
Detecting mosaic structures in DNA sequence alignments

Dirk Husmeier

Biomathematics and Statistics Scotland

Edinburgh, United Kingdom

Email: dirk@bioss.ac.uk

<http://www.bioss.ac.uk/~dirk>

Detecting mosaic structures in DNA sequence alignments

Dirk Husmeier

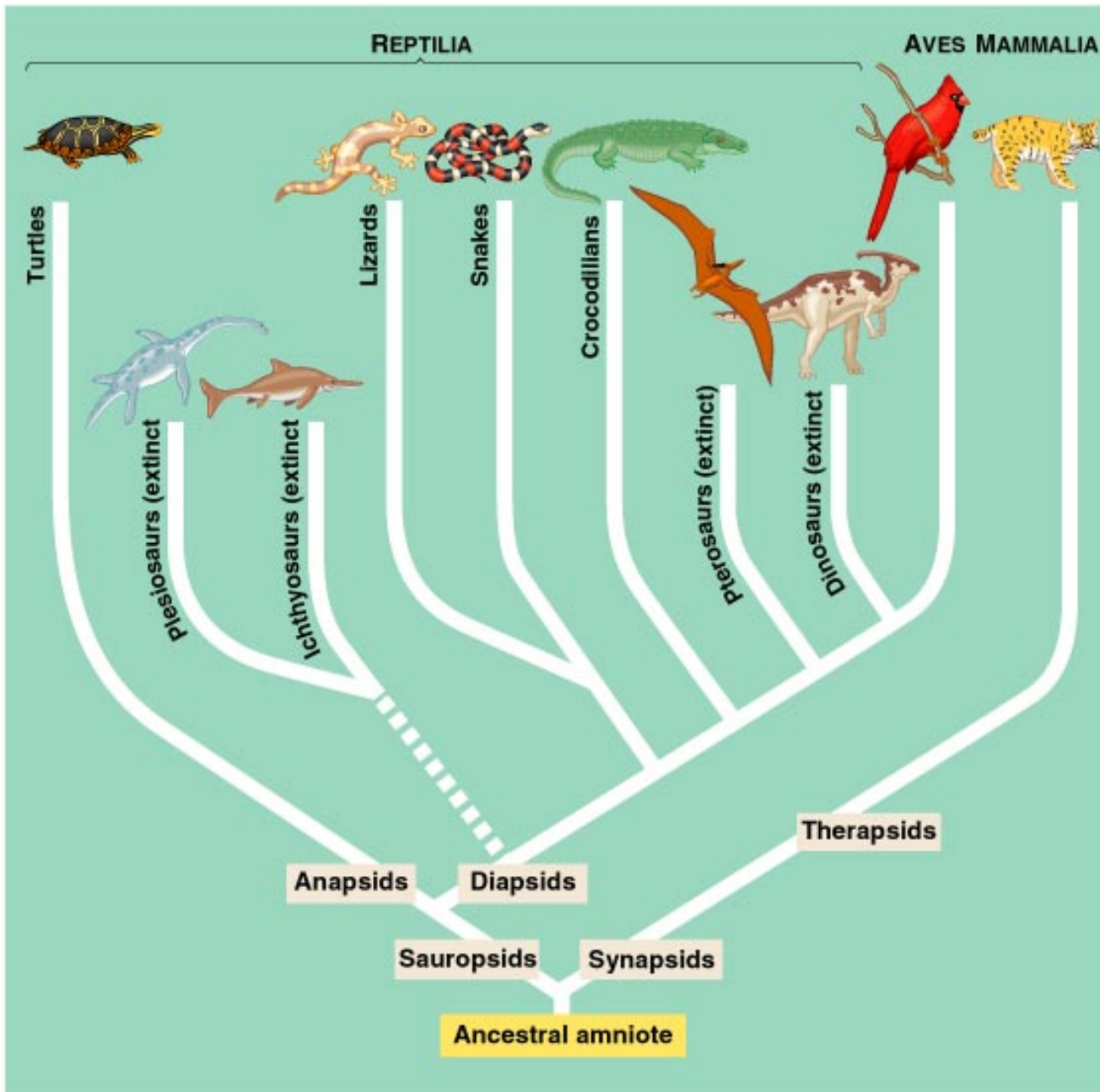
Biomathematics and Statistics Scotland

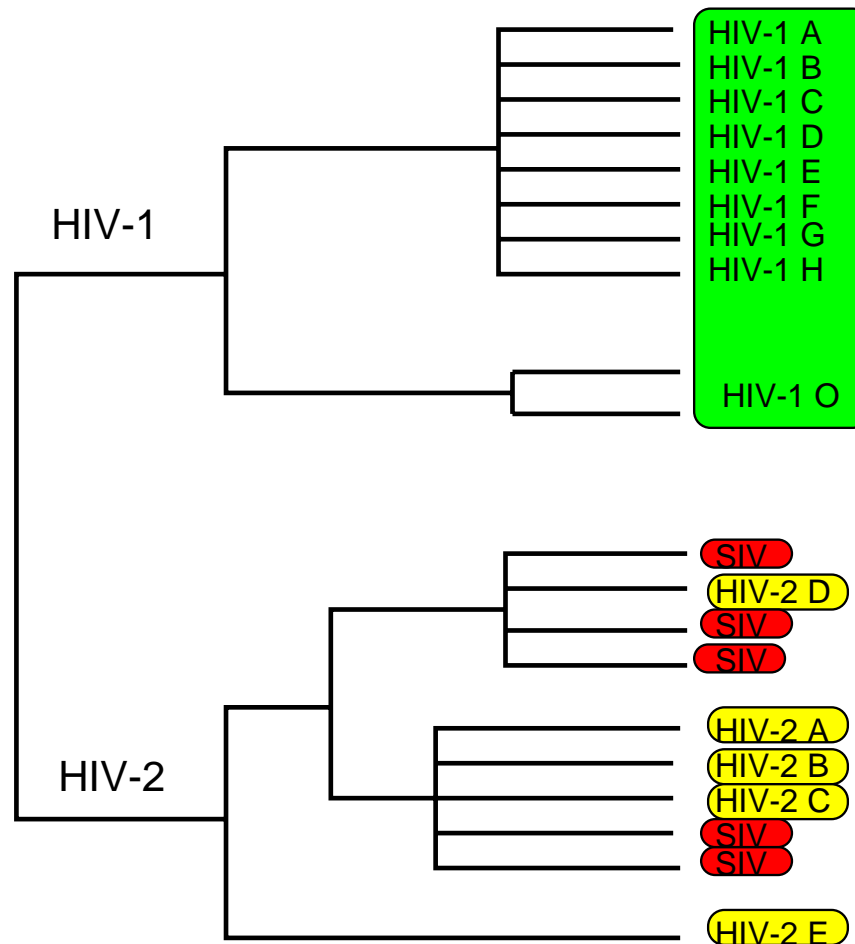
Edinburgh, United Kingdom

Email: dirk@bioss.ac.uk

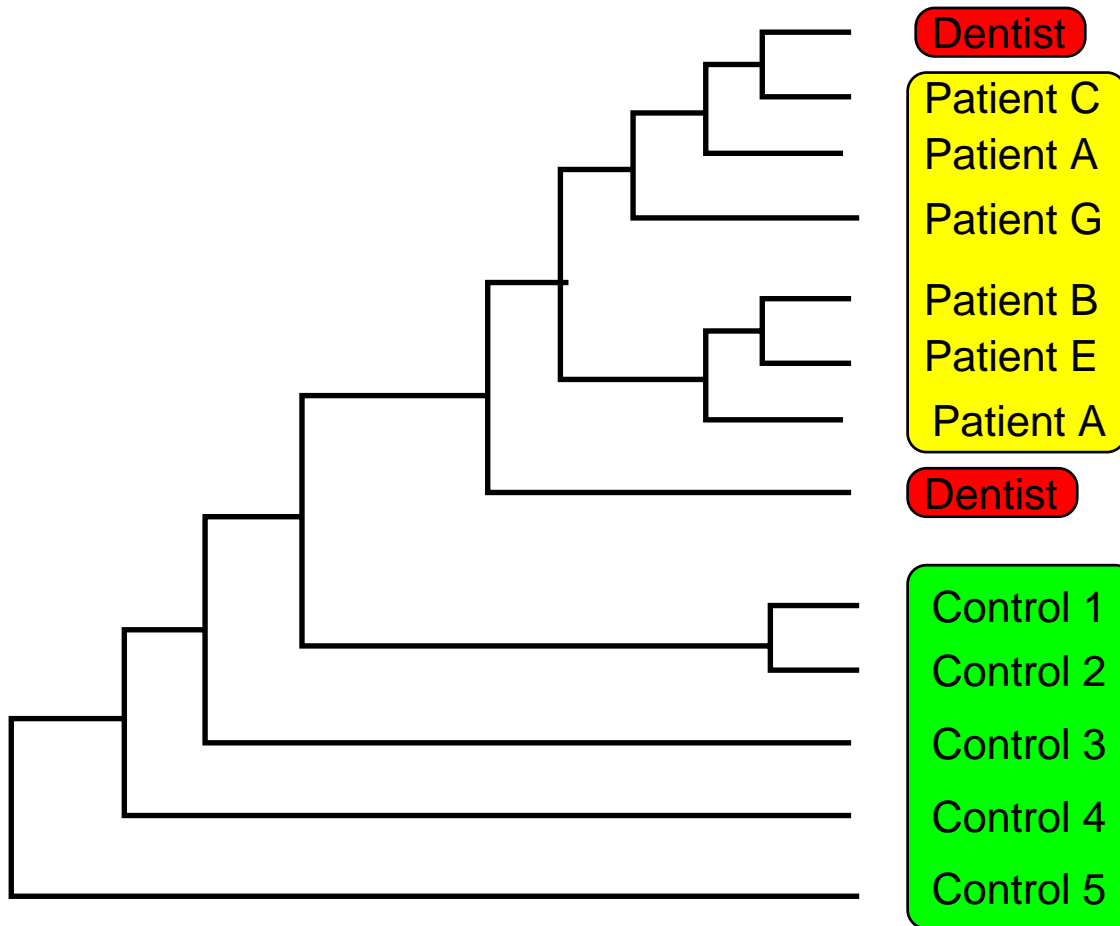
<http://www.bioss.ac.uk/~dirk>

- Phylogenetics
- Recombination





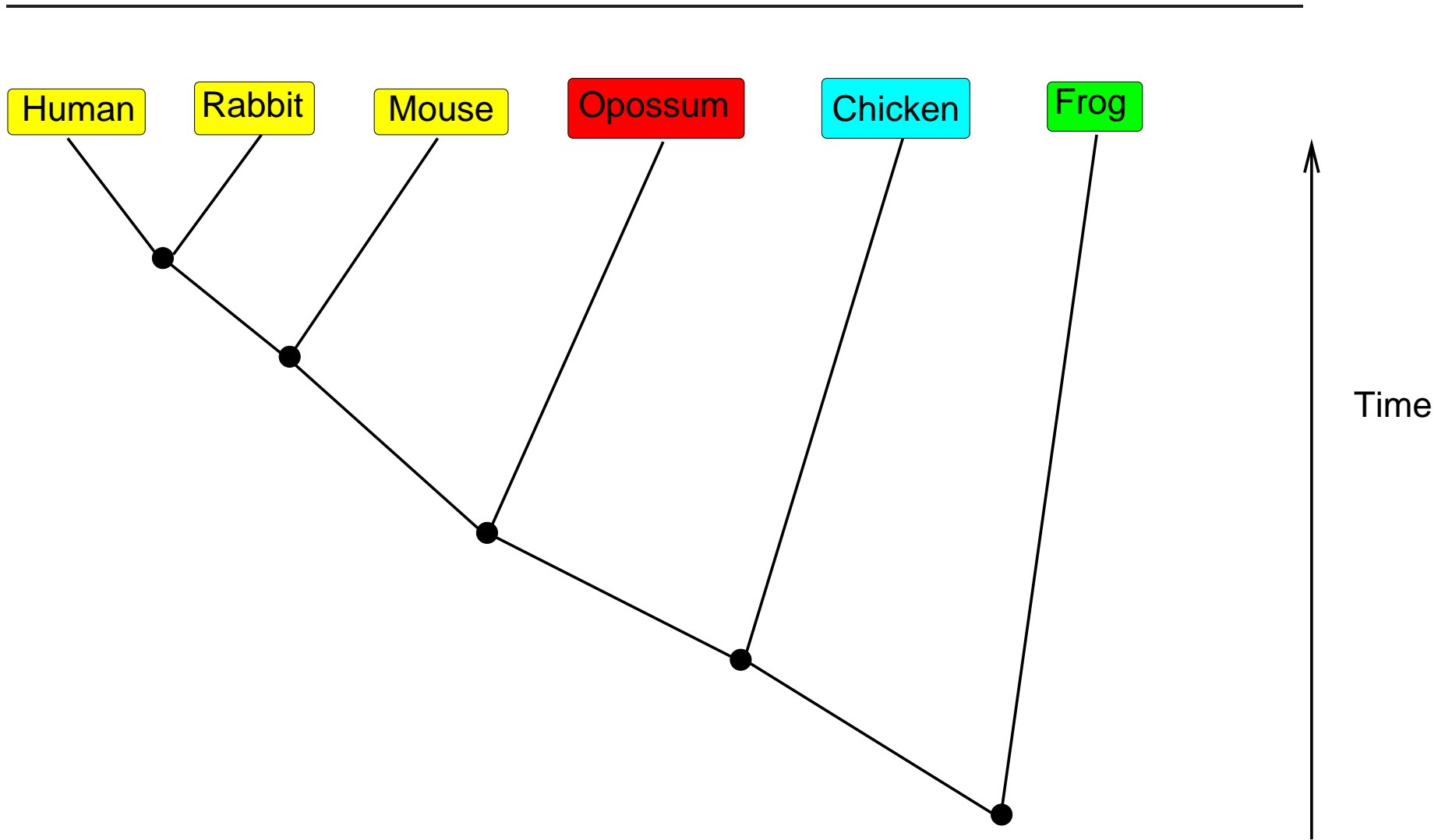
Adapted from Holmes (1998): Evolution in Health and Disease, Oxford University Press



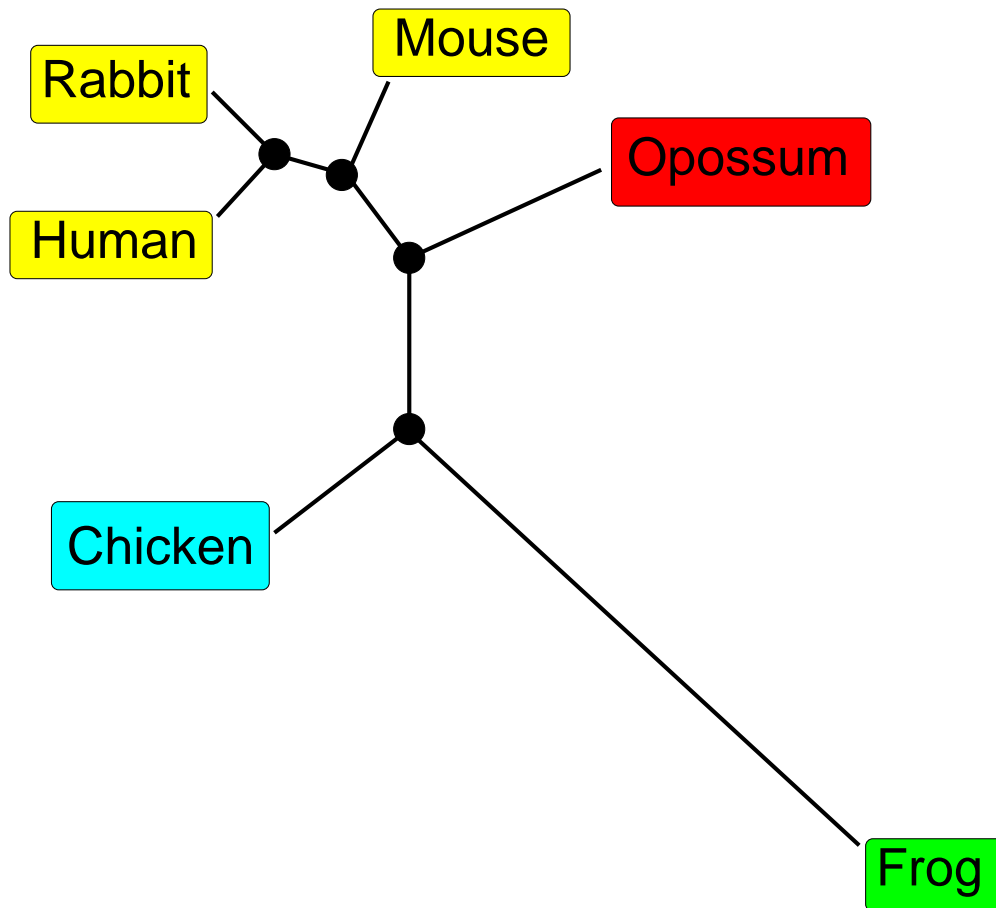
Data from Ou et al. (1992): Science 256, 1165-1171

Tree adapted from Page & Holmes (1998), Blackwell Science

Rooted Phylogenetic Tree



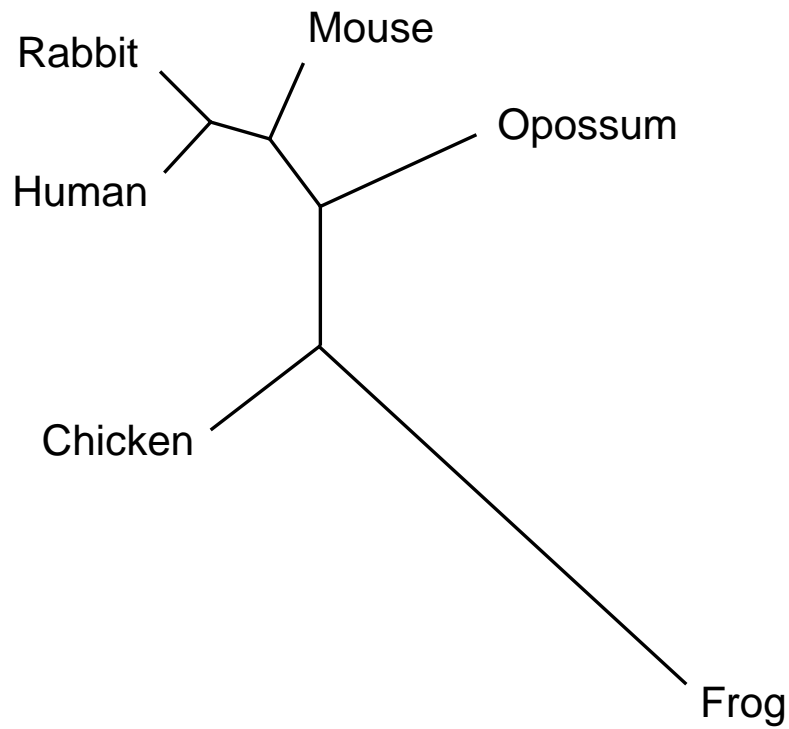
Unrooted Phylogenetic Tree



--> Topology

--> Branch lengths

Frog	G	C	T	T	G	A	C	T	T	C	T	G	A	G	G	T	T
Chicken	G	C	G	T	A	A	C	T	T	C	A	C	A	T	G	A	T
Human	G	C	G	T	C	A	C	T	T	G	A	G	A	C	G	C	T
Rabbit	G	C	G	T	C	A	C	T	T	G	A	G	A	C	G	C	T
Mouse	G	C	G	T	C	A	C	T	T	G	A	C	A	G	G	C	T
Opossum	G	C	G	T	C	A	C	T	T	G	A	G	A	C	G	C	T



--> Topology
 --> Branch lengths

Methods of phylogenetic inference

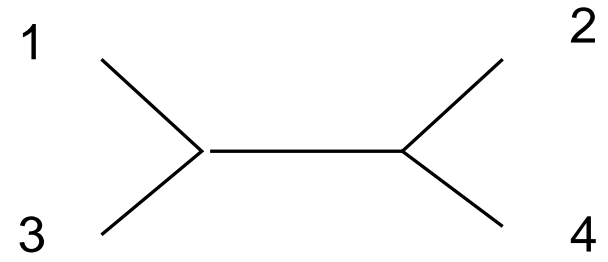
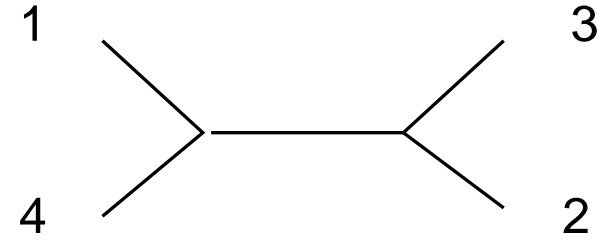
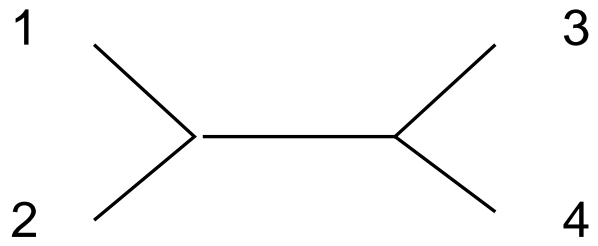
- Parsimony
- Likelihood

Methods of phylogenetic inference

- Parsimony
- Likelihood

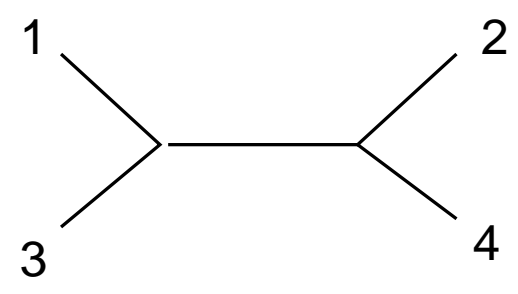
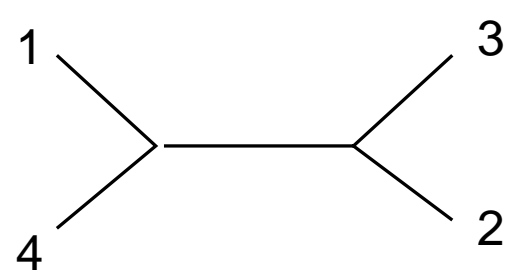
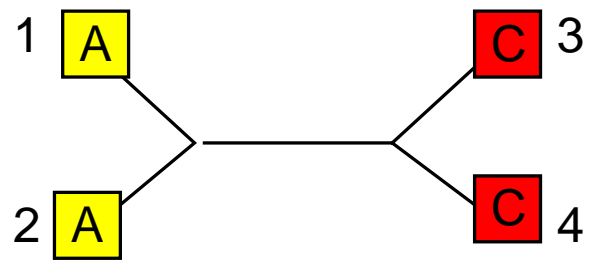
1	A	C	A	C	G
2	A	T	T	C	G
3	C	G	A	G	G
4	C	G	G	C	G

■ ■ ■



∇

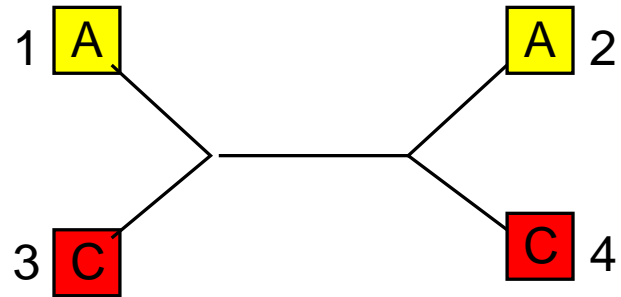
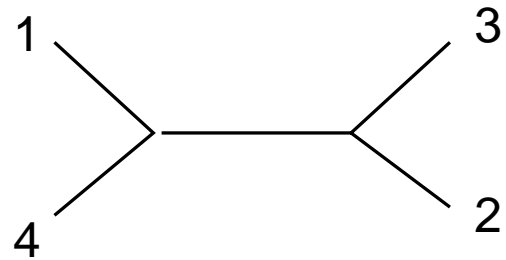
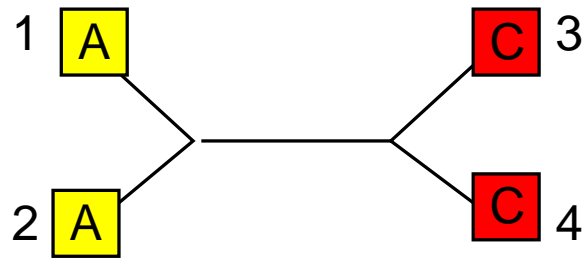
1	A	C	A	C	G			
2	A	T	T	C	G			
3	C	G	A	G	G	▪	▪	▪
4	C	G	G	C	G			



∨

1	A	C	A	C	G
2	A	T	T	C	G
3	C	G	A	G	G
4	C	G	G	C	G

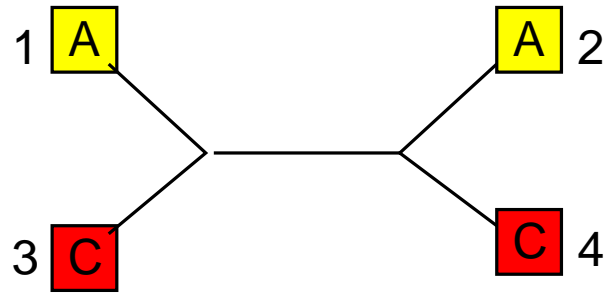
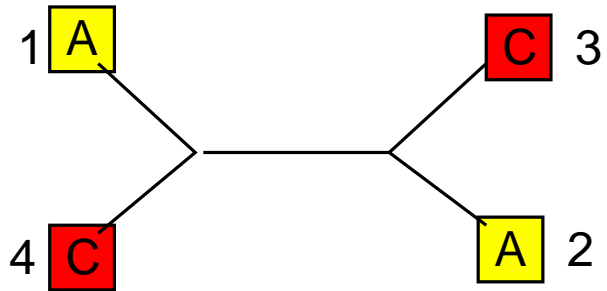
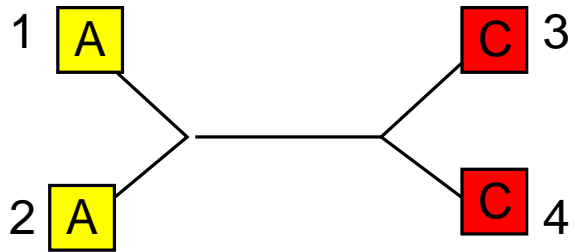
▪ ▪ ▪



∨

1	A	C	A	C	G
2	A	T	T	C	G
3	C	G	A	G	G
4	C	G	G	C	G

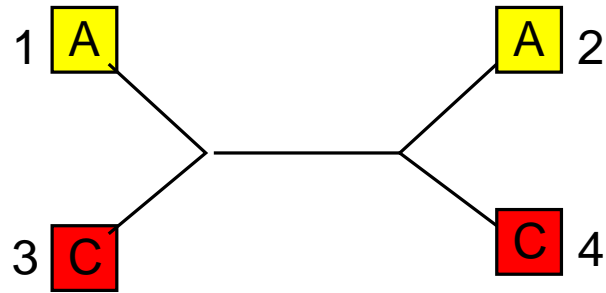
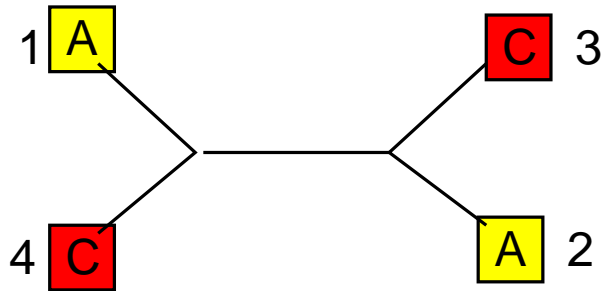
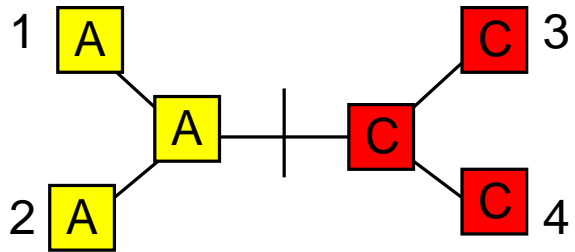
. . .



∨

1	A	C	A	C	G
2	A	T	T	C	G
3	C	G	A	G	G
4	C	G	G	C	G

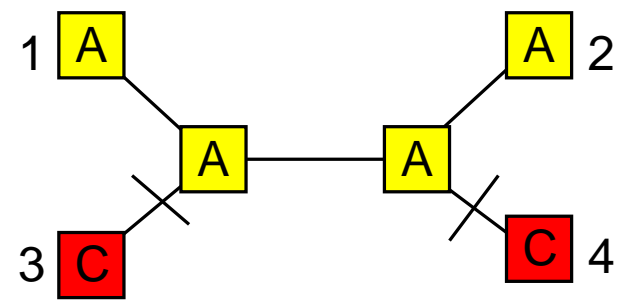
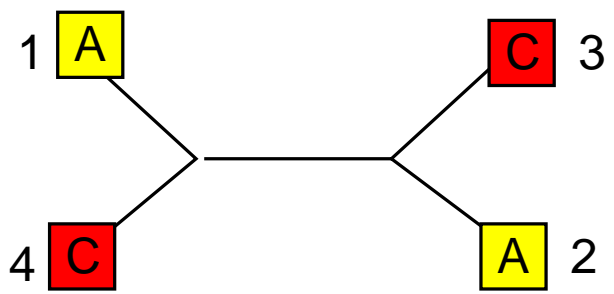
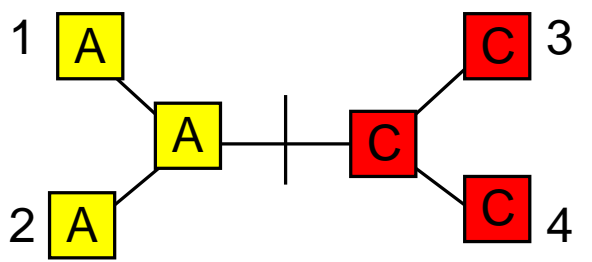
. . .



↓

1	A	C	A	C	G
2	A	T	T	C	G
3	C	G	A	G	G
4	C	G	G	C	G

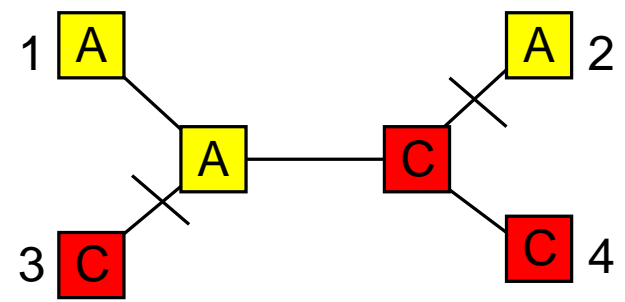
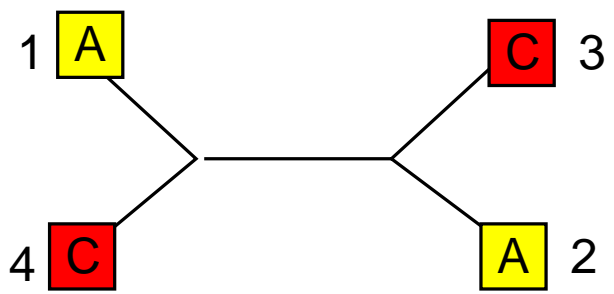
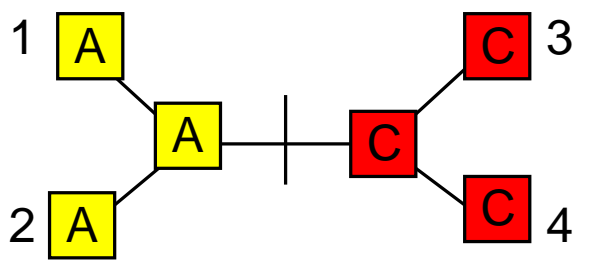
. . .



∇

1	A	C	A	C	G
2	A	T	T	C	G
3	C	G	A	G	G
4	C	G	G	C	G

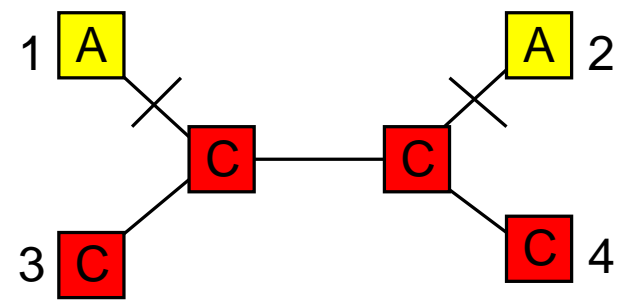
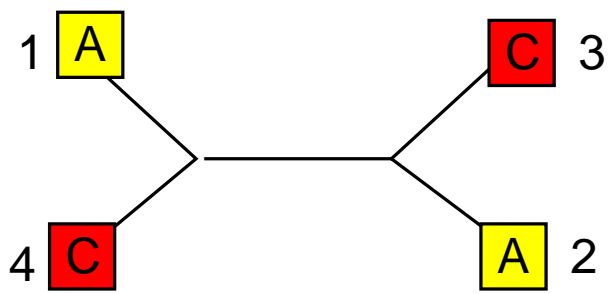
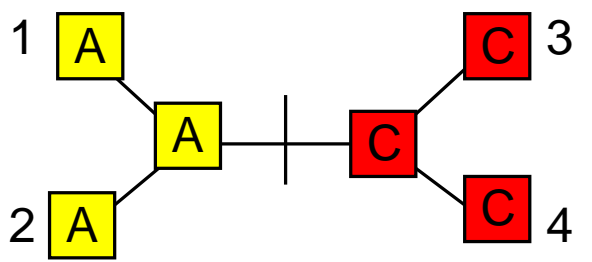
. . .

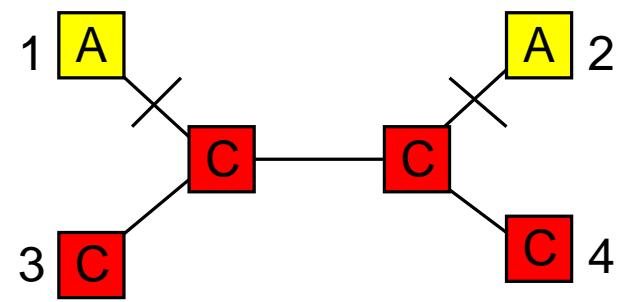
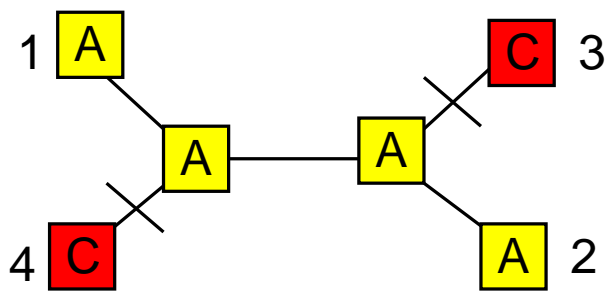
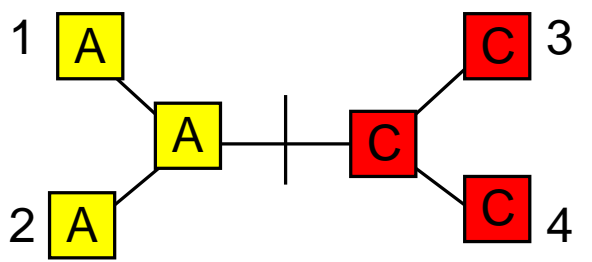
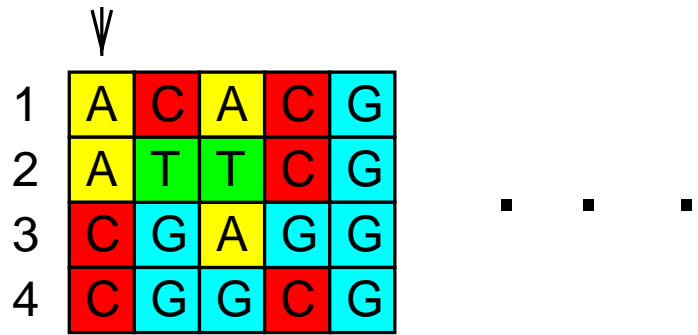


↓

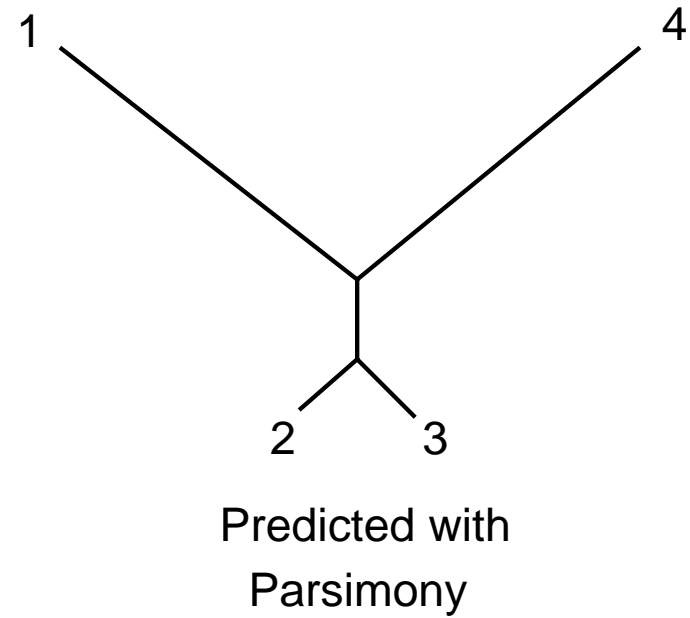
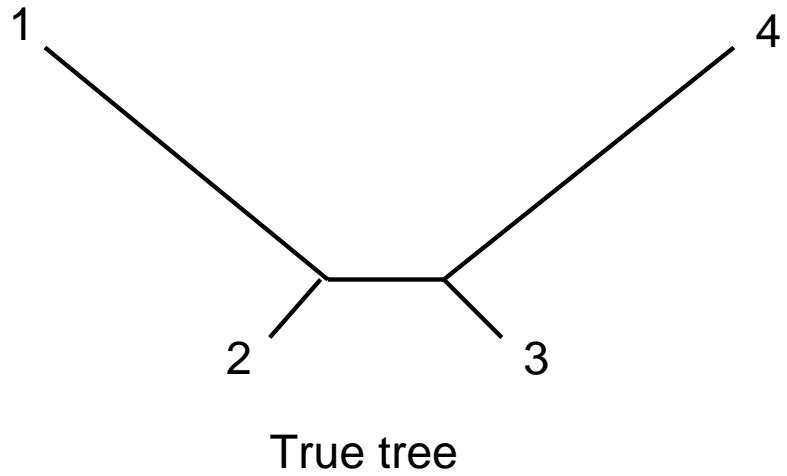
1	A	C	A	C	G
2	A	T	T	C	G
3	C	G	A	G	G
4	C	G	G	C	G

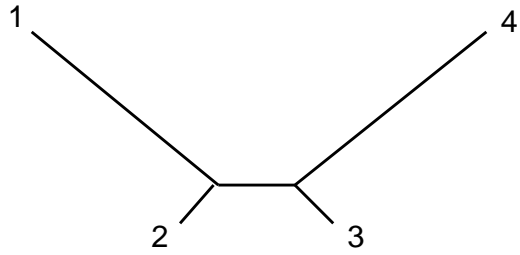
. . .



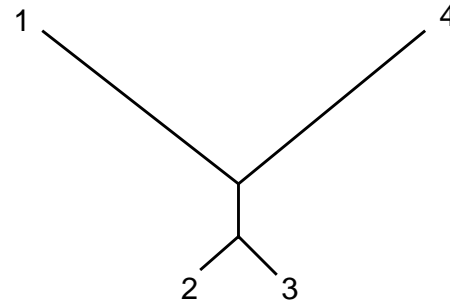


Failure of parsimony

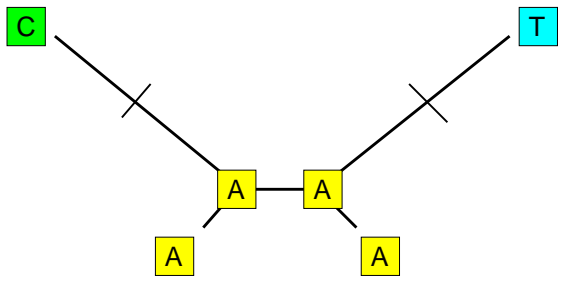




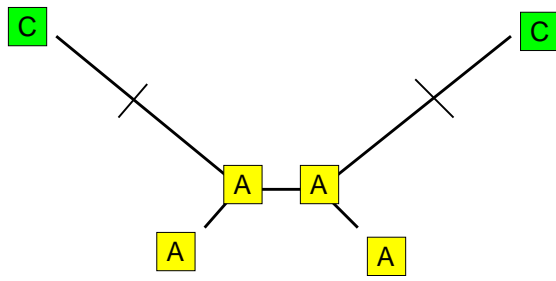
True tree



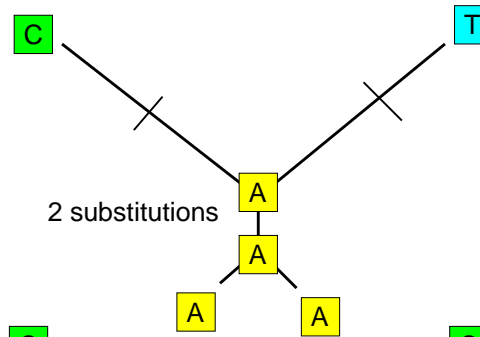
Predicted with Parsimony



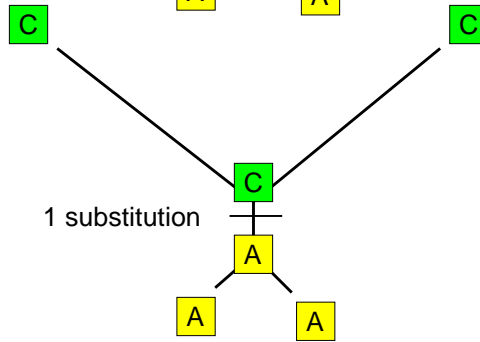
2 substitutions



2 substitutions



2 substitutions

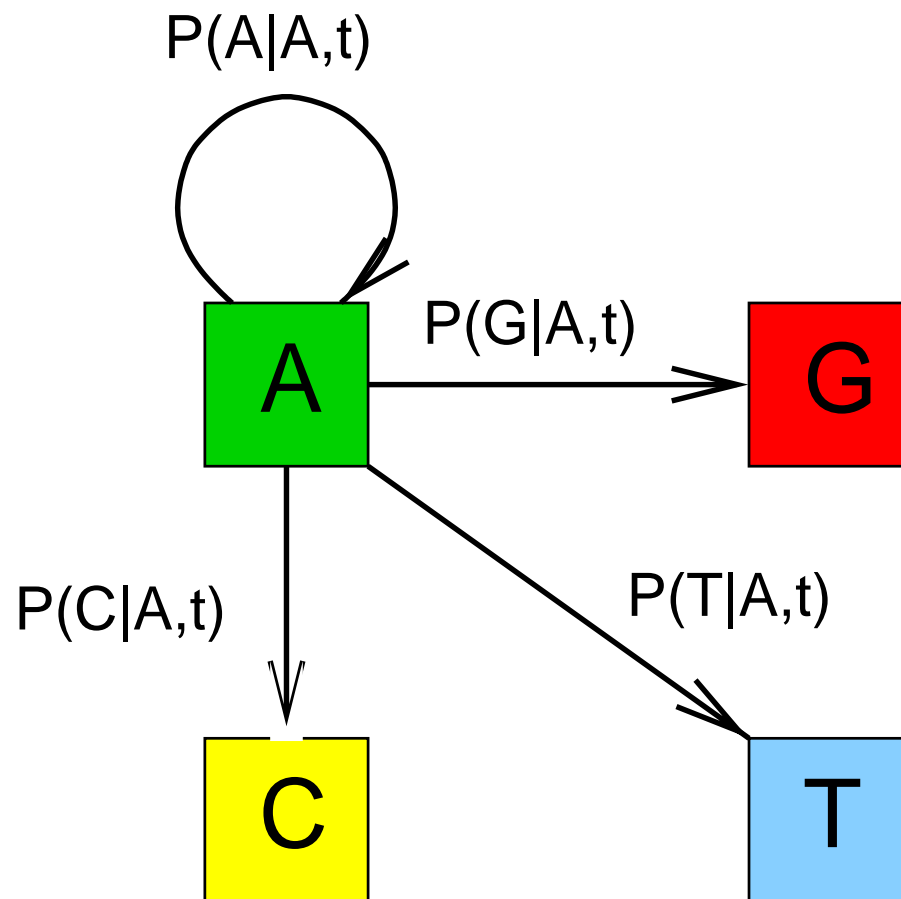


1 substitution

Methods of phylogenetic inference

- Parsimony
- Likelihood

Mutation probabilities



Markov model of evolution

$$\mathbf{P}(t) = \begin{bmatrix} P(y(t) = A|y(0) = A) & P(y(t) = A|y(0) = G) & \dots \\ P(y(t) = G|y(0) = A) & P(y(t) = G|y(0) = G) & \dots \\ P(y(t) = C|y(0) = A) & P(y(t) = C|y(0) = G) & \dots \\ P(y(t) = T|y(0) = A) & P(y(t) = T|y(0) = G) & \dots \end{bmatrix}$$

Markov model of evolution

$$\mathbf{P}(t) = \begin{bmatrix} P(y(t) = A|y(0) = A) & P(y(t) = A|y(0) = G) & \dots \\ P(y(t) = G|y(0) = A) & P(y(t) = G|y(0) = G) & \dots \\ P(y(t) = C|y(0) = A) & P(y(t) = C|y(0) = G) & \dots \\ P(y(t) = T|y(0) = A) & P(y(t) = T|y(0) = G) & \dots \end{bmatrix}$$

- Process is **Markov**:

$$P[y(t + \Delta t)|y(t), y(t - \Delta t), \dots] = P[y(t + \Delta t)|y(t)]$$

Markov model of evolution

$$\mathbf{P}(t) = \begin{bmatrix} P(y(t) = A|y(0) = A) & P(y(t) = A|y(0) = G) & \dots \\ P(y(t) = G|y(0) = A) & P(y(t) = G|y(0) = G) & \dots \\ P(y(t) = C|y(0) = A) & P(y(t) = C|y(0) = G) & \dots \\ P(y(t) = T|y(0) = A) & P(y(t) = T|y(0) = G) & \dots \end{bmatrix}$$

- Process is **Markov**:

$$P[y(t + \Delta t)|y(t), y(t - \Delta t), \dots] = P[y(t + \Delta t)|y(t)]$$

- The Markov process is **homogenous**:

$$P[y(t + t_0)|y(t_0)] = P[y(t)|y(0)]$$

Markov model of evolution

$$\mathbf{P}(t) = \begin{bmatrix} P(y(t) = A|y(0) = A) & P(y(t) = A|y(0) = G) & \dots \\ P(y(t) = G|y(0) = A) & P(y(t) = G|y(0) = G) & \dots \\ P(y(t) = C|y(0) = A) & P(y(t) = C|y(0) = G) & \dots \\ P(y(t) = T|y(0) = A) & P(y(t) = T|y(0) = G) & \dots \end{bmatrix}$$

- Process is **Markov**:

$$P[y(t + \Delta t)|y(t), y(t - \Delta t), \dots] = P[y(t + \Delta t)|y(t)]$$

- The Markov process is **homogenous**:

$$P[y(t + t_0)|y(t_0)] = P[y(t)|y(0)]$$

- The Markov process is the **same for all positions** .

