Combining Statistical Alignment and Phylogenetic Footprinting to Detect Regulatory Elements

COMP 571

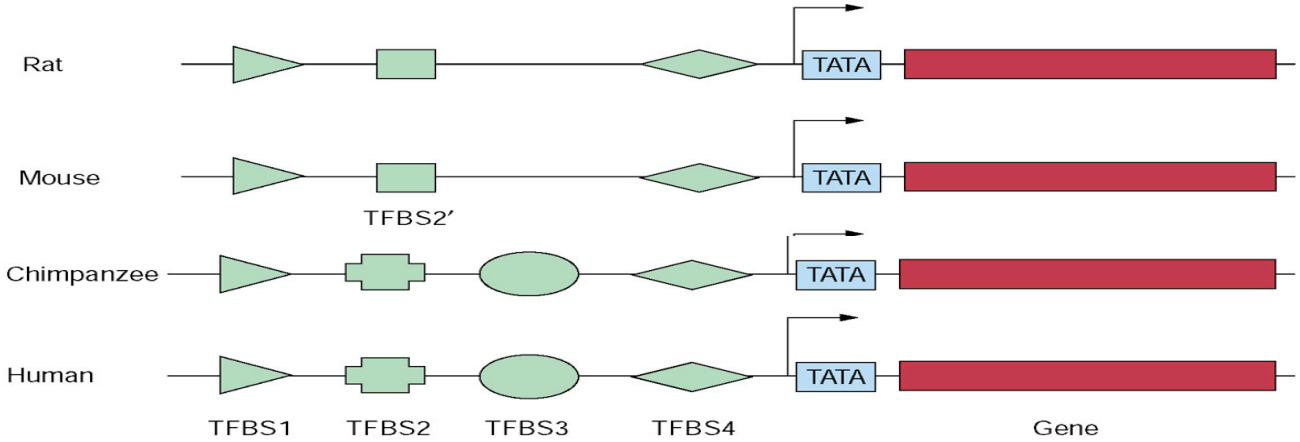
Overview

- Phylogenetic Footprinting
  - HMM based (Siepel et al., 2004))
- Statistical Alignment (Multiple Sequence)
  - Transducers and Branch HMM (Holmes, 2003)
- Statistical Alignment and Phylogenetic Footprinting (SAPF)
- Experimental Results
  - Verification of SAPF results on Drosophila data set
  - Comparison with single sequence HMM PF
- Summary
- Limitations

Phylogenetic Footprinting

Phylogenetic Footprinting

- A technique of identifying regulatory elements:
  - Consider a set of orthologous noncoding sequences from a group of related species
  - Find unusually well conserved regions



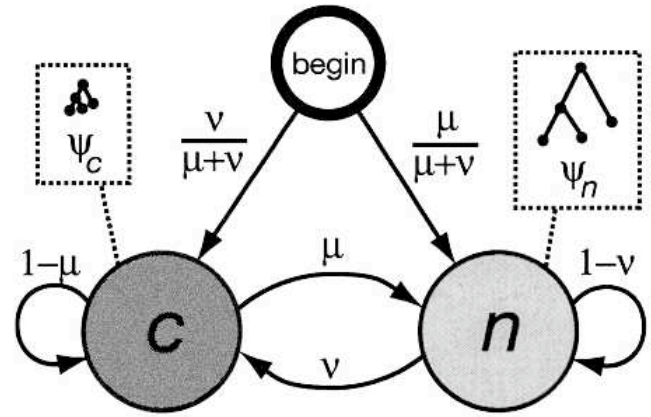
Phylogenetic Footprinting

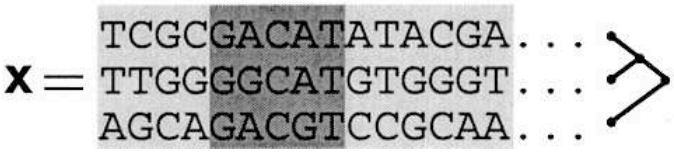
- Formal definition:
  - Given: a set of orthologous sequences S_1, …,S_n from n different species, a guide tree relating these species, and an integer k
  - Problem: Find a set of substrings s_1,…,s_n of S_1,…,S_n, respectively, each of length k, such that the parsimony score of s_1,…,s_n is minimized. The substrings s_1,…,s_n correspond to the region that has undergone the fewest mutation

Phylogenetic Footprinting using HMM

PhastCons (Siepel et al., 2004)

