
Detecting Recombination in DNA Sequence Alignments

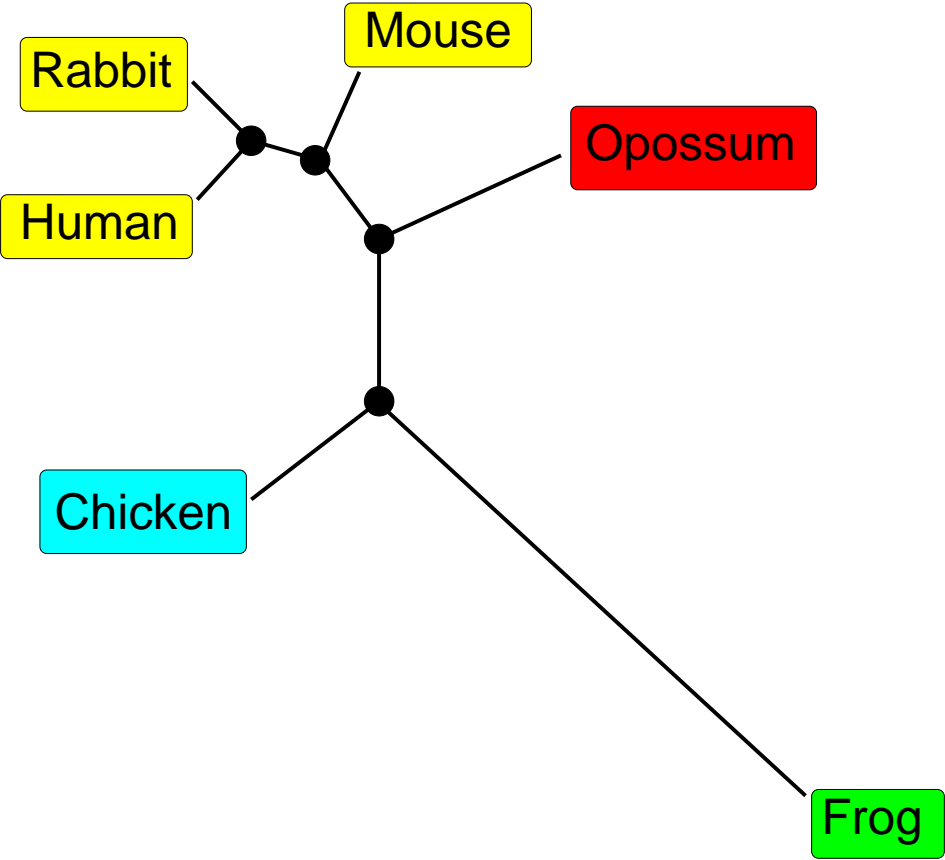
Dirk Husmeier

Biomathematics and Statistics Scotland
SCRI, Invergowrie, Dundee DD2 5DA, UK

Email: dirk@bioinformatics.ac.uk

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

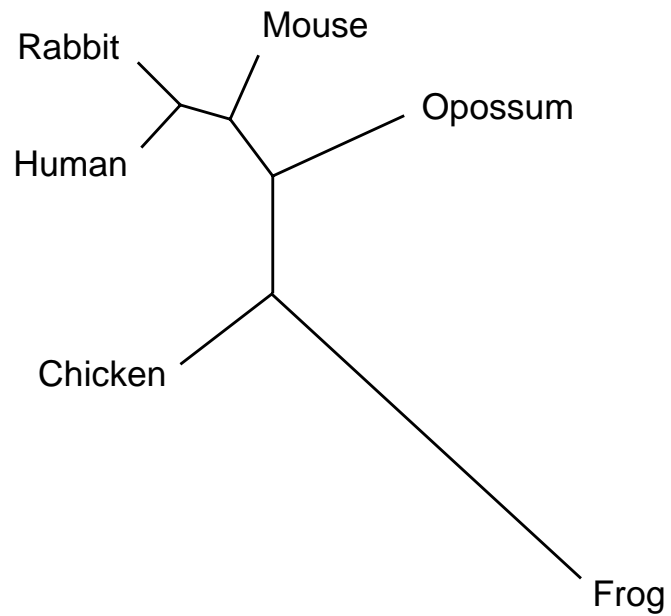
Phylogenetics



--> Topology
--> Branch lengths

Phylogenetics

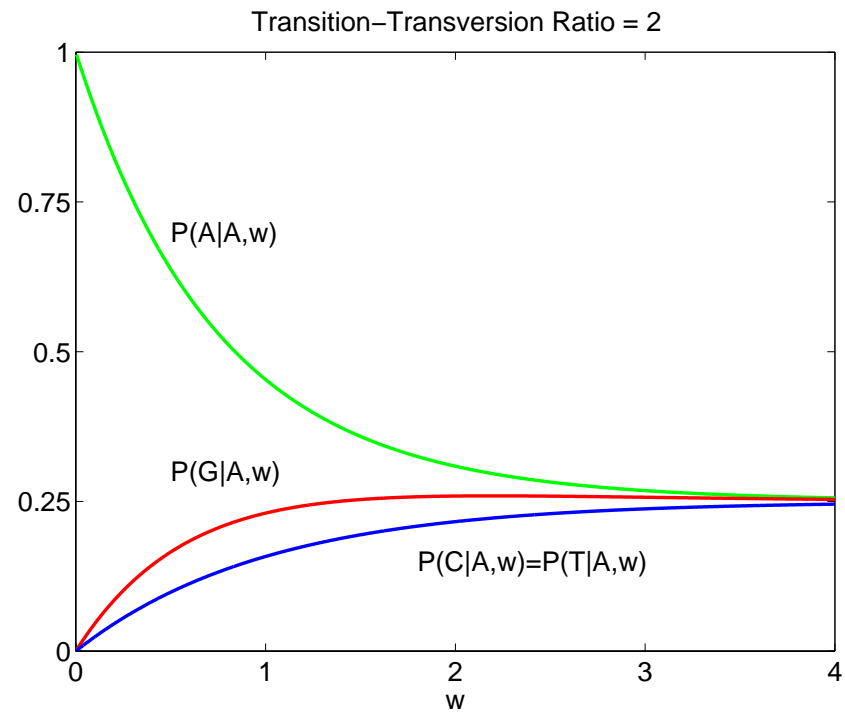
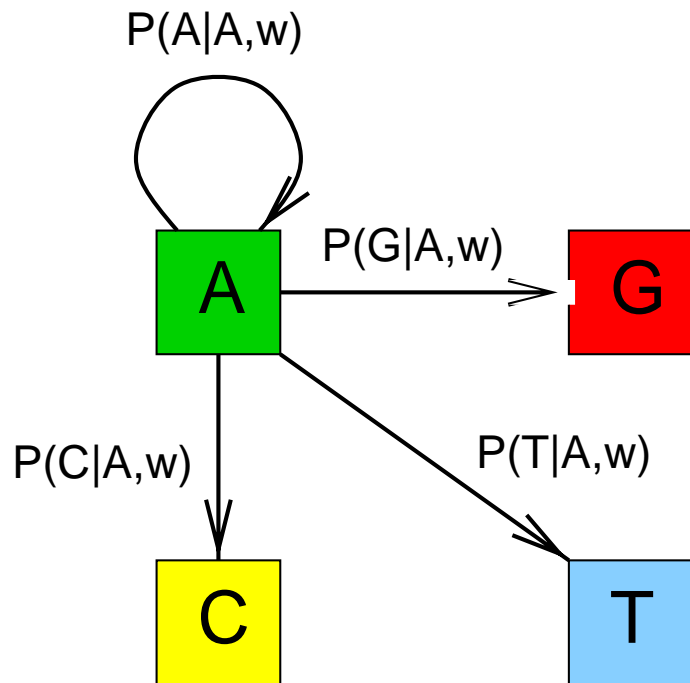
Frog	G	C	T	T	G	A	C	T	T	C	T	G	A	G	G	T	T
Chicken	G	C	G	T	A	A	C	T	T	C	A	C	A	T	G	A	T
Human	G	C	G	T	C	A	C	T	T	G	A	G	A	C	G	C	T
Rabbit	G	C	G	T	C	A	C	T	T	G	A	G	A	C	G	C	T
Mouse	G	C	G	T	C	A	C	T	T	G	A	C	A	G	G	C	T
Opossum	G	C	G	T	C	A	C	T	T	G	A	G	A	C	G	C	T



