


Robust handling of alignment uncertainty when inferring positive selection from divergent sequences.

Benjamin Redelings

February 16, 2013

Site properties

Biological properties of sites


Human **CAG**


Site properties

Biological properties of sites

- ▶ conserved

Human **CAG**

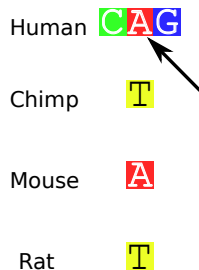


Site properties

Biological properties of sites

- ▶ conserved

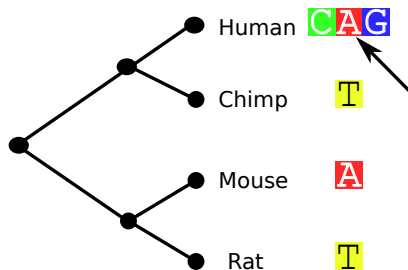
Human	CAG
Chimp	T
Mouse	A
Rat	T



Site properties

Biological properties of sites

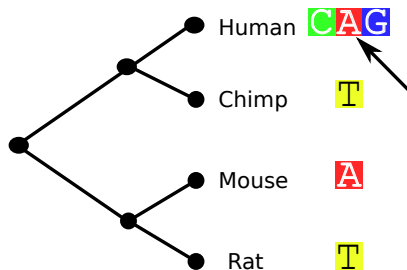
- ▶ conserved



Site properties

Biological properties of sites

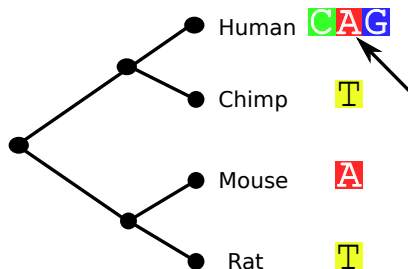
- ▶ conserved
- ▶ hyper-variable



Site properties

Biological properties of sites

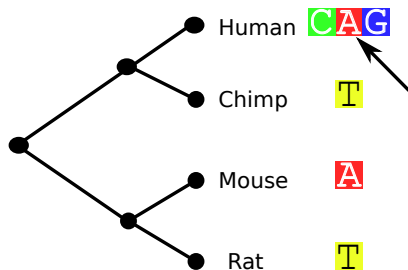
- ▶ conserved
- ▶ hyper-variable
- ▶ positive selection



Site properties

Biological properties of sites

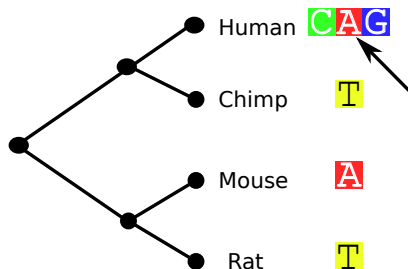
- ▶ conserved
- ▶ hyper-variable
- ▶ positive selection
- ▶ part of a motif



Site properties

Biological properties of sites

- ▶ conserved
- ▶ hyper-variable
- ▶ positive selection
- ▶ part of a motif
- ▶ ...



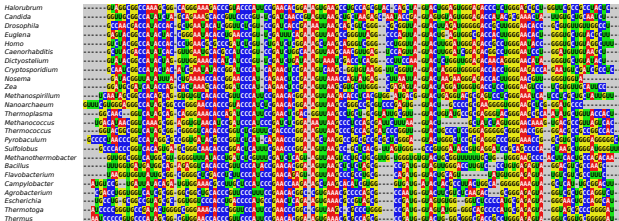
Where do “sites” come from?

Where do “sites” come from?

1. Storks

Where do “sites” come from?

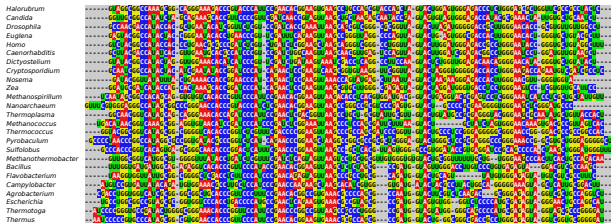
1. Storks
2. Alignment estimates



Clustal W alignment

Where do “sites” come from?

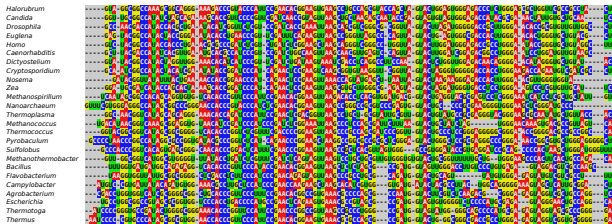
1. Storks
2. Alignment estimates



Muscle alignment

Where do “sites” come from?

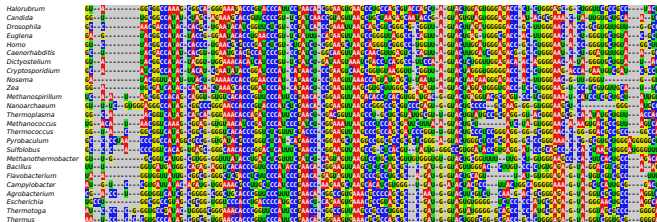
1. Storks
2. Alignment estimates



Muscle alignment

Where do “sites” come from?

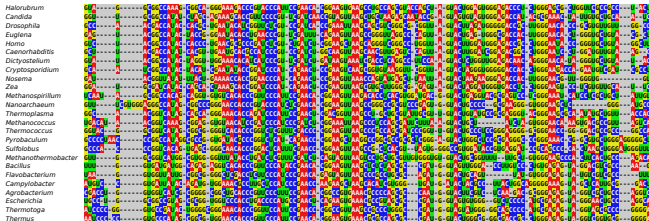
1. Storks
2. Alignment estimates



Probcons alignment

Where do “sites” come from?

1. Storks
2. Alignment estimates



Probcons alignment

Where do “sites” come from?

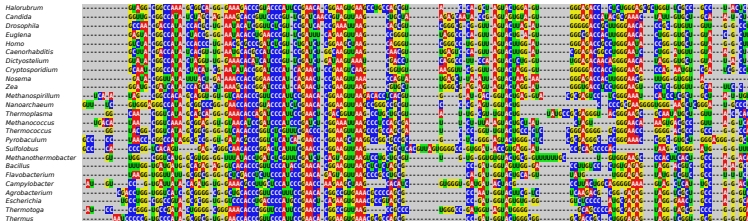
1. Storks
2. Alignment estimates



PRANK alignment

Where do “sites” come from?

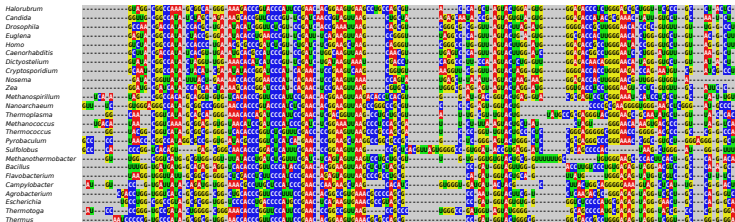
1. Storks
2. Alignment estimates



PRANK alignment

Where do “sites” come from?

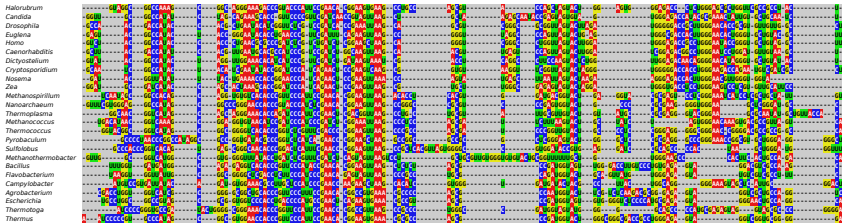
1. Storks
2. Alignment estimates



PRANK alignment

Where do “sites” come from?