Markov model of evolution

$$\mathbf{P}(t) = \begin{bmatrix} P(y(t) = A|y(0) = A) & P(y(t) = A|y(0) = G) & \dots \\ P(y(t) = G|y(0) = A) & P(y(t) = G|y(0) = G) & \dots \\ P(y(t) = C|y(0) = A) & P(y(t) = C|y(0) = G) & \dots \\ P(y(t) = T|y(0) = A) & P(y(t) = T|y(0) = G) & \dots \end{bmatrix}$$

- Process is **Markov**:

$$P[y(t + \Delta t)|y(t), y(t - \Delta t), \dots] = P[y(t + \Delta t)|y(t)]$$

- The Markov process is **homogenous**:

$$P[y(t + t_0)|y(t_0)] = P[y(t)|y(0)]$$

- The Markov process is the **same for all positions** .
- Substitutions at different positions are **independent** of each other:

$$P[(y_1(t), \dots, y_N(t)|y_1(0), \dots, y_N(0))] = \prod_{i=1}^N P[y_i(t)|y_i(0)]$$

Transition Rates

$$\mathbf{P}(0) = \mathbf{I}$$

Transition Rates

$$\mathbf{P}(0) = \mathbf{I}$$

$$\mathbf{P}(dt) - \mathbf{P}(0) = \mathbf{R}dt$$

Transition Rates

$$\mathbf{P}(0) = \mathbf{I}$$

$$\mathbf{P}(dt) - \mathbf{P}(0) = \mathbf{R}dt$$

$$\mathbf{P}(t + dt) = \mathbf{P}(dt)\mathbf{P}(t) = (\mathbf{I} + \mathbf{R}dt)\mathbf{P}(t)$$

Transition Rates

$$\mathbf{P}(0) = \mathbf{I}$$

$$\mathbf{P}(dt) - \mathbf{P}(0) = \mathbf{R}dt$$

$$\mathbf{P}(t + dt) = \mathbf{P}(dt)\mathbf{P}(t) = (\mathbf{I} + \mathbf{R}dt)\mathbf{P}(t)$$

$$\frac{d\mathbf{P}}{dt} = \mathbf{R}\mathbf{P} \implies \mathbf{P}(t) = e^{\mathbf{R}t}$$

Transition Rates

$$\mathbf{P}(0) = \mathbf{I}$$

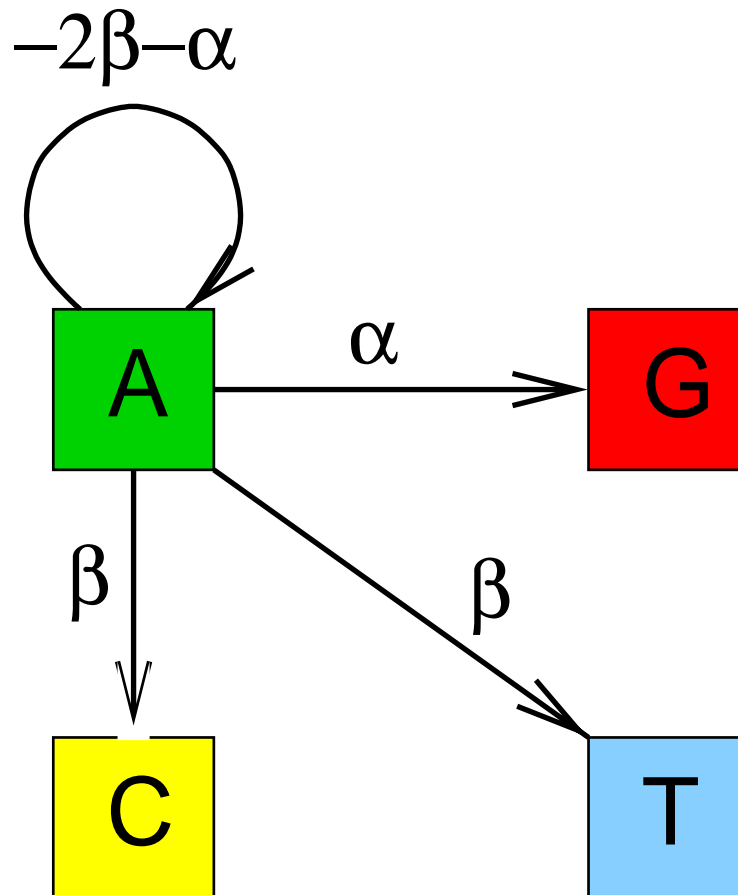
$$\mathbf{P}(dt) - \mathbf{P}(0) = \mathbf{R}dt$$

$$\mathbf{P}(t + dt) = \mathbf{P}(dt)\mathbf{P}(t) = (\mathbf{I} + \mathbf{R}dt)\mathbf{P}(t)$$

$$\frac{d\mathbf{P}}{dt} = \mathbf{R}\mathbf{P} \implies \mathbf{P}(t) = e^{\mathbf{R}t}$$

$$\mathbf{R} = \begin{bmatrix} -2\beta - \alpha & \beta & \alpha & \beta \\ \beta & -2\beta - \alpha & \beta & \alpha \\ \alpha & \beta & -2\beta - \alpha & \beta \\ \beta & \alpha & \beta & -2\beta - \alpha \end{bmatrix}$$

Transition Rates



Transition Probabilities

$$\mathbf{P}(t) = e^{\mathbf{R}t} = \begin{bmatrix} d(t) & f(t) & g(t) & f(t) \\ f(t) & d(t) & f(t) & g(t) \\ g(t) & f(t) & d(t) & f(t) \\ f(t) & g(t) & f(t) & d(t) \end{bmatrix}$$

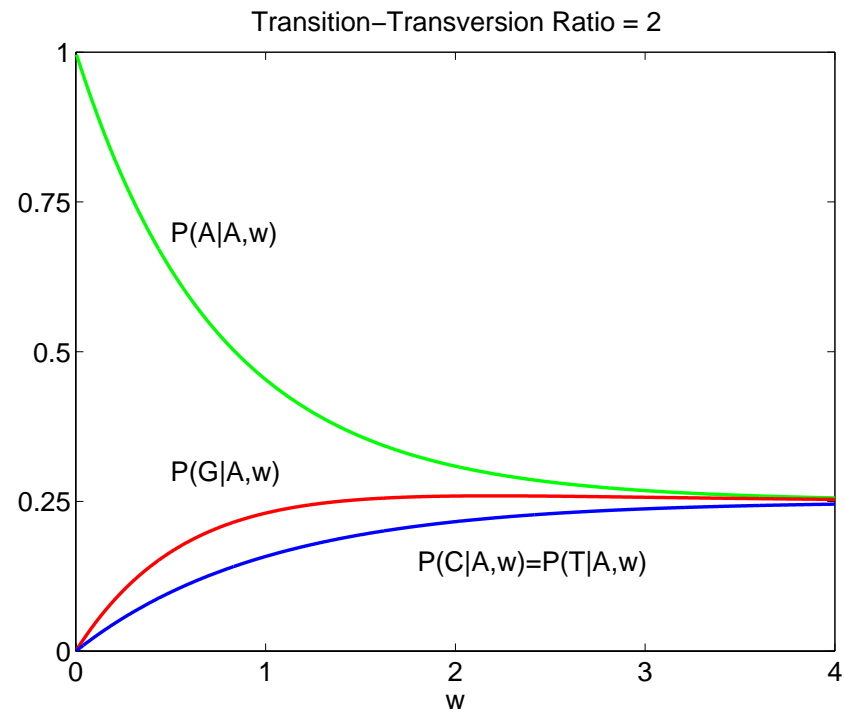
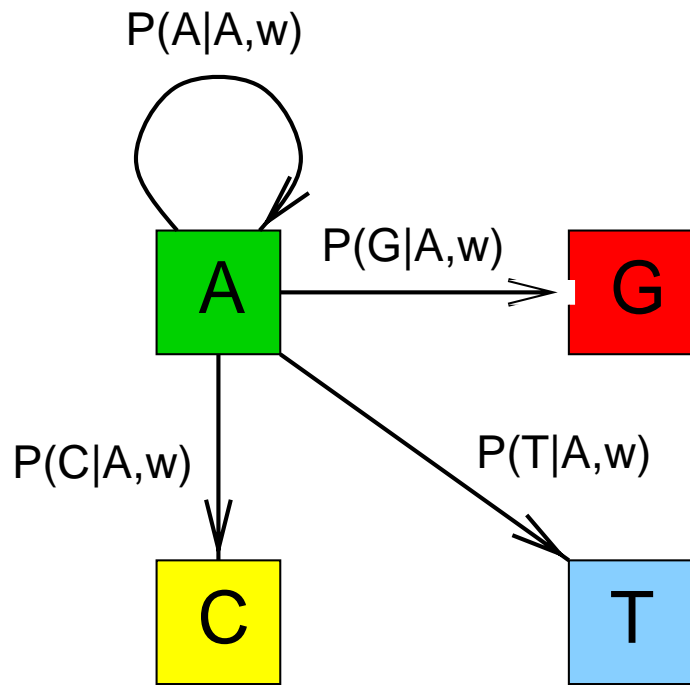
$$f(t) = \frac{1}{4}(1 - e^{-4\beta t}) \quad g(t) = \frac{1}{4}(1 + e^{-4\beta t} - 2e^{-2(\alpha+\beta)t}) \quad d(t) = 1 - 2f(t) - g(t)$$

Molecular time: $w = 4\beta t$

$$\begin{aligned} f(w) &= \frac{1}{4}(1 - e^{-w}) \\ g(w) &= \frac{1}{4}(1 + e^{-w} - 2e^{-\frac{\tau+1}{2}w}) \\ d(w) &= 1 - 2f(w) - g(w) \end{aligned}$$

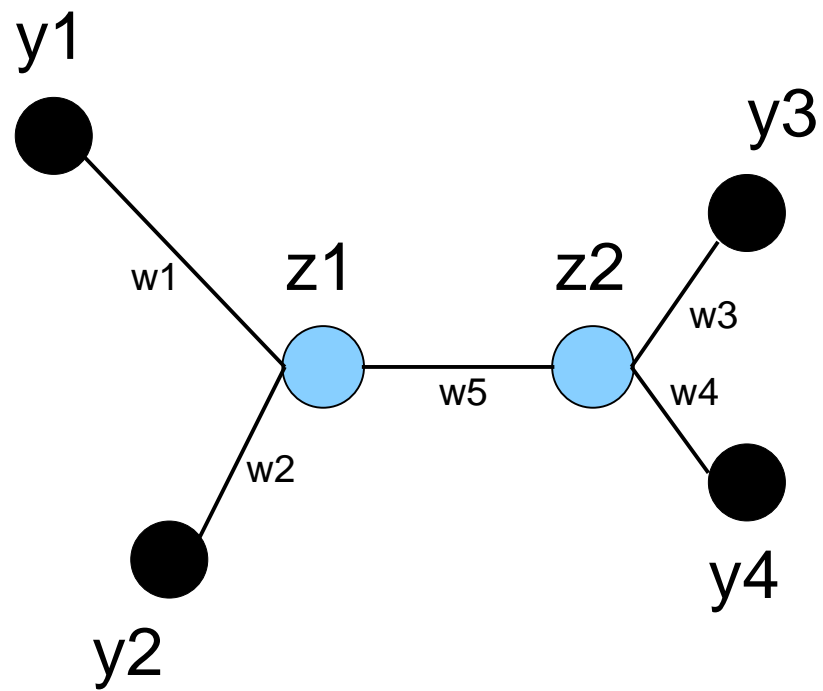
Transition-transversion ratio: $\tau = \frac{\alpha}{\beta}$

Mutation probabilities



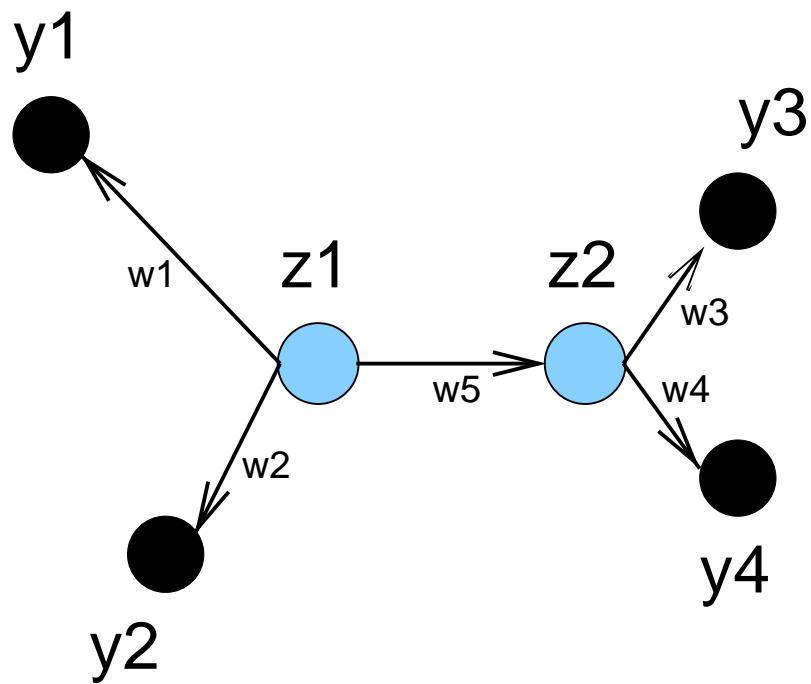
$$w = 4\beta t$$

Phylogenetic tree as an undirected graph



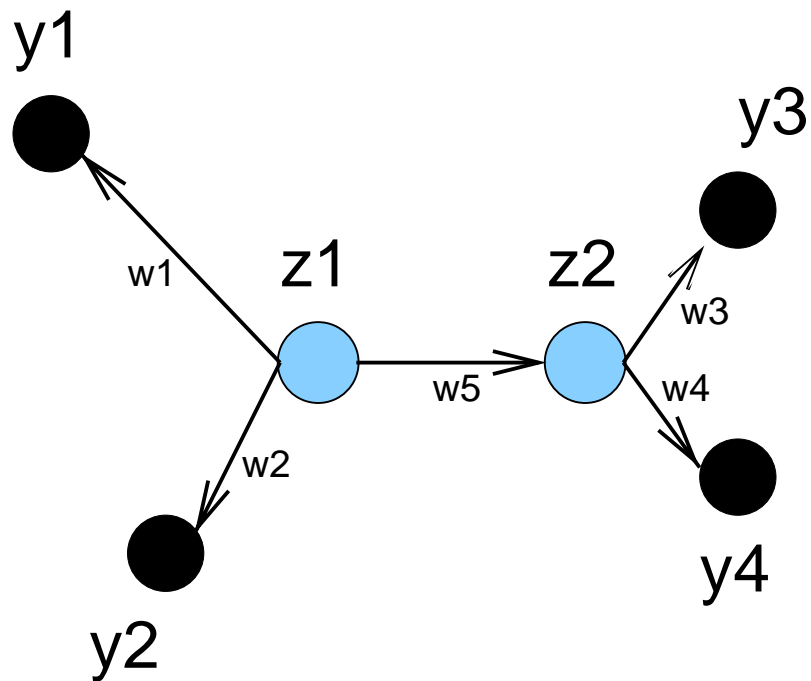
$$P(y_1, y_2, y_3, y_4, z_1, z_2 | \mathbf{w})$$

Phylogenetic tree as a directed graph



$$P(y_1, y_2, y_3, y_4, z_1, z_2 | \mathbf{w})$$

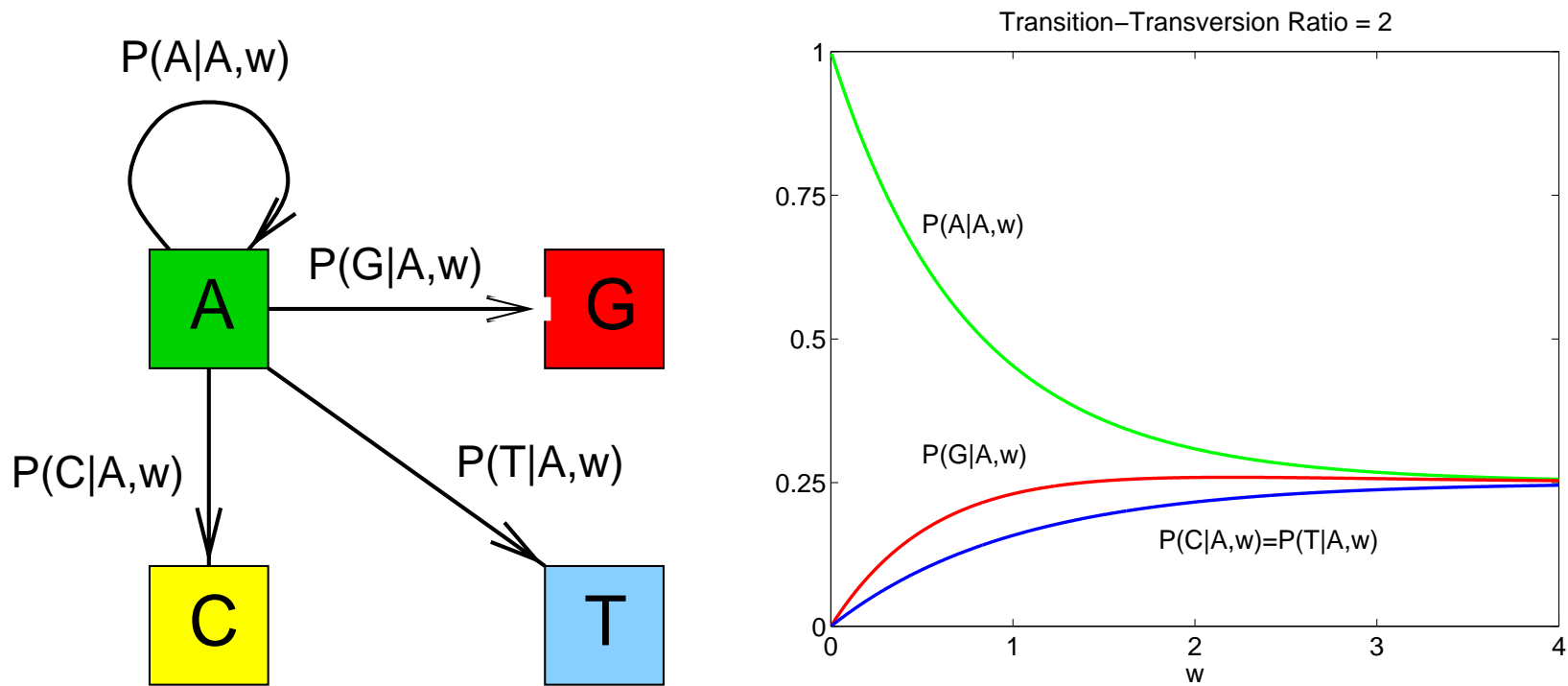
Phylogenetic tree as a directed graph



$$P(y_1, y_2, y_3, y_4, z_1, z_2 | \mathbf{w})$$

$$= P(y_1 | z_1, w_1) P(y_2 | z_1, w_2) P(z_2 | z_1, w_5) P(y_3 | z_2, w_3) P(y_4 | z_2, w_4) P(z_1)$$

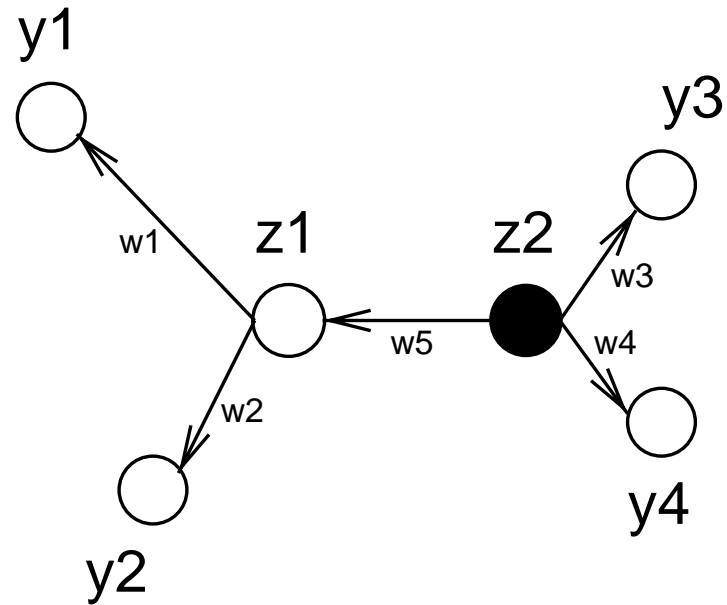
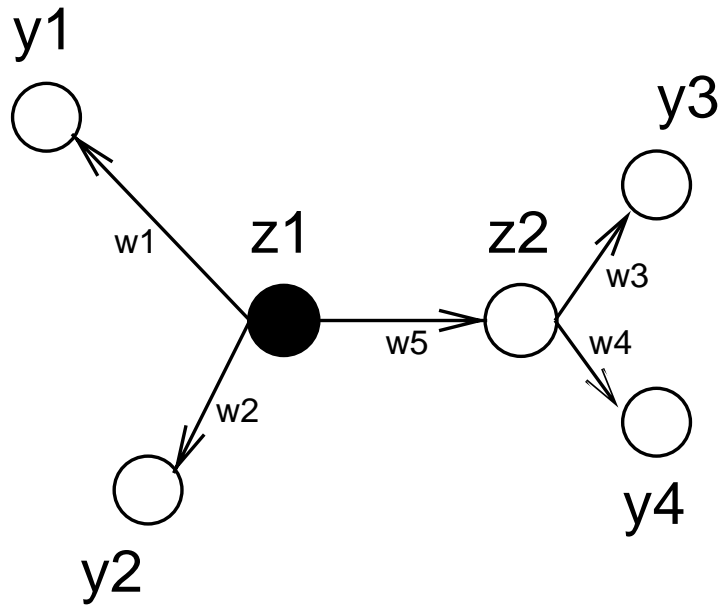
Mutation probabilities



$$w = 4\beta t$$

branch length = mutation rate \times time

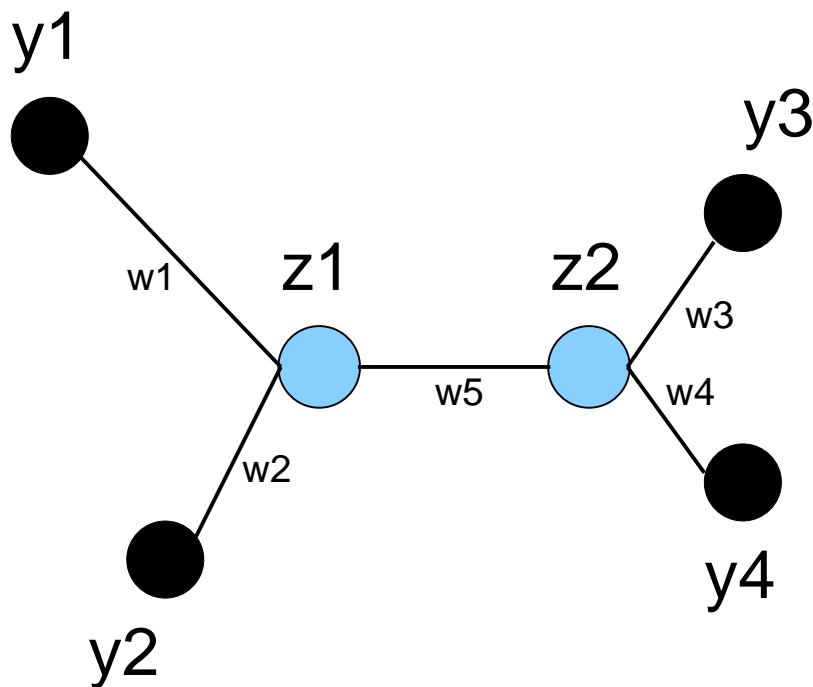
Different directed graphs



Left : $P(y_1, y_2, y_3, y_4, z_1, z_2 | \mathbf{w}) =$
 $P(y_1 | z_1, w_1) P(y_2 | z_1, w_2) P(y_3 | z_2, w_3) P(y_4 | z_2, w_4) P(z_2 | z_1, w_5) P(z_1)$

Right : $P(y_1, y_2, y_3, y_4, z_1, z_2 | \mathbf{w}) =$
 $P(y_1 | z_1, w_1) P(y_2 | z_1, w_2) P(y_3 | z_2, w_3) P(y_4 | z_2, w_4) P(z_1 | z_2, w_5) P(z_2)$

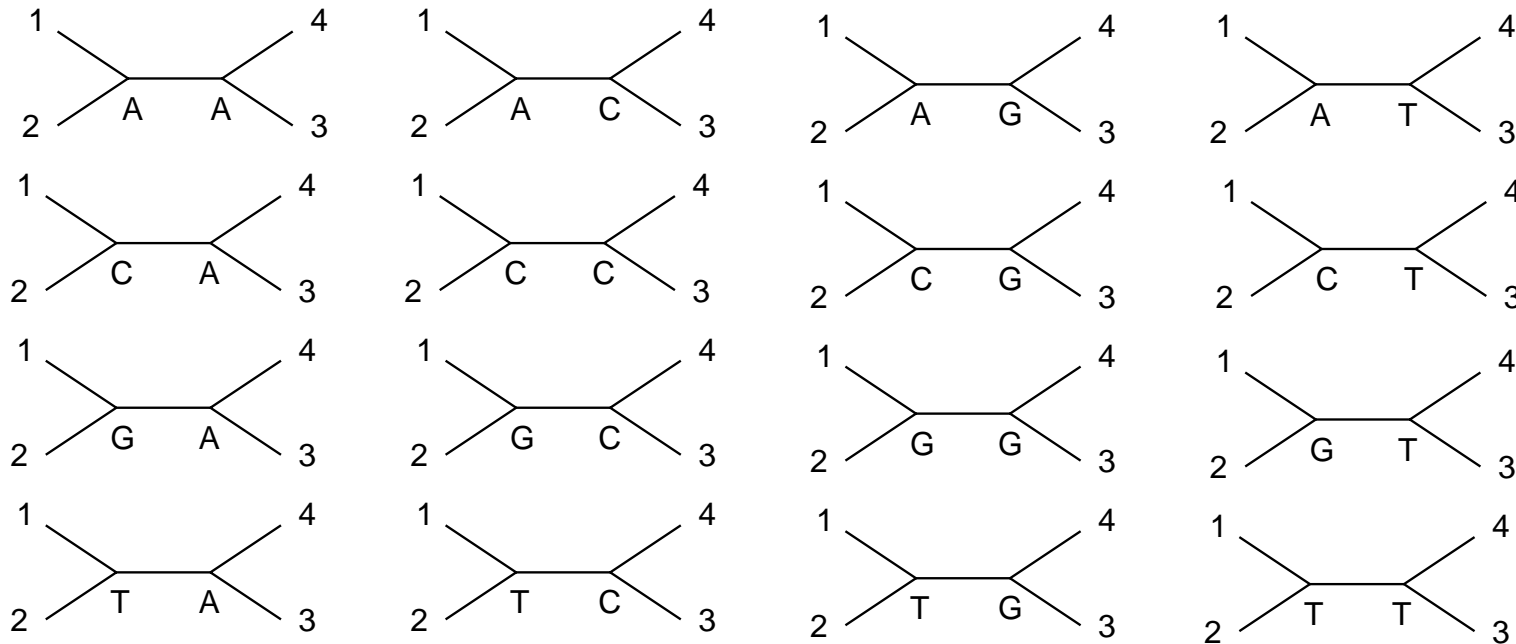
Expansion of the joint probability



$$P(y_1, y_2, y_3, y_4, z_1, z_2 | \mathbf{w})$$

$$= P(y_1 | z_1, w_1) P(y_2 | z_1, w_2) P(z_2 | z_1, w_5) P(y_3 | z_2, w_3) P(y_4 | z_2, w_4) P(z_1)$$

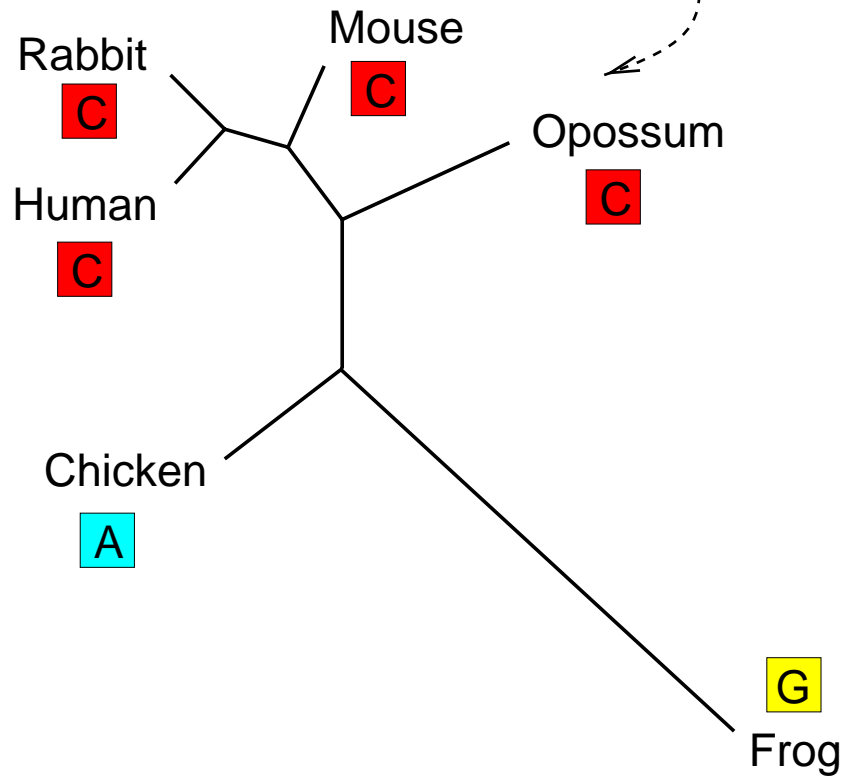
Marginalisation



$$P(y_1, y_2, y_3, y_4 | \mathbf{w}) = \sum_{z_1} \sum_{z_2} P(y_1, y_2, y_3, y_4, z_1, z_2 | \mathbf{w})$$

∇

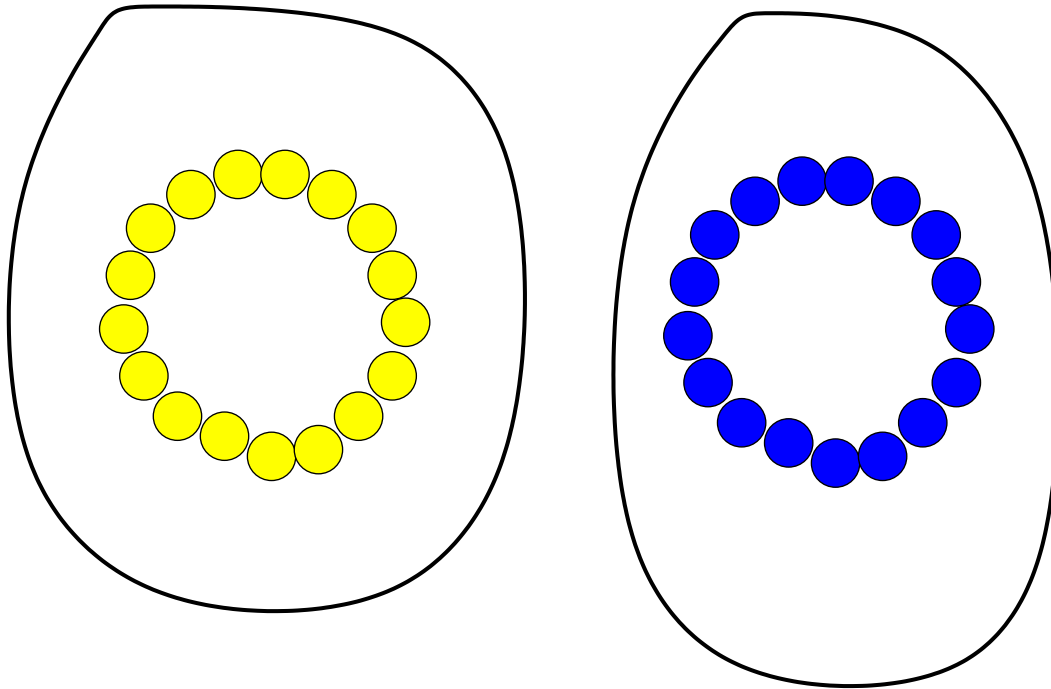
Frog	G	C	T	T	G	A	C	T	T	C	T	G	A	G	G	T	T
Chicken	G	C	G	T	A	A	C	T	T	C	A	C	A	T	G	A	T
Human	G	C	G	T	C	A	C	T	T	G	A	G	A	C	G	C	T
Rabbit	G	C	G	T	C	A	C	T	T	G	A	G	A	C	G	C	T
Mouse	G	C	G	T	C	A	C	T	T	G	A	C	A	G	G	C	T
Opossum	G	C	G	T	C	A	C	T	T	G	A	G	A	C	G	C	T



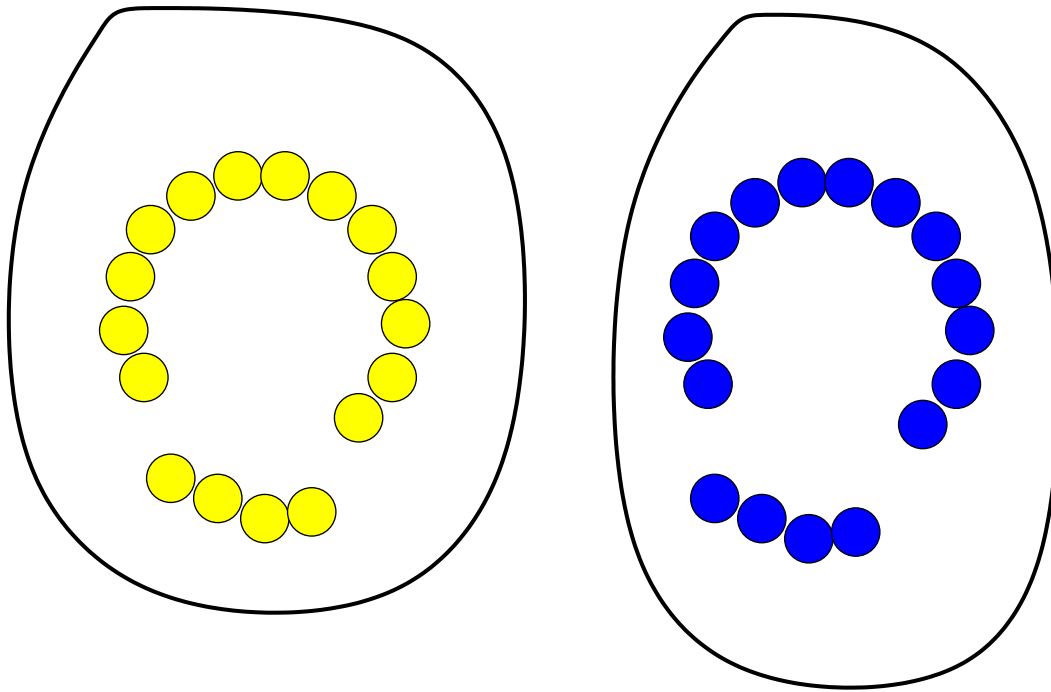
--> Likelihood

Topology
Branch lengths

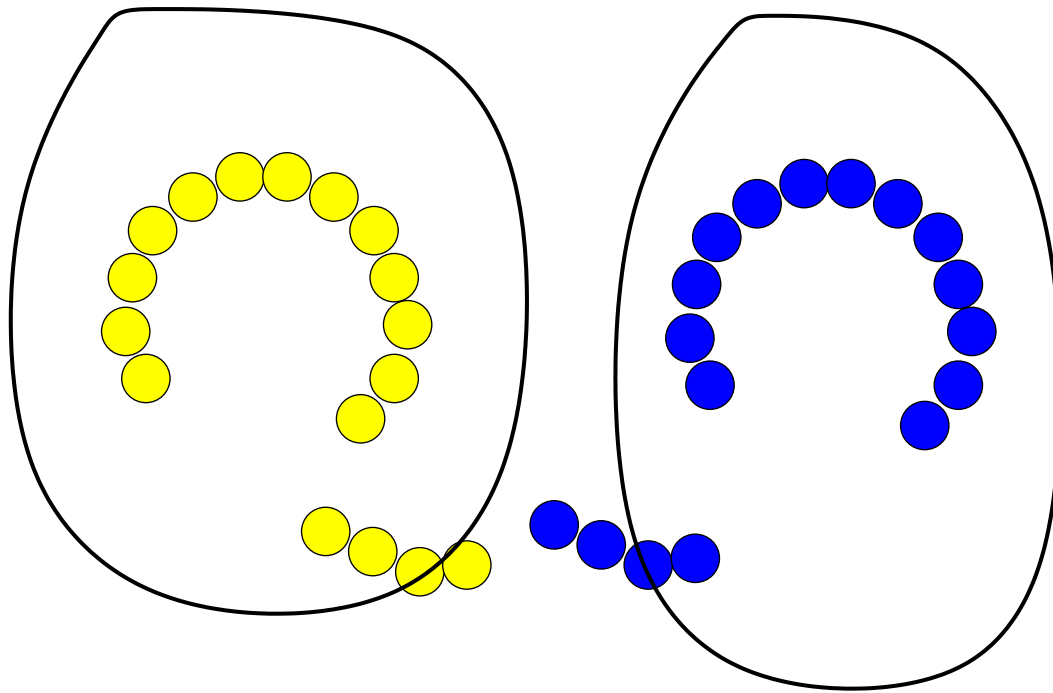
Recombination



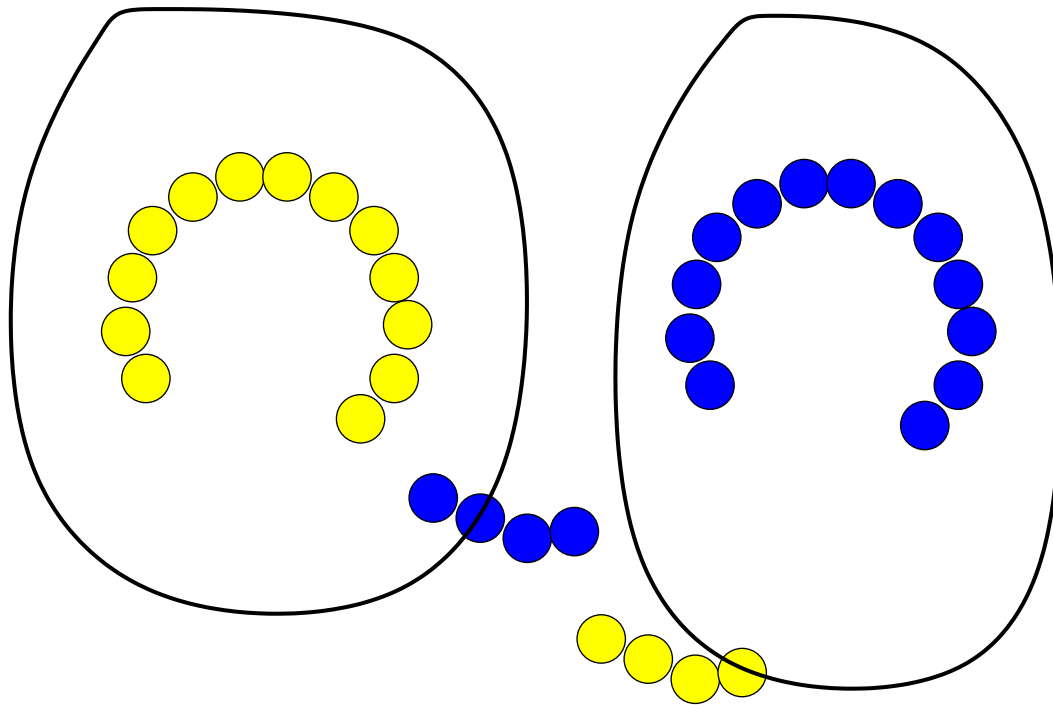
Recombination



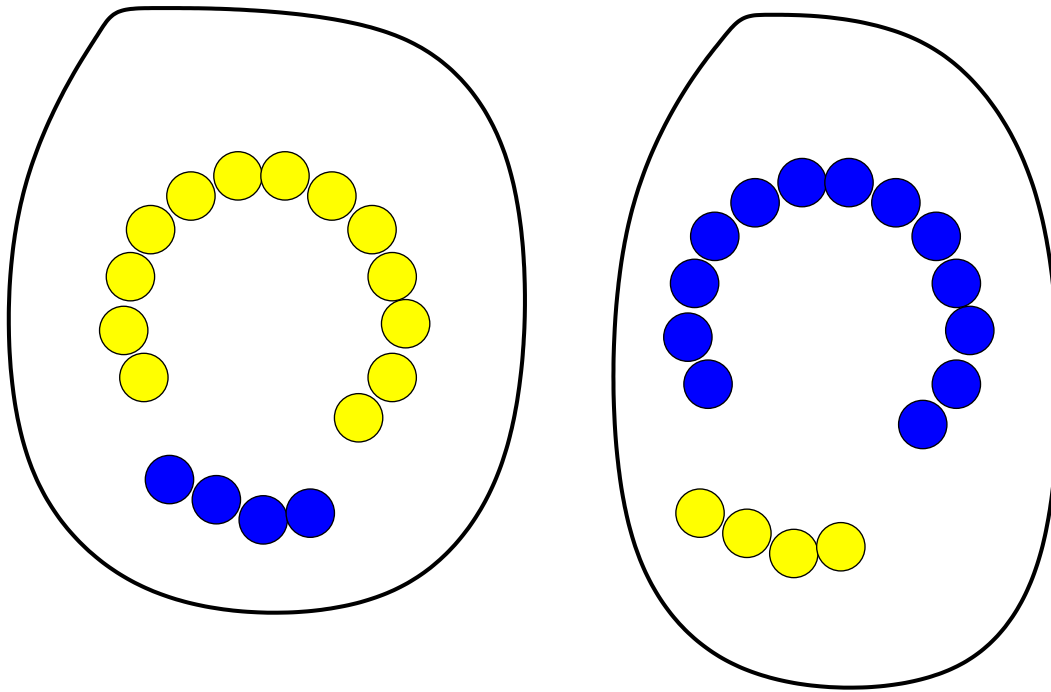
Recombination



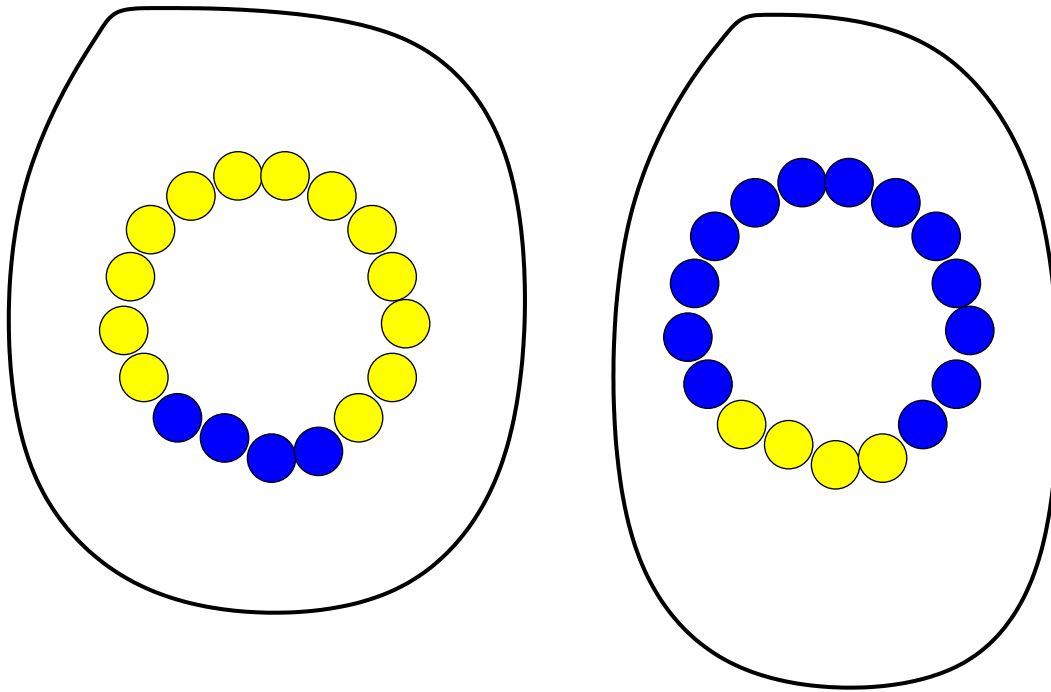
Recombination



Recombination



Recombination



1995

Robertson, Sharp, McCutchan, Hahn

Recombination in HIV-1

Nature 374, pp.124-126

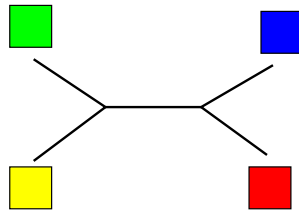
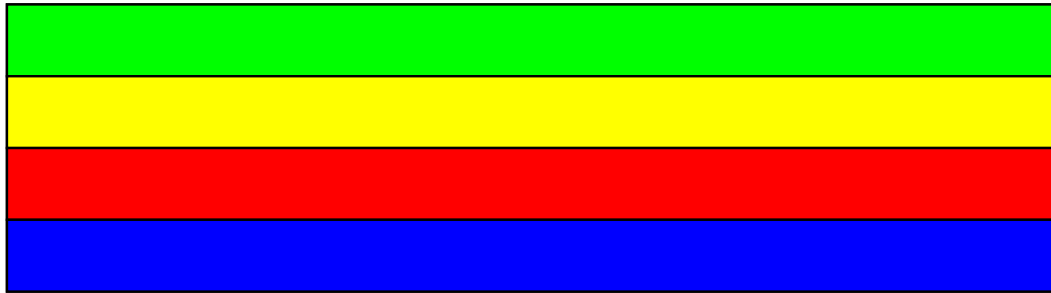
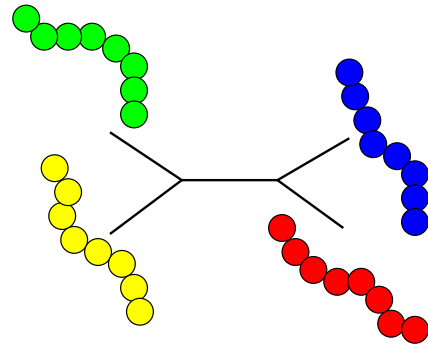
1997

Dennis Blakeslee

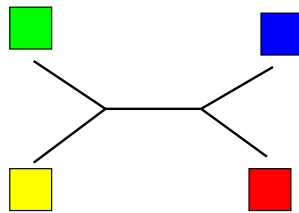
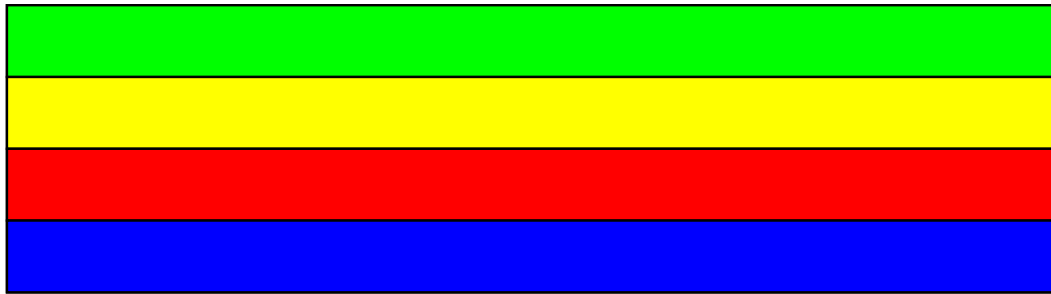
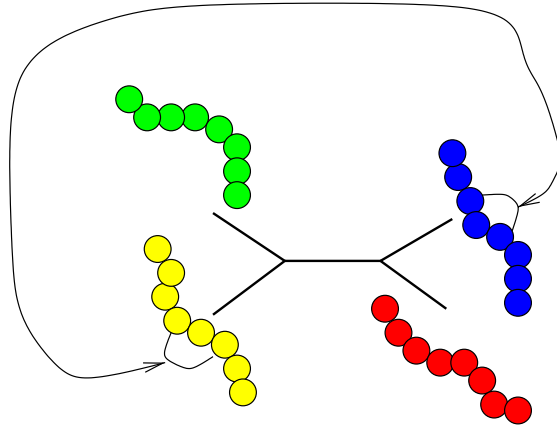
Recombination in HIV: A fast track to resistance?

