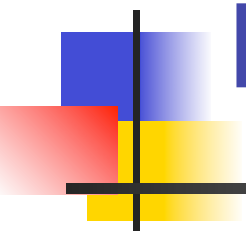
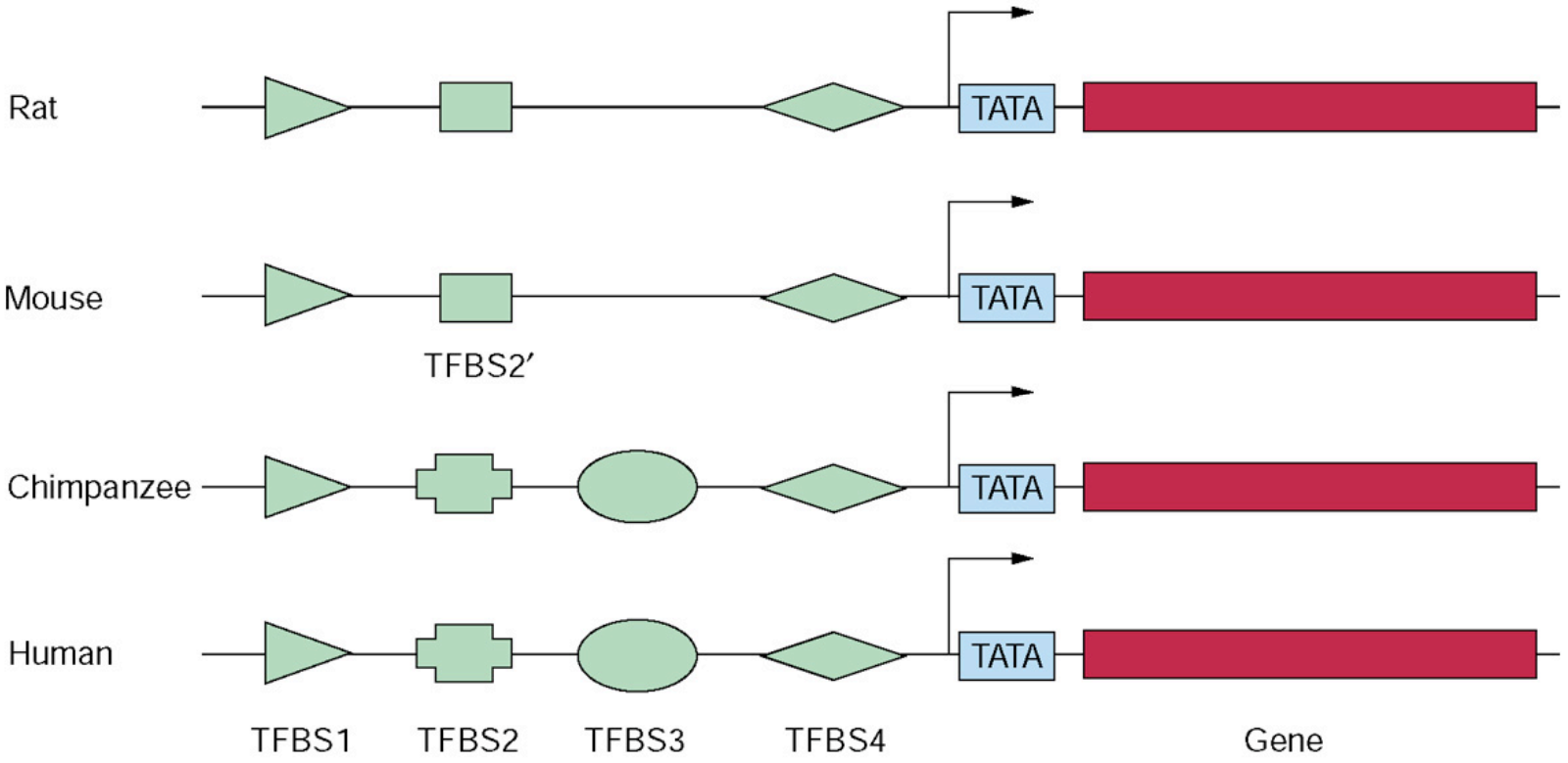


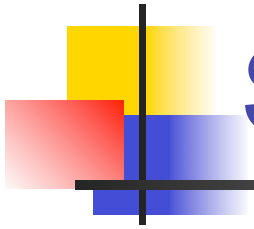
# Detecting Regulatory Elements with SAPF

Rahul Satija
University of Oxford

# Phylogenetic Footprinting

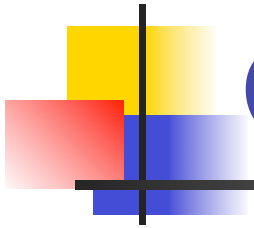



Zhang and Gerstein Journal of Biology 2003 2:11



# Single Alignment Approach

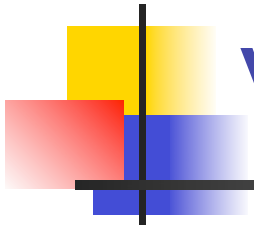
- PhastCons (Siepel et al., 2004)
  - Two-state HMM (fast/slow substitution)
  - Conditioned on a single alignment
  - Emission states are alignment columns
    - Slow state tends to emit more conserved columns
  - Can increase number of states, use indels for inference



# Conventional Challenges

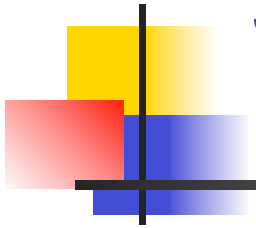
- TFBS not always perfectly conserved
- Single alignment approach
  - *Drosophila* TFBS detection (Stark et al, 2007)
    - 61% agreement from different alignments
  - Pollard et al., 2006
    - Alignment inaccuracies can result in significant errors for evolutionary studies
    - Comparative tools must properly accommodate alignment uncertainty

# Statistical Alignment/Rate Variation



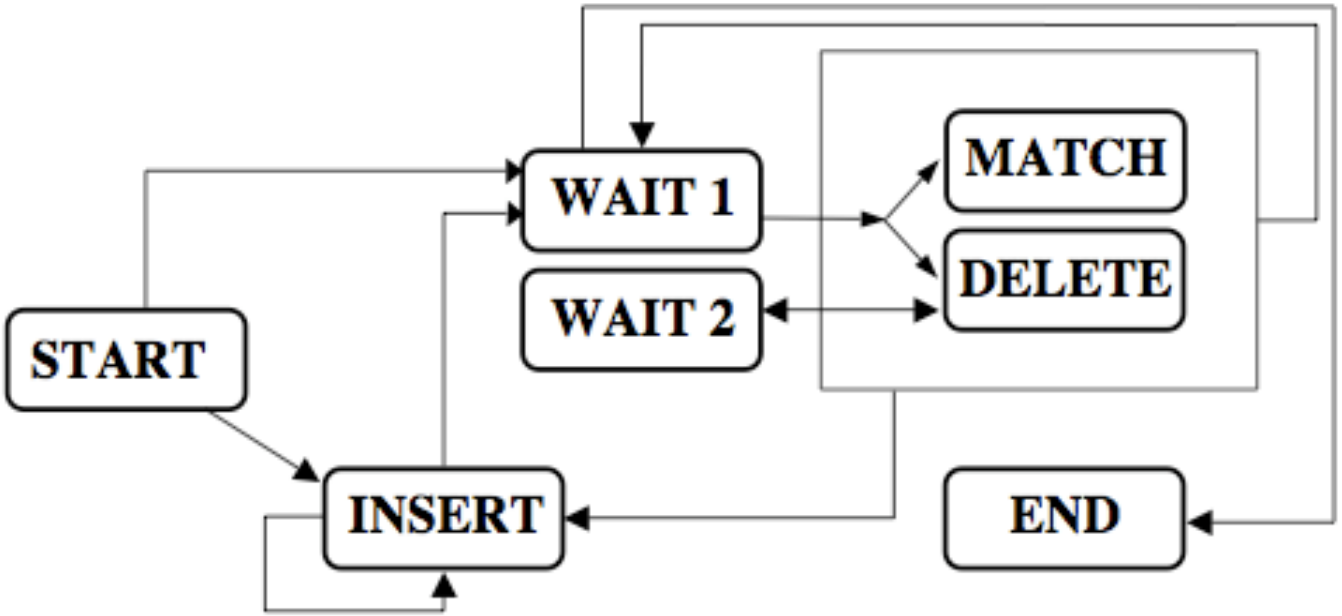
- TKF92: Local substitution rate variation
  - fast/slow fragments
  - only variation in substitution rate
- Arribas-Gil, Metzler, Plouhinec (2007)
  - modified TKF92 HMM implementing fast/slow fragments
    - slow fragments don't contain indels
  - successfully developed MCMC sampler

# Statistical Aligner, Phylogenetic Footprinter (SAPF)

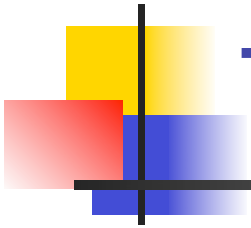


- Neutral evolution vs. purifying selection
  - Fast/slow fragments evolve under same model with rates of substitution, indel
- Analyze multiple species (4-5 max), related by a known phylogeny
  - HMM transducers (Holmes, 2003, 2007)
- Functional element predictions made from *distribution* of alignments
  - Correctly accounts for uncertainty

# SAPF Branch HMM




- Branch HMM represents evolutionary process on each branch
- Second wait state enables delete to self-transition



# The Multiple HMM

- PhyloComposer used to generate multiple sequence HMM
  - Each MHMM state represents collection of branch HMM states
  - Emission states are alignment columns
- Double number of states
  - Corresponds to creating an HMM on the root, alternating between fast/slow
  - Fixes Fast/Slow annotation on a column

