Generalizing the model

Features of $\left\{\begin{array}{l}\text{GTR}\\ \text{GM}\end{array}\right.$ model: independence assumption

gains you data, but likely unrealistic

identically distributed

If aligned sequences are long, perhaps unrealistic to use same parameters on each site.

A common way to address this is with a MIXTURE MODEL

Simple Example: Fix T.

Suppose you have 3 classes of sites (slow, medium, fast evolving) and choose GTR parameters for each of them. Compute the expected pattern frequency arrays

$$\begin{array}{ccc} P^{slow} & P^{medium} & P^{fast} \\ \uparrow & \uparrow & \uparrow \\ \vec{p}_1, Q, \{t_e^1\} & \vec{p}_2, Q_2, \{t_e^2\} & \vec{p}_3, Q_3, \{t_e^3\} \end{array}$$

Additionally, choose weighting or mixture parameter α_1, α_2 $(k=2)$

then the joint frequency array is

$$P = \alpha_1 P_1 + \alpha_2 P_2 + (1-\alpha_1-\alpha_2) P_3$$

Each P_i is called a MIXTURE COMPONENT and the number of parameters to infer is

$$\underset{\substack{\nearrow\\ \#\text{ component}}}{3}\,(\underset{\substack{\uparrow\\ \vec{p}}}{3} + \underset{\substack{\uparrow\\ Q}}{5} + \underset{\substack{\uparrow\\ \text{b.l.}}}{2n-3}) + \underset{\substack{\uparrow\\ \#\text{ of components} -1}}{2}$$

The α_i are called the weights or mixing parameters

Note that for a fixed binary tree on n taxa that if a mixture uses K components, then the number of numerical parameters to be inferred increases roughly by a factor of K. i.e. Increases a lot.

For both biological and practical reasons, mixture models with fewer parameters are used in practice. Some examples include:

- GTR+I ≡ GTR + Invariable sites model

Two classes of sites: those that are free to mutate (GTR)

those that are variable due to perhaps functional constraints (I)

Parameters: GTR : $\vec{p}_Q$, Q $3+6=9$

I : $\vec{p}_I$ $=3$

weights: s $=1$

Excluding branch lengths, this is an 13-parameter model. I.e. 4 additional parameters.

A variation sets $p_I = p_Q$, i.e. assumes the base distribution is the same over all sites, both variable and invariable.

- GTR + rate variation (discrete) Lots of variation

Assumes k classes of sites but that the mutation rate is scaled depending which class you are in.

Example: GTR + rate varation k classes   # of classes chosen by user.

Numerical parameters (excluding branch lengths on T) are

- GTR parameters $\vec{p}, Q$ used for all sites
- classes weights $s_1, s_2, \dots, s_k$ $\sum s_i = 1$, $s_i > 0$ distribution of sites to classes
- rates $r_1, r_2, \dots, r_k$ $r_i \geq 0$ scaling rate for i-th class

The pattern frequency array is

$$P = s_1 P_1 + s_2 P_2 + \dots + s_k P_k$$

where P_i is the expected pattern freq array for the i-th class

P_i is computed using $\vec{p}, \underbrace{r_i Q}$ and branch lengths $\{t_e\}$

scaled version of Q, with scaling factor the rate r_i.

$r_i \geq 1$, $r_i < 1$, etc.

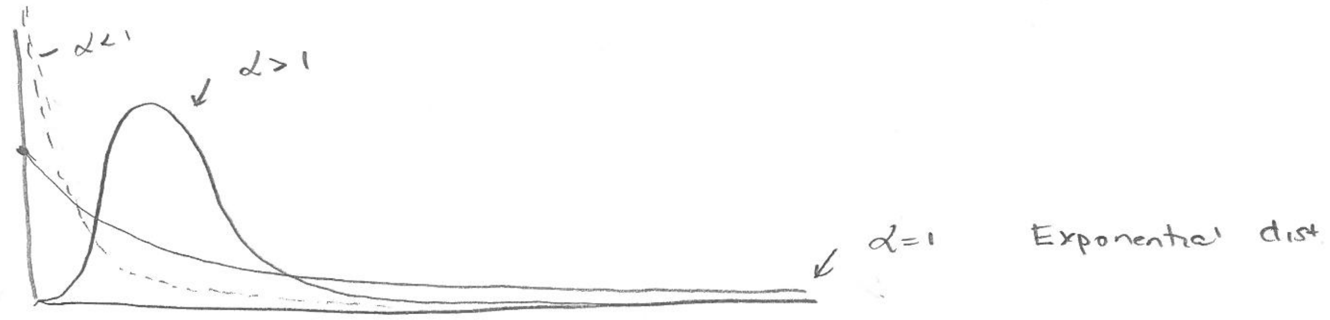
visual effect.

Discuss how to use such a model for simulations ...

Variations: Instead of choosing the rate functions at random, choose them from a distribution. In practice, the Γ-distribution is used or in reality the discrete-Γ.

Gamma distributions in phylogenetics is a 1-parameter family of distributions with the unknown parameter called α = shape parameter.

The densities for various values of α are shown



Notice all $r_i > 0$ are possible rates and the shape of the density says something about the probability of various rates.

See R demo and discuss meaning.