<http://www.ama-assn.org/special/hiv/newsline/conferen/retrocon/recomb.htm>

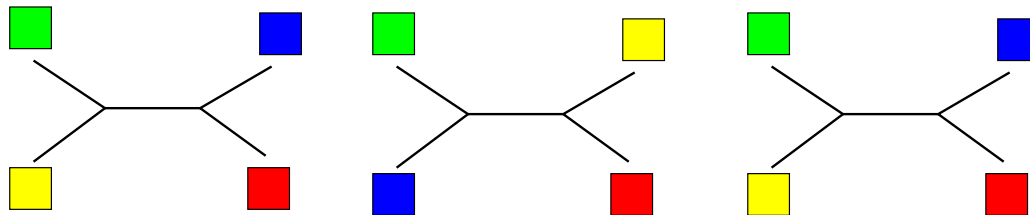
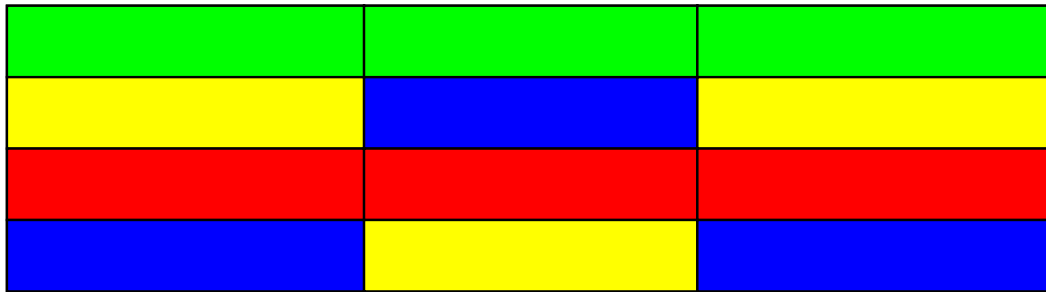
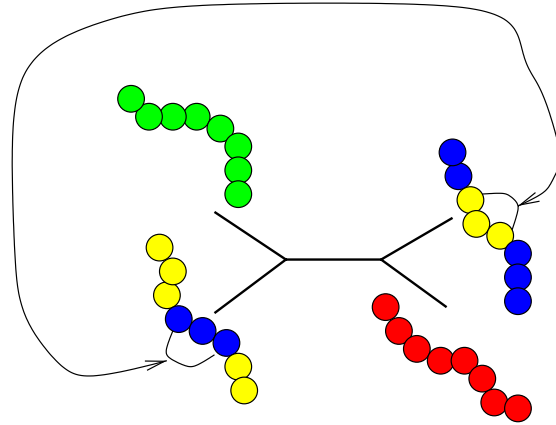
Recombination



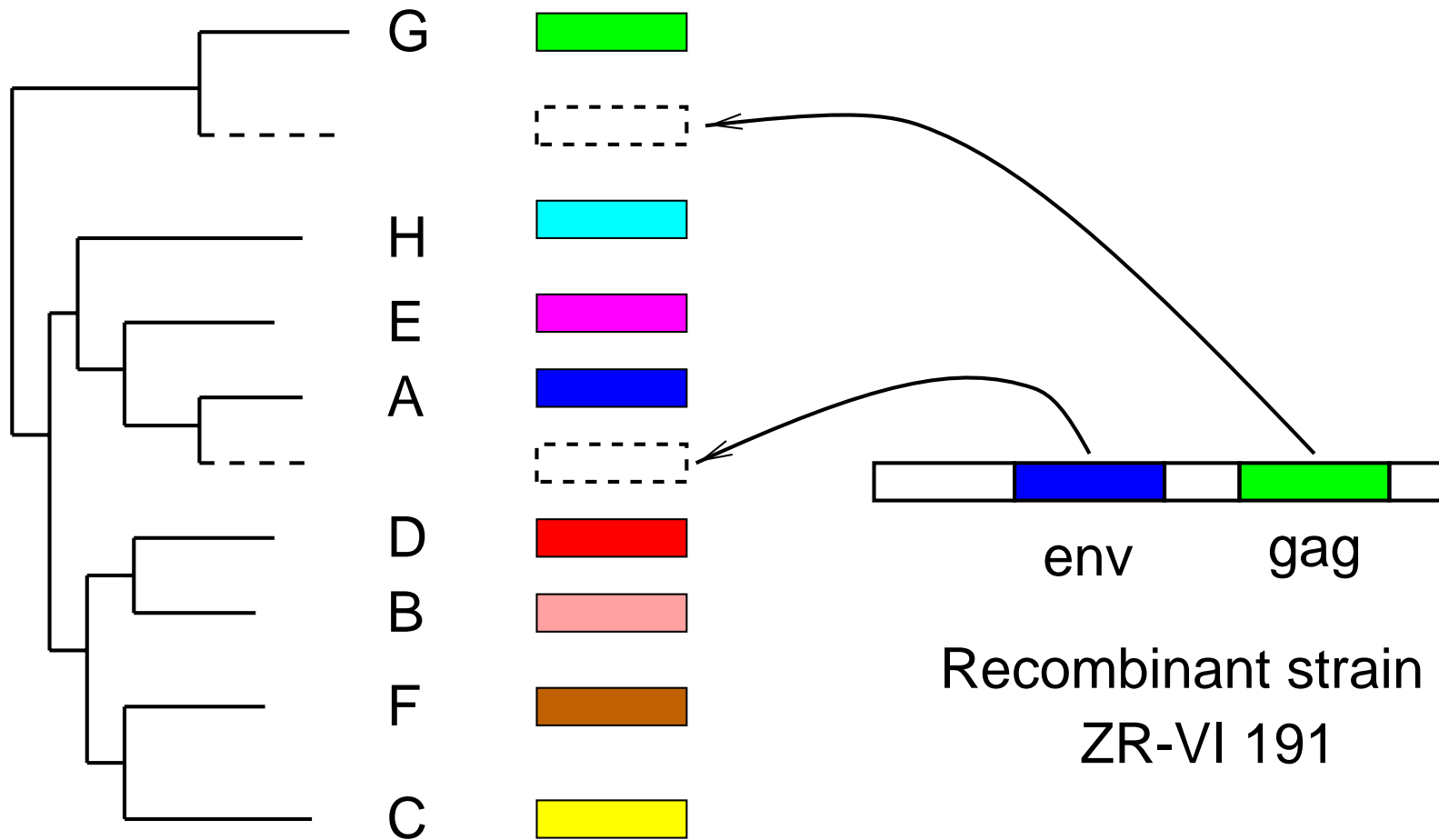
Recombination



Recombination



Recombination in HIV 1

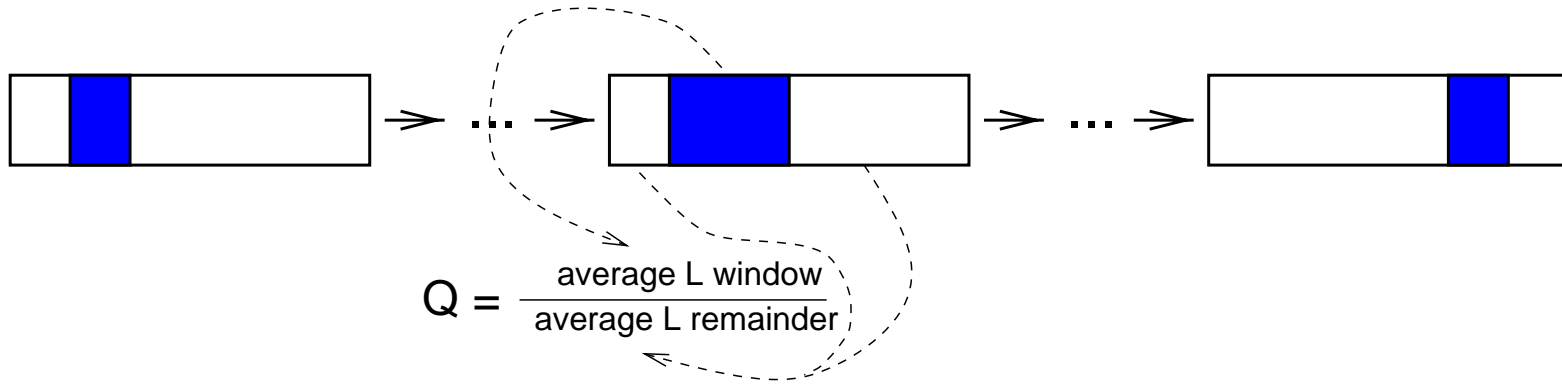


Various recombination detection methods

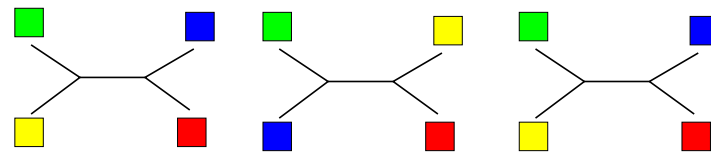
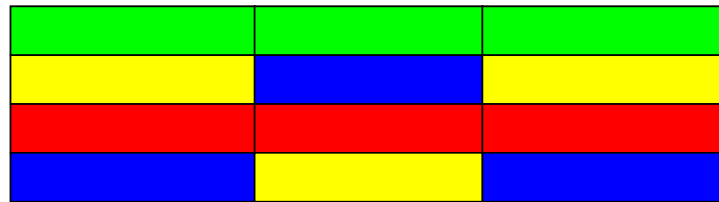
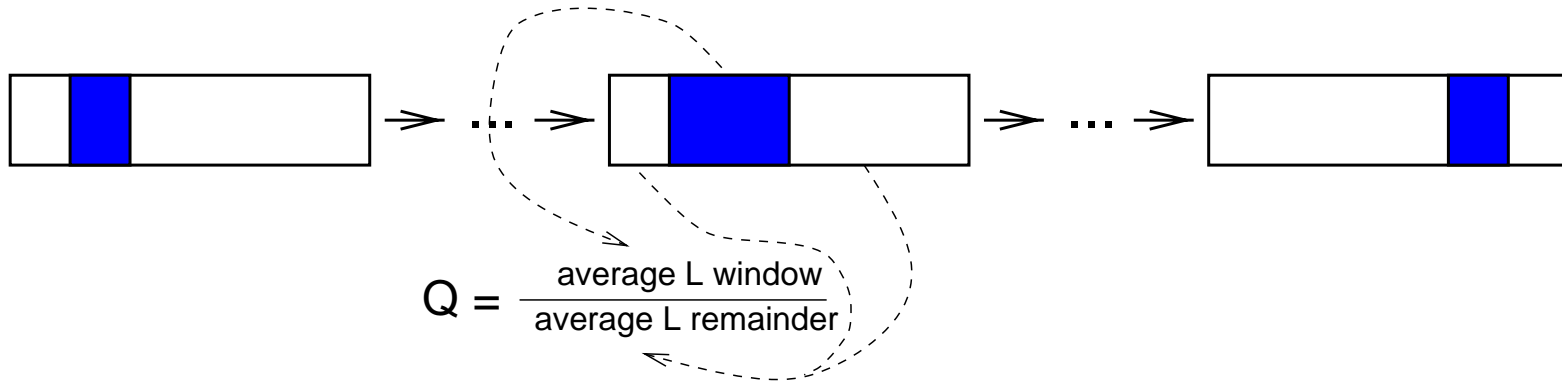
- PLATO
- Window methods
- RecPars
- Phylo-HMMs
- Phylo-FHMMs

-
- PLATO
 - Window methods
 - RecPars
 - Phylo-HMMs
 - Phylo-FHMMs

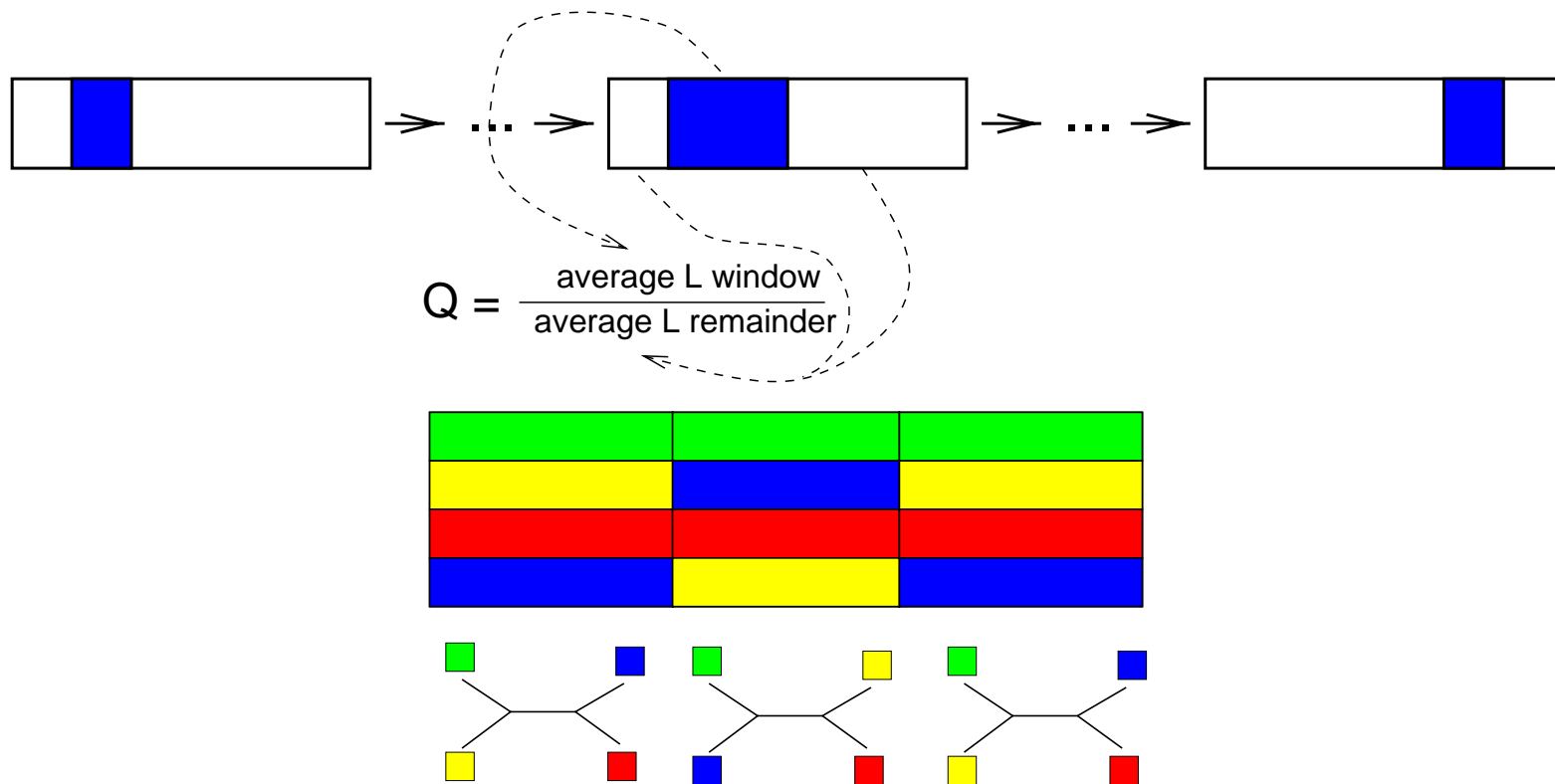
PLATO (Grassly & Holmes, 1997)



PLATO (Grassly & Holmes, 1997)

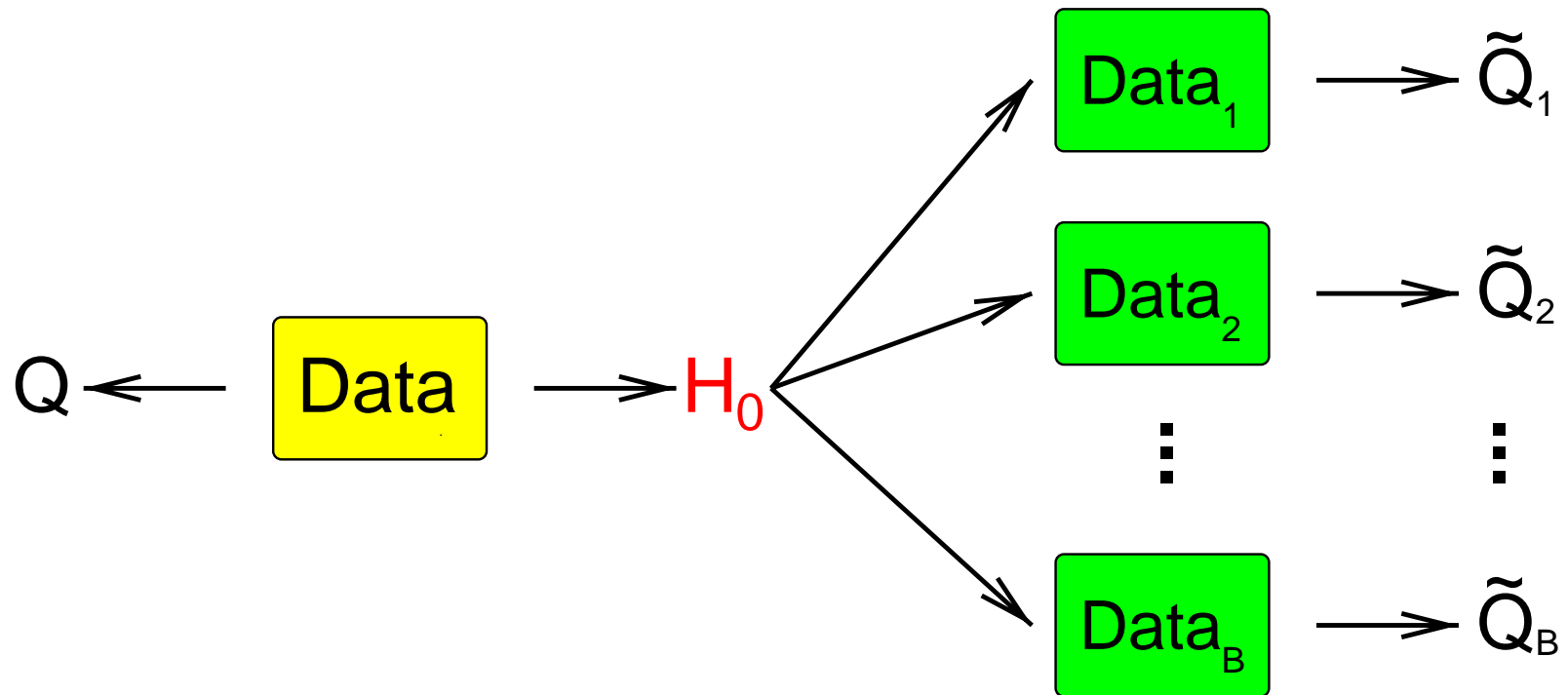


PLATO (Grassly & Holmes, 1997)

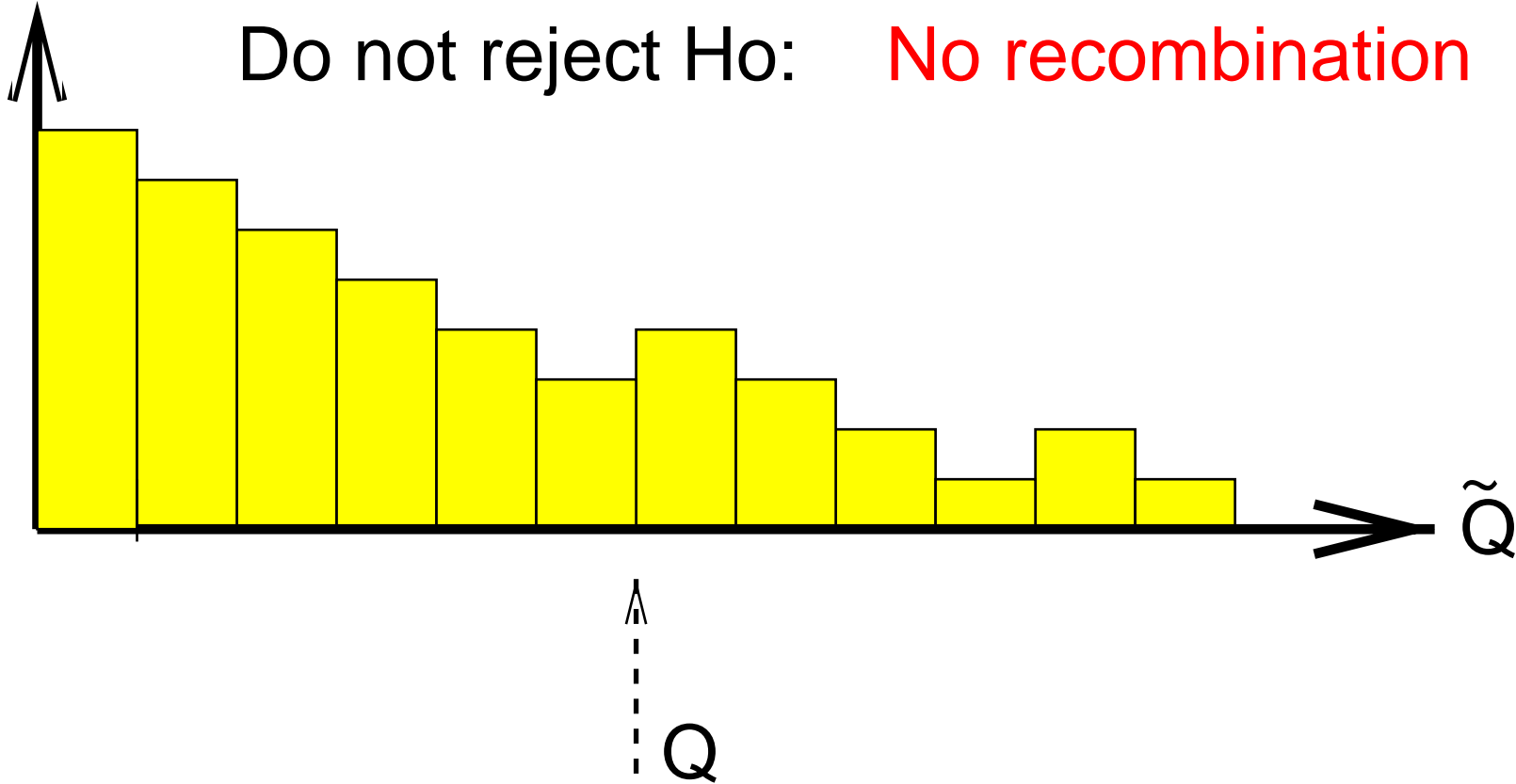


- Find regions with maximum Q values
- Test significance with parametric bootstrapping

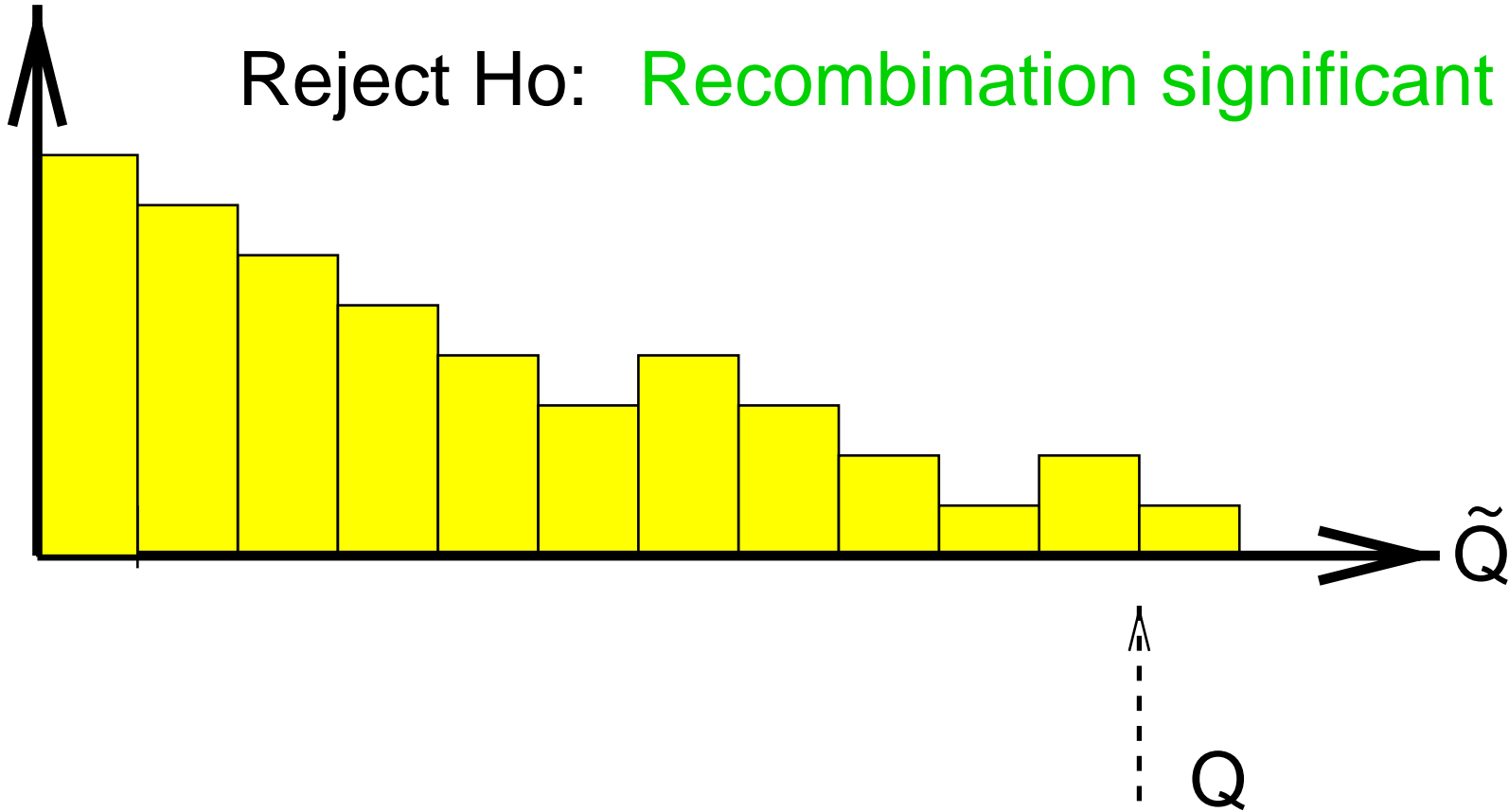
Parametric bootstrapping



Do not reject H_0 : No recombination



Reject H_0 : Recombination significant

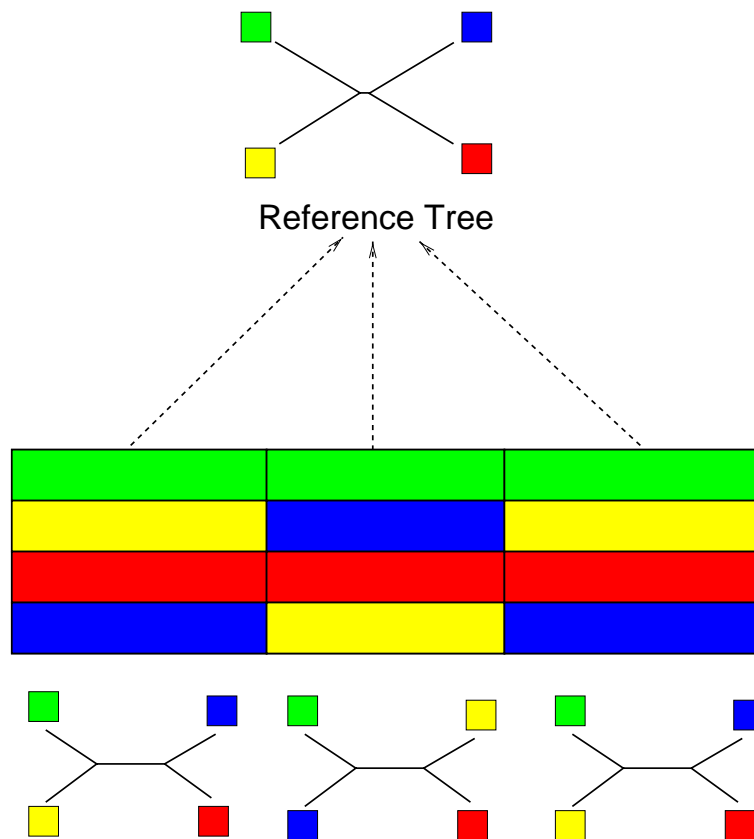


Shortcoming of PLATO

- Need a reference tree
- Obtained with global maximum likelihood

Shortcoming of PLATO

- Need a reference tree
- Obtained with global maximum likelihood

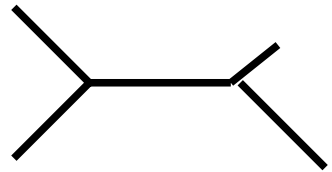
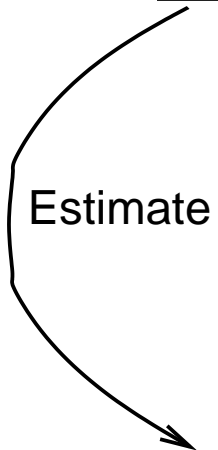
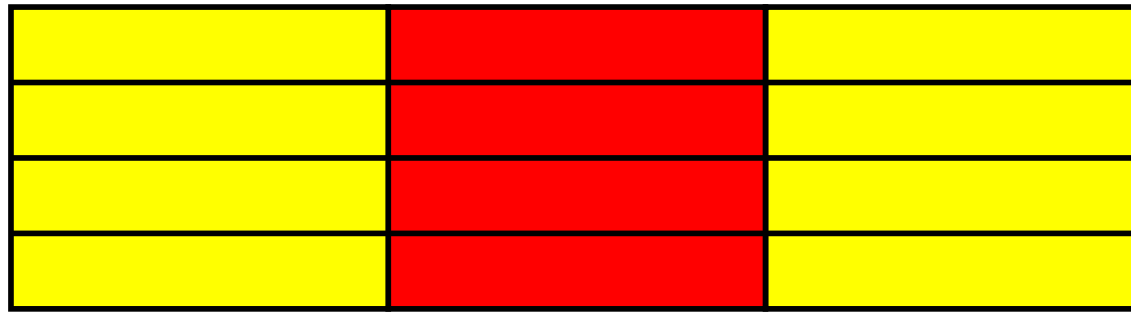


-
- PLATO
 - Window methods
 - RecPars
 - Phylo-HMMs
 - Phylo-FHMMs

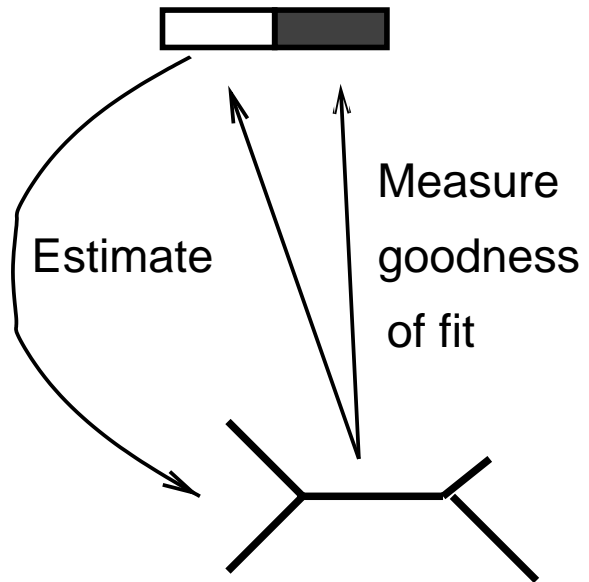
TOPAL (McGuire & Wright, 1997)



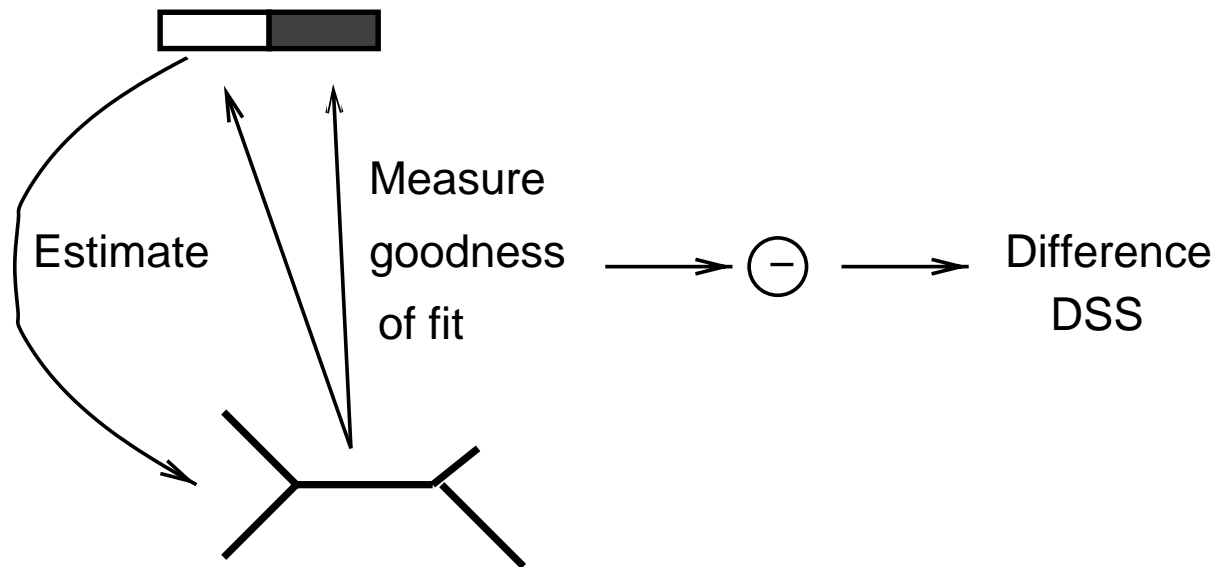
TOPAL (McGuire & Wright, 1997)



TOPAL (McGuire & Wright, 1997)



TOPAL (McGuire & Wright, 1997)



TOPAL (McGuire & Wright, 1997)



small

TOPAL (McGuire & Wright, 1997)



small



large

TOPAL (McGuire & Wright, 1997)



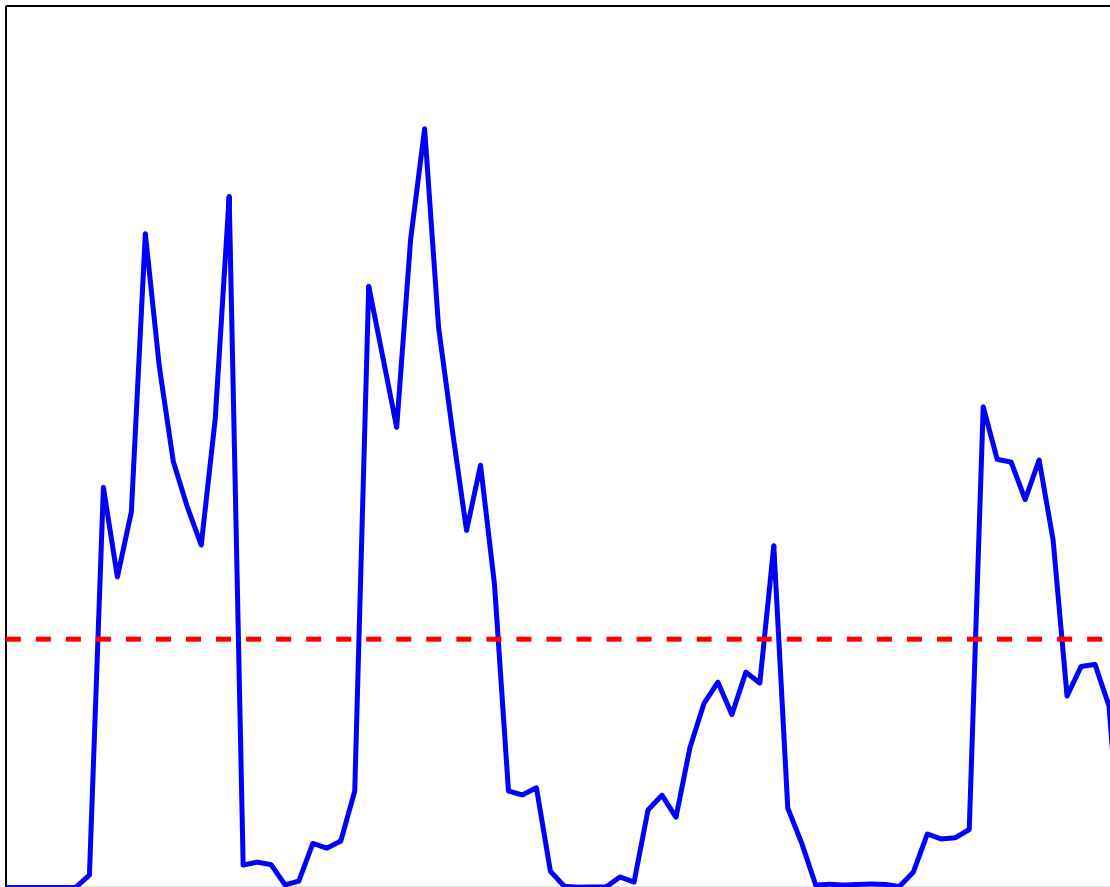
small



large

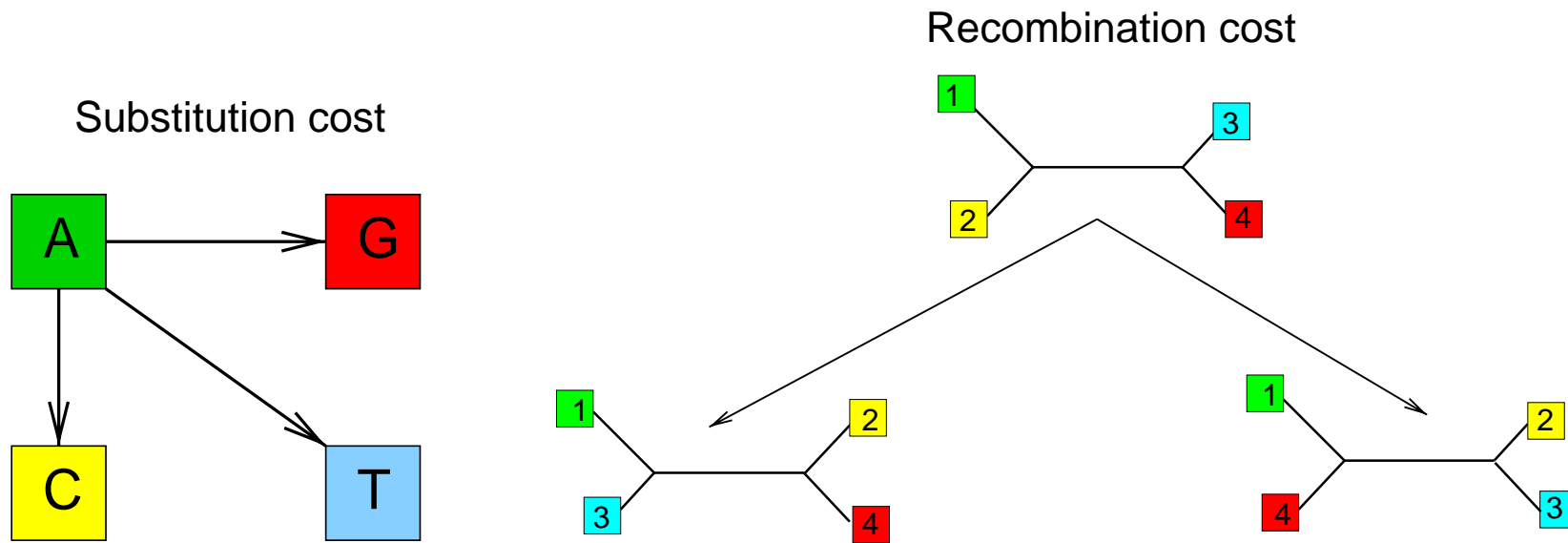
- Detect **significant peaks** of the DSS signal.
- Significance determined with **parametric bootstrapping**.

