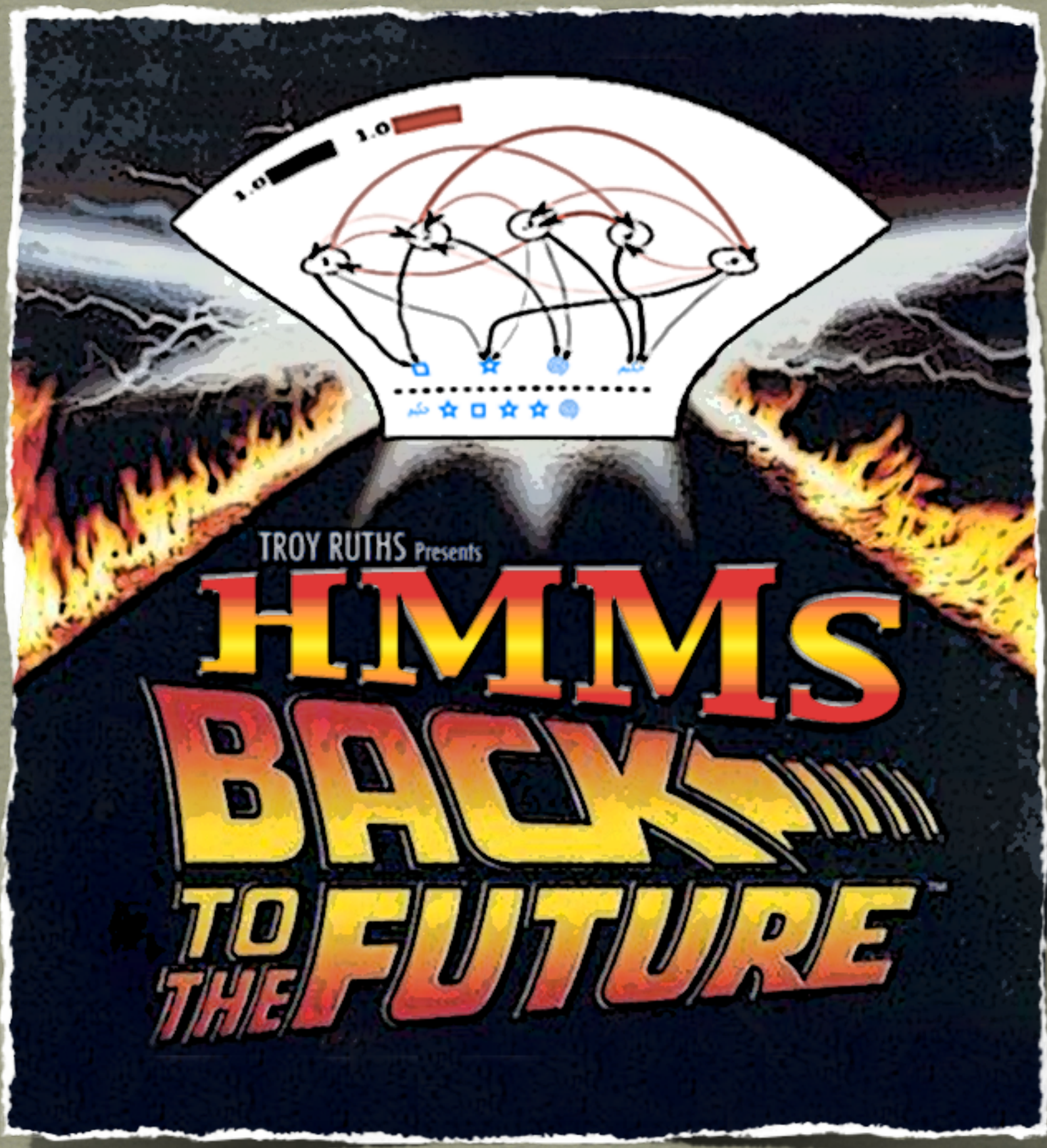

1.0
1.0
TROY RUTHS Presents
HMMs
BACK
TO THE FUTURE

PhyloHMM: the future

- Bring ***back*** the HMM ***to the future***
- Updating the biological model with phylogenetics

Combining Phylogenetic and HMMs in Biosequence Analysis

Adam Siepel & David Haussler (2004)

What we'll Discuss

- Motivation for updating HMMs
- Design of Phylo-HMM
  - Tree Model
  - DNA Substitution
  - Evolutionary rate
  - ➡Categories
  - ➡Higher-order states
- Application to data & results

Recap on HMMs

- Dominant tool in biological sequence analysis
- Gene prediction, homology searching, structure ...