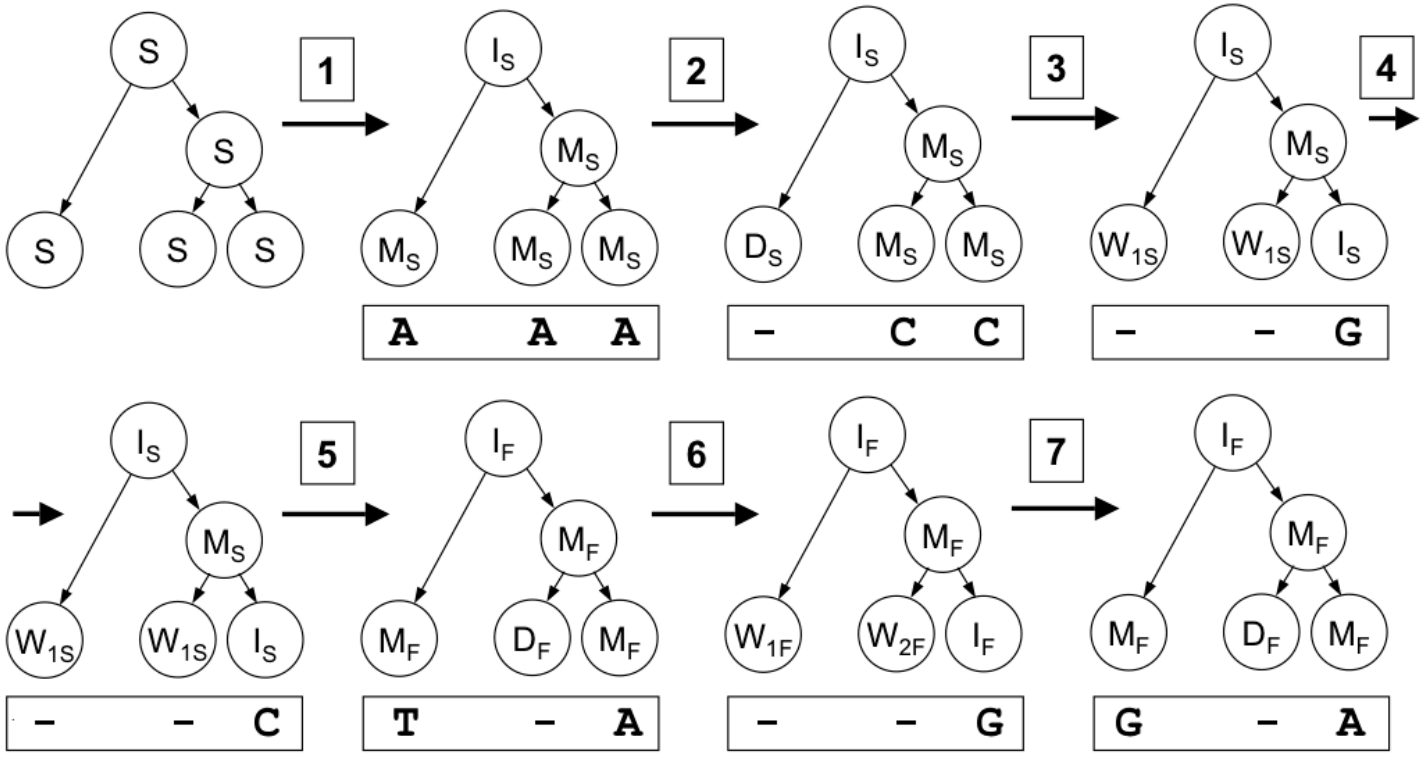
# SAPF HMM Parameters

| Parameters | Description |
|---|---|
| $\lambda_{fast}, \lambda_{slow}$ | Birth rates for links in fast/slow states |
| $\mu_{fast}, \mu_{slow}$ | Death rates for links in fast/slow states |
| $\sigma_{fast}, \sigma_{slow}$ | Insertion state self-transition probability (sets expected indel length) in fast/slow states |
| $s_{fast}, s_{slow}$ | Nucleotide substitution rates for fast/slow states |

- Baum-Welch followed by EM used to calculate ML estimates for all parameters

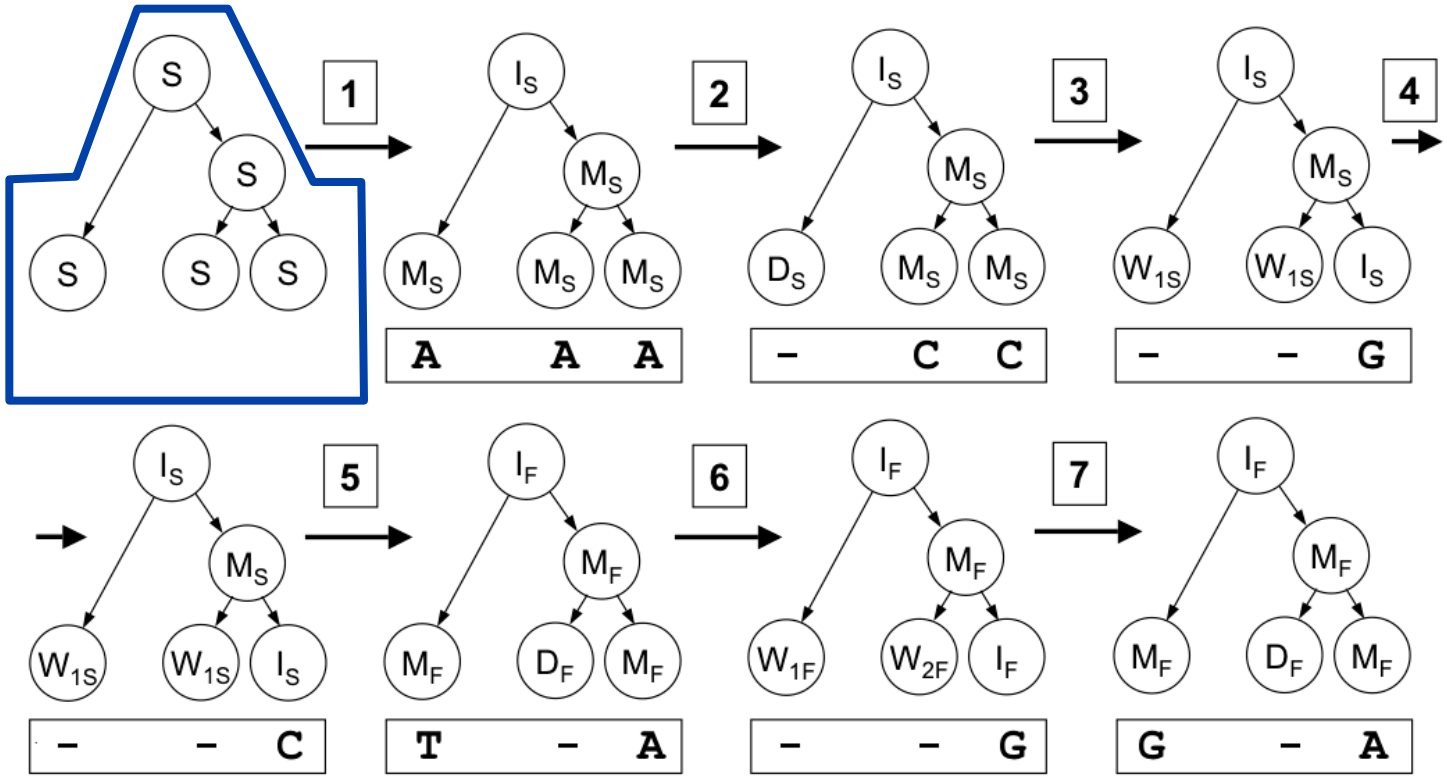
# SAPF HMM

```
Seq1    ACGCAGA
Seq2    AC-----
Seq3    A---T-G
```



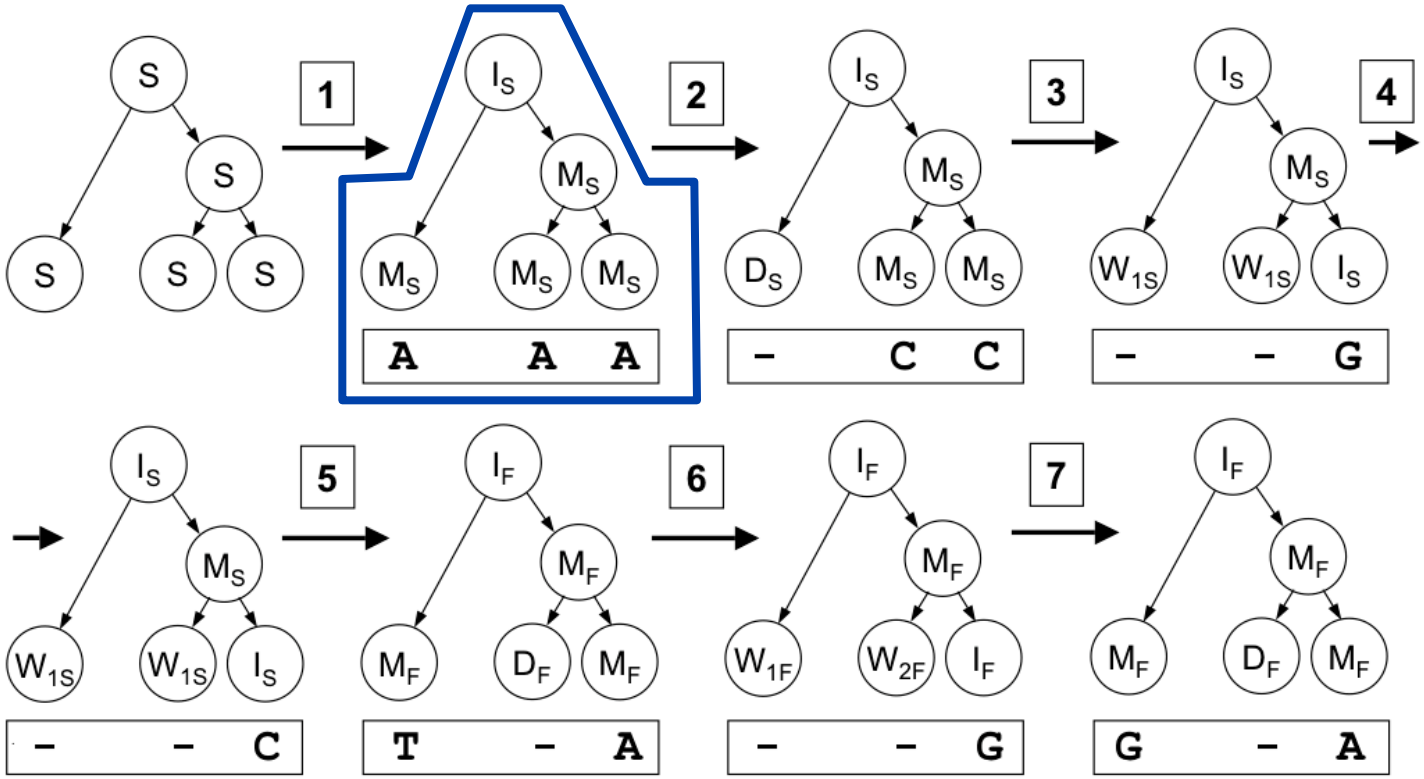
# SAPF HMM

```
Seq1    ACGCAGA
Seq2    AC-----
Seq3    A---T-G
```