Example: TOPAL, window size=200



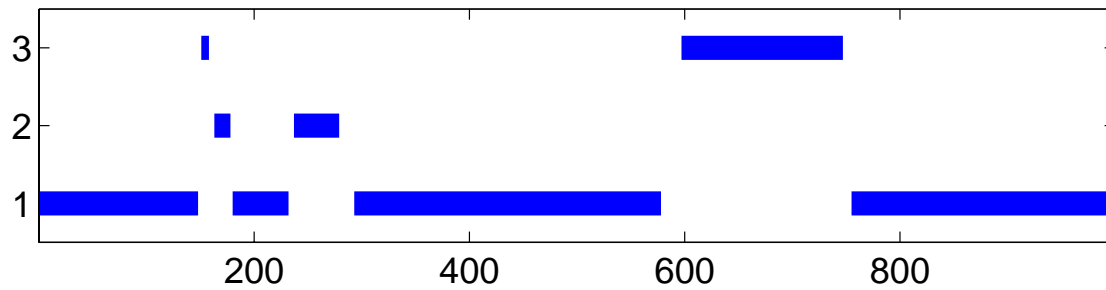
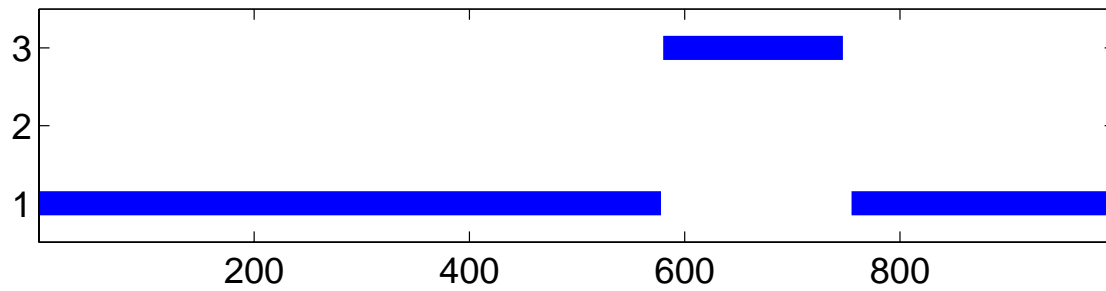
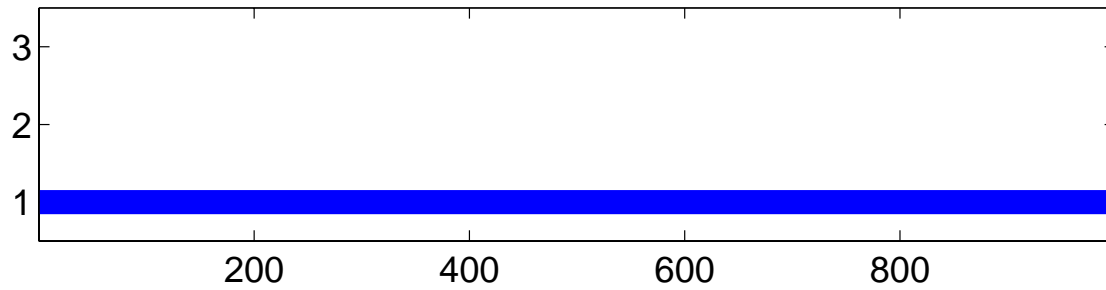
-
- Maximum χ^2
 - PLATO
 - Window methods
 - RecPars
 - Phylo-HMMs
 - Phylo-FHMMs

RecPars (Hein 1993)



$$\Psi = \text{Recombination cost} / \text{substitution cost}$$

Recombination cost / substitution cost = 10, 3, 1.5



-
- PLATO
 - Window methods
 - RecPars
 - **Phylo-HMMs**
 - Phylo-FHMMs

Detecting recombination with HMMs

- Husmeier, Wright (2001)
Journal of Computational Biology 8
- Husmeier, McGuire (2002)
Bioinformatics 18
- Husmeier, McGuire (2003)
Molecular Biology and Evolution 20

Related work

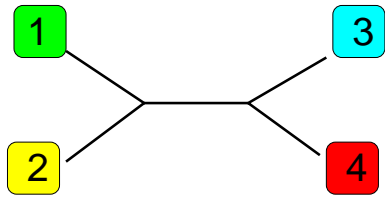
- Felsenstein & Churchill (1996)
Molecular Biology and Evolution 13
- Siepel & Haussler (2004)
Journal of Computational Biology 11

Hidden Markov models (HMMs)

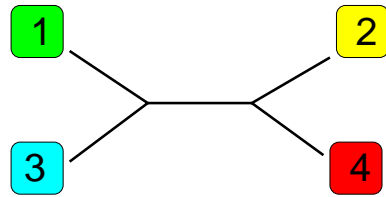
- Probabilistic equivalent to RecPars.
- All parameters can be inferred from the data.

- No window needed.
- More precise location of the breakpoints.

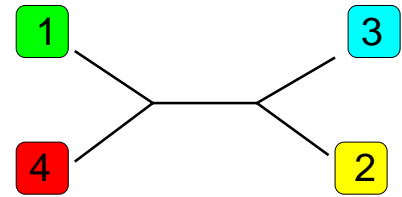
- Can currently only deal with a small number of species.



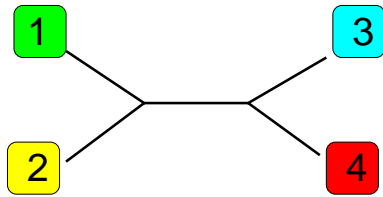
State 1



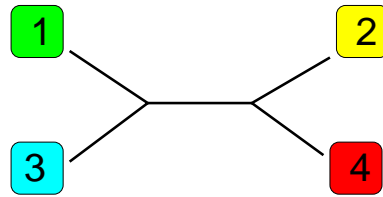
State 2



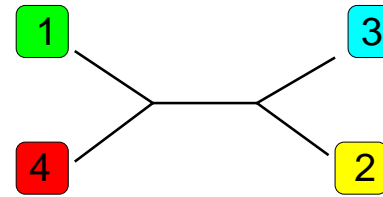
State 3



State 1



State 2



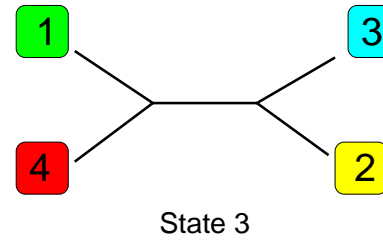
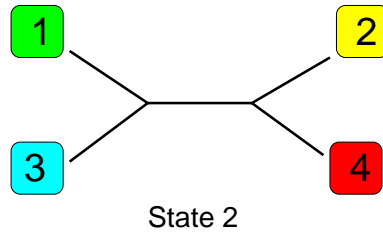
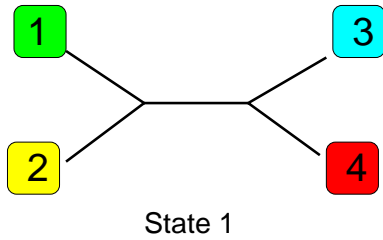
State 3

Naive approach

$$P(S_t | \mathbf{y}_t) = \frac{P(\mathbf{y}_t | S_t) P(S_t)}{P(\mathbf{y}_t)}$$

$$P(S_t) = \text{Const}$$

$$P(S_t | \mathbf{y}_t) = \frac{P(\mathbf{y}_t | S_t)}{\sum_{S'_t} P(\mathbf{y}_t | S'_t)}$$



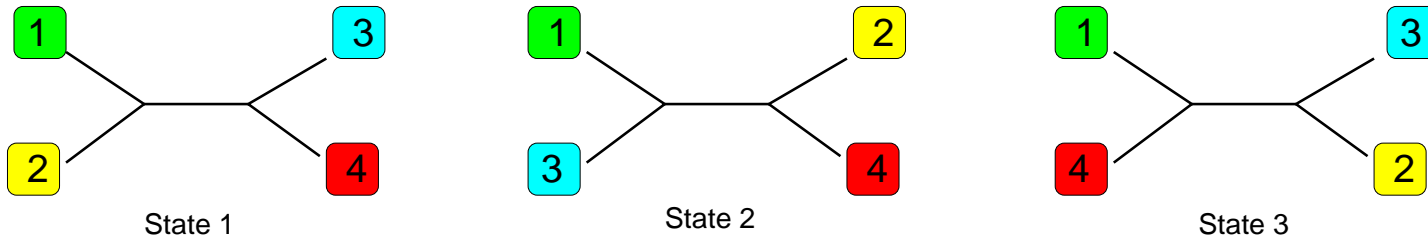
Naive approach

$$P(S_t | \mathbf{y}_t) = \frac{P(\mathbf{y}_t | S_t) P(S_t)}{P(\mathbf{y}_t)}$$

$$P(S_t) = \text{Const}$$

$$P(S_t | \mathbf{y}_t) = \frac{P(\mathbf{y}_t | S_t)}{\sum_{S'_t} P(\mathbf{y}_t | S'_t)}$$

Unrealistic prior on $\mathbf{S} = (S_1, \dots, S_N)$



Naive approach

$$P(S_t | \mathbf{y}_t) = \frac{P(\mathbf{y}_t | S_t) P(S_t)}{P(\mathbf{y}_t)}$$

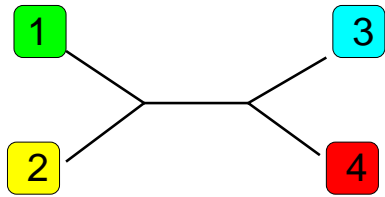
$$P(S_t) = \text{Const}$$

$$P(S_t | \mathbf{y}_t) = \frac{P(\mathbf{y}_t | S_t)}{\sum_{S'_t} P(\mathbf{y}_t | S'_t)}$$

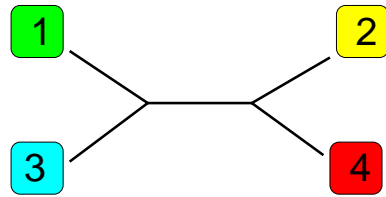
Unrealistic prior on $\mathbf{S} = (S_1, \dots, S_N)$

Introduce first-order **spatial correlations** via a **Markov model**

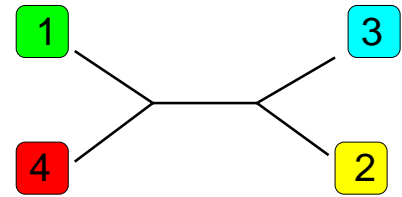
$$P(\mathbf{S}) = \prod_t P(S_t | S_{t-1}) P(S_1)$$



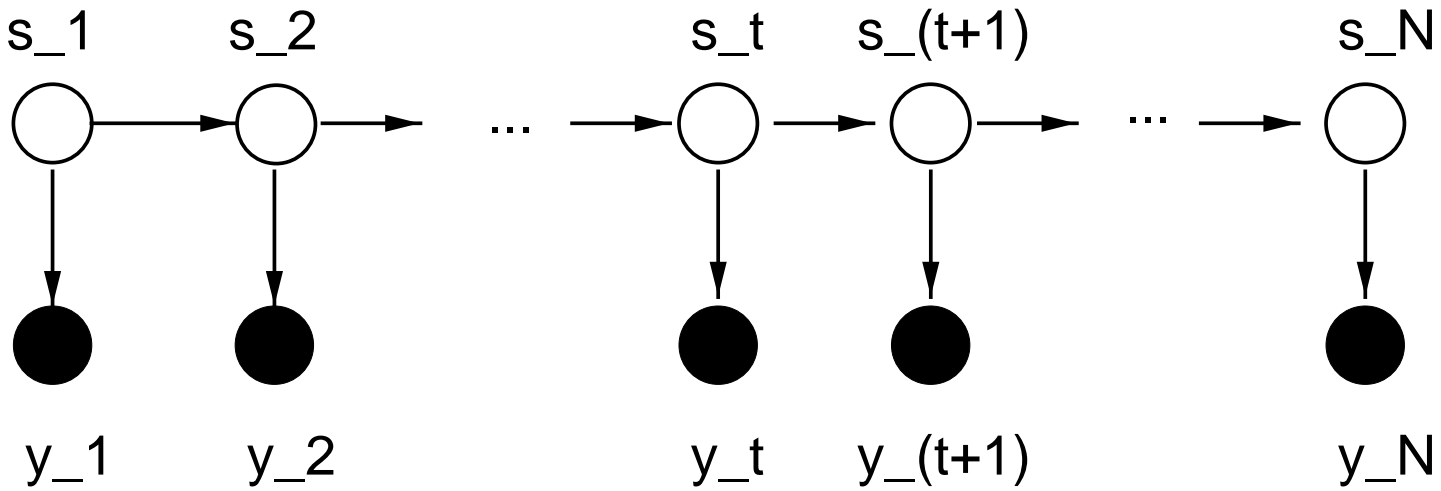
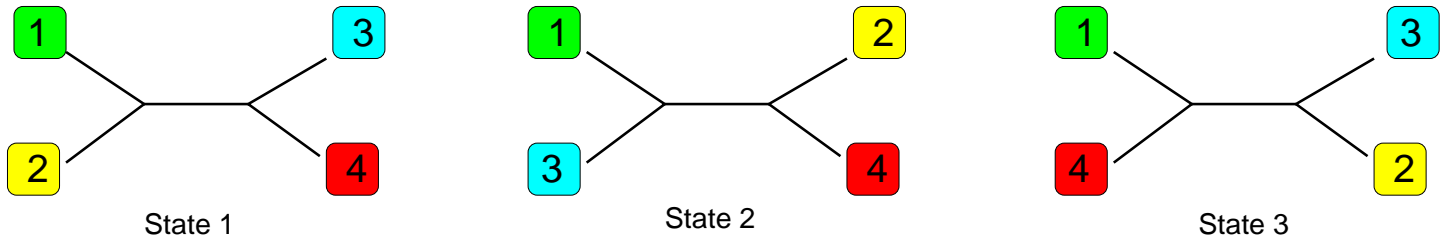
State 1



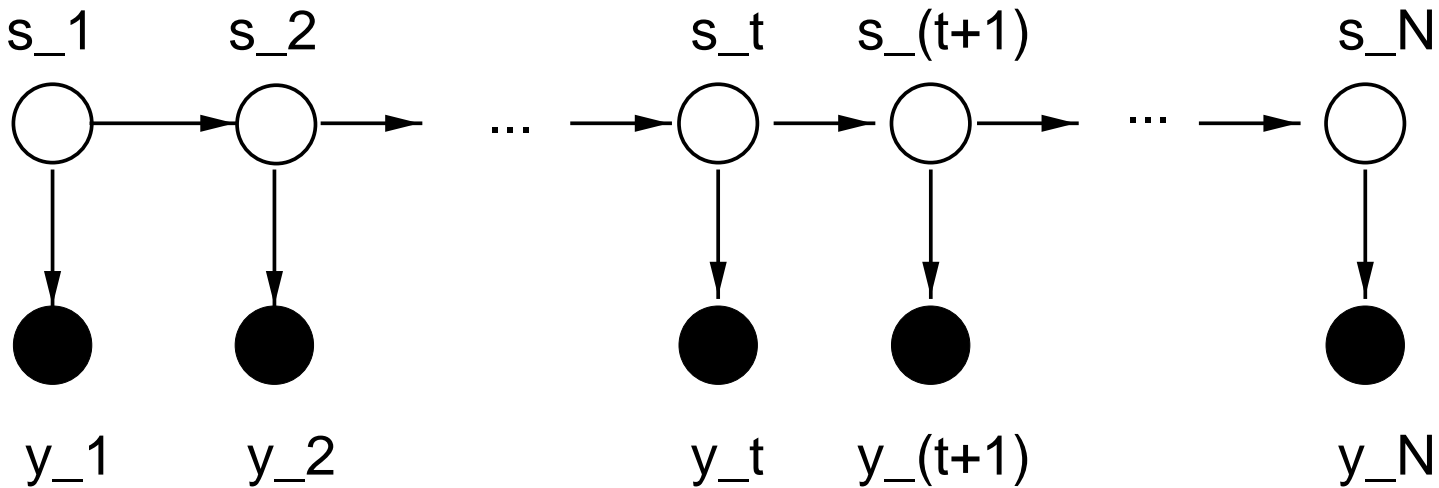
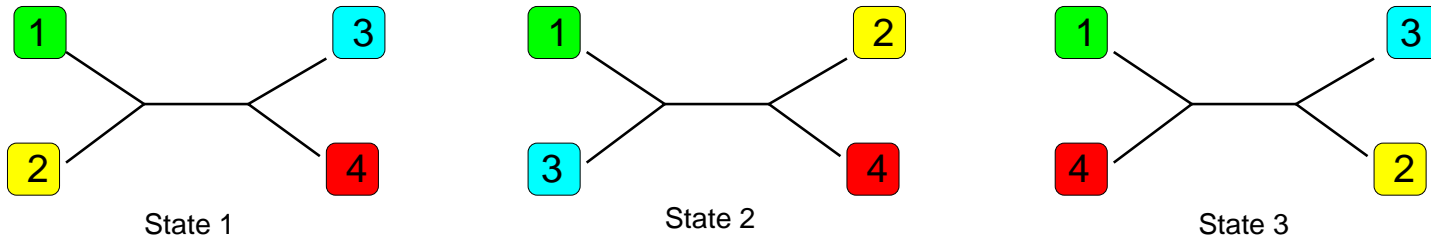
State 2



State 3



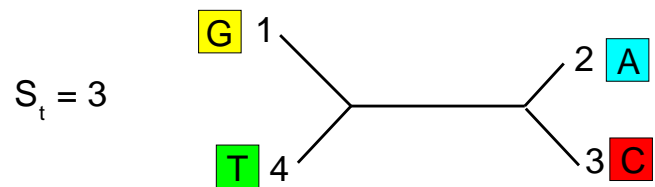
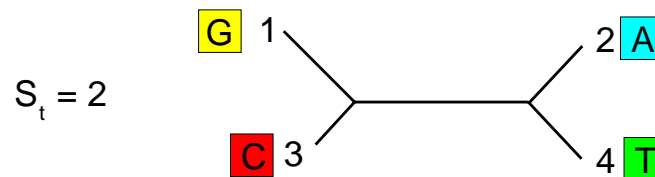
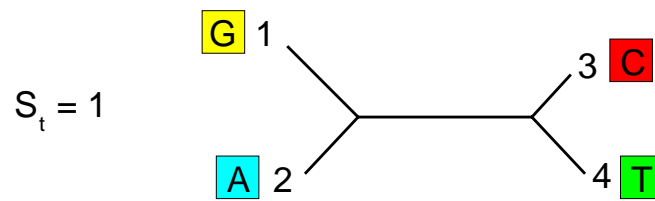
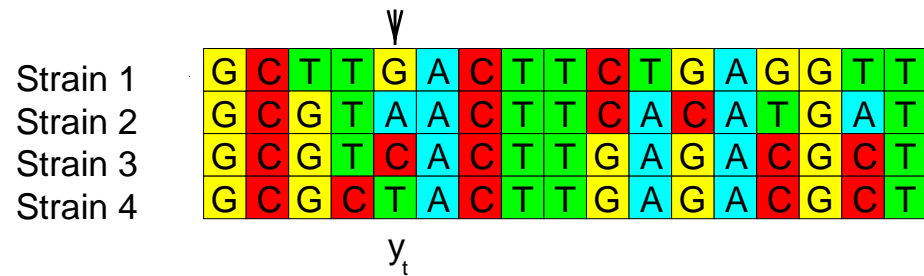
GCATCGTTCTATTTTACCGGCTCCCGA
 GTGTCGCTCAAGATTGCCATCGCGCGT
 GTCGTGGTCTAGATTGCCATCGCGCGT
 GTATCGCTCTAGTTTGCCAGCTCCCGT



GCATCGTTCTATTTTACCGGCTCCCGA
 GTGTCGCTCAAGATTGCCATCGCGCGT
 GTCGTGGTCTAGATTGCCATCGCGCGT
 GTATCGCTCTAGTTTGCCAGCTCCCGT

$$P(\mathcal{D}, \mathbf{S}) = \prod_{t=1}^N P(\mathbf{y}_t | S_t) \prod_{t=2}^N P(S_t | S_{t-1}) P(S_1)$$

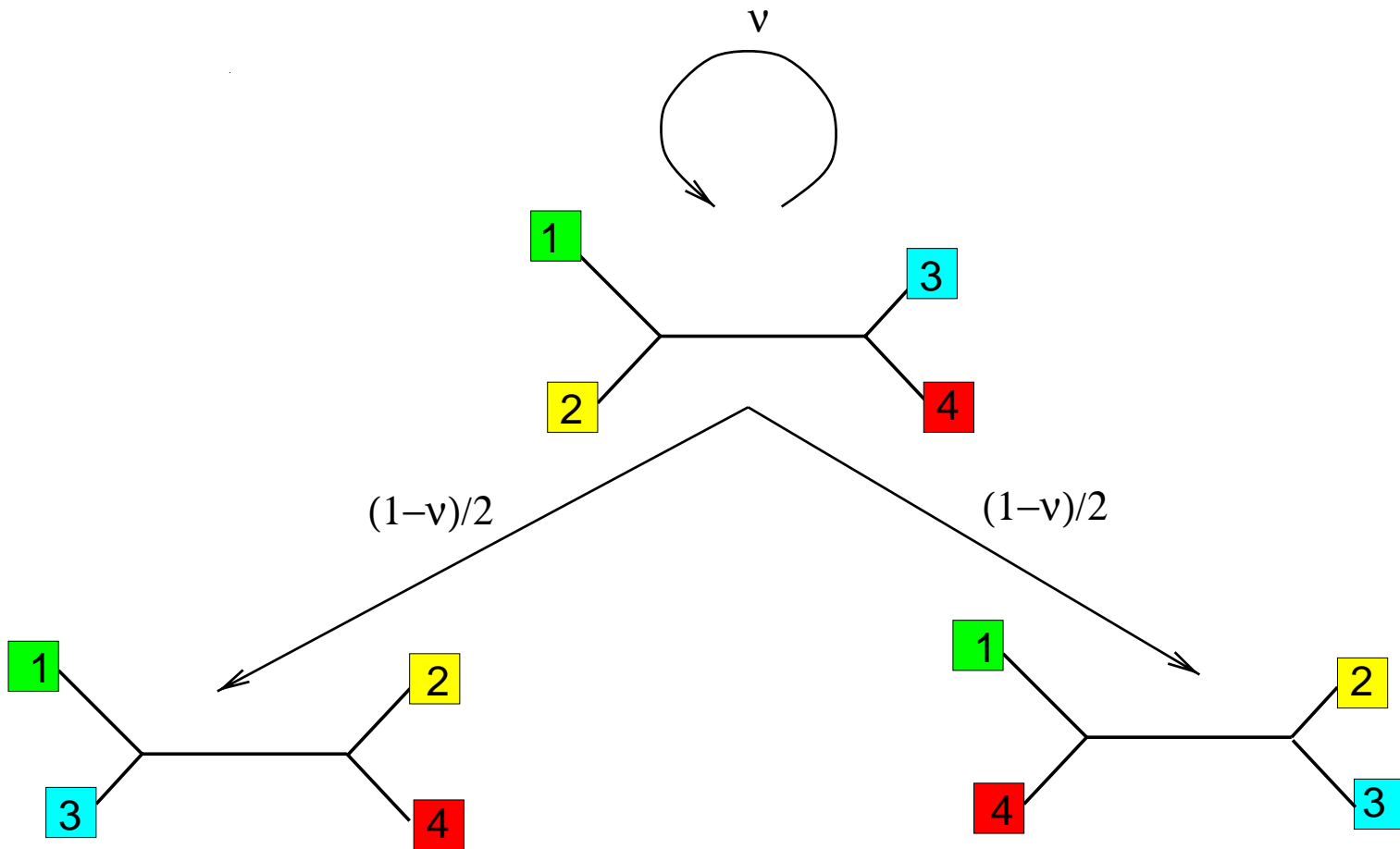
Emission probabilities (vertical arrows)



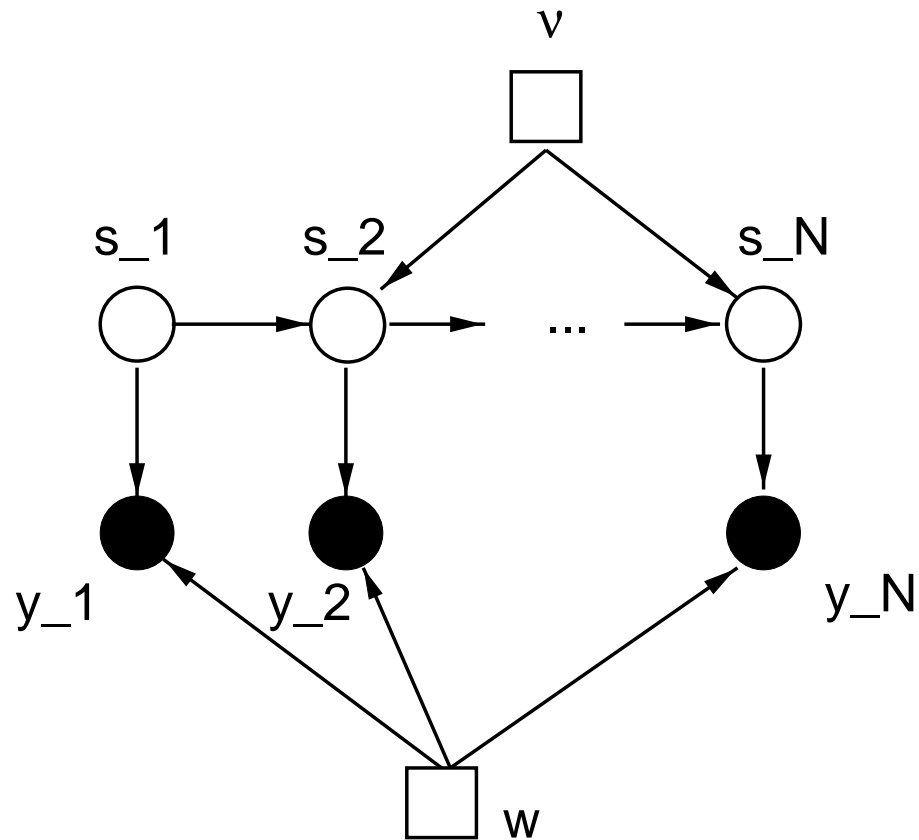
--> $P(y_t | S_t, w)$

Topology	S_t
Branch lengths	w

Transition probabilities (horizontal arrows)



HMM parameters



w \longrightarrow Vector of **branch lengths** of all the trees

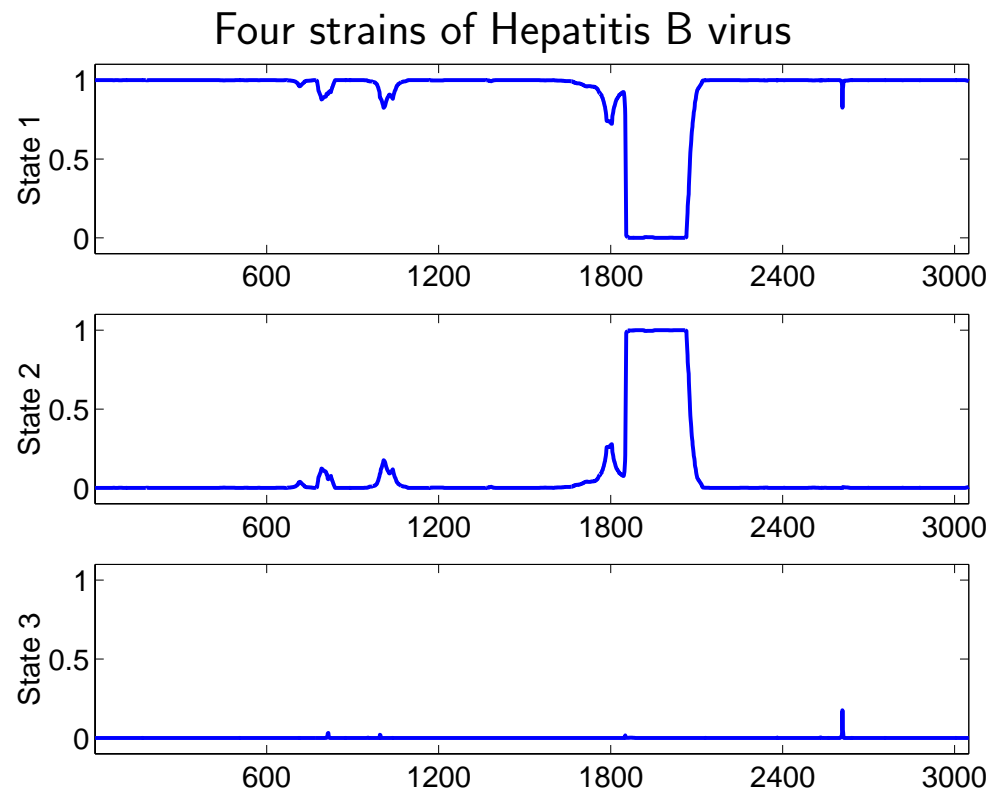
ν \longrightarrow Probability of *not* **changing** the tree **topology**

$$P(\mathbf{S}|\mathcal{D}) = P(S_1, S_2, \dots, S_N|\mathcal{D})$$

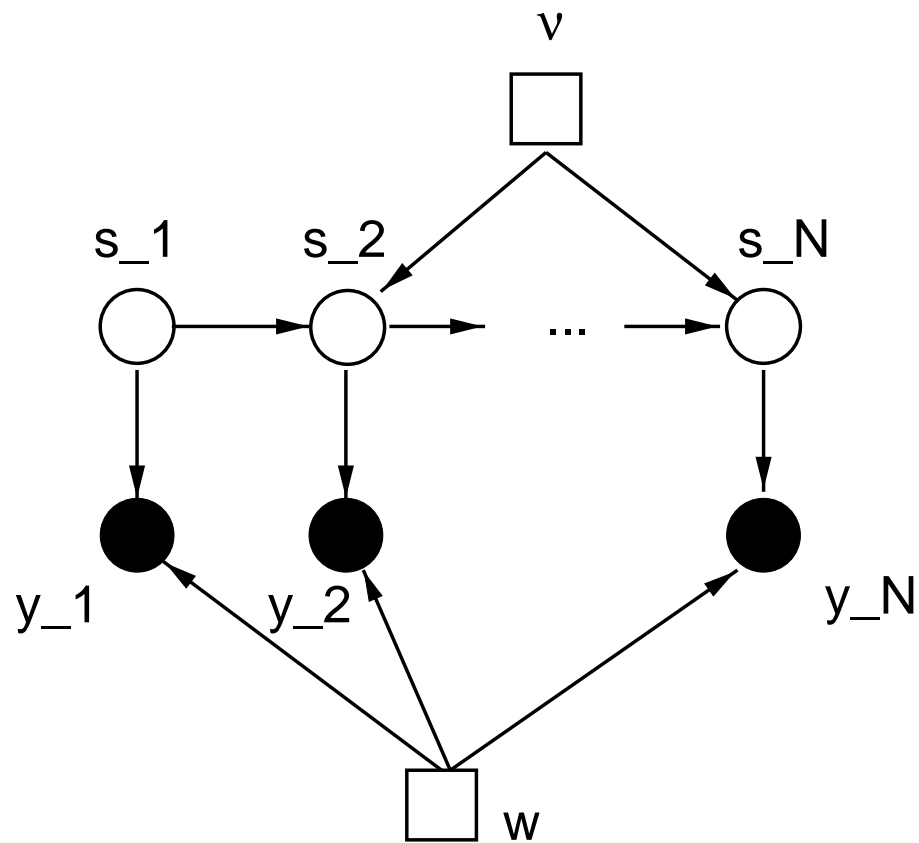
$$P(S_t|\mathcal{D}) = \sum_{S_1, \dots, S_{t-1}, S_{t+1}, \dots, S_N} P(\mathbf{S}|\mathcal{D})$$

$$P(\mathbf{S}|\mathcal{D}) = P(S_1, S_2, \dots, S_N|\mathcal{D})$$

$$P(S_t|\mathcal{D}) = \sum_{S_1, \dots, S_{t-1}, S_{t+1}, \dots, S_N} P(\mathbf{S}|\mathcal{D})$$



HMM parameters



- w \longrightarrow Vector of **branch lengths** of all the trees
- v \longrightarrow Probability of *not* **changing** the tree **topology**

Bayesian approach

Husmeier, McGuire (2002)

Bioinformatics 18, S345-S353

Husmeier, McGuire (2003)

Molecular Biology and Evolution 20, 315-337

Bayesian approach

$$P(\mathbf{S}|\mathcal{D}) = \int P(\mathbf{S}|\mathcal{D}, \mathbf{w}, \nu) P(\mathbf{w}, \nu|\mathcal{D}) d\mathbf{w} d\nu$$

Bayesian approach

$$P(\mathbf{S}|\mathcal{D}) = \int P(\mathbf{S}|\mathcal{D}, \mathbf{w}, \nu) P(\mathbf{w}, \nu|\mathcal{D}) d\mathbf{w} d\nu$$

Posterior $P(\mathbf{w}, \nu|\mathcal{D}) \longleftarrow$ Prior $P(\mathbf{w}, \nu) = \prod_i P(w_i)P(\nu)$

Bayesian approach

$$P(\mathbf{S}|\mathcal{D}) = \int P(\mathbf{S}|\mathcal{D}, \mathbf{w}, \nu) P(\mathbf{w}, \nu|\mathcal{D}) d\mathbf{w} d\nu$$

Posterior $P(\mathbf{w}, \nu|\mathcal{D}) \leftarrow$ Prior $P(\mathbf{w}, \nu) = \prod_i P(w_i)P(\nu)$

$$P(w_i) = \left[\begin{array}{l} 1/\Omega \text{ if } 0 \leq w_i \leq \Omega \\ 0 \text{ otherwise} \end{array} \right]$$

Bayesian approach

$$P(\mathbf{S}|\mathcal{D}) = \int P(\mathbf{S}|\mathcal{D}, \mathbf{w}, \nu) P(\mathbf{w}, \nu|\mathcal{D}) d\mathbf{w} d\nu$$

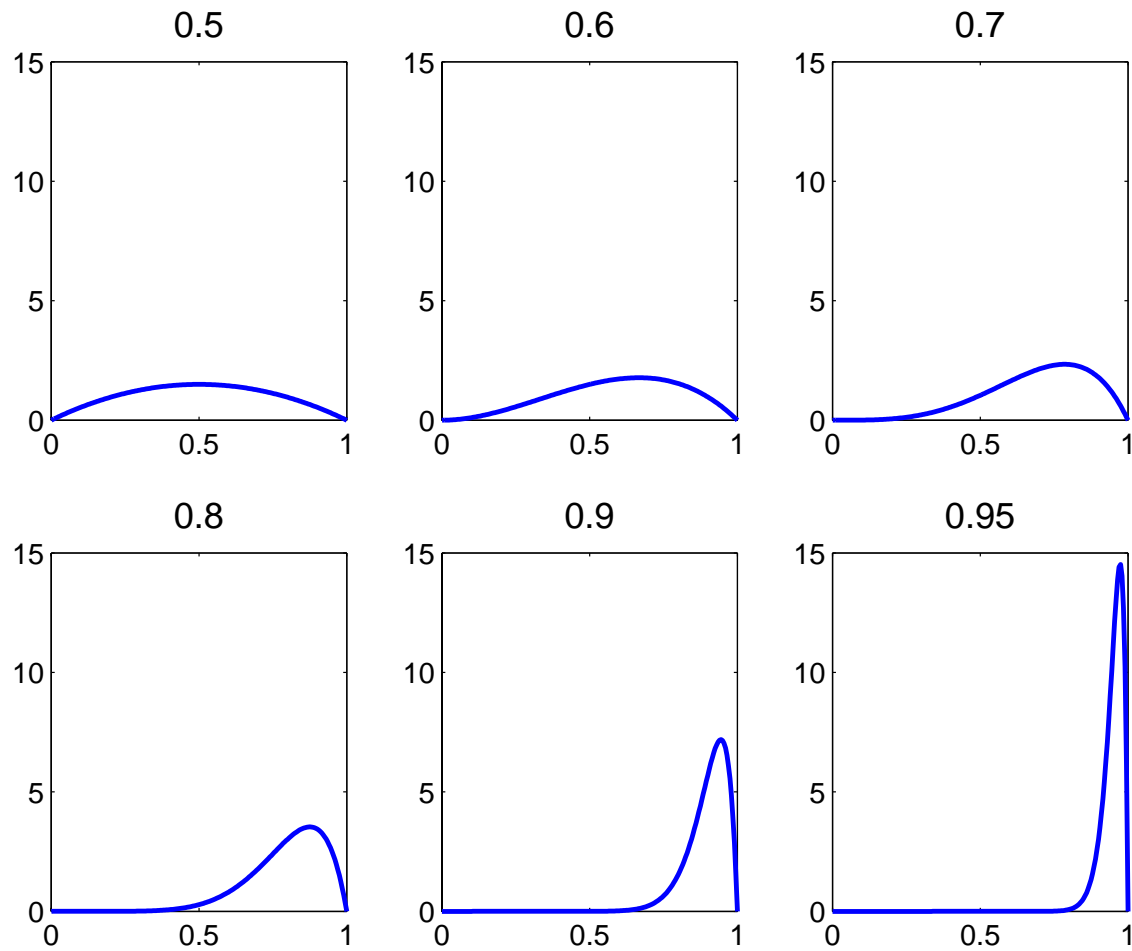
Posterior $P(\mathbf{w}, \nu|\mathcal{D}) \leftarrow$ Prior $P(\mathbf{w}, \nu) = \prod_i P(w_i)P(\nu)$

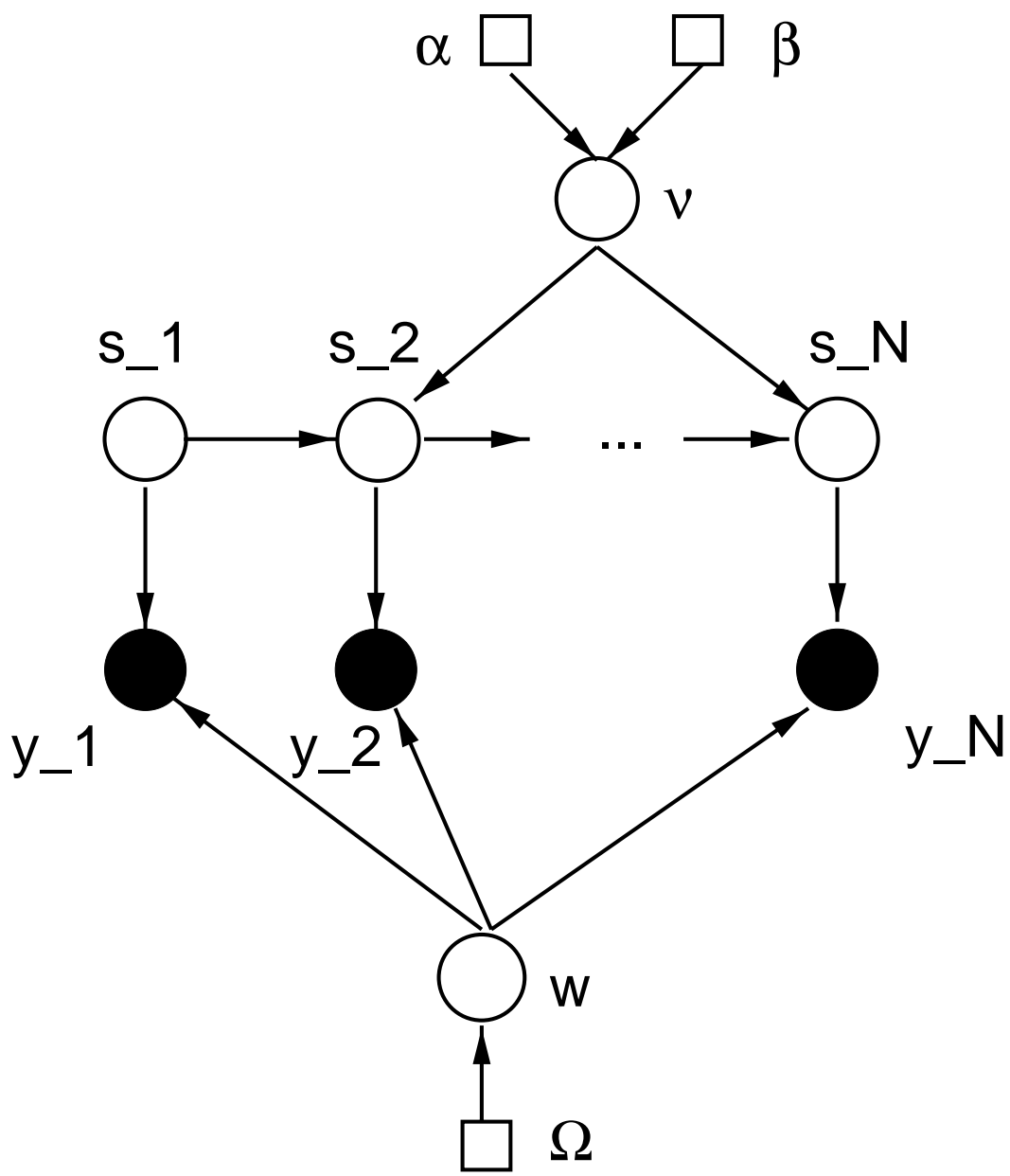
$$P(w_i) = \left[\begin{array}{l} 1/\Omega \text{ if } 0 \leq w_i \leq \Omega \\ 0 \text{ otherwise} \end{array} \right]$$

$$P(\nu) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \nu^{\alpha-1} (1-\nu)^{\beta-1}$$

Conjugate prior: Beta distribution.