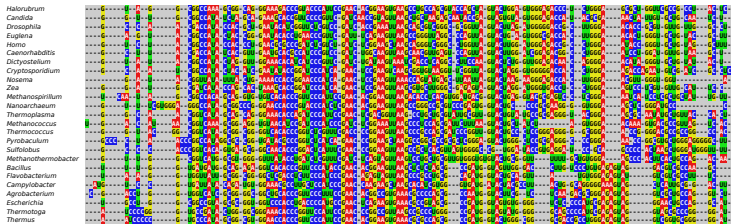
1. Storks
2. Alignment estimates



PRANK alignment

Where do “sites” come from?

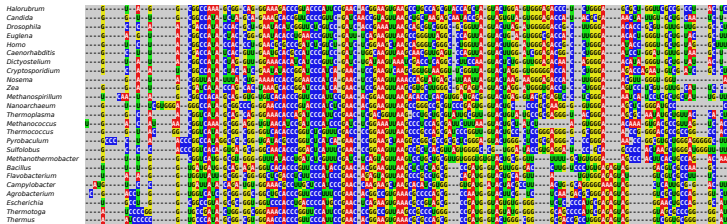
1. Storks
2. Alignment estimates



BALi-Phy alignment

Where do “sites” come from?

1. Storks
2. Alignment estimates

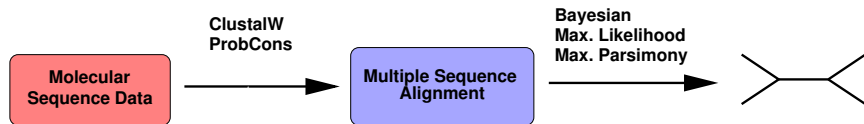


BALI-Phy alignment

Alignment ambiguity is common for divergent sequences.

Sequential Estimation Pipeline

Alignment is often the first stage in a **pipeline**:

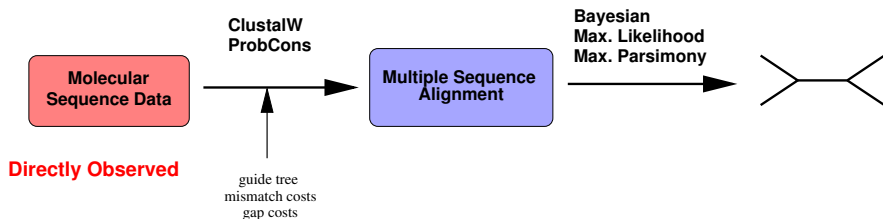


Directly Observed

There are two main sources of alignment ambiguity

Sequential Estimation Pipeline

Alignment is often the first stage in a **pipeline**:

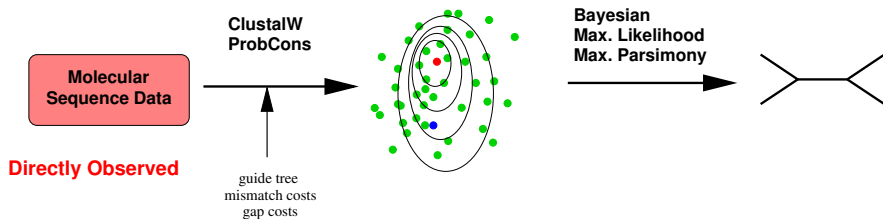


There are two main sources of alignment ambiguity

- ▶ Parameter uncertainty + parameter sensitivity

Sequential Estimation Pipeline

Alignment is often the first stage in a **pipeline**:

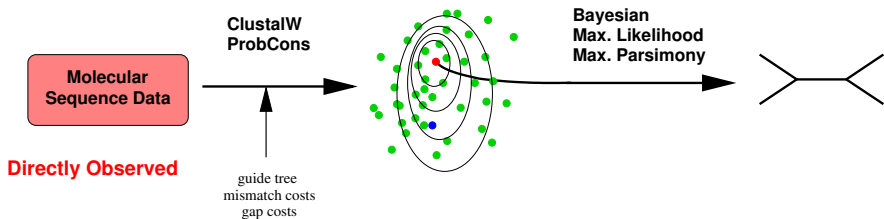


There are two main sources of alignment ambiguity

- ▶ Parameter uncertainty + parameter sensitivity
- ▶ Near-optimal alignments

Sequential Estimation Pipeline

Alignment is often the first stage in a **pipeline**:

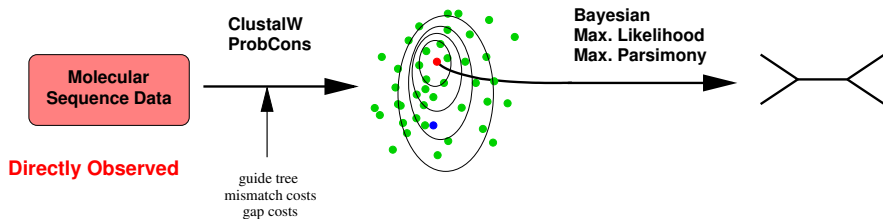


There are two main sources of alignment ambiguity

- ▶ Parameter uncertainty + parameter sensitivity
- ▶ Near-optimal alignments

Sequential Estimation Pipeline

Alignment is often the first stage in a **pipeline**:



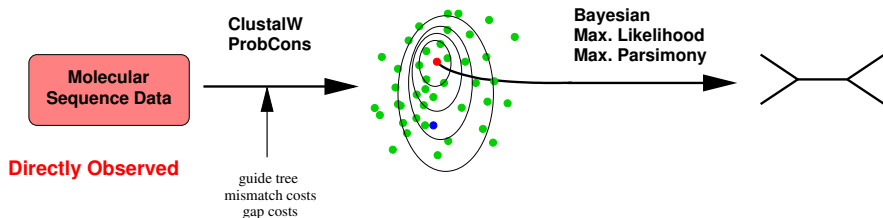
There are two main sources of alignment ambiguity

- ▶ Parameter uncertainty + parameter sensitivity
- ▶ Near-optimal alignments

There are two additional sources of alignment error

Sequential Estimation Pipeline

Alignment is often the first stage in a **pipeline**:



There are two main sources of alignment ambiguity

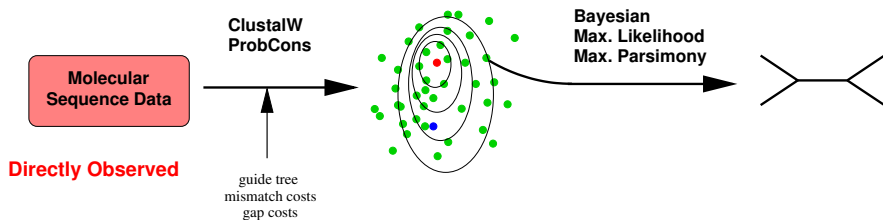
- ▶ Parameter uncertainty + parameter sensitivity
- ▶ Near-optimal alignments

There are two additional sources of alignment error

- ▶ The score function isn't perfect.

Sequential Estimation Pipeline

Alignment is often the first stage in a **pipeline**:



There are two main sources of alignment ambiguity

- ▶ Parameter uncertainty + parameter sensitivity
- ▶ Near-optimal alignments

There are two additional sources of alignment error

- ▶ The score function isn't perfect.
- ▶ Failure to optimize score function.

Positive selection?

(Goldman & Yang, 1994)

Codon sites

CAG

Positive selection?

(Goldman & Yang, 1994)

Codon sites

CAG

Codon substitution model

Positive selection?

(Goldman & Yang, 1994)

Codon sites

CAG

Codon substitution model

- ▶ Must change 1 nucleotide at a time.

Positive selection?

(Goldman & Yang, 1994)

Codon sites

CAG

Codon substitution model

- ▶ Must change 1 nucleotide at a time.

$$Q_{i \rightarrow j} = \pi_j \times \left\{ \begin{array}{ll} 1 & \text{if } \textit{transversion} \\ \kappa & \text{if } \textit{transition} \end{array} \right\} \times \left\{ \begin{array}{ll} 1 & \text{if } \textit{synonymous} \\ \omega & \text{if } \textit{non-synonymous} \end{array} \right\}$$

ω = preference for changes to amino acids.