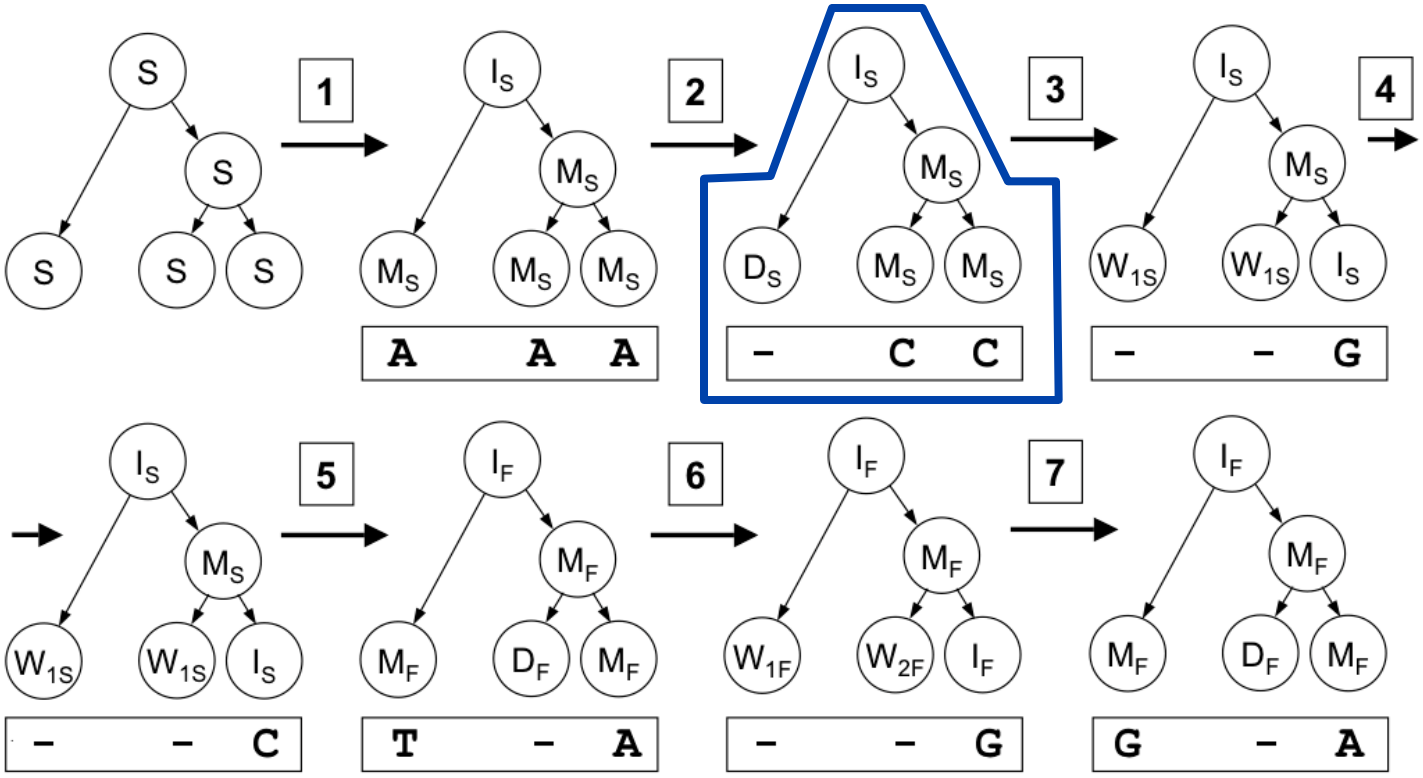
# SAPF HMM

```
Seq1    ACGCAGA
Seq2    AC-----
Seq3    A---T-G
```



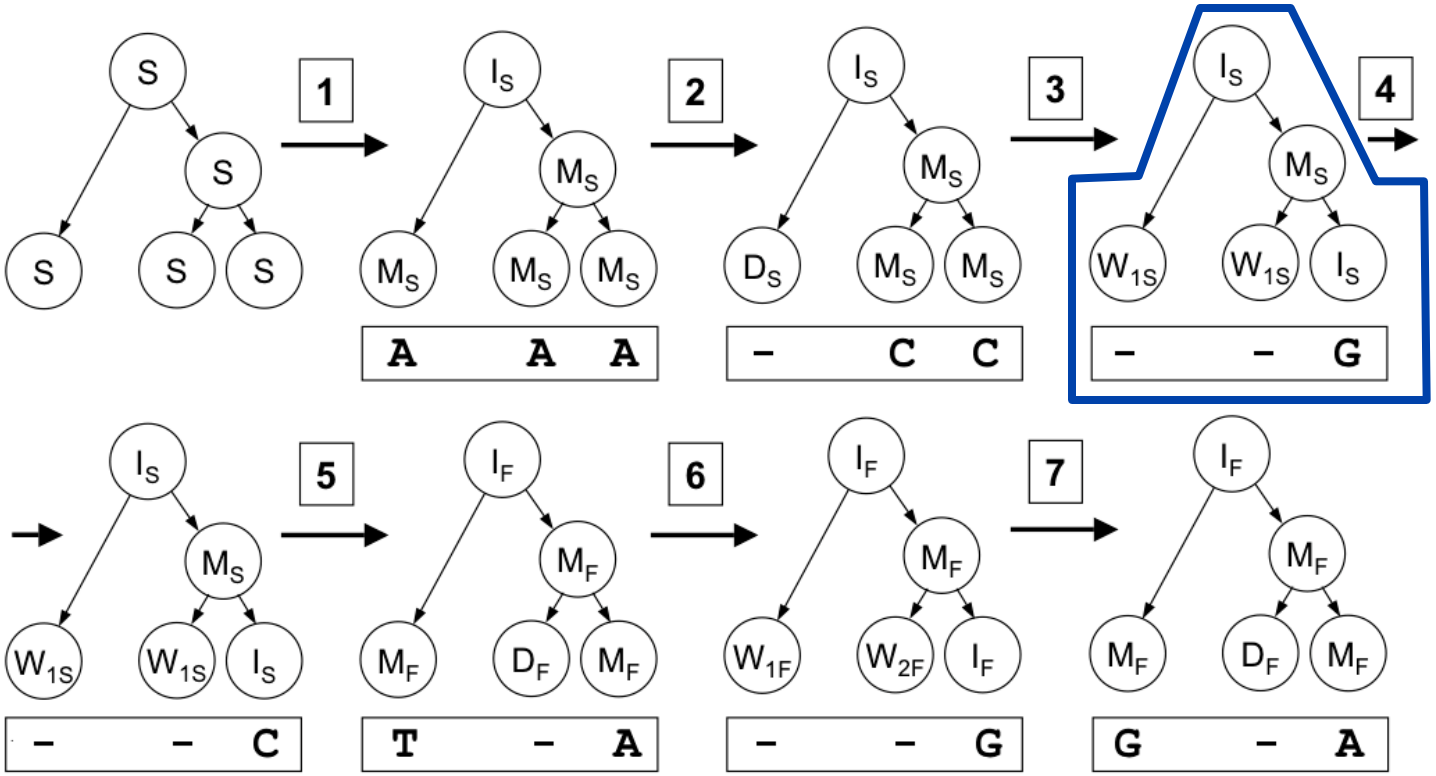
# SAPF HMM

```
Seq1    ACGCAGA
Seq2    AC-----
Seq3    A---T-G
```



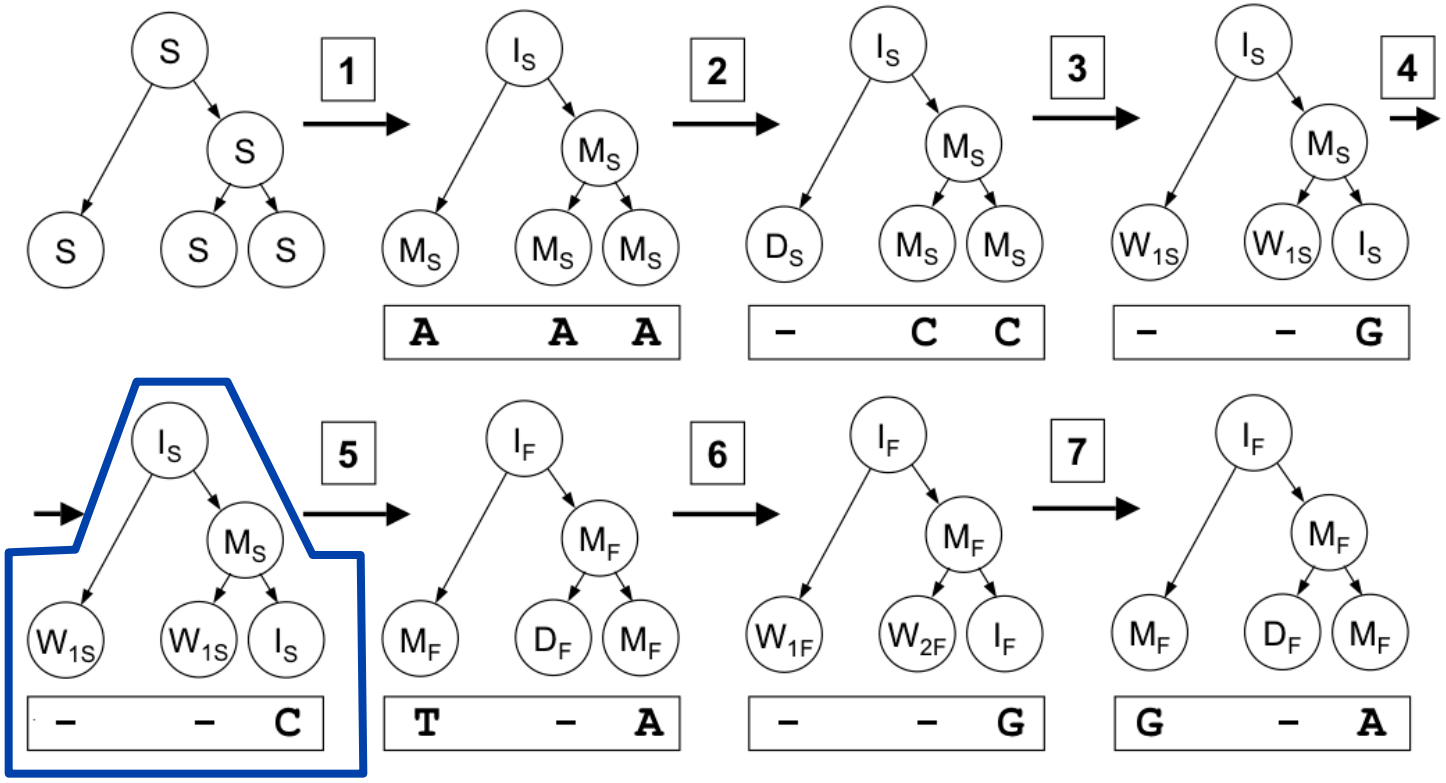
# SAPF HMM

```
Seq1    ACGCAGA
Seq2    AC-----
Seq3    A---T-G
```