➡Balance simplicity and expressiveness

Antiquated HMMs

Your HMMs disregard three decades of sequence evolution research.

…

"CS McFly"

"Biologist Biff"

Antiquated HMMs

- Sites are independent
- Substitutions are homogeneous
- Evolutionary rates are consistent
- Functional categories are disregarded

What's the solution?

Enter Phylogeny

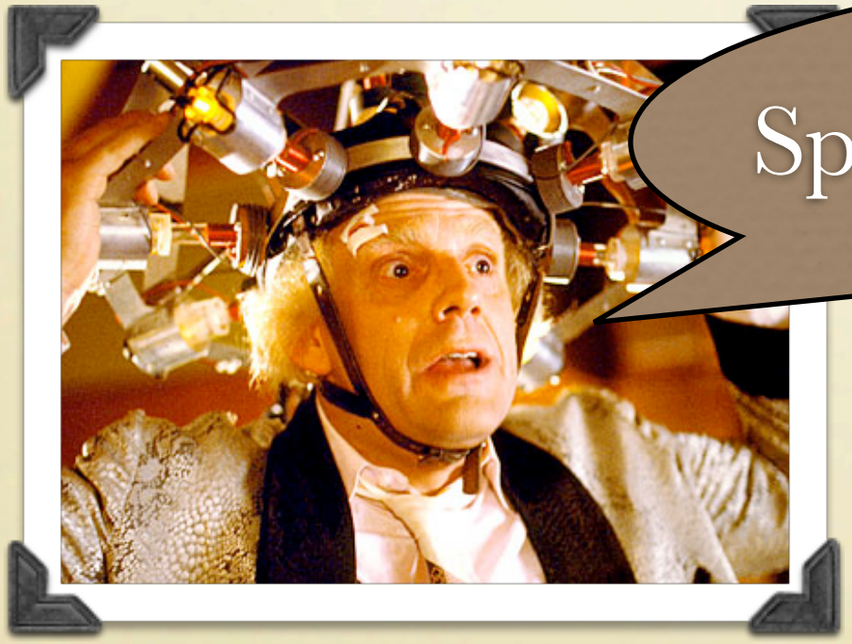
- Provides probabilistic models of evolution
- Based on
  - Topology of tree (relatedness)
  - Lengths of its branches (rates)
  - Pattern of substitution (categories)

➡ ***Time-based***

➡ ***Works across sequences***

HMM += Phylogeny

- Both are built on probabilistic models
- HMM operates along a sequence
- Phylogenetics operate between sequences