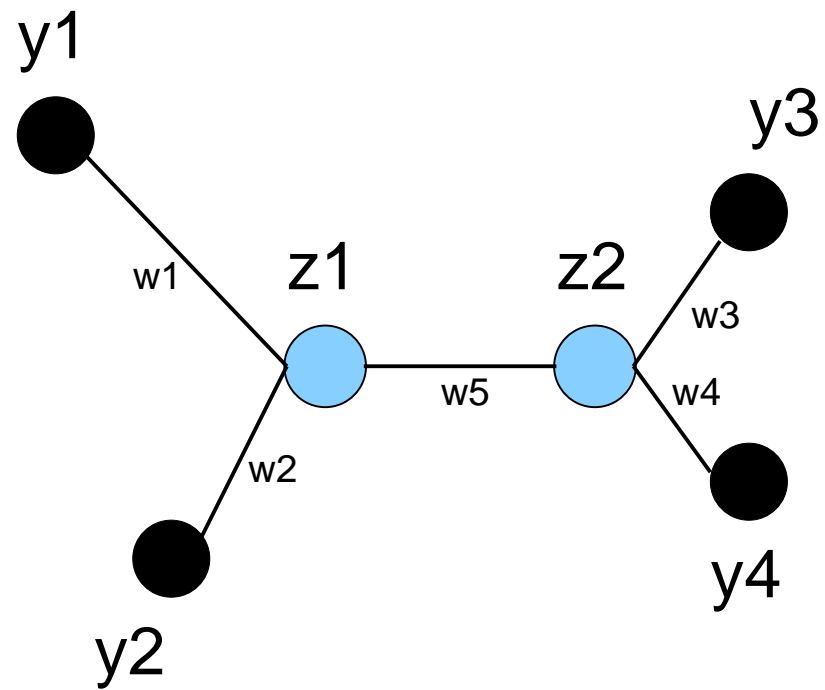
--> Topology

--> Branch lengths

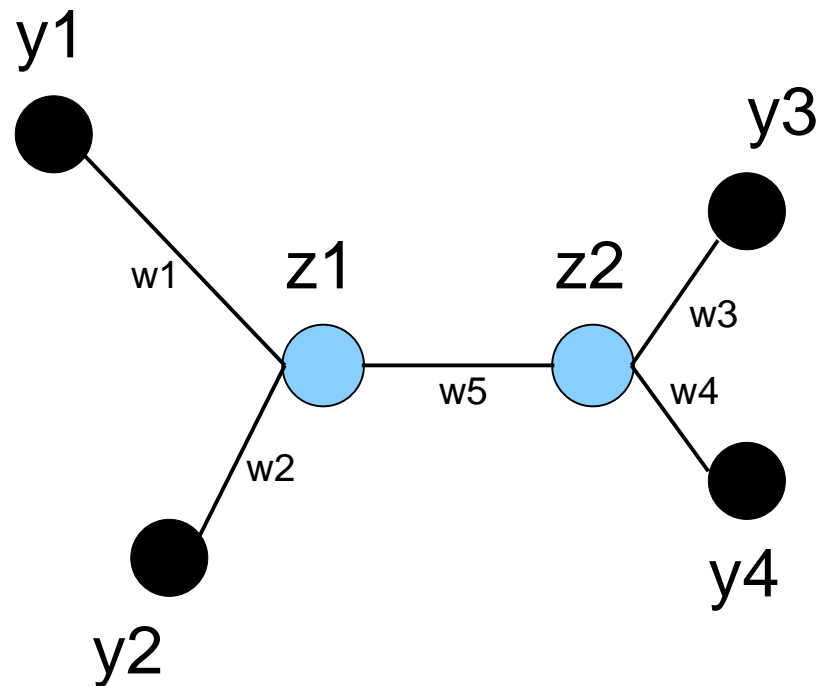
A probabilistic model of evolution



A probabilistic model of evolution



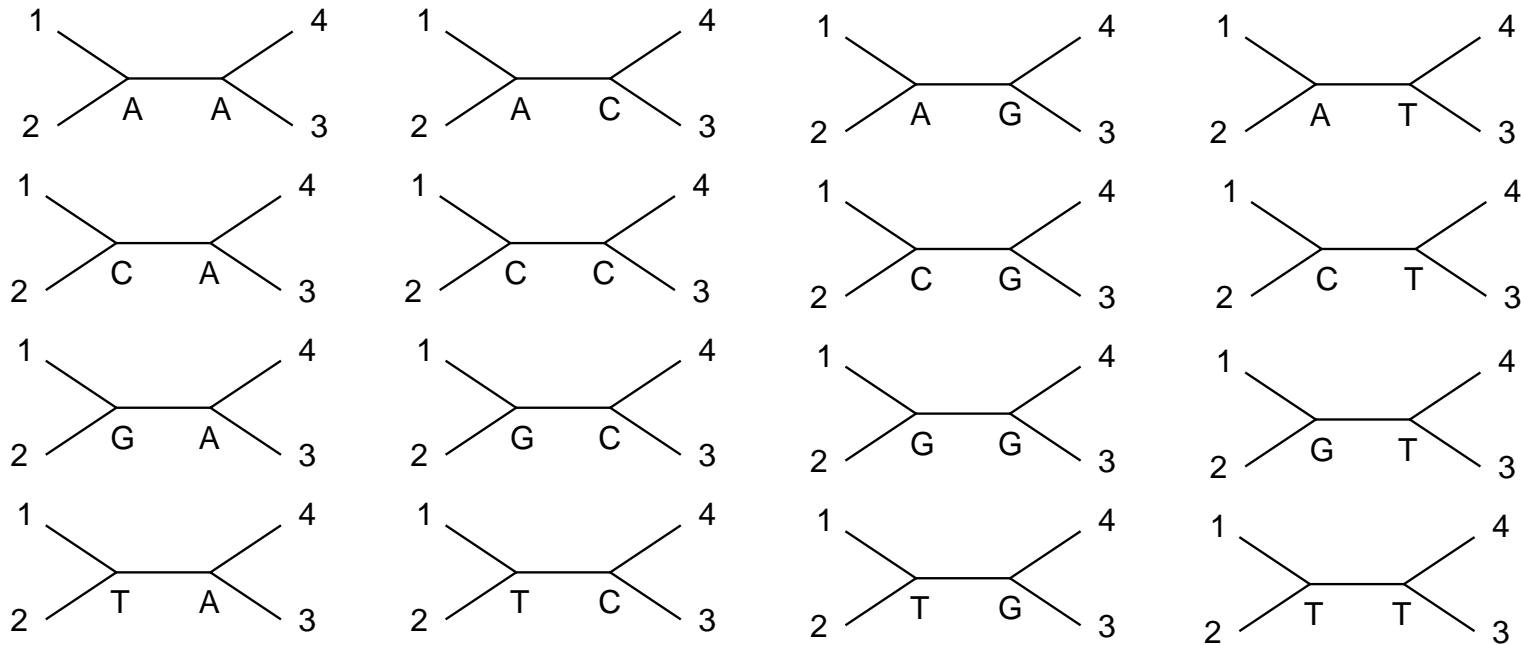