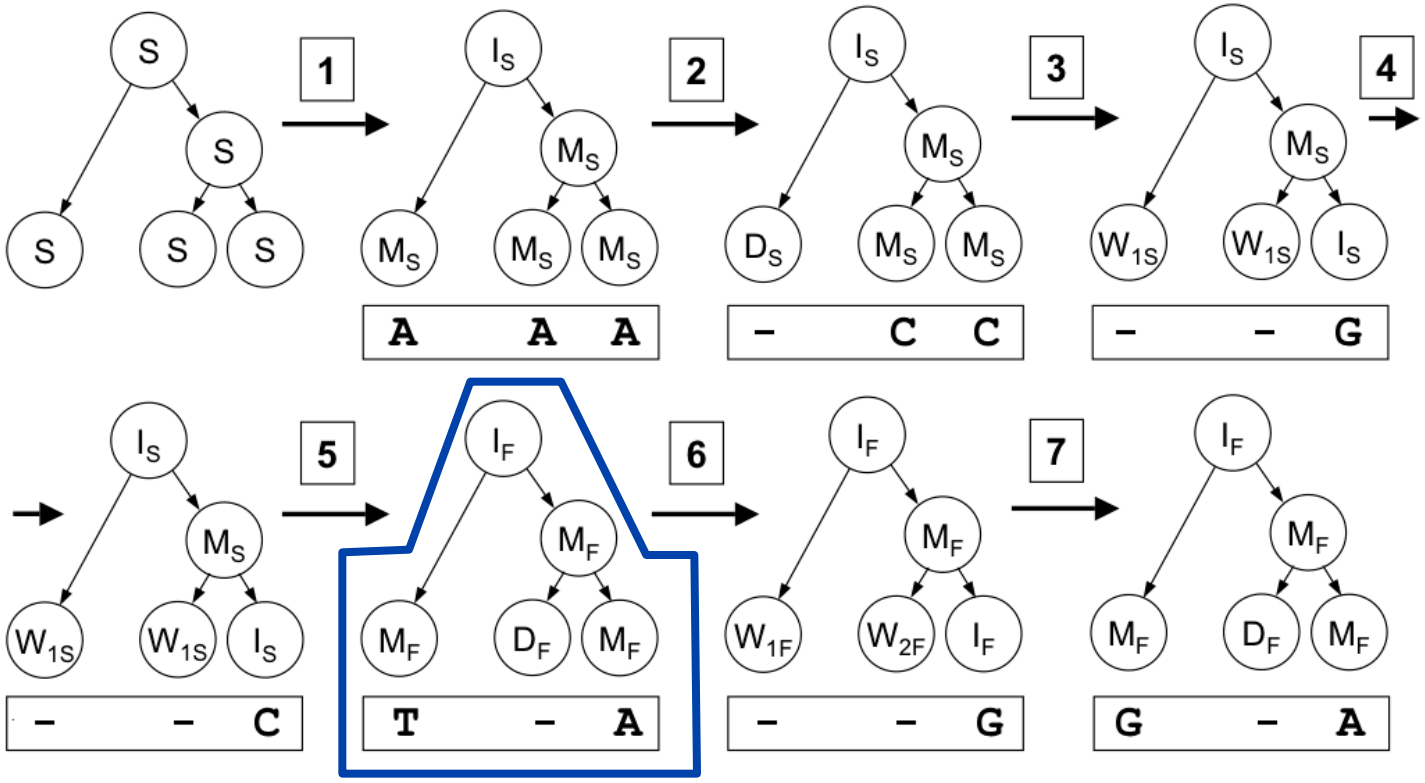
# SAPF HMM

```
Seq1    ACGCAGA
Seq2    AC-----
Seq3    A---T-G
```



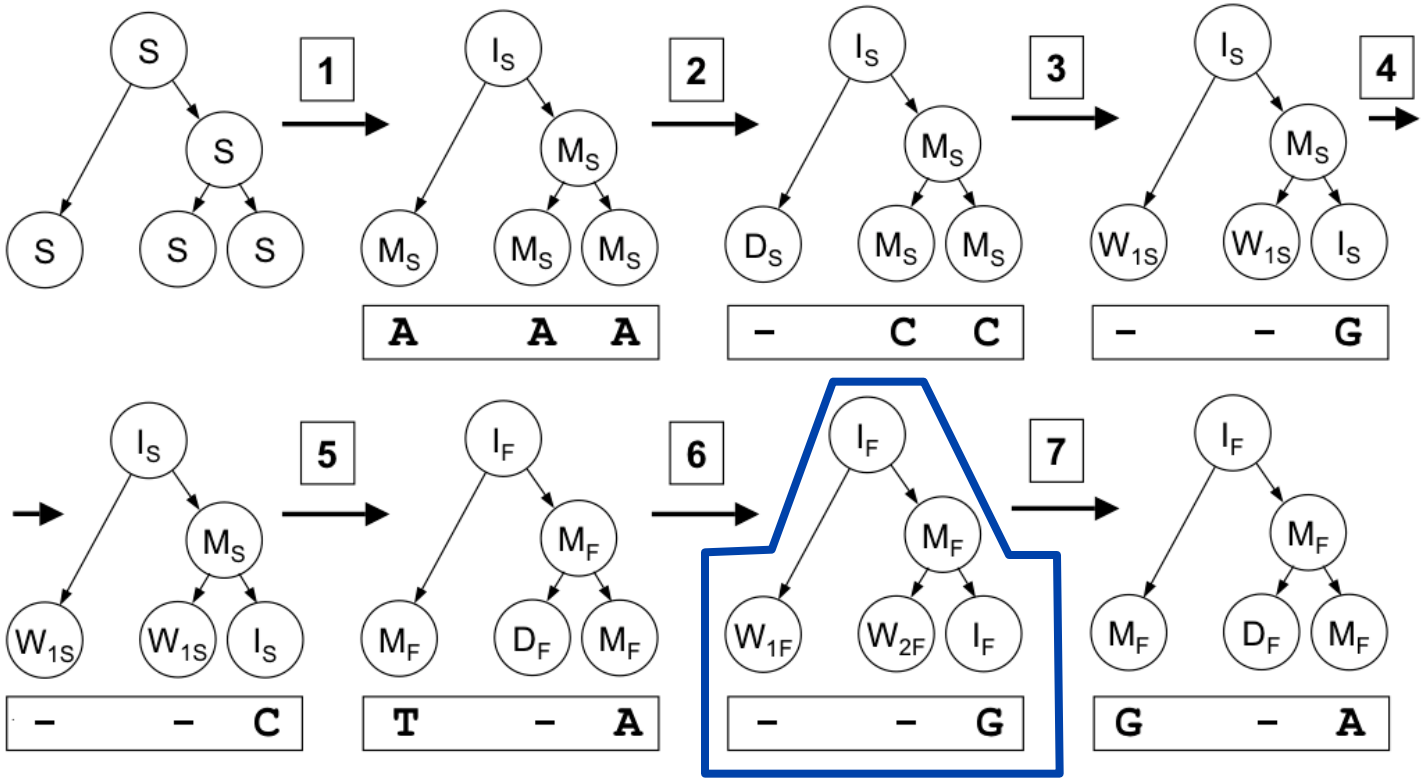
# SAPF HMM

```
Seq1    ACGCAGA
Seq2    AC-----
Seq3    A---T-G
```



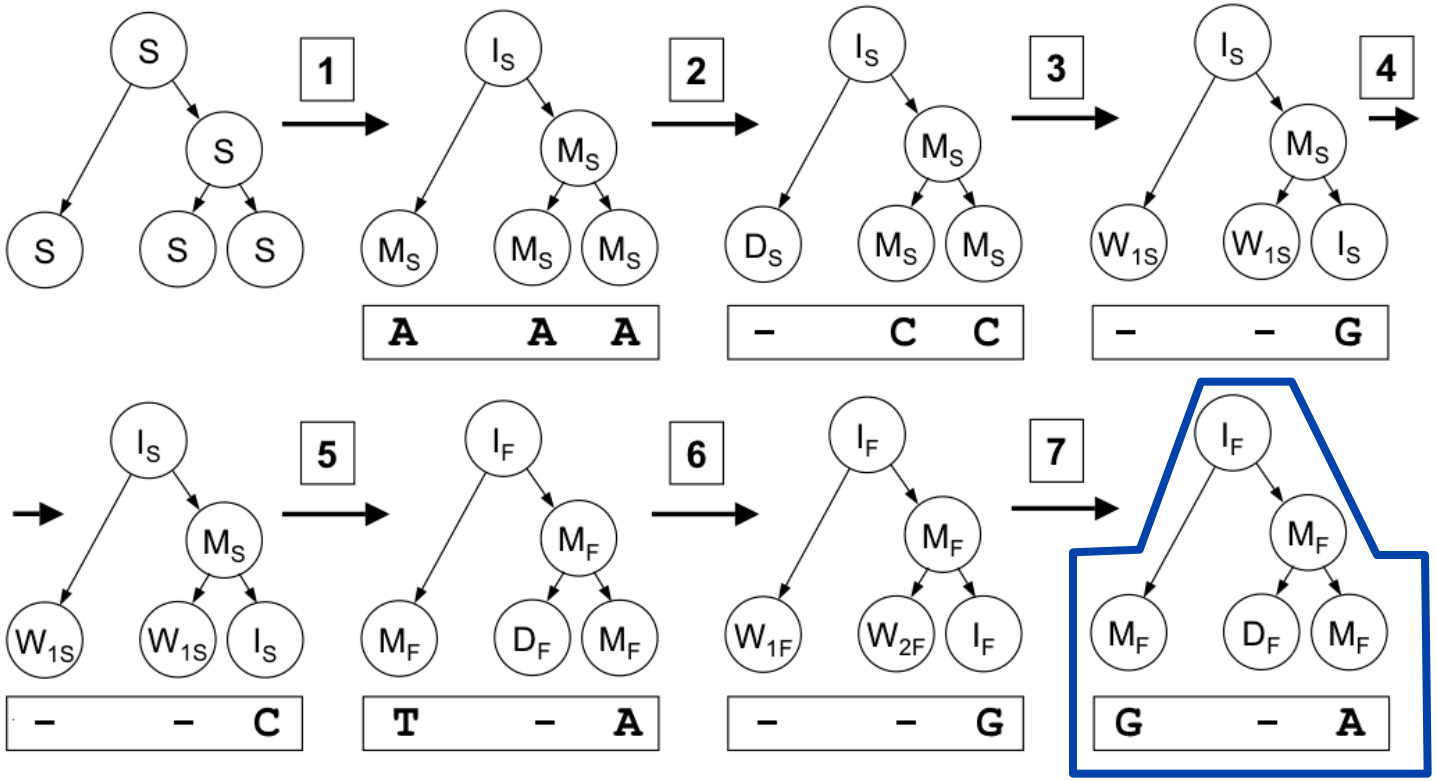
# SAPF HMM

```
Seq1    ACGCAGA
Seq2    AC-----
Seq3    A---T-G
```

