

---

# Detecting Recombination in DNA Sequence Alignments

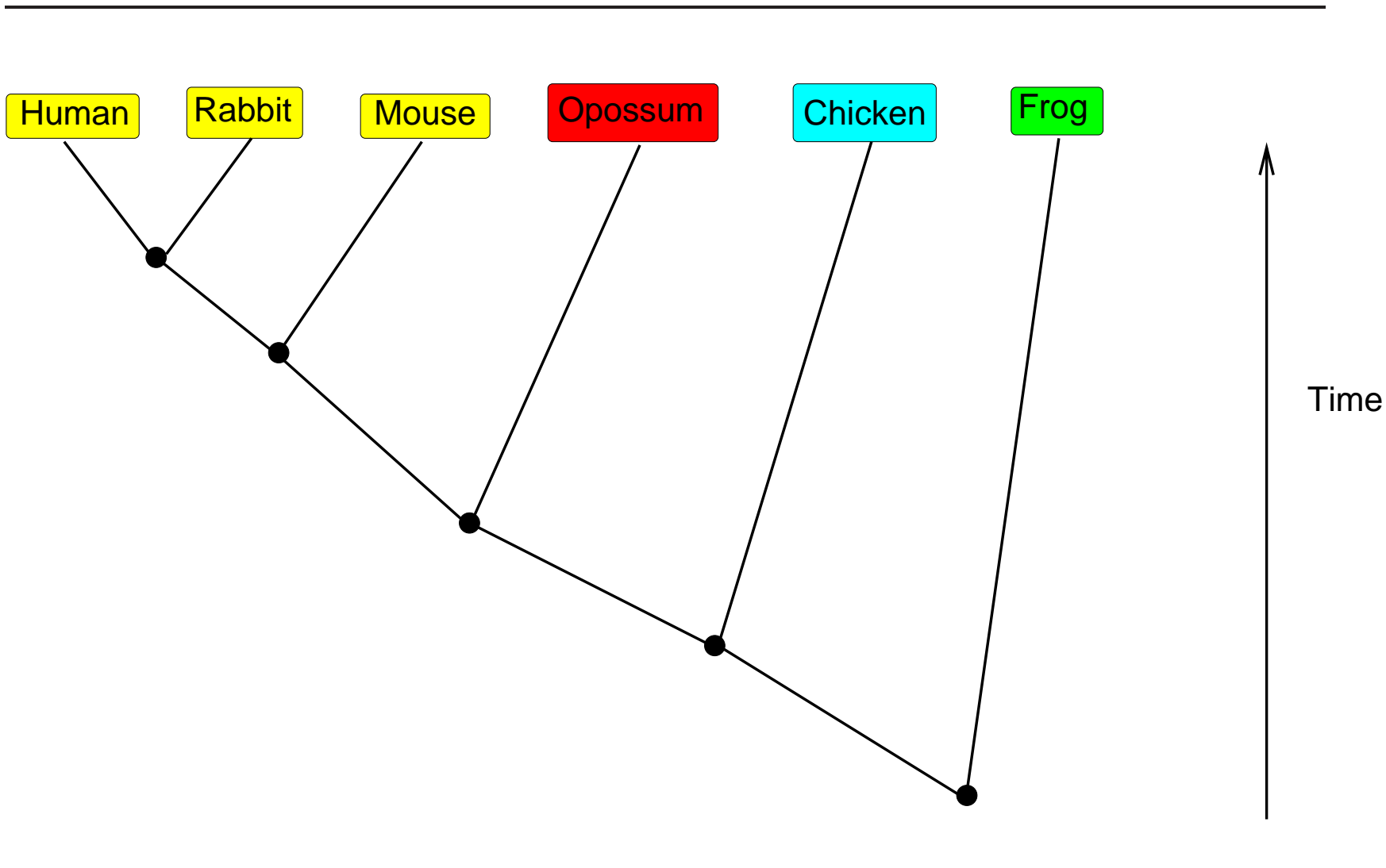
Dirk Husmeier

Biomathematics & Statistics Scotland (BioSS)  
JCMB, The King's Buildings, Edinburgh EH9 3JZ

Email: [dirk@bioess.ac.uk](mailto:dirk@bioess.ac.uk)

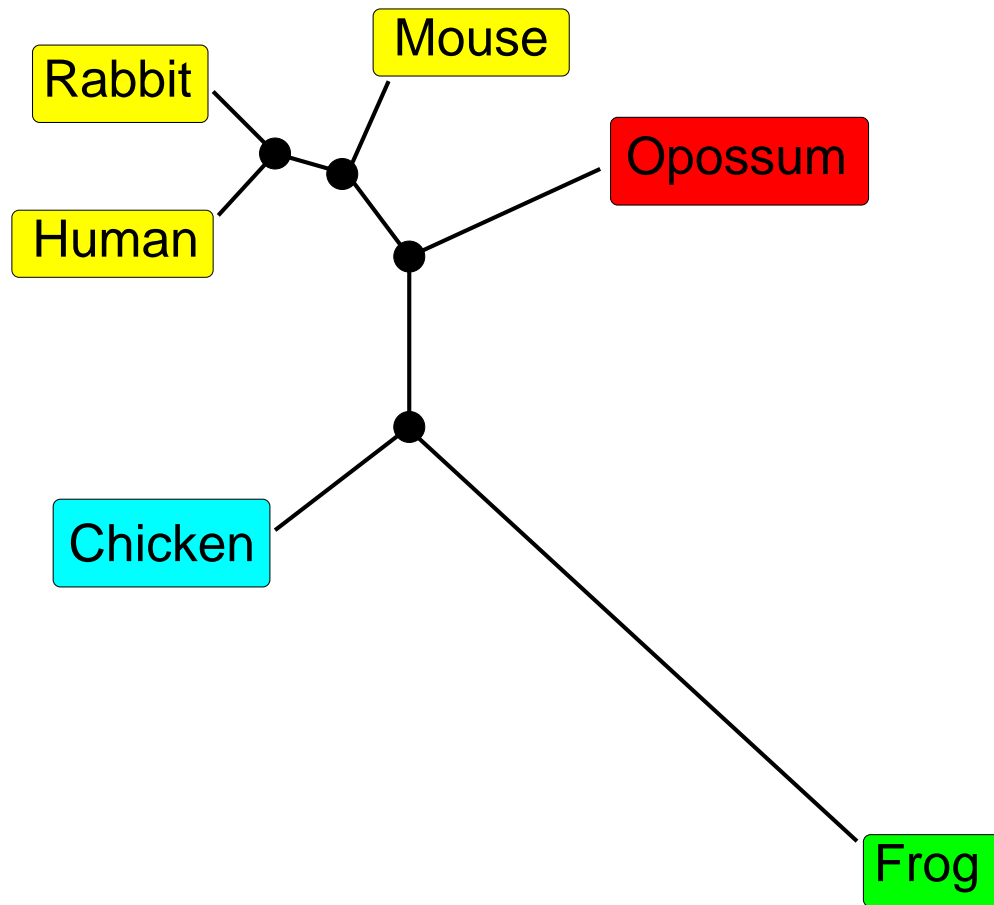
<http://www.bioess.ac.uk/~dirk>

# Phylogenetics



# Phylogenetics

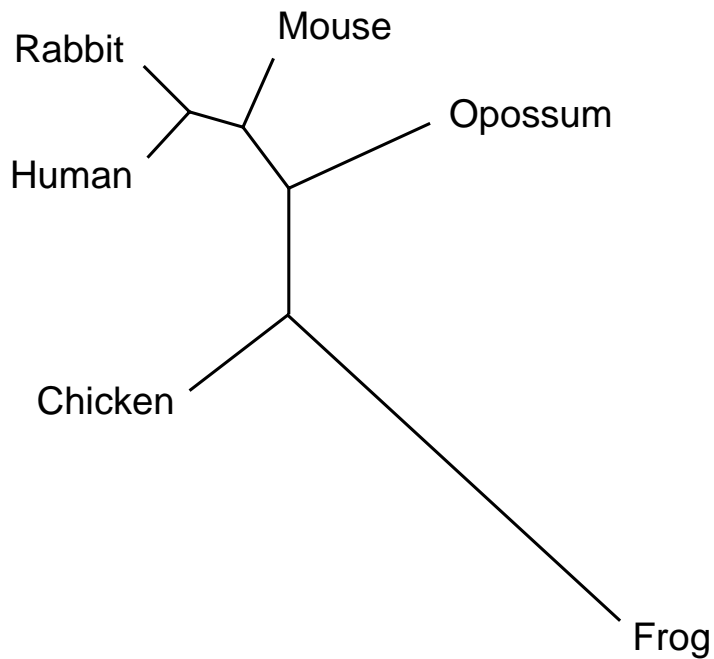
---



--> Topology  
--> Branch lengths

# Phylogenetics

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

---

- Clustering
- Maximum parsimony
- Maximum likelihood

## Inferring phylogeny from pairwise distances

---

Human ... T G T **A** T C G C T C ...  
Rabbit ... T G T **G** T C G C T C ...

---

## Inferring phylogeny from pairwise distances

---

Human ... T G T **A** T C G C T C ...  
Rabbit ... T G T **G** T C G C T C ...

Human ... **T** G T **A** T C G **C** T C ...  
Chicken ... **A** G T **C** T C G **T** T C ...

---

## Inferring phylogeny from pairwise distances

---

Human ... T G T **A** T C G C T C ...  
Rabbit ... T G T **G** T C G C T C ...

Human ... **T** G T **A** T C G **C** T C ...  
Chicken ... **A** G T **C** T C G **T** T C ...

Rabbit ... **T** G T **G** T C G **C** T C ...  
Chicken ... **A** G T **C** T C G **T** T C ...

---

## Inferring phylogeny from pairwise distances

---

Human ... T G T A T C G C T C ...  
 Rabbit ... T G T G T C G C T C ...

Human ... T G T A T C G C T C ...  
 Chicken ... A G T C T C G T T C ...

Rabbit ... T G T G T C G C T C ...  
 Chicken ... A G T C T C G T T C ...

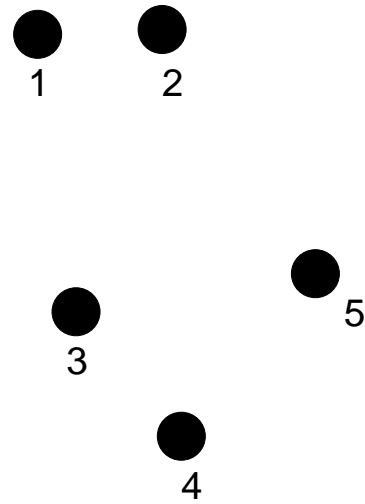
	Rabbit	Chicken
Human	1	3
Rabbit		3

---



## Inferring phylogeny by clustering: UPGMA

---

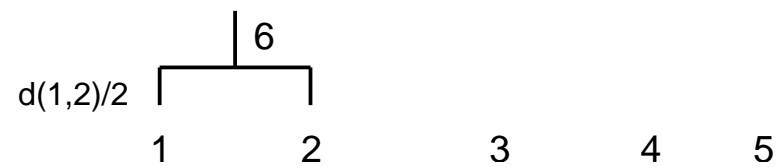
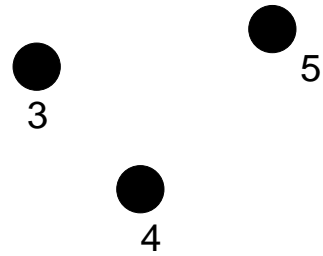
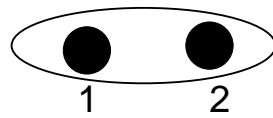


---

1 2 3 4 5

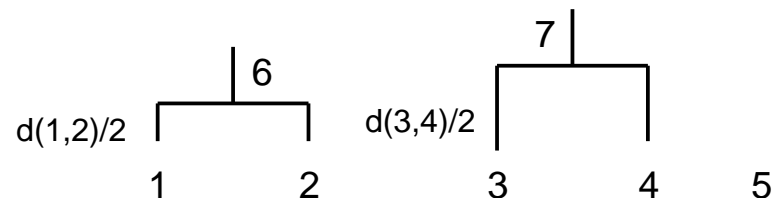
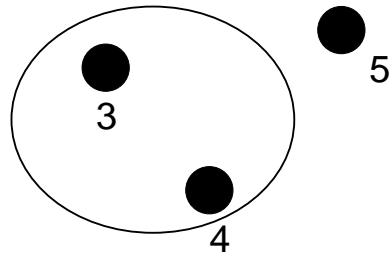
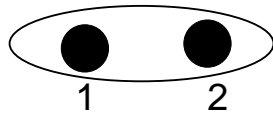
## Inferring phylogeny by clustering: UPGMA

---



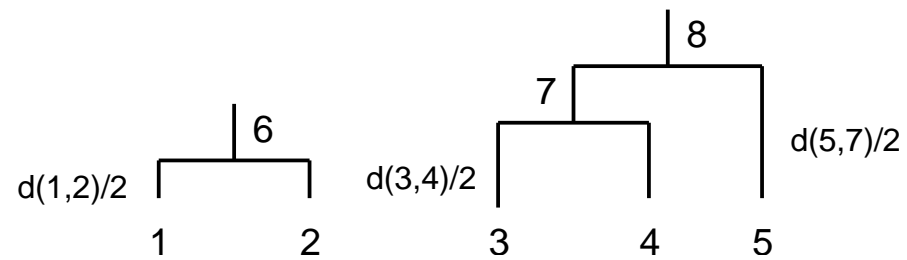
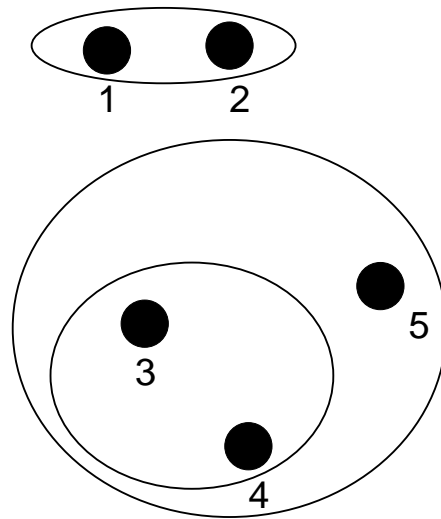
## Inferring phylogeny by clustering: UPGMA

---



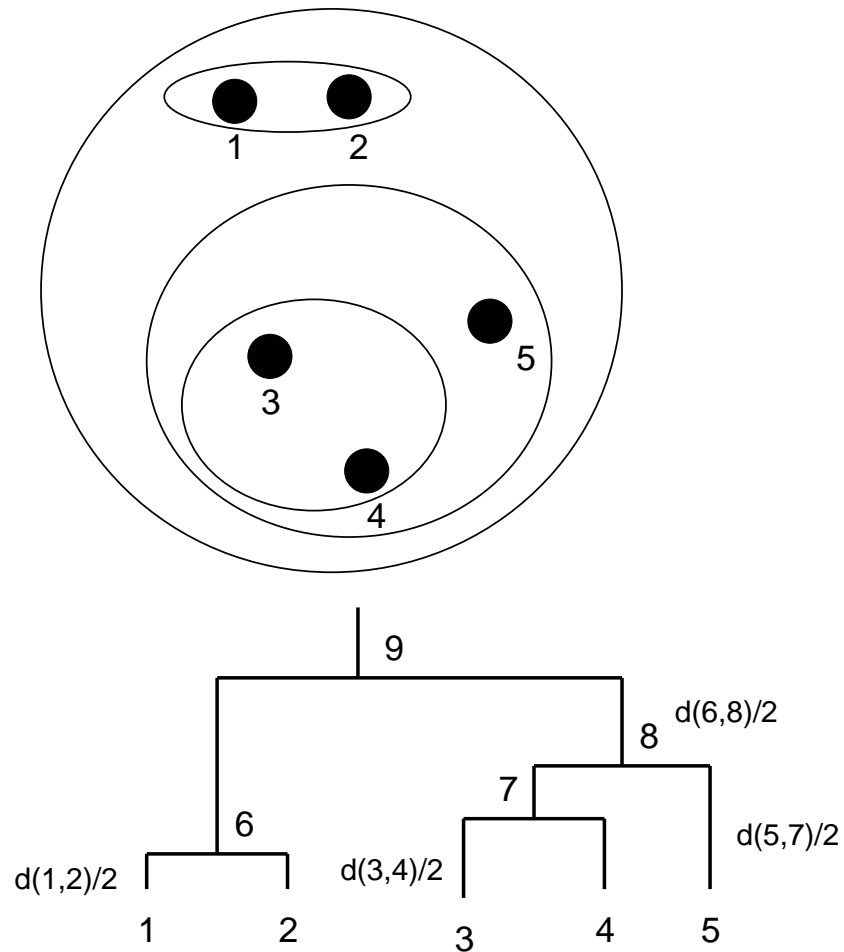
## Inferring phylogeny by clustering: UPGMA

---



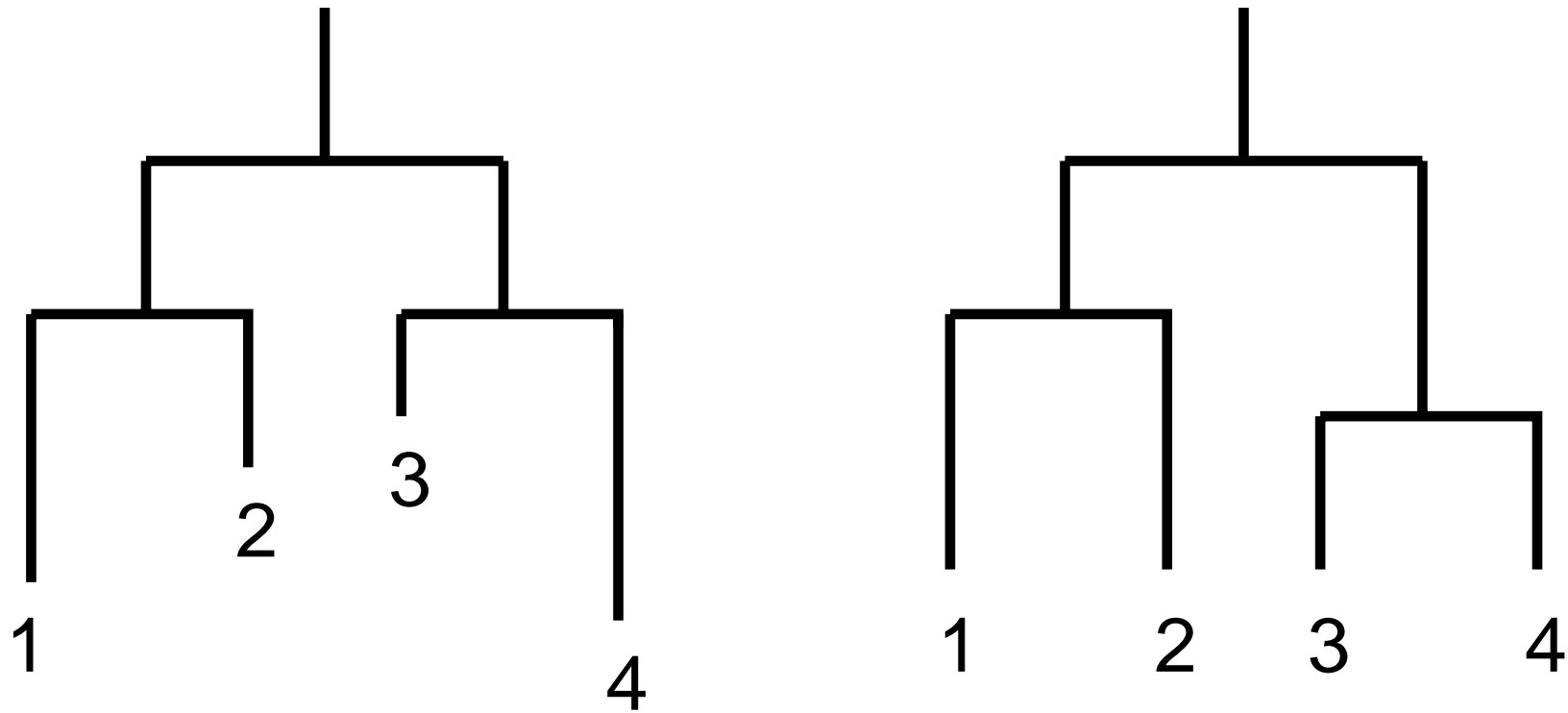
## Inferring phylogeny by clustering: UPGMA

---



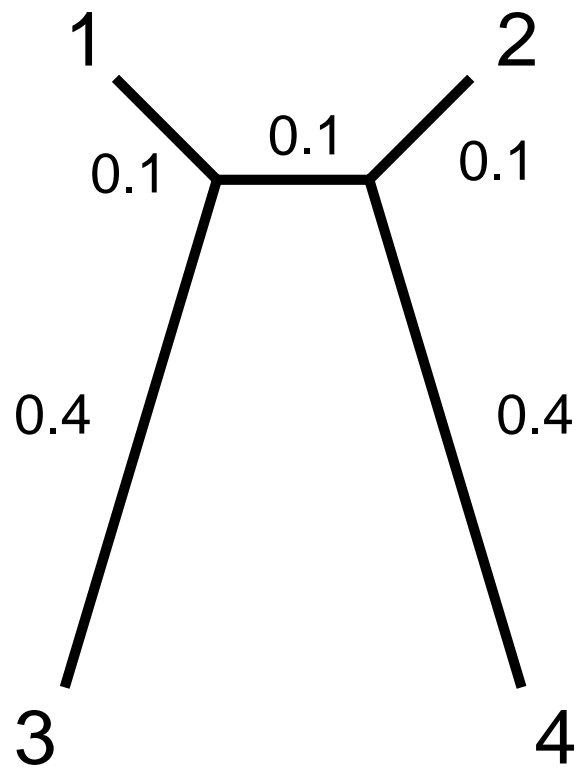
## Molecular clock constraint

---



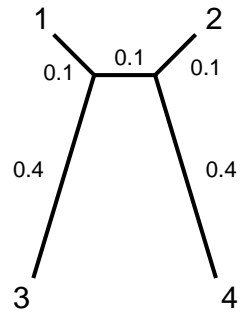
## Shortcomings of clustering

---



## Shortcomings of clustering

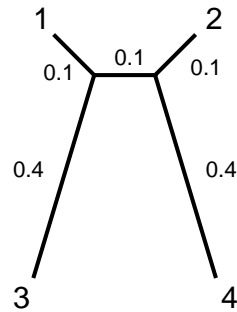
---



- Corrected distance measure (→ neighbour joining).

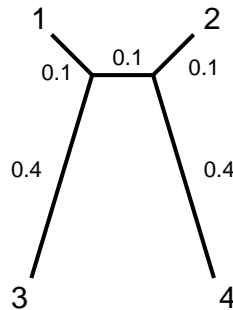
## Shortcomings of clustering

---



- Corrected distance measure (→ neighbour joining).
- Can lead to **negative distances**.

## Shortcomings of clustering

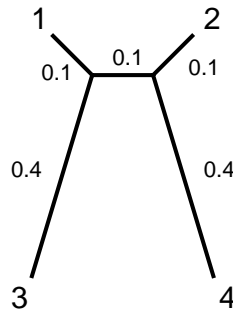


- Corrected distance measure (→ neighbour joining).
- Can lead to **negative distances**.

	Sequences		Distances
1	T T A T T A A C G	→	2
2	A A T T T A A C G		3
3	A A A A A T A C G		5 4
4	A A A A A A T C G		5 4 2
			1 2 3

- **Loss of information**

## Shortcomings of clustering



- Corrected distance measure (→ neighbour joining).
- Can lead to **negative distances**.

Sequences		Distances	
1	T T A T T A A C G	2	3
2	A A T T T A A C G	3	5 4
3	A A A A A T A C G	4	5 4 2
4	A A A A A A T C G		1 2 3

- **Loss of information**
- No **performance measure**.

## Methods of phylogenetic inference

---

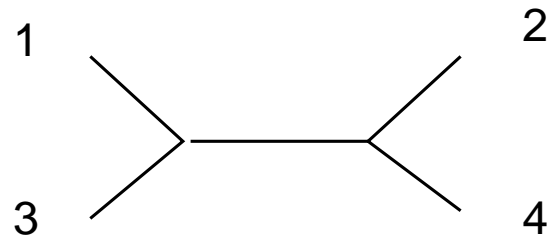
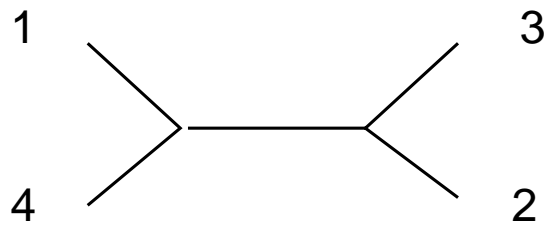
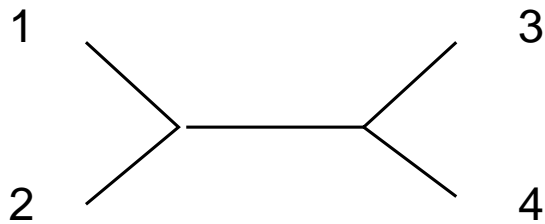
- Clustering
- Maximum parsimony
- Maximum likelihood

# Parsimony

---

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

• • •

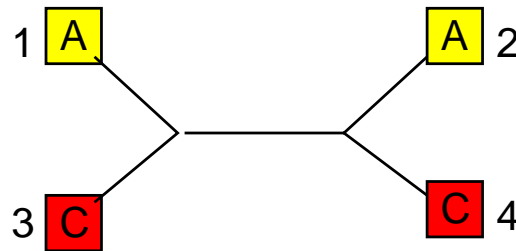
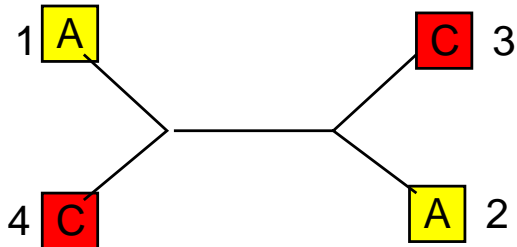
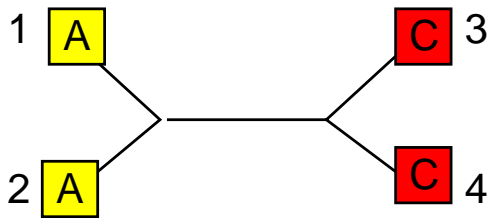


# Parsimony

↓

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

• • •

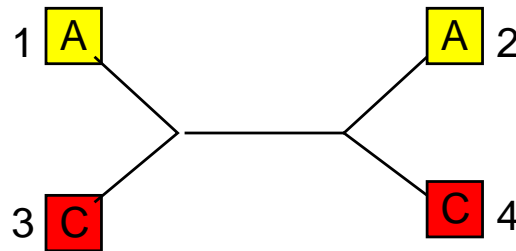
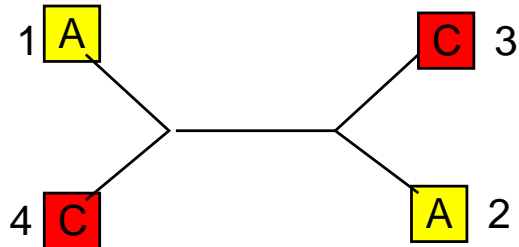
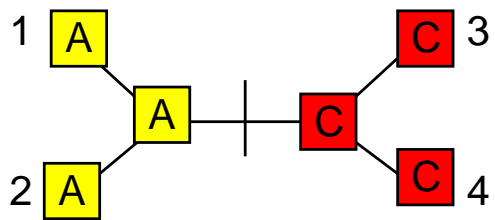


# Parsimony

∨

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

• • •

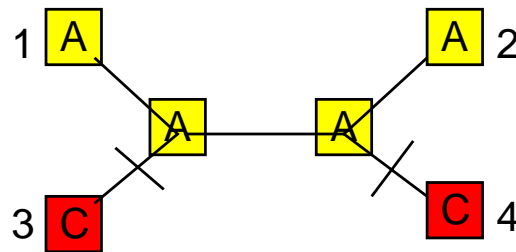
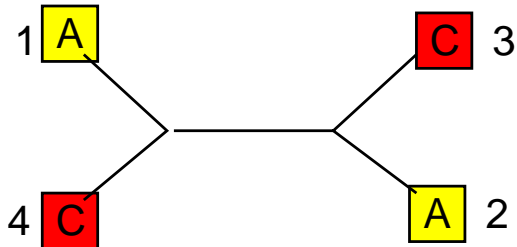
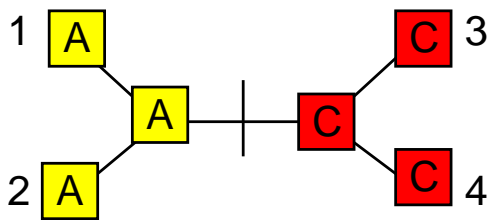


# Parsimony

↓

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

• • •

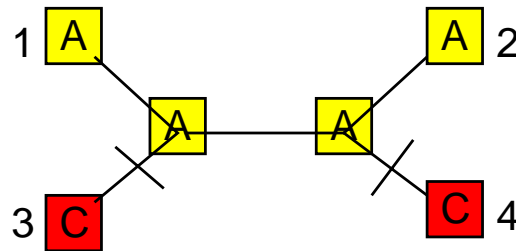
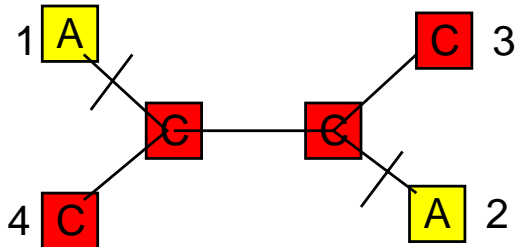
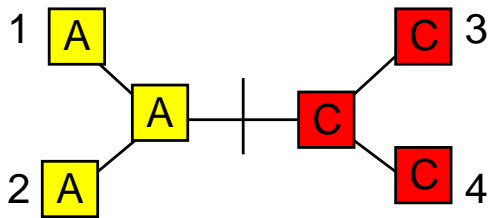


# Parsimony

↓

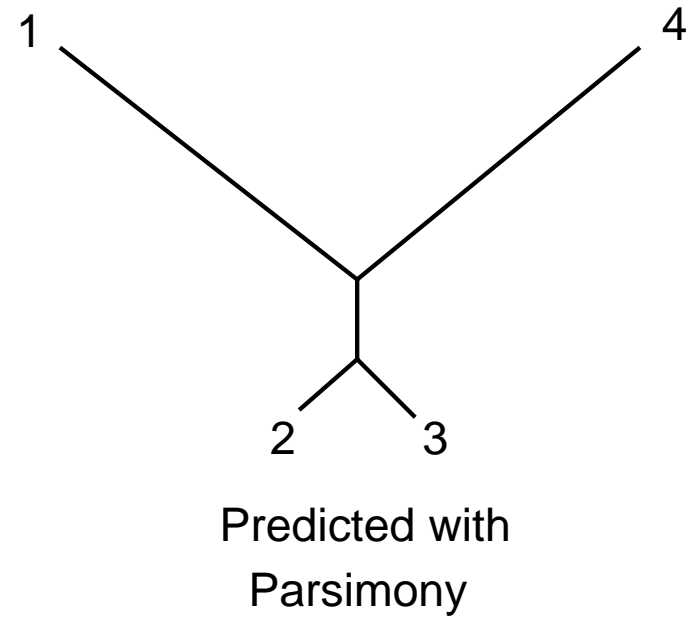
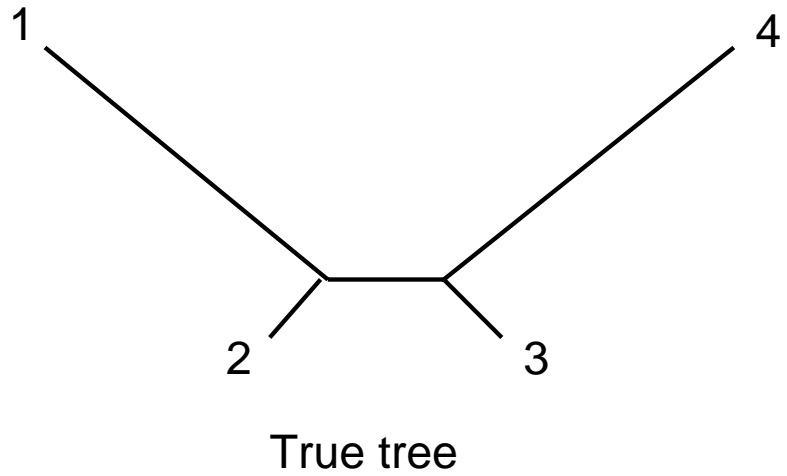
1	A	C	A	C	G
2	A	T	T	C	G
3	C	G	A	G	G
4	C	G	G	C	G

• • •



## Failure of parsimony

---



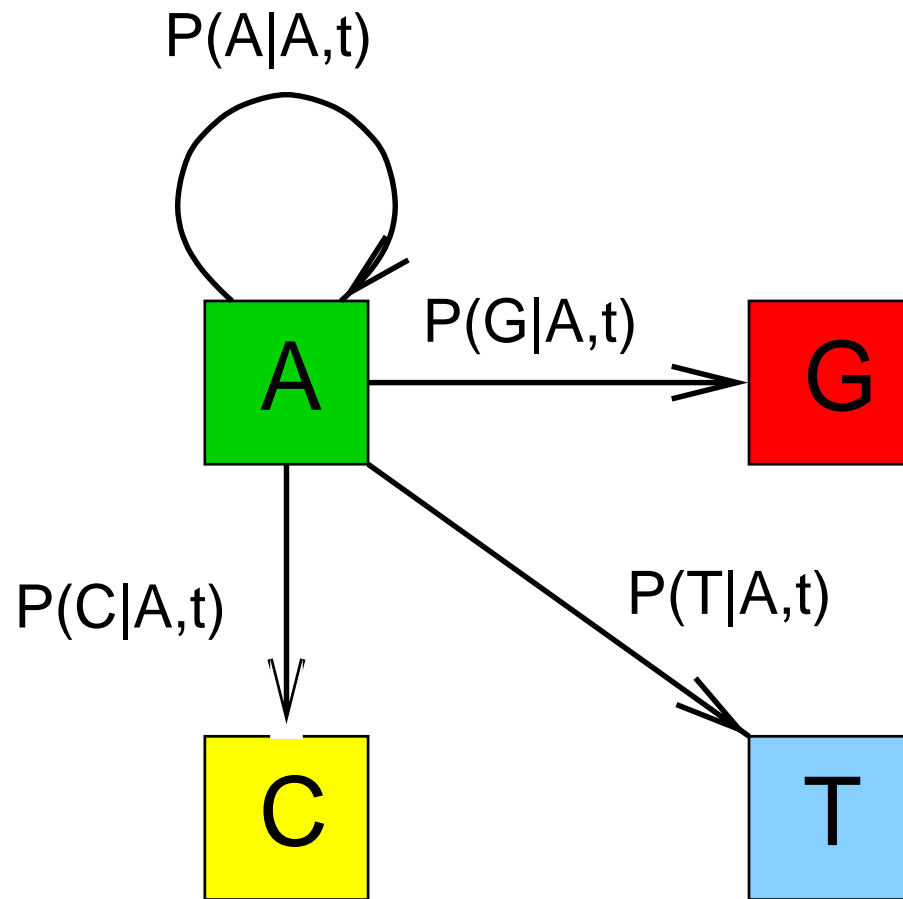
## Methods of phylogenetic inference

---

- Clustering
- Maximum parsimony
- Maximum 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