- Conditioned on a single alignment
- Two-state HMM (fast/slow substitution)
- Emission states are alignment columns
  - Slow state tends to emit more conserved columns





Limitations of PhastCons

- 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

Statistical Alignment

Statistical Multiple Alignment (PhyloComposer, Holmes, 00)

- Hidden Markov Model based multiple alignment
  - Given $\Sigma=\{A,C,T,G\}$ and N sequences, construct an HMM with $(|\Sigma|+1)^N$ number of states; each state corresponds to a column in multiple alignment
  - Emission is an N-dimensional vector in which each entry is a sequence of length 0 or 1
  - t(i,j) is the transition probability from state i to state j

Evolutionary HMM (PhyloComposer, Holmes, 00)

- A multiple HMM constructed using two components: a Guide tree and a Branch HMM (like Predictive Alignment)
- Branch HMM associated with the branch of the guide tree is a two sequence transducer with
  - State types: START, WAIT, INSERT, MATCH, DELETE, END
  - Transition probabilities are function of time

Branch HMM Transducer

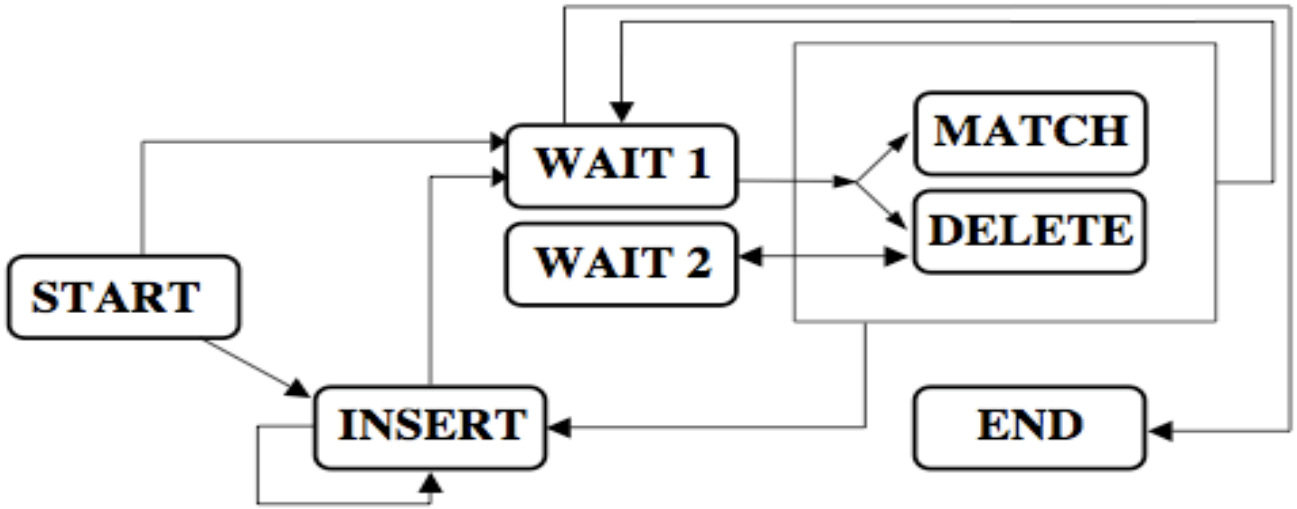
- Consists of an input tape (an ancestral sequence X) and an output tape (a descendant sequence Y)
- The path probability of Π is the conditional likelihood P(Π, Y|X) instead of P(Π,X,Y) in pairwise HMM
- Π in pairwise HMM represents two sequences evolving from a common ancestor where Π in transducer represents the input sequence evolving into the output sequence

Statistical Alignment and Phylogenetic Footprinting

Statistical Aligner, Phylogenetic Footprinter (SAPF)

- Neutral evolution (faster divergence) vs. purifying selection (slower divergence)
  - Fast/slow fragments evolve under same model with rates of substitution, indel
- Analyze multiple species 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