A probabilistic model of evolution



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

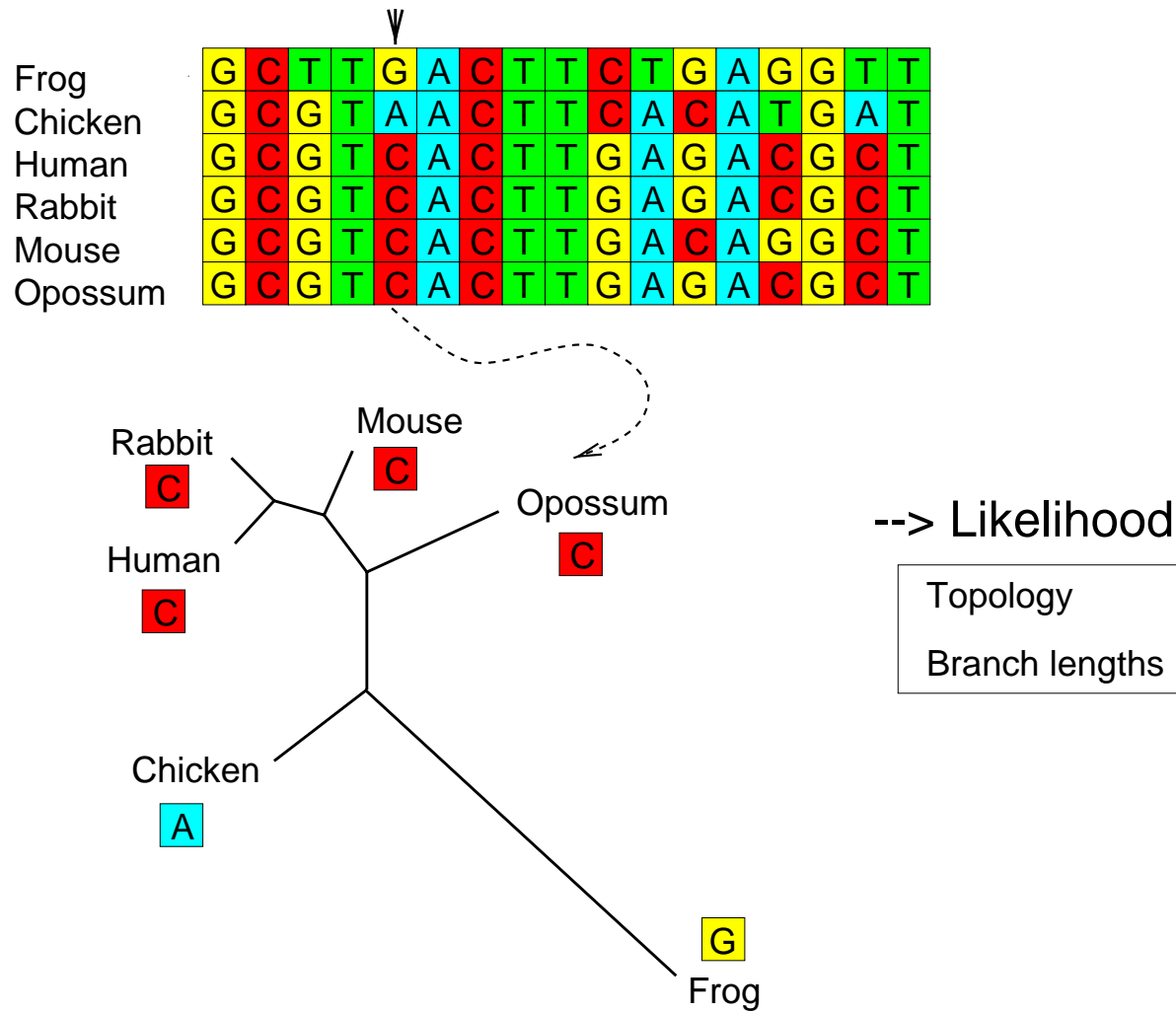
$$= P(y_1 | z_1, w_1) P(y_2 | z_1, w_2) P(z_2 | z_1, w_5) P(y_3 | z_2, w_3) P(y_4 | z_2, w_4)$$