Space + Time = Phylo-HMM

The Method

Input

- n aligned sequences of length L
- Phylogenetic tree relating the n taxa

$$\psi = (Q, \tau, \beta, \pi)$$

Substitution Matrix

Topology

Branch Lengths

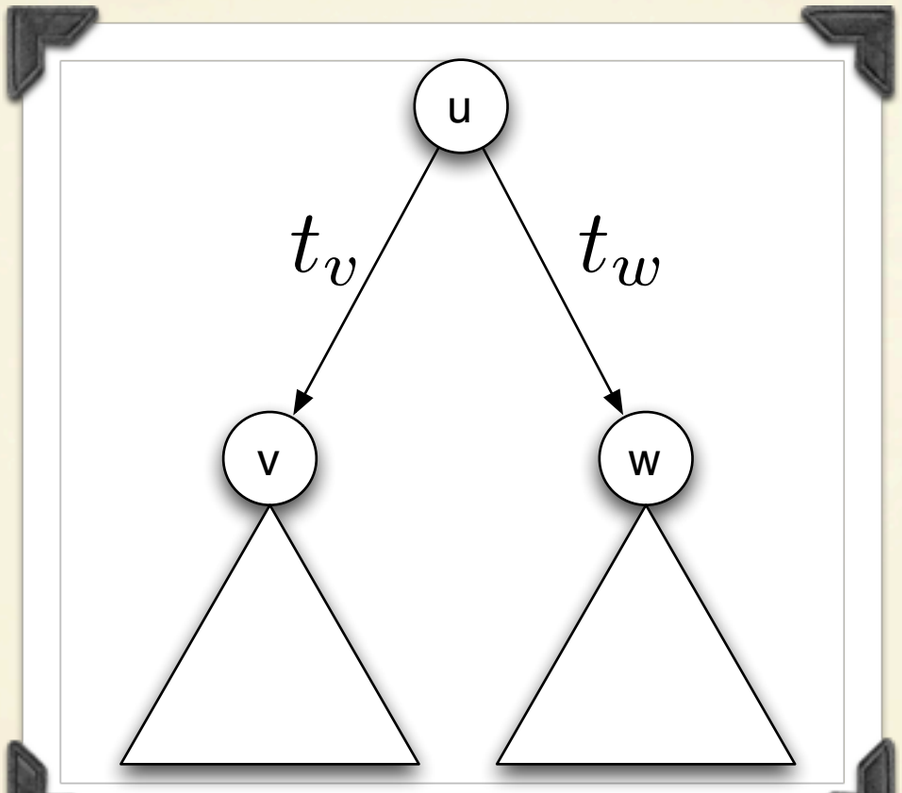
Base Frequencies

Likelihood of a Tree

- Sites of the alignment are assumed independent
- Dynamic programming solution

$$P(X|\psi) = \prod_{i=1}^{L} P(X_i|\psi) = \sum_{\mathbb{L}} P(\mathbb{L}, X_i|\psi)$$

Labeling of ancestral nodes



Recursion

$$P(L_u|a) = \begin{cases} I(a = x_u) & \text{if } u \text{ is a leaf} \\ \sum_b P(b|a, t_v) P(L_v|b) \sum_c P(c|a, t_w) P(L_w|c) & \text{otherwise} \end{cases}$$

Root Call

$$P(X_i|\psi) = \sum_a \pi_a P(L_r|a)$$

DNA Substitution

$$P(b|a,t)$$

- Probability that base b is substituted by base a over a branch of length t

$$\mathbf{Q}_{\mathrm{UNR}} = \begin{pmatrix} - & a & b & c \\ d & - & e & f \\ g & h & - & i \\ j & k & l & - \end{pmatrix} \quad \mathbf{Q}_{\mathrm{REV}} = \begin{pmatrix} - & a\pi_C & b\pi_G & c\pi_T \\ a\pi_A & - & d\pi_G & e\pi_T \\ b\pi_A & d\pi_C & - & f\pi_T \\ c\pi_A & e\pi_C & f\pi_G & - \end{pmatrix}$$

$$\mathbf{Q}_{\mathrm{HKY}} = \begin{pmatrix} - & \pi_C & \kappa\pi_G & \pi_T \\ \pi_A & - & \pi_G & \kappa\pi_T \\ \kappa\pi_A & \pi_C & - & \pi_T \\ \pi_A & \kappa\pi_C & \pi_G & - \end{pmatrix}$$

Evolutionary Rate

- Variate the rate of evolution by scaling the branches
- Discretize the gamma distribution into *k* rates

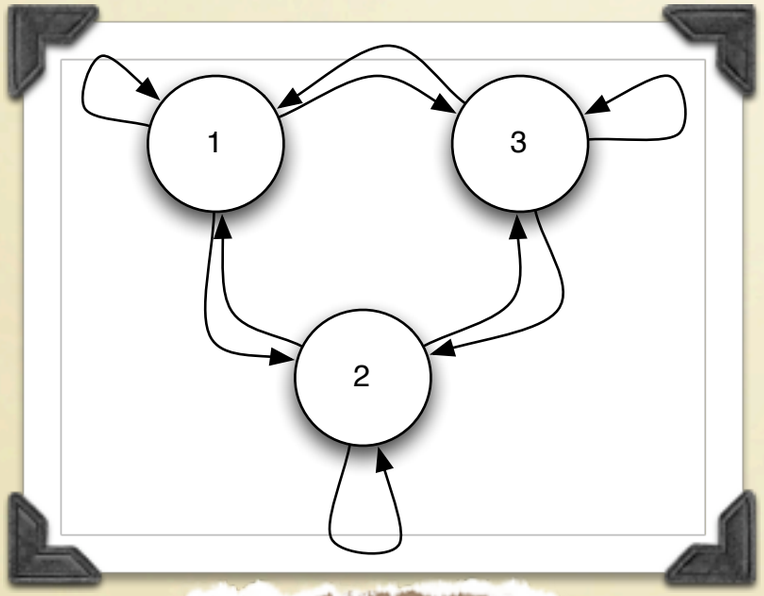
$$P(\mathbf{X}_i|\boldsymbol{\psi}) = \sum_{j=1}^{k} \frac{1}{k} \cdot P(\mathbf{X}_i|(\mathbf{Q}, \boldsymbol{\tau}, r_j\boldsymbol{\beta}, \boldsymbol{\pi}))$$

How do we assign rates?