---

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

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

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

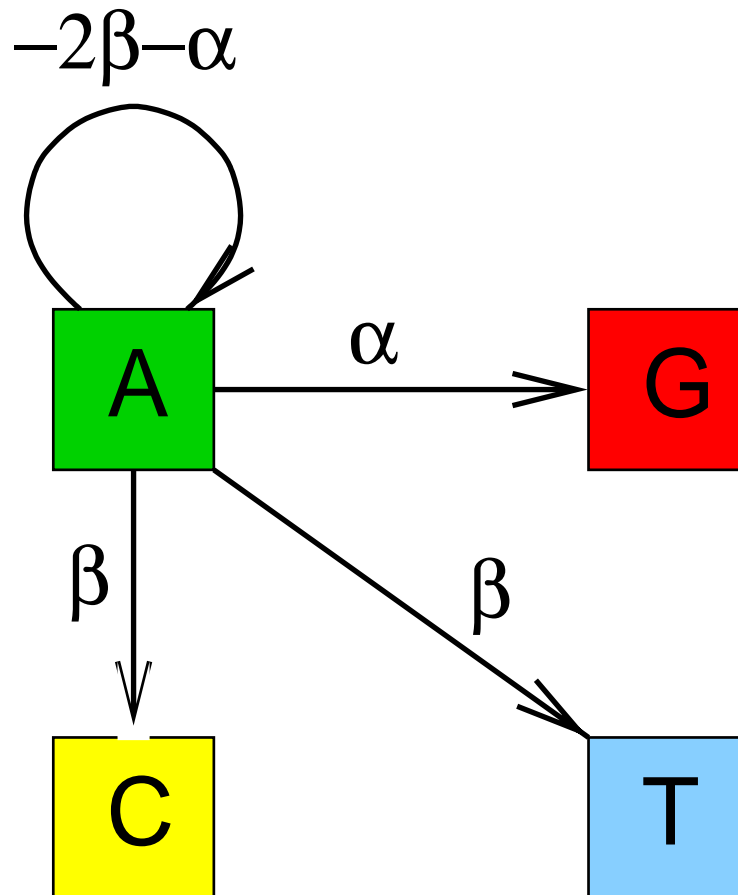
$$\frac{d\mathbf{P}}{dt} = \mathbf{R}\mathbf{P} \implies \mathbf{P}(t) = e^{\mathbf{R}t}$$

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

---

## Transition Rates

---



## Transition Probabilities

---

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

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

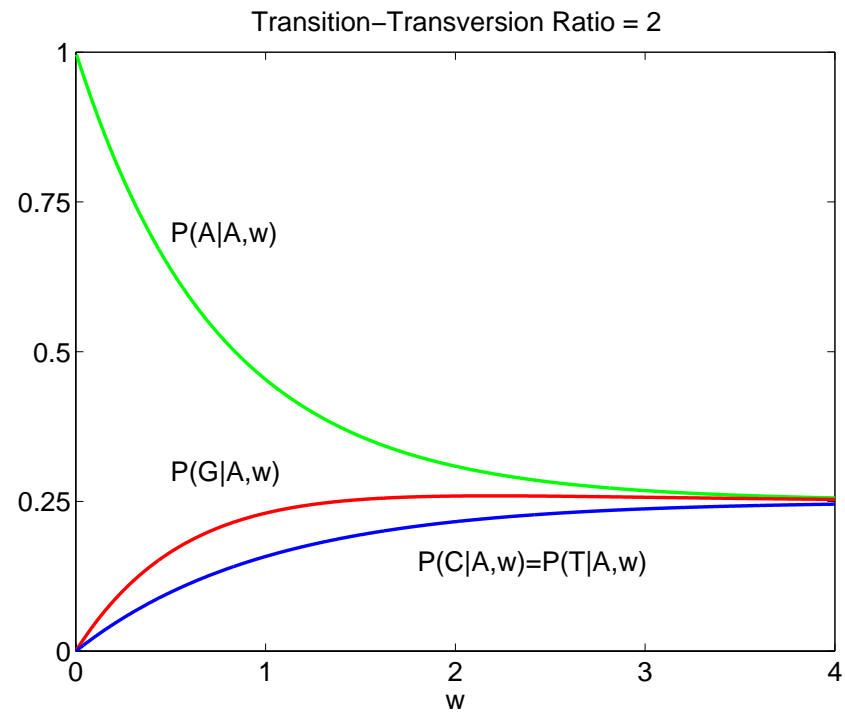
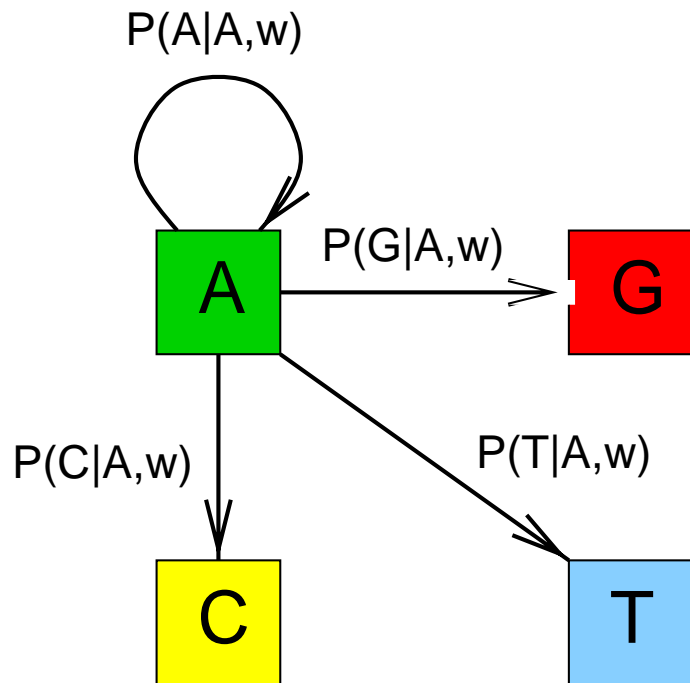
Molecular time:  $w = 4\beta t$

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

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

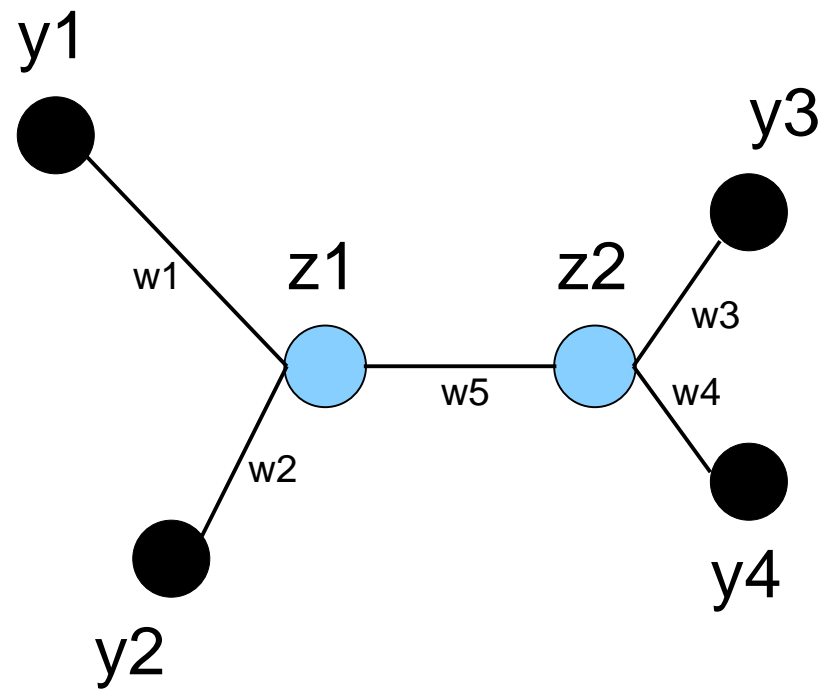
---

## Transition probabilities



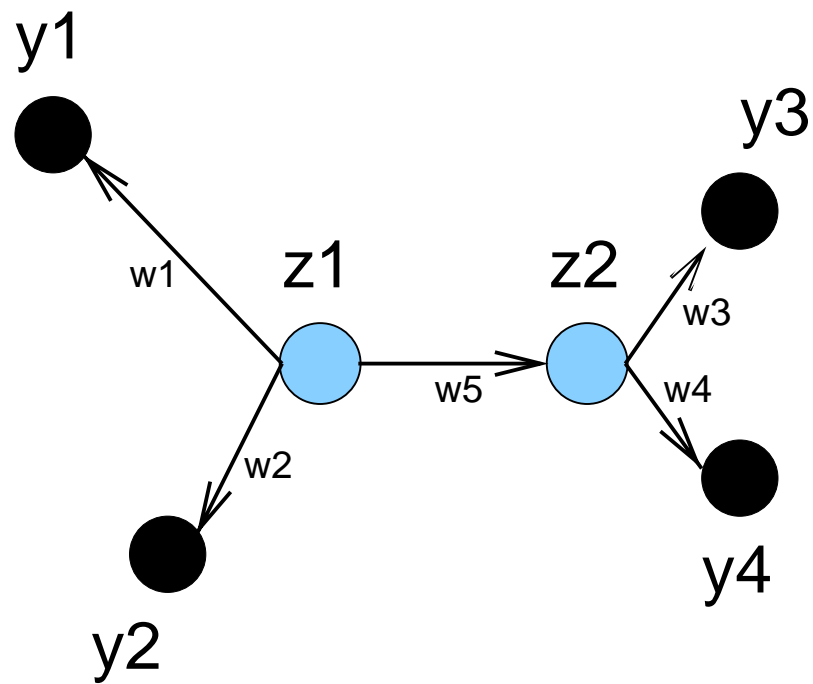
## Phylogenetic tree as an undirected graph

---



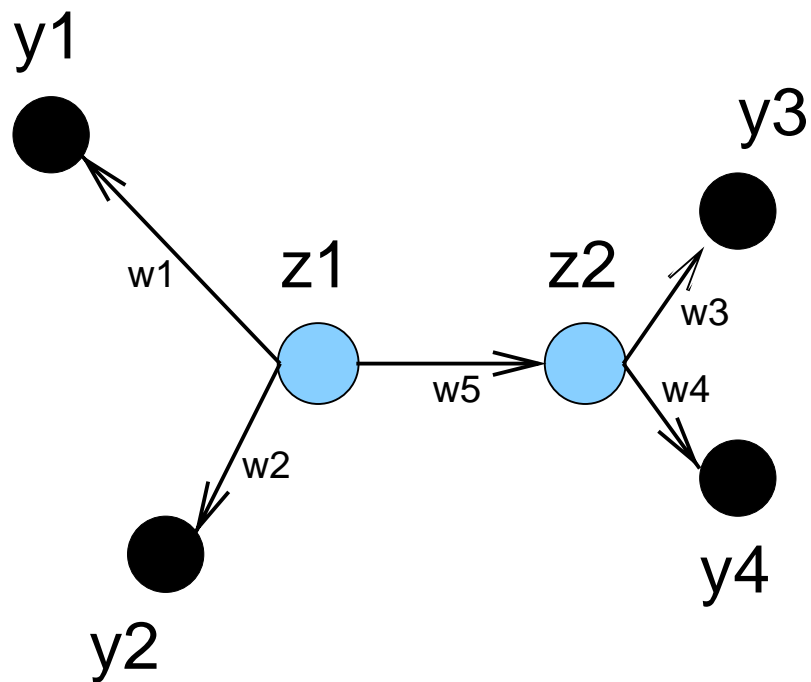
## Phylogenetic tree as a directed graph

---



## Phylogenetic tree as a directed graph

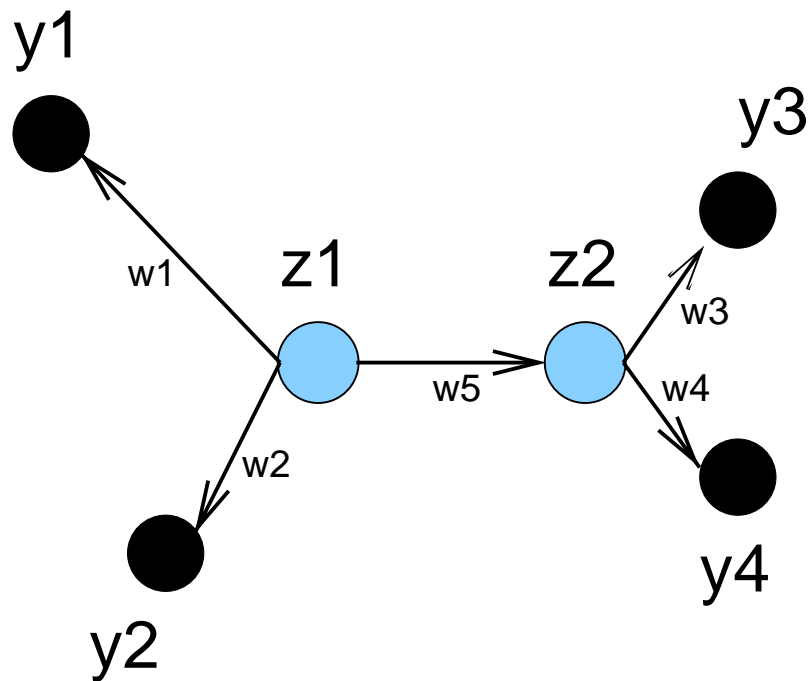
---



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

## Phylogenetic tree as a directed graph

---

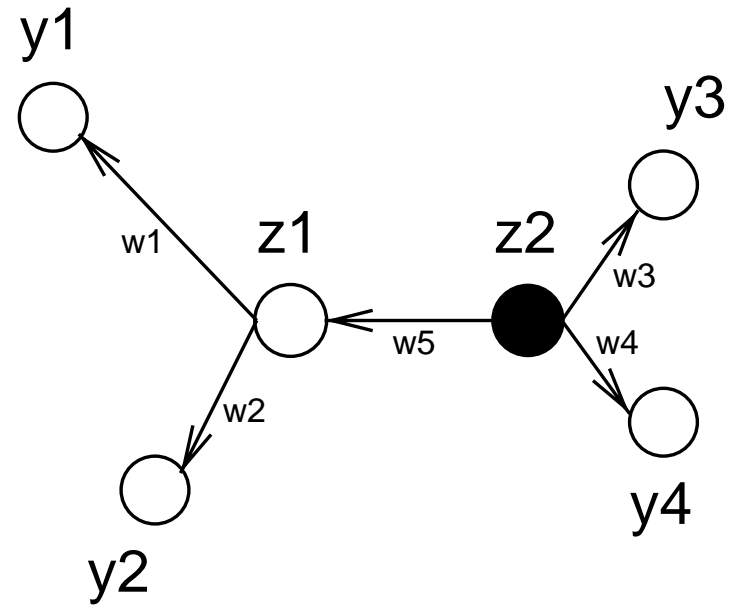
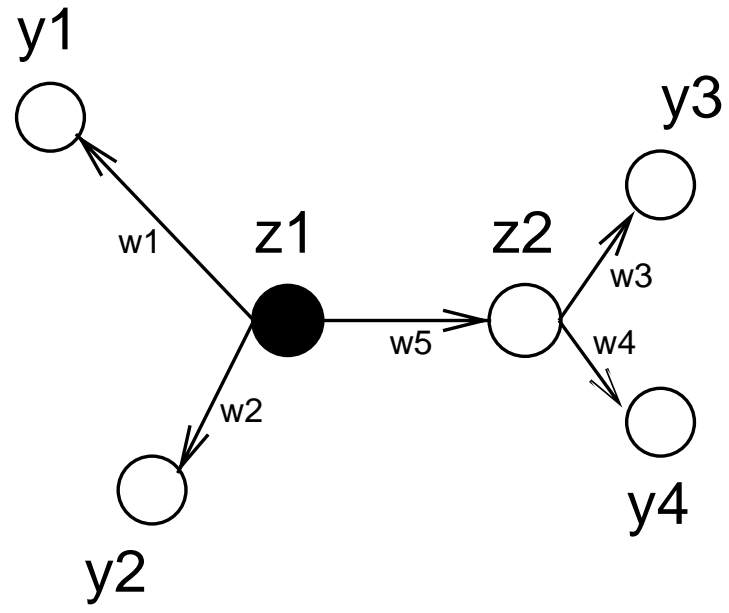


$$P(y_1, y_2, y_3, y_4, z_1, z_2 | \mathbf{w})$$

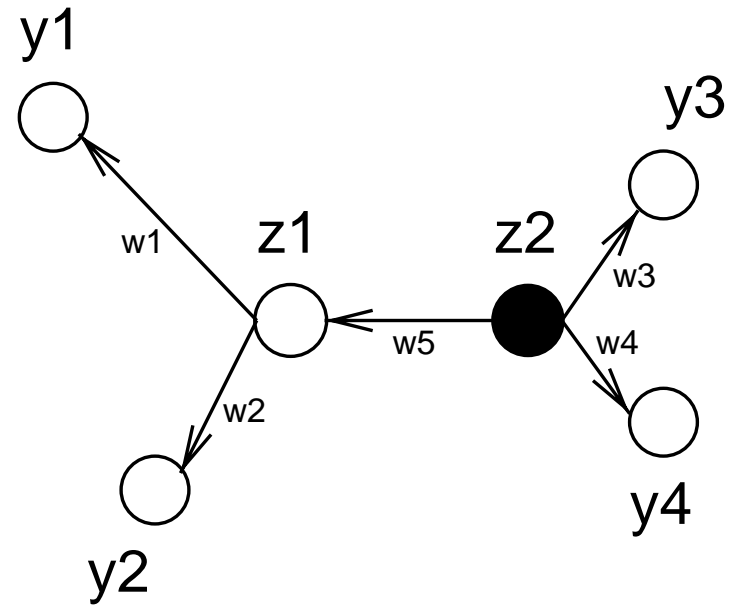
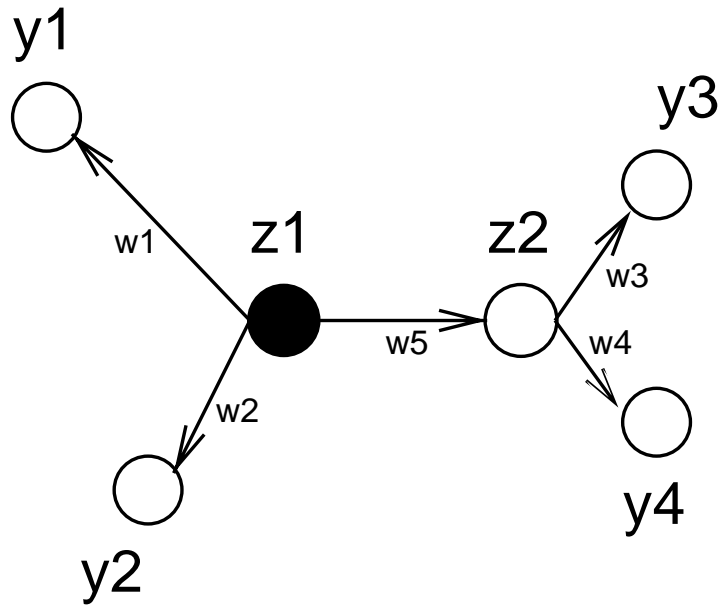
$$= P(y_1 | z_1, w_1) P(y_2 | z_1, w_2) P(z_2 | z_1, w_5) P(y_3 | z_2, w_3) P(y_4 | z_2, w_4) P(z_1)$$

---

## Different directed graphs

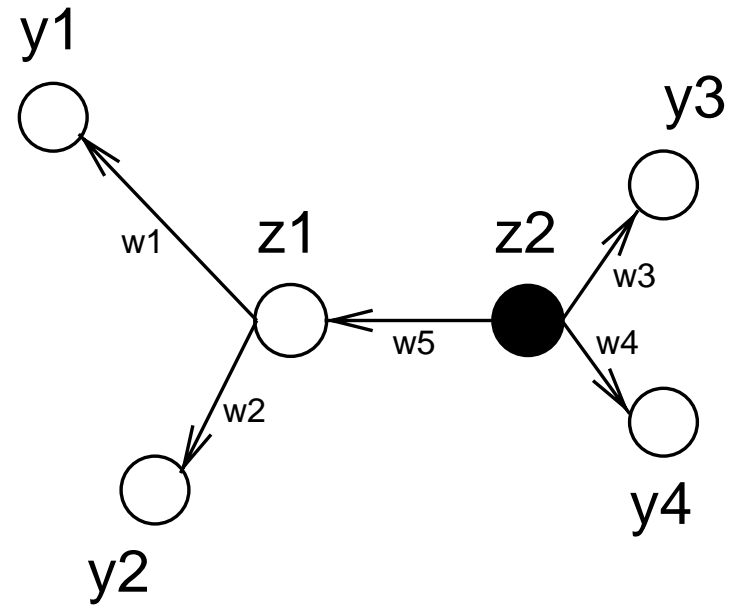
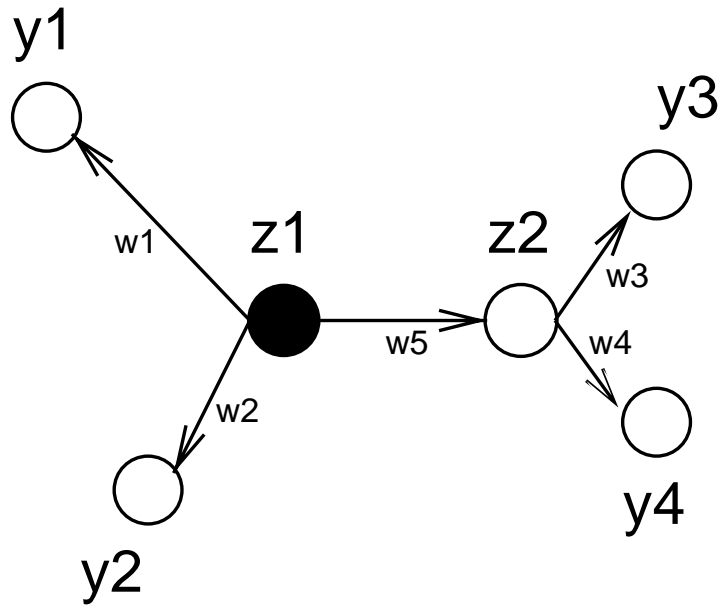


## Different directed graphs



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

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

## Reversibility

---

We can *not* decide on the direction of evolutionary processes.

$$P(z_1|z_2, w_5)P(z_2) = P(z_2|z_1, w_5)P(z_1)$$

## Reversibility

---

We can *not* decide on the direction of evolutionary processes.

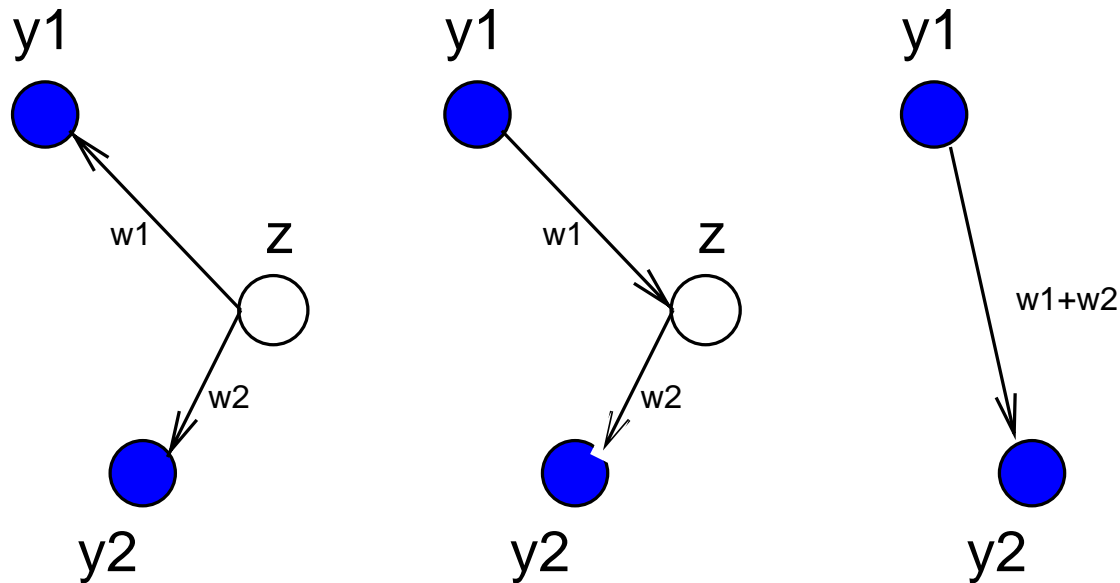
$$P(z_1|z_2, w_5)P(z_2) = P(z_2|z_1, w_5)P(z_1)$$

- Changing the position of the root and the direction of the arcs does not affect the probability.
- All directed graphs are in the same equivalence class.

## Root elimination

**Homogeneous Markov chain**  $\implies$  Multiplicativity of the substitution matrices:

$$\mathbf{P}(w_1)\mathbf{P}(w_2) = \mathbf{P}(w_1 + w_2)$$



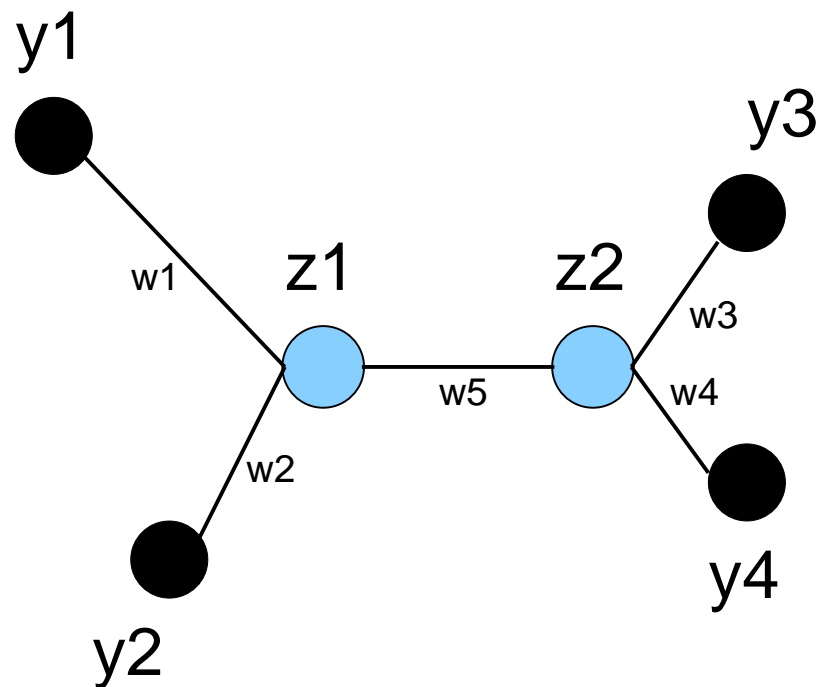
$$\sum_z P(y_2|z, w_2)P(y_1|z, w_1)P(z) \quad \sum_z P(y_2|z, w_2)P(z|y_1, w_1)P(y_1) \quad P(y_2|y_1, w_1+w_2)P(y_1)$$

**Reversibility**  $\implies$  left = middle

**Multiplicativity**  $\implies$  middle = right

## Expansion of the joint probability

---

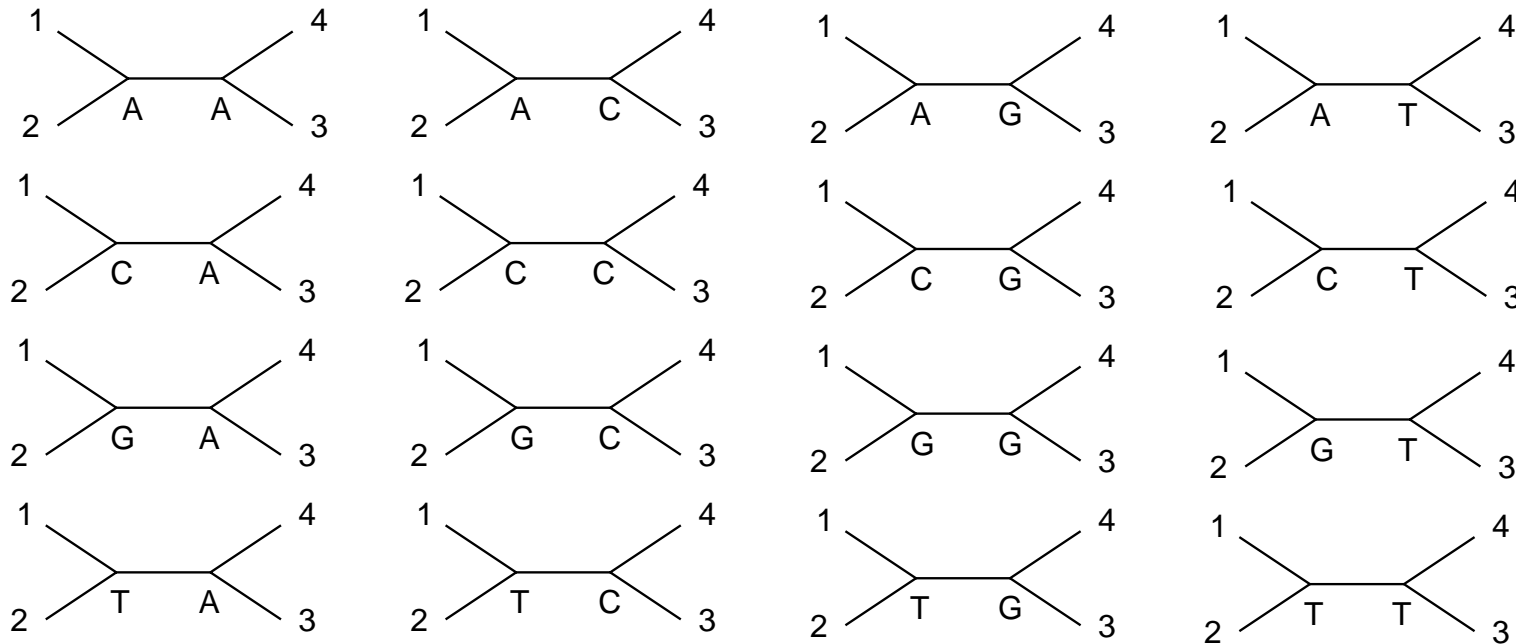


$$P(y_1, y_2, y_3, y_4, z_1, z_2 | \mathbf{w})$$

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

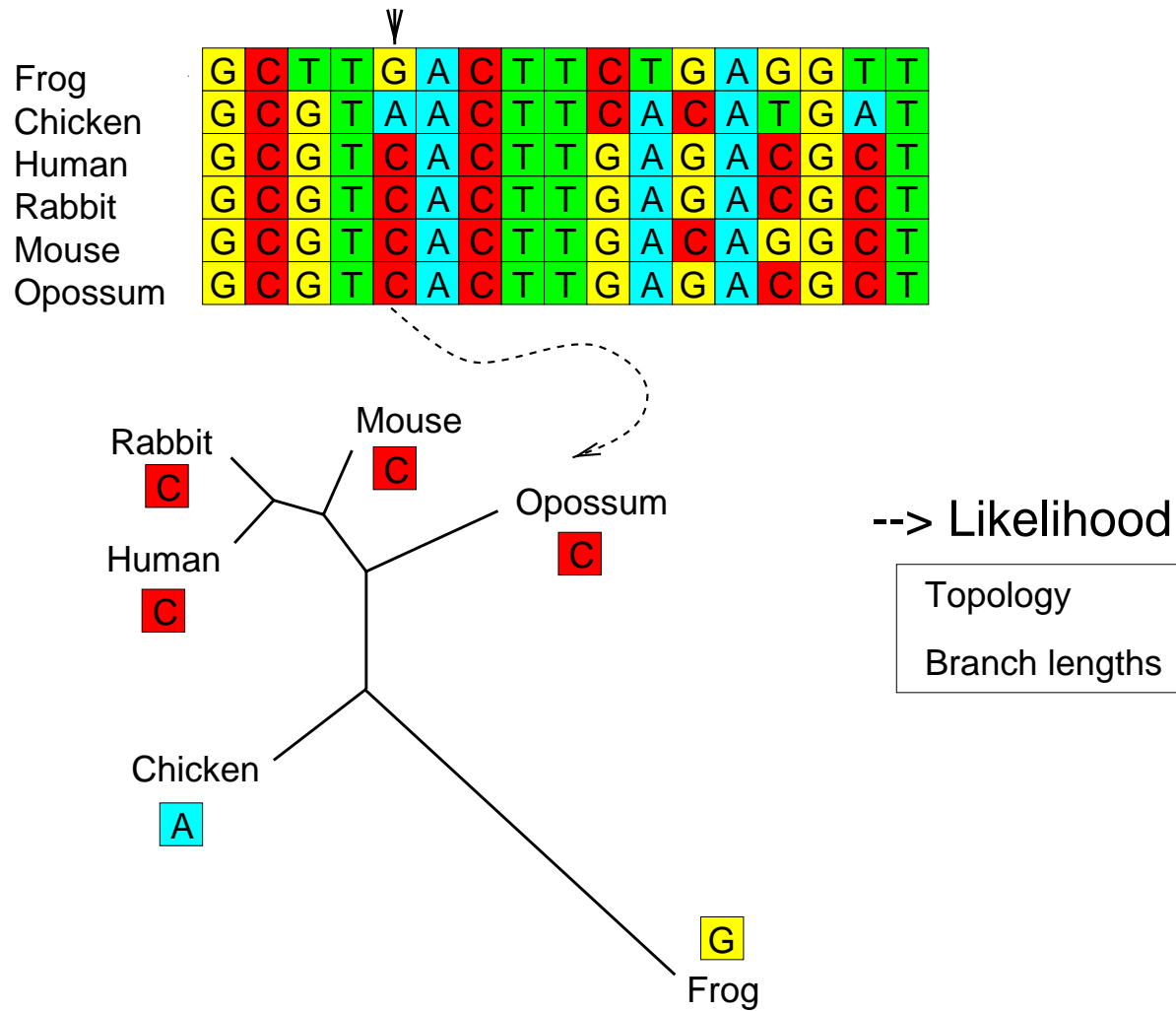
---

## Marginalisation



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

# Statistical approach to phylogenetics



## Maximum likelihood

---

- Find tree topology  $S$  and vector of branch lengths  $\mathbf{w}$  that maximize likelihood  $\ln P(D|S, \mathbf{w})$

## Maximum likelihood

---

- Find tree topology  $S$  and vector of branch lengths  $\mathbf{w}$  that maximize likelihood  $\ln P(D|S, \mathbf{w})$
- No analytic solution.
- Find maximum in a high-dimensional space with a heuristic hill climbing method.
- Given topology  $S$ , optimise branch lengths  $\mathbf{w}$  by gradient ascent:

$$\Delta \mathbf{w} \propto \nabla \ln P(D|S, \mathbf{w})$$



## Maximum likelihood

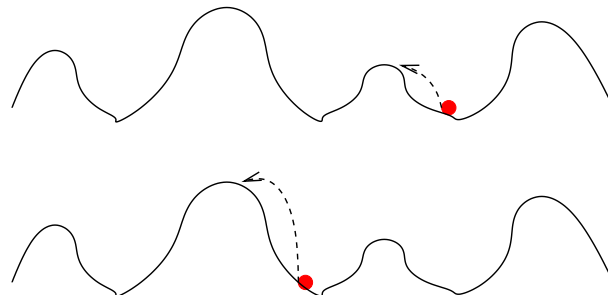
---

- Find tree topology  $S$  and vector of branch lengths  $\mathbf{w}$  that maximize likelihood  $\ln P(D|S, \mathbf{w})$
- No analytic solution.
- Find maximum in a high-dimensional space with a heuristic hill climbing method.
- Given topology  $S$ , optimise branch lengths  $\mathbf{w}$  by gradient ascent:

$$\Delta \mathbf{w} \propto \nabla \ln P(D|S, \mathbf{w})$$



- Repeat for different tree topologies  $S$ .