Codon sites

CAG

Codon substitution model

- ▶ Must change 1 nucleotide at a time.

$$Q_{i \rightarrow j} = \pi_j \times \left\{ \begin{array}{ll} 1 & \text{if } \textit{transversion} \\ \kappa & \text{if } \textit{transition} \end{array} \right\} \times \left\{ \begin{array}{ll} 1 & \text{if } \textit{synonymous} \\ \omega & \text{if } \textit{non-synonymous} \end{array} \right\}$$

ω = preference for changes to amino acids.

Categories

- ▶ $\omega < 1$: amino acid changes happen slowly

Codon sites

CAG

Codon substitution model

- ▶ Must change 1 nucleotide at a time.

$$Q_{i \rightarrow j} = \pi_j \times \left\{ \begin{array}{ll} 1 & \text{if } \textit{transversion} \\ \kappa & \text{if } \textit{transition} \end{array} \right\} \times \left\{ \begin{array}{ll} 1 & \text{if } \textit{synonymous} \\ \omega & \text{if } \textit{non-synonymous} \end{array} \right\}$$

ω = preference for changes to amino acids.

Categories

- ▶ $\omega < 1$: amino acid changes happen slowly
- ▶ $\omega = 1$: neutrality

Codon sites

CAG

Codon substitution model

- ▶ Must change 1 nucleotide at a time.

$$Q_{i \rightarrow j} = \pi_j \times \left\{ \begin{array}{ll} 1 & \text{if transversion} \\ \kappa & \text{if transition} \end{array} \right\} \times \left\{ \begin{array}{ll} 1 & \text{if synonymous} \\ \omega & \text{if non-synonymous} \end{array} \right\}$$

ω = preference for changes to amino acids.

Categories

- ▶ $\omega < 1$: amino acid changes happen slowly
- ▶ $\omega = 1$: neutrality
- ▶ $\omega > 1$: amino acid changes preferred

Codon sites

CAG

Codon substitution model

- ▶ Must change 1 nucleotide at a time.

$$Q_{i \rightarrow j} = \pi_j \times \left\{ \begin{array}{ll} 1 & \text{if transversion} \\ \kappa & \text{if transition} \end{array} \right\} \times \left\{ \begin{array}{ll} 1 & \text{if synonymous} \\ \omega & \text{if non-synonymous} \end{array} \right\}$$

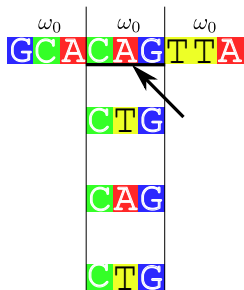
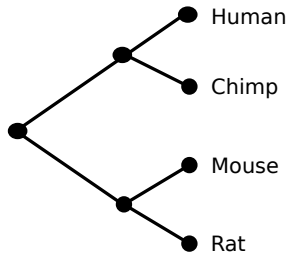
ω = preference for changes to amino acids.

Categories

- ▶ $\omega < 1$: amino acid changes happen slowly
- ▶ $\omega = 1$: neutrality
- ▶ $\omega > 1$: amino acid changes preferred → “positive selection!”

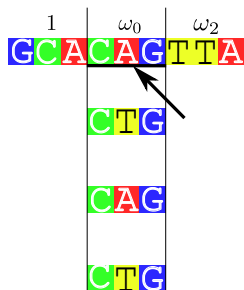
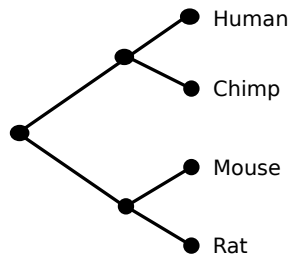
Site models

(Yang et al, 2000)



Site models

(Yang et al, 2000)

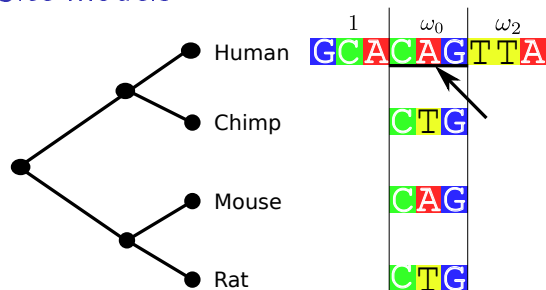


Model:

	Category #1	Category #2	Category #3
ω	$\omega_0 \leq 1$	$\omega_1 = 1$	$\omega_2 \geq 1$
Frequency	p_0	p_1	$1 - p_0 - p_1$

Site models

(Yang et al, 2000)



Model:

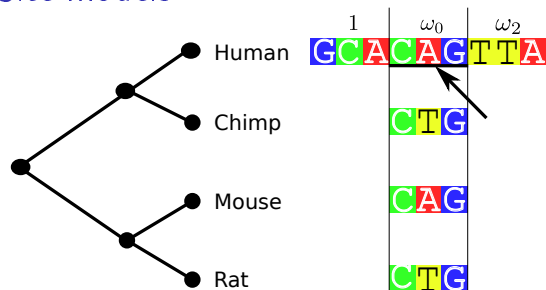
	Category #1	Category #2	Category #3
ω	$\omega_0 \leq 1$	$\omega_1 = 1$	$\omega_2 \geq 1$
Frequency	p_0	p_1	$1 - p_0 - p_1$

Test:

- ▶ Estimate ω_0 , ω_2 , p_0 , p_1 .
- ▶ Compare $H_0 : \omega_2 = 1$ with $H_a : \omega_2 \geq 1$.

Site models

(Yang et al, 2000)



Model:

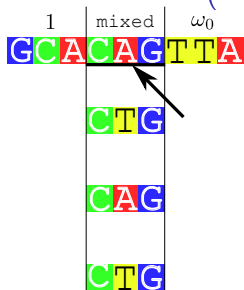
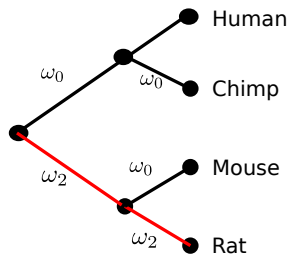
	Category #1	Category #2	Category #3
ω	$\omega_0 \leq 1$	$\omega_1 = 1$	$\omega_2 \geq 1$
Frequency	p_0	p_1	$1 - p_0 - p_1$

Test:

- ▶ Estimate ω_0 , ω_2 , p_0 , p_1 .
- ▶ Compare $H_0 : \omega_2 = 1$ with $H_a : \omega_2 \geq 1$.
- ▶ **Problem:** if any single site is misaligned... *positive selection!*

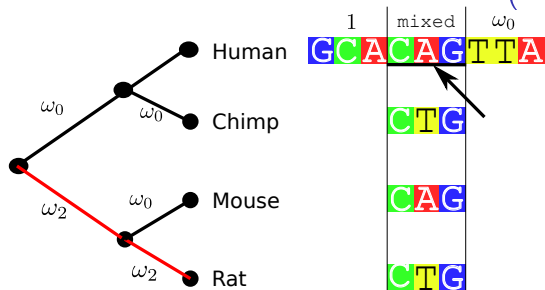
Branch-Site models

(Yang & Nielsen, 2002)



Branch-Site models

(Yang & Nielsen, 2002)

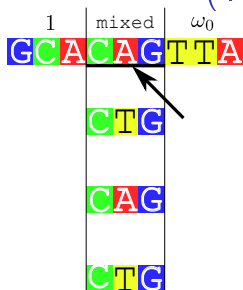
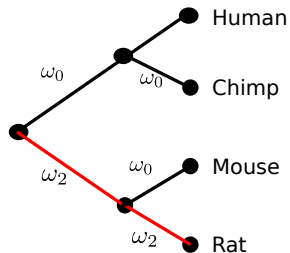


Model:

	Category #1	Category #2	Category #3	Category #4
background ω	ω_0	1	ω_0	1
foreground ω	ω_0	1	ω_2	ω_2
Frequency	p_0	p_1	p_{2a}	p_{2b}

Branch-Site models

(Yang & Nielsen, 2002)