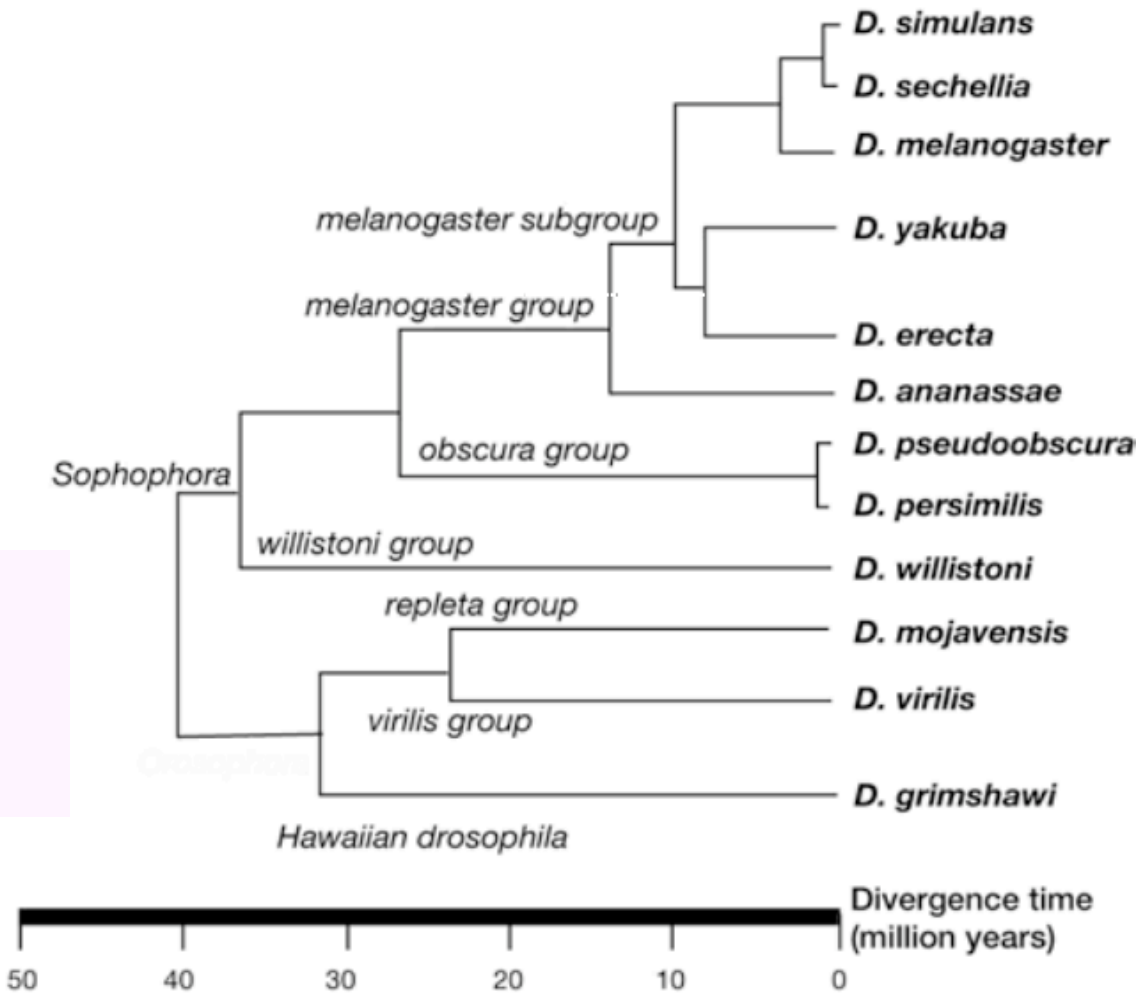


# SAPF HMM

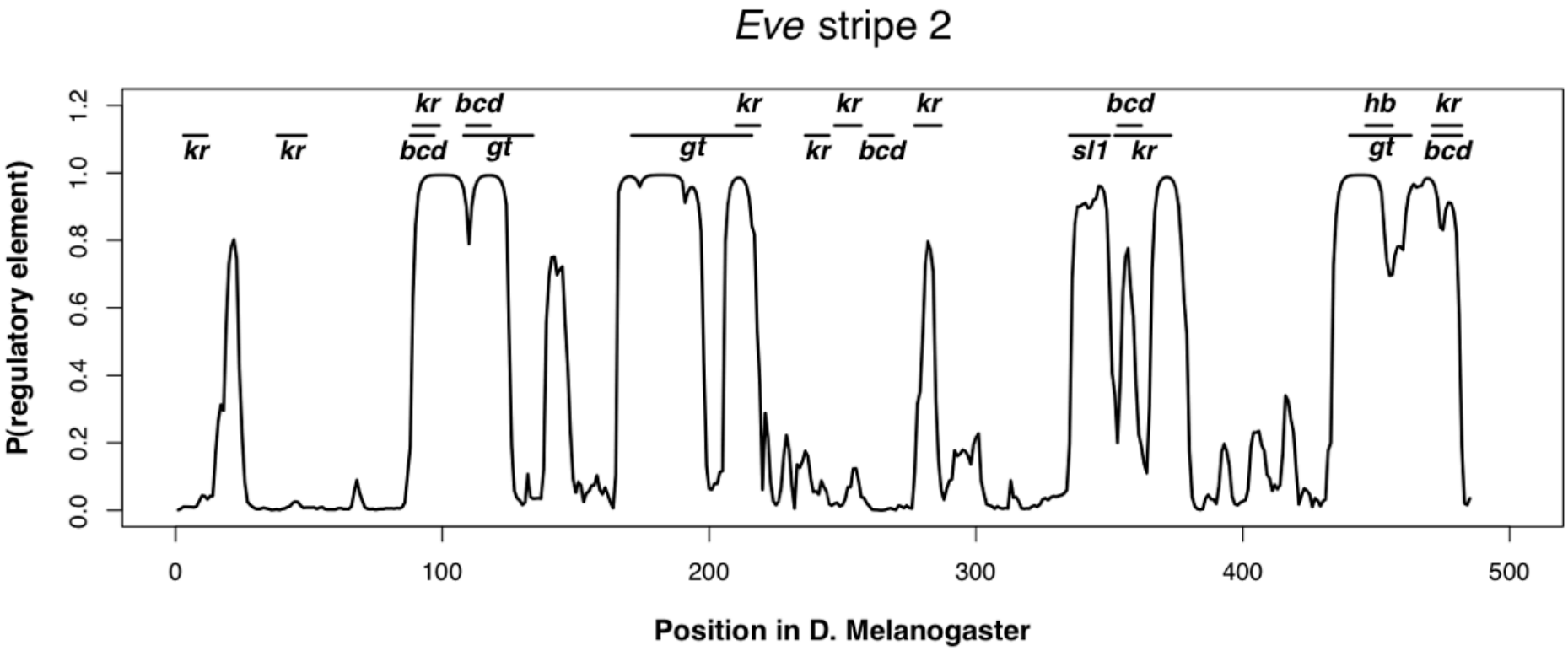
```
Seq1    ACGCAGA
Seq2    AC-----
Seq3    A---T-G
```