Scaling the branches

Rates HMM

- Autocorrelation (site *i* is the same as site *i+1*)
- Used in two step fitting process



$k = 3$

$$c_{j,l} = \frac{1-\lambda}{k}$$

$$c_{j,j} = \lambda + \frac{1-\lambda}{k}$$

Transitions

Category HMMs

- Use tree models for "functional categories"
- Topologies may vary, but are usually the same

$$P(\mathbf{X}, \boldsymbol{\phi}) = \prod_{i=1}^{L} a_{\phi_{i-1},\phi_i} P(\mathbf{X}_i | \boldsymbol{\psi}_{\phi_i})$$

Assignment

Transition

Tree Model

Emission

Category x Rates HMM

- Rate and function are orthogonal
- Create HMM that incorporates both
- Take the cross product of states, transitions

  ➡ scale the tree models

$$\psi'_{i,i'} = r_{i,i'}\psi_i = (\mathbf{Q}_i, \tau_i, r_{i,i'}\beta_i, \pi_i, \alpha_i).$$

Higher-Order States

- Emissions are context-dependent
- Adjust alphabet size to $|\Sigma|^{N+1}$
- Increases complexity
- In practice, N = 2 or 3

Complexity

$$O(nL|\Sigma|^{N+1})$$

$$P(L_u|a_1a_2) = \begin{cases} I(a_1a_2 \text{ matches } x_{u,1}x_{u,2}) & \text{if } u \text{ is a leaf} \\ \sum_{b_1b_2} P(b_1b_2|a_1a_2, t_v) P(L_v|b_1b_2) \sum_{c_1c_2} P(c_1c_2|a_1a_2, t_w) P(L_w|c_1c_2) & \text{otherwise} \end{cases}$$

Summary

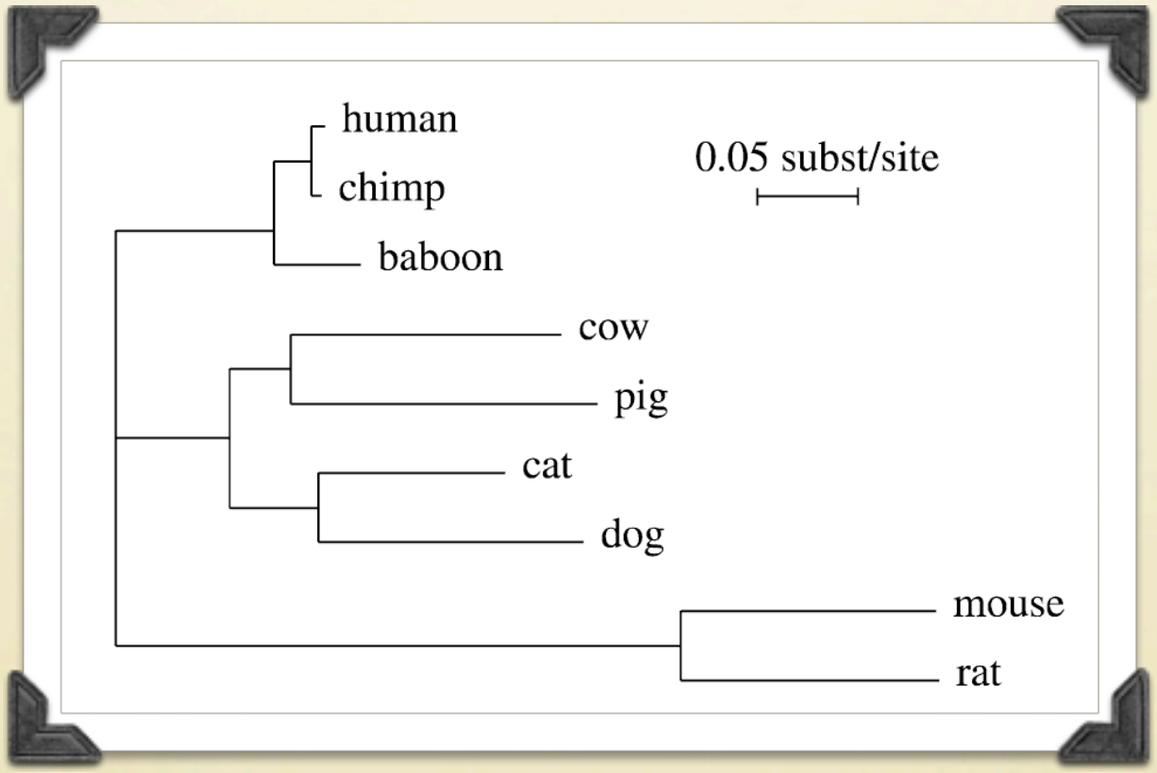
- Assume k rate, q functional categories
- HMM of order N
- Estimate transition probabilities of categories