Model:

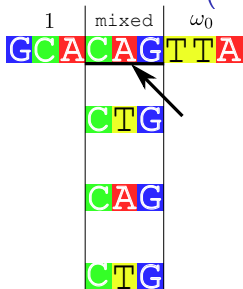
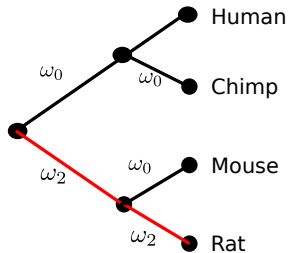
	Category #1	Category #2	Category #3	Category #4
background ω	ω_0	1	ω_0	1
foreground ω	ω_0	1	ω_2	ω_2
Frequency	p_0	p_1	p_{2a}	p_{2b}

Test:

- ▶ Estimate $\omega_0, \omega_2, p_0, p_1, p_2$
- ▶ Compare $H_0 : \omega_2 = 1$ with $H_a : \omega_2 \geq 1$. (Zhang, Nielsen, and Yang, 2005)

Branch-Site models

(Yang & Nielsen, 2002)



Model:

	Category #1	Category #2	Category #3	Category #4
background ω	ω_0	1	ω_0	1
foreground ω	ω_0	1	ω_2	ω_2
Frequency	p_0	p_1	p_{2a}	p_{2b}

Test:

- ▶ Estimate $\omega_0, \omega_2, p_0, p_1, p_2$
- ▶ Compare $H_0 : \omega_2 = 1$ with $H_a : \omega_2 \geq 1$. (Zhang, Nielsen, and Yang, 2005)
- ▶ **Problem:** if any single site is misaligned... *positive selection!*

Conceptually “Nice” way

$$\Pr(\text{unaligned data} | M) \neq \Pr(\text{unaligned data} | M, \hat{\mathbf{A}})$$

Conceptually “Nice” way

$$\Pr(\text{unaligned data}|M) = \sum_{\mathbf{A}} \Pr(\text{unaligned data}, \mathbf{A}|M)$$

Model

τ -tree

\mathbf{A} -alignment

Θ -evolutionary parameters

$$\Pr(data, \mathbf{A}, \tau, \Theta) = \Pr(data|\mathbf{A}, \tau, \Theta) \times \Pr(\mathbf{A}|\tau, \Theta) \times \Pr(\tau) \times \Pr(\Theta).$$

Model

τ -tree

\mathbf{A} -alignment

Θ -evolutionary parameters

$$\Pr(\text{data}, \mathbf{A}, \tau, \Theta) = \Pr(\text{data} | \mathbf{A}, \tau, \Theta) \times \Pr(\mathbf{A} | \tau, \Theta) \times \Pr(\tau) \times \Pr(\Theta).$$

Model

τ -tree

\mathbf{A} -alignment

Θ -evolutionary parameters

$$\Pr(\text{data}, \mathbf{A}, \tau, \Theta) = \Pr(\text{data} | \mathbf{A}, \tau, \Theta) \times \Pr(\mathbf{A} | \tau, \Theta) \times \Pr(\tau) \times \Pr(\Theta).$$

Model

τ -tree

\mathbf{A} -alignment

Θ -evolutionary parameters

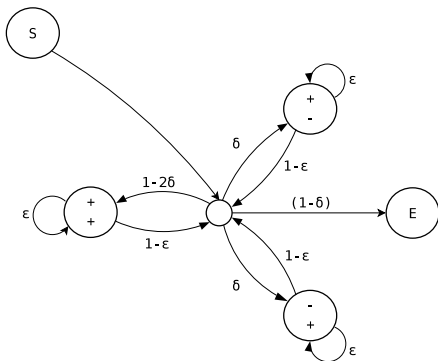
$$\Pr(\text{data}, \mathbf{A}, \tau, \Theta) = \Pr(\text{data} | \mathbf{A}, \tau, \Theta) \times \Pr(\mathbf{A} | \tau, \Theta) \times \Pr(\tau) \times \Pr(\Theta).$$

$\Pr(\mathbf{A}|\tau, \Theta)$: Insertion/Deletion probabilities

RS07 Pairwise alignment distribution on each branch of the tree:

► **Pair HMM** model with 2 parameters:

- indel *rate* is λ
- indel *lengths* are Geometric(ϵ)

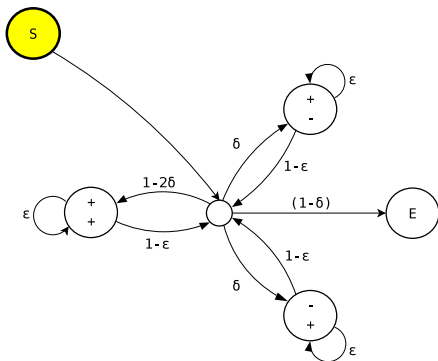


$\Pr(\mathbf{A}|\tau, \Theta)$: Insertion/Deletion probabilities

RS07 Pairwise alignment distribution on each branch of the tree:

► **Pair HMM** model with 2 parameters:

- indel *rate* is λ
- indel *lengths* are Geometric(ϵ)

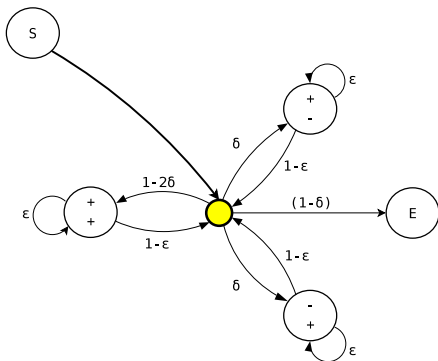


$\Pr(\mathbf{A}|\tau, \Theta)$: Insertion/Deletion probabilities

RS07 Pairwise alignment distribution on each branch of the tree:

► **Pair HMM** model with 2 parameters:

- indel *rate* is λ
- indel *lengths* are Geometric(ϵ)

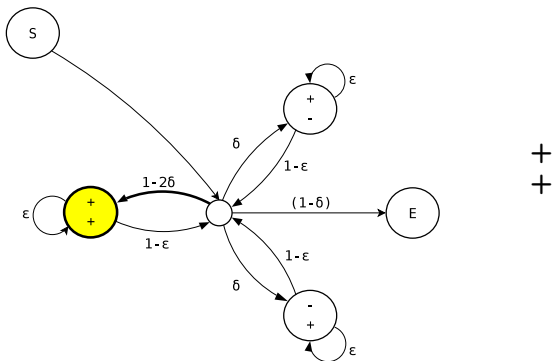


$\Pr(\mathbf{A}|\tau, \Theta)$: Insertion/Deletion probabilities

RS07 Pairwise alignment distribution on each branch of the tree:

► **Pair HMM** model with 2 parameters:

- indel *rate* is λ
- indel *lengths* are Geometric(ϵ)

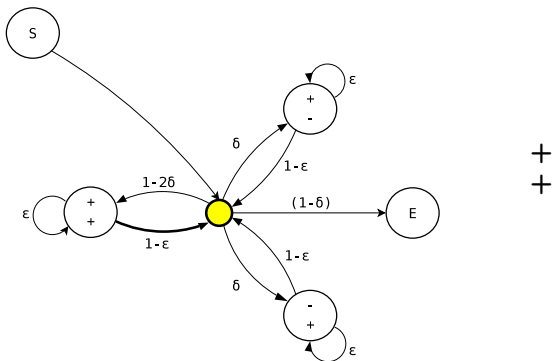


$\Pr(\mathbf{A}|\tau, \Theta)$: Insertion/Deletion probabilities

RS07 Pairwise alignment distribution on each branch of the tree:

► **Pair HMM** model with 2 parameters:

- indel *rate* is λ
- indel *lengths* are Geometric(ϵ)