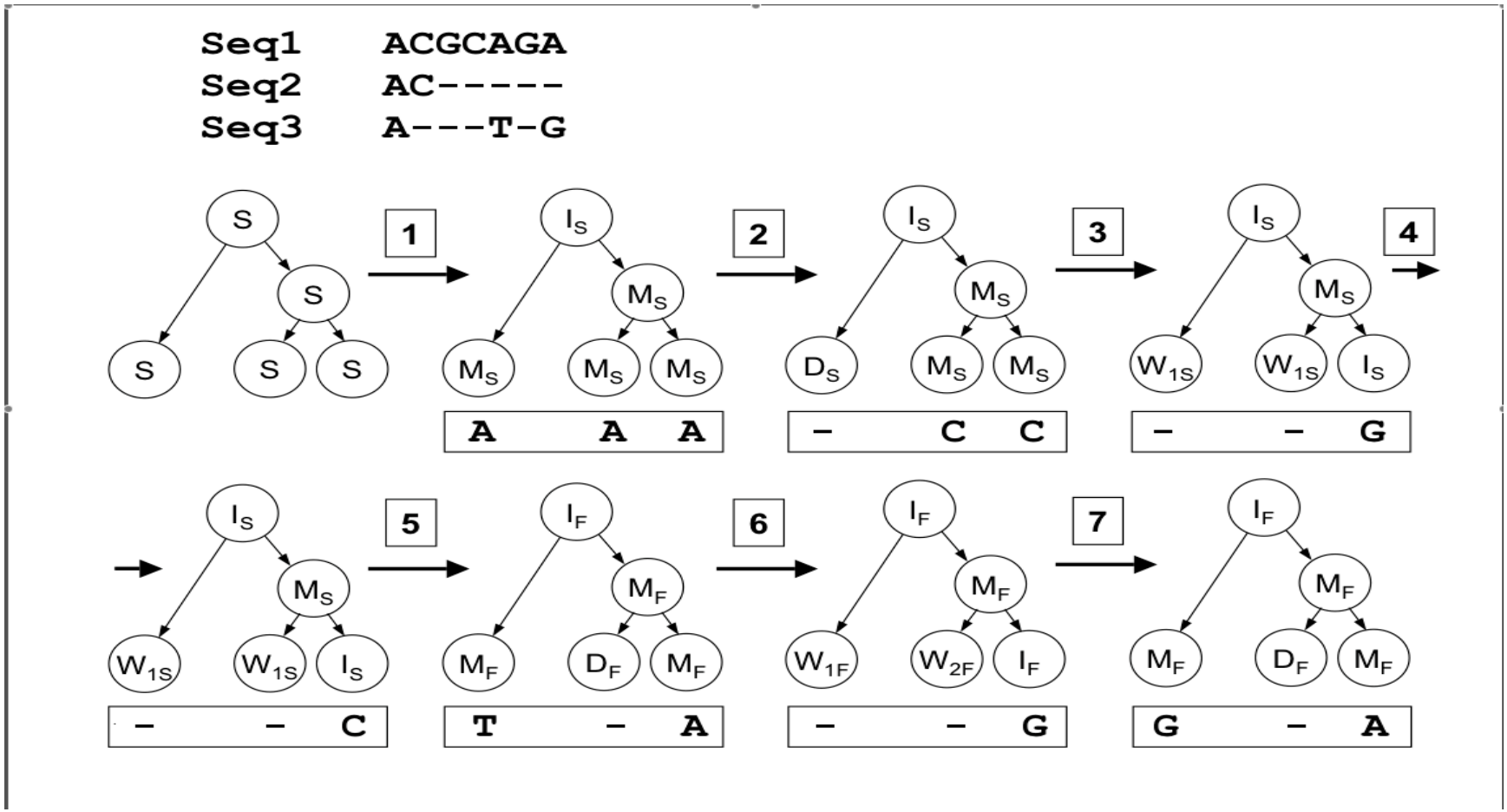


- Allows insertions and deletions of geometrically distributed length
- Second wait state enables delete to self-transition
- Self-transitions result in an expected geometric distribution on the lengths of indel events

SAPF

- Double the number of states
  - Corresponds to creating an HMM on the root, alternating between fast/slow
  - Fixes Fast/Slow annotation on a column
  - Even though fast and slow states have same topology, they have different transition and emission probabilities
- PhyloComposer used to generate MHMM
  - Each MHMM state represents collection of branch HMM states
  - Emission states are alignment columns

Example of SAPF HMM



SAPF HMM parameters

Parameters	Description
$\lambda_{fast}, \lambda_{slow}$	Birth rates for links in fast/slow states
μ_{fast}, μ_{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