- Compute $kq \times L$ emission probabilities
- Train autocorrelation
- Run Viterbi

The Results

Data

- Used portions of huge multiple alignment
- Trained using counting and annotations



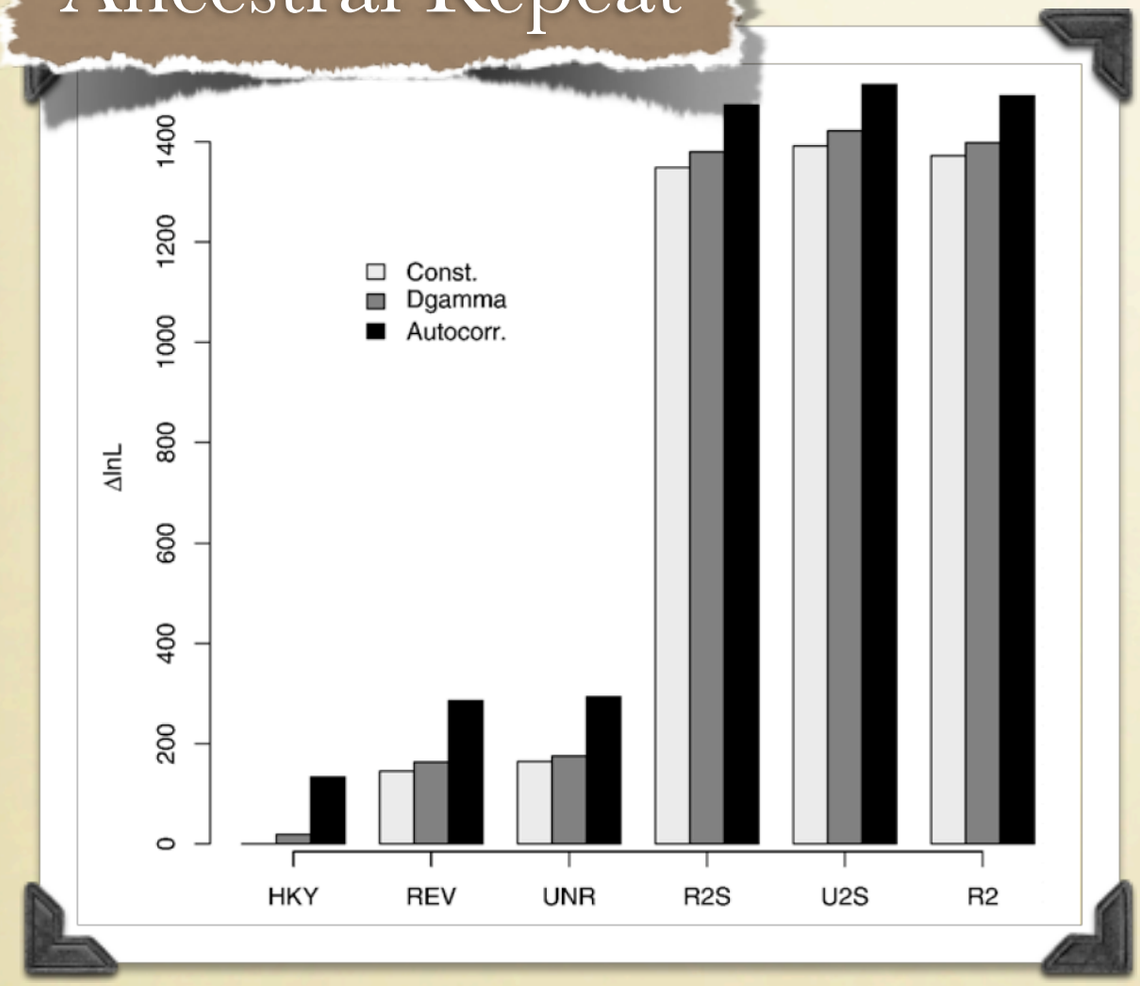
Test

- Compared models with likelihood ratio test (LRT)