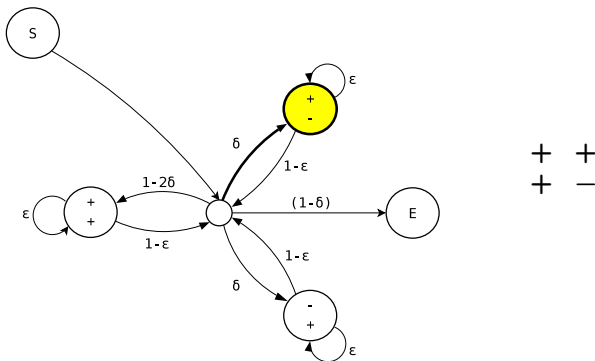


$\Pr(\mathbf{A}|\tau, \Theta)$: Insertion/Deletion probabilities

RS07 Pairwise alignment distribution on each branch of the tree:

► **Pair HMM** model with 2 parameters:

- indel *rate* is λ
- indel *lengths* are Geometric(ϵ)

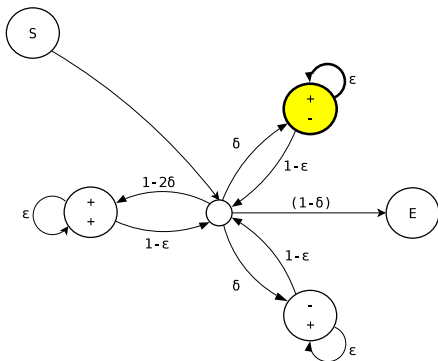


$\Pr(\mathbf{A}|\tau, \Theta)$: Insertion/Deletion probabilities

RS07 Pairwise alignment distribution on each branch of the tree:

► **Pair HMM** model with 2 parameters:

- indel *rate* is λ
- indel *lengths* are Geometric(ϵ)



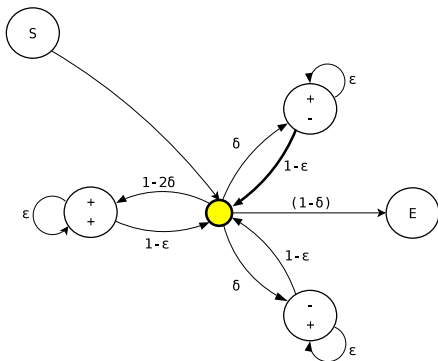
+ + +
+ - -

$\Pr(\mathbf{A}|\tau, \Theta)$: Insertion/Deletion probabilities

RS07 Pairwise alignment distribution on each branch of the tree:

► **Pair HMM** model with 2 parameters:

- indel *rate* is λ
- indel *lengths* are Geometric(ϵ)



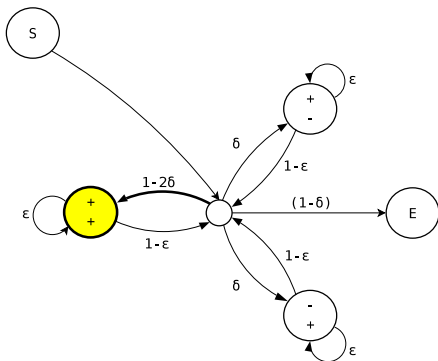
+ + +
+ - -

$\Pr(\mathbf{A}|\tau, \Theta)$: Insertion/Deletion probabilities

RS07 Pairwise alignment distribution on each branch of the tree:

► **Pair HMM** model with 2 parameters:

- indel *rate* is λ
- indel *lengths* are Geometric(ϵ)



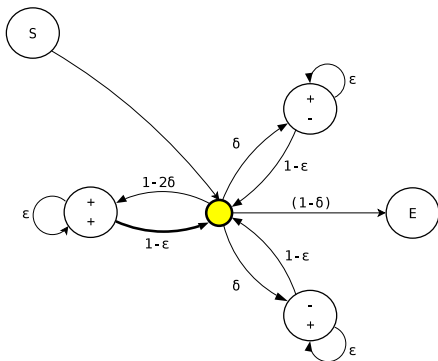
+ + + +
+ - - +

$\Pr(\mathbf{A}|\tau, \Theta)$: Insertion/Deletion probabilities

RS07 Pairwise alignment distribution on each branch of the tree:

► **Pair HMM** model with 2 parameters:

- indel *rate* is λ
- indel *lengths* are Geometric(ϵ)



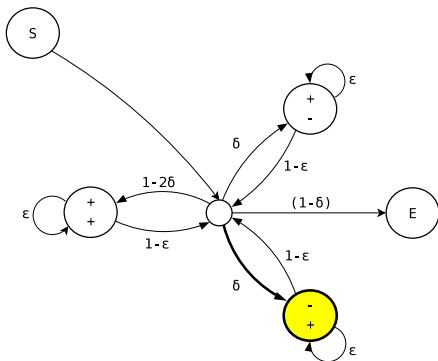
+ + + +
+ - - +

$\Pr(\mathbf{A}|\tau, \Theta)$: Insertion/Deletion probabilities

RS07 Pairwise alignment distribution on each branch of the tree:

► **Pair HMM** model with 2 parameters:

- indel *rate* is λ
- indel *lengths* are Geometric(ϵ)



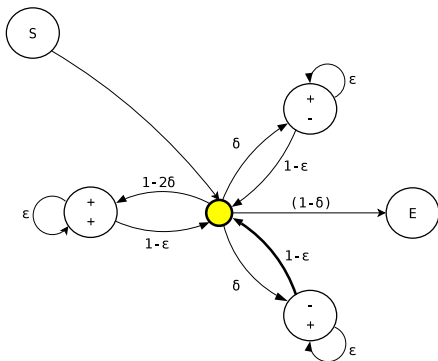
+ + + + -
+ - - + +

$\Pr(\mathbf{A}|\tau, \Theta)$: Insertion/Deletion probabilities

RS07 Pairwise alignment distribution on each branch of the tree:

► **Pair HMM** model with 2 parameters:

- indel *rate* is λ
- indel *lengths* are Geometric(ϵ)



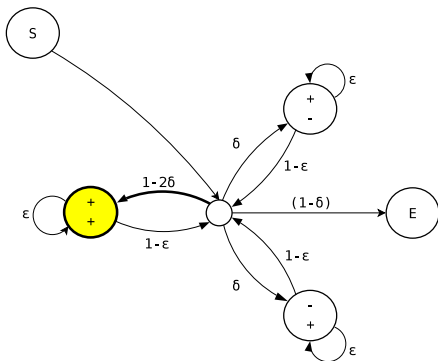
+ + + + -
+ - - + +

$\Pr(\mathbf{A}|\tau, \Theta)$: Insertion/Deletion probabilities

RS07 Pairwise alignment distribution on each branch of the tree:

► **Pair HMM** model with 2 parameters:

- indel *rate* is λ
- indel *lengths* are Geometric(ϵ)



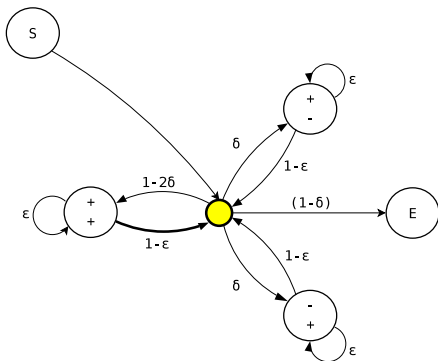
+ + + + - +
+ - - + + +

$\Pr(\mathbf{A}|\tau, \Theta)$: Insertion/Deletion probabilities

RS07 Pairwise alignment distribution on each branch of the tree:

► **Pair HMM** model with 2 parameters:

- indel *rate* is λ
- indel *lengths* are Geometric(ϵ)



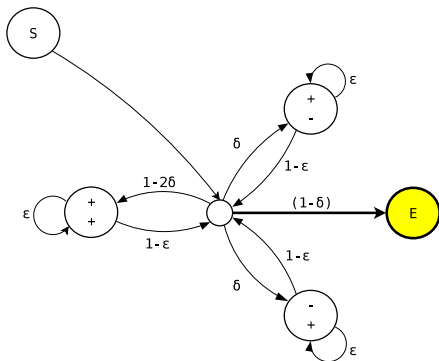
+ + + + - +
+ - - + + +

$\Pr(\mathbf{A}|\tau, \Theta)$: Insertion/Deletion probabilities

RS07 Pairwise alignment distribution on each branch of the tree:

► **Pair HMM** model with 2 parameters:

- indel *rate* is λ
- indel *lengths* are Geometric(ϵ)



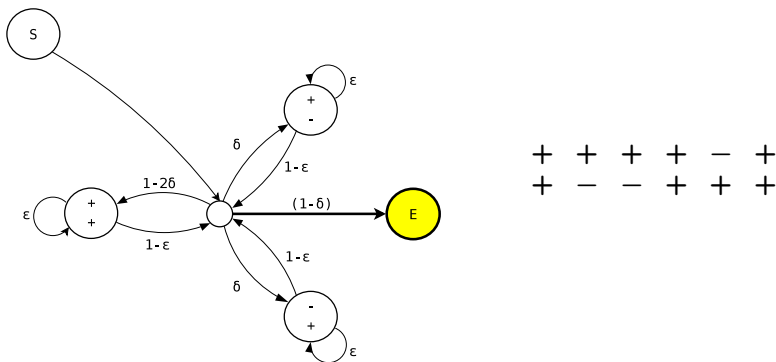
+ + + + - +
+ - - + + +

$\Pr(\mathbf{A}|\tau, \Theta)$: Insertion/Deletion probabilities

RS07 Pairwise alignment distribution on each branch of the tree:

► **Pair HMM** model with 2 parameters:

- indel *rate* is λ
- indel *lengths* are Geometric(ϵ)



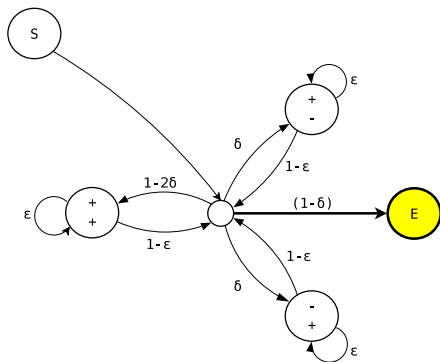
Probability of 1 gap $\approx \delta \times \epsilon^{(L-1)} \times (1 - \epsilon)$

$\Pr(\mathbf{A}|\tau, \Theta)$: Insertion/Deletion probabilities

RS07 Pairwise alignment distribution on each branch of the tree:

► **Pair HMM** model with 2 parameters:

- indel *rate* is λ
- indel *lengths* are Geometric(ϵ)



+ + + + - +
+ - - + + +

Affine gap penalty $\approx [\log \delta] + (L - 1) \times [\log \epsilon]$

Algorithm: Markov chain Monte Carlo (MCMC)

Goal: Sample $(A, \Theta, M | \tau, data)$ from posterior distribution

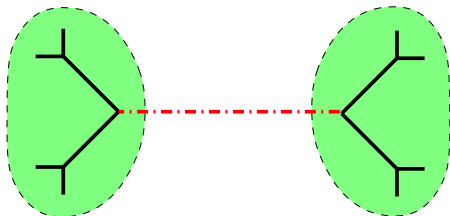
- ▶ $M = 1$: positive selection
- ▶ $M = 0$: no positive selection.

Algorithm: Markov chain Monte Carlo (MCMC)

Goal: Sample $(A, \Theta, M | \tau, data)$ from posterior distribution

- ▶ $M = 1$: positive selection
- ▶ $M = 0$: no positive selection.

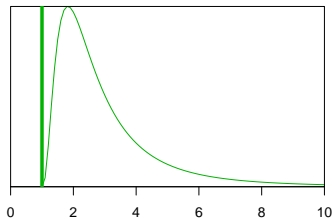
Method: Randomly alter alignment, parameters



Bayesian Tests

Priors

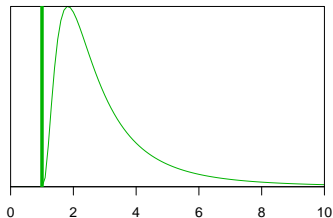
- ▶ $\frac{p_0}{p_0+p_1} \sim \text{Uniform}(0, 1)$
- ▶ $p_2 \sim \text{Beta}(1, 10)$
- ▶ $\log \omega_2 \sim \text{Gamma}(4, 0.25)$



Bayesian Tests

Priors

- ▶ $\frac{p_0}{p_0+p_1} \sim \text{Uniform}(0, 1)$
- ▶ $p_2 \sim \text{Beta}(1, 10)$
- ▶ $\log \omega_2 \sim \text{Gamma}(4, 0.25)$



Bayes Factor

$$BF = \frac{\Pr(\text{data}|\omega_2 > 1)}{\Pr(\text{data}|\omega_2 = 1)}$$

Simulation Parameters

(Fletcher & Yang, 2010)

Simulation Parameters

(Fletcher & Yang, 2010)

- ▶ **Software:** INDELible (Fletcher & Yang, 2010)

Simulation Parameters

(Fletcher & Yang, 2010)

- ▶ **Software:** INDELible (Fletcher & Yang, 2010)
- ▶ **Length** = 300 codons

Simulation Parameters

(Fletcher & Yang, 2010)

- ▶ **Software:** INDELible (Fletcher & Yang, 2010)
- ▶ **Length** = 300 codons
- ▶ **Indel rate** = (substitution rate) \times 0.05

Simulation Parameters

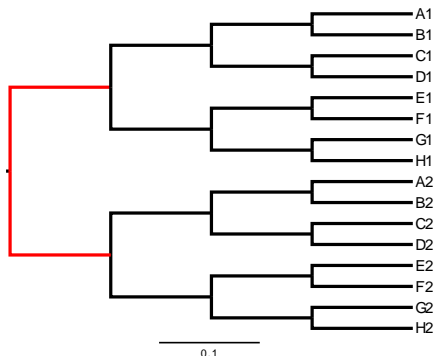
(Fletcher & Yang, 2010)

- ▶ **Software:** INDELible (Fletcher & Yang, 2010)
- ▶ **Length** = 300 codons
- ▶ **Indel rate** = (substitution rate) \times 0.05
- ▶ ω : Scheme #1:
 - ▶ **10** categories: all **neutral** or **conserved**.

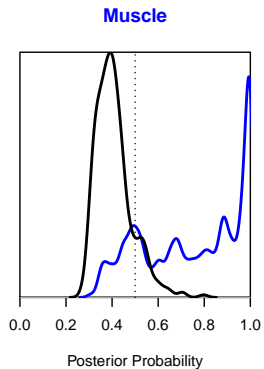
Simulation Parameters

(Fletcher & Yang, 2010)

- ▶ **Software:** INDELible (Fletcher & Yang, 2010)
- ▶ **Length** = 300 codons
- ▶ **Indel rate** = (substitution rate) \times 0.05
- ▶ ω : Scheme #1:
 - ▶ 10 categories: all **neutral** or **conserved**.
- ▶ ω : Scheme #2:
 - ▶ Foreground branch has some positive selection