Marginalisation

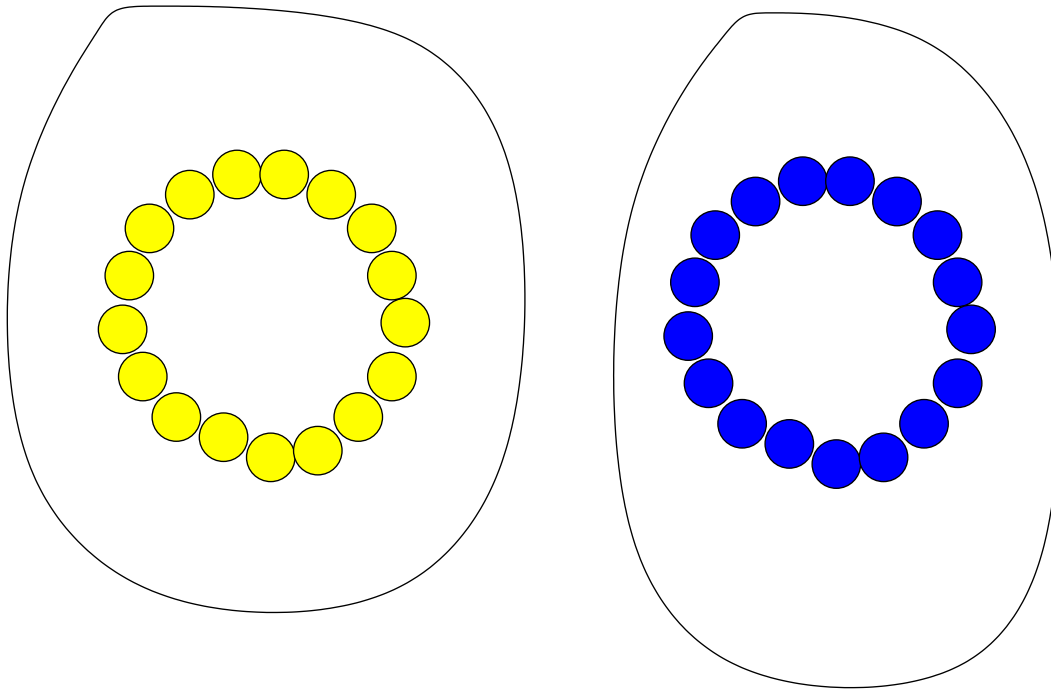


$$P(y_1, y_2, y_3, y_4 | \mathbf{w}) = \sum_{z_1} \sum_{z_2} P(y_1, y_2, y_3, y_4, z_1, z_2 | \mathbf{w})$$