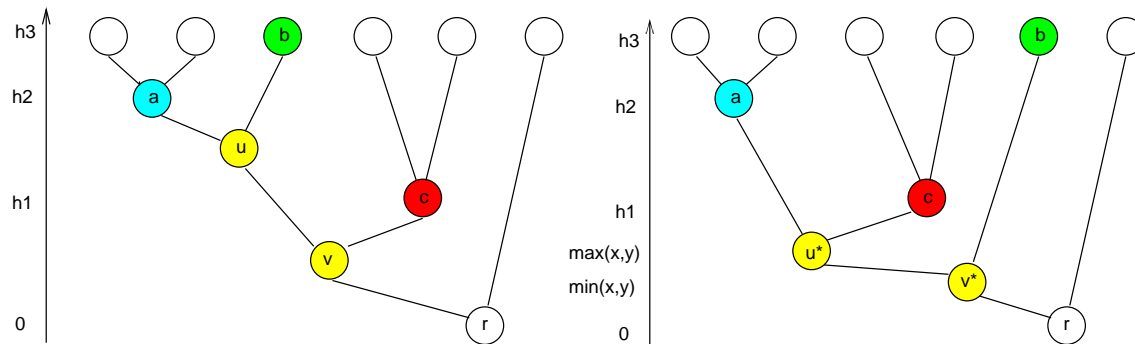
## NP hard problem

---

- For  $M$  taxa, there are  $(2M - 5)!!$  unrooted trees.
- $M = 4 \longrightarrow 3!! = 3$
- $M = 6 \longrightarrow 7!! = 7 \times 5 \times 3 = 105$
- $M = 10 \longrightarrow \approx 2 \times 10^6$
- $M = 20 \longrightarrow \approx 2 \times 10^{20}$
- $M$  large  $\longrightarrow$  Exhaustive search impossible.

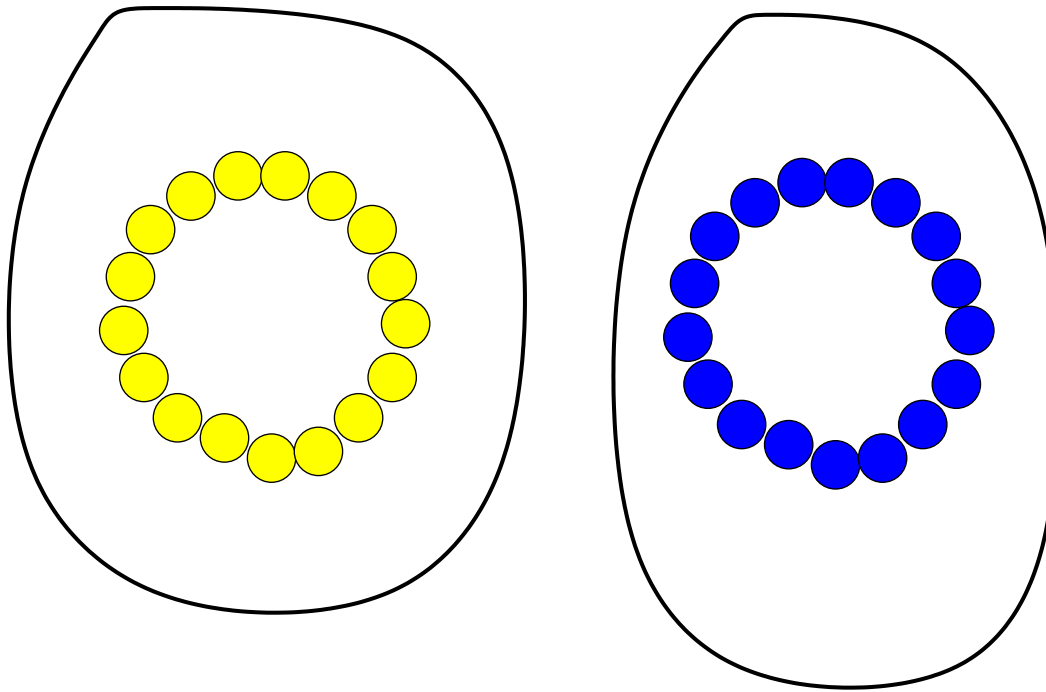
## NP hard problem

- For  $M$  taxa, there are  $(2M - 5)!!$  unrooted trees.
- $M = 4 \longrightarrow 3!! = 3$
- $M = 6 \longrightarrow 7!! = 7 \times 5 \times 3 = 105$
- $M = 10 \longrightarrow \approx 2 \times 10^6$
- $M = 20 \longrightarrow \approx 2 \times 10^{20}$
- $M$  large  $\longrightarrow$  Exhaustive search impossible.
- Heuristic search methods.



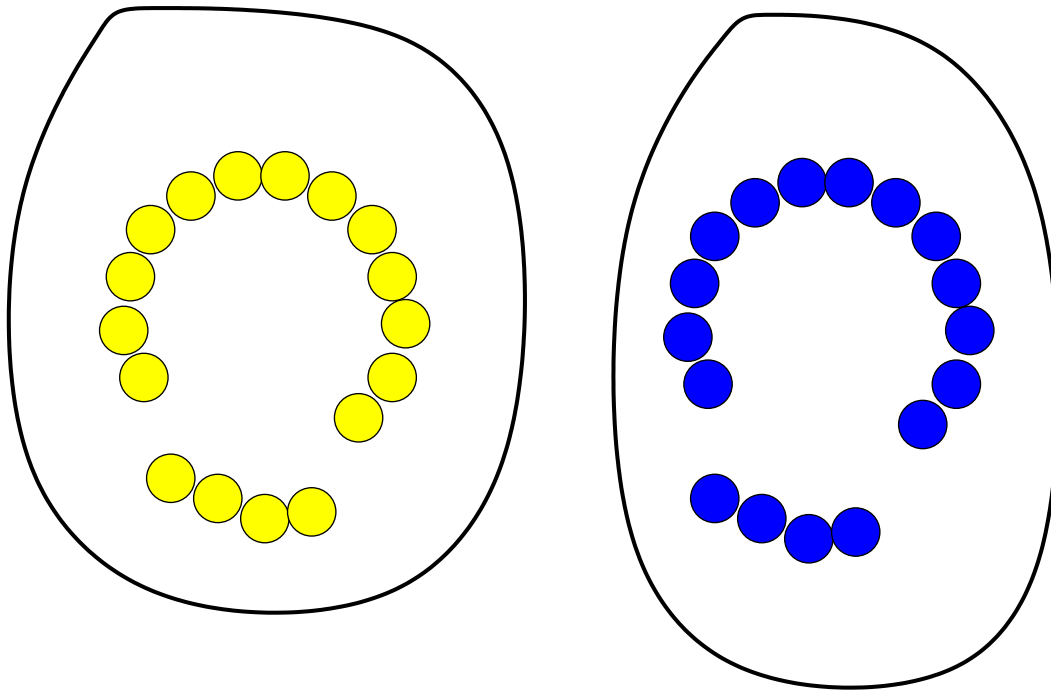
*Neisseria*

---



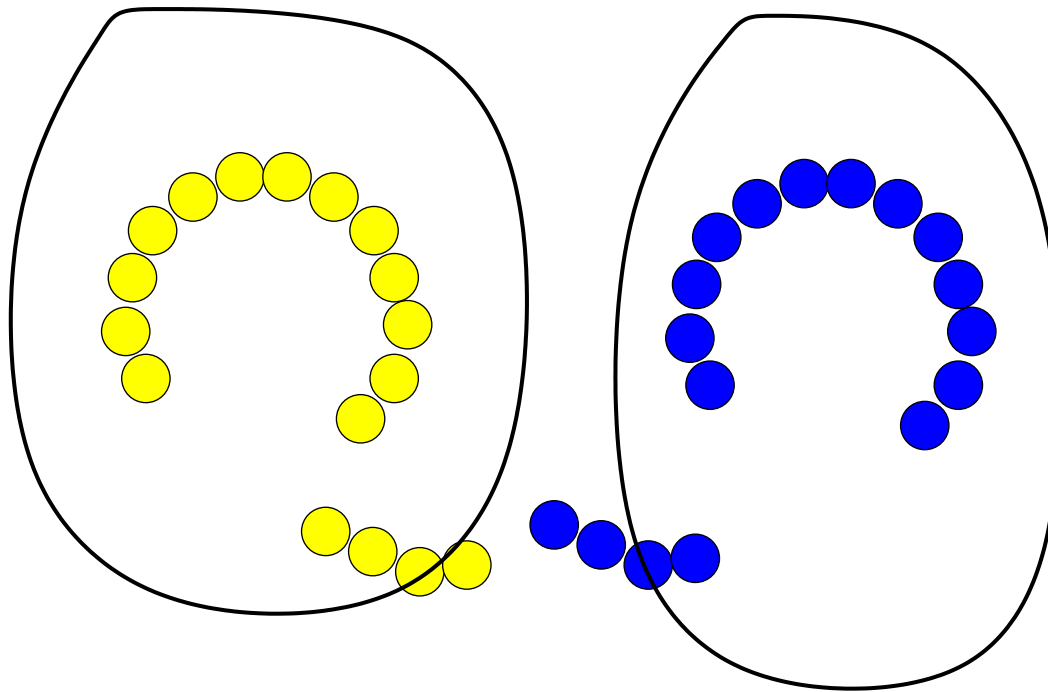
# Recombination

---



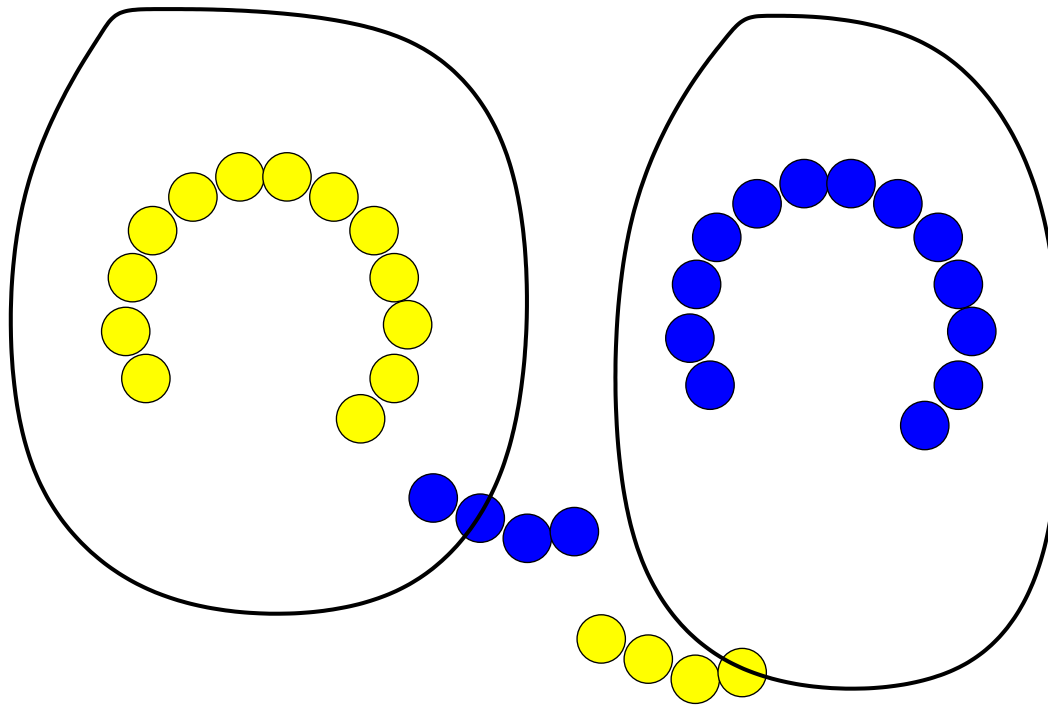
# Recombination

---



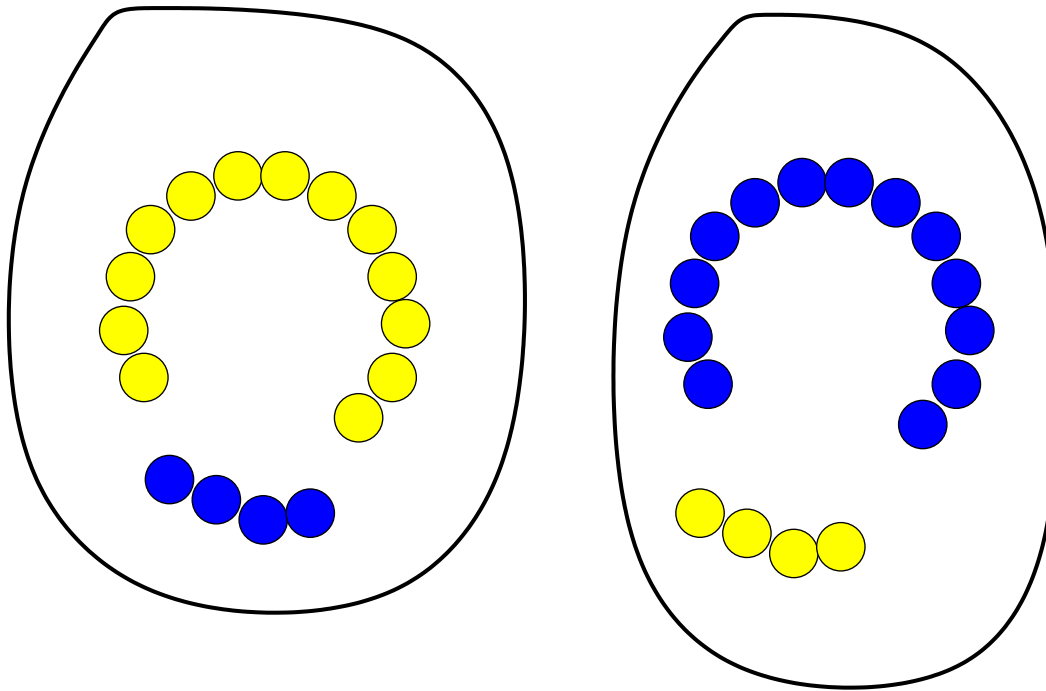
# Recombination

---



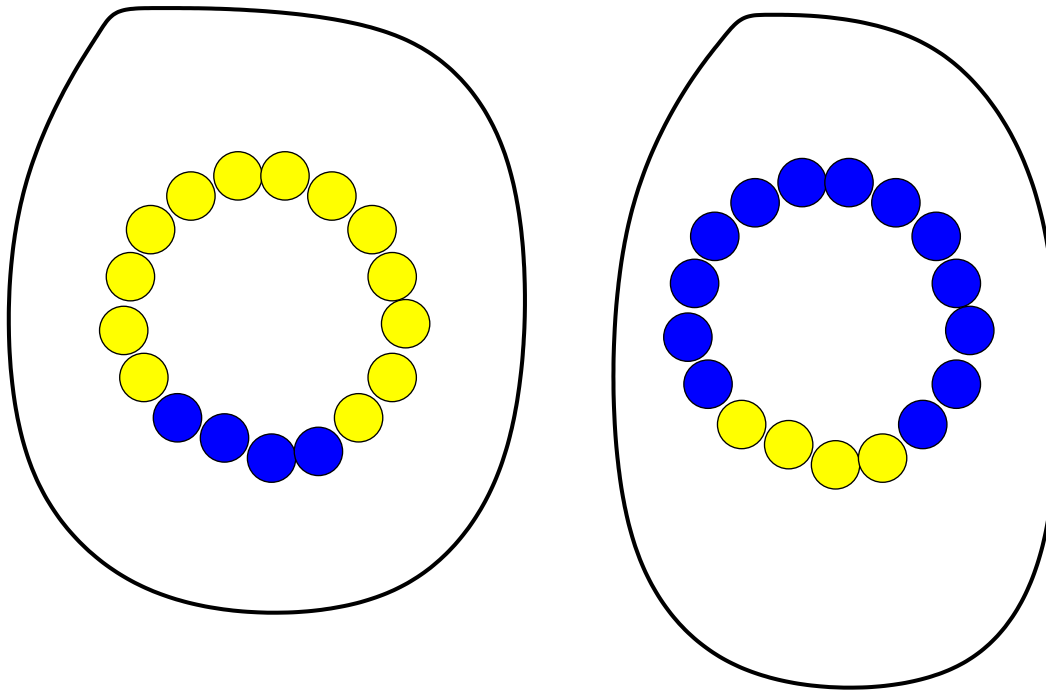
# Recombination

---



## Recombination

---



---

1995

Robertson, Sharp, McCutchan, Hahn

Recombination in HIV-1

Nature 374, pp.124-126

1997

Dennis Blakeslee

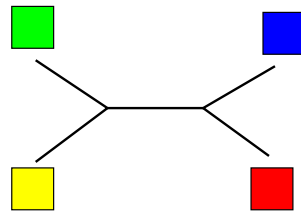
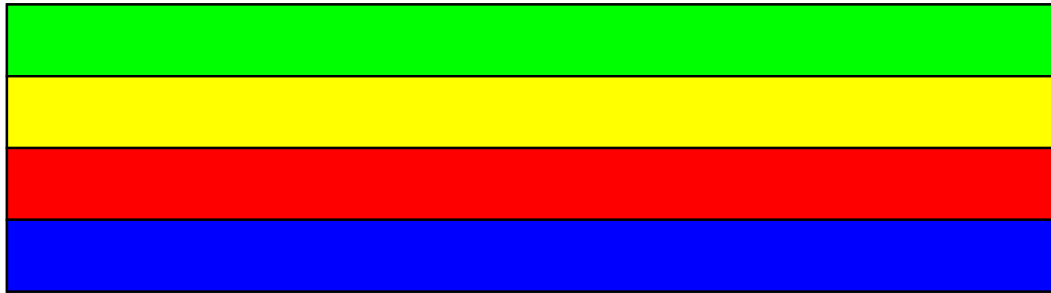
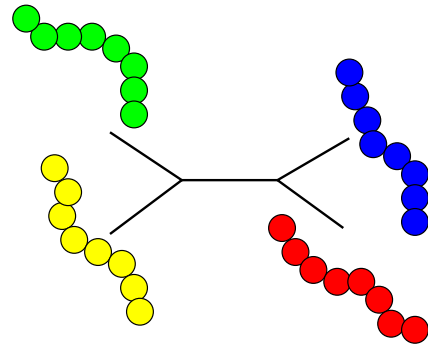
Recombination in HIV: A fast track to resistance?

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

---

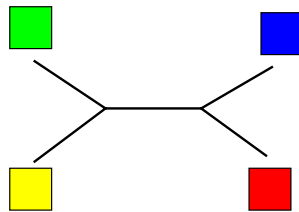
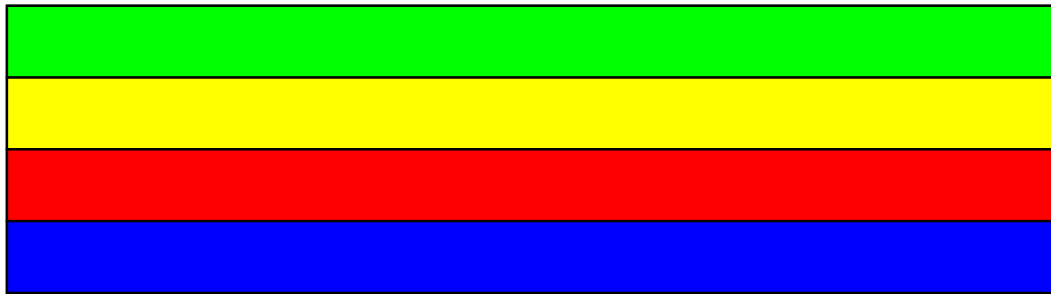
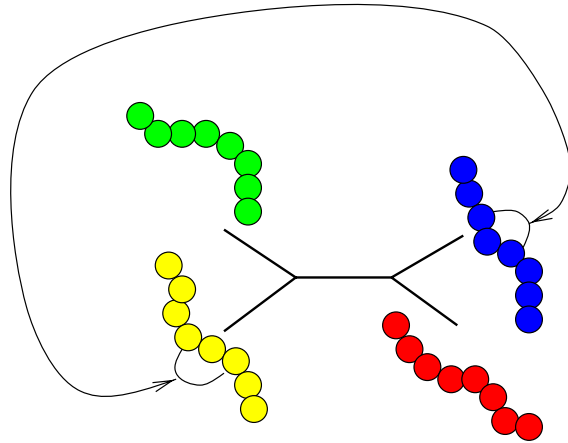
# Recombination

---



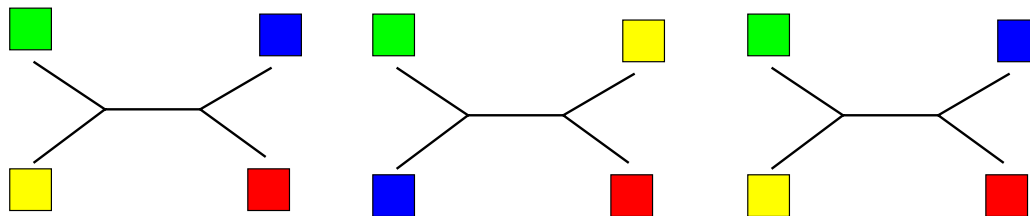
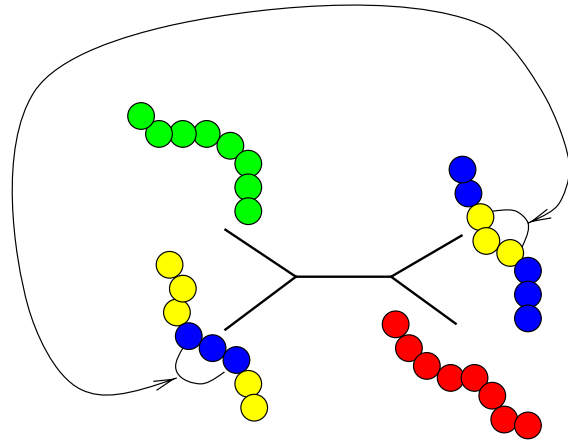
# Recombination

---

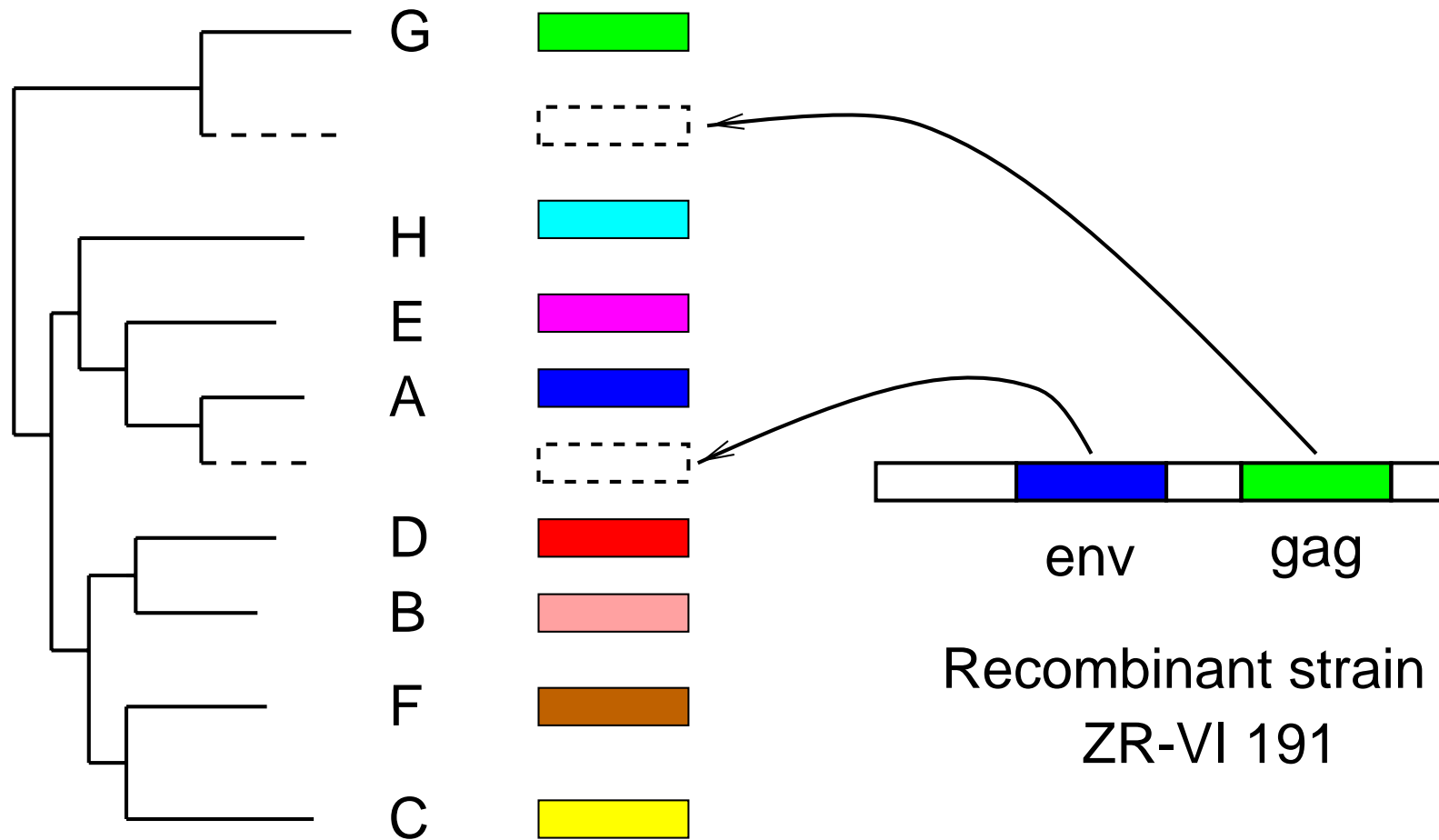


# Recombination

---



## Recombination in HIV 1



## Detecting recombination with window methods

---

---

## Detecting recombination with window methods

---

- Slide a **window** across the alignment.
- Look for **subregions** that are **significantly different** from the rest.

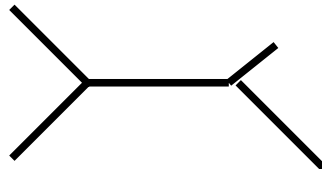
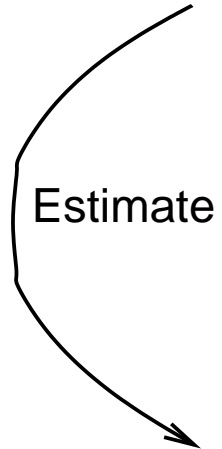
## TOPAL (McGuire & Wright, 1997)

---



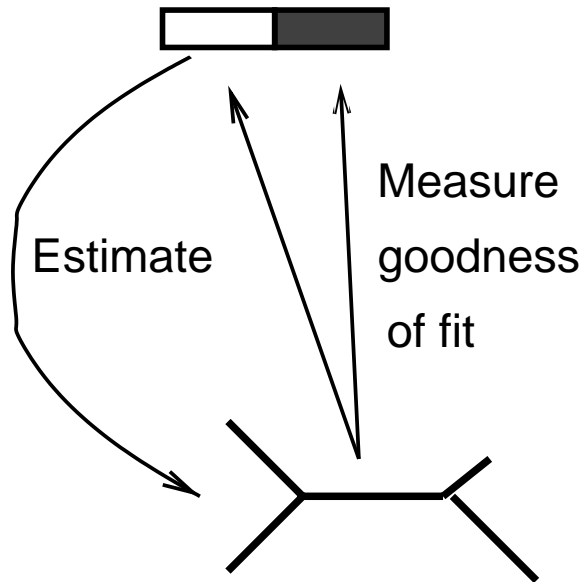
## TOPAL (McGuire & Wright, 1997)

---

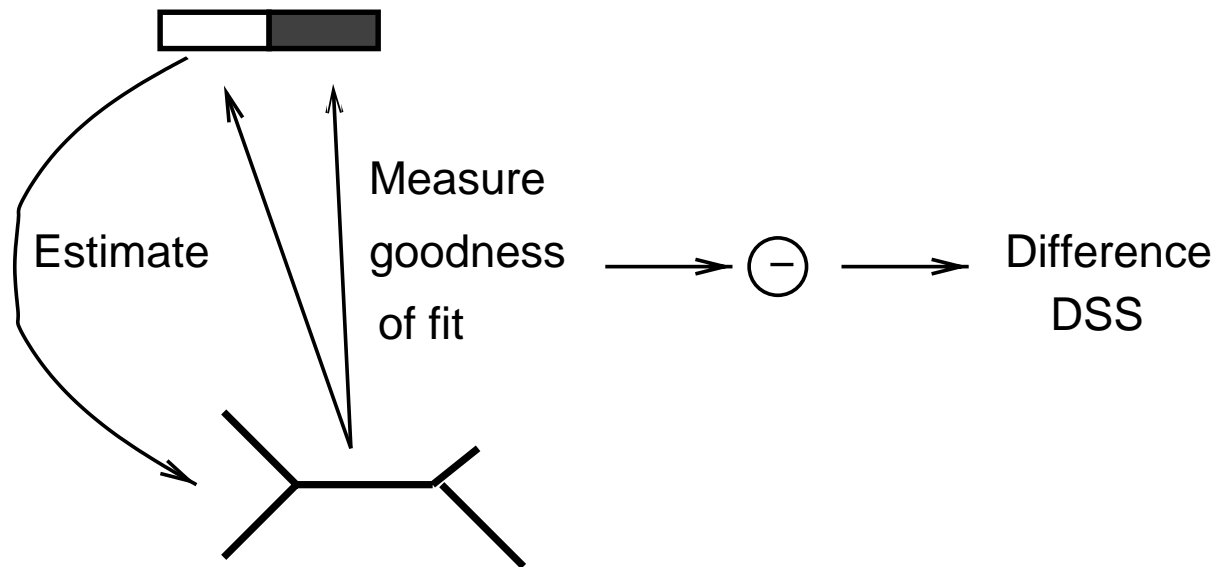


## TOPAL (McGuire & Wright, 1997)

---



## TOPAL (McGuire & Wright, 1997)



## TOPAL (McGuire & Wright, 1997)

---



small

TOPAL (McGuire & Wright, 1997)

---



small



large

---

## TOPAL (McGuire & Wright, 1997)

---



small

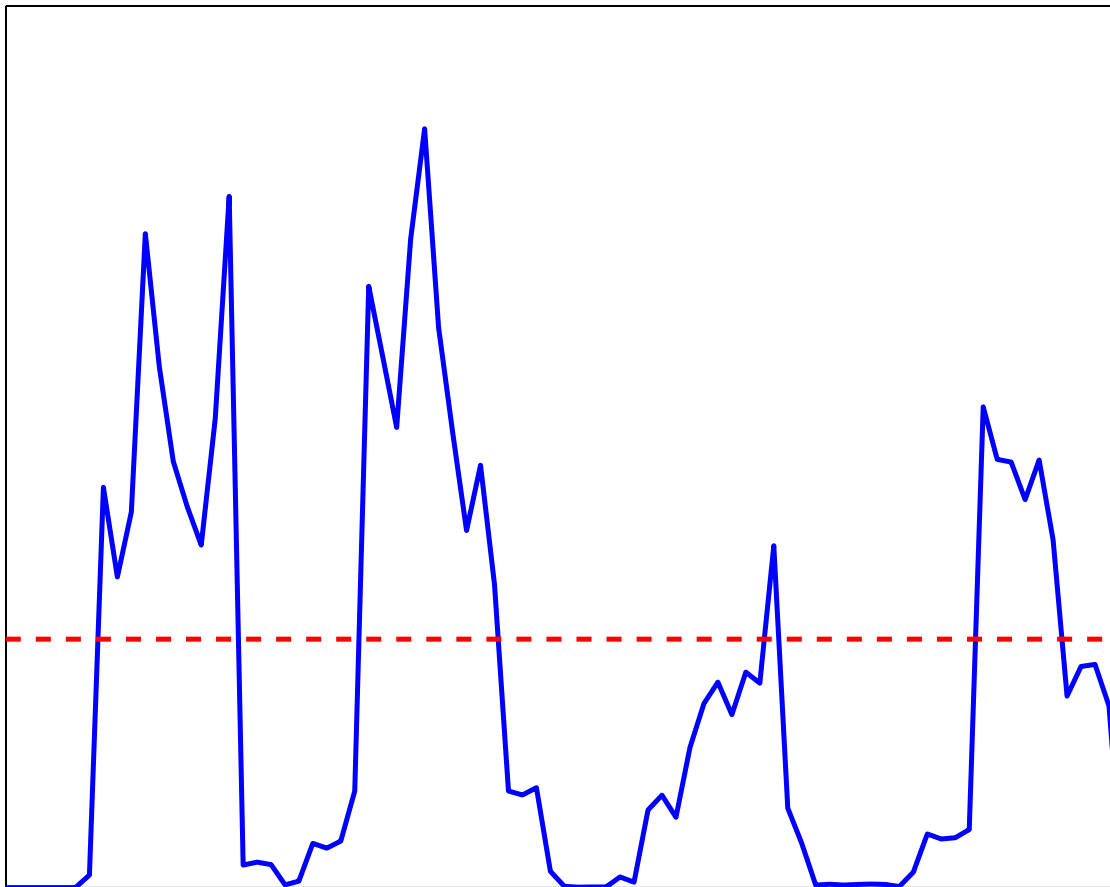


large

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

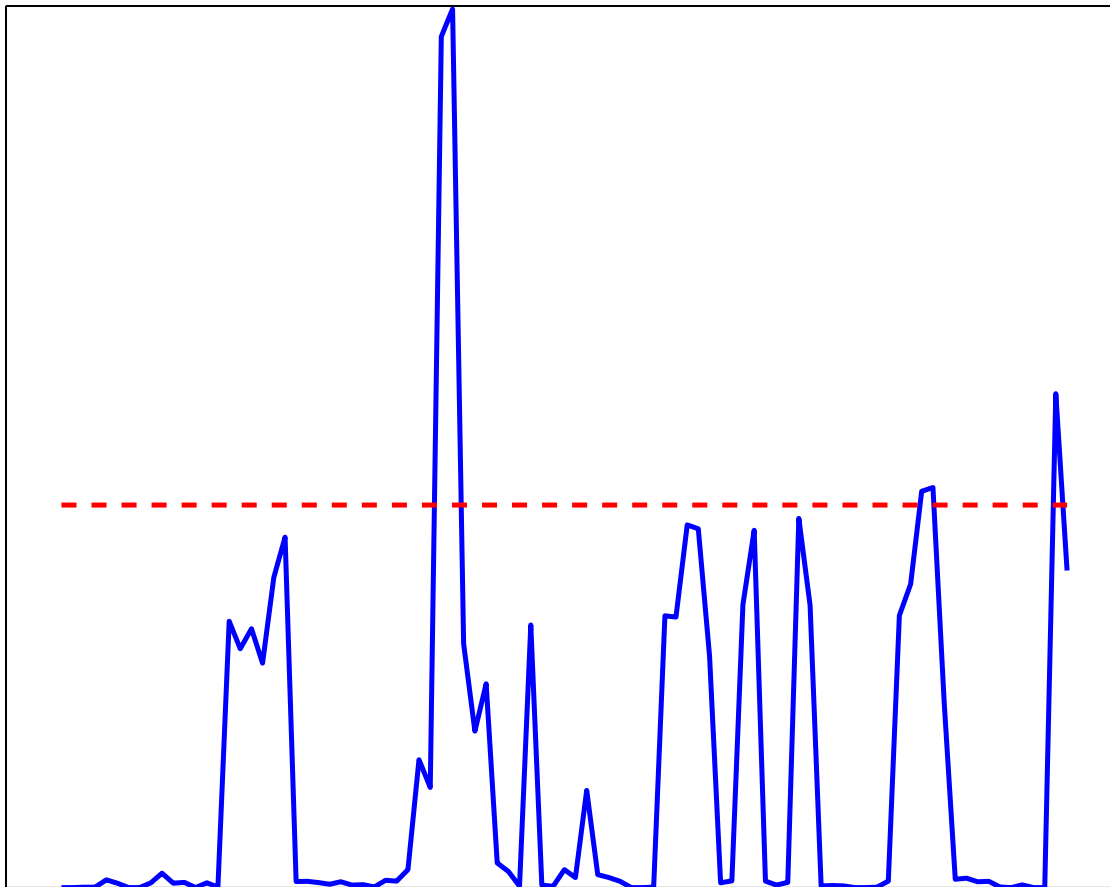
Example: TOPAL, window size=200

---



Example: TOPAL, window size=100

---



## Hidden Markov models (HMMs)

---

---

## Hidden Markov models (HMMs)

---

- No window needed.
- More precise location of the breakpoints.
- All parameters inferred from the data.