- 5 substitution models (includes higher order)
  - REV, HKY, UNR, R2, R2S, U2S
- 3 rate variations -constant, gamma, autocorrelation

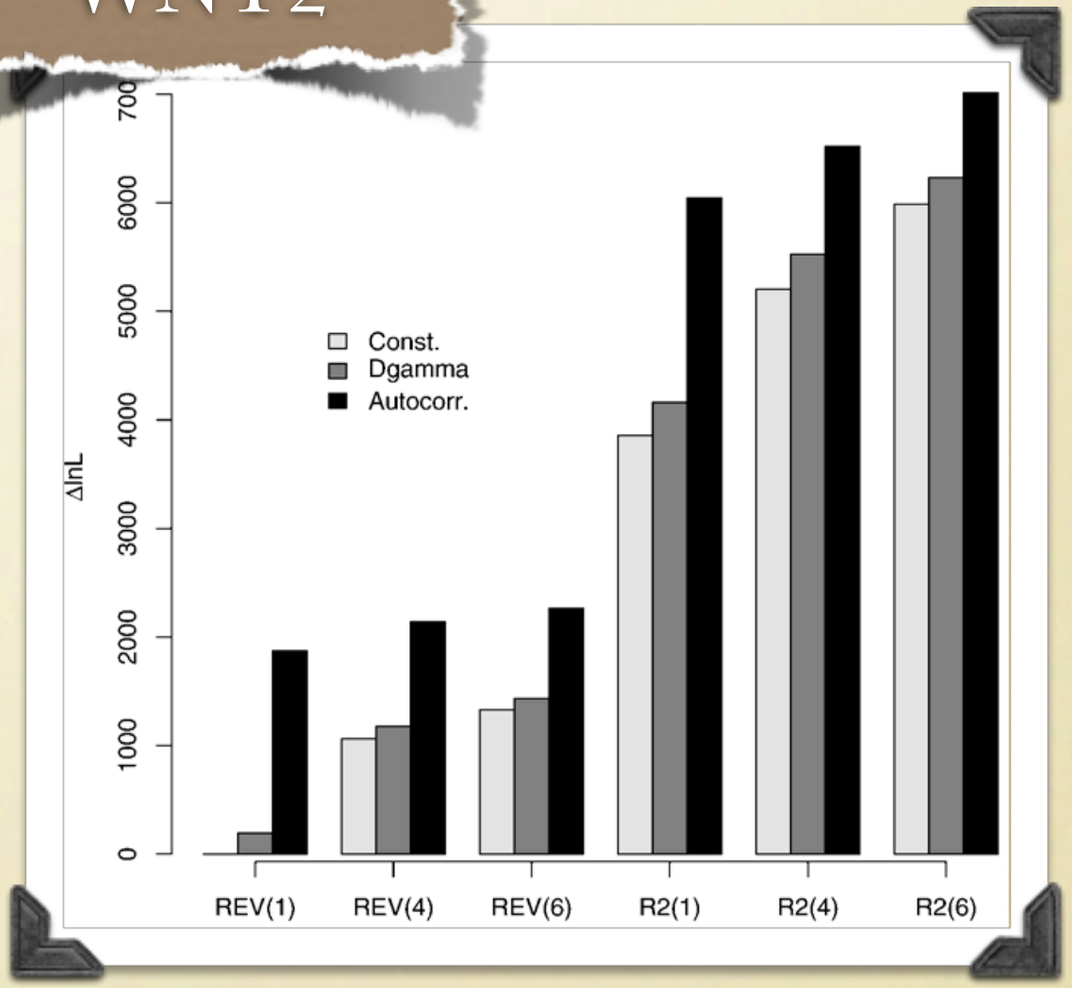
Results

- Higher states give largest boost

Ancestral Repeat



WNT2



Questions?