Predicting Functional Element

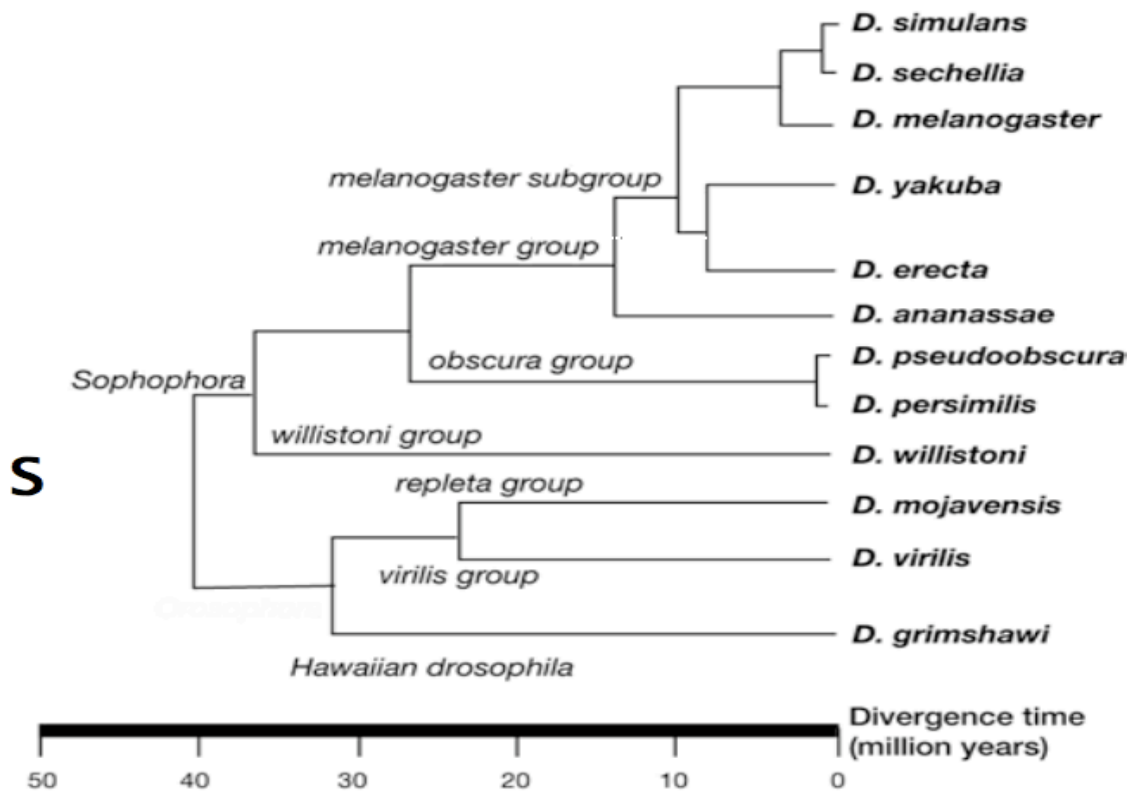
- Use forward and backward algorithm to calculate probability distribution of alignments (represents homology)
- This also computes the probability that any alignment column was generated by either a fast or a slow state
- Make predictions by calculating and summing over distribution of many possible alignments (due to lack of data)