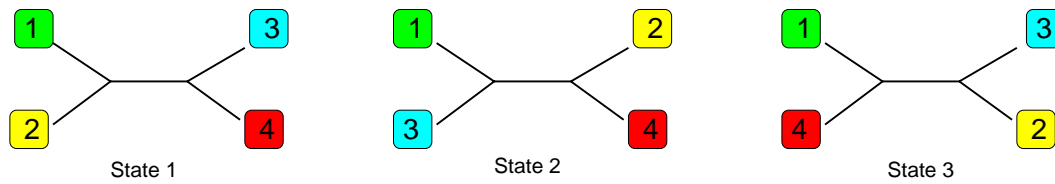
## Hidden Markov models (HMMs)

---

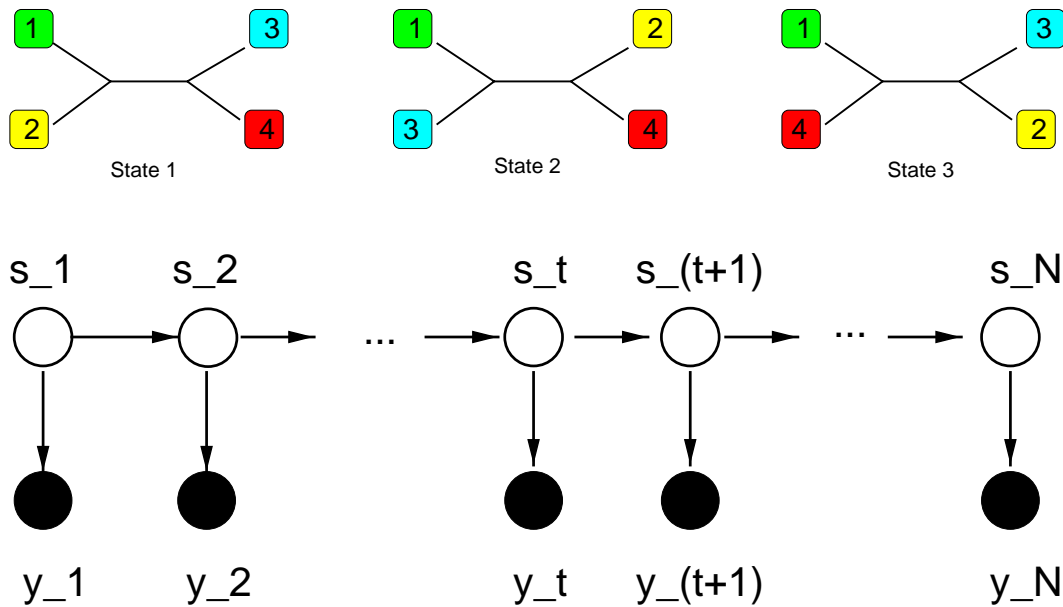
- No window needed.
- More precise location of the breakpoints.
- All parameters inferred from the data.
  
- Can currently only deal with a small number of species.

## Modelling recombination with HMMs

---

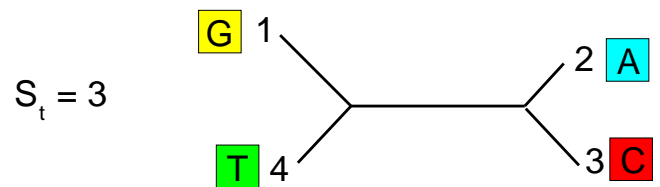
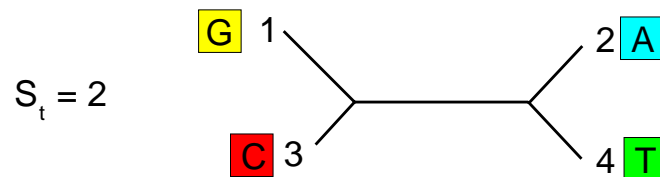
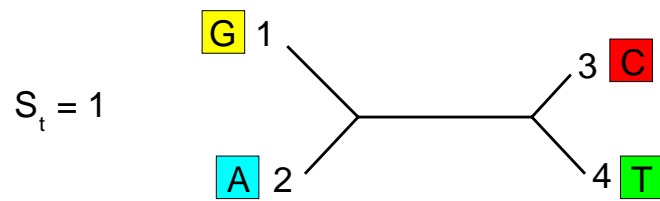
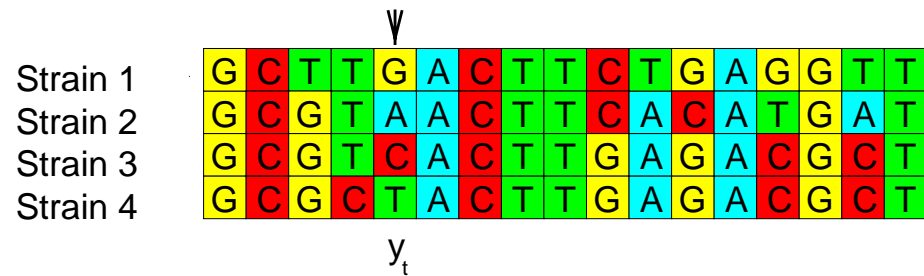


## Modelling recombination with HMMs



AGCATCGTTCTATTTTACCGGCTCCCG  
TGTGTCGCTCAAGATTGCCATCGCGCG  
TGTGTCGTGGTCTAGATTGCCATCGCGCG  
TGTATCGCTCTAGTTTGCCAGCTCCCG

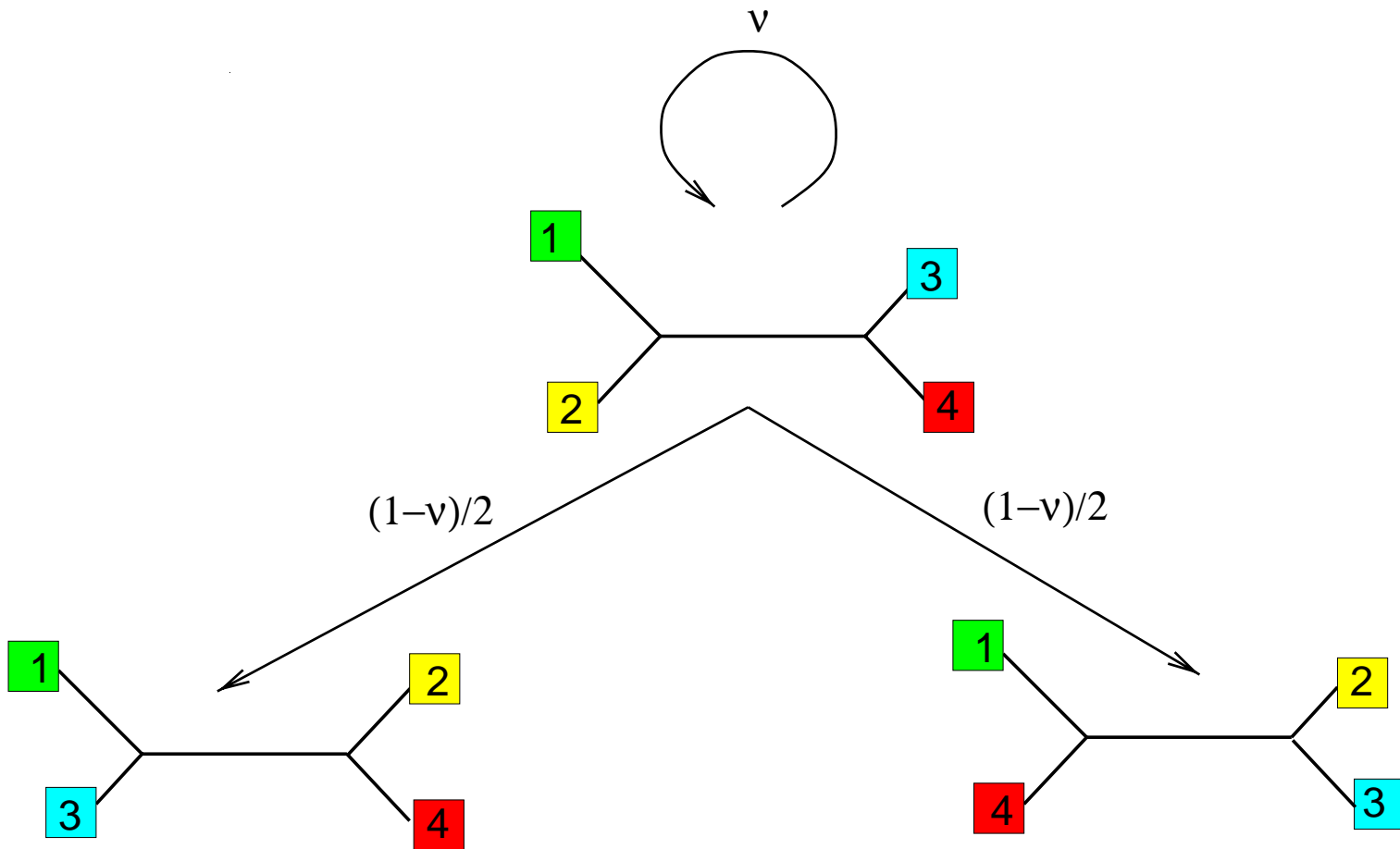
## Emission probabilities (vertical arrows)



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

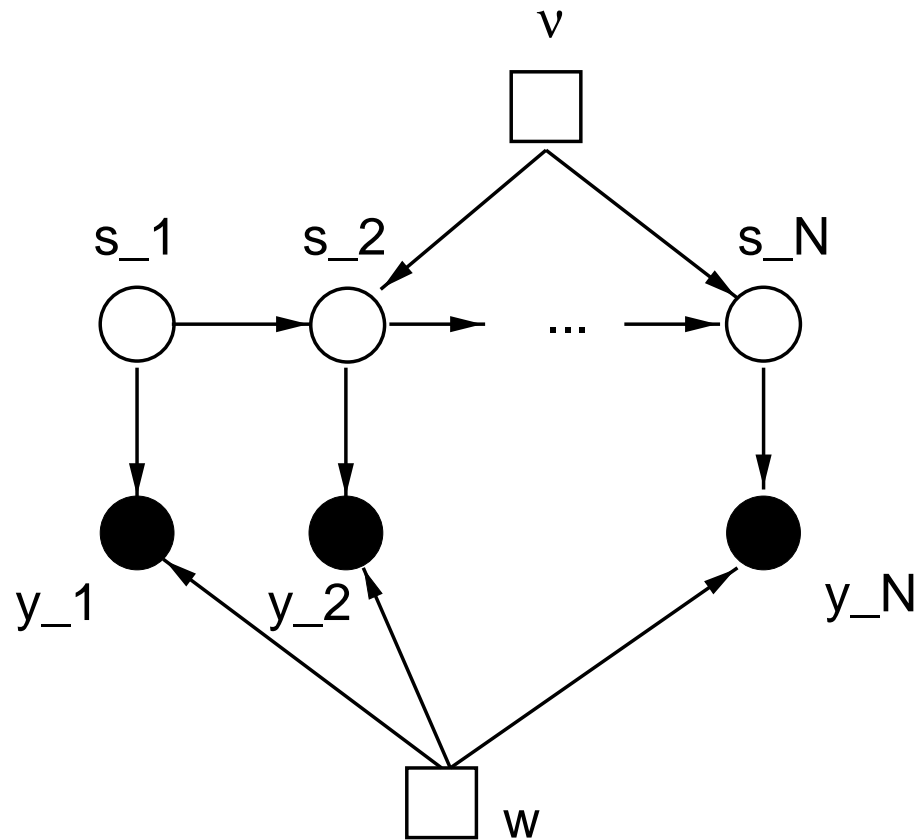
Topology	$S_t$
Branch lengths	$w$

## Transition probabilities (horizontal arrows)



## HMM parameters

---

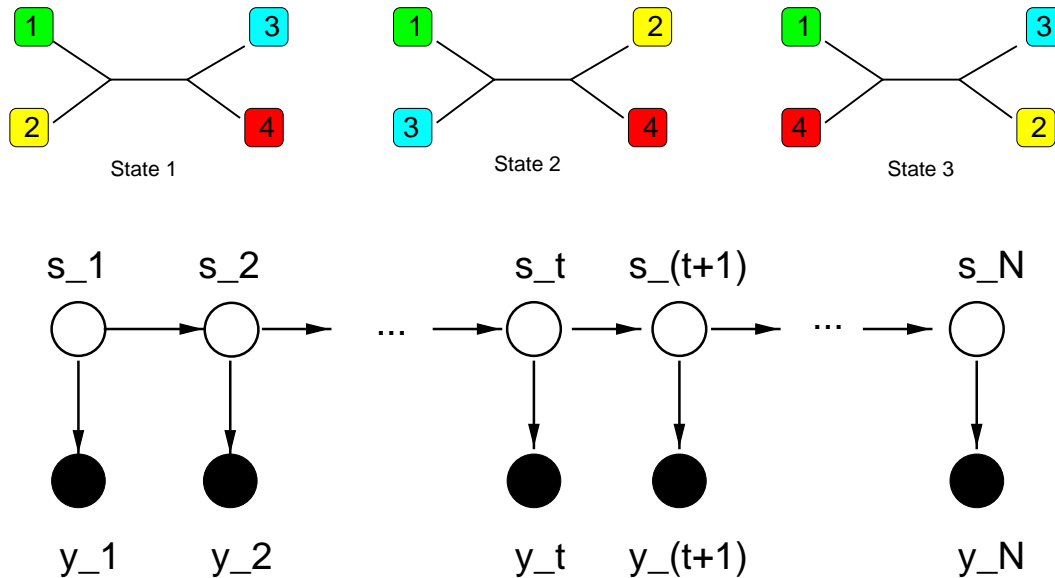


$w$   $\longrightarrow$  Vector of **branch lengths** of all the trees

$\nu$   $\longrightarrow$  Probability of *not* **changing** the tree **topology**

---

## Modelling recombination with HMMs

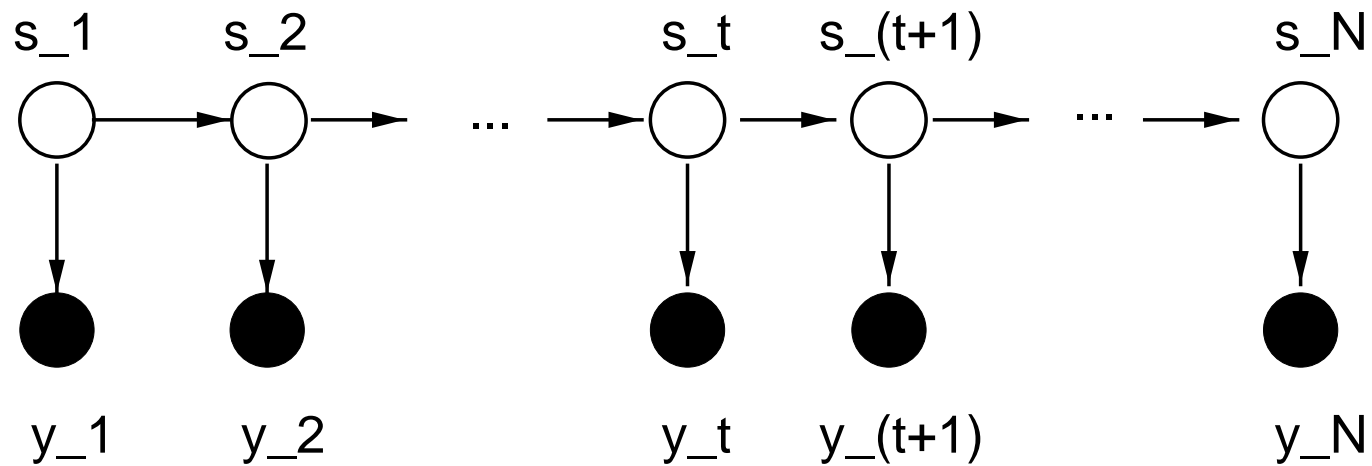


AGCATCGTTCTATTTTACCGGCTCCCG  
 TGTGTCGCTCAAGATTGCCATCGCGCG  
 TGTGTTGGTCTAGATTGCCATCGCGCG  
 TGTATCGCTCTAGTTTGCCAGCTCCCG

Find **optimal sequence**  $S_1, S_2, \dots, S_N \longrightarrow$  **Maximise**  $P(S_1, S_2, \dots, S_N | \mathcal{D})$

## Factorisation in HMMs

---

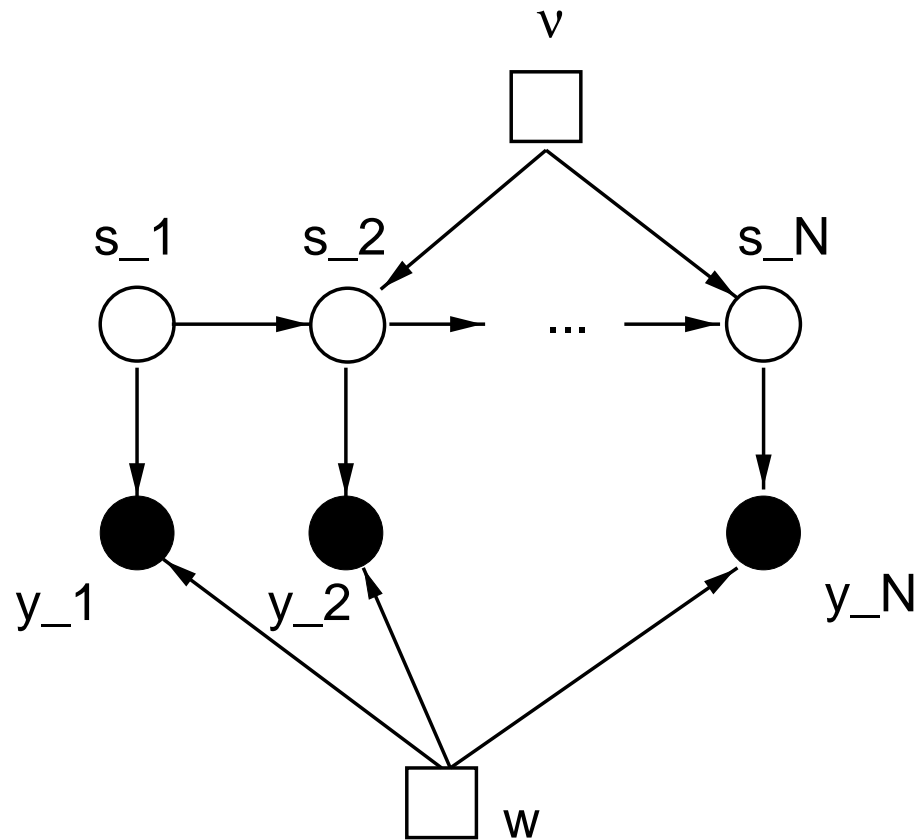


$$\begin{aligned} P(y_1, \dots, y_N, S_1, \dots, S_N) &= \prod_{t=1}^N P(y_t | S_t) \prod_{t=2}^N P(S_t | S_{t-1}) P(S_1) \\ &:= \prod_{t=1}^N P(y_t | S_t) \prod_{t=1}^N P(S_t | S_{t-1}) \end{aligned}$$

---

## HMM parameters

---



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

$v$   $\longrightarrow$  Probability of *not* **changing** the tree **topology**

---

# Parameter estimation

---

---

## Parameter estimation

---

- **Heuristic method**

McGuire, Wright, Prentice (2000)

**Journal of Computational Biology 7**

## Parameter estimation

---

- **Heuristic method**

McGuire, Wright, Prentice (2000)

**Journal of Computational Biology 7**

- **Maximum likelihood (EM algorithm)**

Husmeier, Wright (2001)

**Journal of Computational Biology 8**

---

## Parameter estimation

---

- **Heuristic method**

McGuire, Wright, Prentice (2000)

**Journal of Computational Biology 7**

- **Maximum likelihood (EM algorithm)**

Husmeier, Wright (2001)

**Journal of Computational Biology 8**

- **Bayesian approach**

Husmeier, McGuire (2002)

**Bioinformatics, *to appear***

---

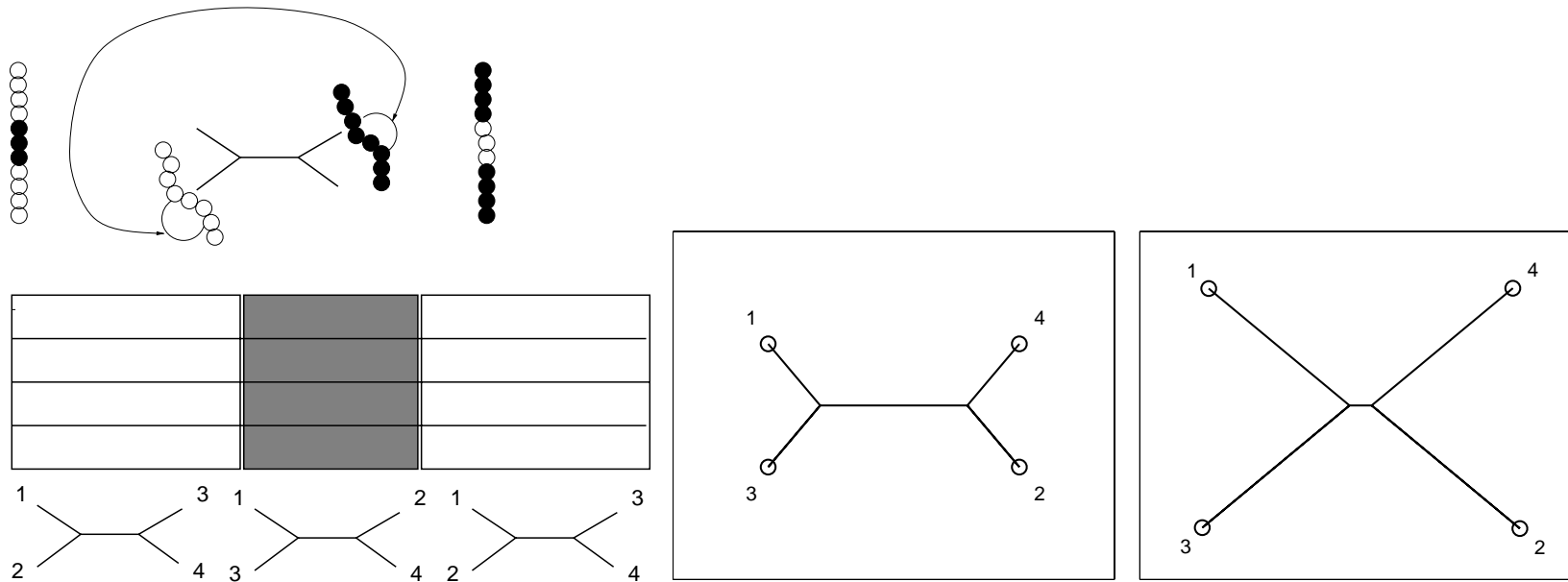
## Heuristic method

---

Optimise the branch lengths  $\mathbf{w}_S$  for each tree topology  $S$  *separately* from the whole alignment.

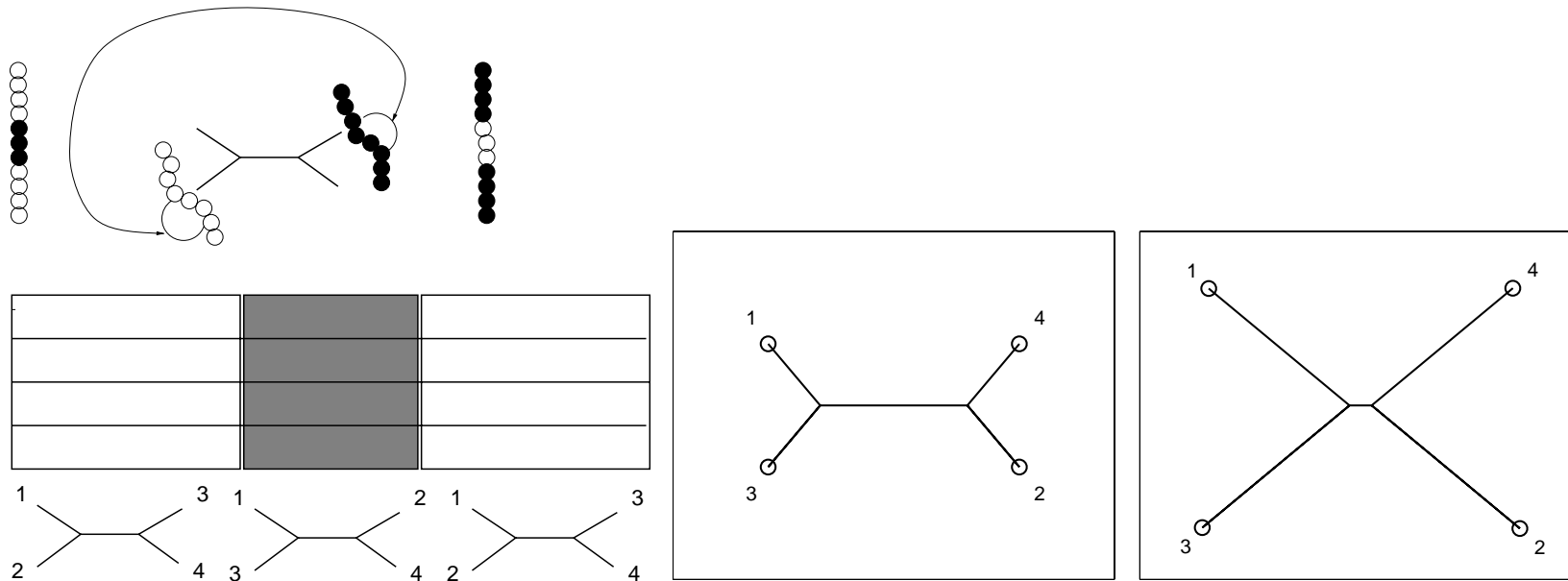
## Heuristic method

Optimise the branch lengths  $\mathbf{w}_S$  for each tree topology  $S$  *separately* from the whole alignment.



## Heuristic method

Optimise the branch lengths  $\mathbf{w}_S$  for each tree topology  $S$  *separately* from the whole alignment.



No optimisation of  $\nu$ .

## Maximum likelihood

---

Likelihood:

$$L(\mathbf{w}, \nu) = \ln P(\mathcal{D}|\mathbf{w}, \nu) = \ln \sum_{\mathbf{S}} P(\mathcal{D}, \mathbf{S}|\mathbf{w}, \nu)$$

## Maximum likelihood

---

Likelihood:

$$L(\mathbf{w}, \nu) = \ln P(\mathcal{D}|\mathbf{w}, \nu) = \ln \sum_{\mathbf{S}} P(\mathcal{D}, \mathbf{S}|\mathbf{w}, \nu)$$

- Requires **marginalisation** over all state sequences  $\mathbf{S} = (S_1, S_2, \dots, S_N)$ .
-

## Maximum likelihood

---

Likelihood:

$$L(\mathbf{w}, \nu) = \ln P(\mathcal{D}|\mathbf{w}, \nu) = \ln \sum_{\mathbf{S}} P(\mathcal{D}, \mathbf{S}|\mathbf{w}, \nu)$$

- Requires **marginalisation** over all state sequences  $\mathbf{S} = (S_1, S_2, \dots, S_N)$ .
  - $K$  states, DNA sequence alignment of length  $N$   
→  $K^N$  state sequences.
-

## Maximum likelihood: EM algorithm

---

$$F(\mathbf{w}, \nu) = \sum_{\mathbf{S}} Q(\mathbf{S}) \ln \frac{P(\mathcal{D}, \mathbf{S} | \mathbf{w}, \nu)}{Q(\mathbf{S})} = \sum_{\mathbf{S}} Q(\mathbf{S}) \ln \frac{P(\mathbf{S} | \mathcal{D}, \mathbf{w}, \nu)}{Q(\mathbf{S})} + \ln P(\mathcal{D} | \mathbf{w}, \nu)$$

## Maximum likelihood: EM algorithm

---

$$F(\mathbf{w}, \nu) = \sum_{\mathbf{S}} Q(\mathbf{S}) \ln \frac{P(\mathcal{D}, \mathbf{S} | \mathbf{w}, \nu)}{Q(\mathbf{S})} = \sum_{\mathbf{S}} Q(\mathbf{S}) \ln \frac{P(\mathbf{S} | \mathcal{D}, \mathbf{w}, \nu)}{Q(\mathbf{S})} + \ln P(\mathcal{D} | \mathbf{w}, \nu)$$

$$L(\mathbf{w}, \nu) = F(\mathbf{w}, \nu) + KL[Q, P]$$

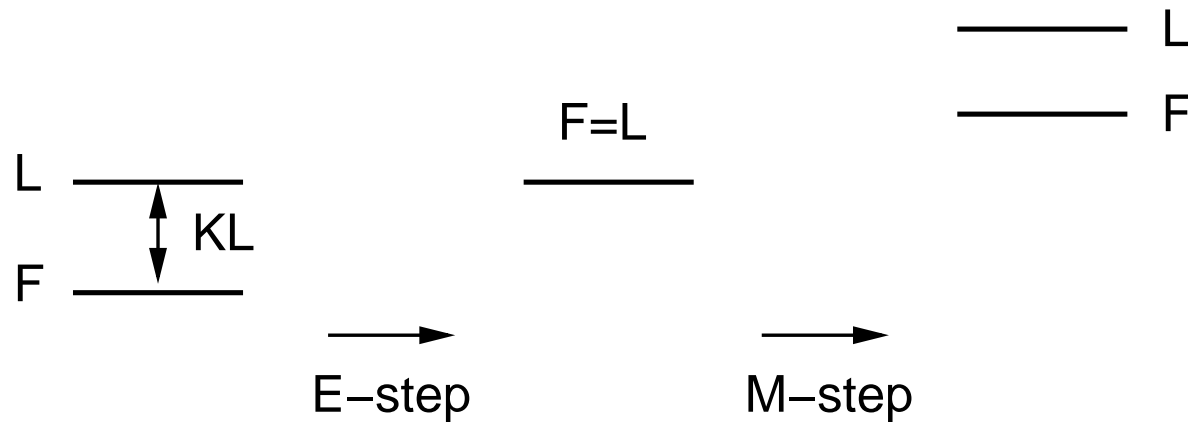
---

## Maximum likelihood: EM algorithm

---

$$F(\mathbf{w}, \nu) = \sum_{\mathbf{S}} Q(\mathbf{S}) \ln \frac{P(\mathcal{D}, \mathbf{S} | \mathbf{w}, \nu)}{Q(\mathbf{S})} = \sum_{\mathbf{S}} Q(\mathbf{S}) \ln \frac{P(\mathbf{S} | \mathcal{D}, \mathbf{w}, \nu)}{Q(\mathbf{S})} + \ln P(\mathcal{D} | \mathbf{w}, \nu)$$

$$L(\mathbf{w}, \nu) = F(\mathbf{w}, \nu) + KL[Q, P]$$

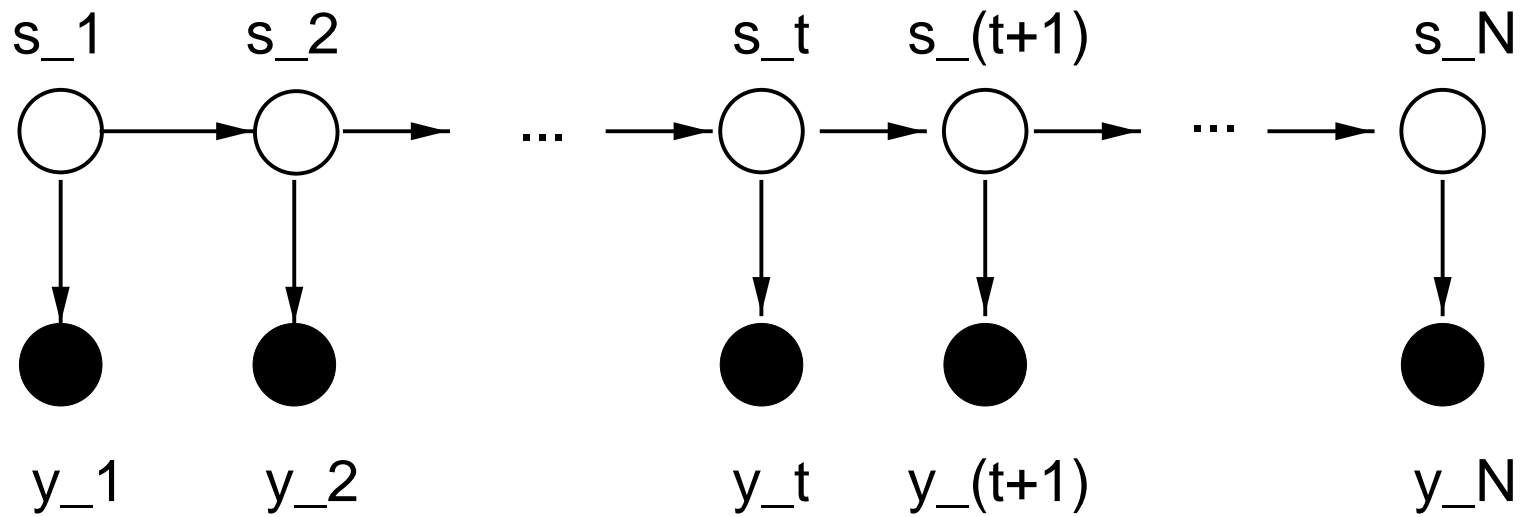


---

**E-step**  $\longrightarrow Q(\mathbf{S}) = P(\mathbf{S} | \mathcal{D}, \mathbf{w}, \nu)$   
**M-step**  $\longrightarrow$  Maximise  $F(\mathbf{w}, \nu)$