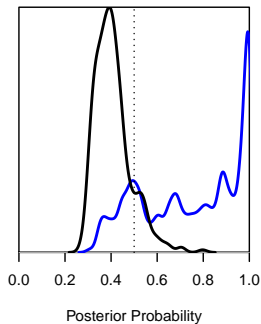


Histogram of Posterior Probabilities

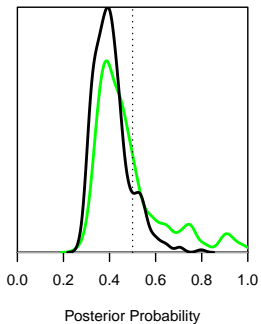


Histogram of Posterior Probabilities

Muscle

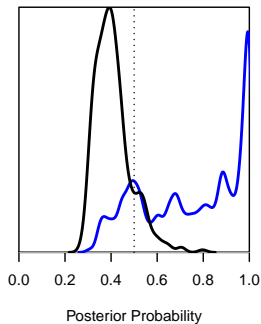


Prank

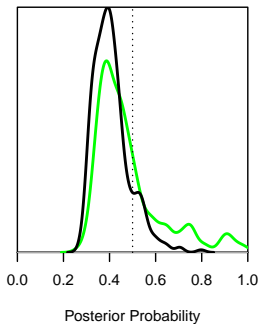


Histogram of Posterior Probabilities

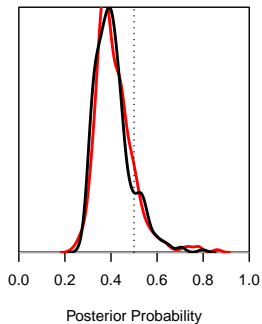
Muscle

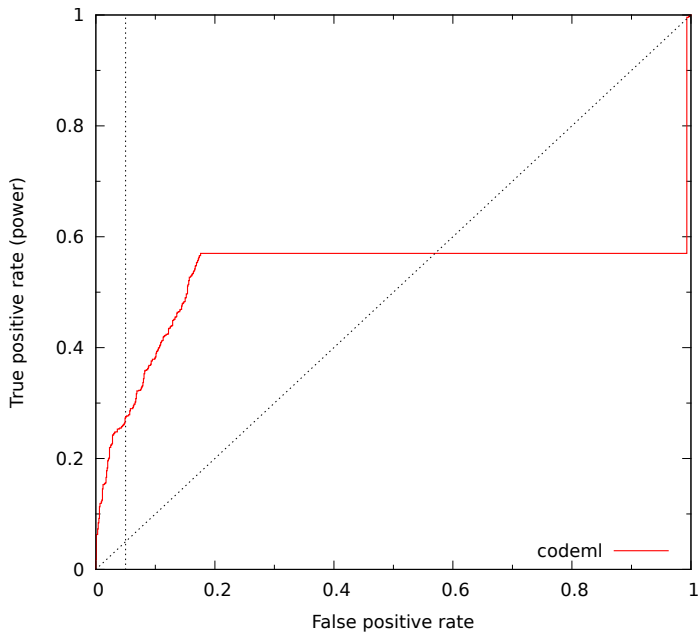


Prank

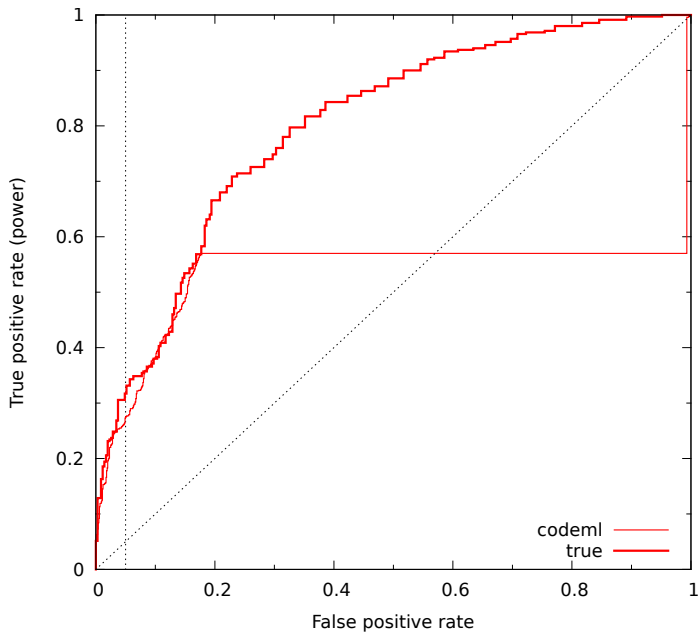


Joint Estimation

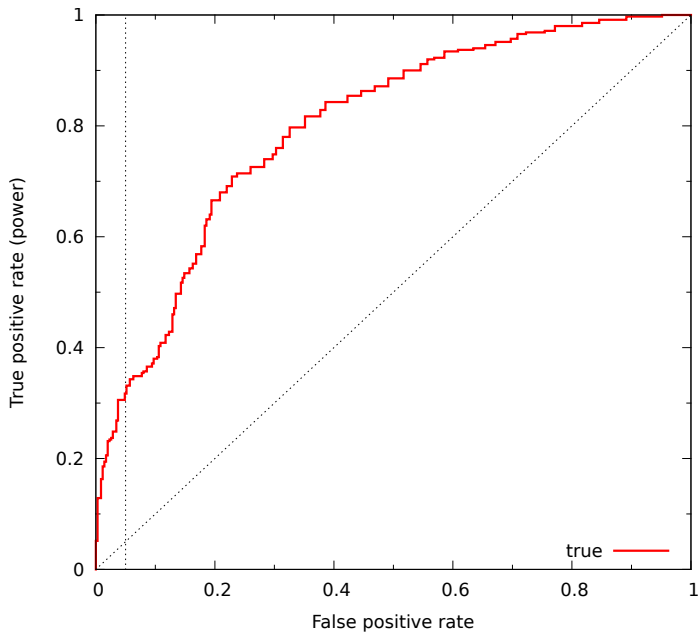




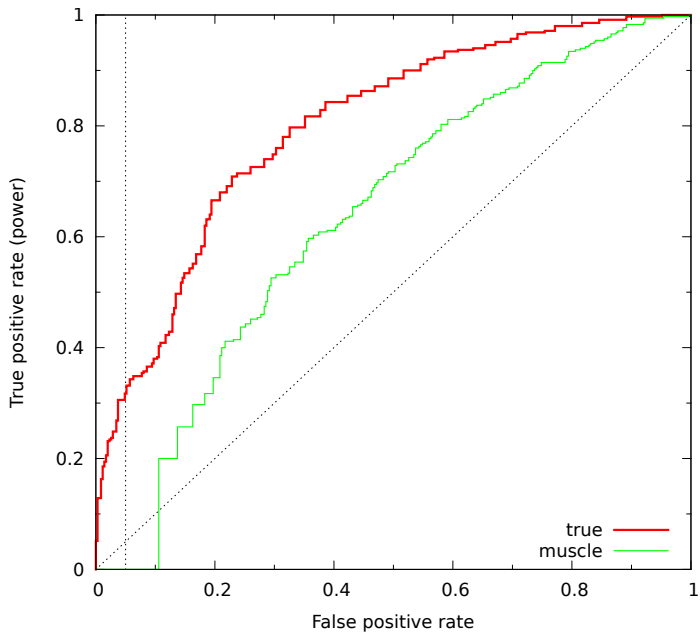
Signal Detection



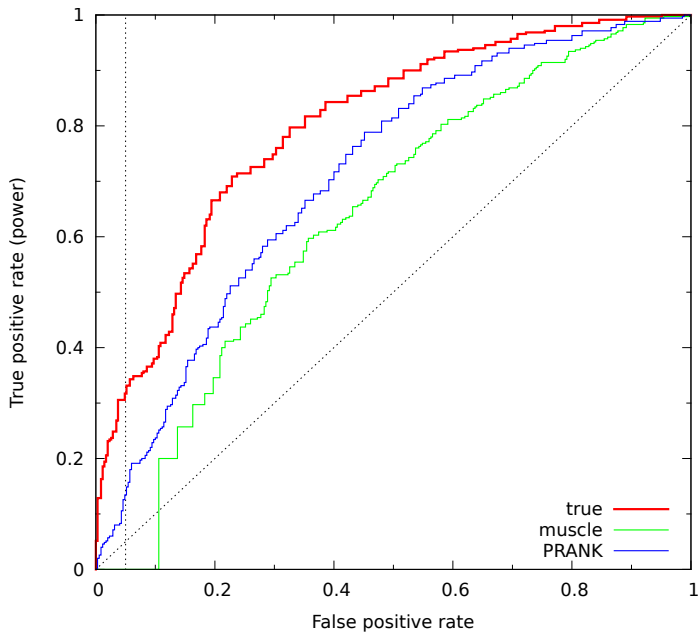
Signal Detection



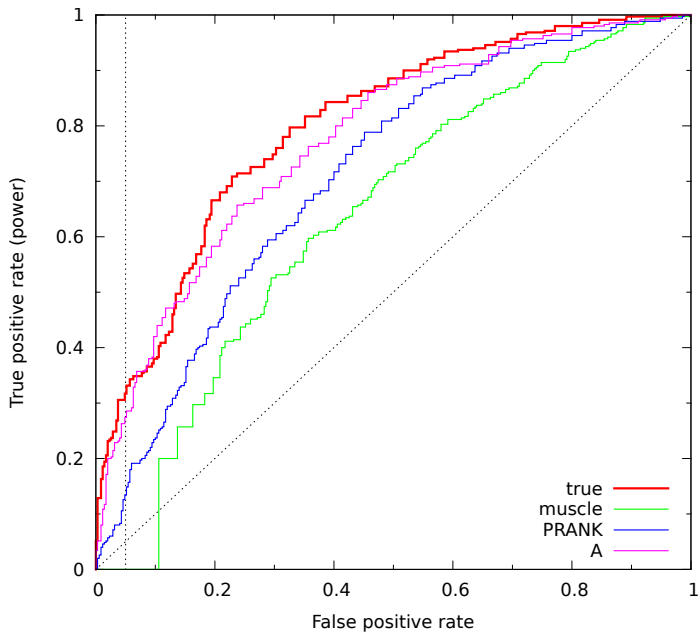
Signal Detection



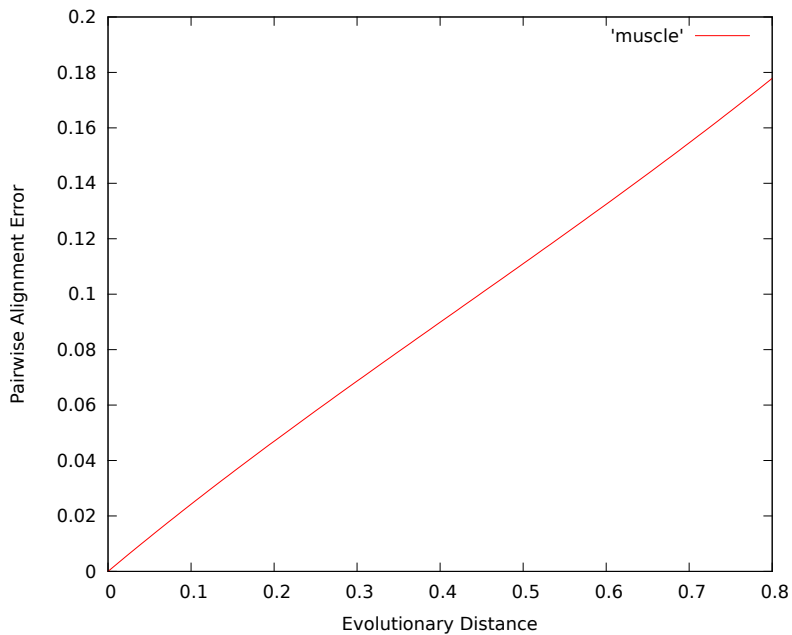
Signal Detection



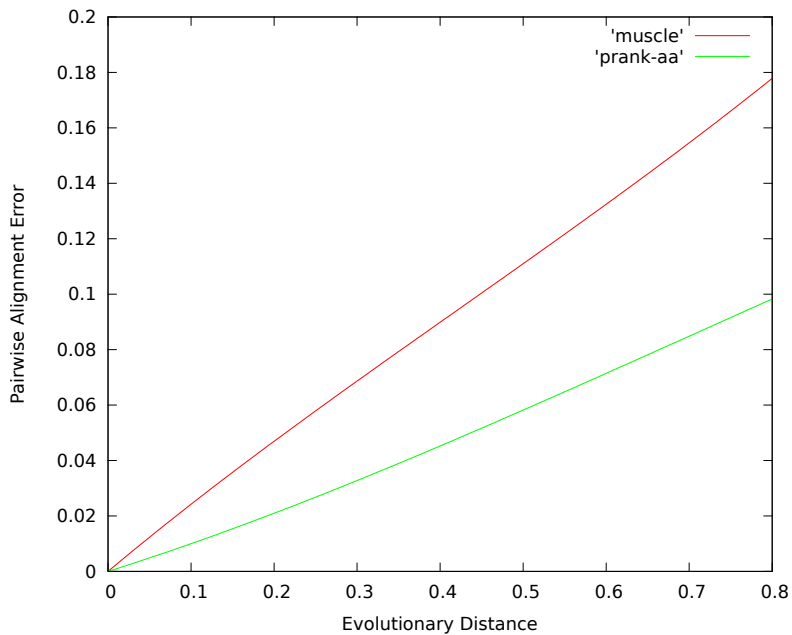
Signal Detection



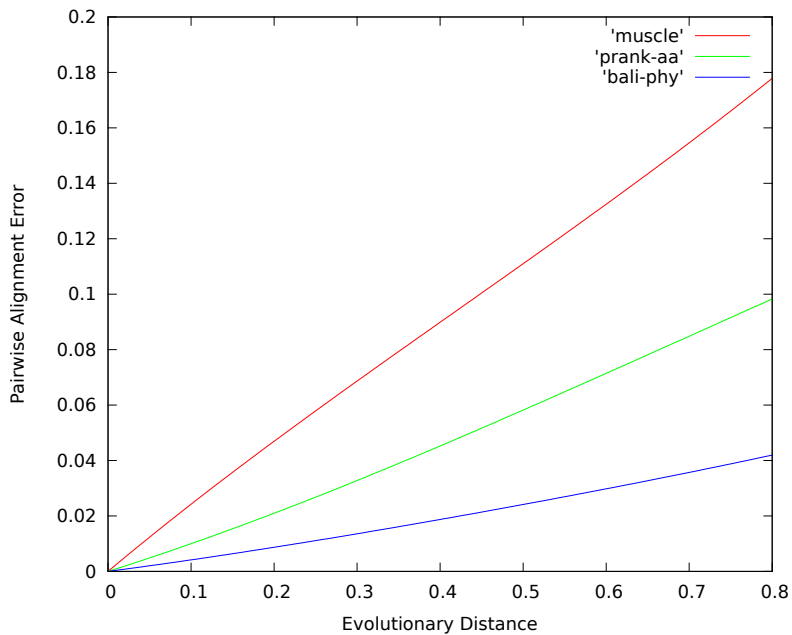
Alignment error



Alignment error



Alignment error



Merits of different approaches?

1. Fixed A (*Muscle*)

Merits of different approaches?

1. Fixed A (*Muscle*)
2. Fixed A, **tree-based** (*Prank*+codeml)

Merits of different approaches?