Beta Prior, $\beta = 2$, $\mu = \alpha / (\alpha + \beta)$





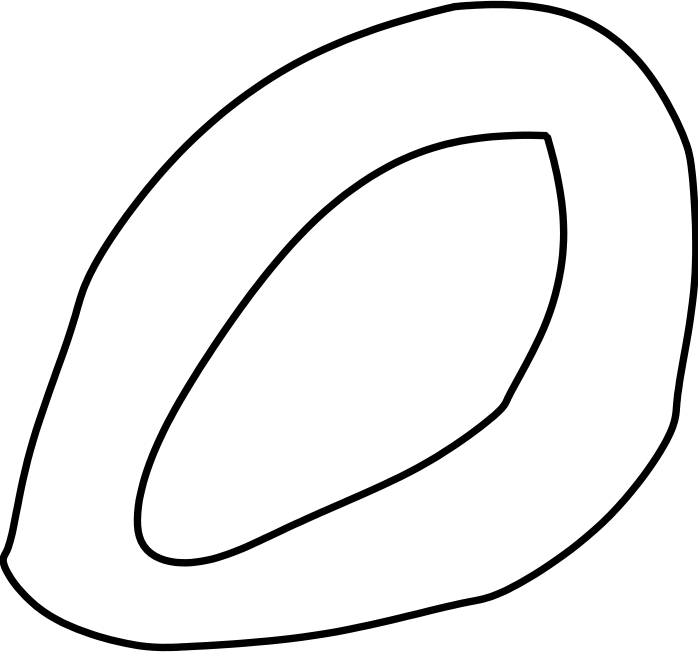
Sampling from the posterior distribution

- Sampling from $P(\mathbf{S}, \mathbf{w}, \nu | \mathcal{D})$

Sampling from the posterior distribution

- Sampling from $P(\mathbf{S}, \mathbf{w}, \nu | \mathcal{D})$
- Gibbs sampling

y

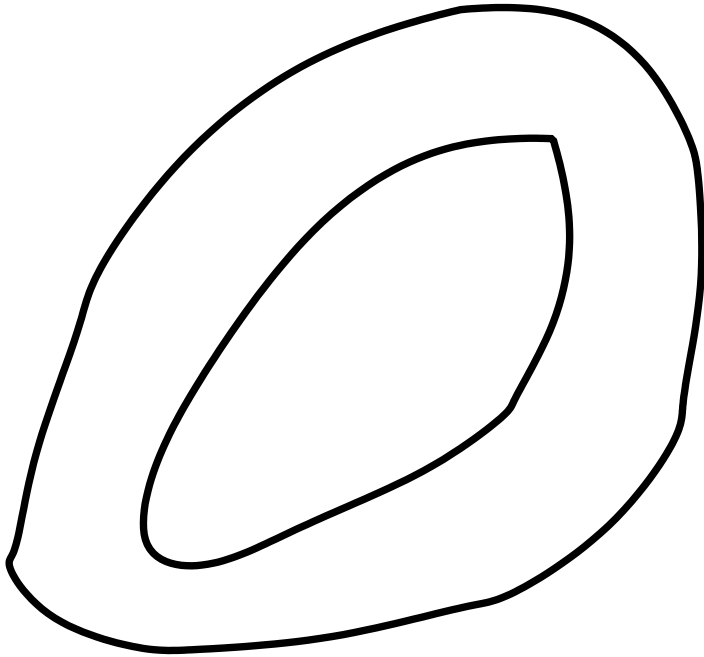


$P(x,y)$



x

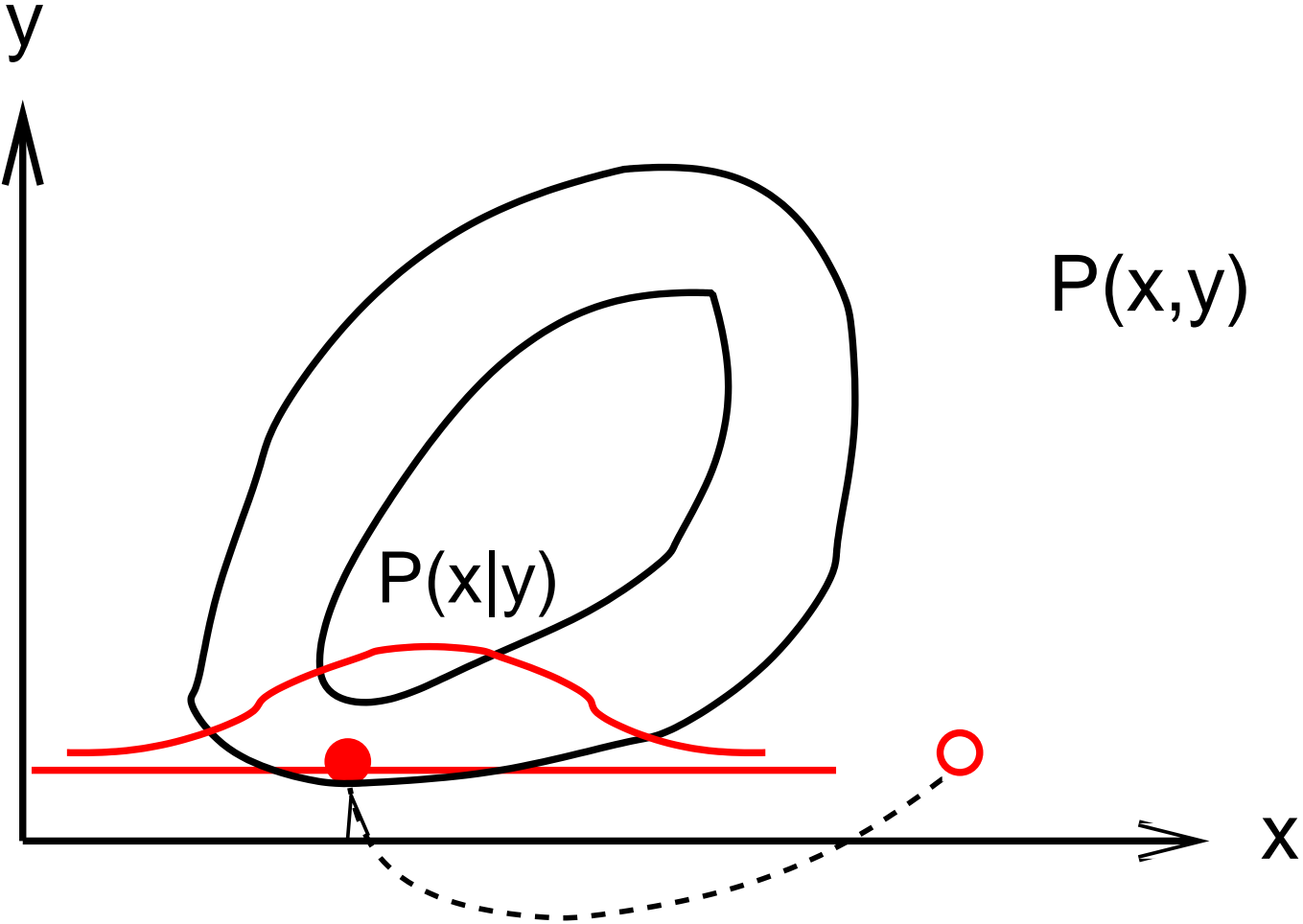
y

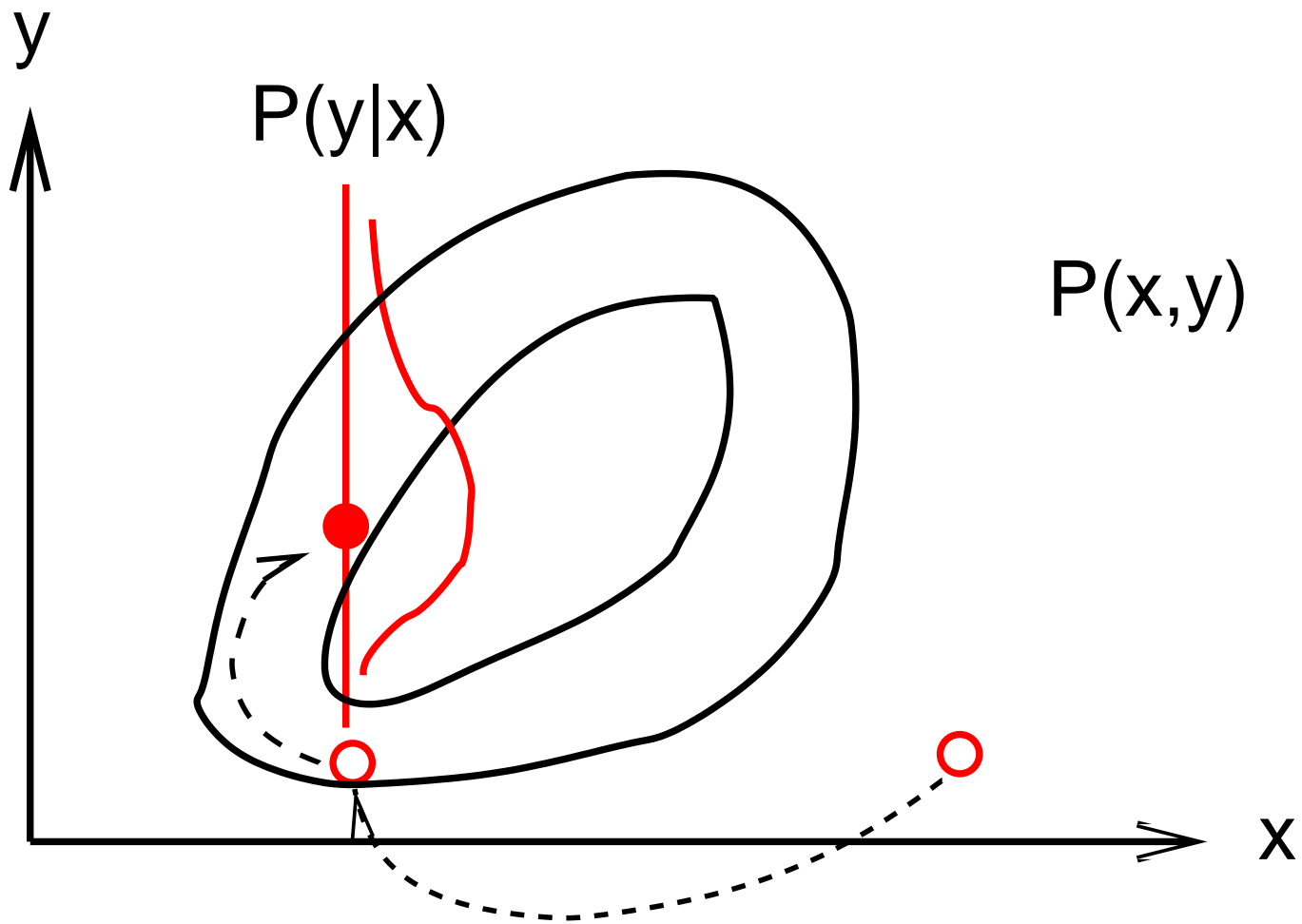


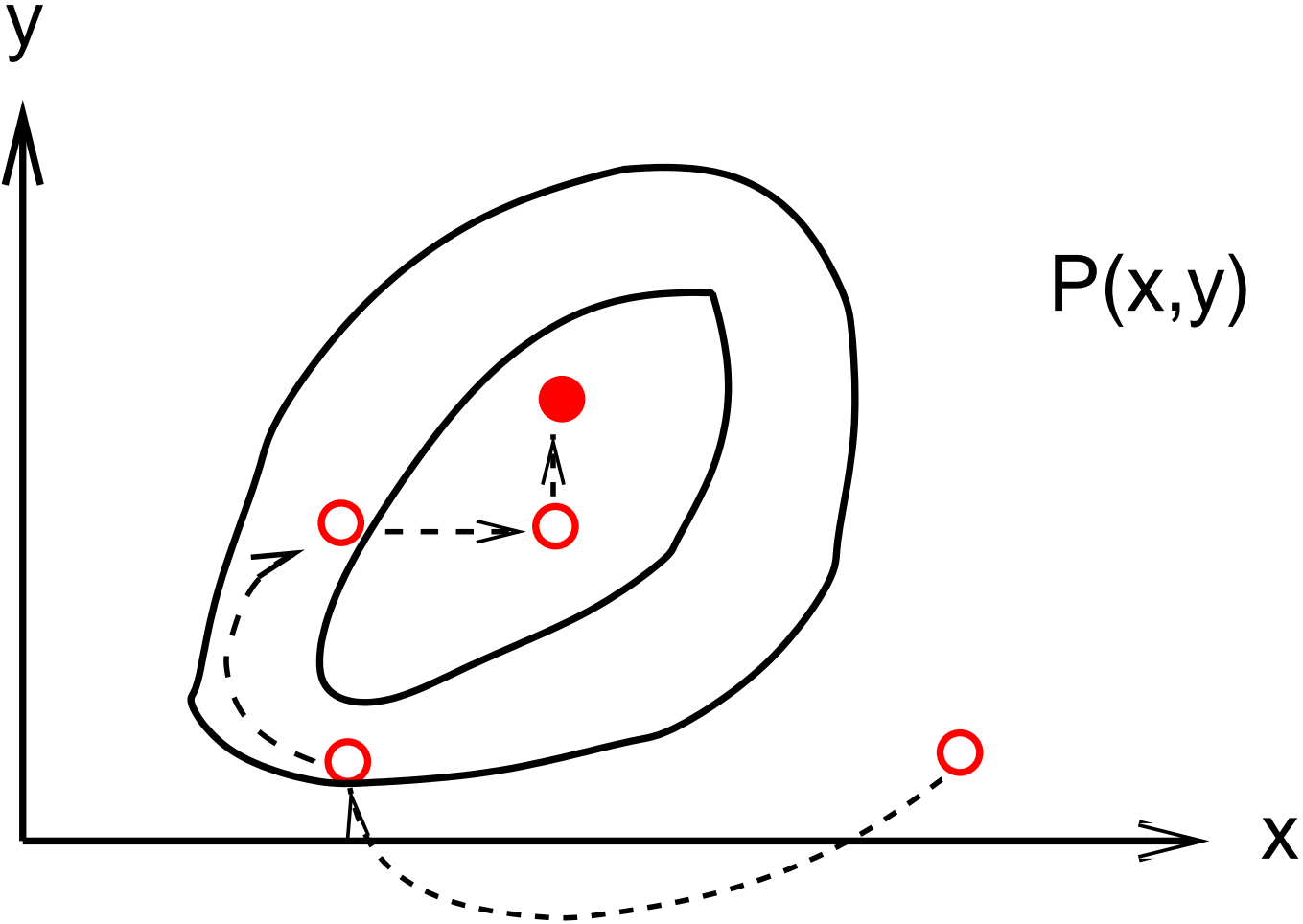
$P(x,y)$



x







Sampling from the posterior distribution

- Sampling from $P(\mathbf{S}, \mathbf{w}, \nu | \mathcal{D})$
- Gibbs-like approach:

Sampling from the posterior distribution

- Sampling from $P(\mathbf{S}, \mathbf{w}, \nu | \mathcal{D})$
- Gibbs-like approach:
 - $\mathbf{S} \sim P(\mathbf{S} | \mathbf{w}, \nu, \mathcal{D})$

Sampling from the posterior distribution

- Sampling from $P(\mathbf{S}, \mathbf{w}, \nu | \mathcal{D})$
- Gibbs-like approach:
 - $\mathbf{S} \sim P(\mathbf{S} | \mathbf{w}, \nu, \mathcal{D})$
 - $\mathbf{w} \sim P(\mathbf{w} | \mathbf{S}, \nu, \mathcal{D})$

Sampling from the posterior distribution

- Sampling from $P(\mathbf{S}, \mathbf{w}, \nu | \mathcal{D})$
- Gibbs-like approach:
 - $\mathbf{S} \sim P(\mathbf{S} | \mathbf{w}, \nu, \mathcal{D})$
 - $\mathbf{w} \sim P(\mathbf{w} | \mathbf{S}, \nu, \mathcal{D})$
 - $\nu \sim P(\nu | \mathbf{S}, \mathbf{w}, \mathcal{D})$

Sampling from the posterior distribution

- Sampling from $P(\mathbf{S}, \mathbf{w}, \nu | \mathcal{D})$
- Gibbs-like approach:
 - $\mathbf{S} \sim P(\mathbf{S} | \mathbf{w}, \nu, \mathcal{D})$
 - $\mathbf{w} \sim P(\mathbf{w} | \mathbf{S}, \nu, \mathcal{D})$
 - $\nu \sim P(\nu | \mathbf{S}, \mathbf{w}, \mathcal{D})$
- ν : Sample from Beta distribution

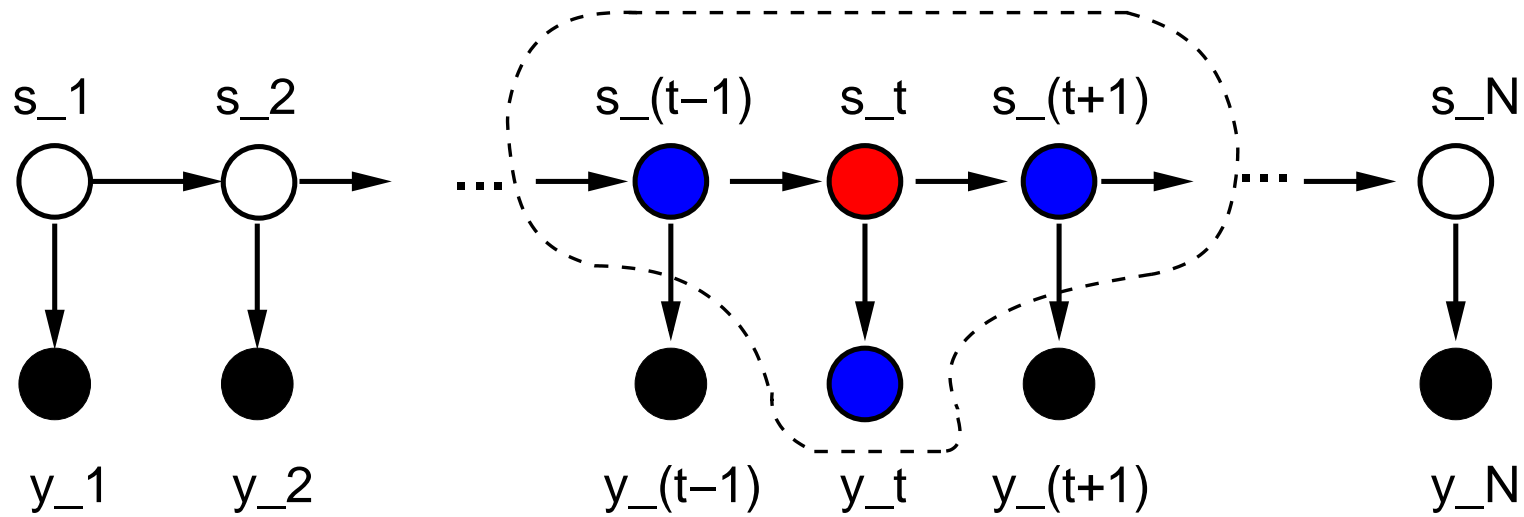
Sampling from the posterior distribution

- Sampling from $P(\mathbf{S}, \mathbf{w}, \nu | \mathcal{D})$
- Gibbs-like approach:
 - $\mathbf{S} \sim P(\mathbf{S} | \mathbf{w}, \nu, \mathcal{D})$
 - $\mathbf{w} \sim P(\mathbf{w} | \mathbf{S}, \nu, \mathcal{D})$
 - $\nu \sim P(\nu | \mathbf{S}, \mathbf{w}, \mathcal{D})$
- ν : Sample from Beta distribution
- \mathbf{w} : Metropolis-Hastings

Sampling from the posterior distribution

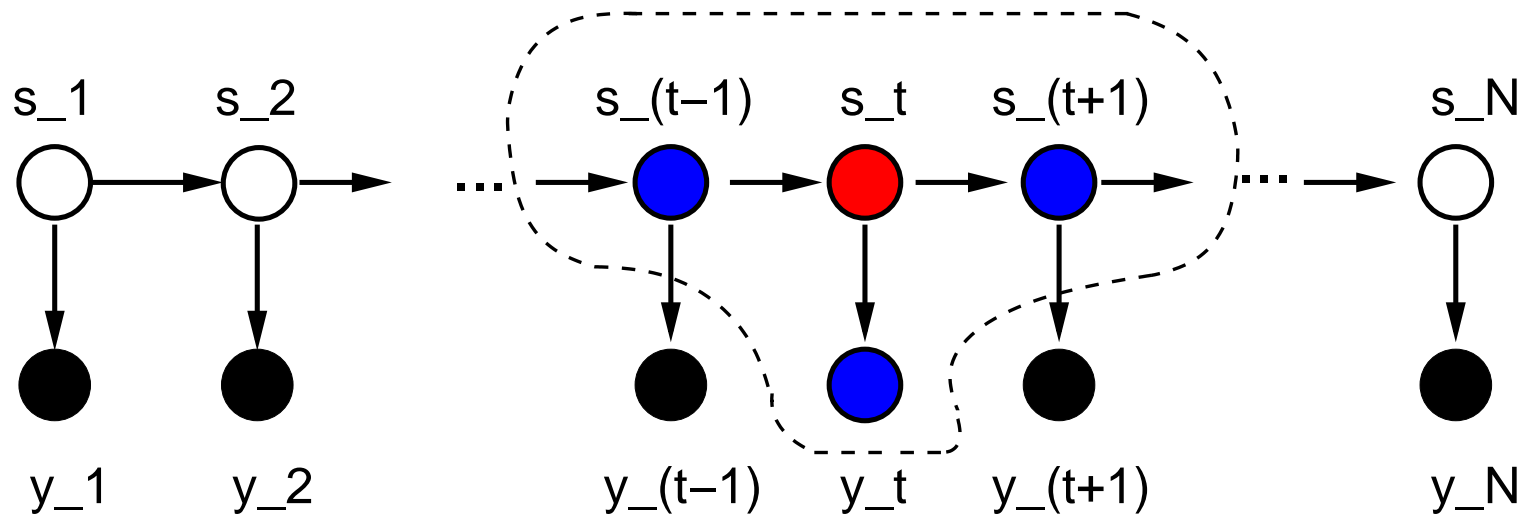
- Sampling from $P(\mathbf{S}, \mathbf{w}, \nu | \mathcal{D})$
- Gibbs-like approach:
 - $\mathbf{S} \sim P(\mathbf{S} | \mathbf{w}, \nu, \mathcal{D})$
 - $\mathbf{w} \sim P(\mathbf{w} | \mathbf{S}, \nu, \mathcal{D})$
 - $\nu \sim P(\nu | \mathbf{S}, \mathbf{w}, \mathcal{D})$
- ν : Sample from Beta distribution
- \mathbf{w} : Metropolis-Hastings
- \mathbf{S} : Gibbs sampling
 - $S_t \sim P(S_t | S_1, \dots, S_{t-1}, S_{t+1}, \dots, S_N, \mathcal{D}, \mathbf{w}, \nu)$

Sampling from the posterior distribution



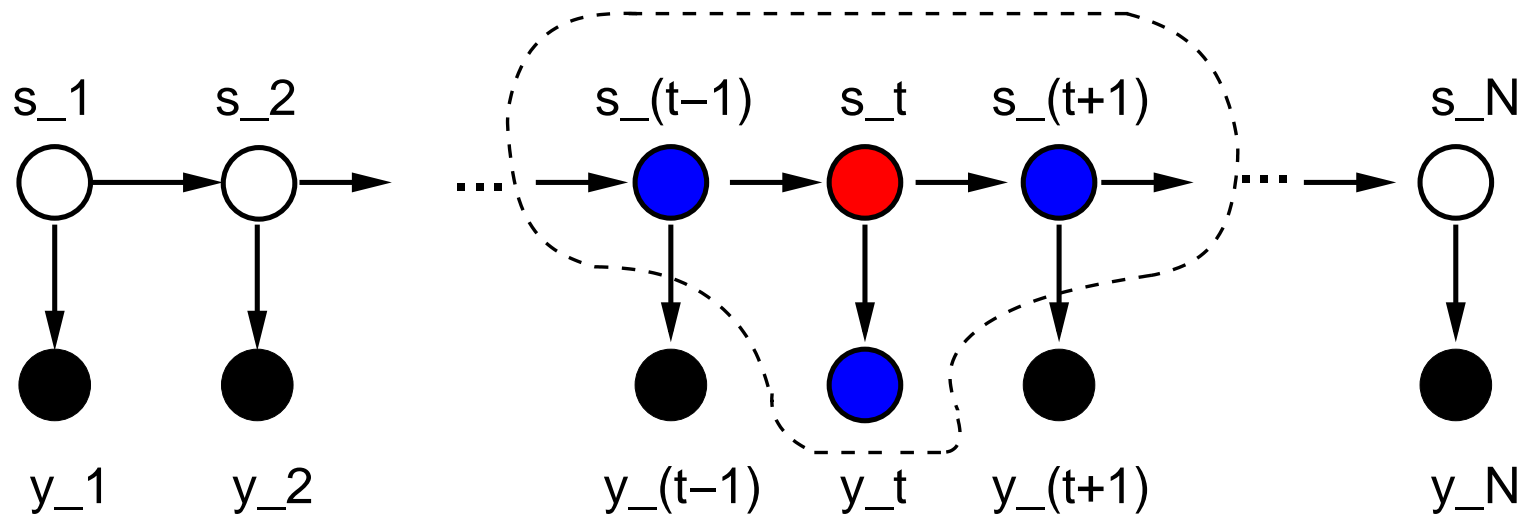
$$P(S_t | S_1, \dots, S_{t-1}, S_{t+1}, \dots, S_N, \mathcal{D}, \mathbf{w}, \nu)$$

Sampling from the posterior distribution



$$\begin{aligned} P(S_t | S_1, \dots, S_{t-1}, S_{t+1}, \dots, S_N, \mathcal{D}, \mathbf{w}, \nu) \\ = P(S_t | S_{t-1}, S_{t+1}, y_t, \mathbf{w}, \nu) \end{aligned}$$

Sampling from the posterior distribution



$$\begin{aligned} P(S_t | S_1, \dots, S_{t-1}, S_{t+1}, \dots, S_N, \mathcal{D}, \mathbf{w}, \nu) \\ &= P(S_t | S_{t-1}, S_{t+1}, \mathbf{y}_t, \mathbf{w}, \nu) \\ &\propto P(S_{t+1} | S_t, \nu) P(S_t | S_{t-1}, \nu) P(\mathbf{y}_t | S_t, \mathbf{w}) \end{aligned}$$

Gibbs-within-Gibbs scheme

Robert, Celeux, Diebolt (1993)
Statistics & Probability Letters 16, 77-83

Robert, Ryden, Titterington (2000)
J. R. Statist. Soc. B, 62, 57-75

Husmeier, McGuire (2003)
Molecular Biology and Evolution 20, 315-337

Gibbs-within-Gibbs scheme

Robert, Celeux, Diebolt (1993)
Statistics & Probability Letters 16, 77-83

Robert, Ryden, Titterington (2000)
J. R. Statist. Soc. B, 62, 57-75

Husmeier, McGuire (2003)
Molecular Biology and Evolution 20, 315-337

Slow mixing and convergence

Gibbs-within-Gibbs scheme

Robert, Celeux, Diebolt (1993)
Statistics & Probability Letters 16, 77-83

Robert, Ryden, Titterton (2000)
J. R. Statist. Soc. B, 62, 57-75

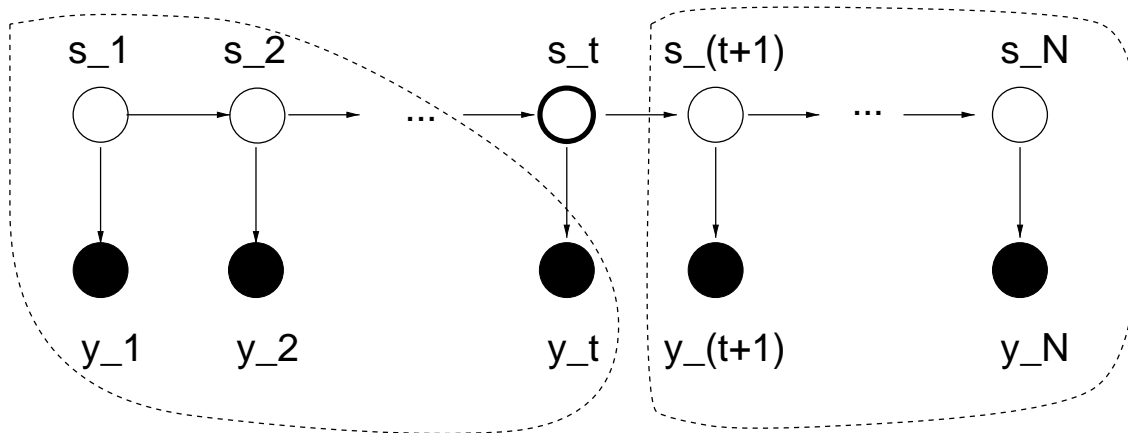
Husmeier, McGuire (2003)
Molecular Biology and Evolution 20, 315-337

Slow mixing and convergence

Boys, Henderson, Wilkinson (2000)
Applied Statistics 49, 269-285

Simultaneous sampling of the states from $P(\mathbf{S}|\mathbf{w}, \nu, \mathcal{D})$

$$\begin{aligned} & P(S_t | S_{t+1}, \dots, S_N, \mathbf{y}_1, \dots, \mathbf{y}_N) \\ \propto & P(S_t, S_{t+1}, \dots, S_N, \mathbf{y}_1, \dots, \mathbf{y}_N) \\ = & P(\mathbf{y}_{t+1}, \dots, \mathbf{y}_N, S_{t+1}, \dots, S_N | S_t, \mathbf{y}_1, \dots, \mathbf{y}_t) P(S_t, \mathbf{y}_1, \dots, \mathbf{y}_t) \\ = & P(\mathbf{y}_{t+1}, \dots, \mathbf{y}_N, S_{t+1}, \dots, S_N | S_t) \alpha_t(S_t) \\ = & P(\mathbf{y}_{t+1}, \dots, \mathbf{y}_N, S_{t+2}, \dots, S_N | S_{t+1}) P(S_{t+1} | S_t) \alpha_t(S_t) \\ \propto & P(S_{t+1} | S_t) \alpha_t(S_t) \end{aligned}$$



Stochastic forward–backward algorithm

- Run the **forward algorithm** to obtain

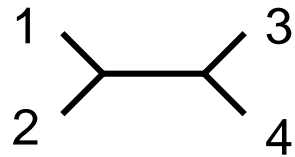
$$\alpha_t(S_t) = P(S_t, \mathbf{y}_1, \dots, \mathbf{y}_t)$$

- Sample S_N from $P(S_N = k | \mathbf{y}_1, \dots, \mathbf{y}_N) = \frac{\alpha_N(S_N=k)}{\sum_i \alpha_N(S_N=i)}$

- Sample the remaining states S_{N-1}, \dots, S_1 recursively from

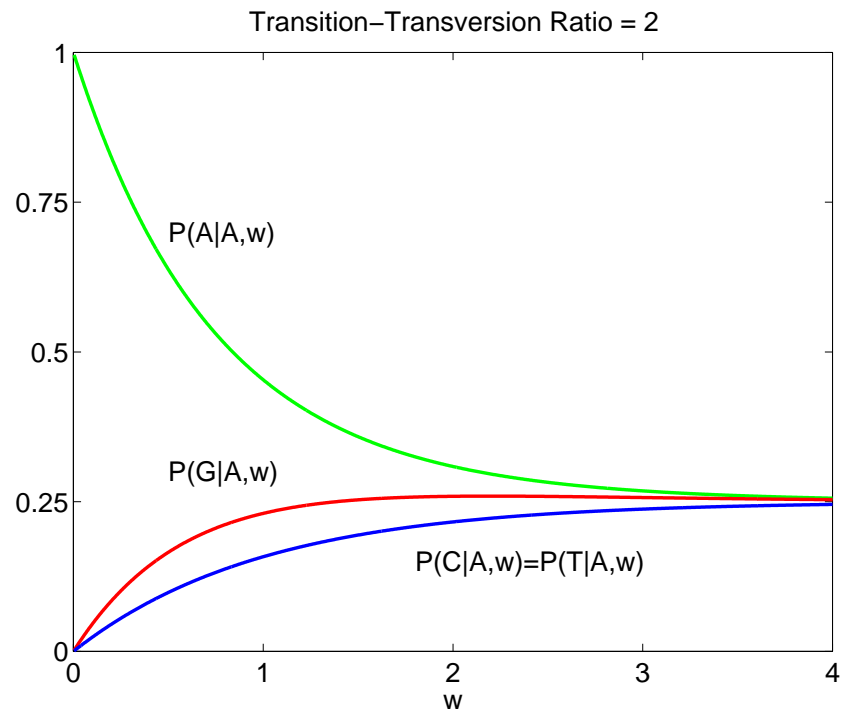
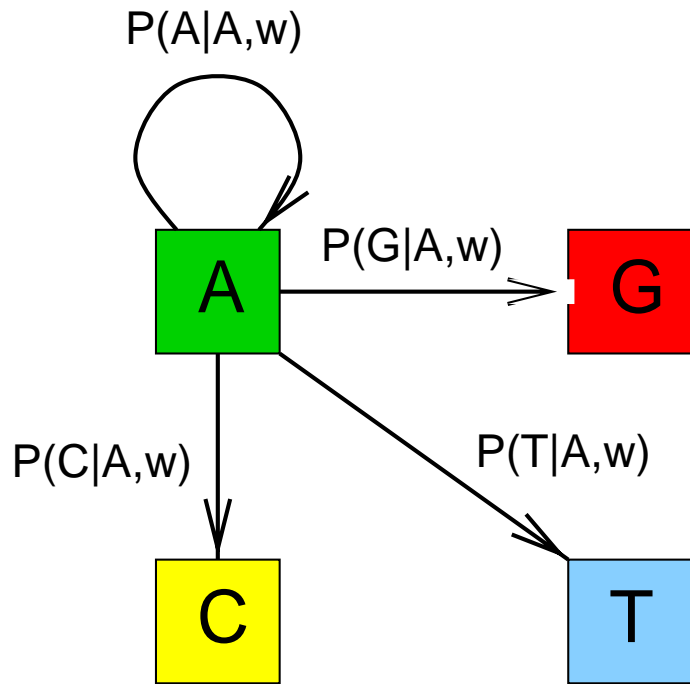
$$P(S_t = k | S_{t+1}, \dots, S_N, \mathbf{y}_1, \dots, \mathbf{y}_N) = \frac{P(S_{t+1} | S_t=k) \alpha_t(S_t=k)}{\sum_i P(S_{t+1} | S_t=i) \alpha_t(S_t=i)}$$

Synthetic example



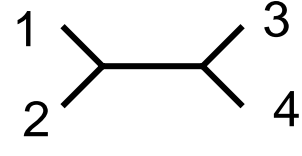
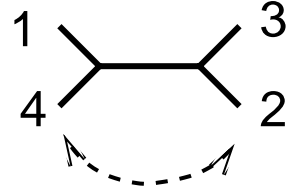
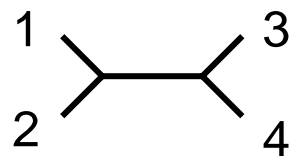
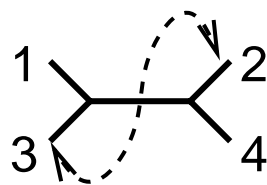
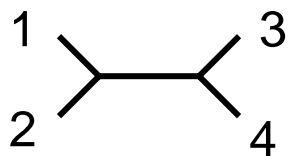
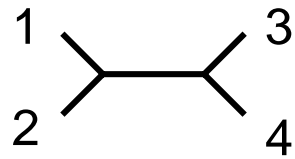
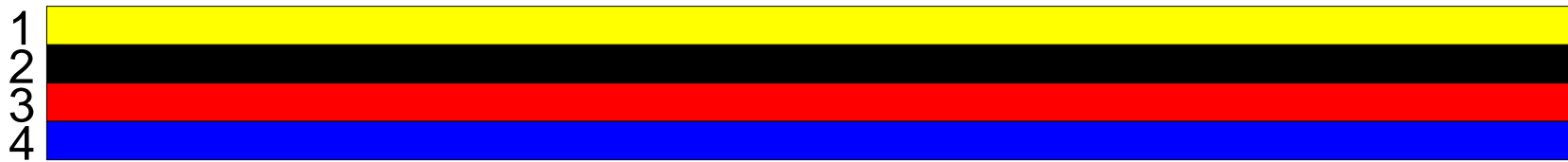
- Model of nucleotide substitution: Kimura 2-parameter, $\tau = 2$.
- Alignment of length $N = 1000$ nucleotides.

Mutation probabilities

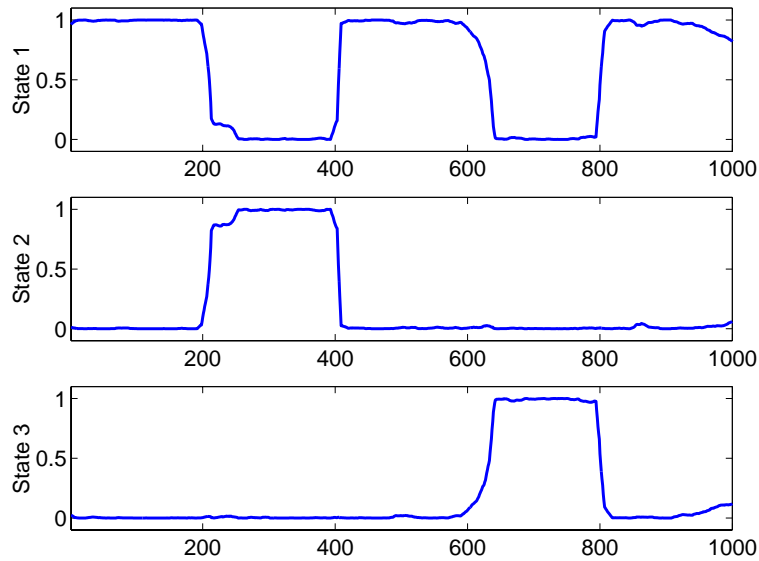
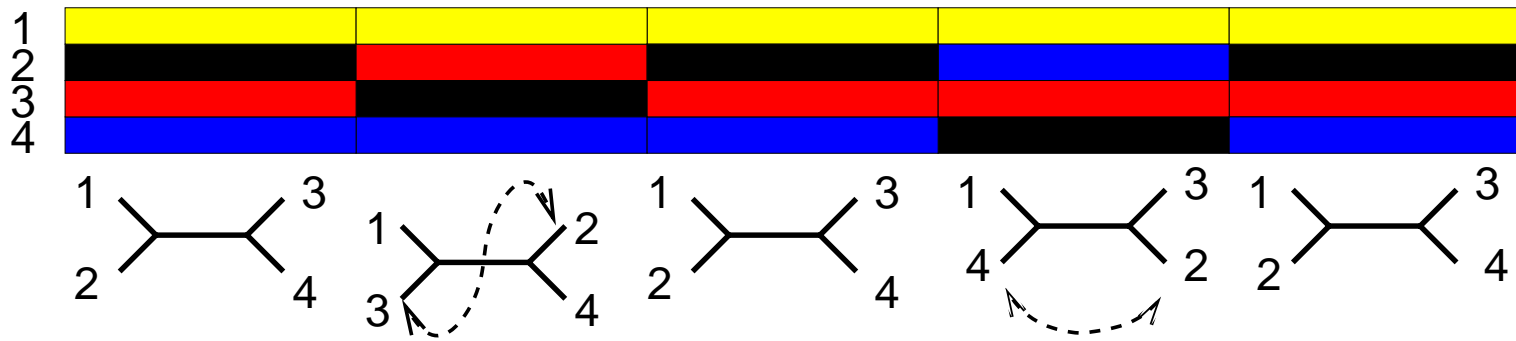


branch length = mutation rate \times time

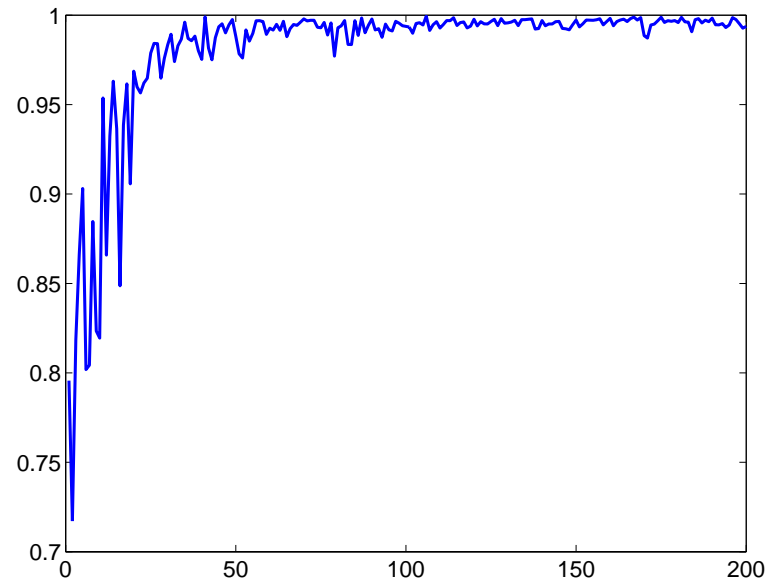
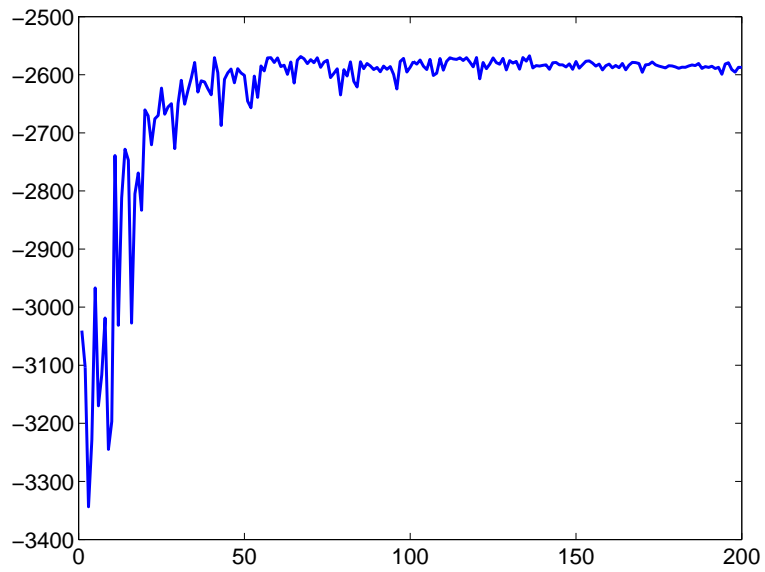
Synthetic example



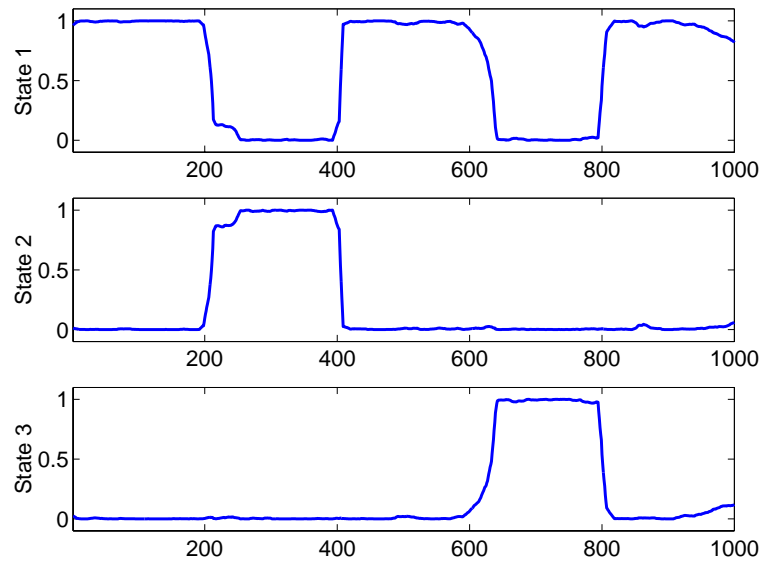
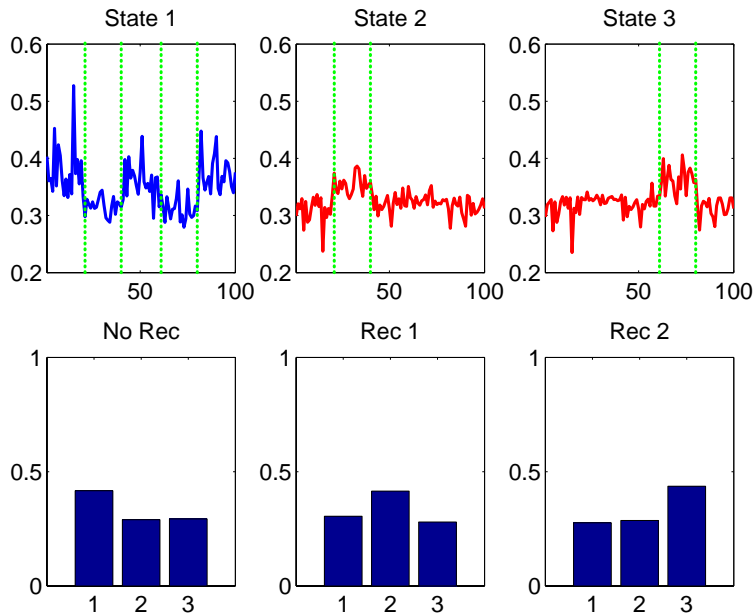
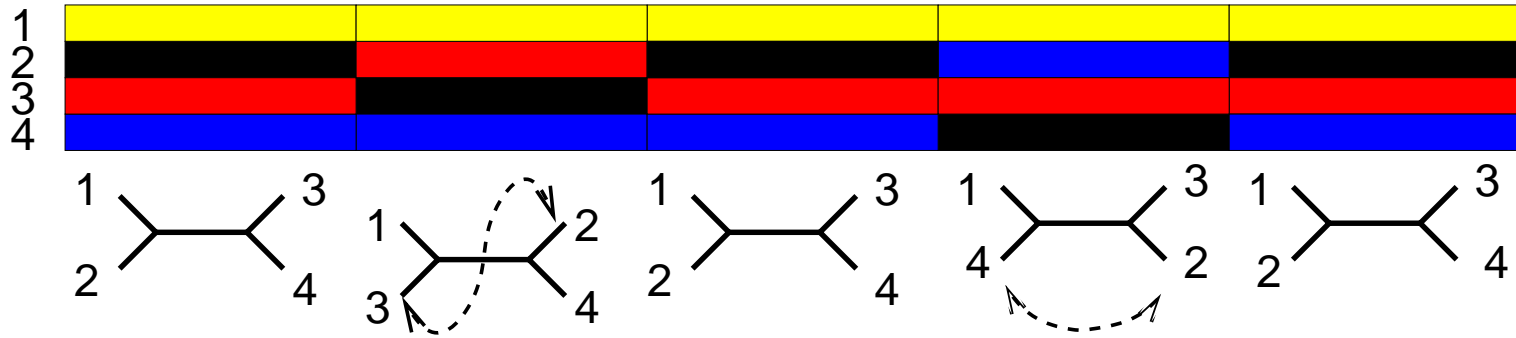
$P(S_t|\mathcal{D})$: Marginal posterior probability