---

## HMM: Factorisation



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

---

M-step (for  $w$ )

---

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

$$F(\mathbf{w}, \nu) = \sum_{\mathbf{S}} Q(\mathbf{S}) \ln \frac{P(\mathcal{D}, \mathbf{S} | \mathbf{w}, \nu)}{Q(\mathbf{S})}$$

---

M-step (for  $\mathbf{w}$ )

---

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

$$F(\mathbf{w}, \nu) = \sum_{\mathbf{S}} Q(\mathbf{S}) \ln \frac{P(\mathcal{D}, \mathbf{S} | \mathbf{w}, \nu)}{Q(\mathbf{S})}$$

$$F(\mathbf{w}, \nu) = \sum_{\mathbf{S}} Q(\mathbf{S}) \sum_{t=1}^N \ln P(\mathbf{y}_t | S_t, \mathbf{w}) + C$$

---

M-step (for  $\mathbf{w}$ )

---

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

$$F(\mathbf{w}, \nu) = \sum_{\mathbf{S}} Q(\mathbf{S}) \ln \frac{P(\mathcal{D}, \mathbf{S} | \mathbf{w}, \nu)}{Q(\mathbf{S})}$$

$$F(\mathbf{w}, \nu) = \sum_{\mathbf{S}} Q(\mathbf{S}) \sum_{t=1}^N \ln P(\mathbf{y}_t | S_t, \mathbf{w}) + C$$

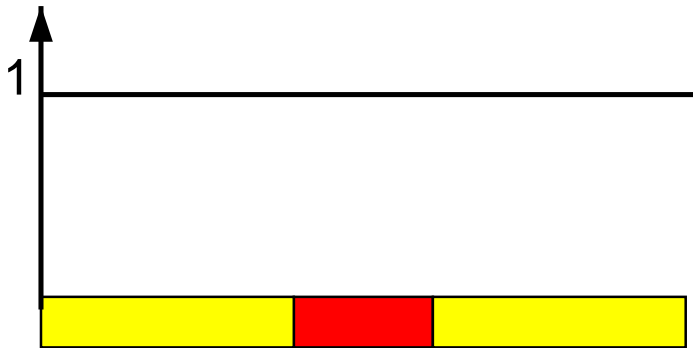
$$F(\mathbf{w}, \nu) = \sum_{t=1}^N \sum_{S_t=1}^K Q(S_t) \ln P(\mathbf{y}_t | S_t, \mathbf{w}) + C$$

---

## Illustration: EM

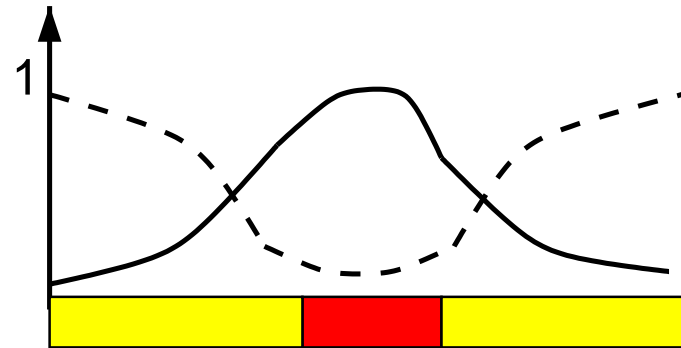
---

$Q(s_t)$



Standard

$Q(s_t)$

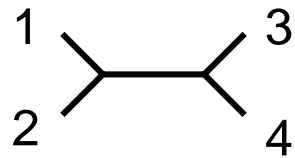


EM

E-step:  $Q(S_t) \longrightarrow P(S_t | \mathcal{D}, \mathbf{w}, \nu)$

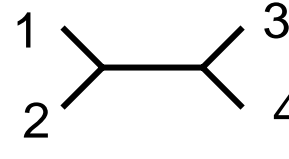
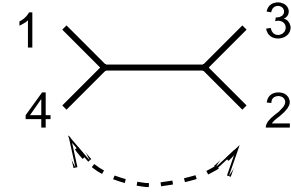
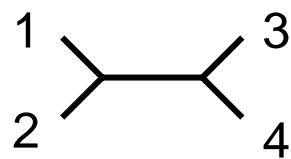
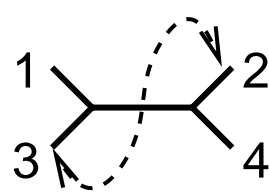
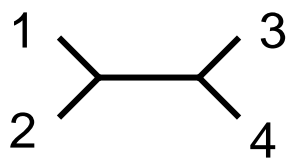
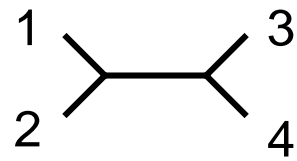
---

## Synthetic example

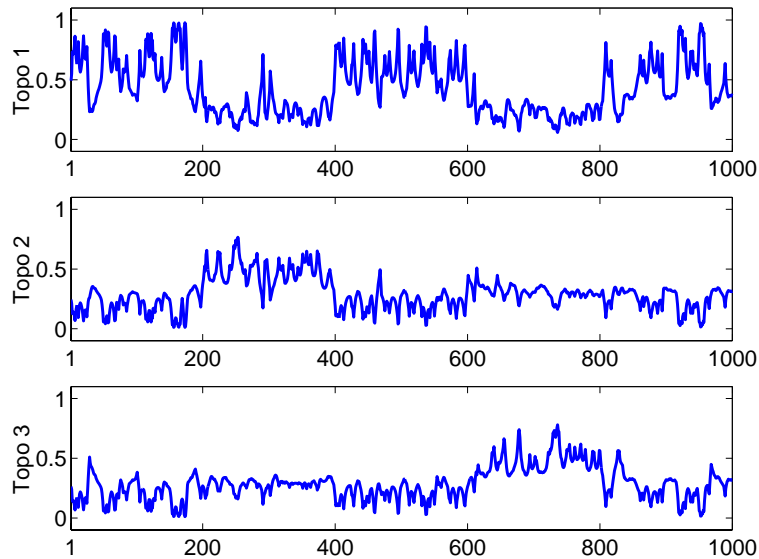
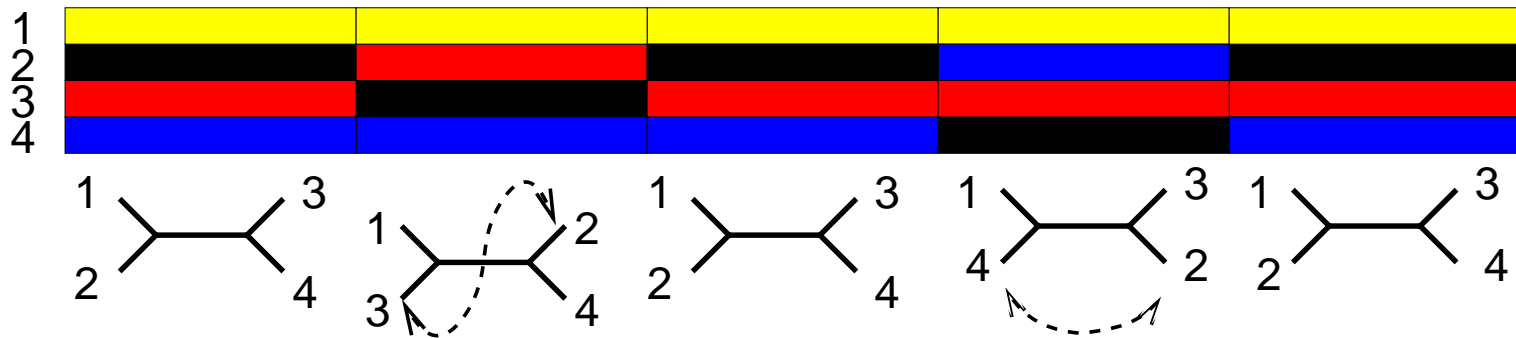


- Model of nucleotide substitution: Kimura 2-parameter,  $\tau = 2$ .
- Alignment of length  $N = 1000$  nucleotides.

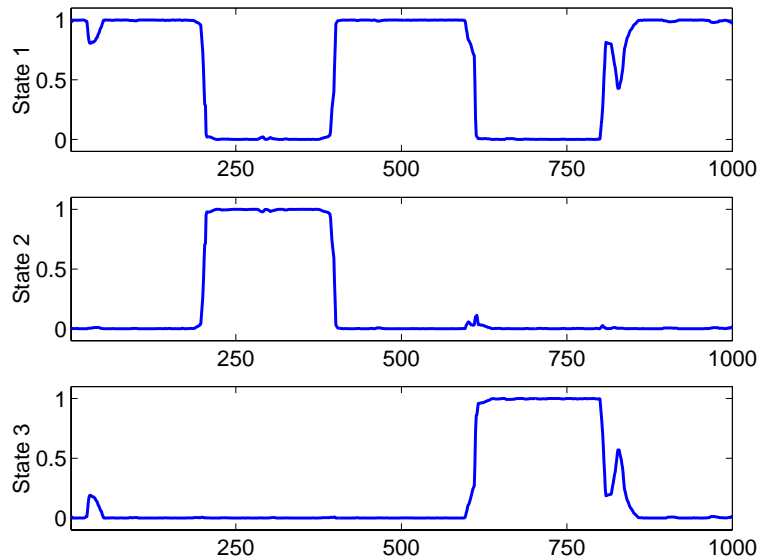
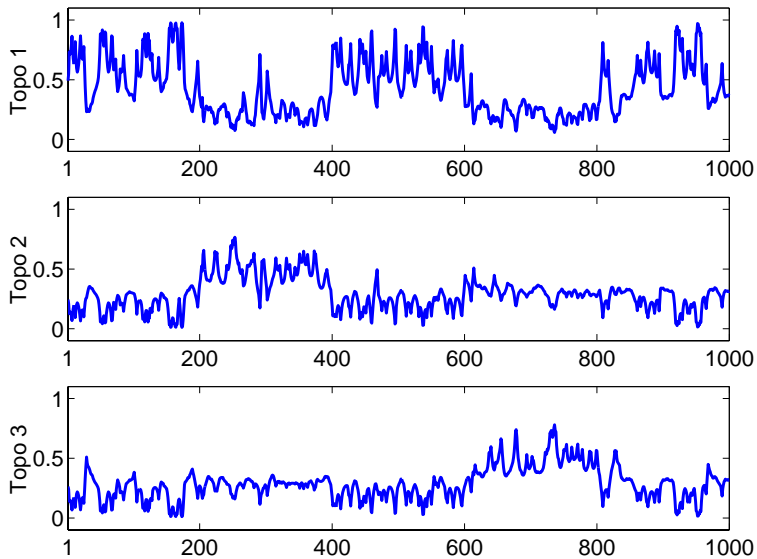
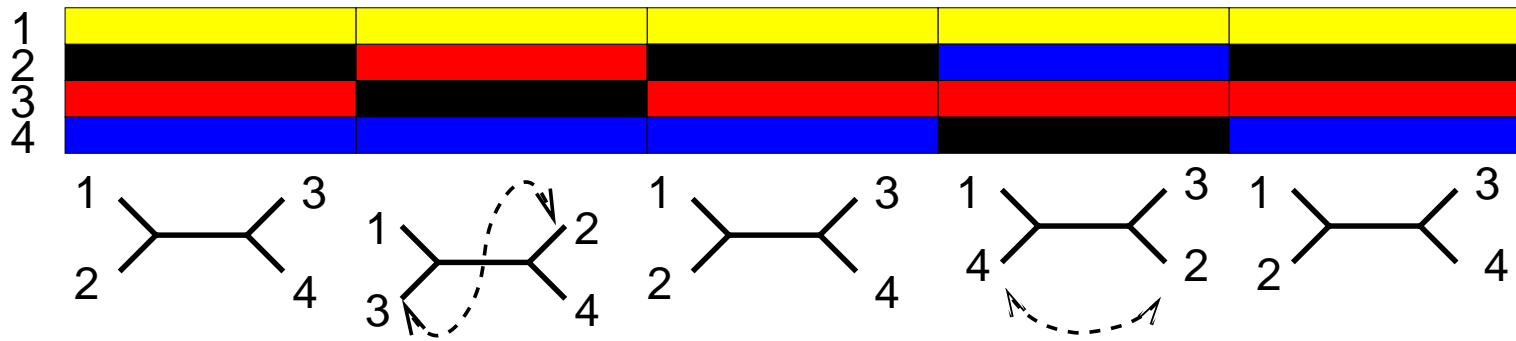
# Synthetic example



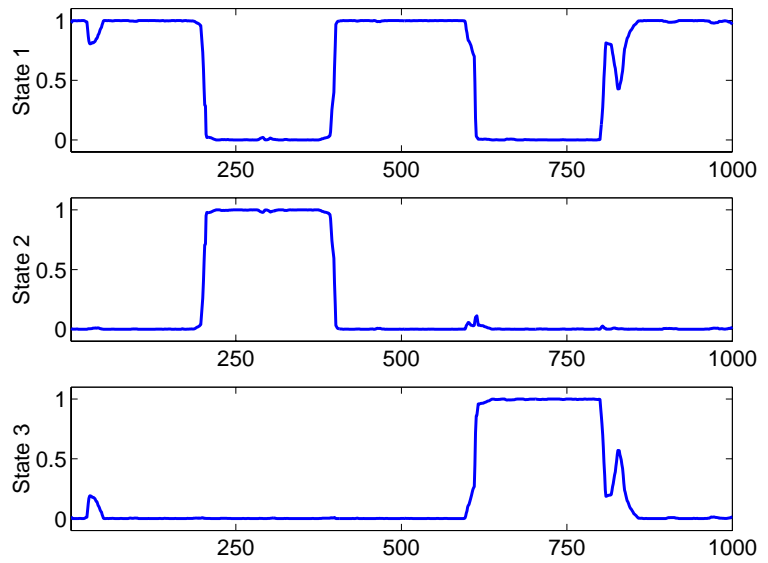
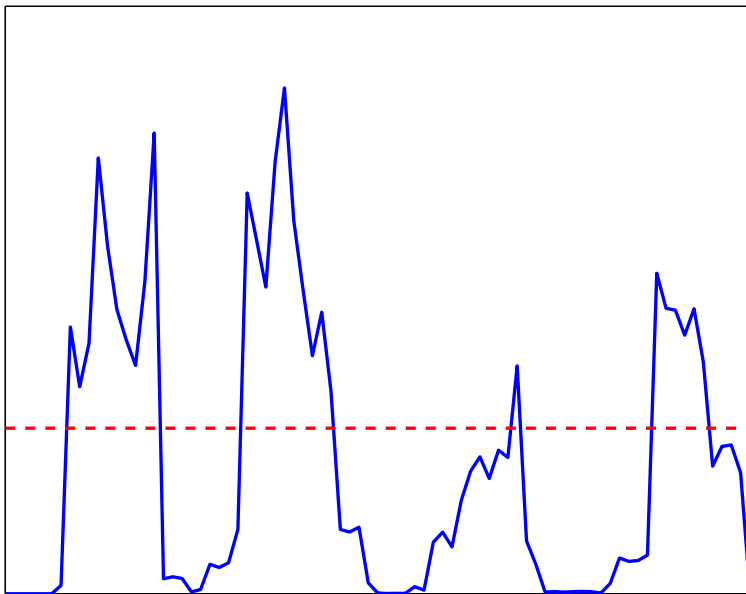
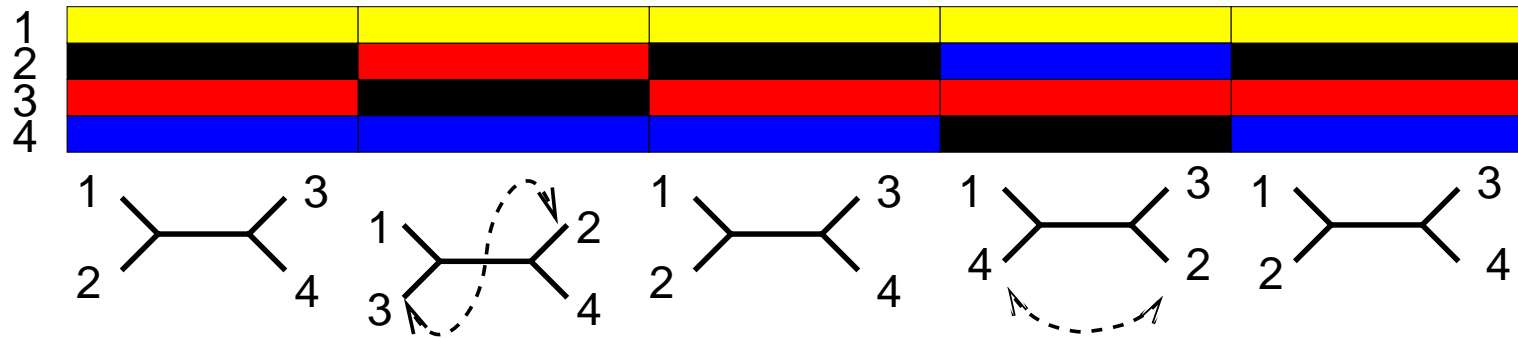
$P(S_t|\mathcal{D})$ : Heuristic method ( $\nu = 0.8$ )



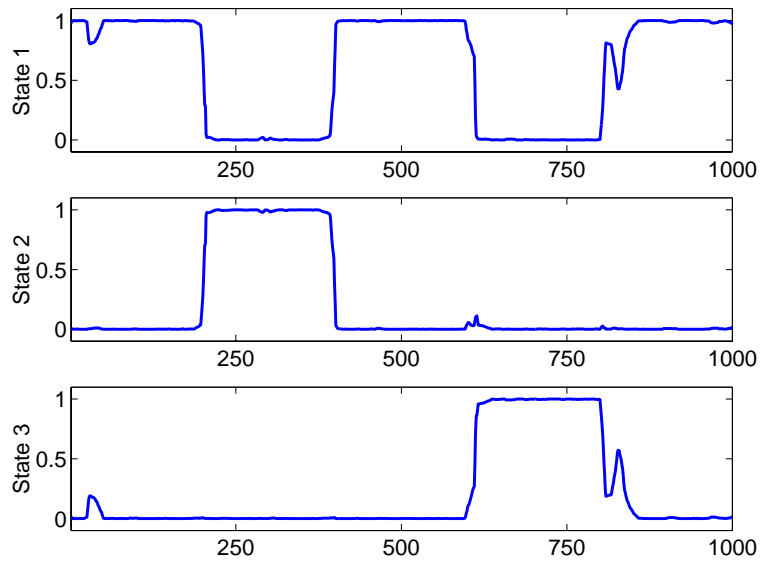
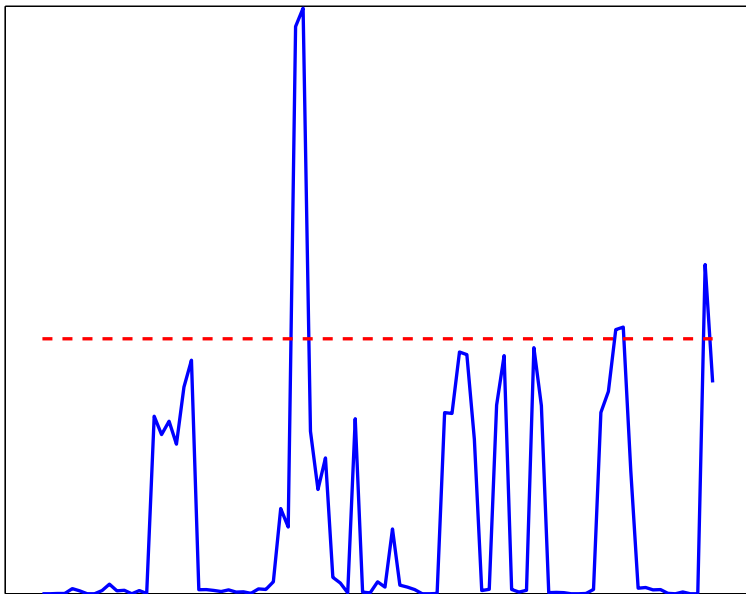
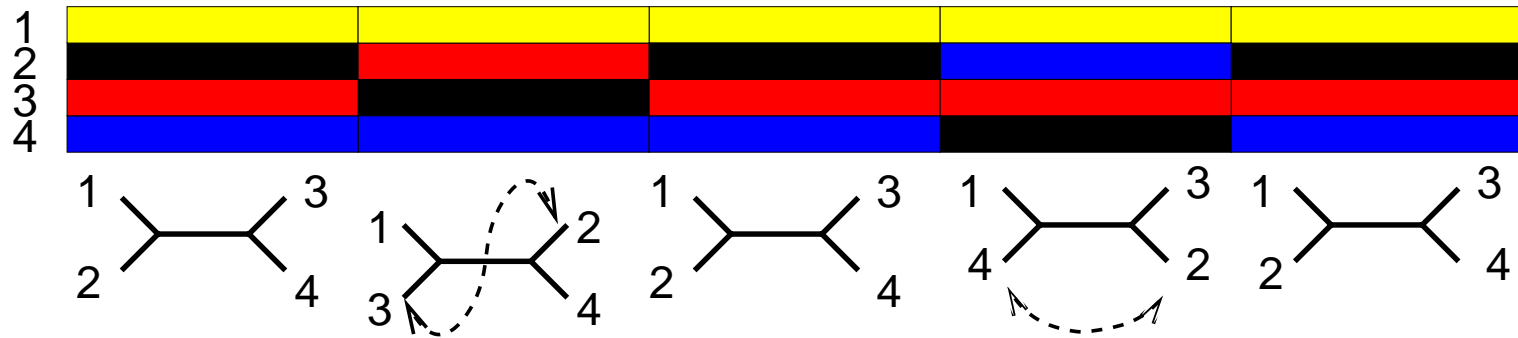
# $P(S_t|\mathcal{D})$ : Heuristic method vs. maximum likelihood



# HMM-ML vs. Topal, window size=200



# HMM-ML vs. Topal, window size=100



## Disadvantages of maximum likelihood

---

---

## Disadvantages of maximum likelihood

---

- ML:  $P(\mathbf{S}|\mathcal{D}, \hat{\mathbf{w}}, \hat{\nu})$

## Disadvantages of maximum likelihood

---

- ML:  $P(\mathbf{S}|\mathcal{D}, \hat{\mathbf{w}}, \hat{\nu})$
- Possibility of over-fitting.
- Separate hypothesis testing required, e.g., using parametric bootstrapping.

## Disadvantages of maximum likelihood

---

- ML:  $P(\mathbf{S}|\mathcal{D}, \hat{\mathbf{w}}, \hat{\nu})$
- Possibility of over-fitting.
- Separate hypothesis testing required, e.g., using parametric bootstrapping.

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

## Bayesian approach

---

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

## 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**  $P(\mathbf{w}, \nu|\mathcal{D}) \longleftarrow$  **Prior**  $P(\mathbf{w}, \nu) = \prod_i P(w_i)P(\nu)$
- $P(w_i) = \left[ \begin{array}{l} 1/\Omega \text{ if } 0 \leq w_i \leq \Omega \\ 0 \text{ otherwise} \end{array} \right|$

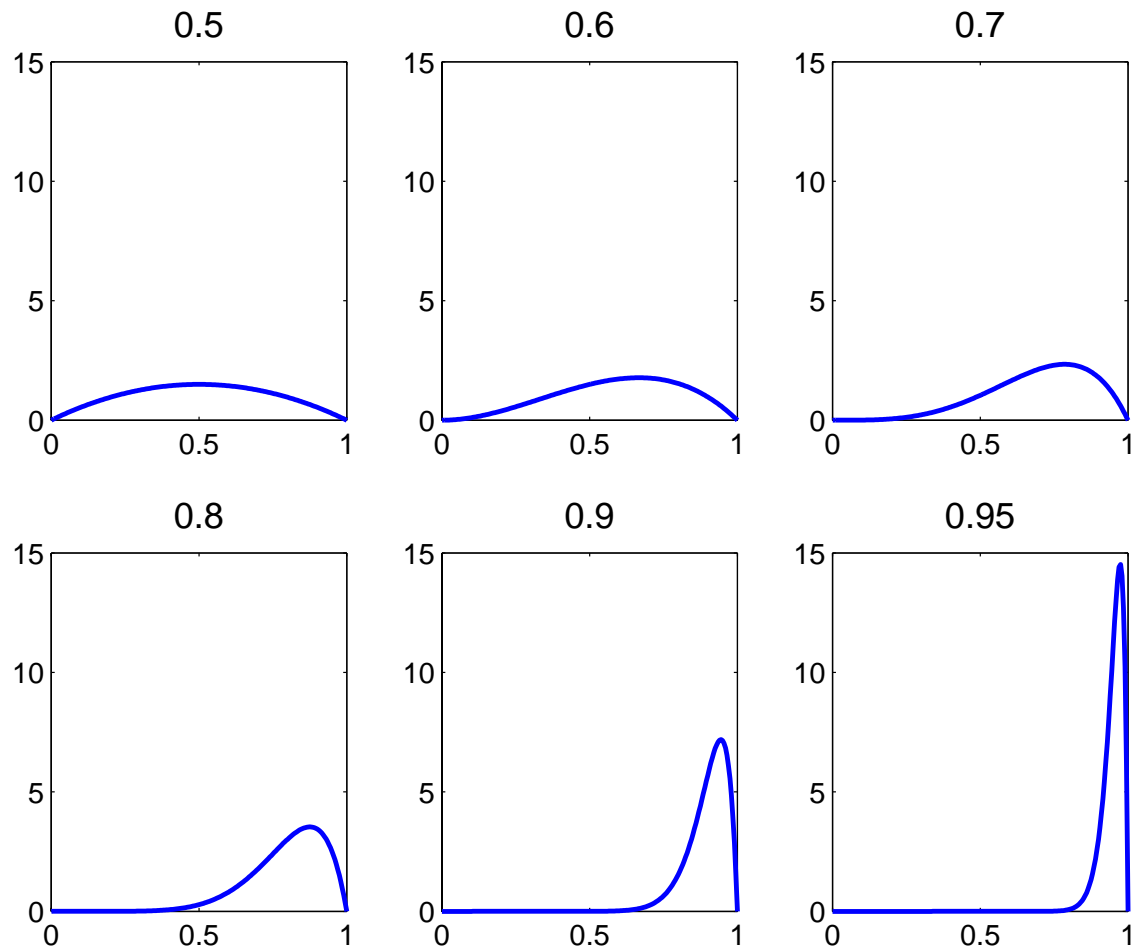
## Bayesian approach

---

- $P(\mathbf{S}|\mathcal{D}) = \int P(\mathbf{S}|\mathcal{D}, \mathbf{w}, \nu)P(\mathbf{w}, \nu|\mathcal{D})d\mathbf{w}d\nu$
  - **Posterior**  $P(\mathbf{w}, \nu|\mathcal{D}) \longleftarrow$  **Prior**  $P(\mathbf{w}, \nu) = \prod_i P(w_i)P(\nu)$
  - $P(w_i) = \begin{cases} 1/\Omega & \text{if } 0 \leq w_i \leq \Omega \\ 0 & \text{otherwise} \end{cases}$
  - $P(\nu) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\nu^{\alpha-1}(1-\nu)^{\beta-1}$   
Conjugate prior: Beta distribution.
-

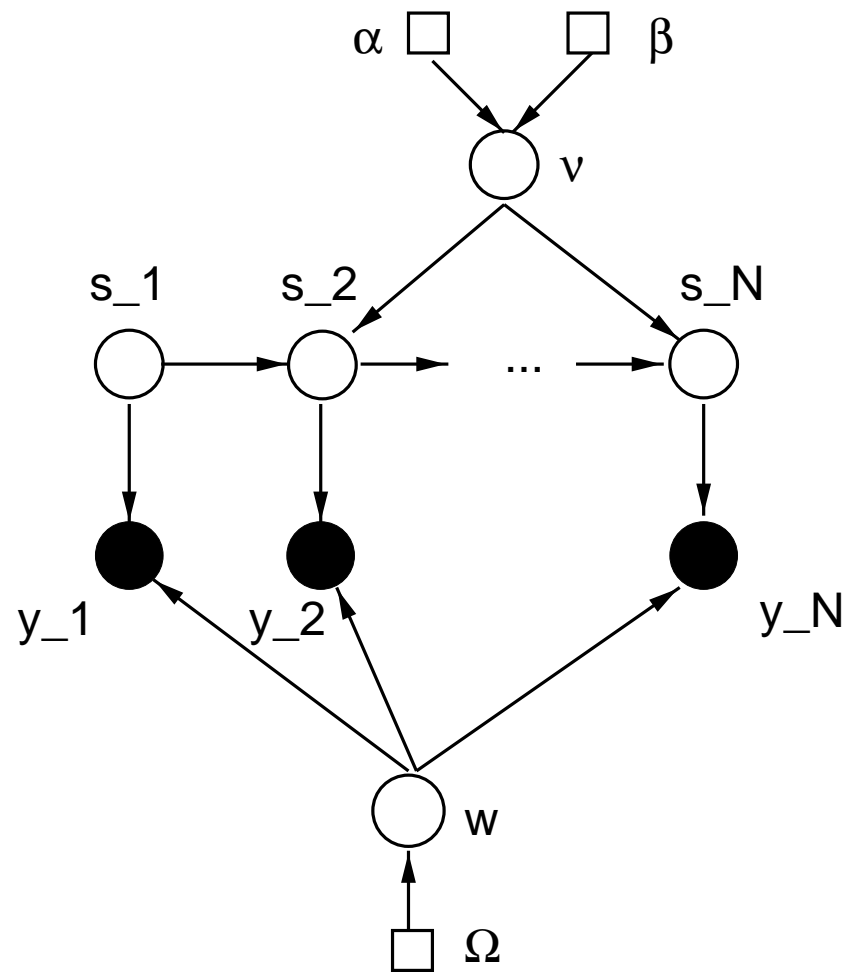
Beta Prior,  $\beta = 2$ ,  $\mu = \alpha / (\alpha + \beta)$

---



## Bayesian approach

---



## Sampling from the posterior distribution

---

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

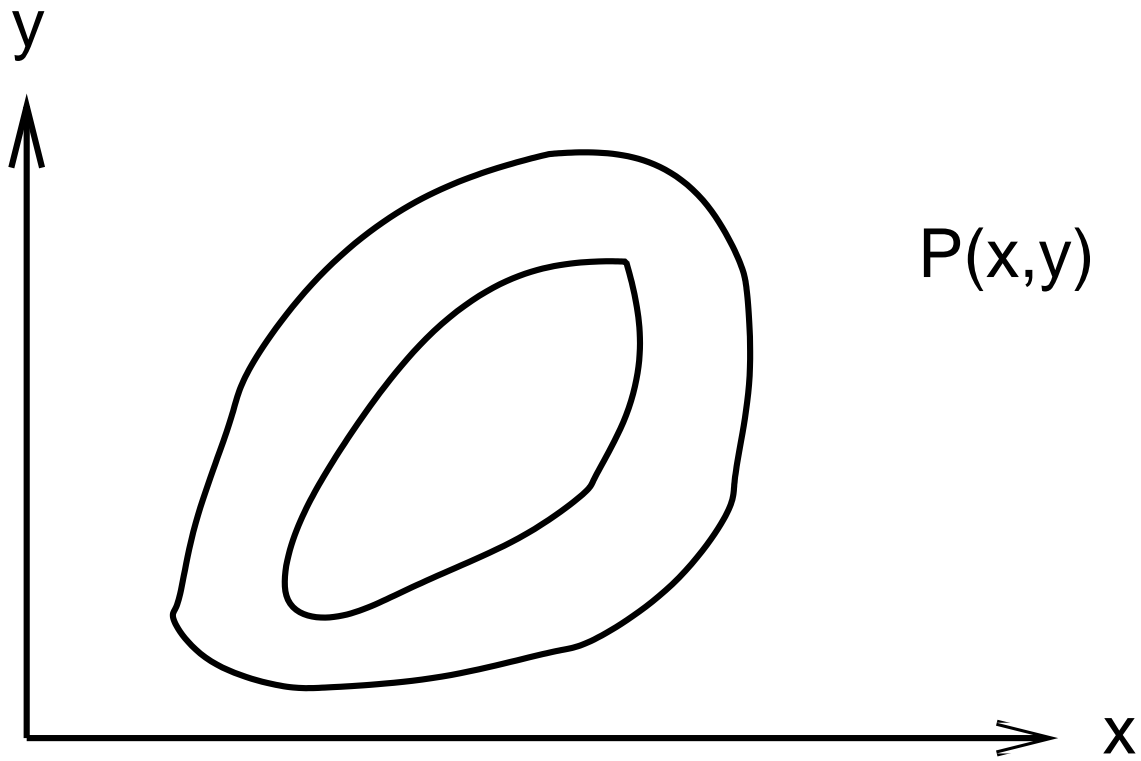
## Sampling from the posterior distribution