1. Fixed A (*Muscle*)
2. Fixed A, tree-based (*Prank*+codeml)
3. Average A, tree-based (*Prank*+codeml)

Merits of different approaches?

1. Fixed A (*Muscle*)
2. Fixed A, tree-based (*Prank*+codeml)
3. Average A, tree-based (*Prank*+codeml)
4. Average A, tree-based, MCMC (*bali-phy*+codeml)

Merits of different approaches?

1. Fixed A (*Muscle*)
2. Fixed A, tree-based (*Prank*+codeml)
3. Average A, tree-based (*Prank*+codeml)
4. Average A, tree-based, MCMC (*bali-phy*+codeml)
5. Joint Estimation (*bali-phy*)

Merits of different approaches?

1. Fixed A (*Muscle*)
2. Fixed A, tree-based (*Prank*+codeml)
3. Average A, tree-based (*Prank*+codeml)
4. Average A, tree-based, MCMC (*bali-phy*+codeml)
5. Joint Estimation (*bali-phy*)
 - ▶ Integrate alignment estimation into the inside of the test.

Generic Model Framework

```
bali-phy file.fasta --tree=file.tree --alphabet=Codons  
--smodel=branch-site --disable=topology
```

$(\text{map } \lambda x.M_0(x) \ [\omega_0, 1, \omega_2]) \quad \rightarrow \quad [M_0(\omega_0), M_0(1), M_0(\omega_2)]$

Acknowledgments

National Evolutionary Synthesis Center