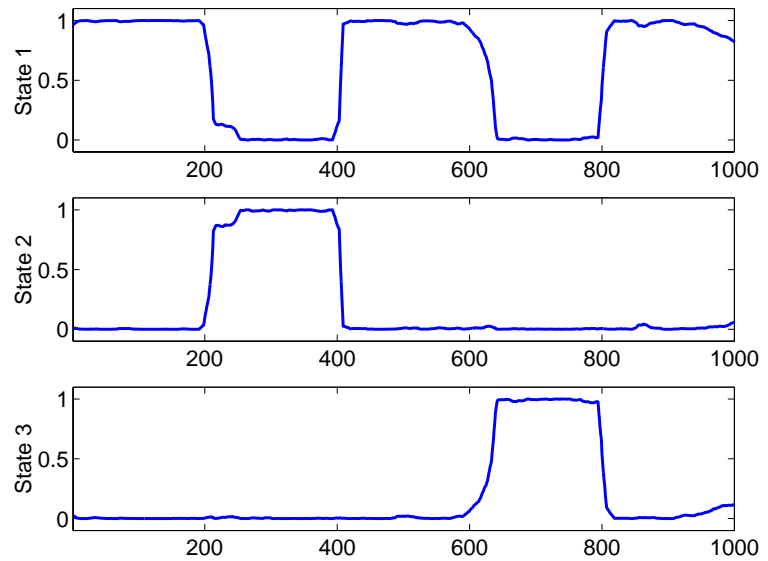
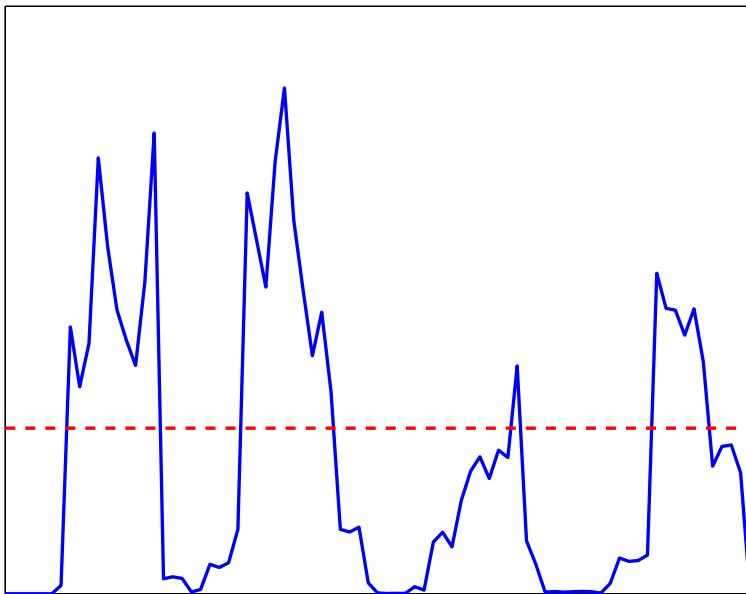
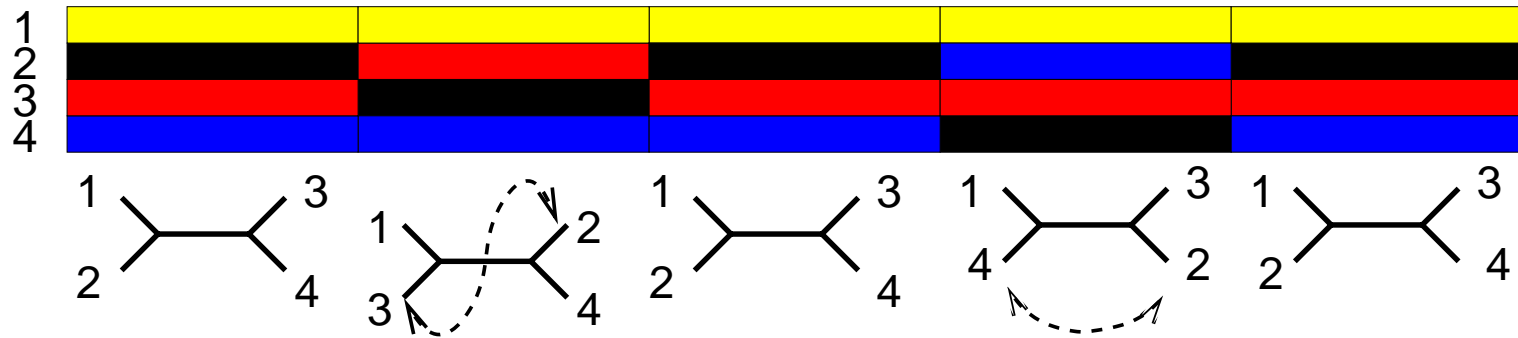
Trace plots of the **log likelihood** (left) and ν (right)



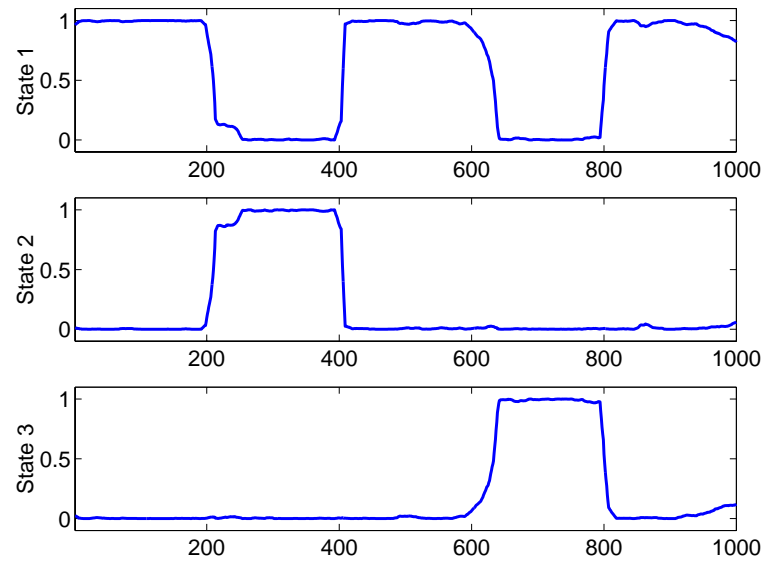
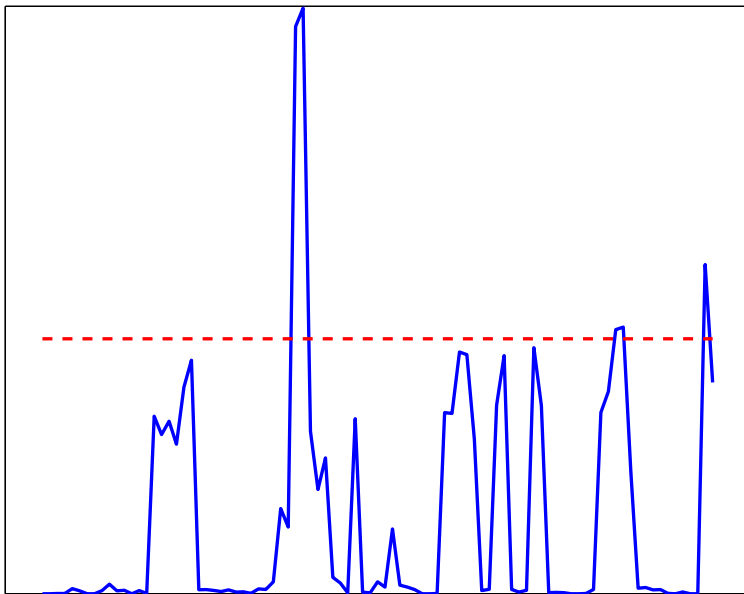
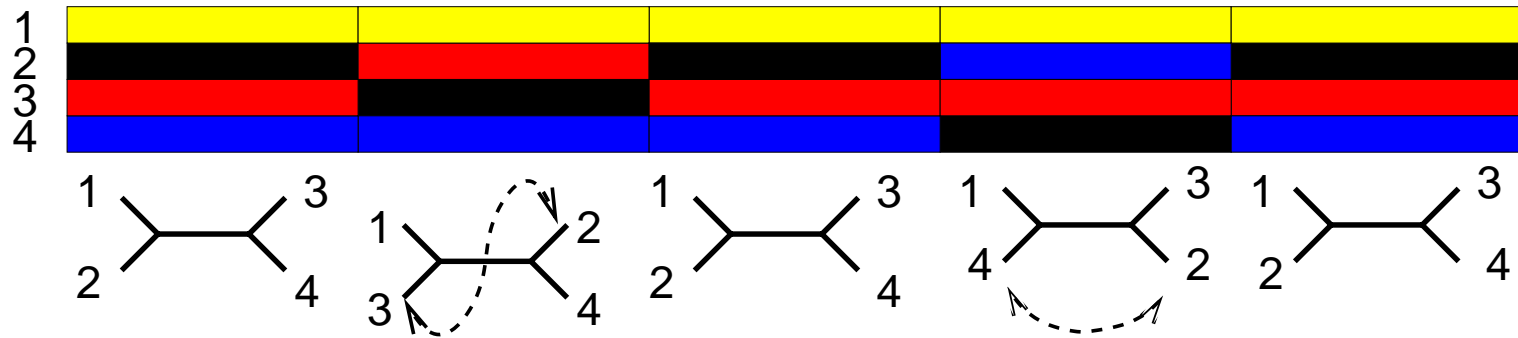
Phylo-HMM vs. naive method



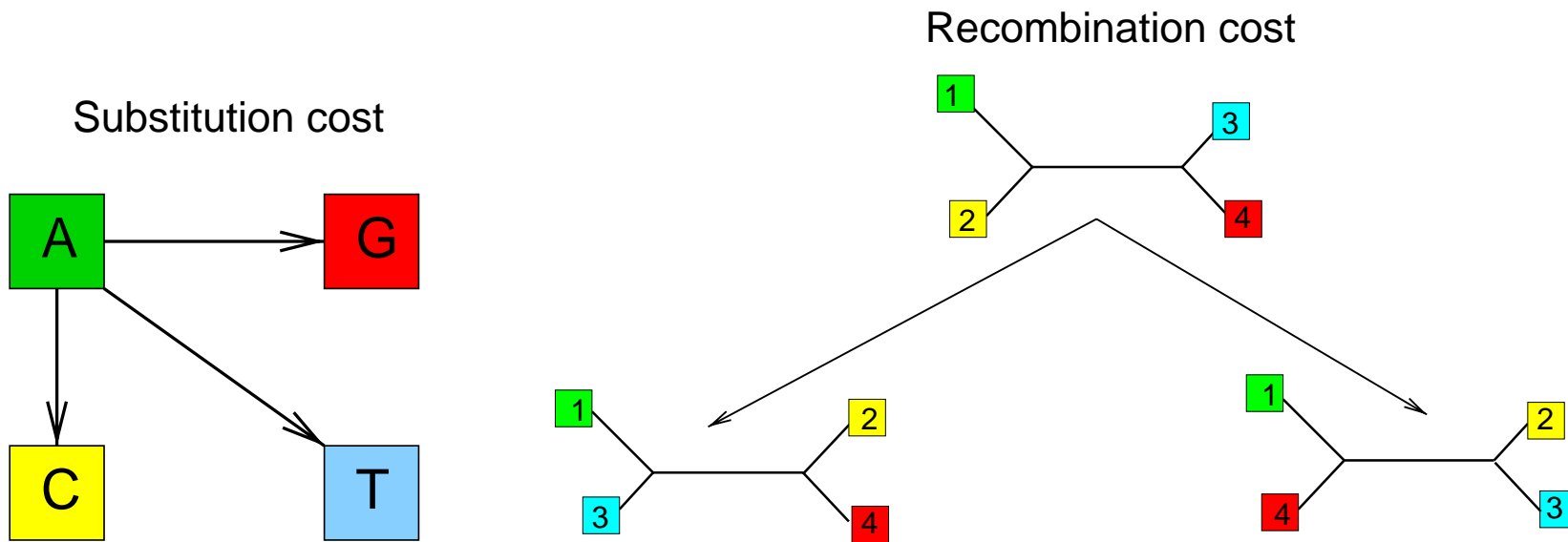
Phylo-HMM vs. Topal, window size=200



Phylo-HMM vs. Topal, window size=100



Comparison with RecPars (Jotun Hein 1993)



$$\Psi = \text{Recombination cost} / \text{substitution cost}$$

The most likely state sequence: Viterbi algorithm

$$\begin{aligned}\max_{S_1, \dots, S_N} P(S_1, \dots, S_N | y_1, \dots, y_N) &= \max_{S_1, \dots, S_N} P(S_1, \dots, S_N, y_1, \dots, y_N) \\ &= \max_{S_N} \gamma_N(S_N)\end{aligned}$$

$$\begin{aligned}\gamma_n(S_n) &= \max_{S_1, \dots, S_{n-1}} P(y_1, \dots, y_n, S_1, \dots, S_n) \\ &= \max_{S_1, \dots, S_{n-1}} \prod_{t=1}^n P(y_t | S_t) P(S_t | S_{t-1}) \\ &= \max_{S_1, \dots, S_{n-1}} P(y_n | S_n) P(S_n | S_{n-1}) \prod_{t=1}^{n-1} P(y_t | S_t) P(S_t | S_{t-1}) \\ &= P(y_n | S_n) \max_{S_{n-1}} P(S_n | S_{n-1}) \max_{S_1, \dots, S_{n-2}} \prod_{t=1}^{n-1} P(y_t | S_t) P(S_t | S_{t-1}) \\ &= P(y_n | S_n) \max_{S_{n-1}} P(S_n | S_{n-1}) \gamma_{n-1}(S_{n-1})\end{aligned}$$

Relation to RECPARS (Hein 1993)

$$\gamma_n(S_n) = P(y_n|S_n) \max_{S_{n-1}} P(S_n|S_{n-1}) \gamma_{n-1}(S_{n-1})$$

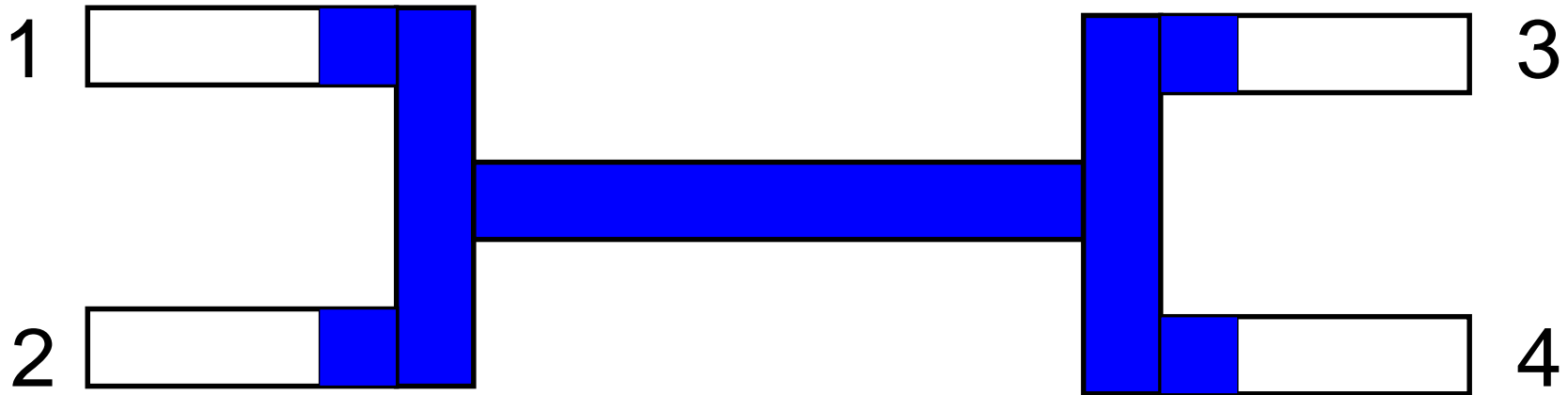
$$\log \gamma_n(S_n) = \log P(y_n|S_n) + \max_{S_{n-1}} [\log P(S_n|S_{n-1}) + \log \gamma_{n-1}(S_{n-1})]$$

- Parsimony cost: $E(S_n) = -\log \gamma_n(S_n)$
- Mutation cost: $M(y_n|S_n) = -\log P(y_n|S_n)$
- Recombination cost: $R(S_n|S_{n-1}) = -\log P(S_n|S_{n-1})$

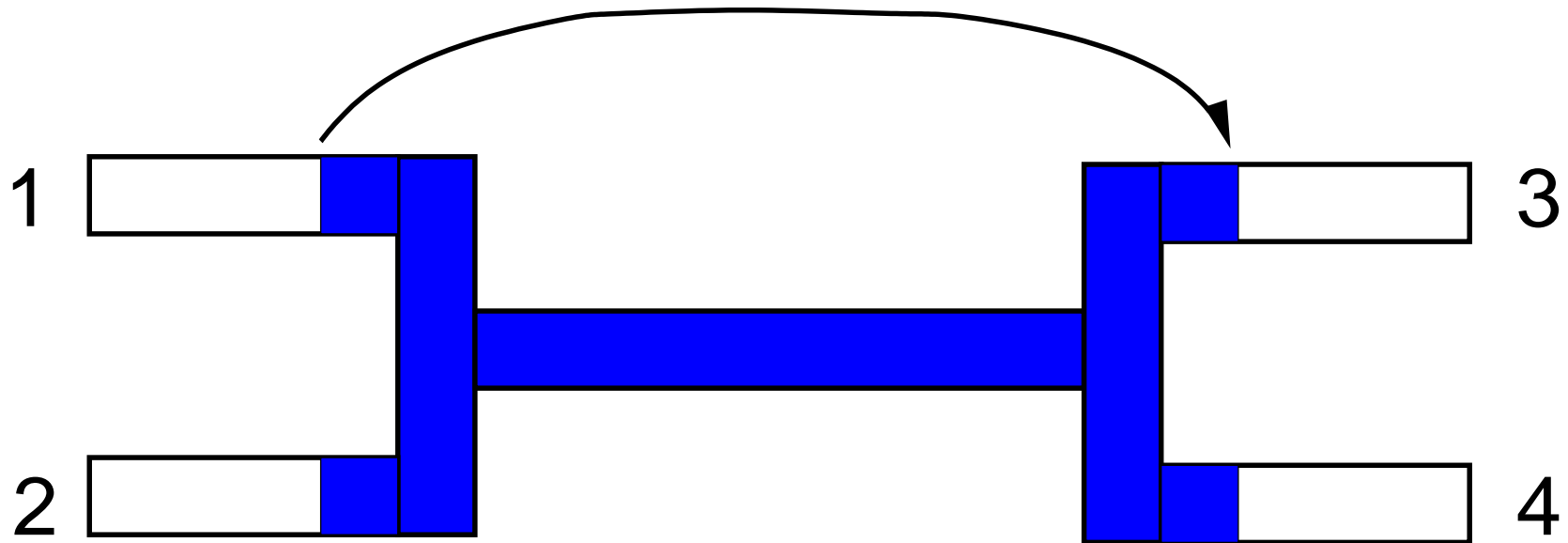
$$E(S_n) = M(y_n|S_n) + \min_{S_{n-1}} [R(S_n|S_{n-1}) + E(S_{n-1})]$$

Find sequence S_1, \dots, S_N that minimises E

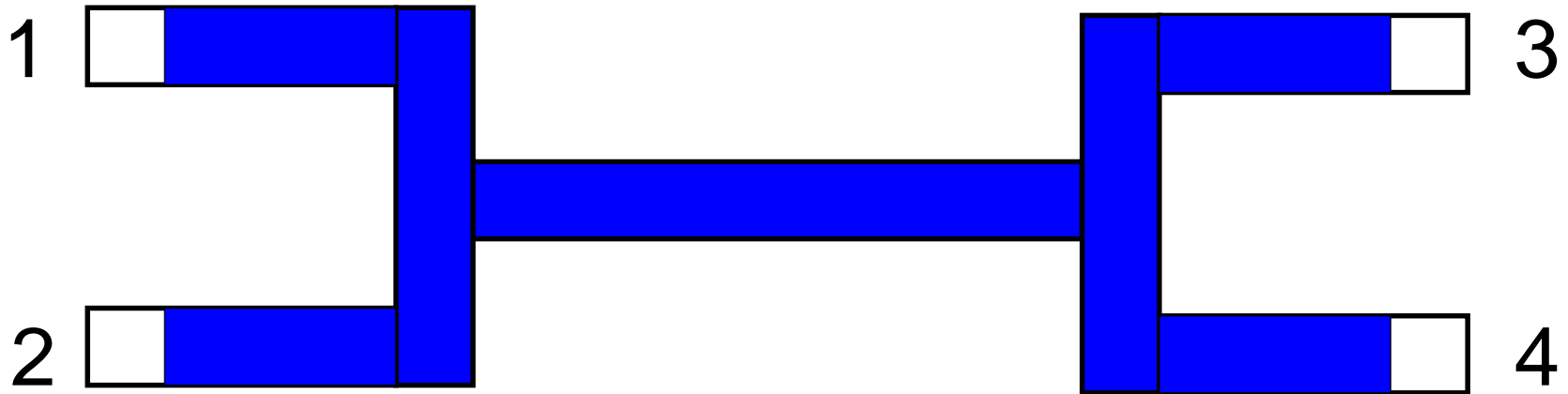
Simulation of recombination



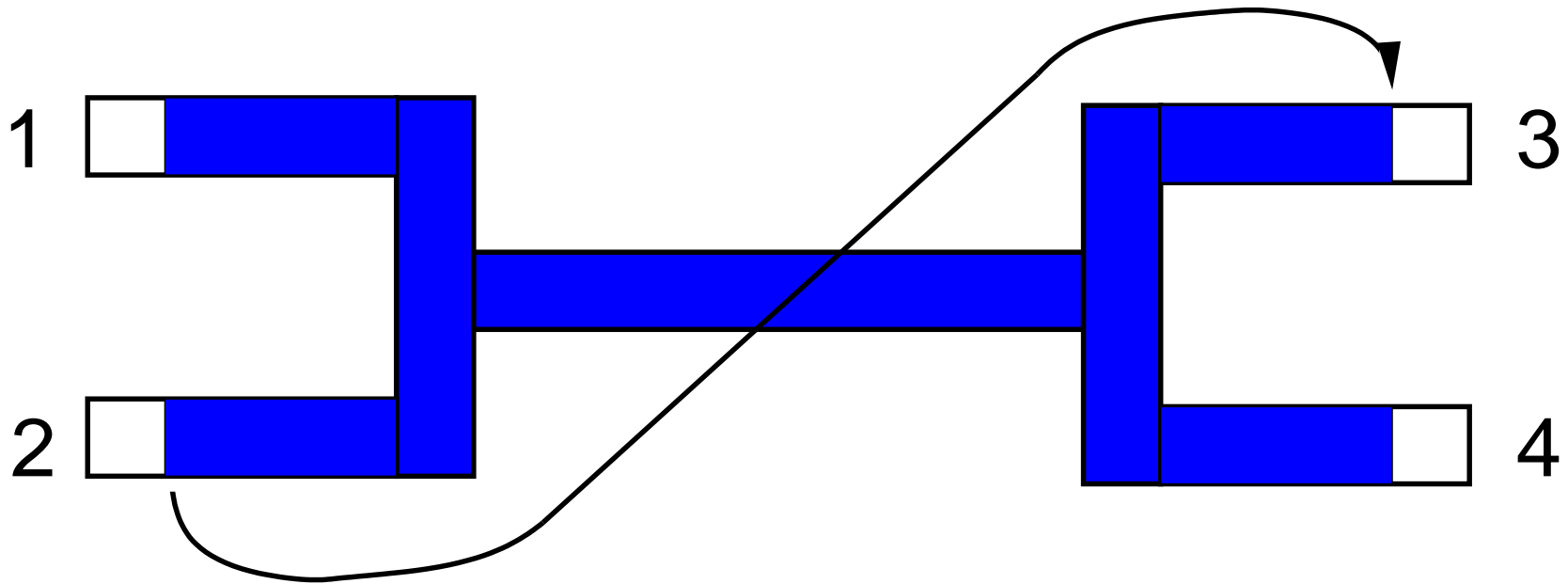
Simulation of recombination



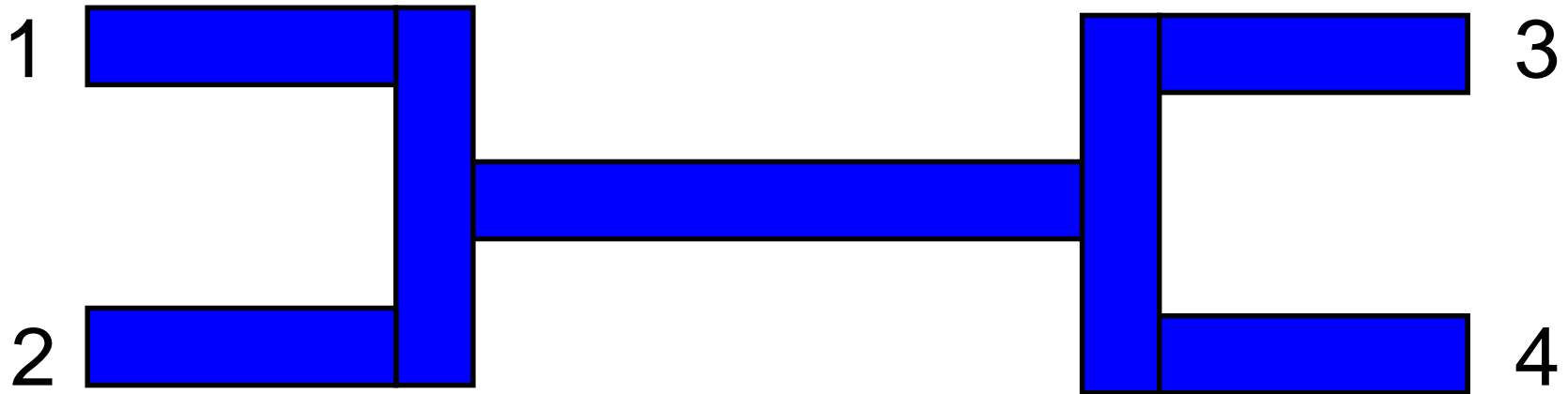
Simulation of recombination



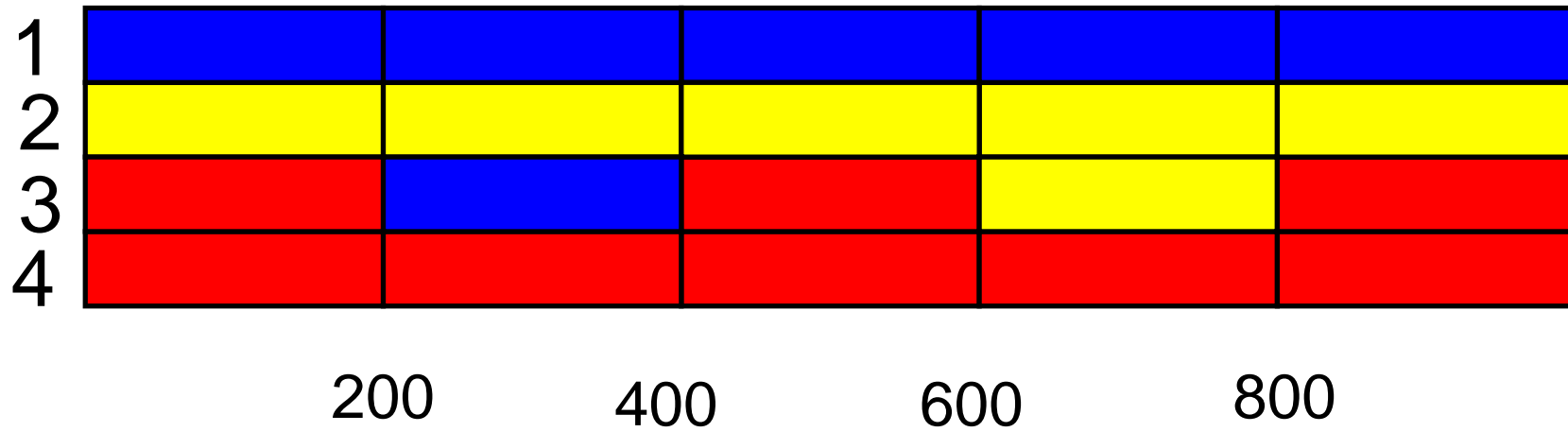
Simulation of recombination



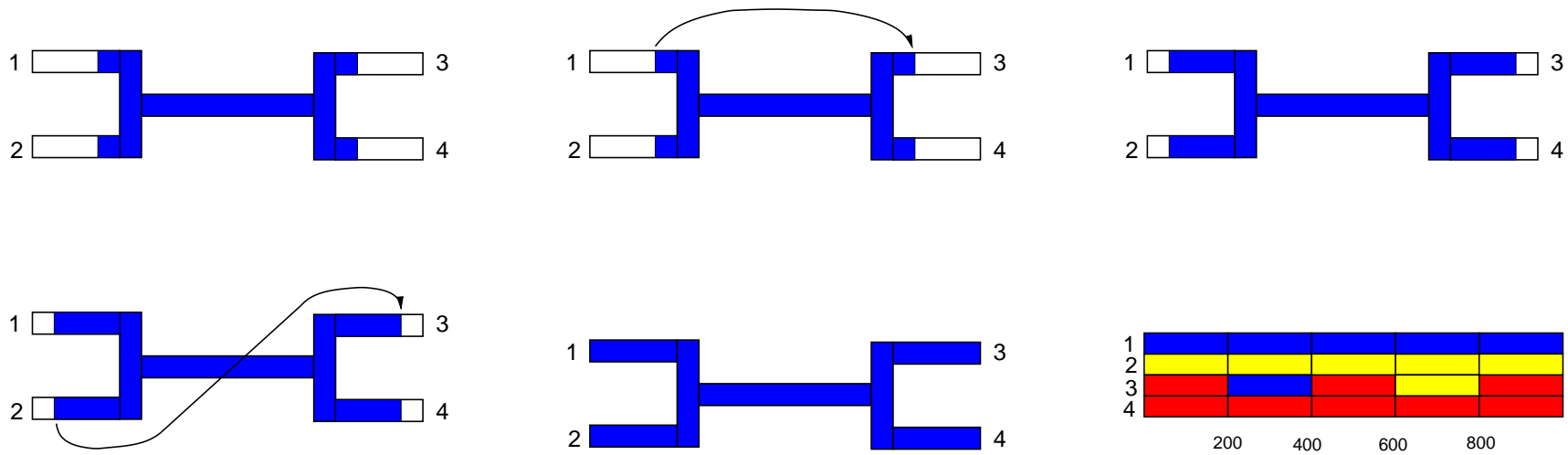
Simulation of recombination



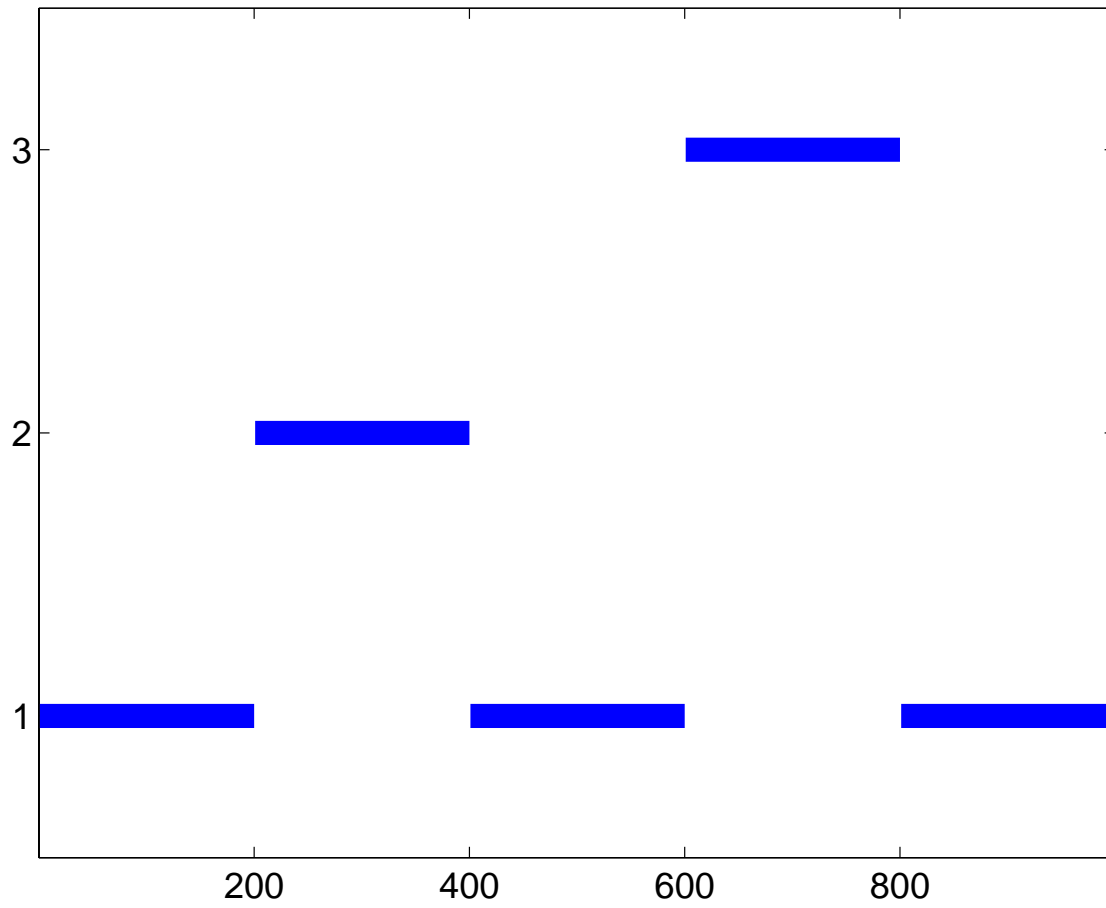
Simulation of recombination



Synthetic simulation study

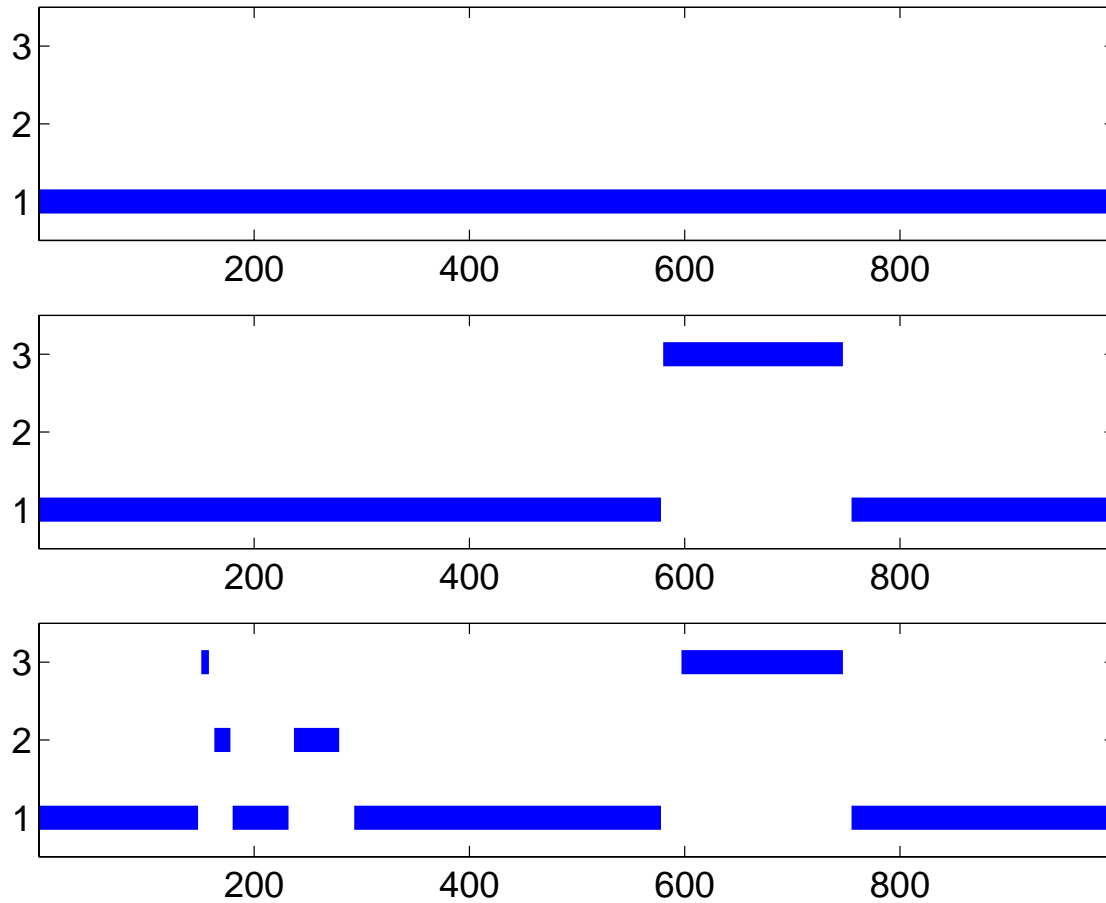


True mosaic structure

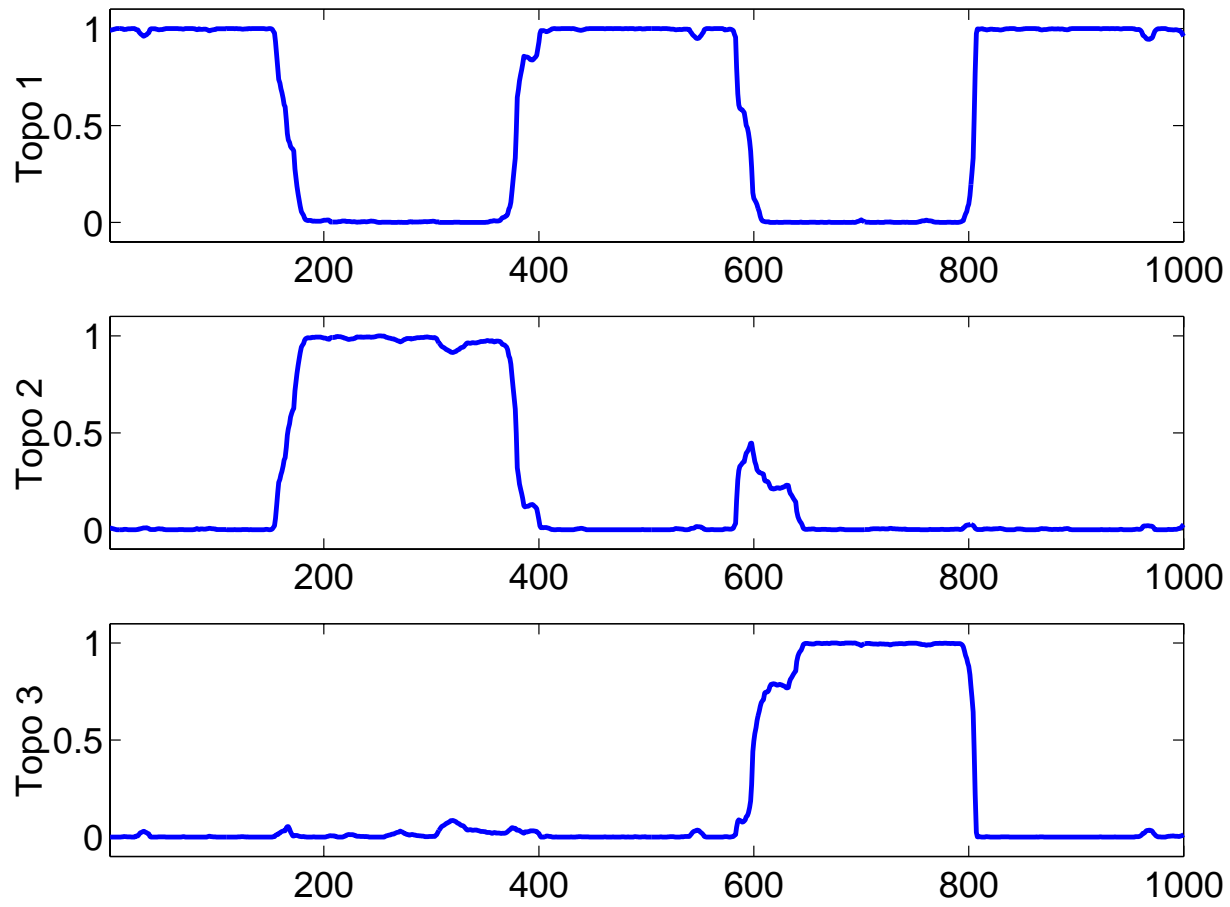


Prediction with RECPARS (Hein, 1993)

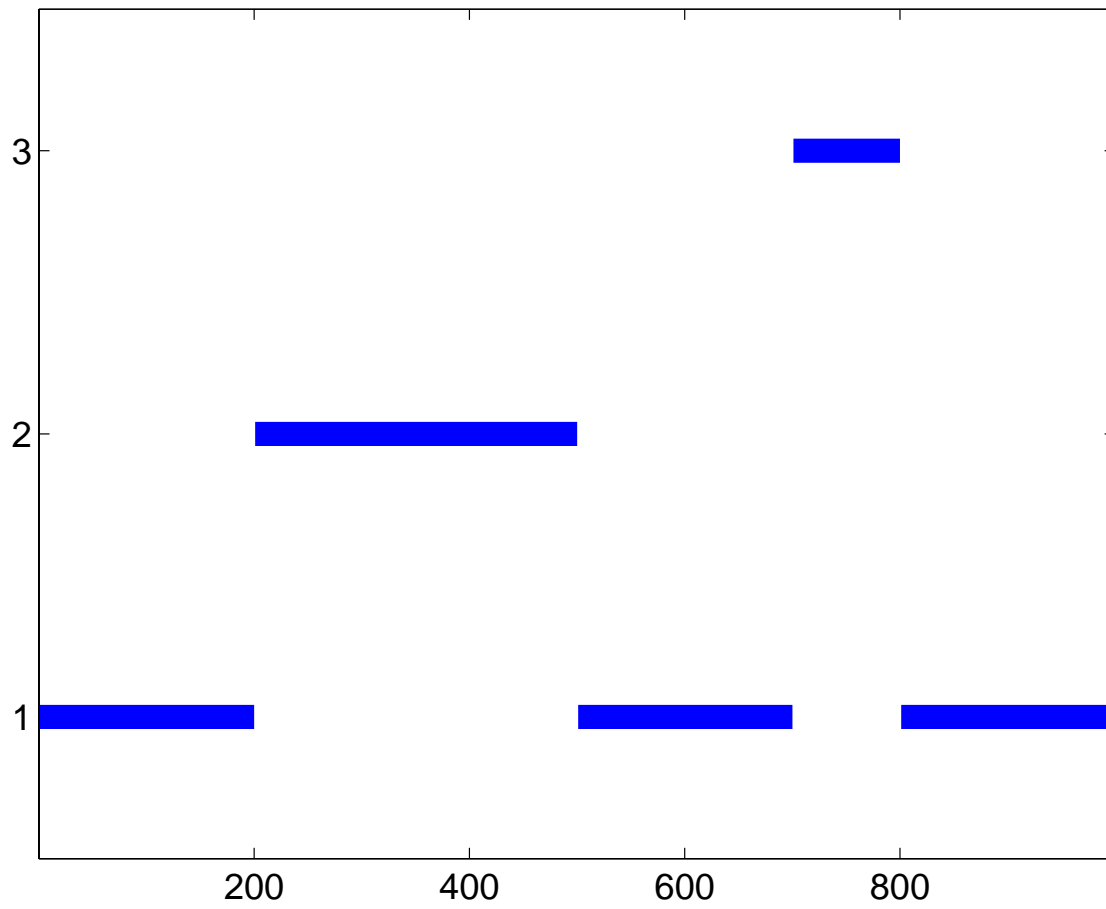
Top: $C_{recomb}/C_{mut} = 10.0$ Middle: $C_{recomb}/C_{mut} = 3.0$ Right: $C_{recomb}/C_{mut} = 1.5$



Prediction with HMM-Bytes

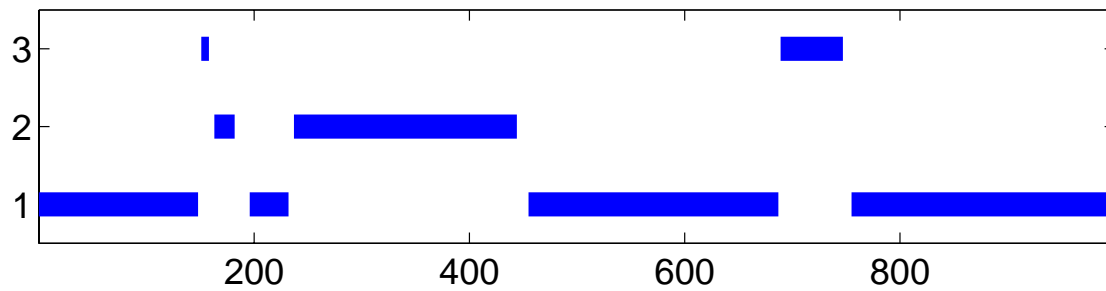
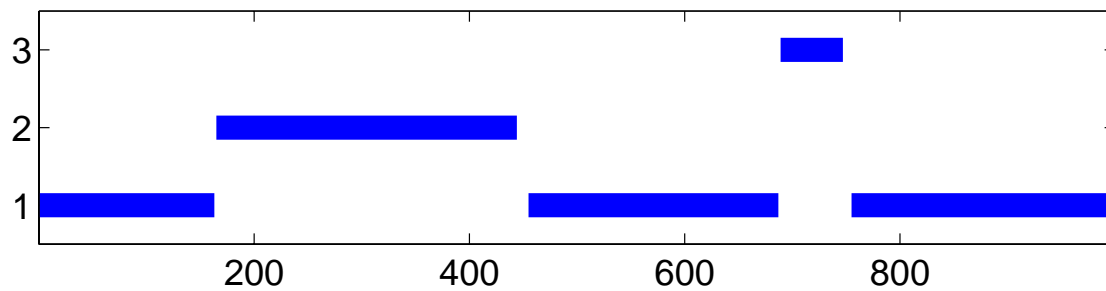
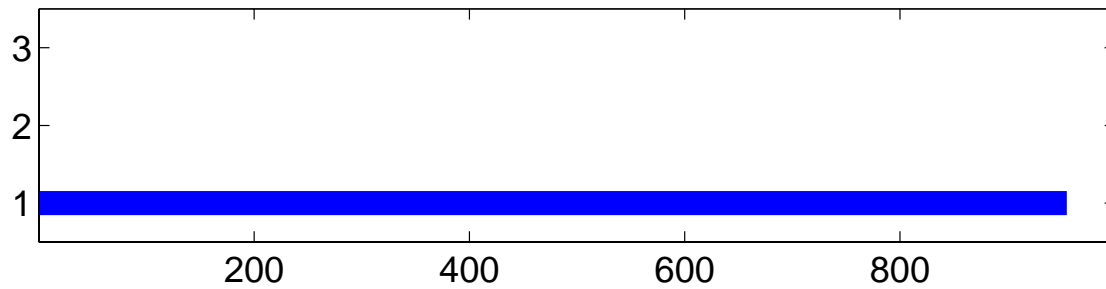