---

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

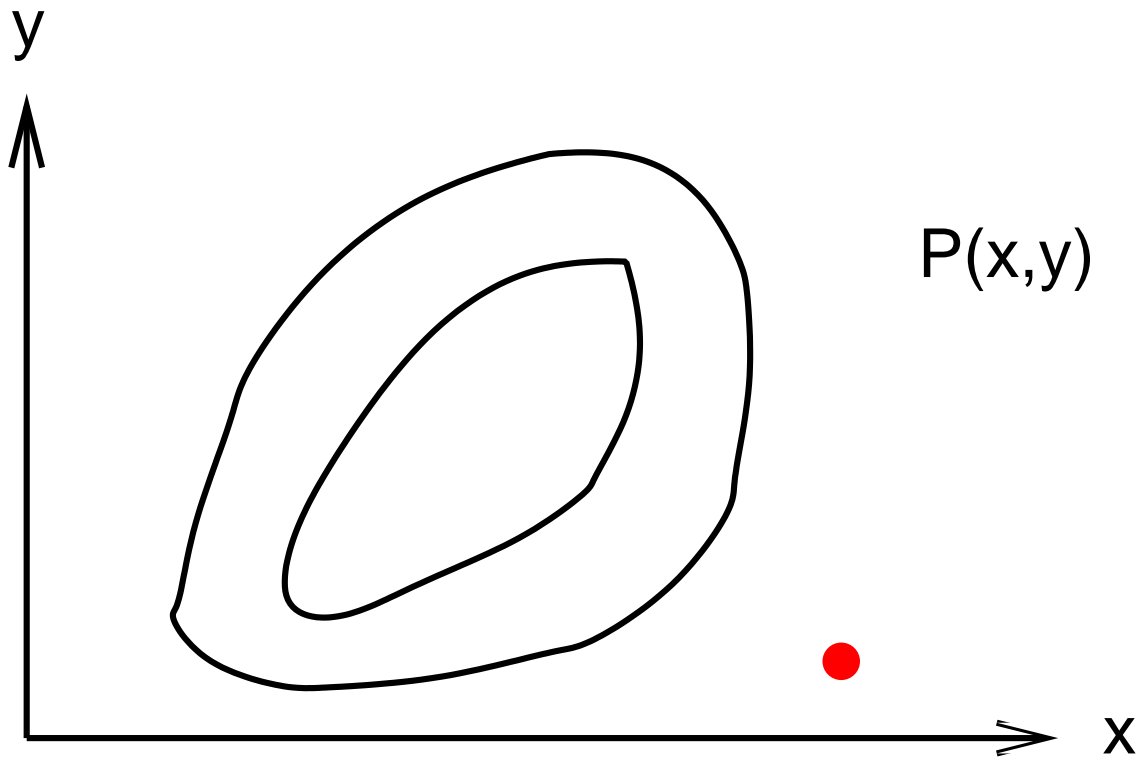
# Gibbs sampling

---

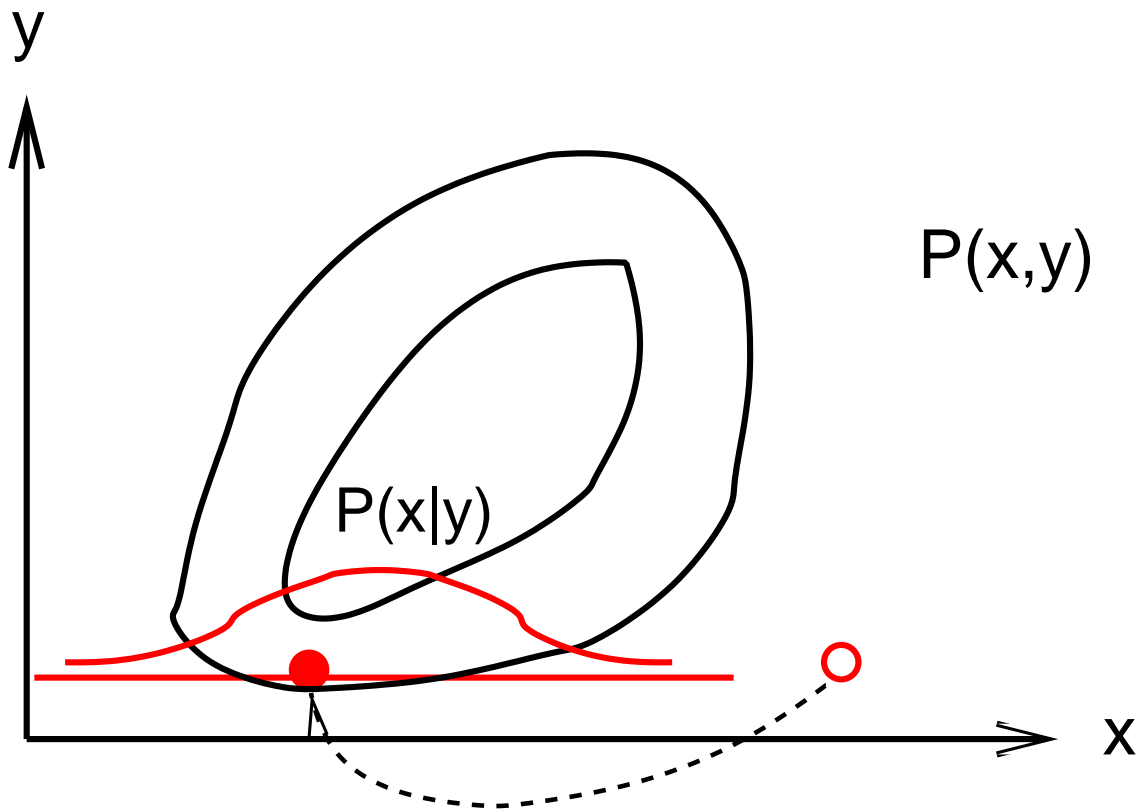


# Gibbs sampling

---

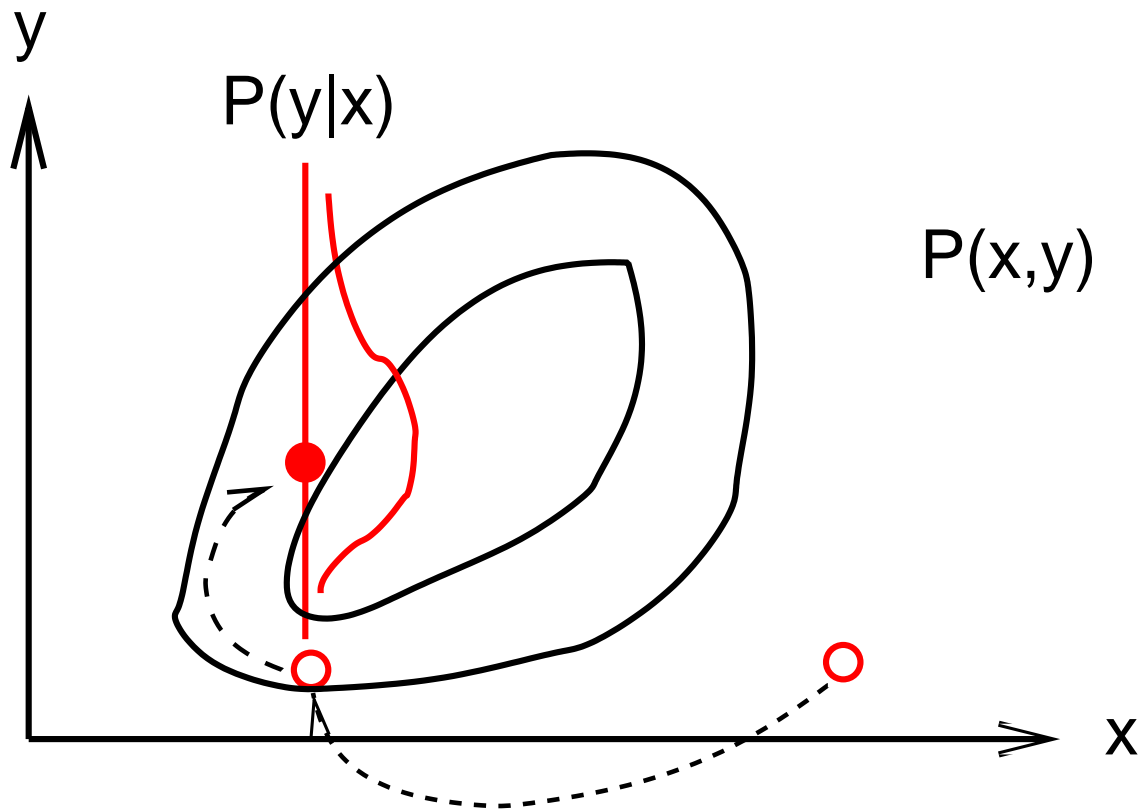


# Gibbs sampling



# Gibbs sampling

---





## 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:
  - $\mathbf{S} \sim P(\mathbf{S} | \mathbf{w}, \nu, \mathcal{D})$
  - $\mathbf{w} \sim P(\mathbf{w} | \mathbf{S}, \nu, \mathcal{D})$
  - $\nu \sim P(\nu | \mathbf{S}, \mathbf{w}, \mathcal{D})$
- $\nu$ : Sample from Beta distribution

## Sampling from the posterior distribution

---

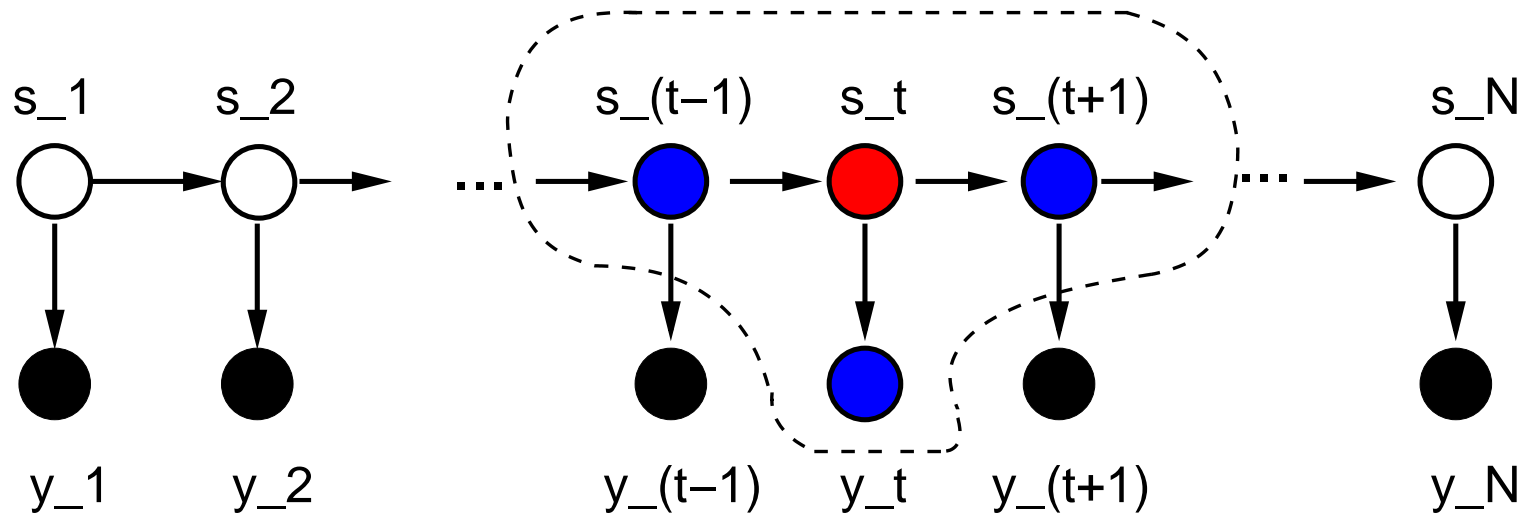
- Sampling from  $P(\mathbf{S}, \mathbf{w}, \nu | \mathcal{D})$
  - Gibbs-like approach:
    - $\mathbf{S} \sim P(\mathbf{S} | \mathbf{w}, \nu, \mathcal{D})$
    - $\mathbf{w} \sim P(\mathbf{w} | \mathbf{S}, \nu, \mathcal{D})$
    - $\nu \sim P(\nu | \mathbf{S}, \mathbf{w}, \mathcal{D})$
  - $\nu$ :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

## Sampling from the posterior distribution

---

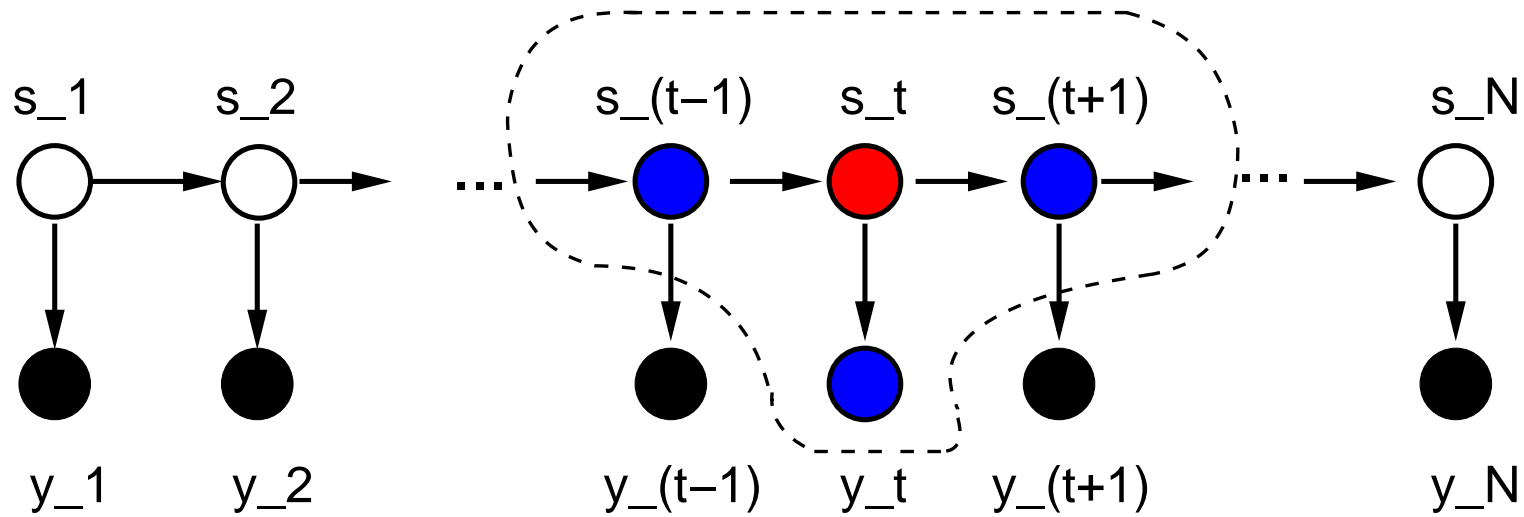
- Sampling from  $P(\mathbf{S}, \mathbf{w}, \nu | \mathcal{D})$
  - Gibbs-like approach:
    - $\mathbf{S} \sim P(\mathbf{S} | \mathbf{w}, \nu, \mathcal{D})$
    - $\mathbf{w} \sim P(\mathbf{w} | \mathbf{S}, \nu, \mathcal{D})$
    - $\nu \sim P(\nu | \mathbf{S}, \mathbf{w}, \mathcal{D})$
  - $\nu$ :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



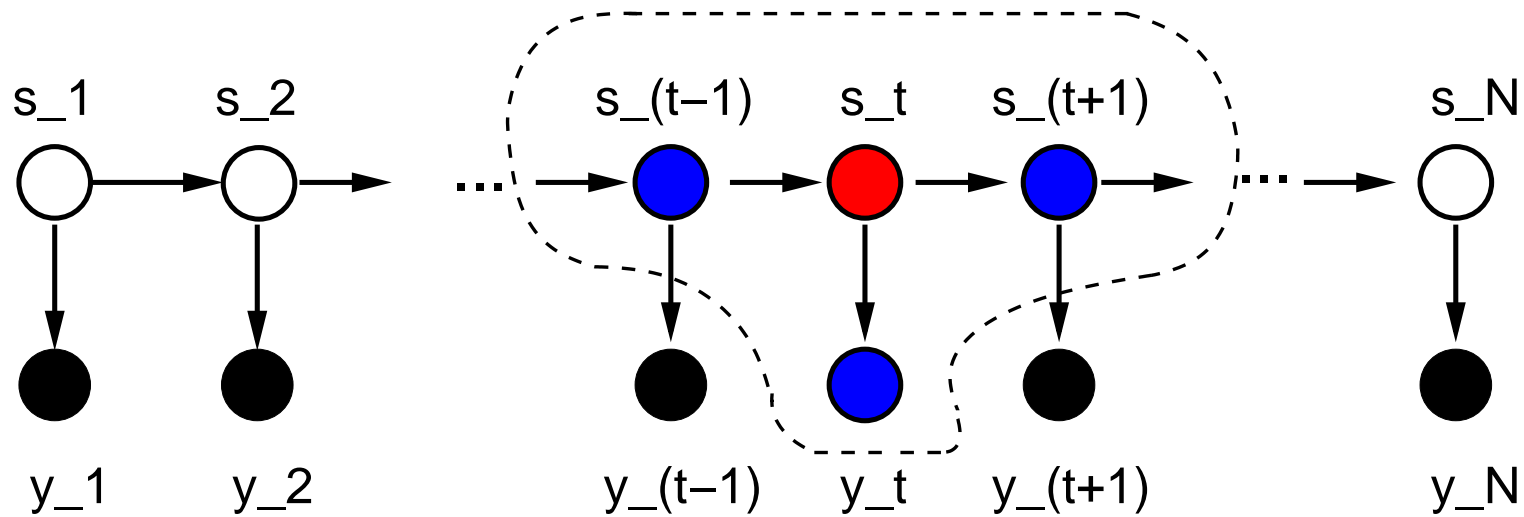
$$P(S_t | S_1, \dots, S_{t-1}, S_{t+1}, \dots, S_N, \mathcal{D}, \mathbf{w}, \nu)$$

## Sampling from the posterior distribution



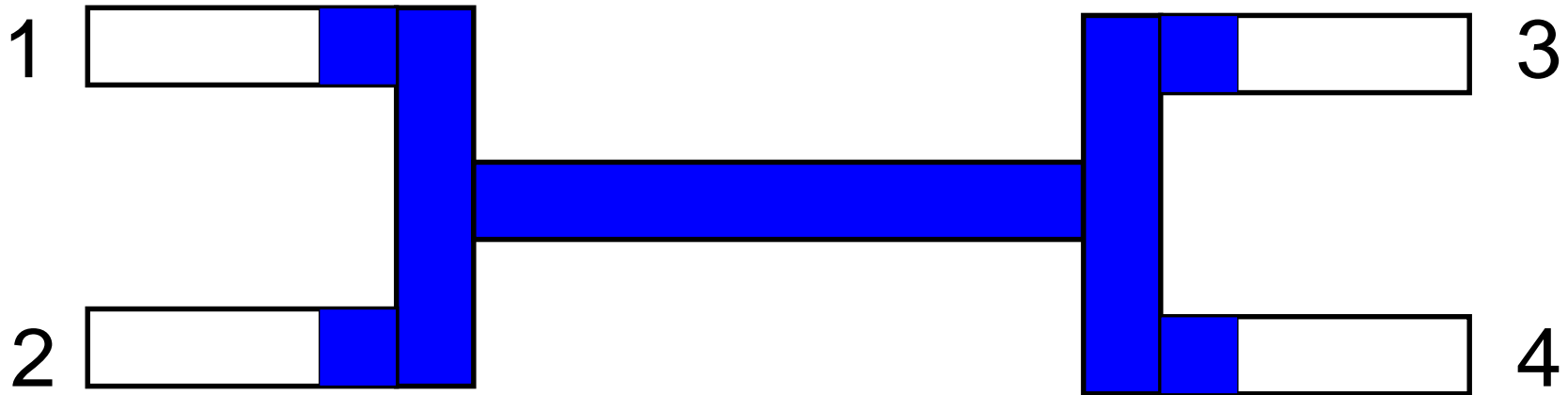
$$\begin{aligned} P(S_t | S_1, \dots, S_{t-1}, S_{t+1}, \dots, S_N, \mathcal{D}, \mathbf{w}, \nu) \\ = P(S_t | S_{t-1}, S_{t+1}, \mathbf{y}_t, \mathbf{w}, \nu) \end{aligned}$$

## Sampling from the posterior distribution

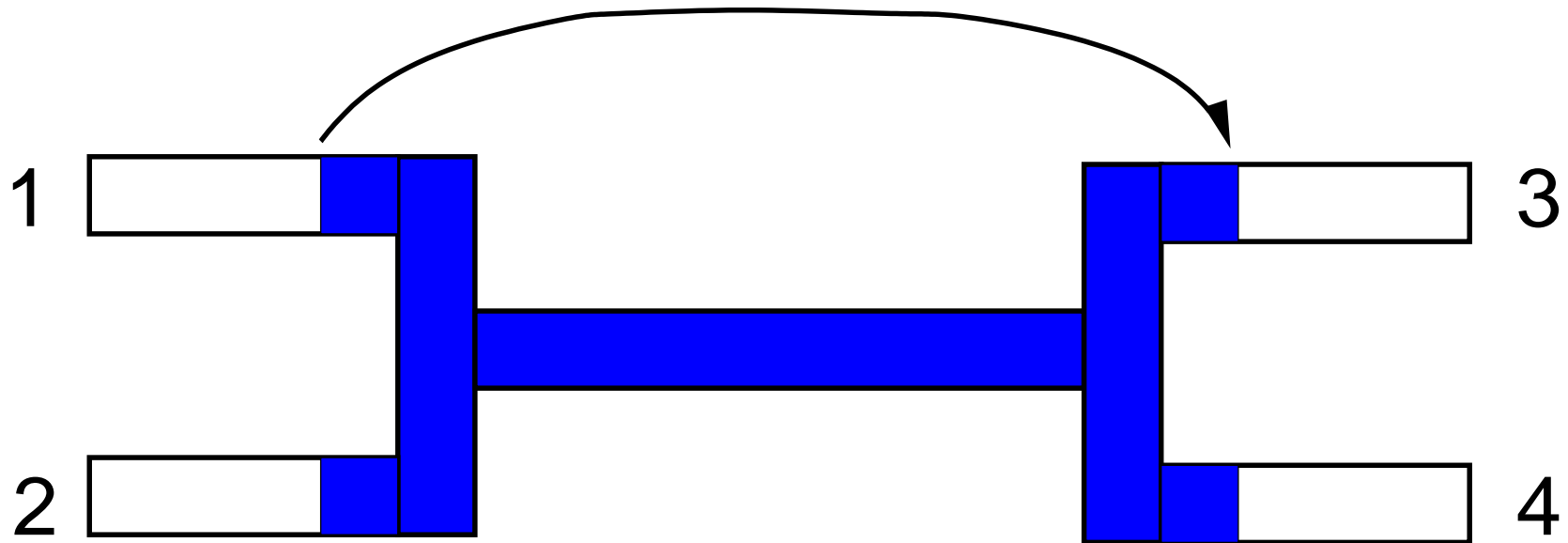


$$\begin{aligned} P(\mathbf{S}_t | S_1, \dots, S_{t-1}, S_{t+1}, \dots, S_N, \mathcal{D}, \mathbf{w}, \nu) \\ &= P(\mathbf{S}_t | S_{t-1}, S_{t+1}, \mathbf{y}_t, \mathbf{w}, \nu) \\ &\propto P(S_{t+1} | S_t, \nu) P(S_t | S_{t-1}, \nu) P(\mathbf{y}_t | S_t, \mathbf{w}) \end{aligned}$$