Experimental Results

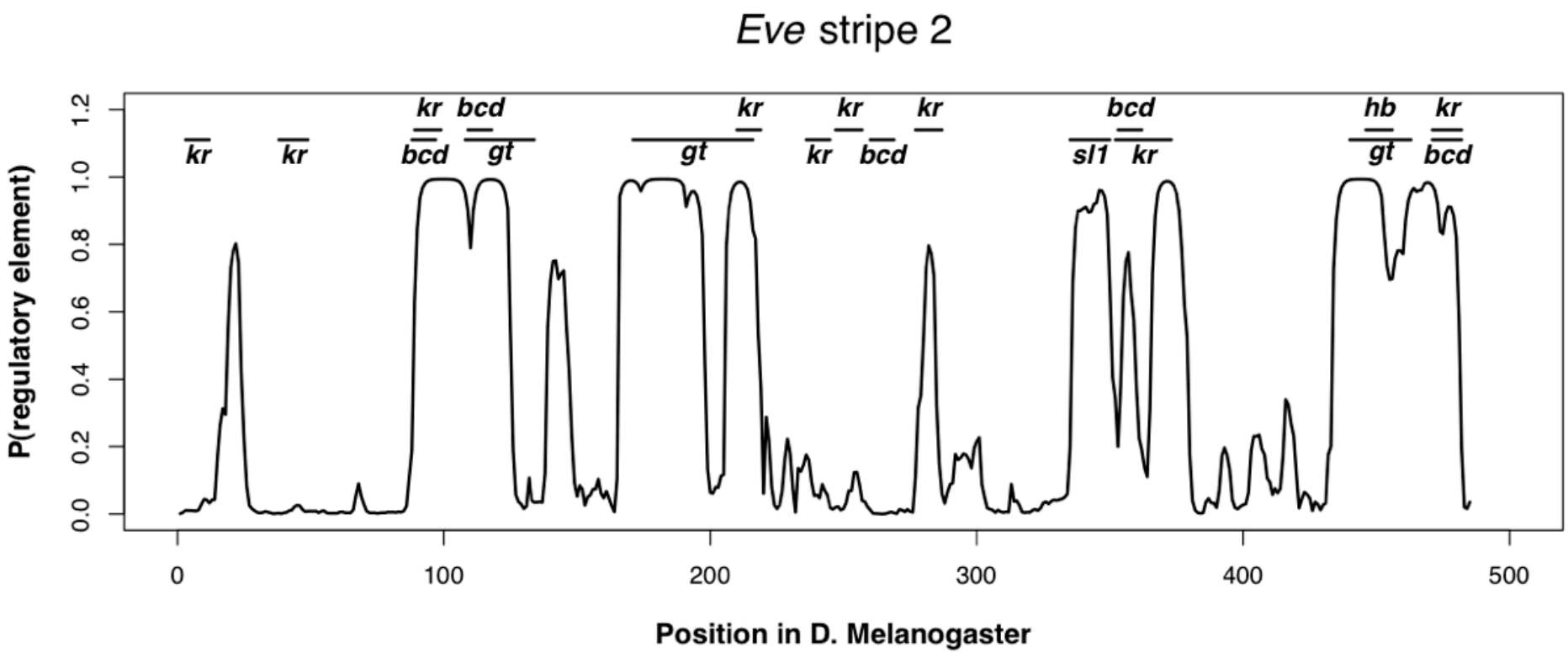
Experimental Setup

- Run SAPF to predict functional elements in Drosophila whole genome sequences
- Drosophila sequences exhibit large evolutionary distances and is ideal for phylogenetic footprinting tests
  - Significant annotations available for TFBS and CRM
- The homeodomain encoding eve protein is crucial in early development in Drosophila and is available in seven transverse stripes whose TFBS is exactly annotated