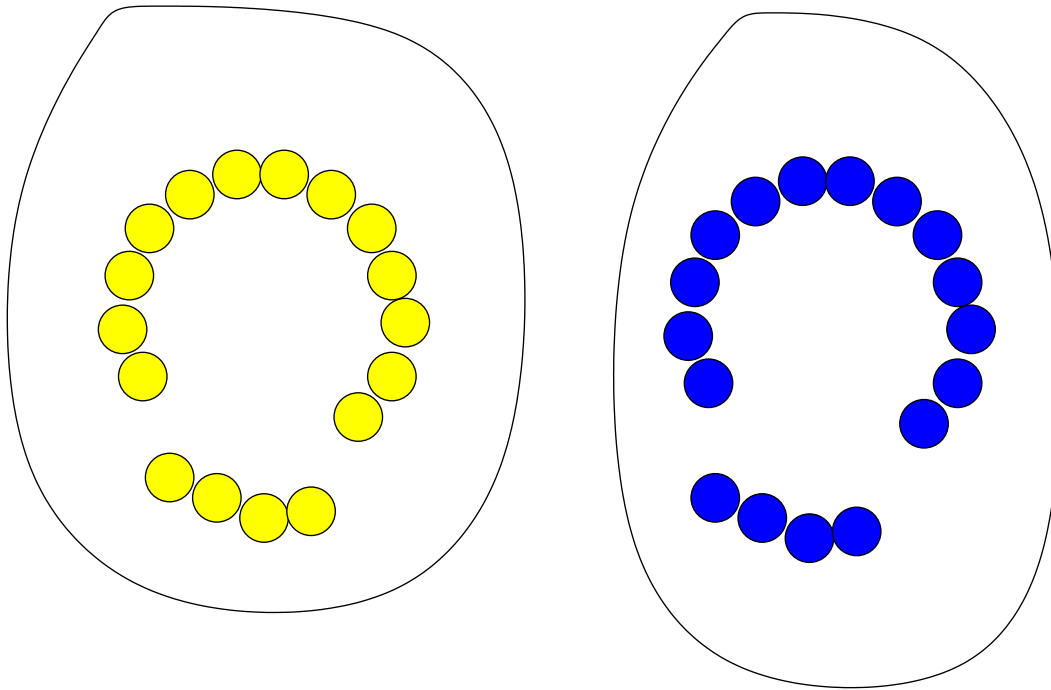
Statistical approach to phylogenetics



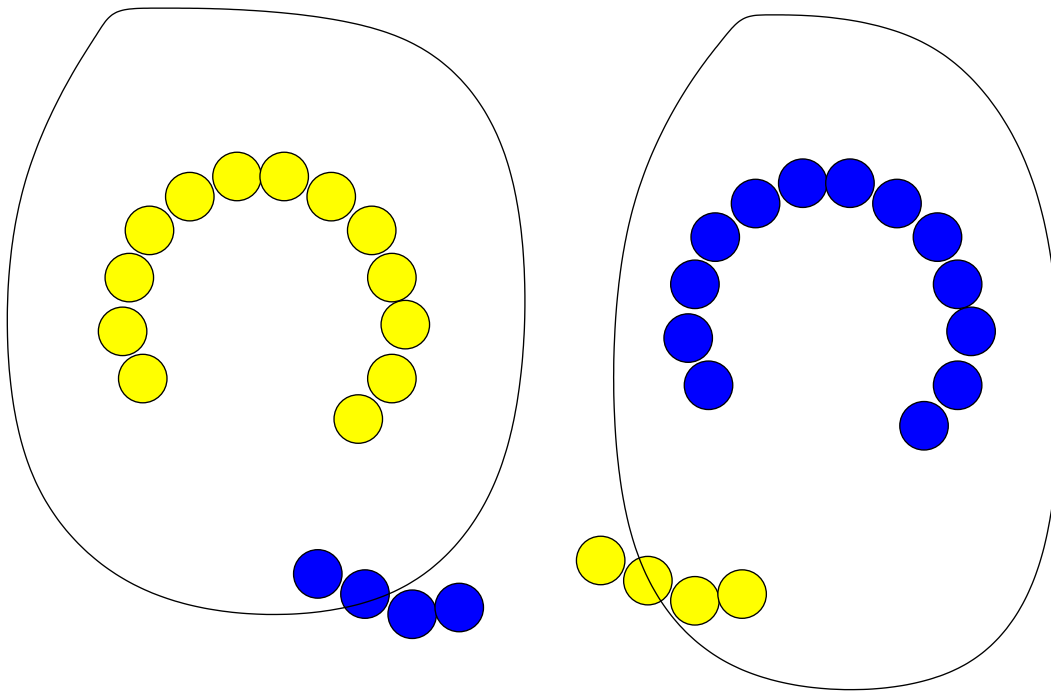
Recombination



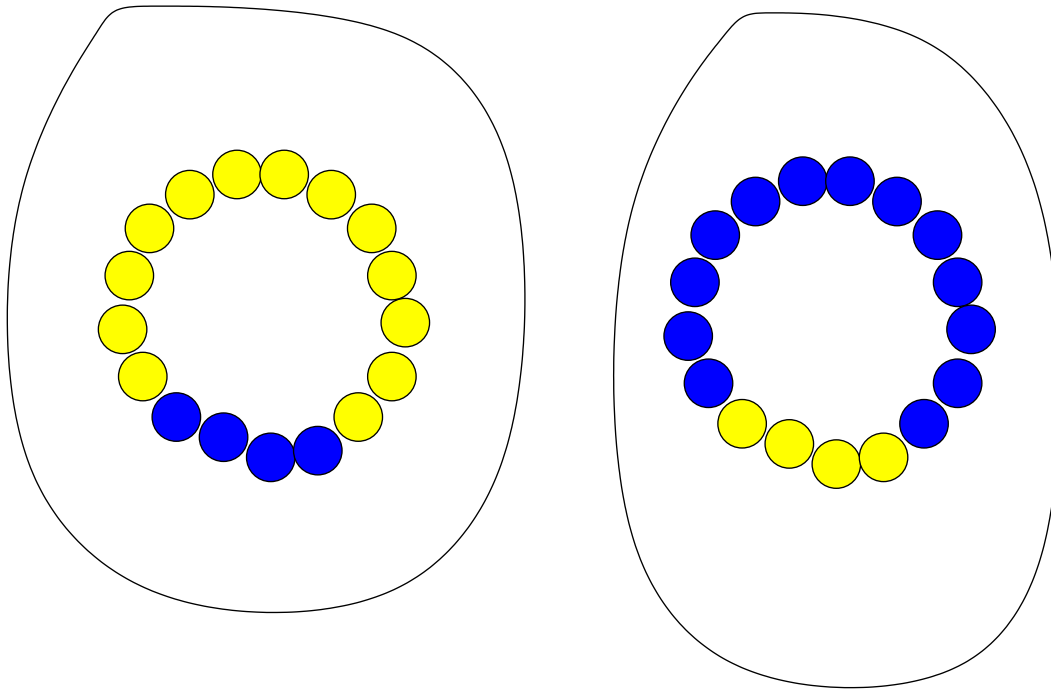