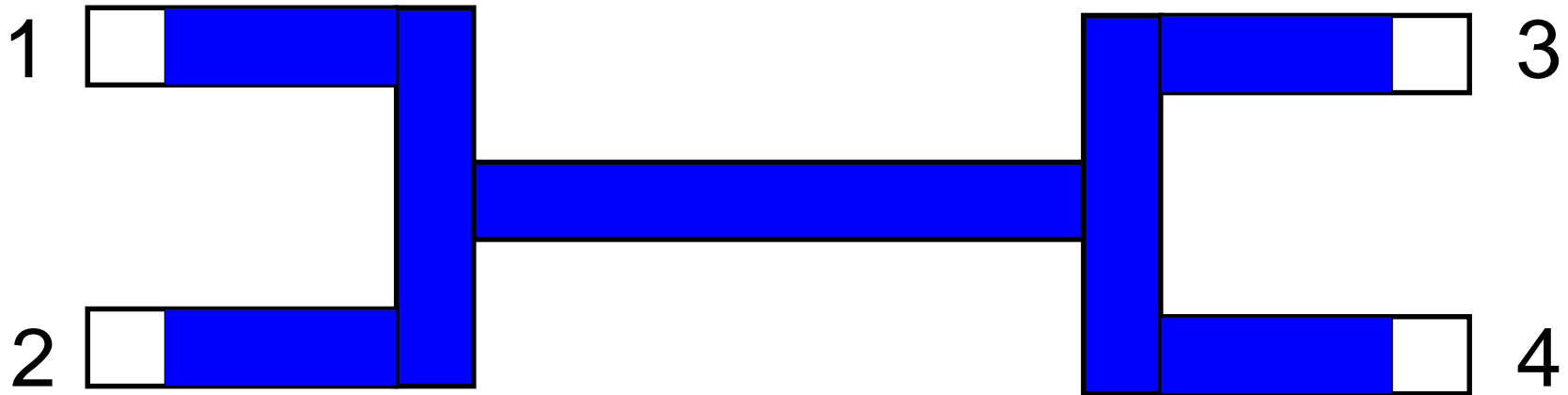
## Simulation of recombination



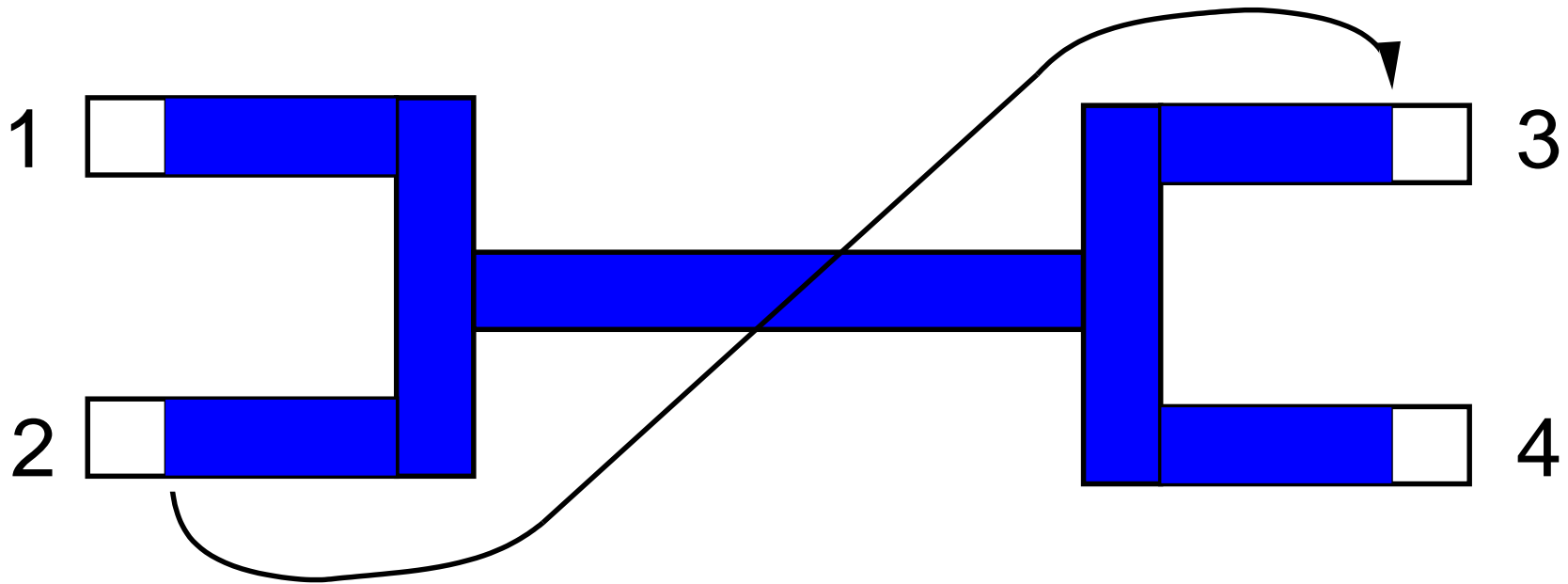
## Simulation of recombination



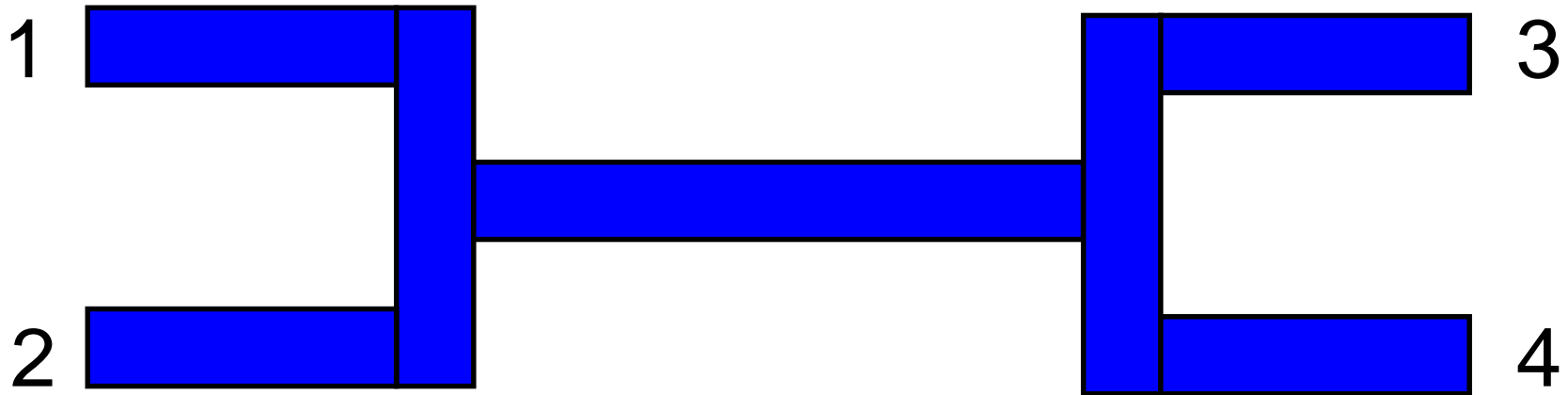
## Simulation of recombination



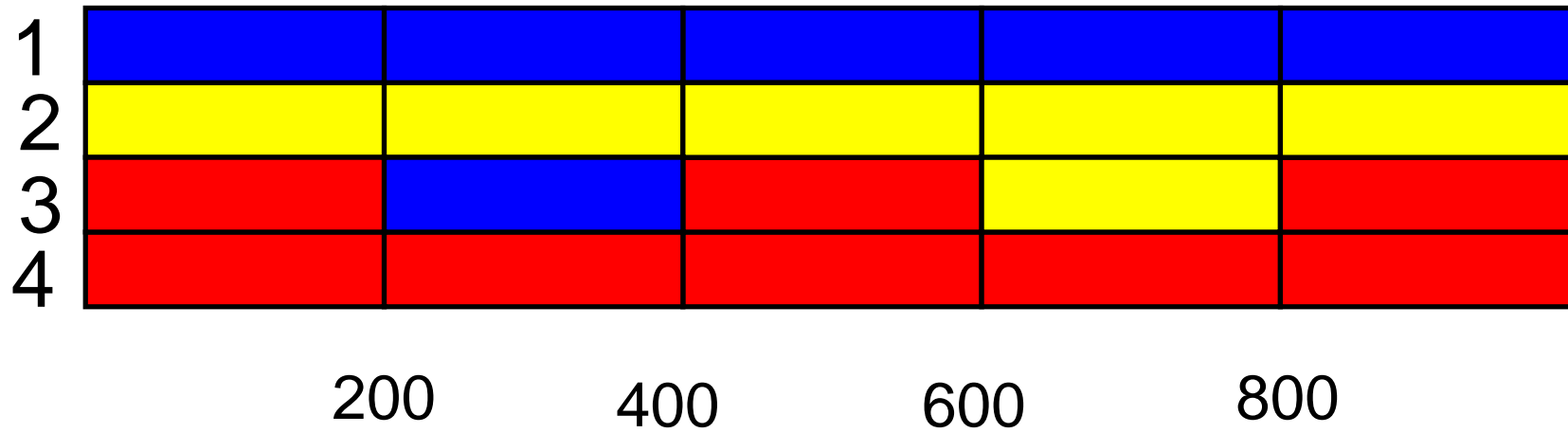
## Simulation of recombination



## Simulation of recombination

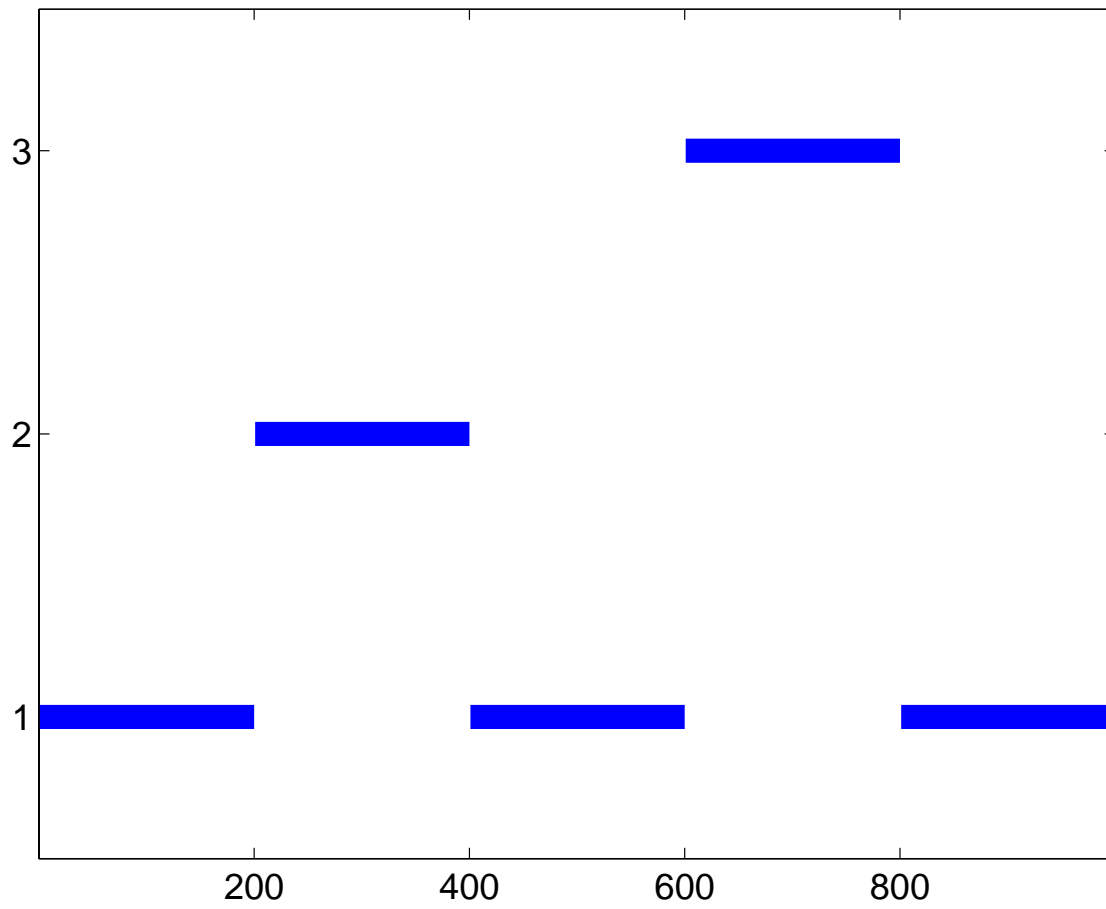


## Simulation of recombination



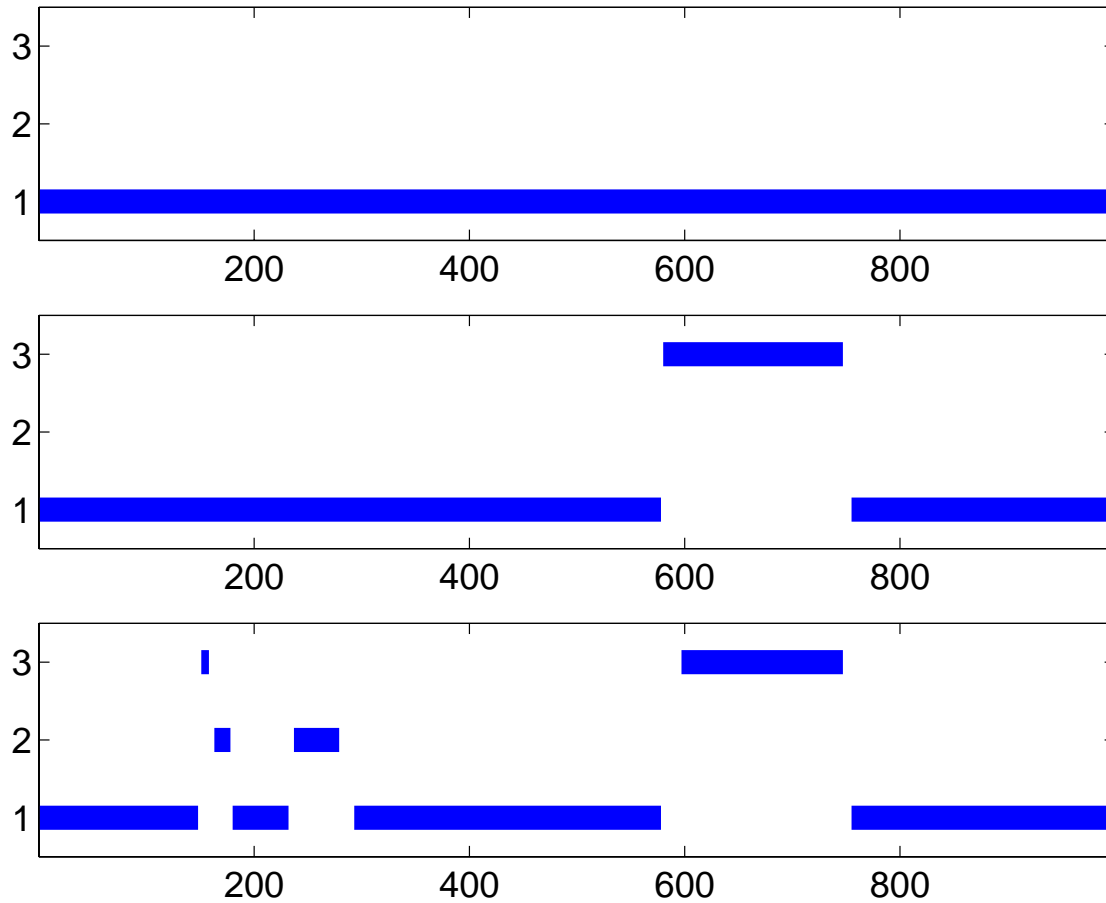
## True mosaic structure

---



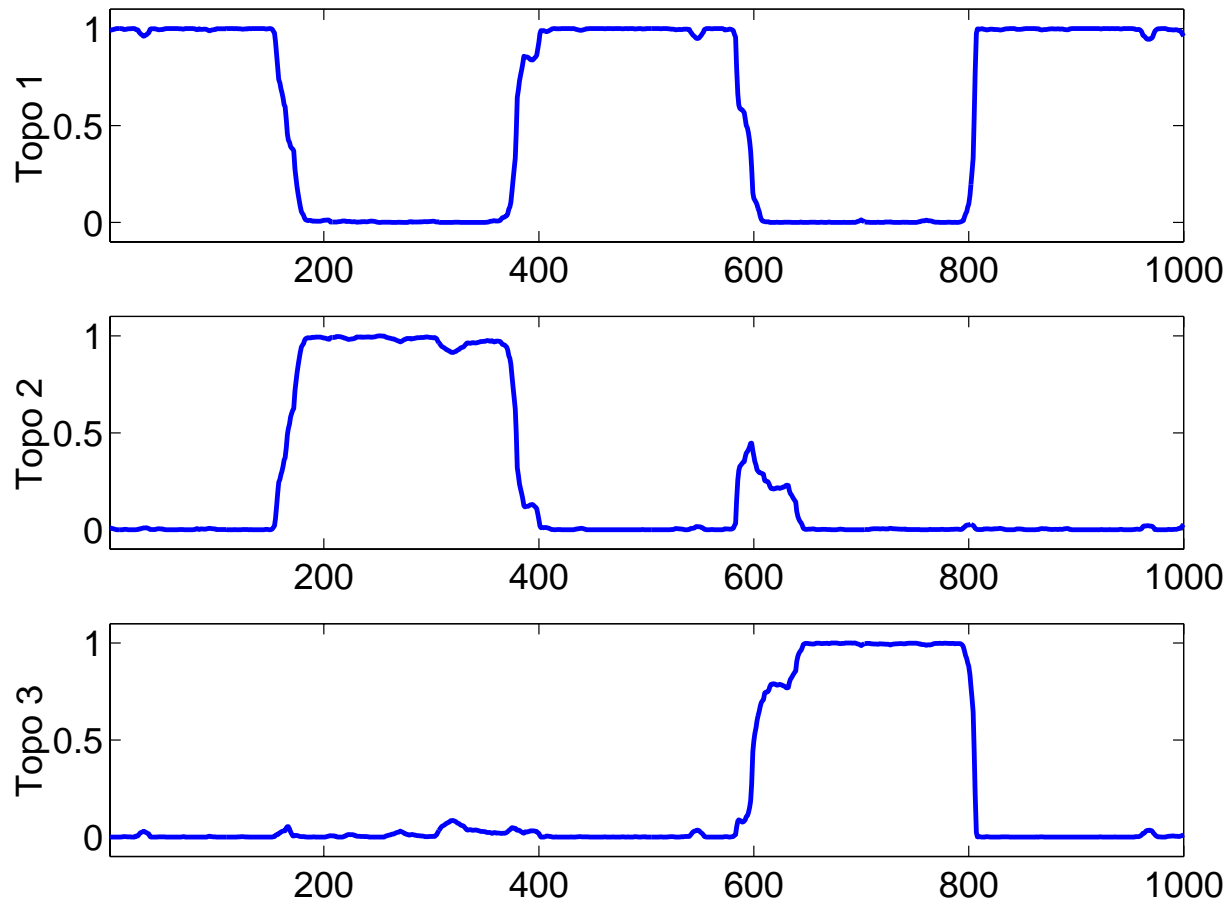
## Prediction with RECPARS (Hein, 1993)

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



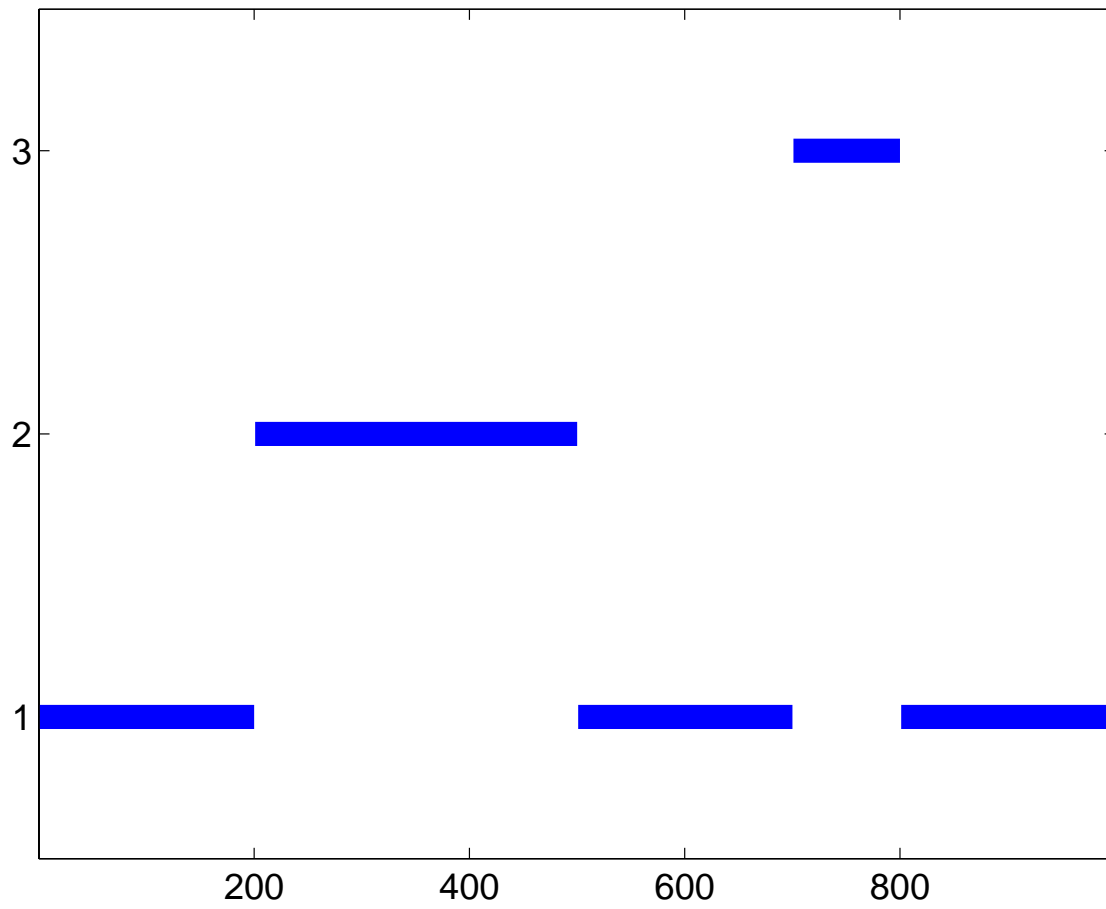
## Prediction with HMM-Bytes

---



## True mosaic structure

---

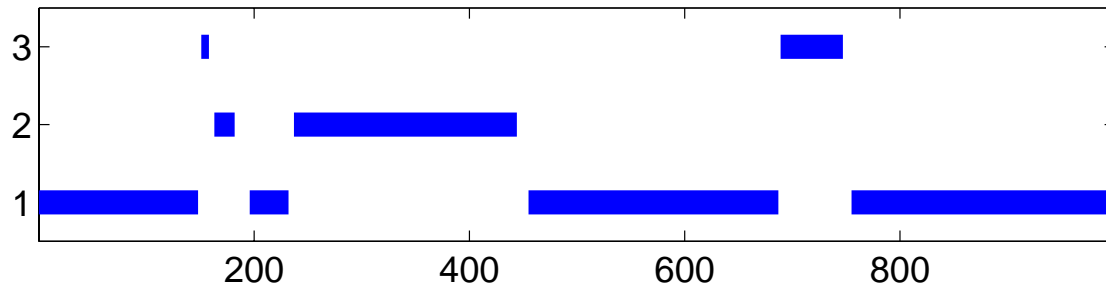
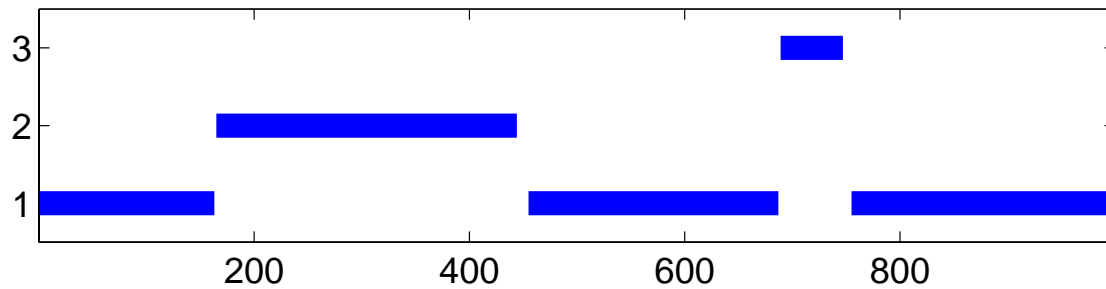
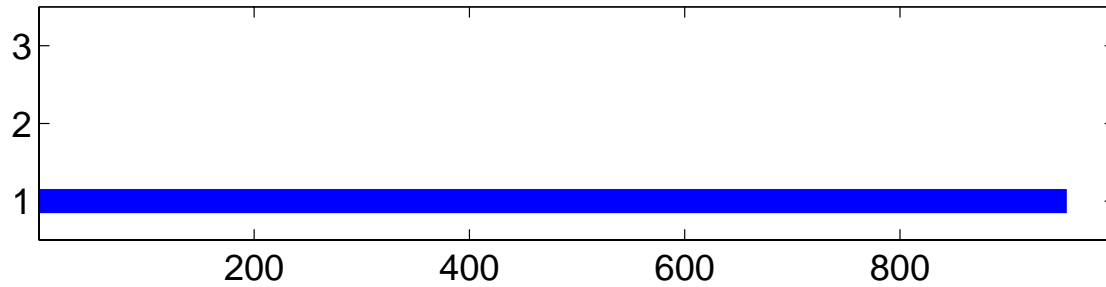


## Prediction with RECPARS

Top:  $C_{recomb}/C_{mut} = 10.0$

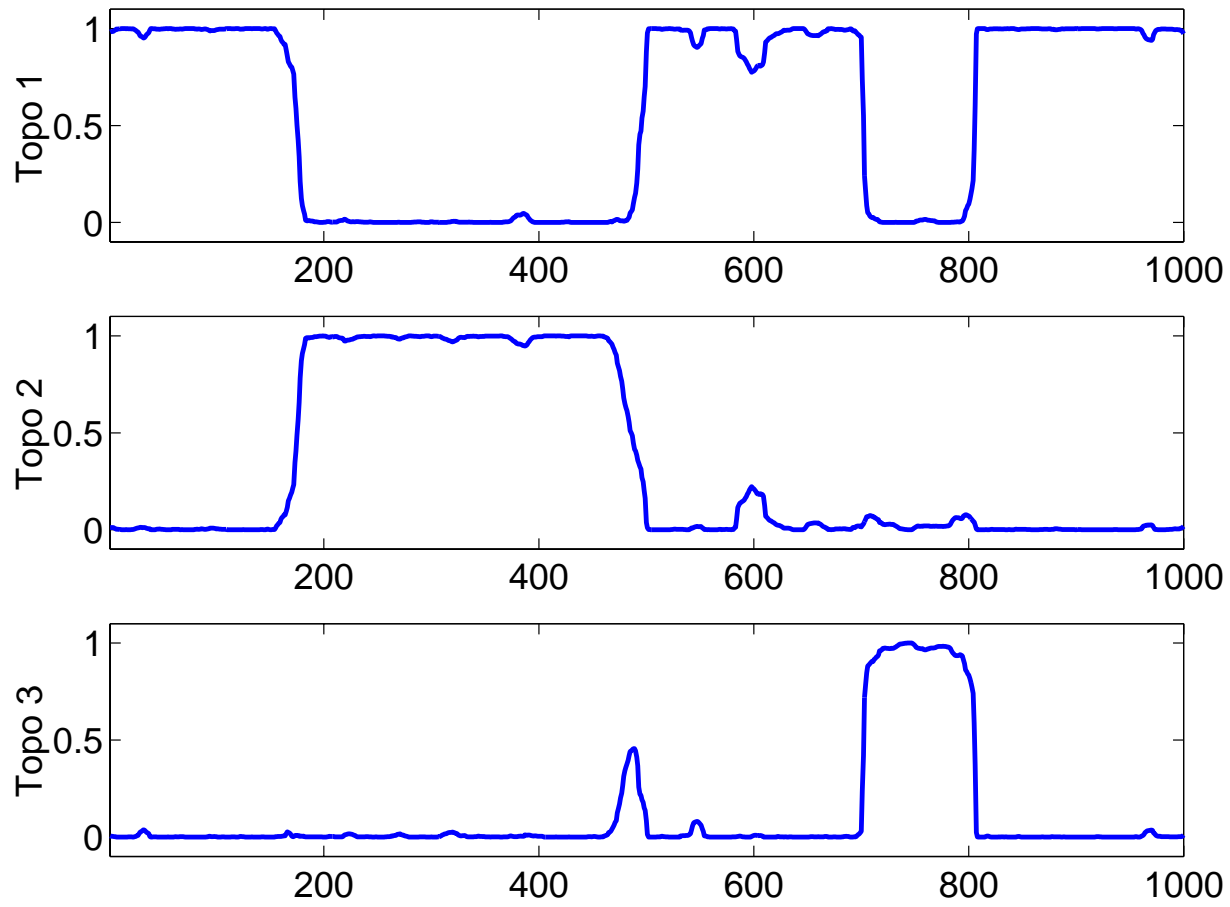
Middle:  $C_{recomb}/C_{mut} = 3.0$

Right:  $C_{recomb}/C_{mut} = 1.5$



## Prediction with HMM-Bayes

---



## Comparison between RECPARS and HMM-Bayes

---

- **RECPARS:** Direct classification of each site.
- **HMM-Bayes:** Assign each site to the mode of the posterior distribution.

## Comparison between RECPARS and HMM-Bayes

---

- **RECPARS**: Direct classification of each site.
- **HMM-Bayes**: Assign each site to the mode of the posterior distribution.
- **Sensitivity**: Percentage of correctly classified recombinant sites.
- **Specificity**: Percentage of correctly classified non-recombinant sites.

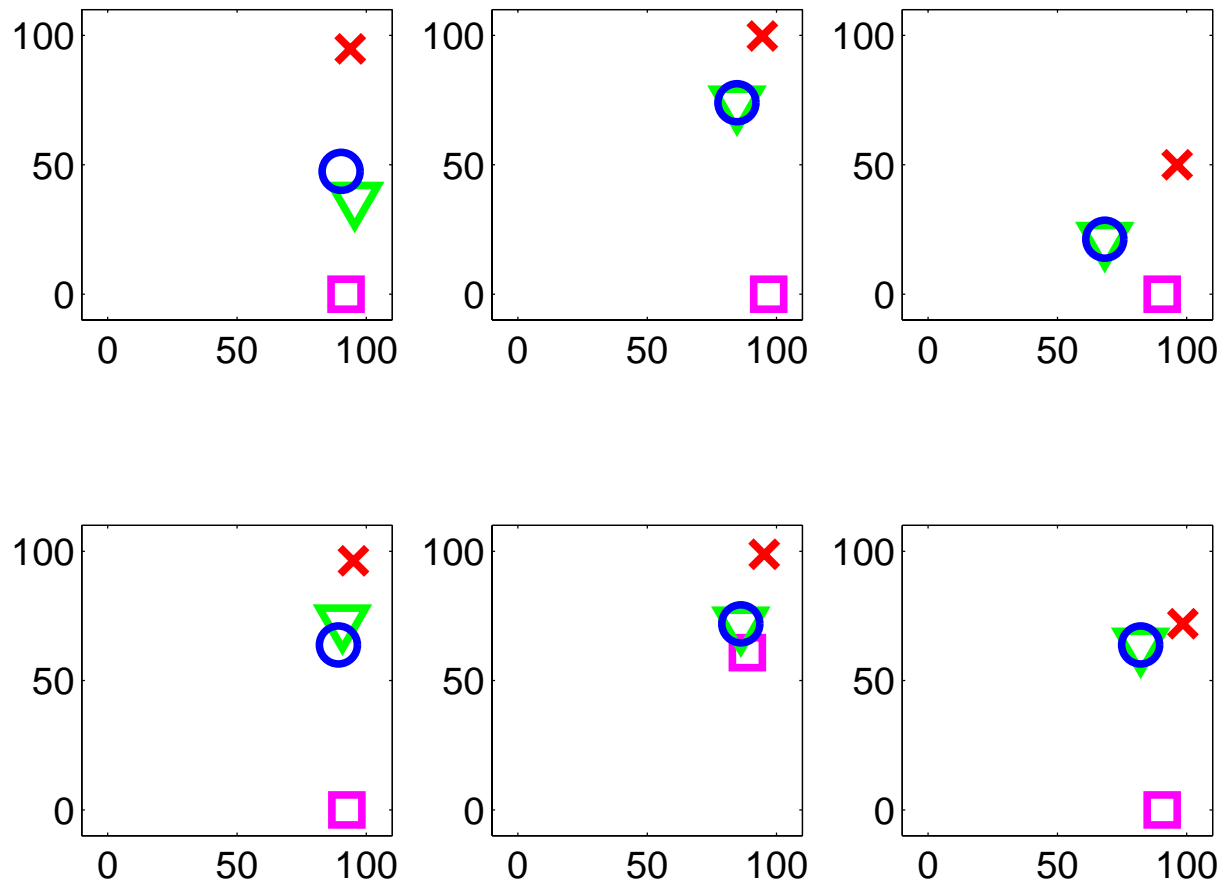
Scatterplot: **sensitivity** (vertical axis) against **specificity** (horizontal axis)

Crosses	HMM-Bayes
Squares	RECPARS, $C_{recomb}/C_{mut} = 10.0$
Circles	RECPARS, $C_{recomb}/C_{mut} = 3.0$
Triangles	RECPARS, $C_{recomb}/C_{mut} = 1.5$

---

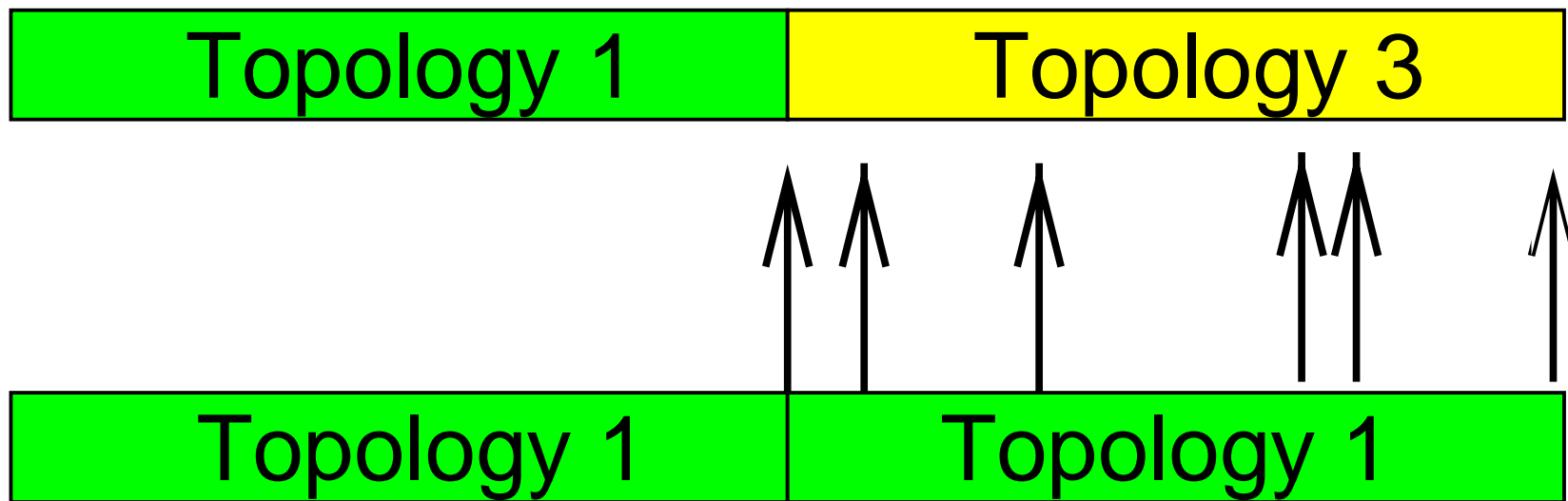
## Comparison between RECPARS and HMM-Bayes

---



## Comparison between HMM-ML and HMM-Bayes

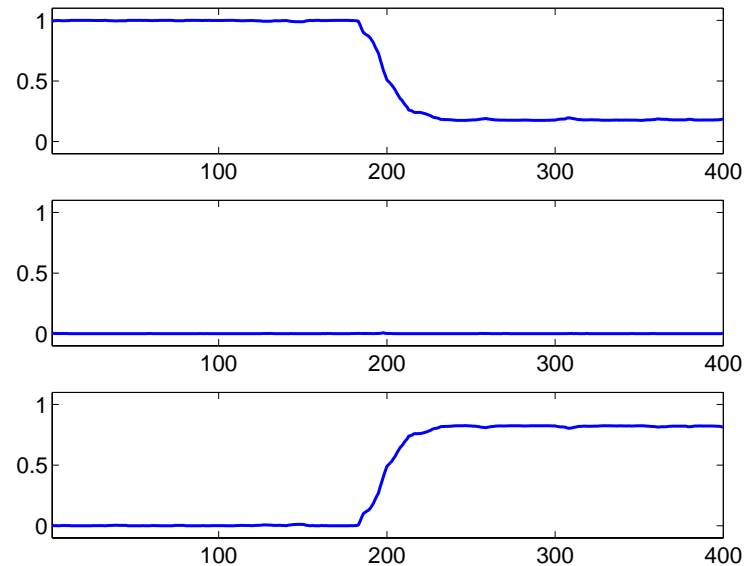
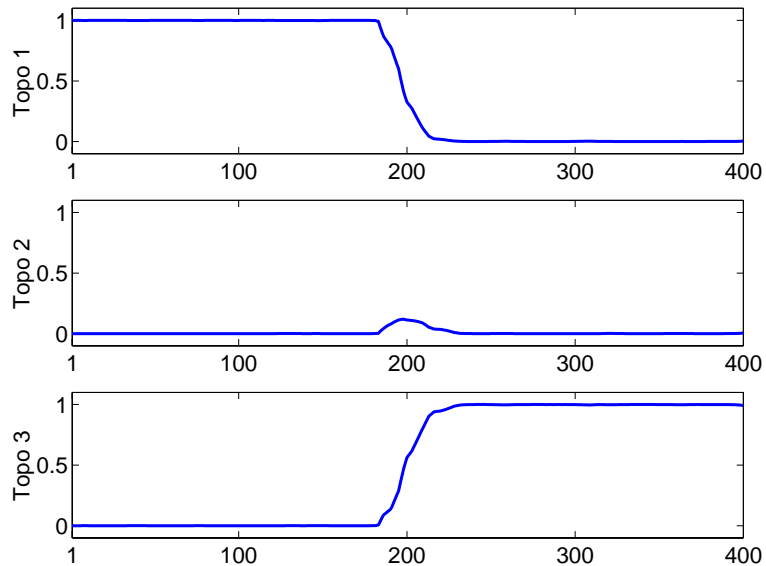
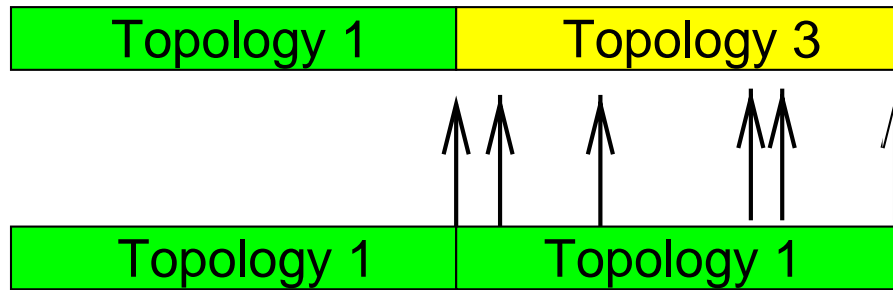
---



20 % of the sites exchanged

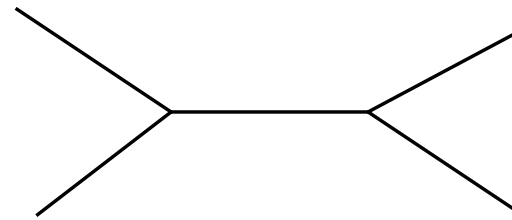
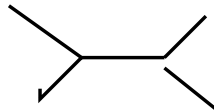
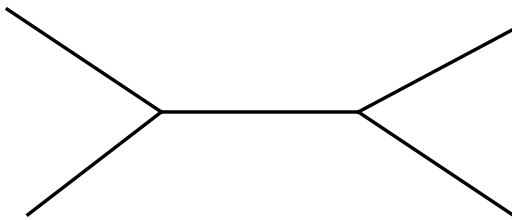
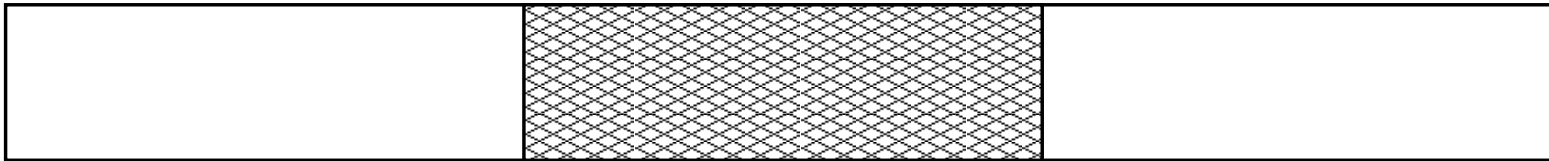
---

# HMM-ML (left) versus HMM-Bayes (right)

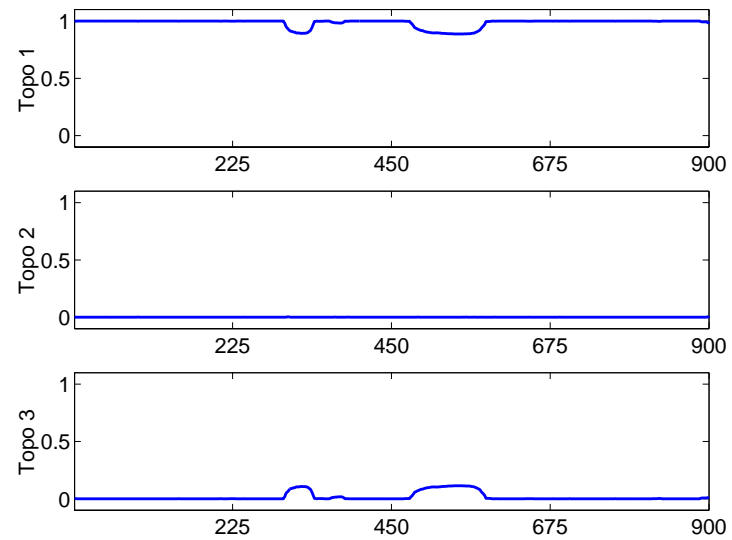
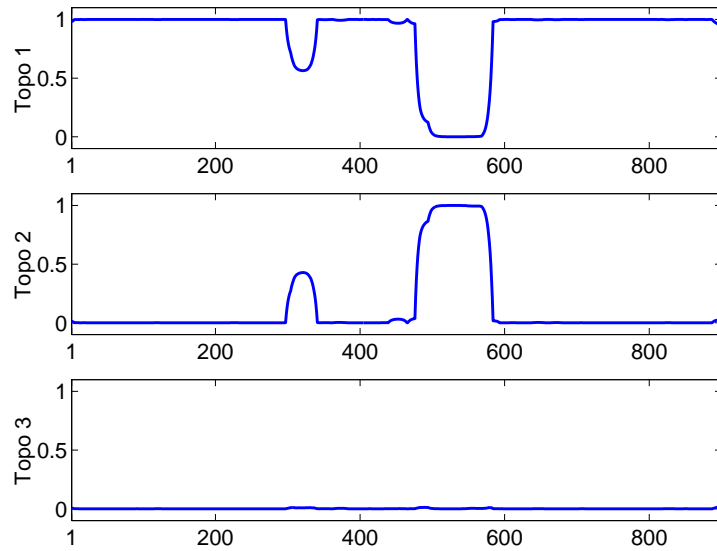
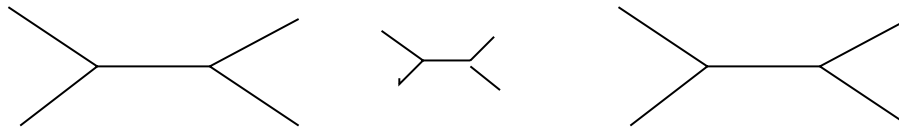
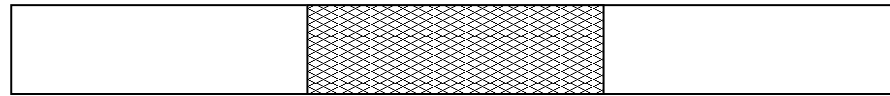


## Comparison between RECPARS and HMM-Bayes

---



# HMM-ML (left) versus HMM-Bayes (right)

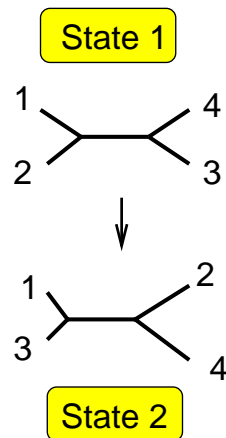
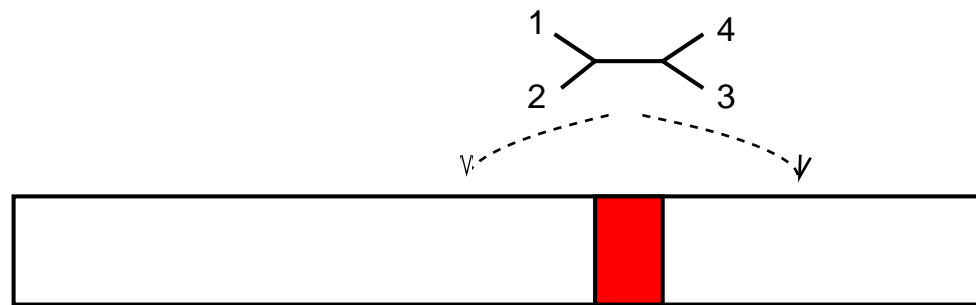