True mosaic structure

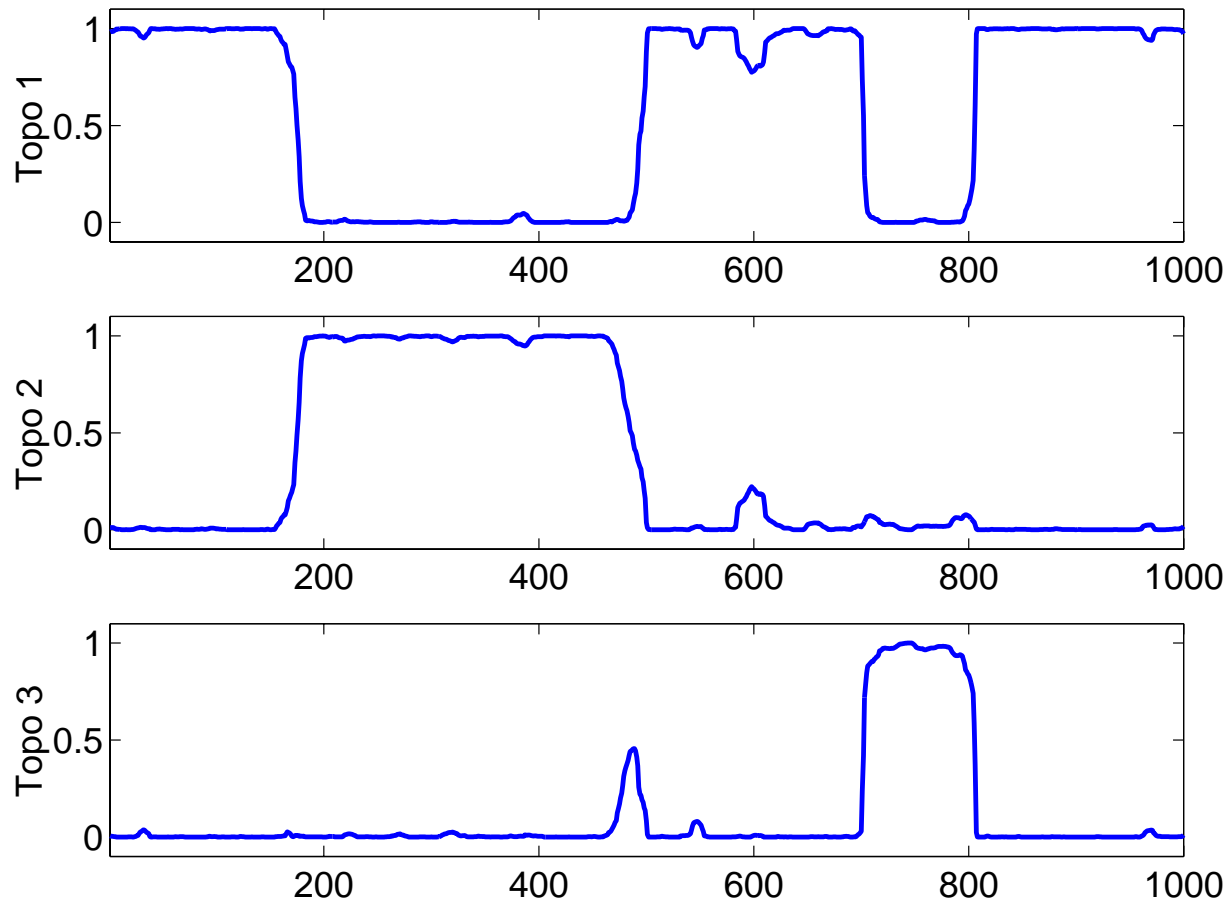


Prediction with RECPARS

Top: $C_{recomb}/C_{mut} = 10.0$ Middle: $C_{recomb}/C_{mut} = 3.0$ Right: $C_{recomb}/C_{mut} = 1.5$



Prediction with HMM-Bayes



Hepatitis B Virus (Bollyky et al. 1995)

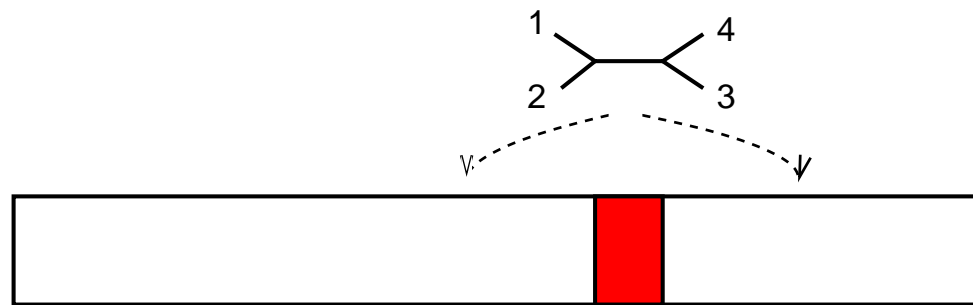
DNA alignment, 3049 nucleotides

1) HPBADW1

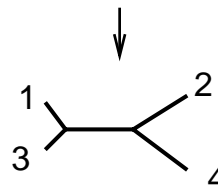
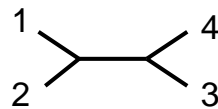
2) HPBADW2

3) HPBADWZCG

4) HPBADRC



State 1



State 2

$P(S_t|\mathcal{D})$: Marginal posterior probability

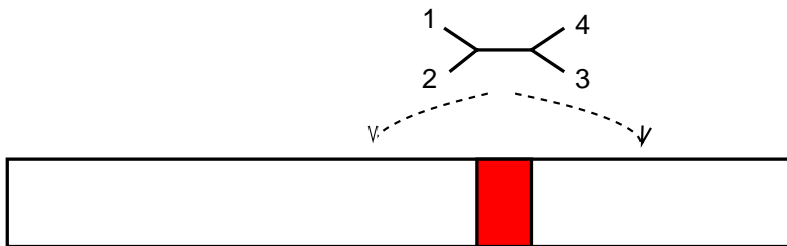
DNA alignment, 3049 nucleotides

1) HPBADW1

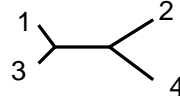
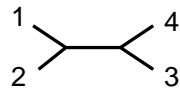
2) HPBADW2

3) HPBADWZCG

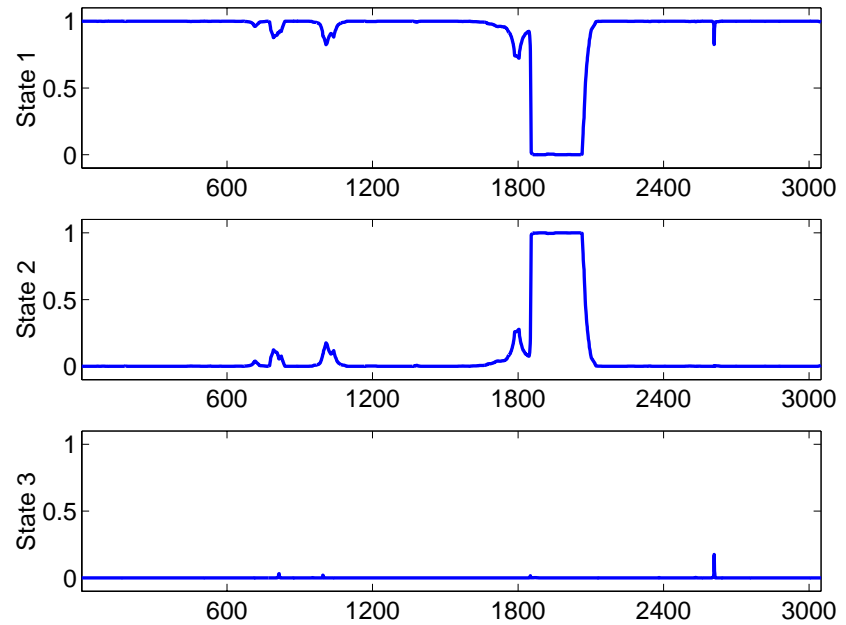
4) HPBADRC



State 1



State 2



Gibbs-within-Gibbs sampling

Burn-in:	10^6 Gibbs steps
Sampling:	10^6 Gibbs steps
Computational costs:	> 6 hours (on SUN Ultra-60)

Gibbs-within-Gibbs sampling

Burn-in:	10^6 Gibbs steps
Sampling:	10^6 Gibbs steps
Computational costs:	> 6 hours (on SUN Ultra-60)

By how much can we reduce the computational costs with the
modified forward-backward algorithm?

Gibbs-within-Gibbs sampling

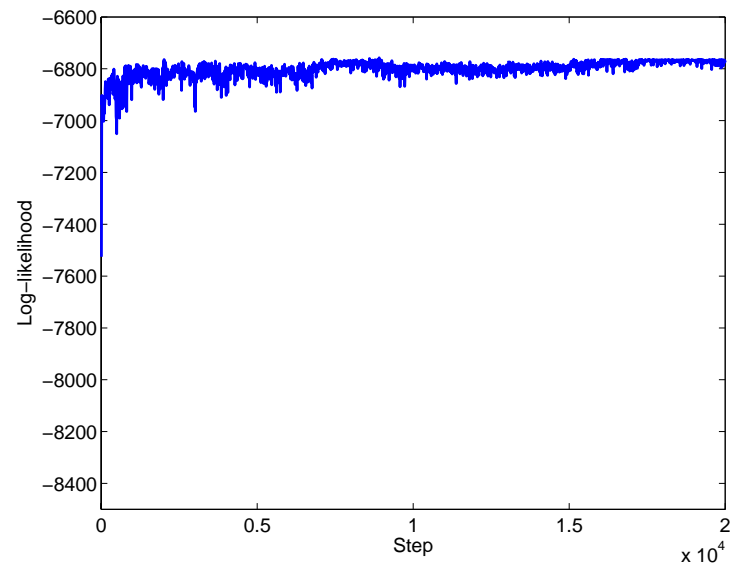
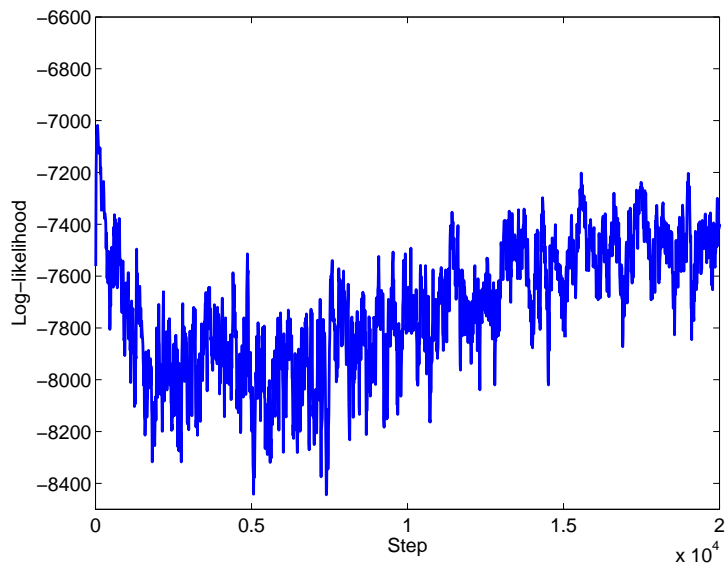
Burn-in:	10^6 Gibbs steps
Sampling:	10^6 Gibbs steps
Computational costs:	> 6 hours (on SUN Ultra-60)

By how much can we reduce the computational costs with the
modified forward-backward algorithm?

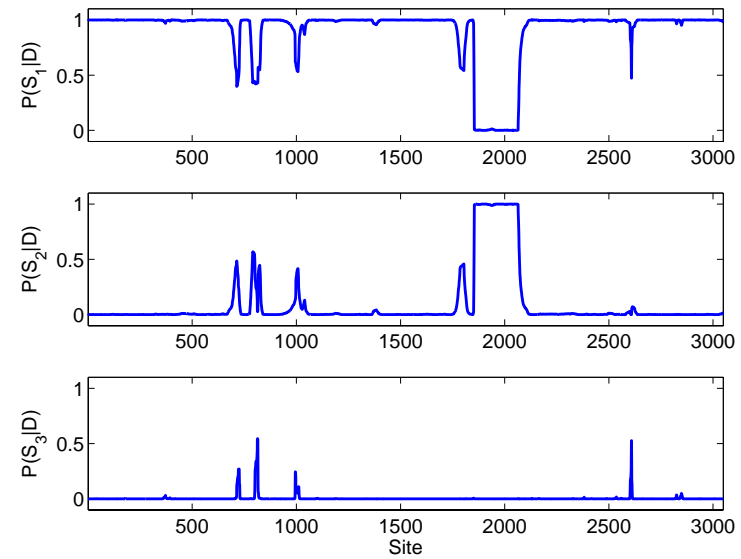
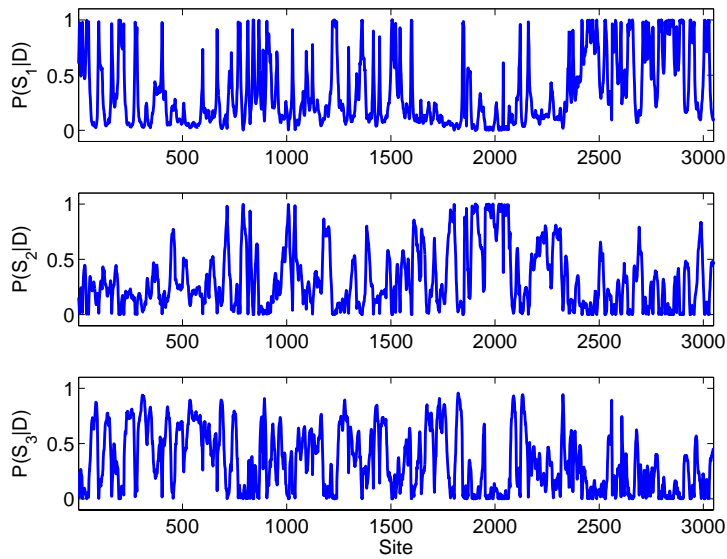
Burn-in:	10^4 Gibbs steps
Sampling:	10^4 Gibbs steps
Computational costs:	a few minutes (on SUN Ultra-60)

Adriano Werhli

Comparison Gibbs-within-Gibbs versus stochastic forward-backward algorithm



Comparison Comparison Gibbs-within-Gibbs versus stochastic forward-backward algorithm



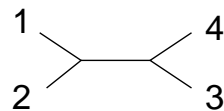
Gene conversion in maize (Moniz de Sa, Drouin, 1996)

Actin genes: DNA alignment of 1008 nucleotides

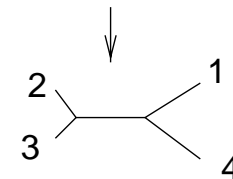
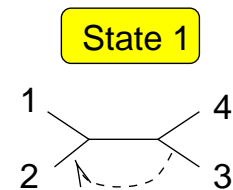
- 1) Maz56
- 2) Maz63
- 3) Maz63
- 4) Maz89

875 bases

133 bases



State 1

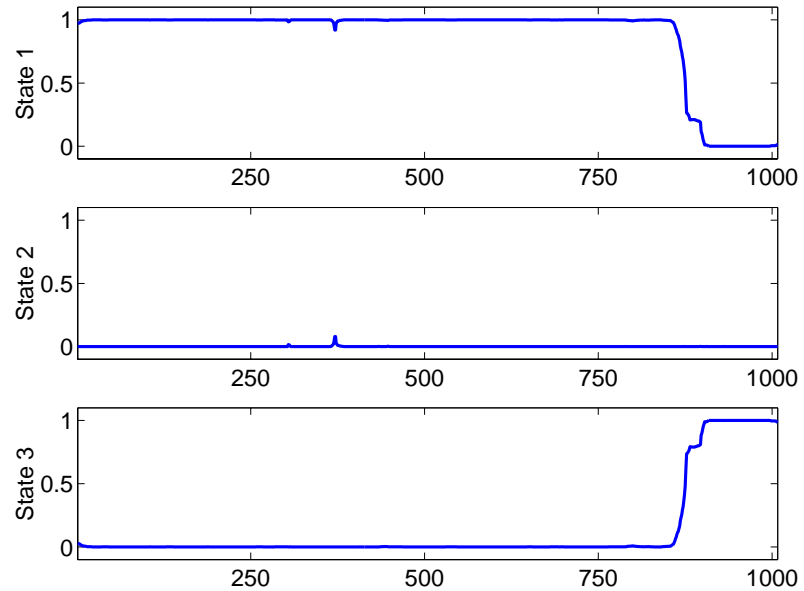
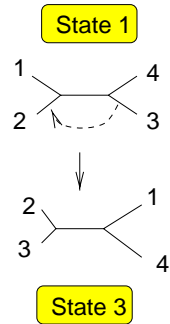
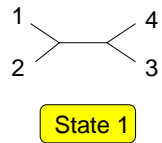


State 3

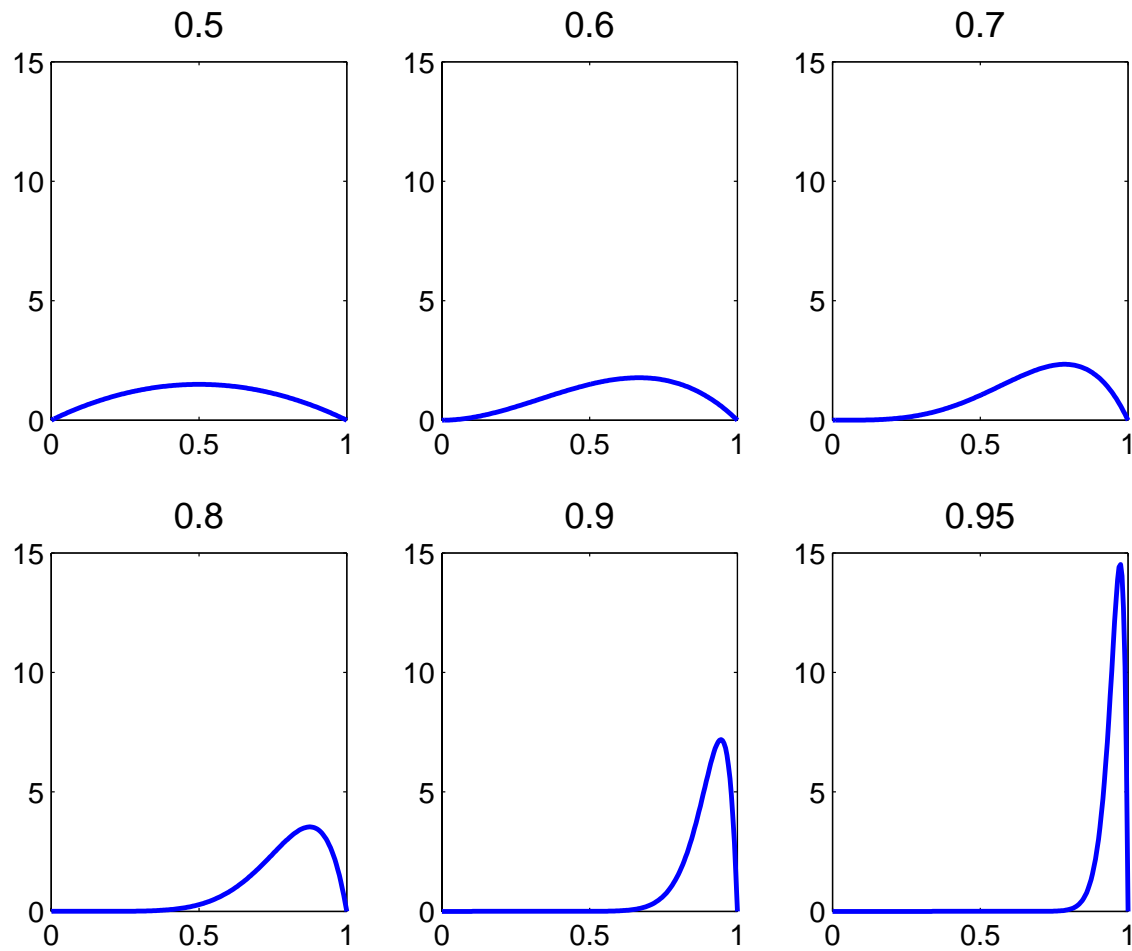
$P(S_t|\mathcal{D})$: Marginal posterior probability

Actin genes: DNA alignment of 1008 nucleotides

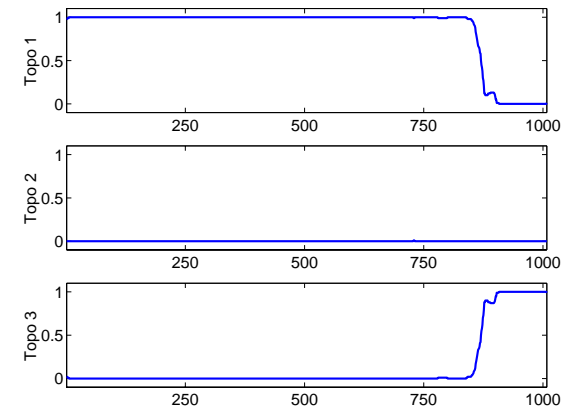
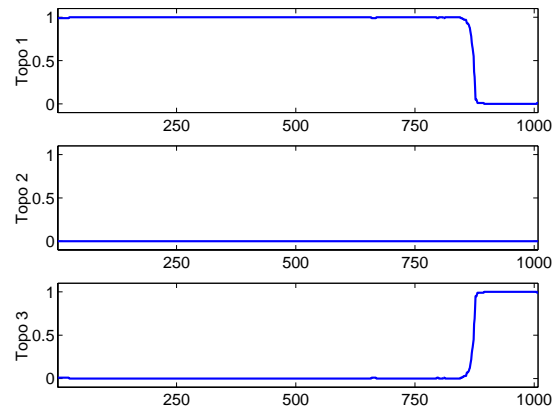
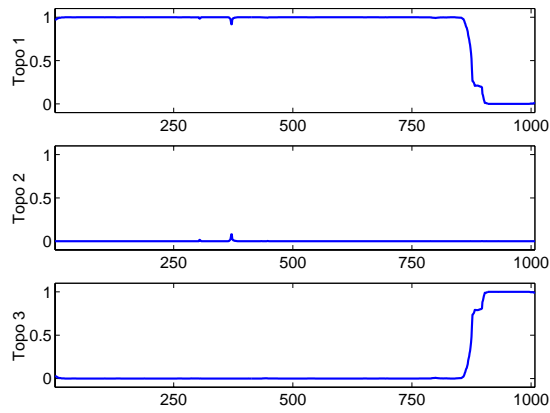
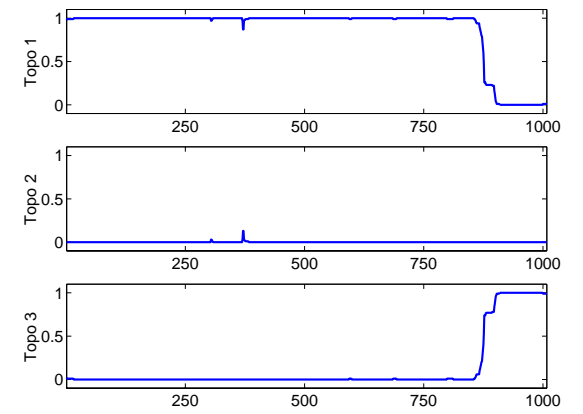
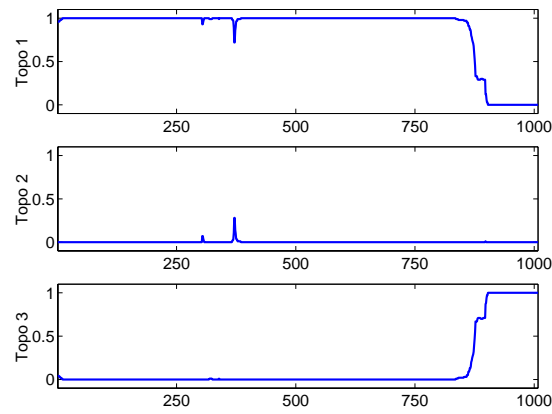
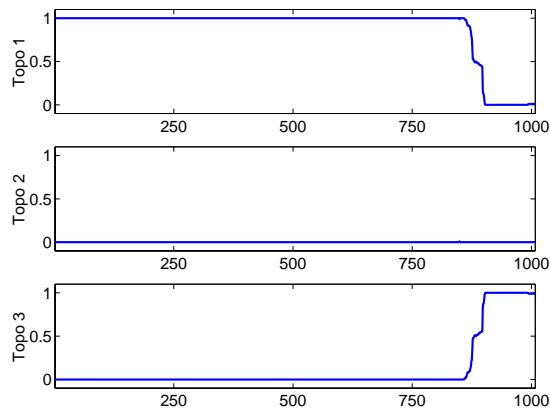
- 1) Maz56
- 2) Maz63
- 3) Maz63
- 4) Maz89



Beta Prior, $\beta = 2$, $\mu = \alpha / (\alpha + \beta)$



Dependence on the prior and the initialization



Neisseria (Zhou & Spratt, 1992)

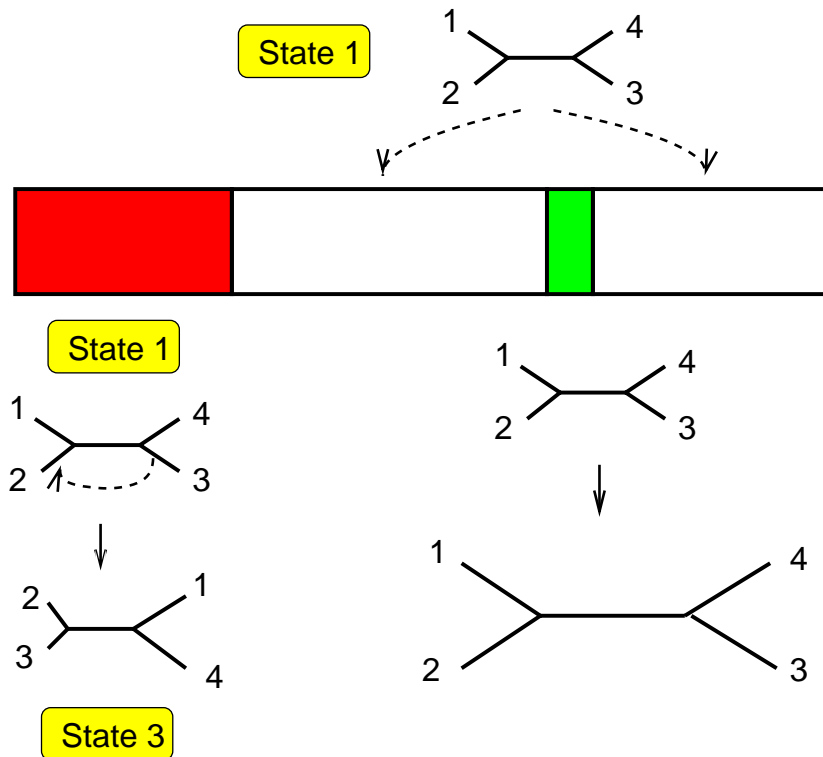
DNA alignment, 787 nucleotides (argF gene)

- | | |
|----------------------------------|-----------------------------|
| 1) Neisseria gonorrhoeae | 3) Neisseria cinerea |
| 2) Neisseria meningitidis | 4) Neisseria mucosa |

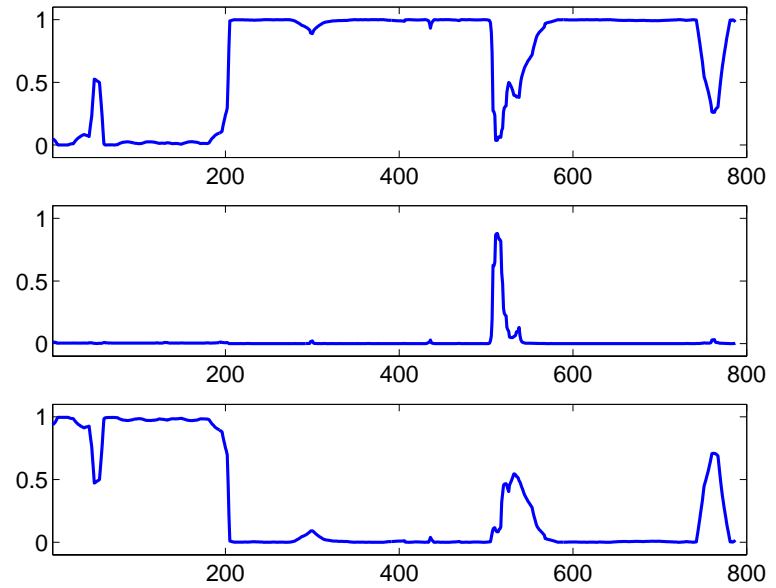
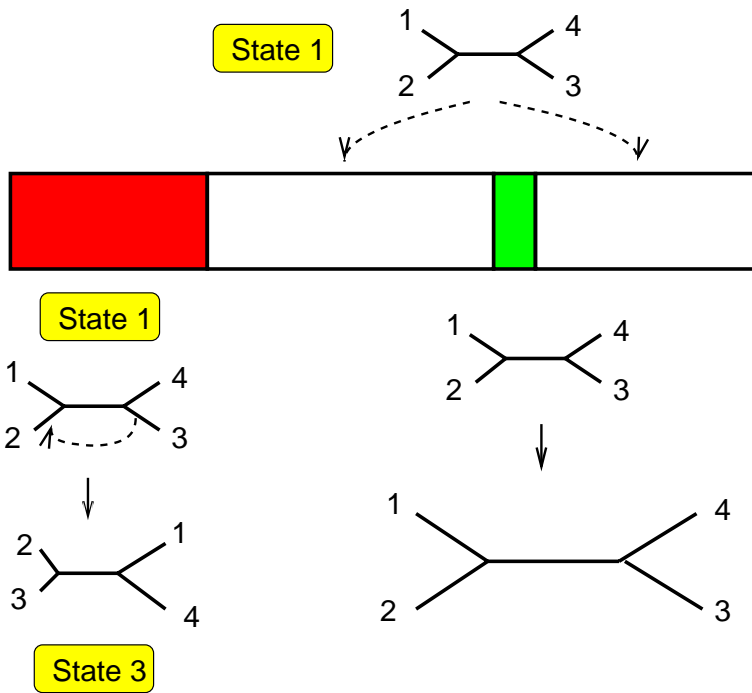
Neisseria (Zhou & Spratt, 1992)

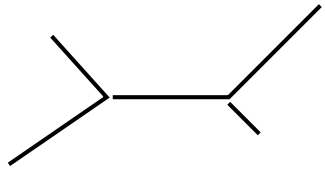
DNA alignment, 787 nucleotides (argF gene)

- 1) Neisseria **gonorrhoeae**
- 2) Neisseria **meningitidis**
- 3) Neisseria **cinerea**
- 4) Neisseria **mucosa**

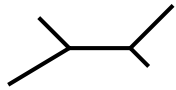


$P(S_t|\mathcal{D})$: Marginal posterior probability





$$w = \alpha t$$

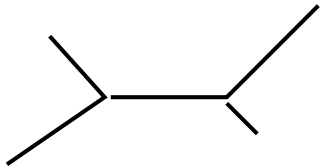


$$\alpha \rightarrow r^- \alpha$$

negative selective pressure

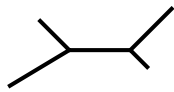
$$w \rightarrow r^- w$$

$$0 < r^- < 1$$



$$w = \alpha t$$

reference ("neutral") state

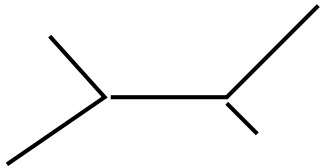


$$\alpha \rightarrow r^- \alpha$$

negative selective pressure

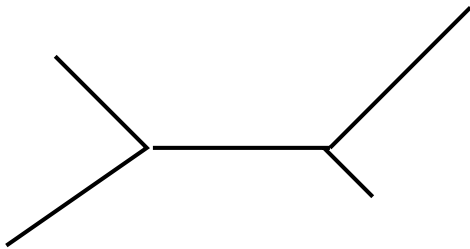
$$w \rightarrow r^- w$$

$$0 < r^- < 1$$



$$w = \alpha t$$

reference ("neutral") state



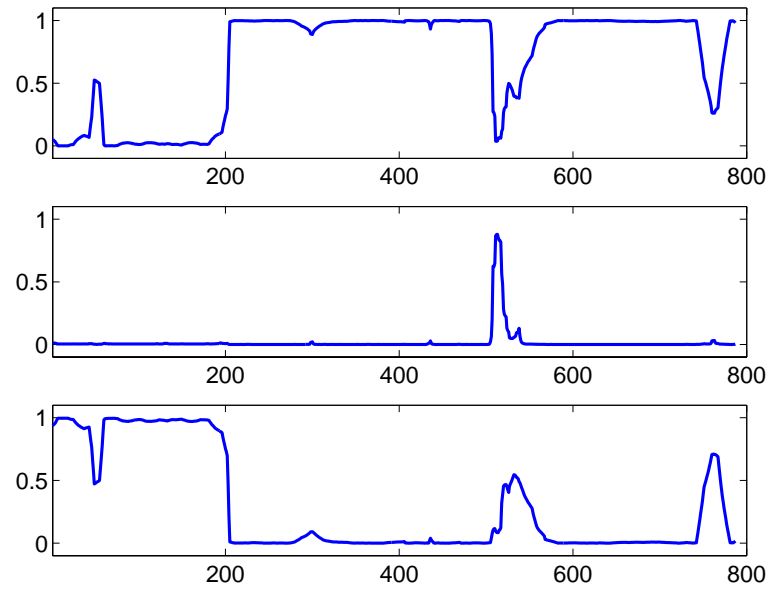
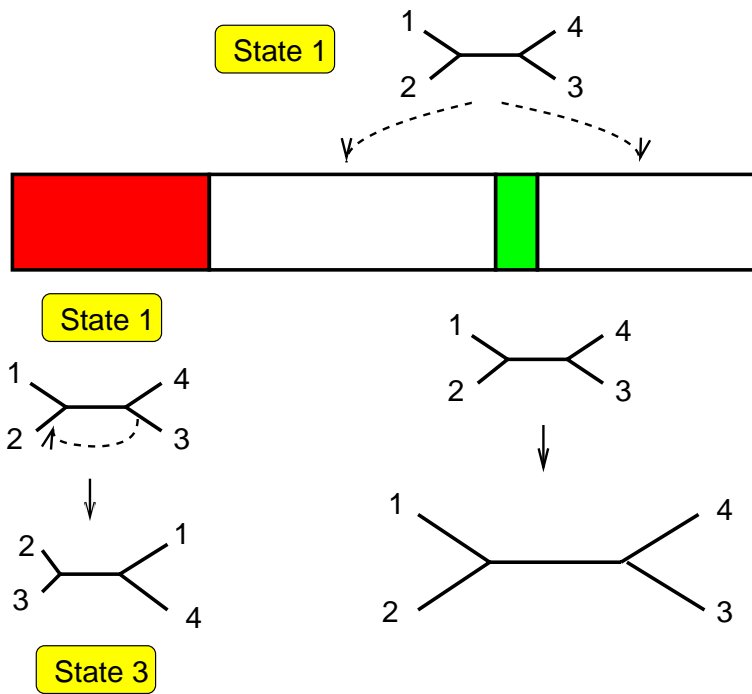
$$\alpha \rightarrow r^+ \alpha$$

positive selective pressure

$$w \rightarrow r^+ w$$

$$r^+ > 1$$

$P(S_t|\mathcal{D})$: Marginal posterior probability



Problem:

Model cannot distinguish between **recombination** and **rate variation** .

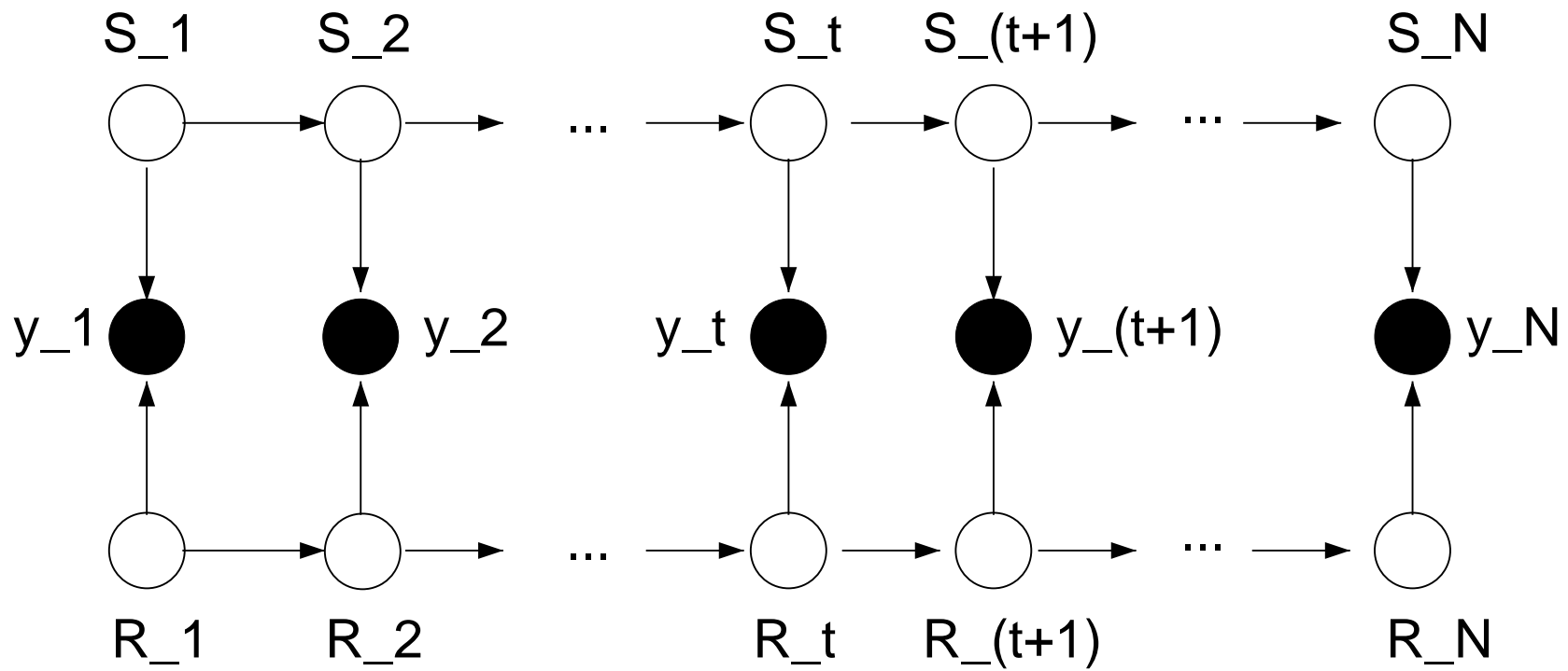
Challenge

Distinguish between
recombination
and
rate heterogeneity

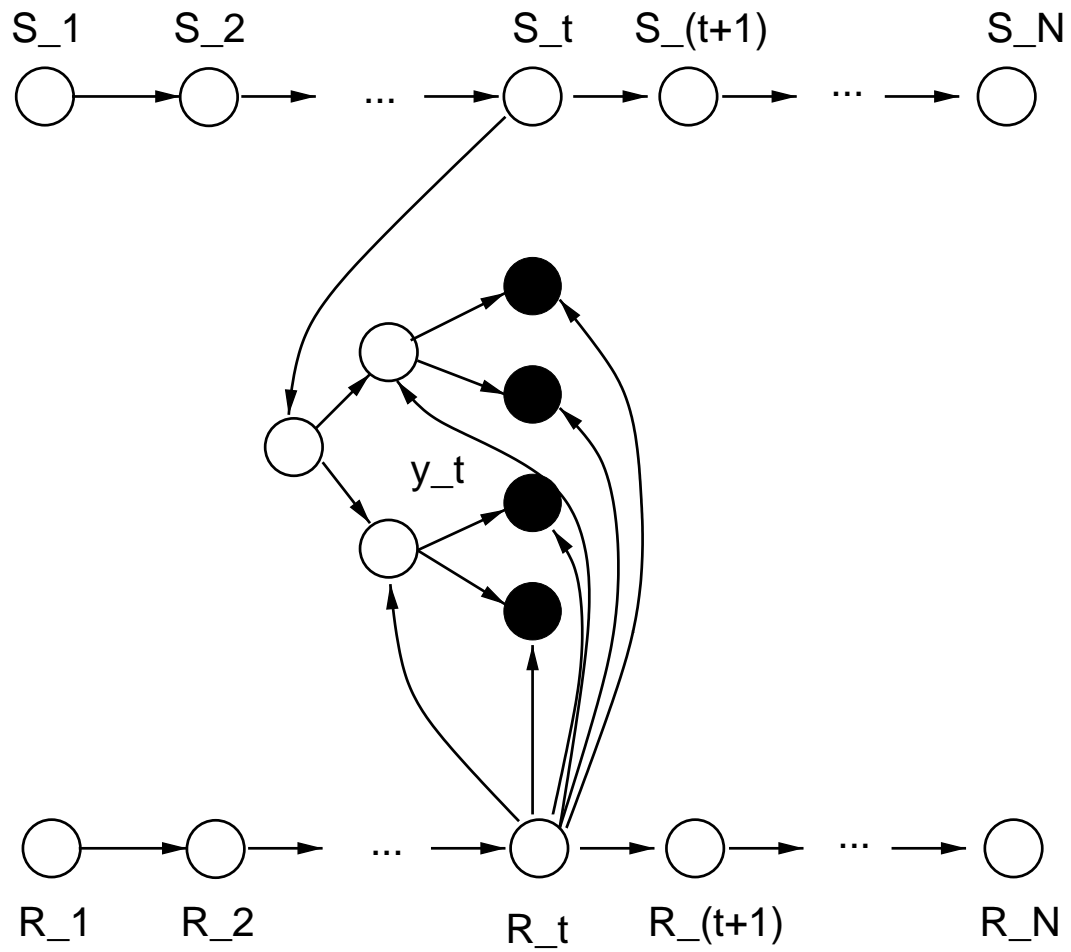
-
- PLATO
 - Window methods
 - RecPars
 - Phylo-HMMs
 - Phylo-FHMMs

Distinguishing between
recombination
and
rate variation
with
factorial hidden Markov models (FHMMs)

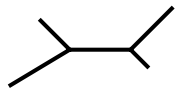
Factorial hidden Markov model (FHMM)



Phylo-FHMM



Rate states

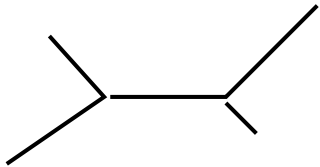


$$R = R^-$$

negative selective pressure

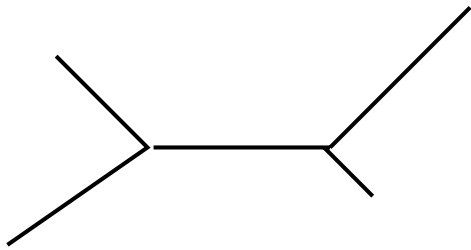
$$w \rightarrow r^- w$$

$$0 < r^- < 1$$



$$R = R^0$$

reference ("neutral") state



$$R = R^+$$

positive selective pressure

$$w \rightarrow r^+ w$$

$$r^+ > 1$$

Parameters

- Topology state sequences:

$$\mathbf{S} = (S_1, \dots, S_N)$$

- Rate state sequences:

$$\mathbf{R} = (R_1, \dots, R_N)$$

- Rate variation parameters:

$$\mathbf{r} = (r_1, \dots, r_N)$$

- Branch lengths:

$$\mathbf{w}$$

- Transition probability parameters:

$$\nu_S, \nu_R$$

Sampling from the posterior distribution

- Sampling from

$$P(\mathbf{S}, \mathbf{R}, \mathbf{r}, \mathbf{w}, \nu_S, \nu_R | \mathcal{D})$$

Sampling from the posterior distribution

- Sampling from
 $P(\mathbf{S}, \mathbf{R}, \mathbf{r}, \mathbf{w}, \nu_S, \nu_R | \mathcal{D})$
- Gibbs sampling

Sampling from the posterior distribution

- Sampling from

$$P(\mathbf{S}, \mathbf{R}, \mathbf{r}, \mathbf{w}, \nu_S, \nu_R | \mathcal{D})$$

- Gibbs sampling

- $\mathbf{S} \sim P(\mathbf{S} | \mathbf{R}, \mathbf{r}, \mathbf{w}, \nu_S, \nu_R, \mathcal{D})$

Sampling from the posterior distribution

- Sampling from

$$P(\mathbf{S}, \mathbf{R}, \mathbf{r}, \mathbf{w}, \nu_S, \nu_R | \mathcal{D})$$

- Gibbs sampling

- $\mathbf{S} \sim P(\mathbf{S} | \mathbf{R}, \mathbf{r}, \mathbf{w}, \nu_S, \nu_R, \mathcal{D})$

- $\mathbf{R} \sim P(\mathbf{R} | \mathbf{S}, \mathbf{r}, \mathbf{w}, \nu_S, \nu_R, \mathcal{D})$

Sampling from the posterior distribution

- Sampling from $P(\mathbf{S}, \mathbf{R}, \mathbf{r}, \mathbf{w}, \nu_S, \nu_R | \mathcal{D})$
- Gibbs sampling
 - $\mathbf{S} \sim P(\mathbf{S} | \mathbf{R}, \mathbf{r}, \mathbf{w}, \nu_S, \nu_R, \mathcal{D})$
 - $\mathbf{R} \sim P(\mathbf{R} | \mathbf{S}, \mathbf{r}, \mathbf{w}, \nu_S, \nu_R, \mathcal{D})$
 - ...

