrametric	uniform	none	none	17515.95
2	ultrametric	uniform	none	p_{inv}	17518.39
3	ultrametric	uniform	Gamma	none	17517.89
4	ultrametric	uniform	Gamma	p_{inv}	17519.75
5	ultrametric	empirical	none	none	16114.66
6	ultrametric	empirical	none	p_{inv}	16056.85
7	ultrametric	empirical	Gamma	none	15997.16
8	ultrametric	empirical	Gamma	p_{inv}	16022.21
9	ultrametric	ML	none	none	16034.96
10	ultrametric	ML	none	p_{inv}	16058.83
11	ultrametric	ML	Gamma	none	15981.94
12	ultrametric	ML	Gamma	p_{inv}	16009.90
13	unconstrained	uniform	none	none	17492.73
14	unconstrained	uniform	none	p_{inv}	17494.73
15	unconstrained	uniform	Gamma	none	17494.73
16	unconstrained	uniform	Gamma	p_{inv}	17496.73
17	unconstrained	empirical	none	none	16106.52
18	unconstrained	empirical	none	p_{inv}	16049.28
19	unconstrained	empirical	Gamma	none	16033.21
20	unconstrained	empirical	Gamma	p_{inv}	16011.38
21	unconstrained	ML	none	none	16102.04
22	unconstrained	ML	none	p_{inv}	16051.27
23	unconstrained	ML	Gamma	none	16025.99
24	unconstrained	ML	Gamma	p_{inv}	16001.00

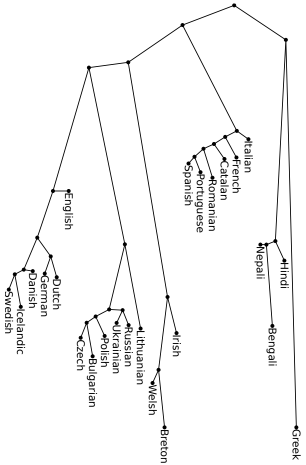
Tree search

- ML computation gives us likelihood of a tree topology, given data and a model
- ML tree:
  - heuristic search to find the topology maximizing likelihood
  - optimize branch lengths to maximize likelihood for that topology
- computationally very demanding!
- *for the 25 taxa in our running example, ML tree search for the full model requires several hours on a single processor; parallelization helps*
- ideally, one would want to do 24 heuristic tree searches, one for each model specification, and pick the tree+model with lowest AIC
- in practice one has to make compromises

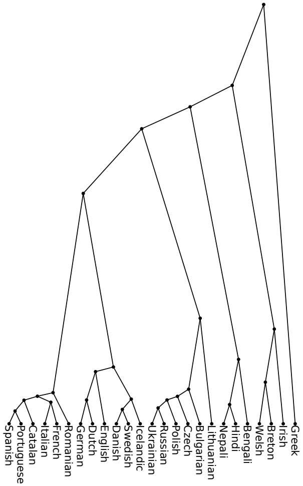
Running example

Running example: cognacy data

unconstrained branch lengths:
AIC = 7929

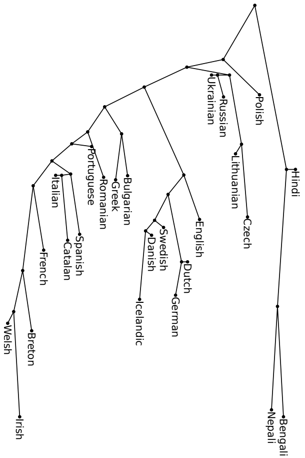


ultrametric:
AIC = 7972

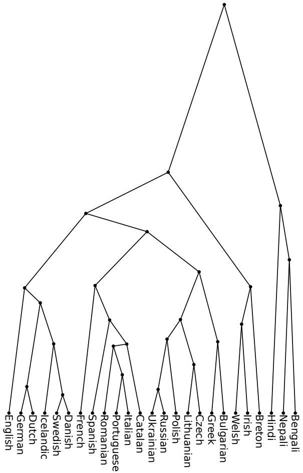


Running example: WALS data

unconstrained branch lengths:
AIC = 2752

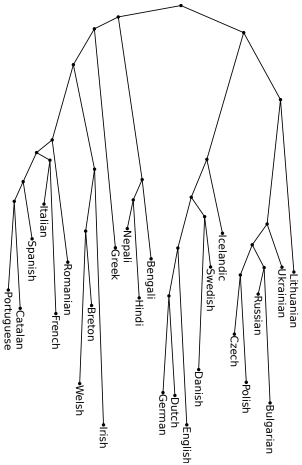


ultrametric:
AIC = 2828

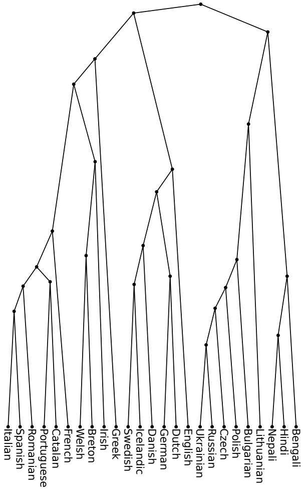


Running example: phonetic data

unconstrained branch lengths:
AIC = 89871



ultrametric:
AIC = 90575



Wrapping up

- ML is conceptually superior to MP (let alone distance methods)
  - different mutation rates for different characters are inferred from the data
  - possibility of multiple mutations are taken into account — depending on branch lengths
  - side effect of likelihood computation: probability distribution over character states at each internal node can be read off
- disadvantages:
  - computationally demanding
  - many parameter settings makes model selection difficult
    (note that the ultrametric trees in our example are sometimes better even though they have higher AIC)
  - ultrametric constraint makes branch lengths optimization computationally more expensive ⇒ not feasible for larger data sets

Ewens, W. and G. Grant (2005). *Statistical Methods in Bioinformatics: An Introduction*. Springer, New York.