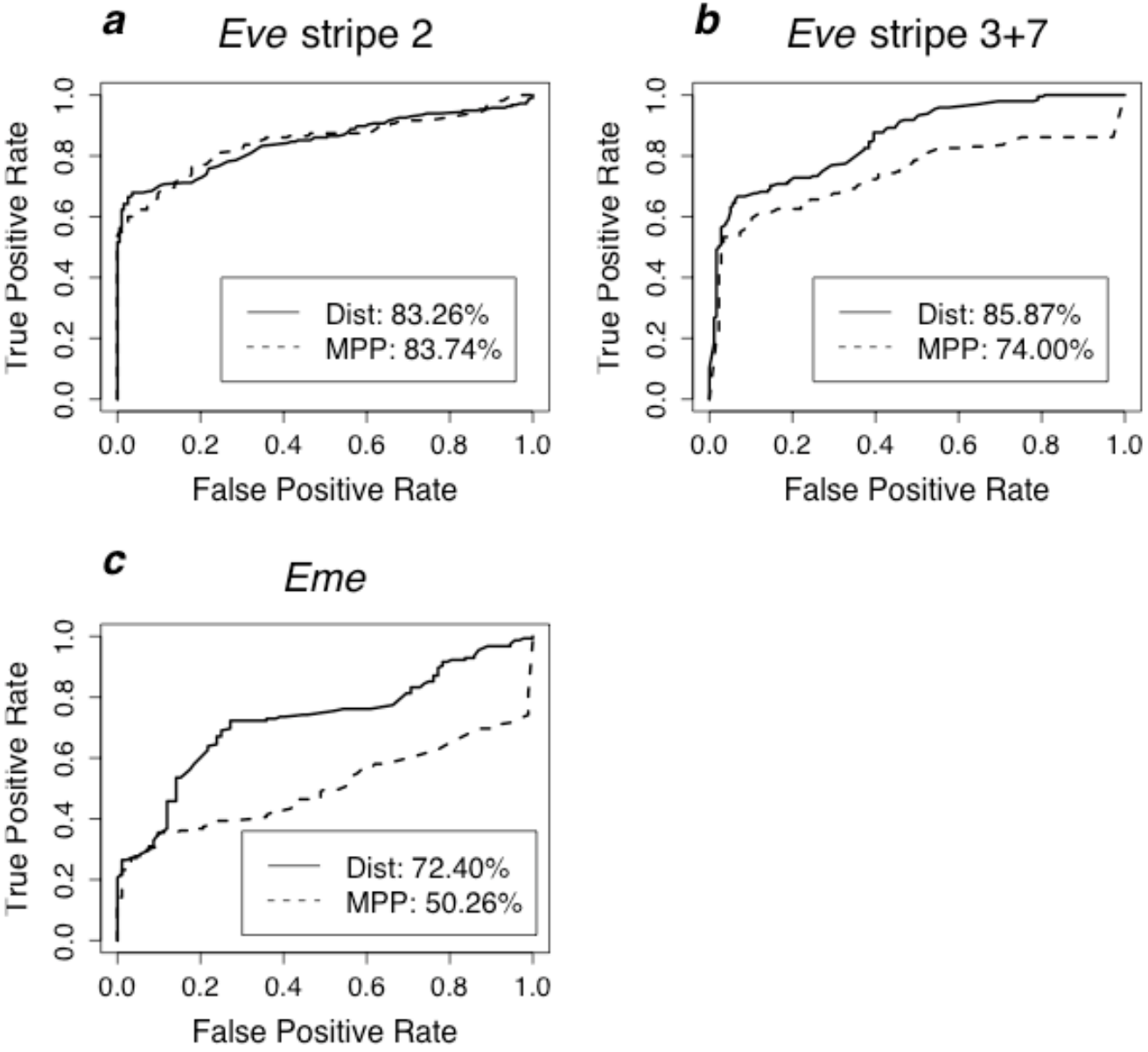
# *Drosophila* Genome Data



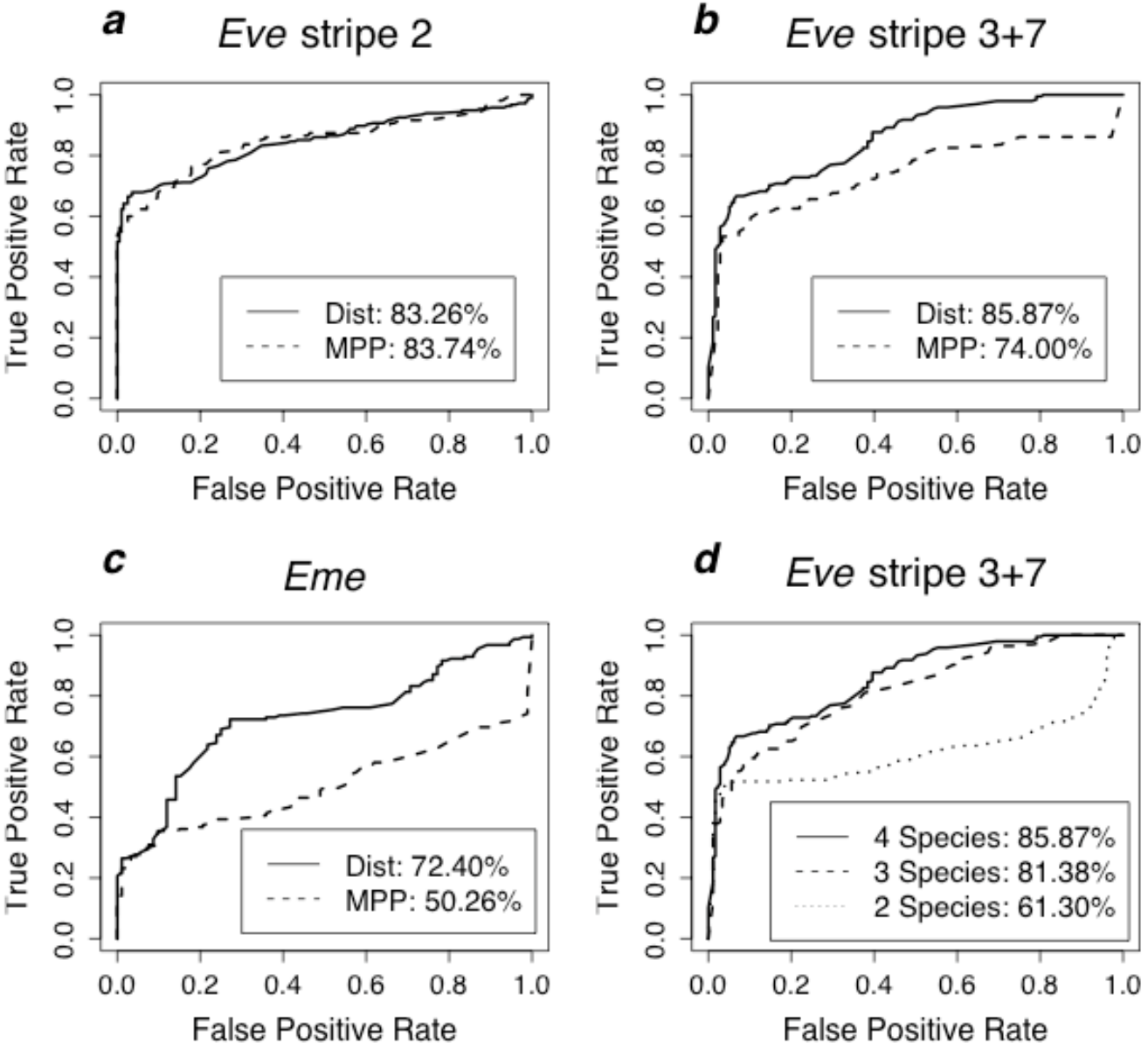
# SAPF Results



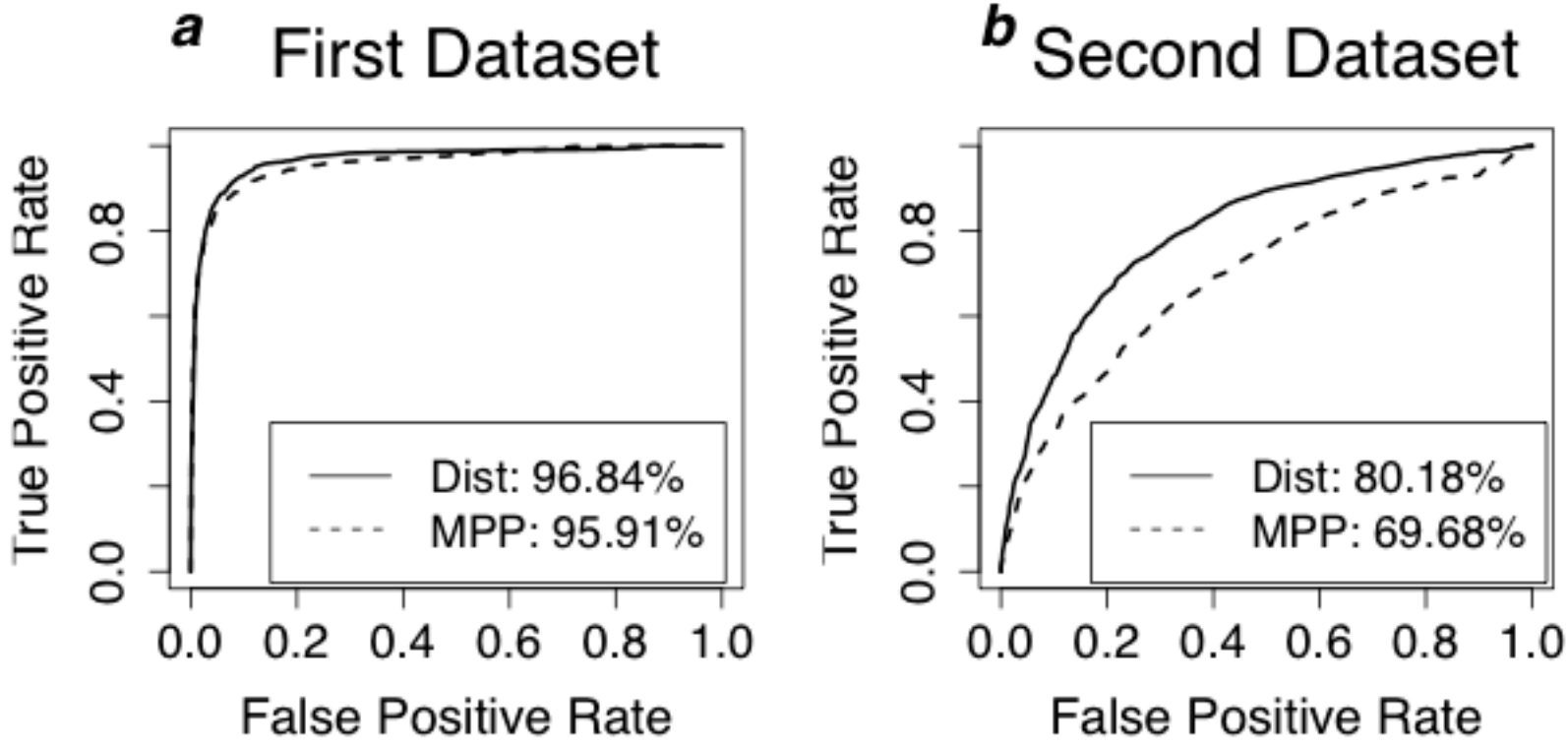
# SAPF Results



# SAPF Results

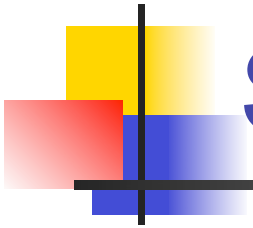


# SAPF Results



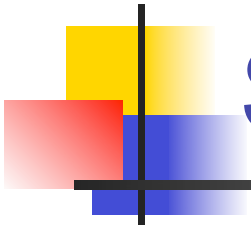
# Alignment Uncertainty

**Hunchback binding site, *eve* stripe 3+7 CRM**



# Summary

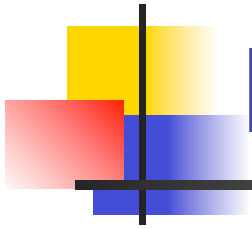
- Transducer framework allows for multiple sequence analysis
- State doubling enables PF
- Integrating over alignments can improve performance
- Increase speed, analyze more data
  - Aim is to analyze 12-16 species
  - MCMC approach
    - Collaboration with Istvan Miklos, Adam Novak



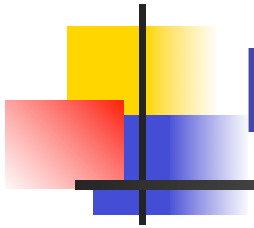
# StatAlign Package

- Bayesian co-estimation of alignment, phylogeny
  - partial alignment sampler
    - TKF92 represents each branch alignment
    - Kimura3, Jukes-Cantor, etc.
    - all parameters are sampled
  - phylogeny sampling
    - branch lengths, tree topology
  - Java GUI

# Adding Rate Heterogeneity



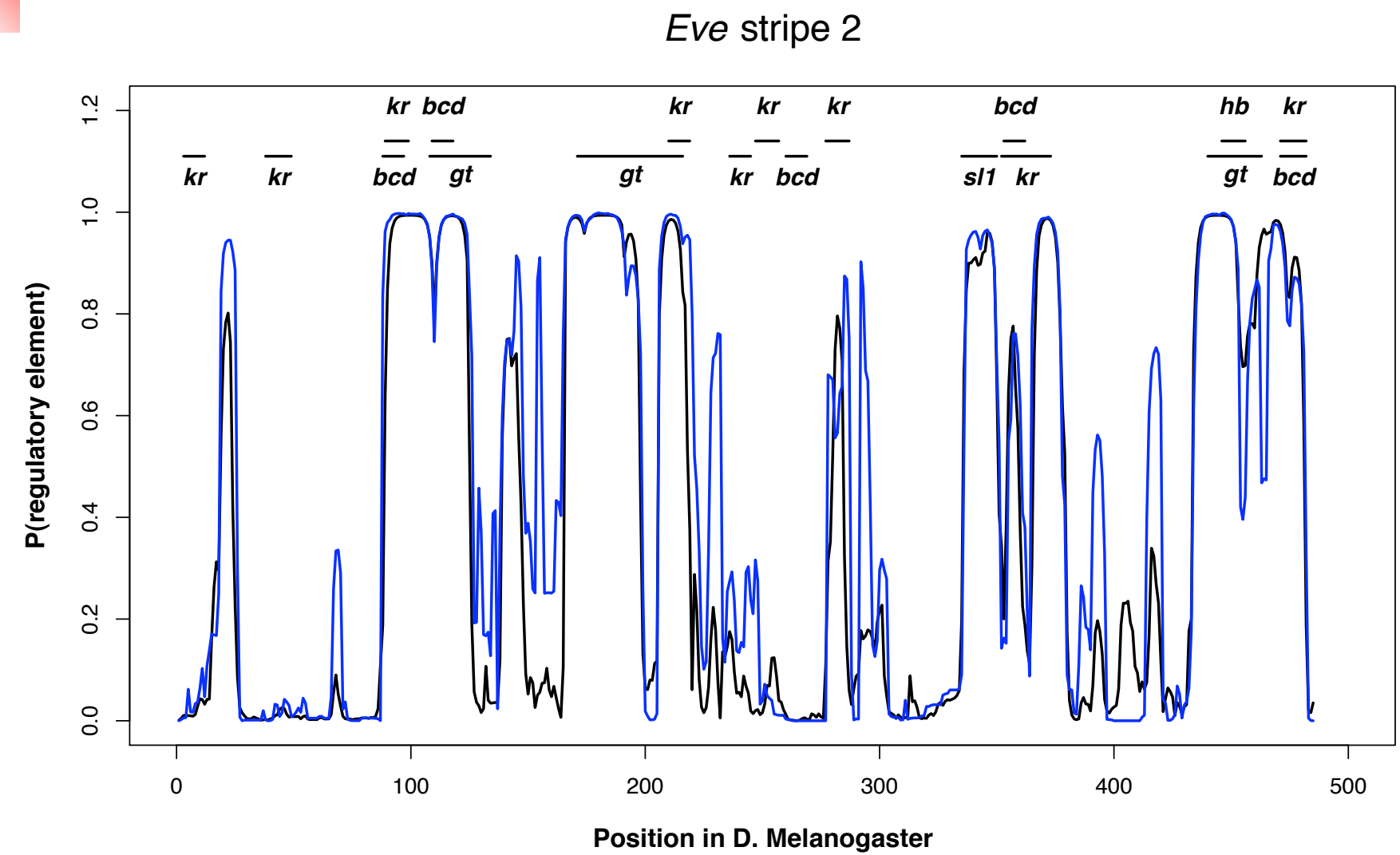
- Fix phylogeny
- Split into fast and slow fragments
  - Add root HMM
  - Sample new parameters
    - expected fragment length
    - branch length scaling (substitution, indel)
  - Sample fragment locations
    - fragment split (create new fragment boundary)
    - fragment merge (delete fragment boundary)
    - adjust fragment boundary