Evolutionary distance of Drosophila



SAPF Result - 1



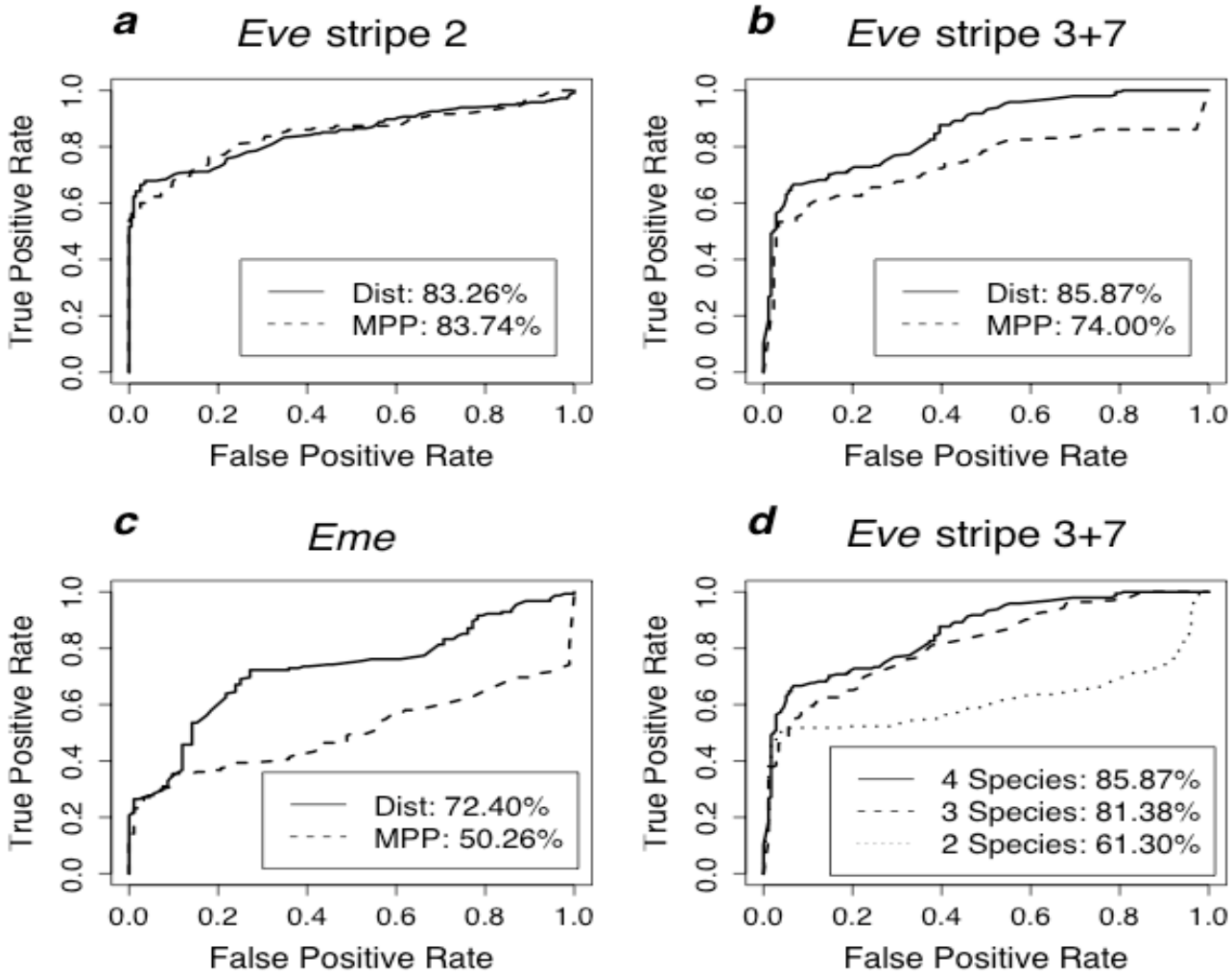
SAPF Result - 1

- SAPF correctly annotated the bases as functional with high probabilities
  - 12 binding sites contain bases assigned a posterior functional probability of greater than 95%; two others contained bases with 80% probability
  - 5 binding sites were incorrectly annotated as neutral indicating that functional orthologs do not exist in all species – Earlier they were characterized as “low-affinity” kr binding sites
  - Predicts two binding sites that were not previously annotated as functional regions!

SAPF Result - 2

- ROC Curve: to predict accuracy of the results that accounts to specificity and sensitivity
  - Run SAPF to predict functional elements, estimate parameters, and construct a single summarized alignment(Maximum Posterior Probability - MPP)
  - Compare MPP with summing over a distribution of alignment using SAPF (Dist)

SAPF Result - 2



Summary

- Transducer framework allows for multiple sequence analysis
- State doubling enables Phylogenetic Footprinting
- The benefits of SAPF increases as the uncertainty in the alignment of functional regions increases

Limitations

- Inability to analyze more than 4 species due to the large number of states
- The algorithm is slow and the authors are considering other methods (like MCMC simulation techniques) to approximate alignment probability distribution
- The annotation of fast or slow is fixed in all species in an alignment column and the model is unable to properly model gain or loss of functional sequence in a single sequence or in a partial group of sequences

Acknowledgement

I would like to acknowledge the authors of the paper **Jotun Hein** and **Rahul Satija** for providing various materials including lecture notes, slides and the supplementary material of the paper.

Questions?

BACKUP

Details of Combining States

- State space for fast HMM, $\phi_f = (S_f, E_f, \Psi^1_f, \ldots, \Psi^n_f)$ and for slow HMM, $\phi_s = (S_s, E_s, \Psi^1_s, \ldots, \Psi^n_s)$
- Combining:
  - Merge both start states; remove end states; Combined state space is $\phi_f = (S_f, \Psi^1_c, \ldots, \Psi^n_c)$
  - Linking of fast states to slow states is done by two transitions:
    - Fast to slow :Transition from Ψ^x_f to End fast state followed by transition from Start state to Ψ^y_s
    - Slow to fast: Transition from Ψ^x_s to End slow state followed by transition from Start state to Ψ^y_f

Predicting Functional Elements

Since laboratory experiments are usually only available for one **reference species** in a closely related group (for example, the *D. melanogaster genome* is the reference for all *Drosophila species), we have chosen to collapse our* results onto one axis and report posterior probabilities for one species, as in (Wasserman *et al., 2000). This is accomplished by* ***grouping together all*** **alignment columns containing the same nucleotide in the reference species, and summing over the group to calculate the overall probability** that the reference nucleotide was generated from a slow state.