## Hepatitis B Virus (Bollyky et al. 1995)

---

DNA alignment, 3049 nucleotides

- 1) HPBADW1      2) HPBADW2      3) HPBADWZCG      4) HPBADRC



# TOPAL, window size = 100

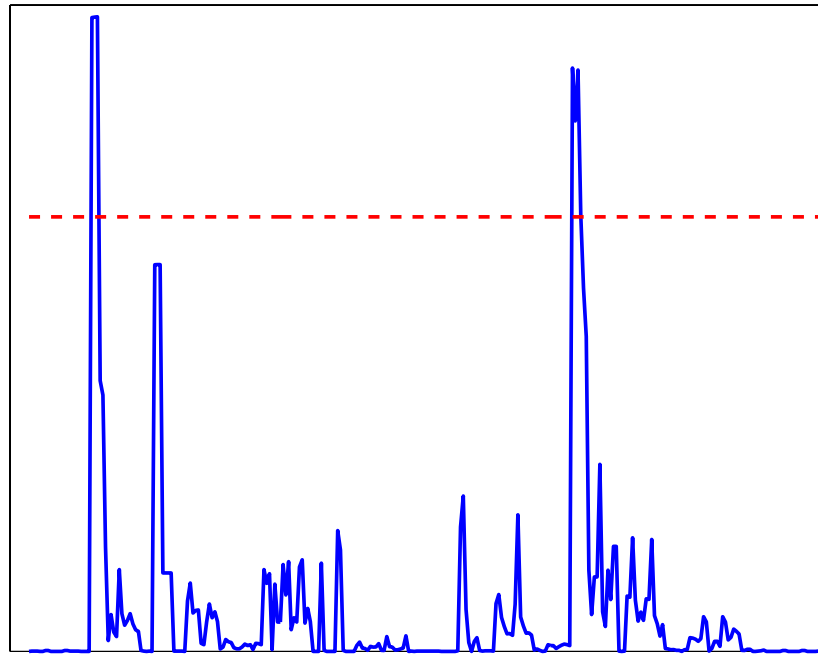
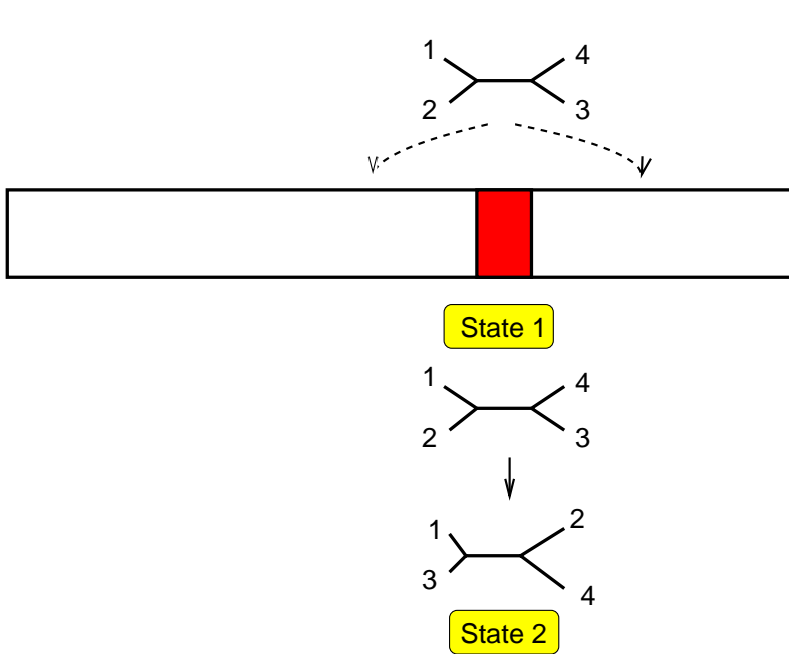
DNA alignment, 3049 nucleotides

1) HPBADW1

2) HPBADW2

3) HPBADWZCG

4) HPBADRC



# TOPAL, window size = 200

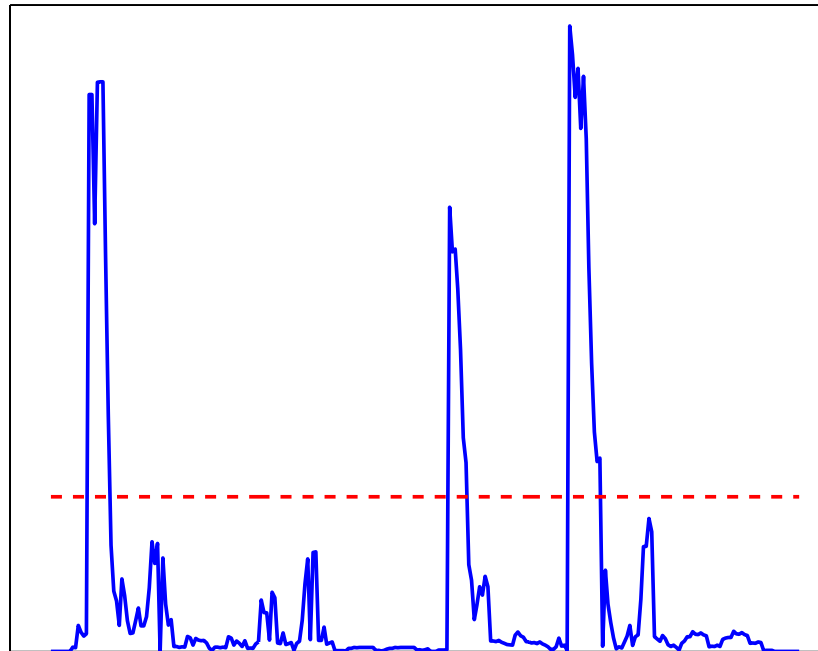
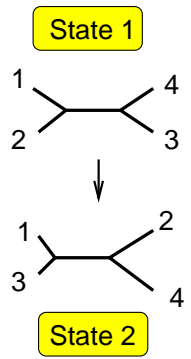
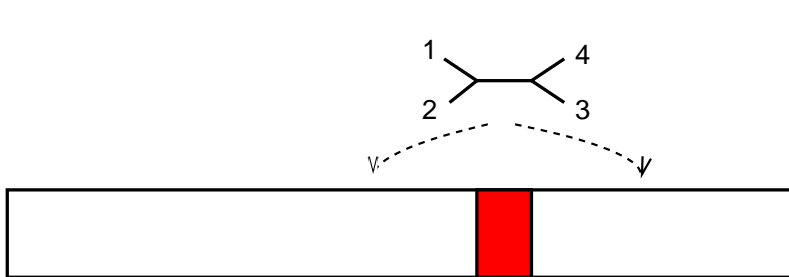
DNA alignment, 3049 nucleotides

1) HPBADW1

2) HPBADW2

3) HPBADWZCG

4) HPBADRC



# RECPARS

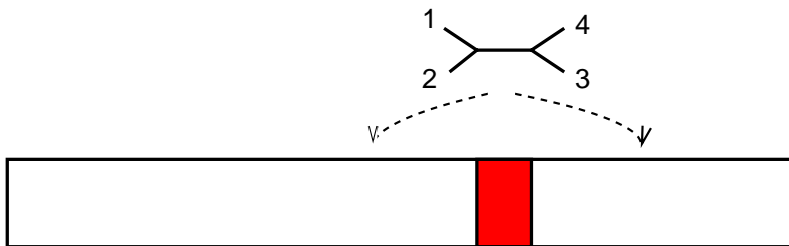
DNA alignment, 3049 nucleotides

1) HPBADW1

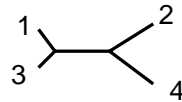
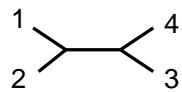
2) HPBADW2

3) HPBADWZCG

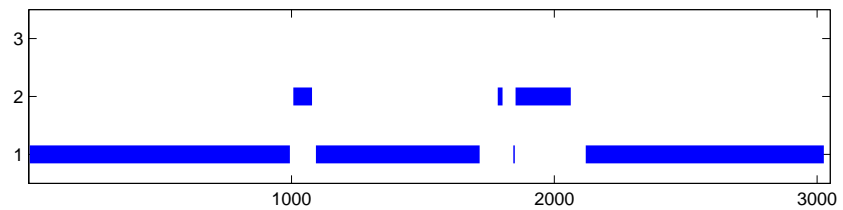
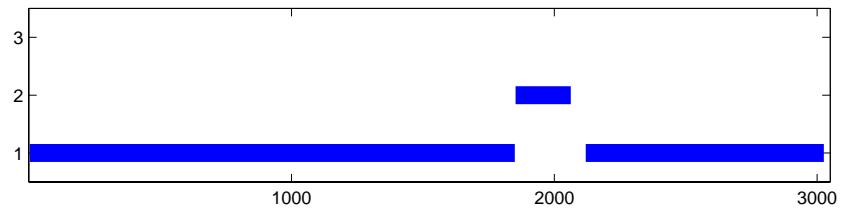
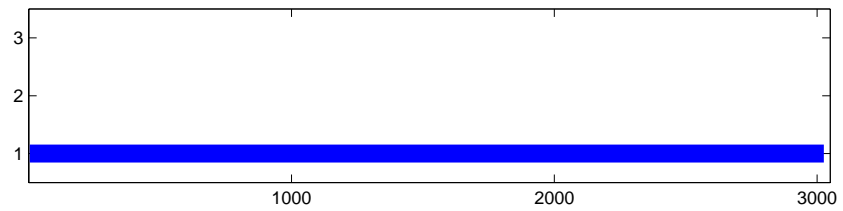
4) HPBADRC



State 1



State 2



# HMM-Bayes

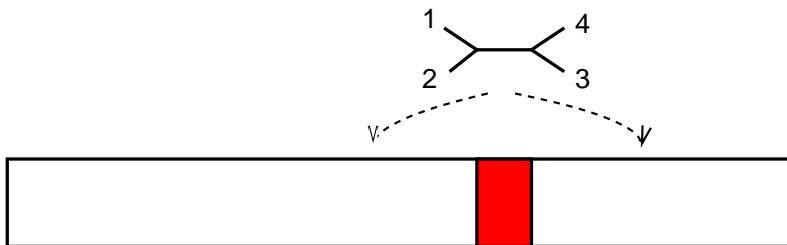
DNA alignment, 3049 nucleotides

1) HPBADW1

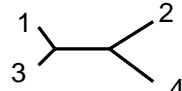
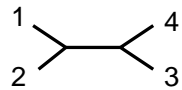
2) HPBADW2

3) HPBADWZCG

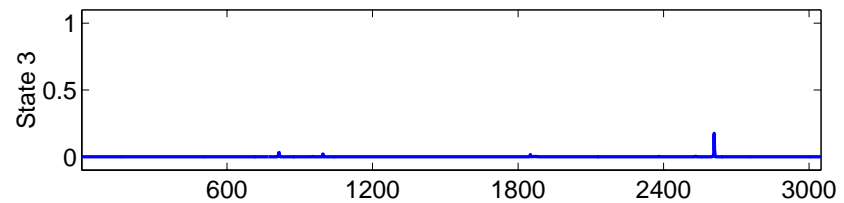
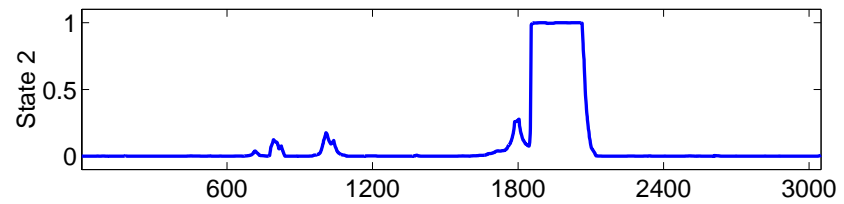
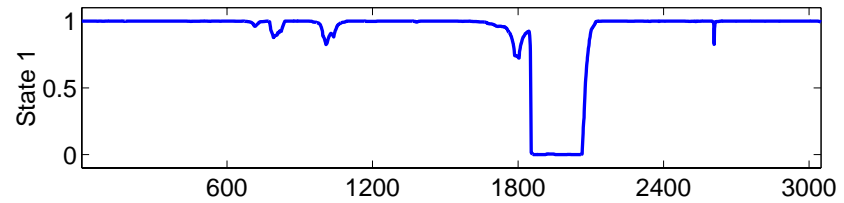
4) HPBADRC



State 1



State 2



## Neisseria (Zhou & Spratt, 1992)

---

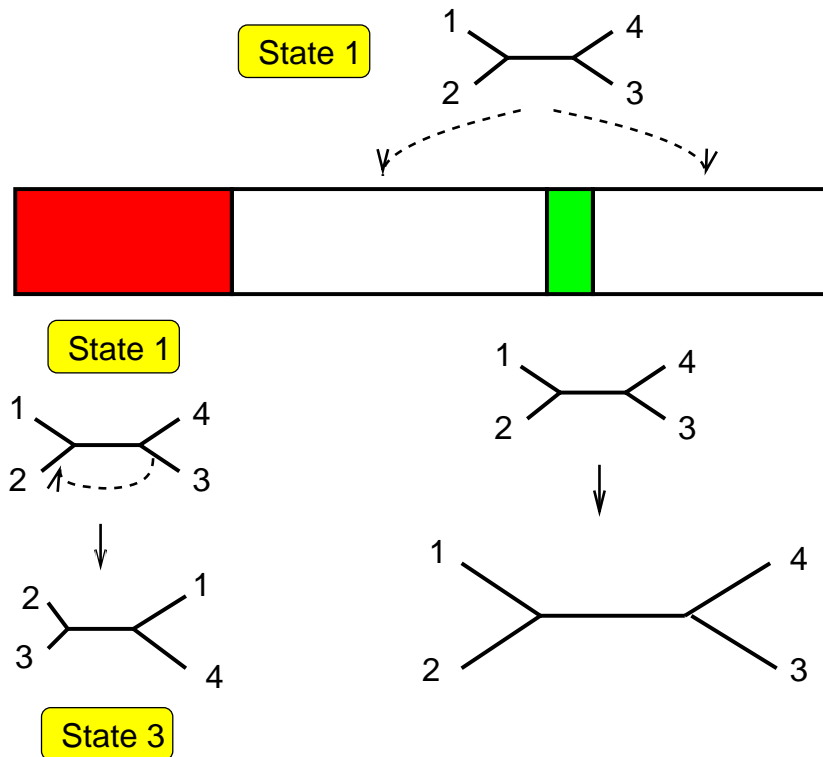
DNA alignment, 787 nucleotides (argF gene)

- |                                  |                             |
|----------------------------------|-----------------------------|
| 1) Neisseria <b>gonorrhoeae</b>  | 3) Neisseria <b>cinerea</b> |
| 2) Neisseria <b>meningitidis</b> | 4) Neisseria <b>mucosa</b>  |

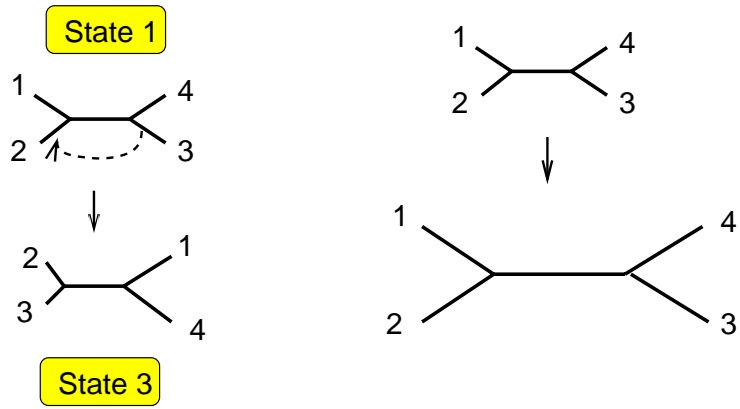
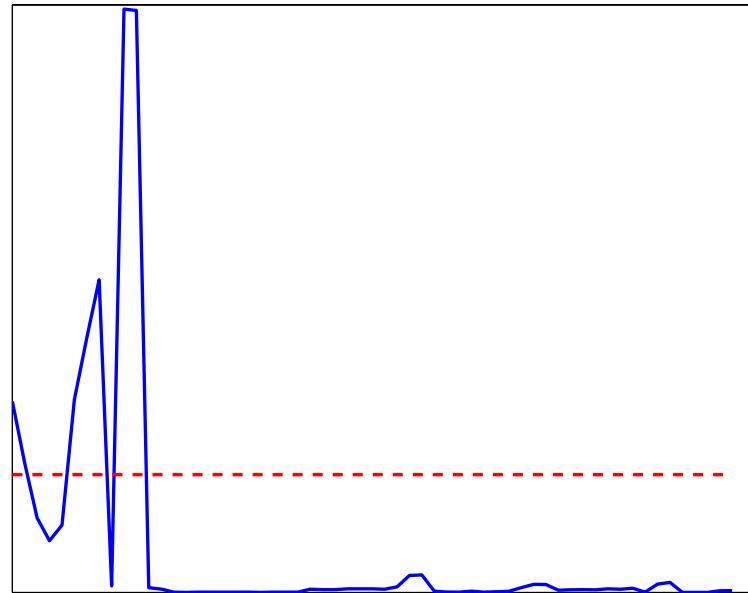
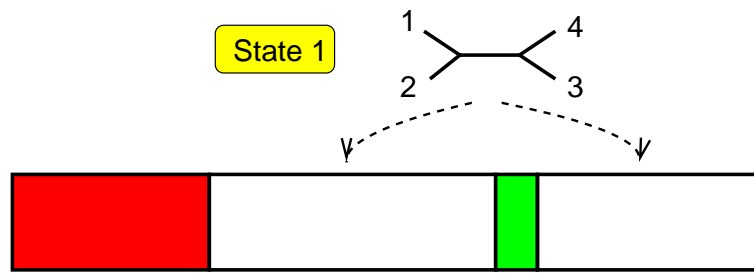
## Neisseria (Zhou & Spratt, 1992)

DNA alignment, 787 nucleotides (argF gene)

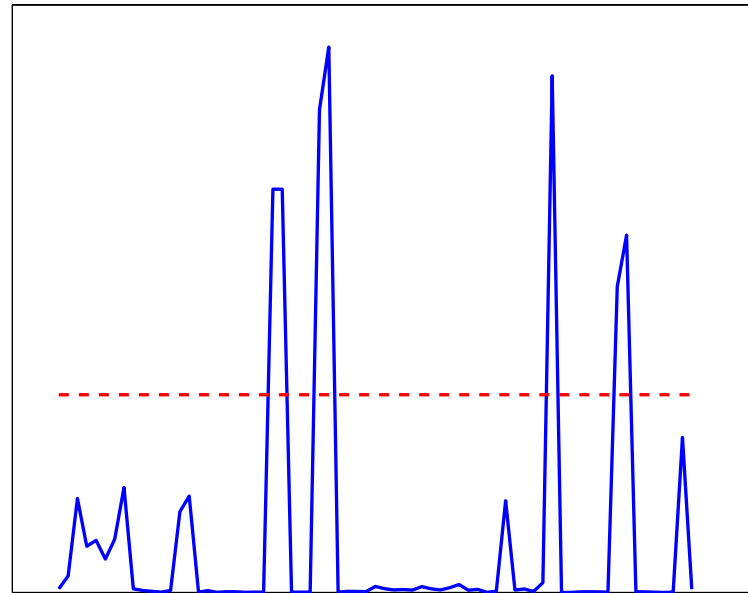
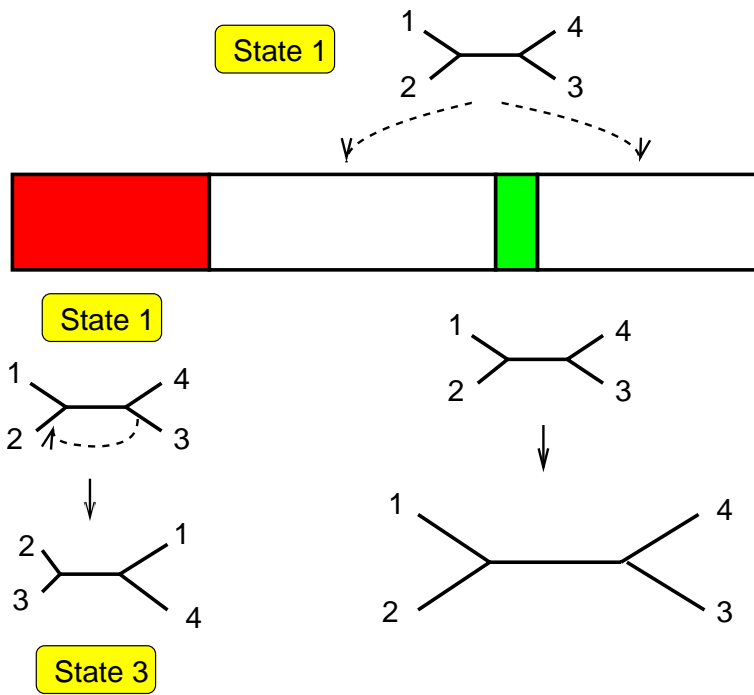
- 1) Neisseria **gonorrhoeae**
- 2) Neisseria **meningitidis**
- 3) Neisseria **cinerea**
- 4) Neisseria **mucosa**



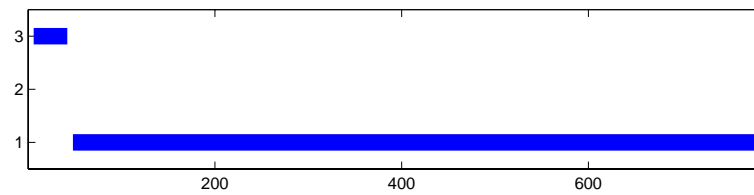
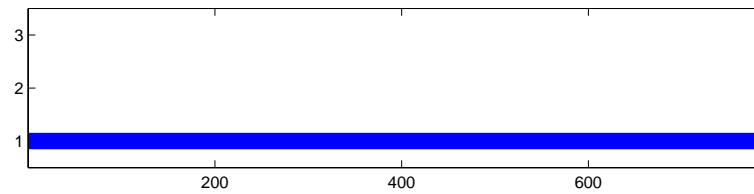
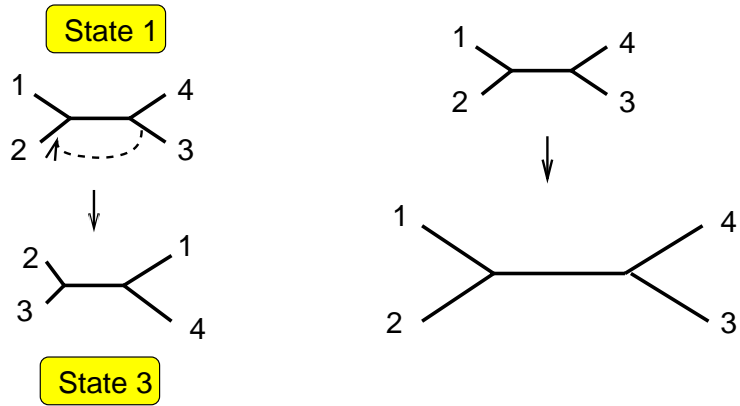
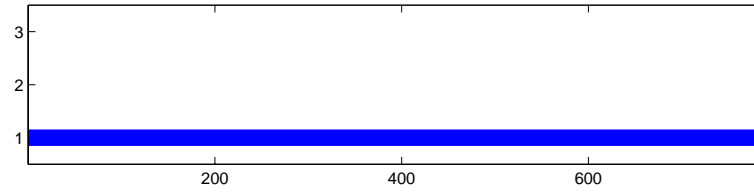
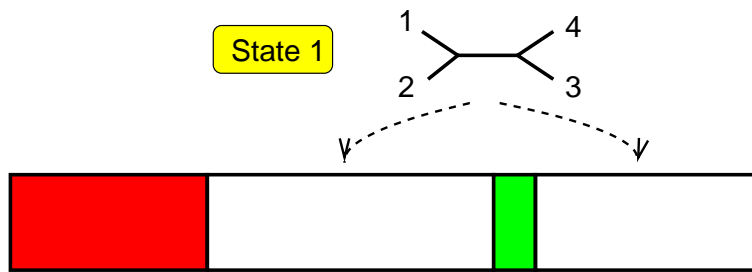
# Topal, window size 200



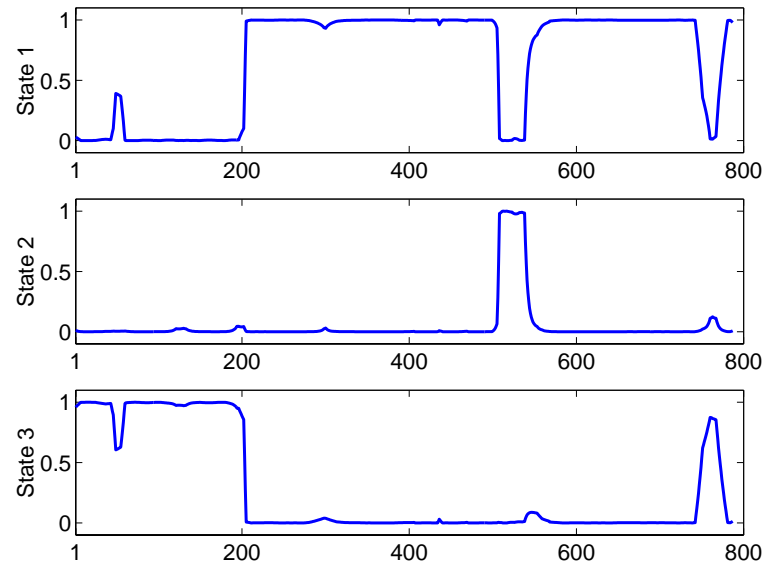
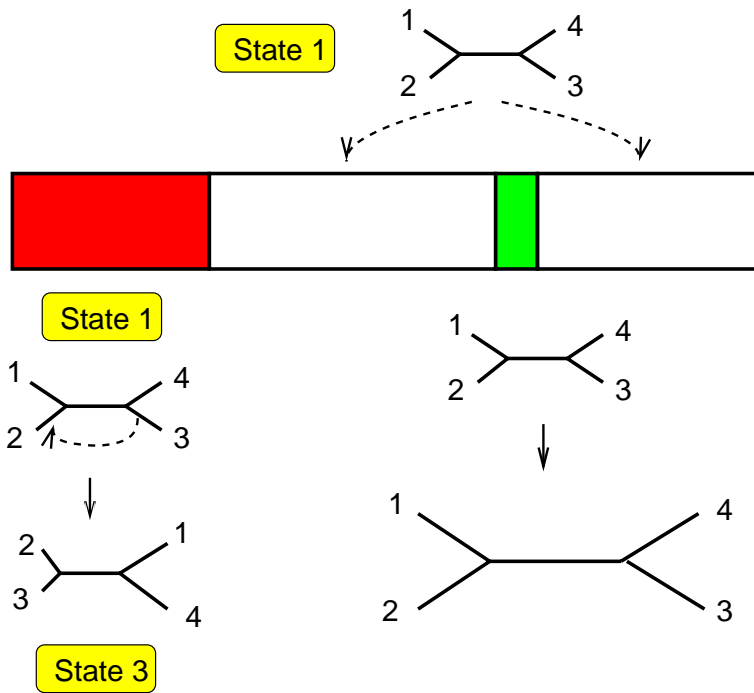
# Topal, window size 100



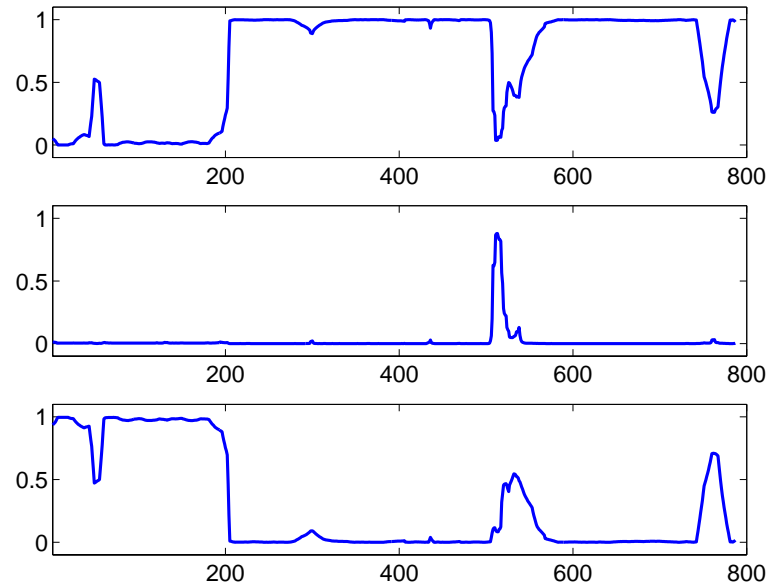
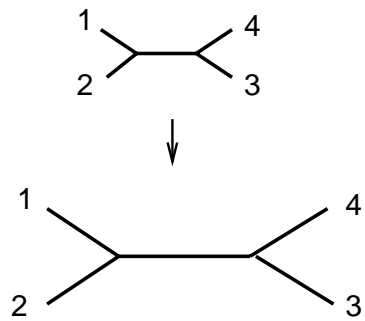
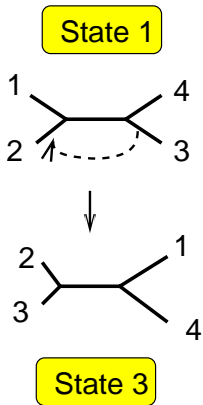
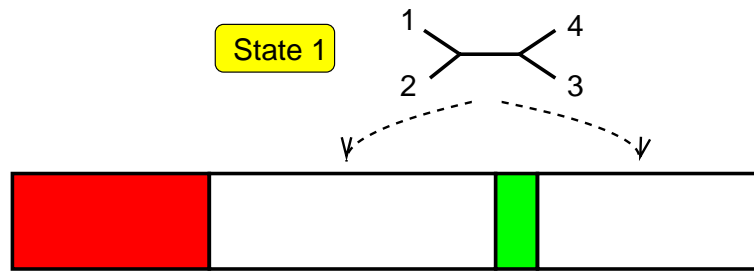
# RECPARS



# Prediction of $P(S_t|\mathcal{D})$ with ML



## Prediction of $P(S_t|\mathcal{D})$ with Bayes



## Discussion and future work

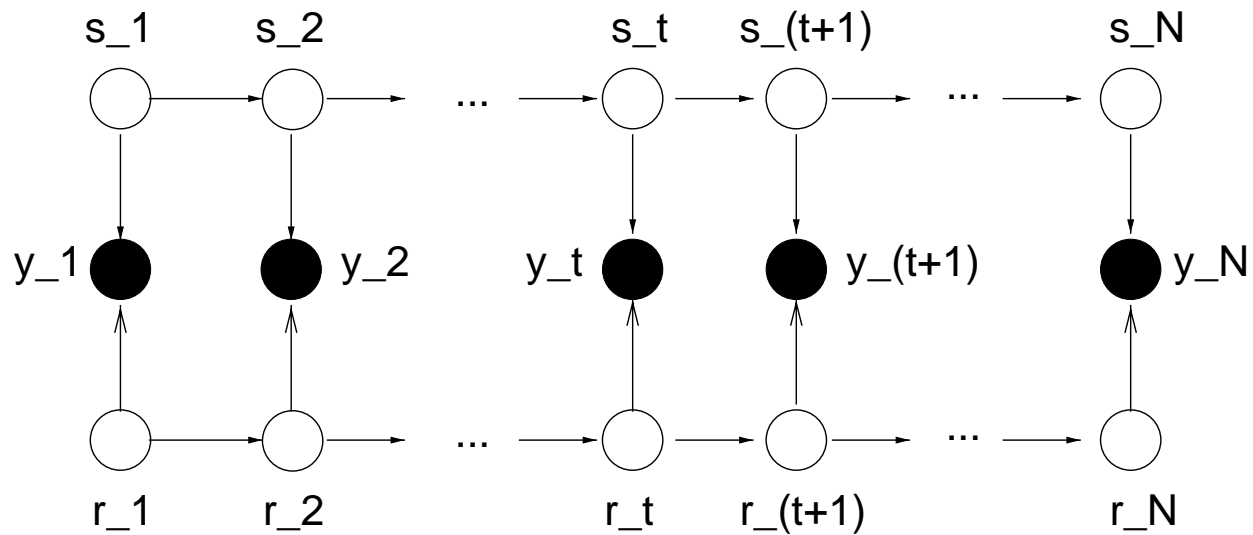
---

- Problem: rate heterogeneity

## Discussion and future work

---

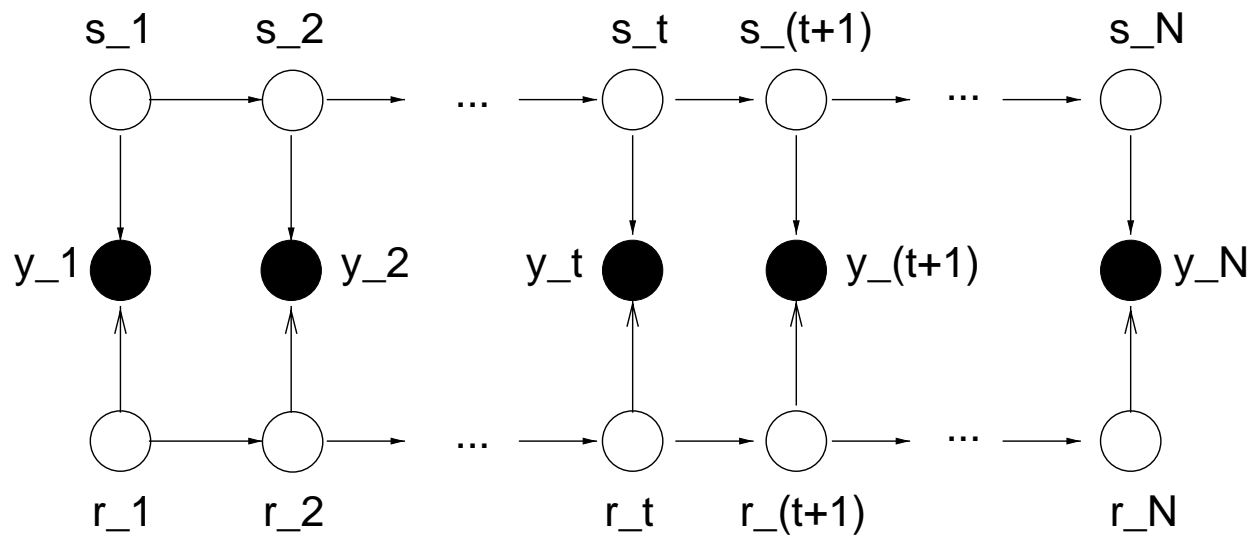
- Problem: rate heterogeneity
- Future work: factorial HMM



## Discussion and future work

---

- Problem: rate heterogeneity
- Future work: factorial HMM



- Limited in the number of different tree topologies.
-

## Acknowledgements

---

### Collaboration

Frank Wright  
Gráinne McGuire

### Funding

BBSRC  
SEERAD

---