Recombination



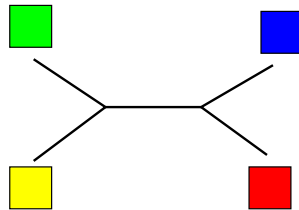
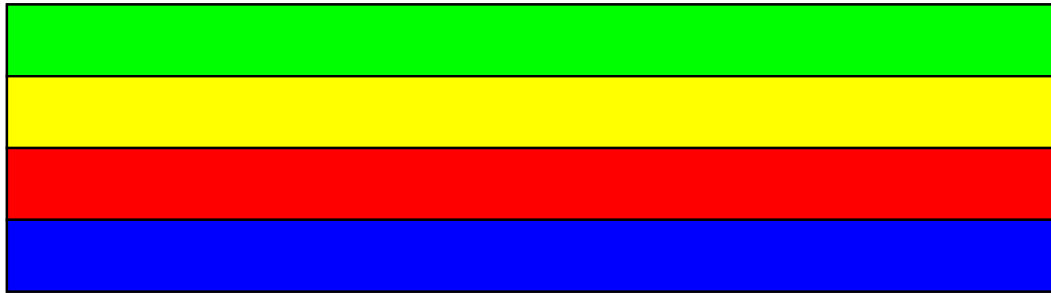
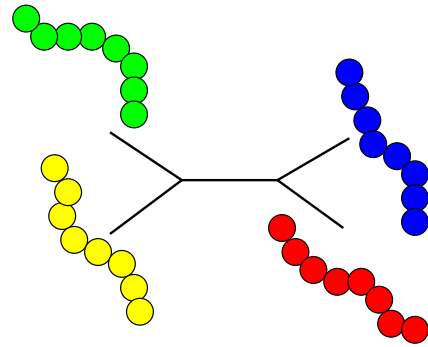
Recombination