Sampling from the posterior distribution

- Sampling from $P(\mathbf{S}, \mathbf{R}, \mathbf{r}, \mathbf{w}, \nu_S, \nu_R | \mathcal{D})$
- Gibbs sampling
 - $\mathbf{S} \sim P(\mathbf{S} | \mathbf{R}, \mathbf{r}, \mathbf{w}, \nu_S, \nu_R, \mathcal{D})$
 - $\mathbf{R} \sim P(\mathbf{R} | \mathbf{S}, \mathbf{r}, \mathbf{w}, \nu_S, \nu_R, \mathcal{D})$
 - ...
- ν_S, ν_R : Sample from Beta distribution

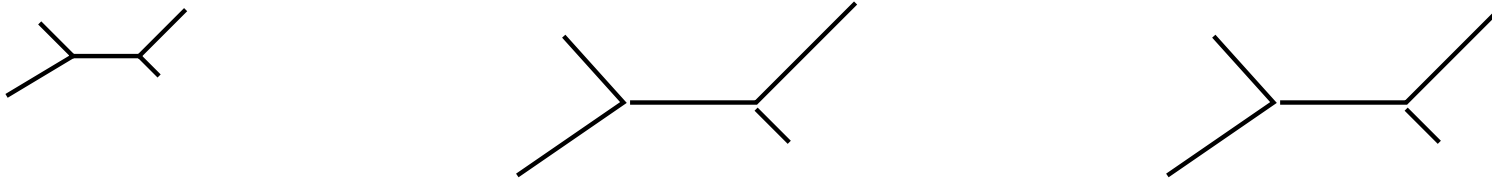
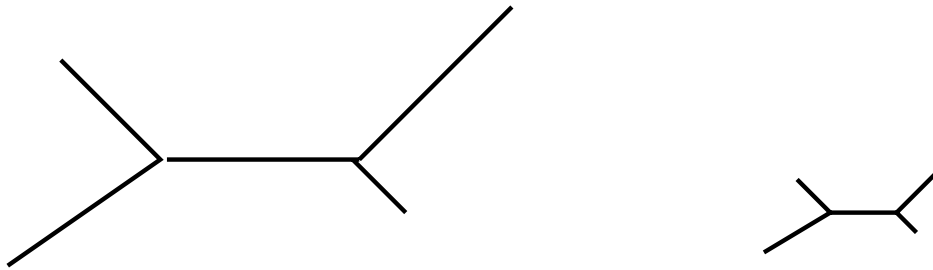
Sampling from the posterior distribution

- Sampling from $P(\mathbf{S}, \mathbf{R}, \mathbf{r}, \mathbf{w}, \nu_S, \nu_R | \mathcal{D})$
- Gibbs sampling
 - $\mathbf{S} \sim P(\mathbf{S} | \mathbf{R}, \mathbf{r}, \mathbf{w}, \nu_S, \nu_R, \mathcal{D})$
 - $\mathbf{R} \sim P(\mathbf{R} | \mathbf{S}, \mathbf{r}, \mathbf{w}, \nu_S, \nu_R, \mathcal{D})$
 - ...
- ν_S, ν_R : Sample from Beta distribution
- \mathbf{S}, \mathbf{R} : Stochastic forward–backward algorithm

Sampling from the posterior distribution

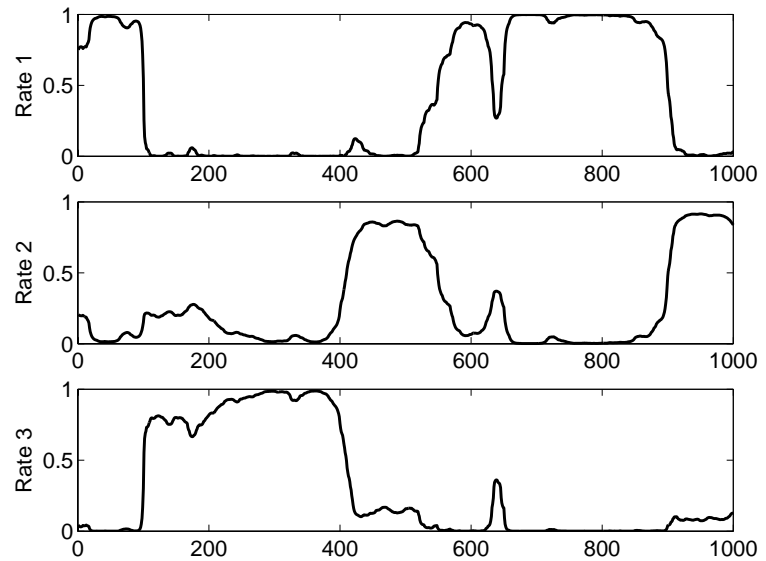
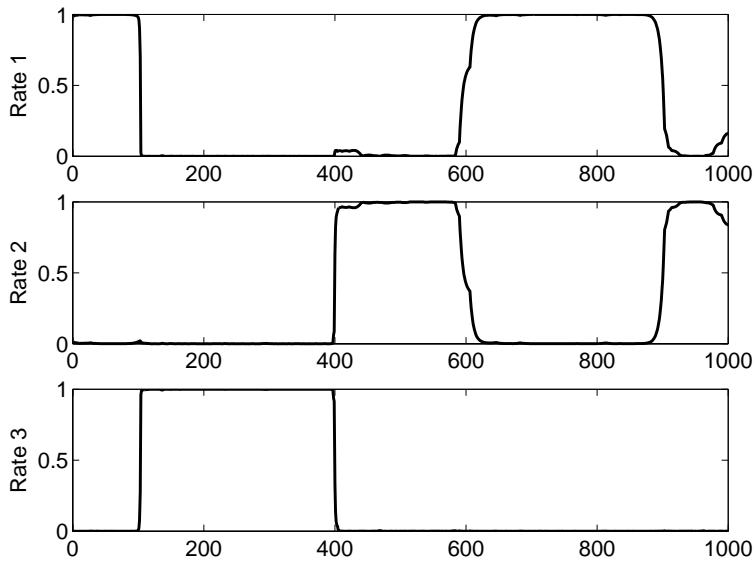
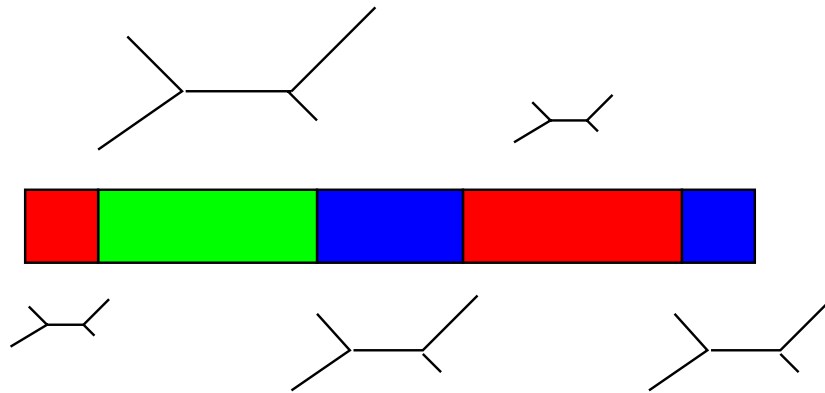
- Sampling from $P(\mathbf{S}, \mathbf{R}, \mathbf{r}, \mathbf{w}, \nu_S, \nu_R | \mathcal{D})$
- Gibbs sampling
 - $\mathbf{S} \sim P(\mathbf{S} | \mathbf{R}, \mathbf{r}, \mathbf{w}, \nu_S, \nu_R, \mathcal{D})$
 - $\mathbf{R} \sim P(\mathbf{R} | \mathbf{S}, \mathbf{r}, \mathbf{w}, \nu_S, \nu_R, \mathcal{D})$
 - ...
- ν_S, ν_R : Sample from Beta distribution
- \mathbf{S}, \mathbf{R} : Stochastic forward–backward algorithm
- \mathbf{w}, \mathbf{r} : Metropolis-Hastings

Mark Chiang: Simulated rate variation



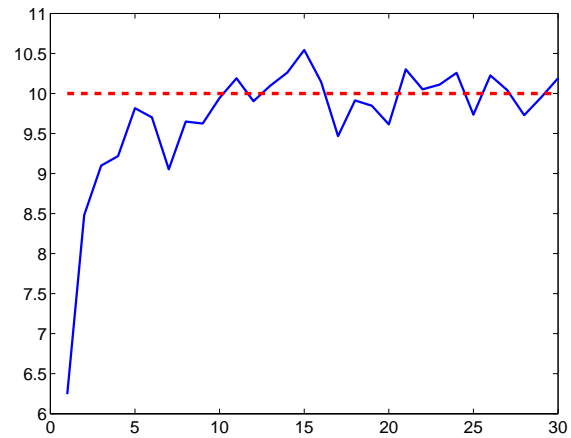
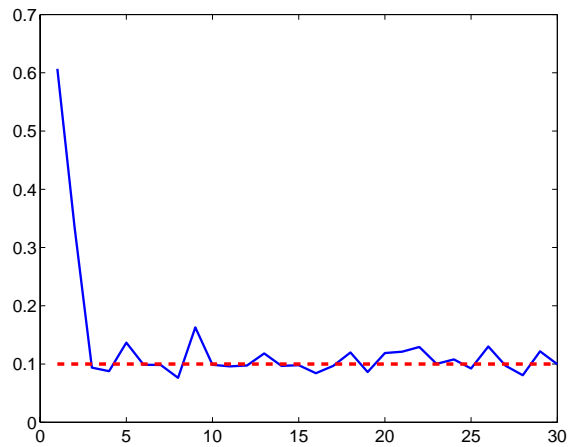
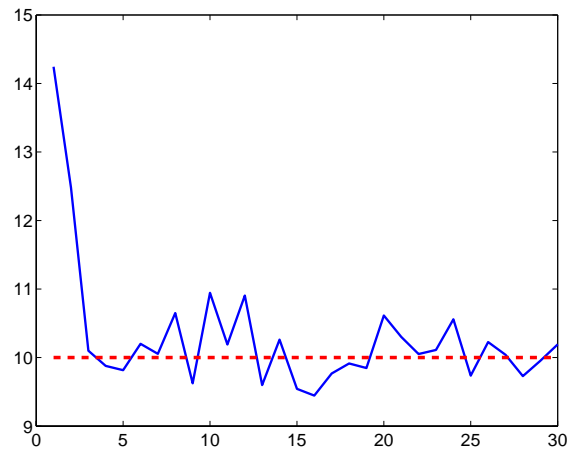
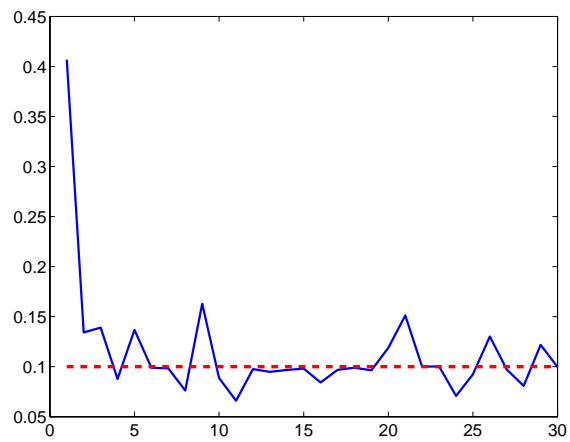
	Negative selection (R^-)	Positive selection (R^+)
Simulation 1	$r^- = 0.1$	$r^+ = 10$
Simulation 2	$r^- = 0.5$	$r^+ = 2$

Predicted rate states: $P(R_t|D)$. Left: $r = (0.1, 10)$. Right: $r = (0.5, 2)$



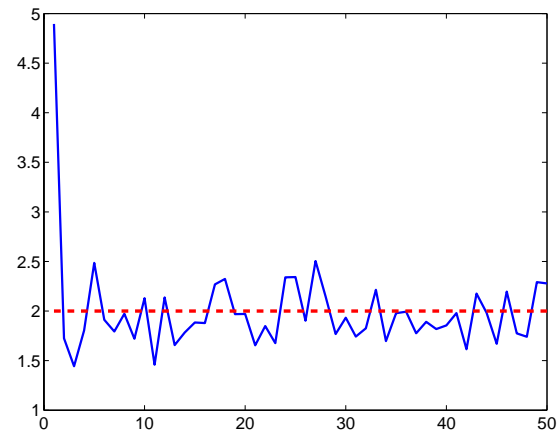
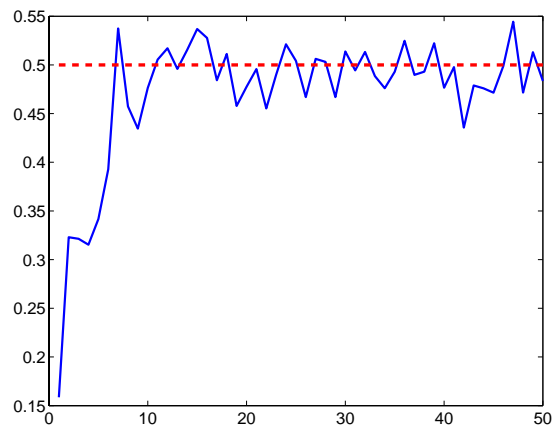
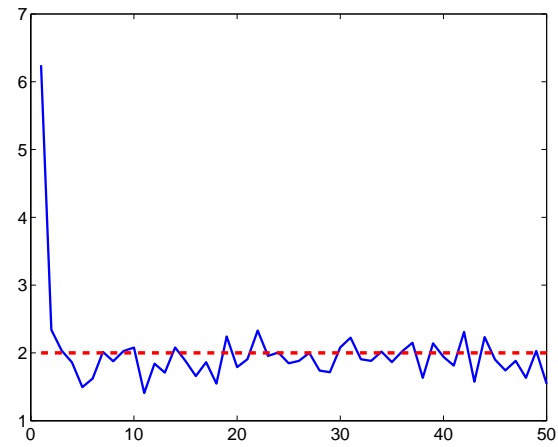
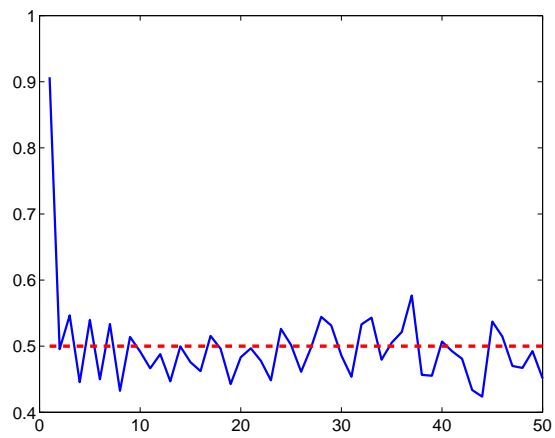
Evolution of rate factors r^- , r^+

30,000 MCMC steps. True values: $r^- = 0.1$, $r^+ = 10$

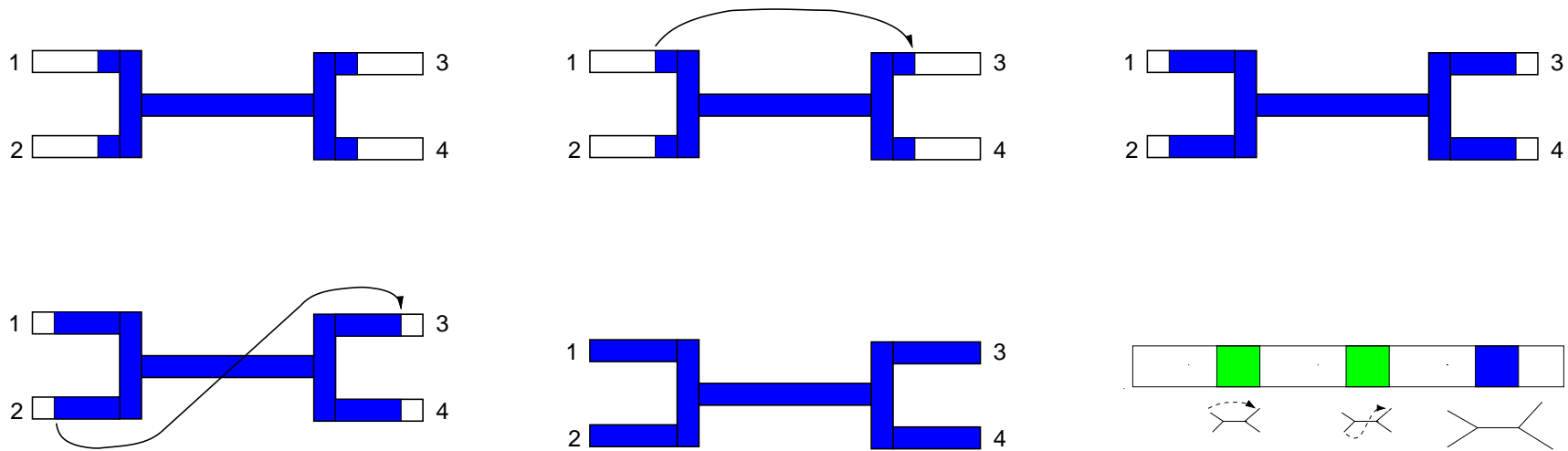


Evolution of rate factors r^- , r^+

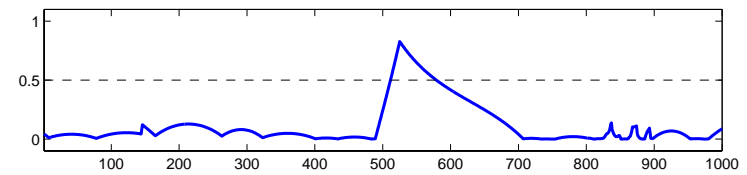
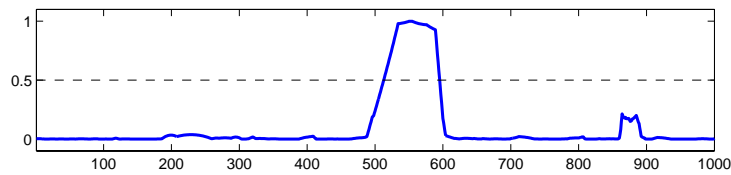
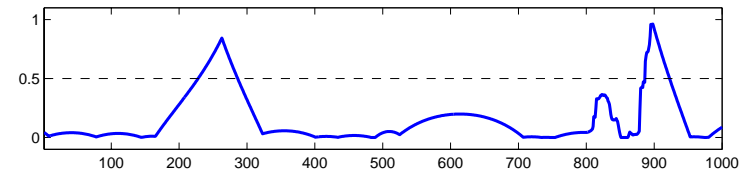
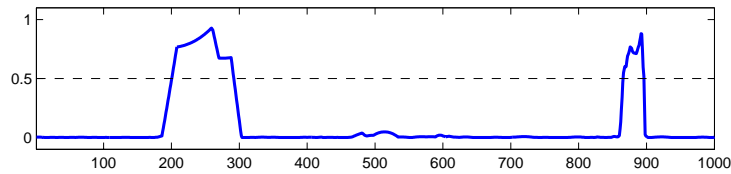
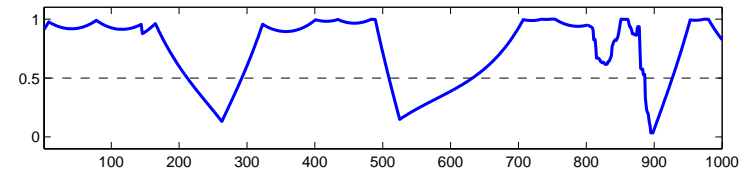
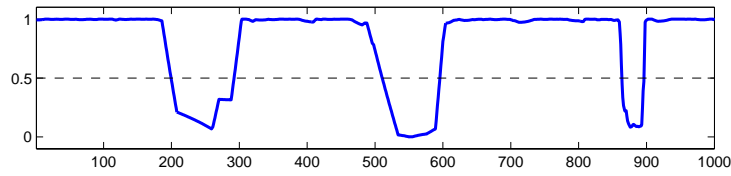
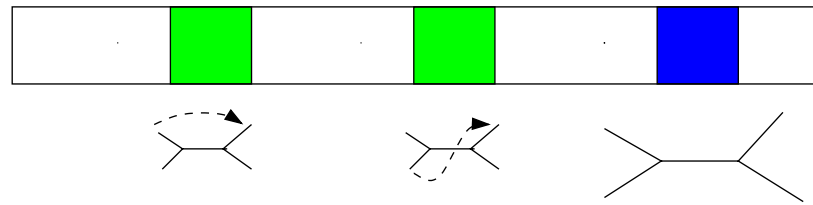
50,000 MCMC steps. True values: $r^- = 0.5$, $r^+ = 2$



Synthetic simulation study



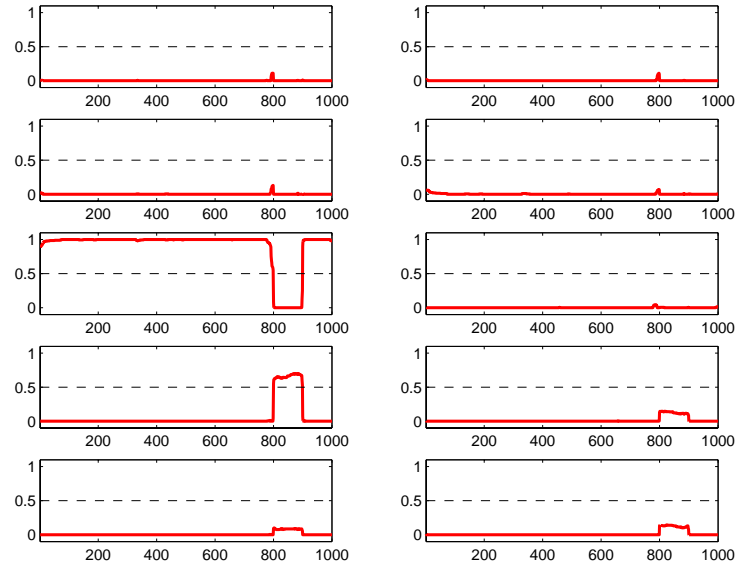
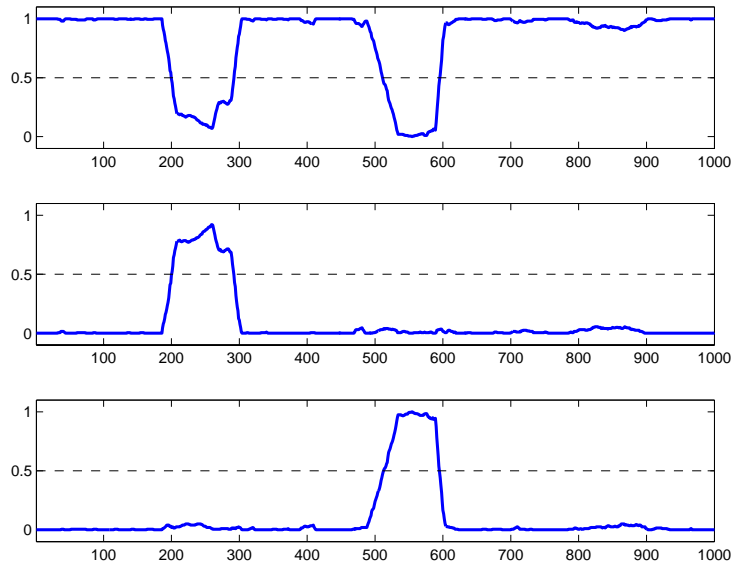
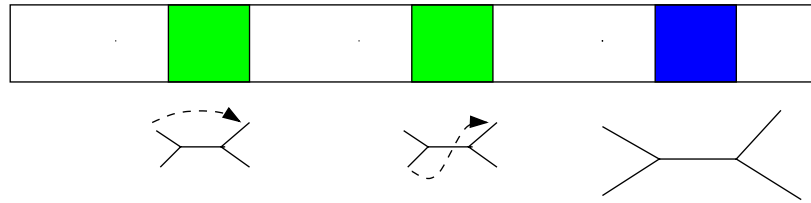
HMM, synthetic data. Left: long branches. Right: short branches



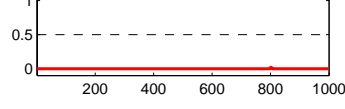
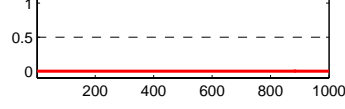
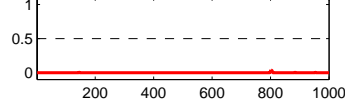
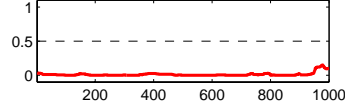
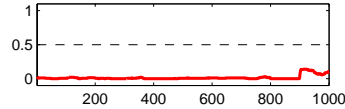
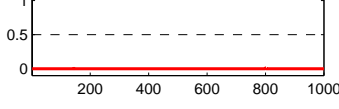
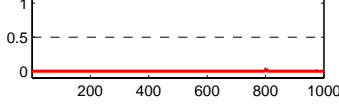
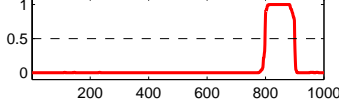
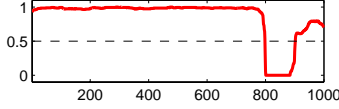
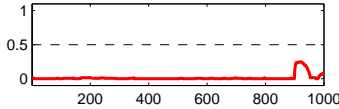
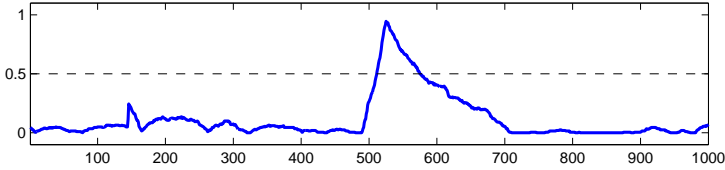
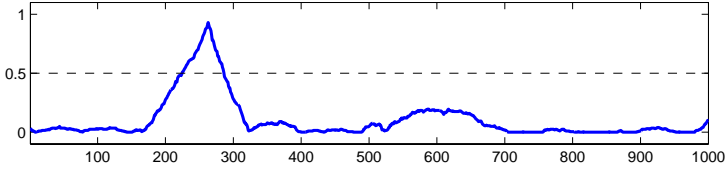
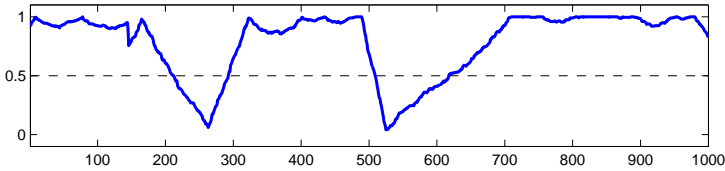
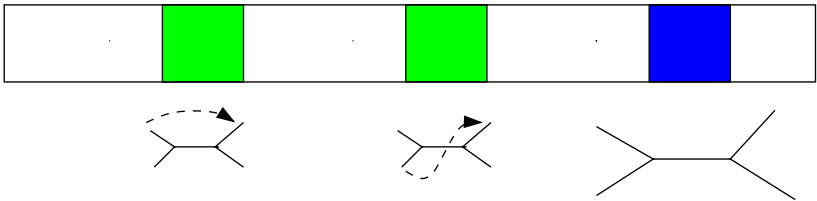
FHMM

- 10 rate states
- Fixed rate factors
- Between $r = 0.001$ and $r = 100$ approximately uniform on a log scale

FHMM, synthetic data, long branch lengths



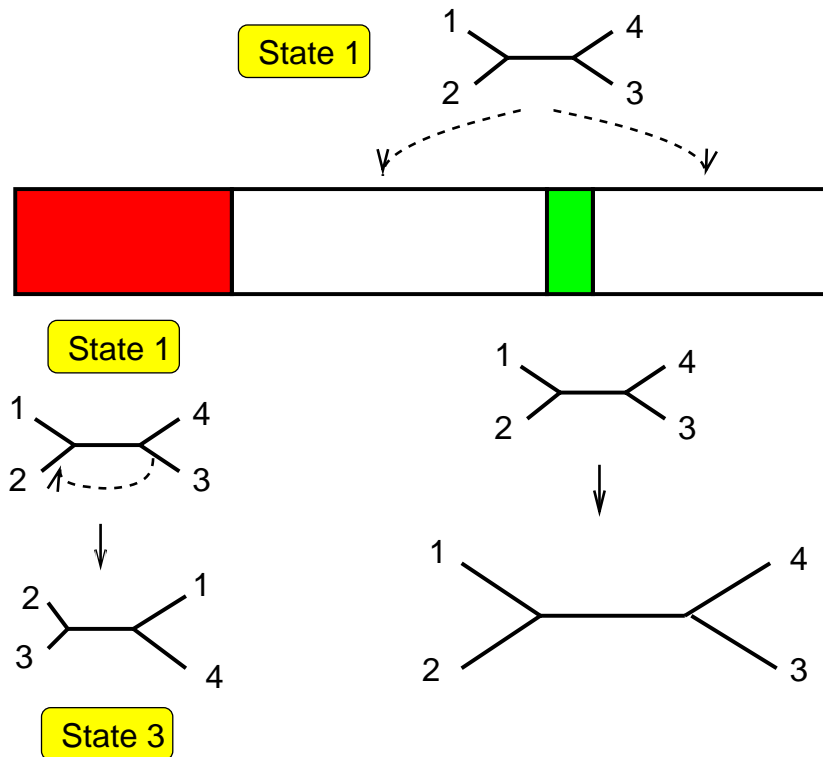
FHMM, synthetic data, short branch lengths



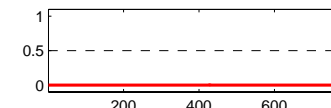
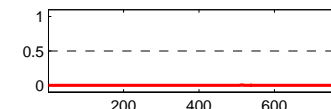
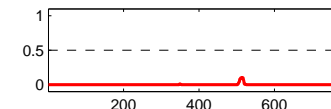
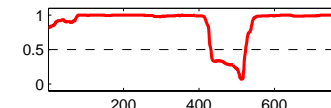
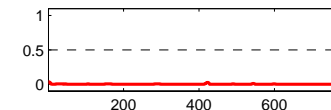
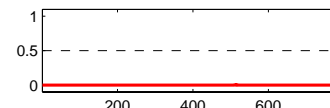
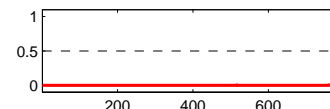
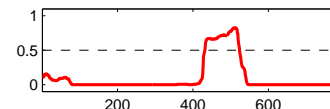
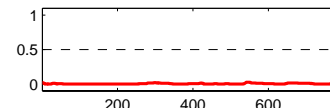
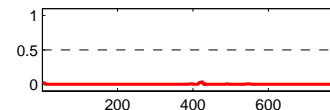
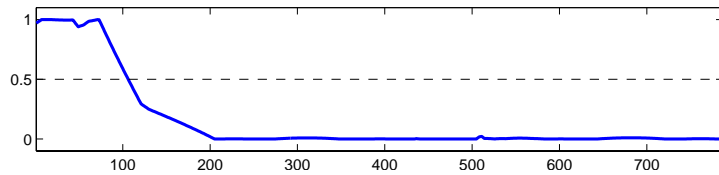
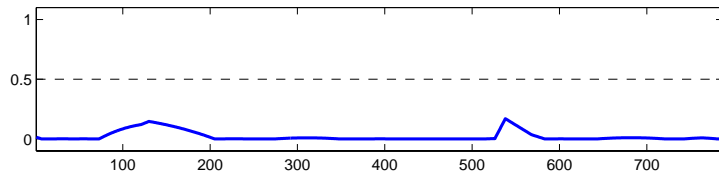
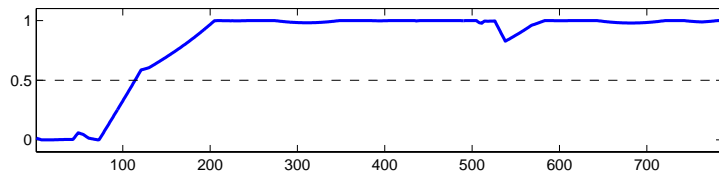
Neisseria (Zhou & Spratt, 1992)

DNA alignment, 787 nucleotides (argF gene)

- 1) Neisseria **gonorrhoeae**
- 2) Neisseria **meningitidis**
- 3) Neisseria **cinerea**
- 4) Neisseria **mucosa**



FHMM, Neisseria



Future work:

RJMCMC

Future work:

RJMCMC

Green (1995)
Biometrika 82, 711-732

Robert, Ryden, Titterington (2000)
J. R. Statist. Soc. B, 62, 57-7

Boys, Henderson (2001)
Comp. Sci. and Statist., 33, 35-49

Suchard, Weiss, Dormin, Sinsheimer (2003)
J. Am. Statist. Assoc. 98, 427-437

RJMCMC

Before sampling a new sequence of rate states \mathbf{R} ,
do one of three possible moves:

RJMCMC

Before sampling a new sequence of rate states \mathbf{R} ,
do one of three possible moves:

1) Birth of a new rate state

RJMCMC

Before sampling a new sequence of rate states \mathbf{R} ,
do one of three possible moves:

- 1) Birth of a new rate state
- 2) Death of an existing rate state

RJMCMC

Before sampling a new sequence of rate states \mathbf{R} ,
do one of three possible moves:

- 1) Birth of a new rate state
- 2) Death of an existing rate state
- 3) Relocation of a rate factor

RJMCMC

Before sampling a new sequence of rate states \mathbf{R} ,
do one of three possible moves:

- 1) Birth of a new rate state
- 2) Death of an existing rate state
- 3) Relocation of a rate factor

Acceptance probability=

$$\begin{aligned} & \text{Likelihood ratio} \times \\ & \text{Prior ratio} \times \\ & \text{Inverse proposal probability ratio} \times \\ & \text{Jacobian} \end{aligned}$$

Prior probabilities

Prior probabilities

Number of rate states k : Truncated Poisson distribution

$$P(k) \propto \frac{\lambda^k}{k!} I(k \leq k_{max}) \Rightarrow \frac{P(k+1)}{P(k)} = \frac{\lambda}{k+1} I(k+1 \leq k_{max})$$

Prior probabilities

Number of rate states k : Truncated Poisson distribution

$$P(k) \propto \frac{\lambda^k}{k!} I(k \leq k_{max}) \Rightarrow \frac{P(k+1)}{P(k)} = \frac{\lambda}{k+1} I(k+1 \leq k_{max})$$

Rate factors

Define $\rho = \log r$. Given k , the logarithmic rate factors are distributed as the even-numbered order statistics from $2k+1$ points uniformly distributed on $[L_{min}, L_{max}]$. $L = L_{max} - L_{min}$.

Prior probabilities

Number of rate states k : Truncated Poisson distribution

$$P(k) \propto \frac{\lambda^k}{k!} I(k \leq k_{max}) \Rightarrow \frac{P(k+1)}{P(k)} = \frac{\lambda}{k+1} I(k+1 \leq k_{max})$$

Rate factors

Define $\rho = \log r$. Given k , the logarithmic rate factors are distributed as the even-numbered order statistics from $2k+1$ points uniformly distributed on $[L_{min}, L_{max}]$. $L = L_{max} - L_{min}$.

Relocation move $\rho_i \rightarrow \rho'_i$:
$$\frac{P(\boldsymbol{\rho}')}{P(\boldsymbol{\rho})} = \frac{(\rho_{i+1} - \rho'_i)(\rho'_i - \rho_{i-1})}{(\rho_{i+1} - \rho_i)(\rho_i - \rho_{i-1})}$$

Birth move $\boldsymbol{\rho} \rightarrow \boldsymbol{\rho}^*$:
$$\frac{P(\boldsymbol{\rho}^*)}{P(\boldsymbol{\rho})} = \frac{2(k+1)(2k+3)}{L^2} \frac{(\rho_{i+1} - \rho^*)(\rho^* - \rho_i)}{\rho_{i+1} - \rho_i}$$

Proposal probabilities

Proposal probabilities

Birth move:

$$b_k = c \min\left\{1, \frac{P(k+1)}{P(k)}\right\}$$

New log rate sampled uniformly from $[L_{min}, L_{max}]$

$$\pi_b(\rho^*) = \frac{1}{L}$$

Proposal probabilities

Birth move:

$$b_k = c \min\left\{1, \frac{P(k+1)}{P(k)}\right\}$$

New log rate sampled uniformly from $[L_{min}, L_{max}]$

$$\pi_b(\rho^*) = \frac{1}{L}$$

Death move:

$$d_k = c \min\left\{1, \frac{P(k)}{P(k+1)}\right\}$$

Deleted rate factor chosen uniformly
from the set of existing rate factors:

$$\pi_d(\rho^*) = \frac{1}{k+1}$$

Proposal probabilities

Birth move:

$$b_k = c \min\left\{1, \frac{P(k+1)}{P(k)}\right\}$$

New log rate sampled uniformly from $[L_{min}, L_{max}]$

$$\pi_b(\rho^*) = \frac{1}{L}$$

Death move:

$$d_k = c \min\left\{1, \frac{P(k)}{P(k+1)}\right\}$$

Deleted rate factor chosen uniformly from the set of existing rate factors:

$$\pi_d(\rho^*) = \frac{1}{k+1}$$

Ratio:
$$\frac{d_{k+1}\pi_d(\rho^*)}{b_k\pi_b(\rho^*)} = \frac{P(k)L}{P(k+1)(k+1)}$$

Acceptance probability for a birth move

$$\text{Likelihood ratio: } \frac{P(\mathcal{D}, \mathbf{S}, \mathbf{R}^* | k+1, \boldsymbol{\rho}^*, \nu_S, \nu_R)}{P(\mathcal{D}, \mathbf{S}, \mathbf{R} | k, \boldsymbol{\rho}, \nu_S, \nu_R)}$$

$$\text{Prior ratio: } \frac{P(k+1)}{P(k)} \frac{P(\boldsymbol{\rho}^*)}{P(\boldsymbol{\rho})} \frac{(k+1)!}{k!}$$

$$\text{Inverse proposal probability ratio: } \frac{d_{k+1} \pi_d(r^*)}{b_k \pi_b(r^*)} \frac{P(\mathbf{R} | \mathcal{D}, \mathbf{S}, k+1, \boldsymbol{\rho}, \nu_S, \nu_R)}{P(\mathbf{R}^* | \mathcal{D}, \mathbf{S}, k, \boldsymbol{\rho}^*, \nu_S, \nu_R)}$$

$$\text{Jacobian} = 1$$

Acceptance probability for a birth move

$$\text{Likelihood ratio: } \frac{P(\mathcal{D}, \mathbf{S}, \mathbf{R}^* | k+1, \boldsymbol{\rho}^*, \nu_S, \nu_R)}{P(\mathcal{D}, \mathbf{S}, \mathbf{R} | k, \boldsymbol{\rho}, \nu_S, \nu_R)}$$

$$\text{Prior ratio: } \frac{P(k+1)}{P(k)} \frac{P(\boldsymbol{\rho}^*)}{P(\boldsymbol{\rho})} \frac{(k+1)!}{k!}$$

$$\text{Inverse proposal probability ratio: } \frac{d_{k+1} \pi_d(r^*)}{b_k \pi_b(r^*)} \frac{P(\mathbf{R} | \mathcal{D}, \mathbf{S}, k+1, \boldsymbol{\rho}, \nu_S, \nu_R)}{P(\mathbf{R}^* | \mathcal{D}, \mathbf{S}, k, \boldsymbol{\rho}^*, \nu_S, \nu_R)}$$

$$\text{Jacobian} = 1$$

Acceptance probability for a birth move:

$$\frac{P(\mathcal{D}, \mathbf{S} | k+1, \boldsymbol{\rho}^*, \nu_S, \nu_R)}{P(\mathcal{D}, \mathbf{S} | k, \boldsymbol{\rho}, \nu_S, \nu_R)}$$

Acceptance probability for a birth move

$$\text{Likelihood ratio: } \frac{P(\mathcal{D}, \mathbf{S}, \mathbf{R}^* | k+1, \boldsymbol{\rho}^*, \nu_S, \nu_R)}{P(\mathcal{D}, \mathbf{S}, \mathbf{R} | k, \boldsymbol{\rho}, \nu_S, \nu_R)}$$

$$\text{Prior ratio: } \frac{P(k+1)}{P(k)} \frac{P(\boldsymbol{\rho}^*)}{P(\boldsymbol{\rho})} \frac{(k+1)!}{k!}$$

$$\text{Inverse proposal probability ratio: } \frac{d_{k+1} \pi_d(r^*)}{b_k \pi_b(r^*)} \frac{P(\mathbf{R} | \mathcal{D}, \mathbf{S}, k+1, \boldsymbol{\rho}, \nu_S, \nu_R)}{P(\mathbf{R}^* | \mathcal{D}, \mathbf{S}, k, \boldsymbol{\rho}^*, \nu_S, \nu_R)}$$

$$\text{Jacobian} = 1$$

Acceptance probability for a birth move:

$$\frac{P(\mathcal{D}, \mathbf{S} | k+1, \boldsymbol{\rho}^*, \nu_S, \nu_R)}{P(\mathcal{D}, \mathbf{S} | k, \boldsymbol{\rho}, \nu_S, \nu_R)}$$

Acceptance probability for a birth move

$$\text{Likelihood ratio: } \frac{P(\mathcal{D}, \mathbf{S}, \mathbf{R}^* | k+1, \boldsymbol{\rho}^*, \nu_S, \nu_R)}{P(\mathcal{D}, \mathbf{S}, \mathbf{R} | k, \boldsymbol{\rho}, \nu_S, \nu_R)}$$

$$\text{Prior ratio: } \frac{P(k+1)}{P(k)} \frac{P(\boldsymbol{\rho}^*)}{P(\boldsymbol{\rho})} \frac{(k+1)!}{k!}$$

$$\text{Inverse proposal probability ratio: } \frac{d_{k+1} \pi_d(r^*)}{b_k \pi_b(r^*)} \frac{P(\mathbf{R} | \mathcal{D}, \mathbf{S}, k+1, \boldsymbol{\rho}, \nu_S, \nu_R)}{P(\mathbf{R}^* | \mathcal{D}, \mathbf{S}, k, \boldsymbol{\rho}^*, \nu_S, \nu_R)}$$

$$\text{Jacobian} = 1$$

Acceptance probability for a birth move:

$$\frac{P(\mathcal{D}, \mathbf{S} | k+1, \boldsymbol{\rho}^*, \nu_S, \nu_R)}{P(\mathcal{D}, \mathbf{S} | k, \boldsymbol{\rho}, \nu_S, \nu_R)} \frac{2(k+1)(2k+3)}{L} \frac{(\rho_{i+1} - \rho^*)(\rho^* - \rho_i)}{\rho_{i+1} - \rho_i}$$

Acceptance probability for a birth move

$$\text{Likelihood ratio: } \frac{P(\mathcal{D}, \mathbf{S}, \mathbf{R}^* | k+1, \boldsymbol{\rho}^*, \nu_S, \nu_R)}{P(\mathcal{D}, \mathbf{S}, \mathbf{R} | k, \boldsymbol{\rho}, \nu_S, \nu_R)}$$

$$\text{Prior ratio: } \frac{P(k+1)}{P(k)} \frac{P(\boldsymbol{\rho}^*)}{P(\boldsymbol{\rho})} \frac{(k+1)!}{k!}$$

$$\text{Inverse proposal probability ratio: } \frac{d_{k+1} \pi_d(r^*)}{b_k \pi_b(r^*)} \frac{P(\mathbf{R} | \mathcal{D}, \mathbf{S}, k+1, \boldsymbol{\rho}, \nu_S, \nu_R)}{P(\mathbf{R}^* | \mathcal{D}, \mathbf{S}, k, \boldsymbol{\rho}^*, \nu_S, \nu_R)}$$

$$\text{Jacobian} = 1$$

Acceptance probability for a birth move:

$$\frac{P(\mathcal{D}, \mathbf{S} | k+1, \boldsymbol{\rho}^*, \nu_S, \nu_R)}{P(\mathcal{D}, \mathbf{S} | k, \boldsymbol{\rho}, \nu_S, \nu_R)} \frac{2(k+1)(2k+3)}{L} \frac{(\rho_{i+1} - \rho^*)(\rho^* - \rho_i)}{\rho_{i+1} - \rho_i}$$

Conclusion: Phylo-HMMs and Phylo-FHMMs

- Probabilistic equivalent to RecPars.
- All parameters can be inferred from the data.
- No window needed.
- More precise location of the breakpoints.
- Phylo-FHMM can distinguish between recombination and rate variation.
- Can currently only deal with a small number of species.

Husmeier D., Dybowski R., and Roberts S. (2005)

Probabilistic Modeling in Bioinformatics
and Medical Informatics

Springer Verlag, New York