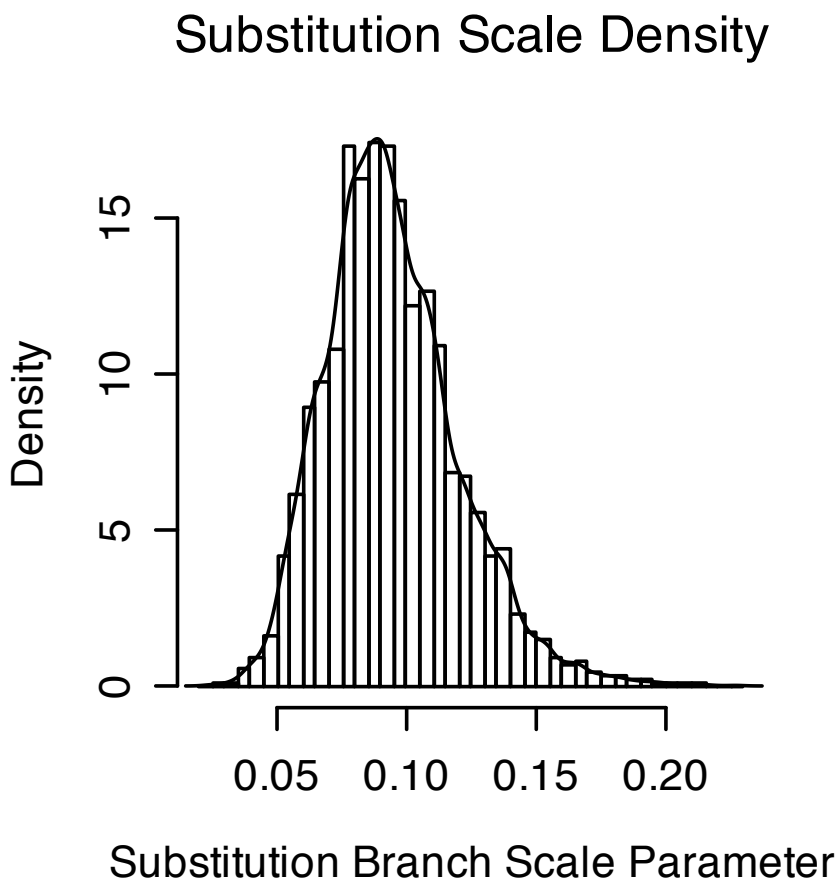
# MCMC Challenges

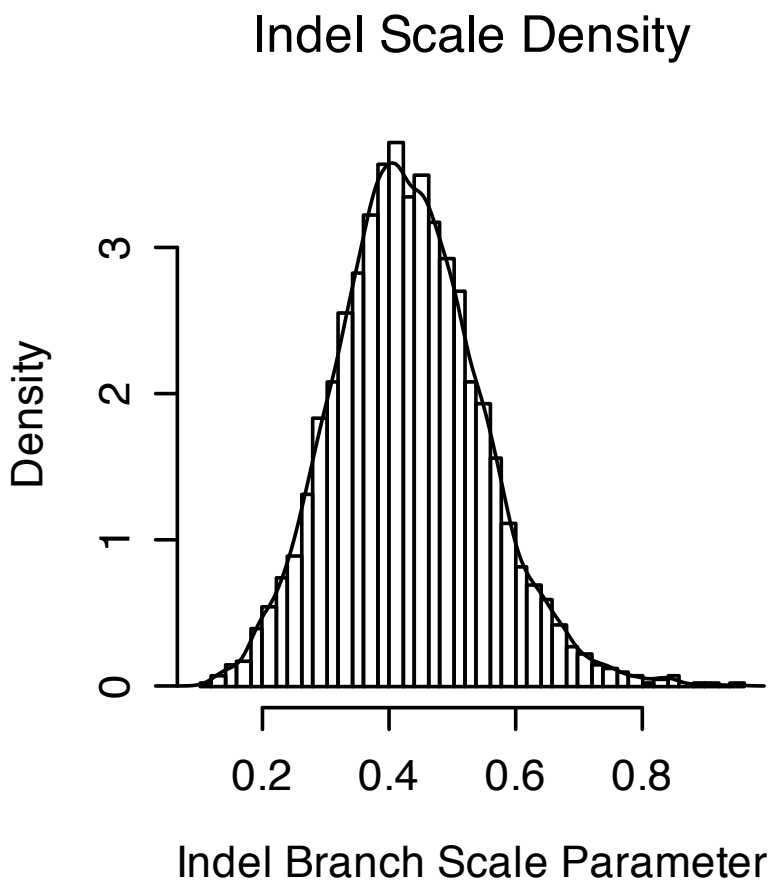
- Low probability regions
  - alignment, fragment boundaries
  - must visit occasionally, not often
    - visiting rarely causes skewed MH ratios
- Mixing
  - Allowing homology between sequences in different fragments