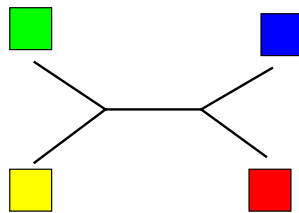
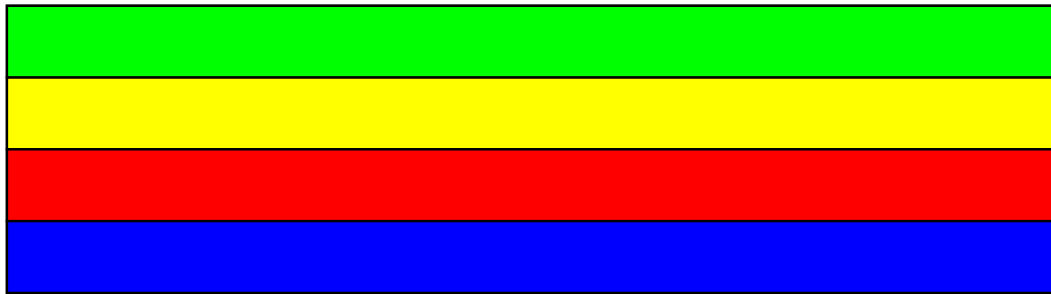
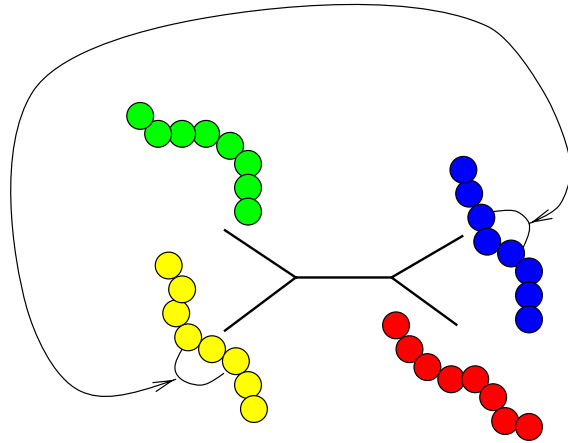
Recombination



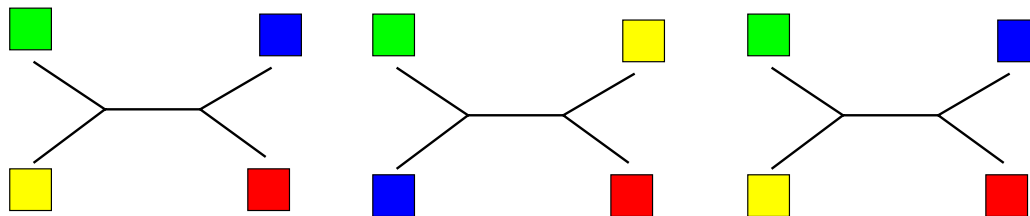
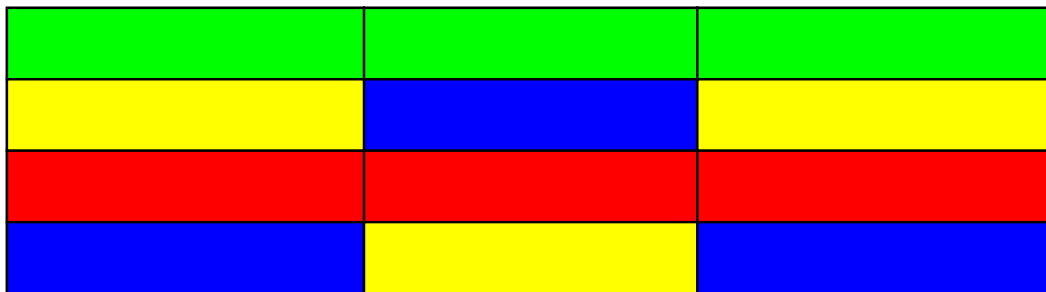
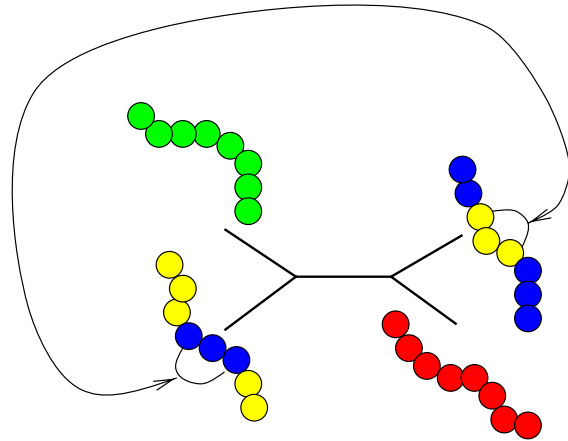
Recombination