# MCMC Results (4 species)

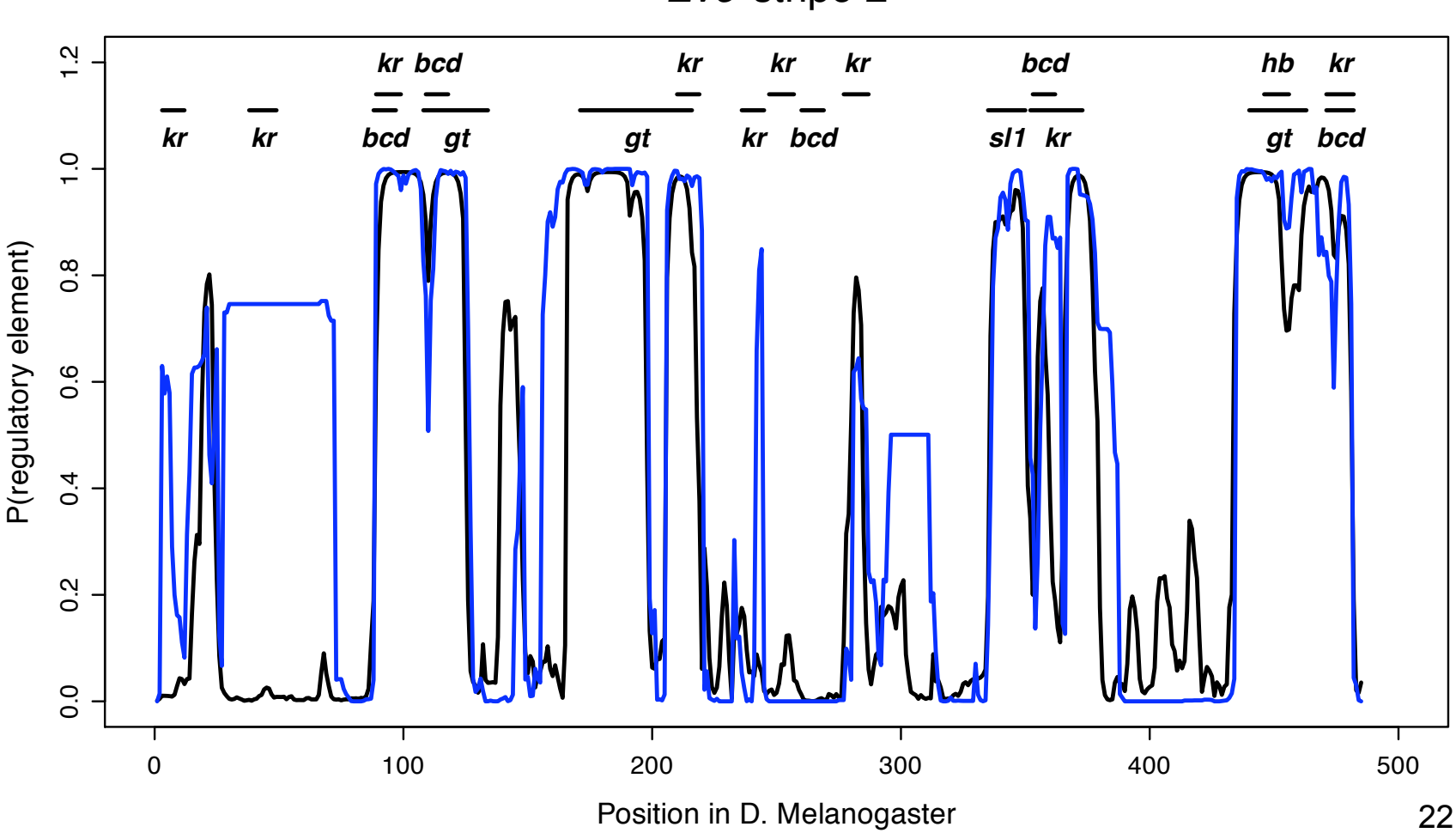


# Scale Parameters



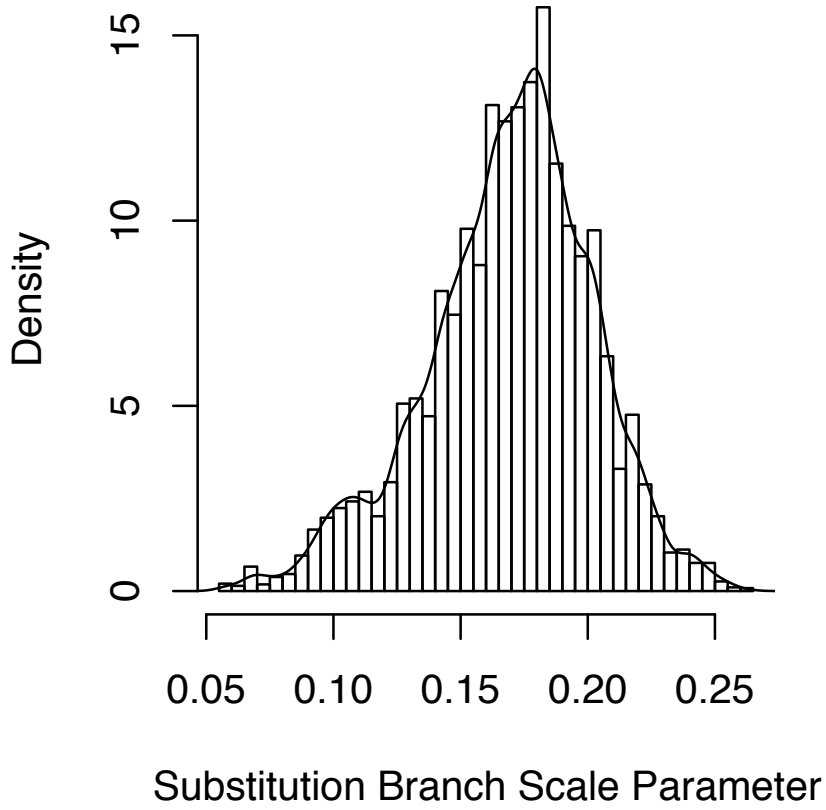



# MCMC Results (10 Species)

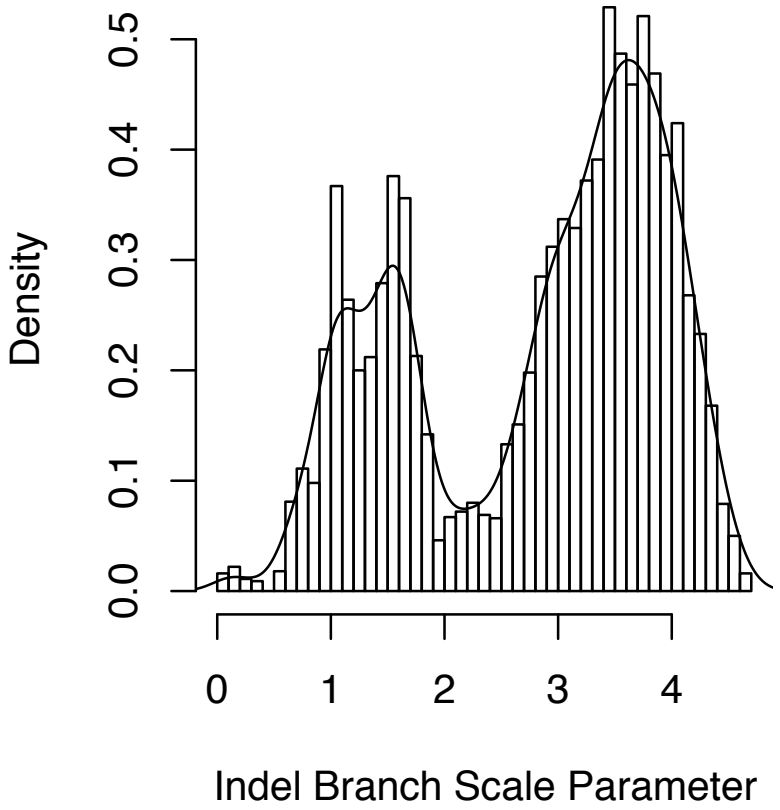


# Scale Parameters

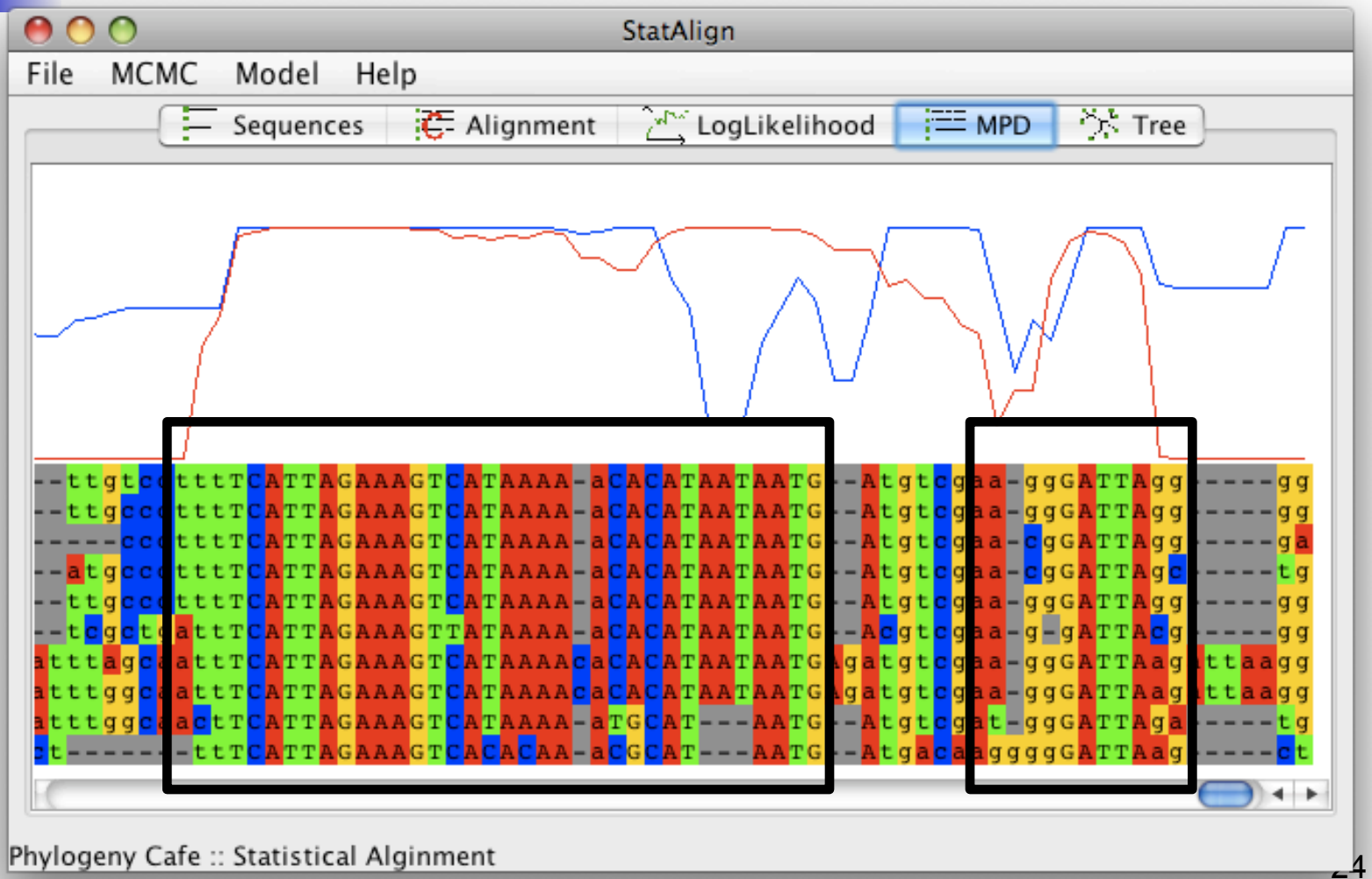
**Substitution Scale Density**



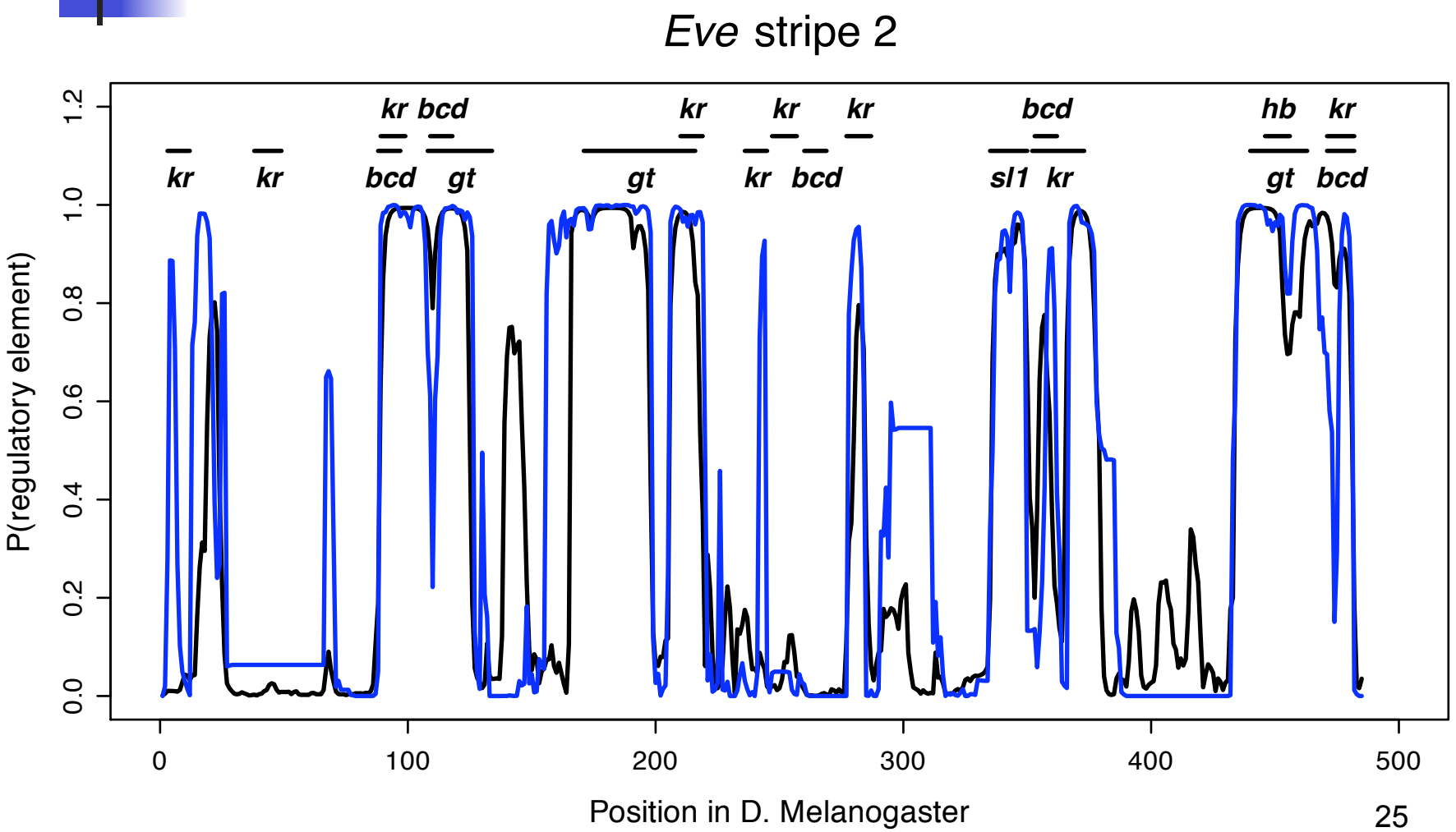

**Indel Scale Density**

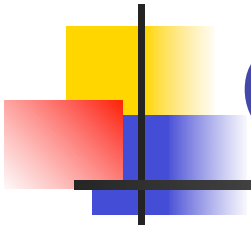


# Indels in binding site regions



# MCMC Results (10 Species)



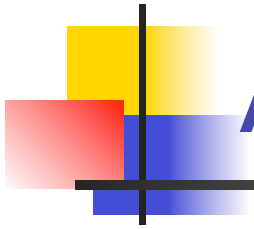


# Conclusions

- Modified StatAlign to incorporate rate heterogeneity
  - Gives similar results to dynamic programming
  - Hopefully enables us to analyze more data
- Indels in fast/slow fragments
  - Both frequency and length distributions are important in inference
  - Especially important with larger numbers of species

# Advanced Models of TFBS

- Position state weight matrix
  - Accurate modeling of known binding sites
  - Incorporate into SAPF HMM
    - Improve quality of predictions, alignment
- Accurately model binding site gain/loss
  - Better understanding of TFBS evolution



# Acknowledgements


- University of Oxford: Jotun Hein, Rune Lyngsoe, Gerton Lunter
  - Istvan Miklos, Adam Novak
- UC Berkeley: Lior Pachter, Ian Holmes