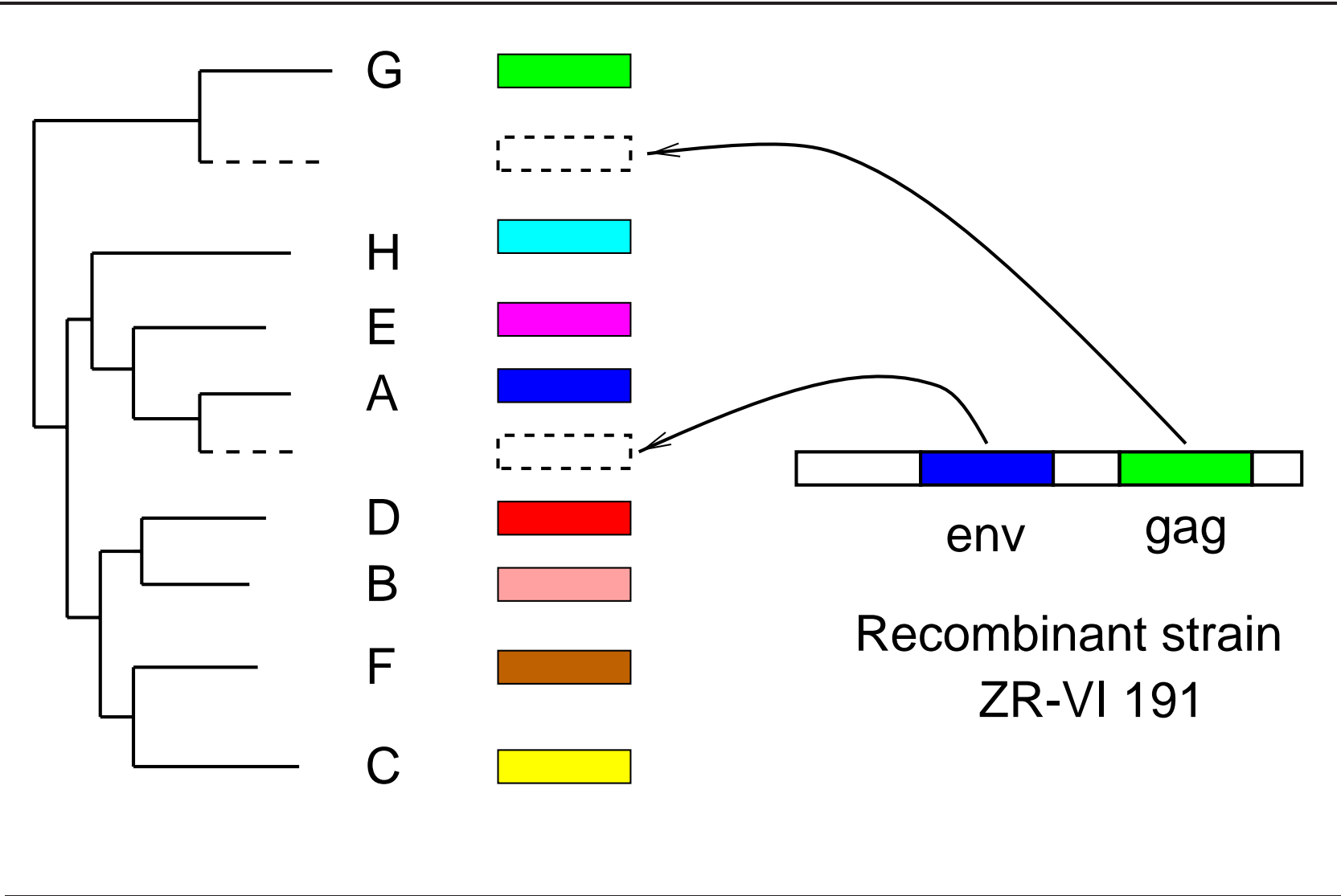
Recombination



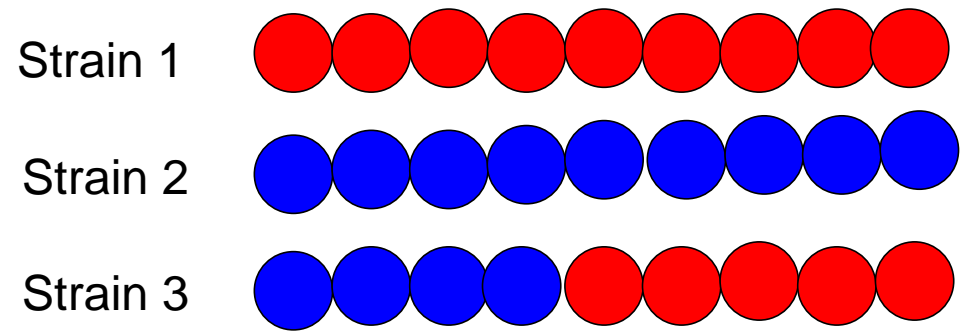
Recombination



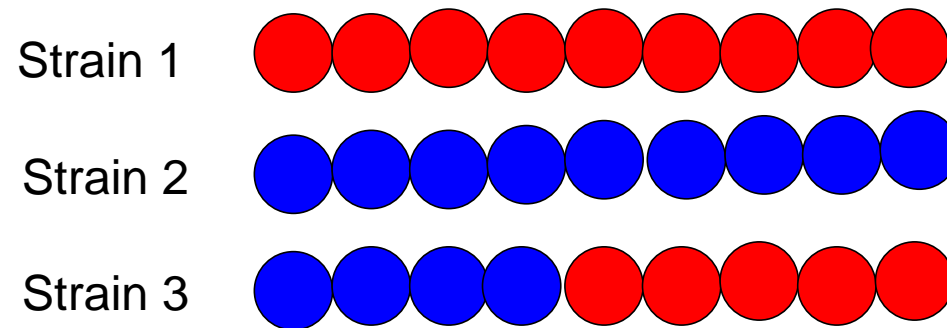
Recombination in HIV 1



Maximum χ^2 method (J. M. Smith, 1992)

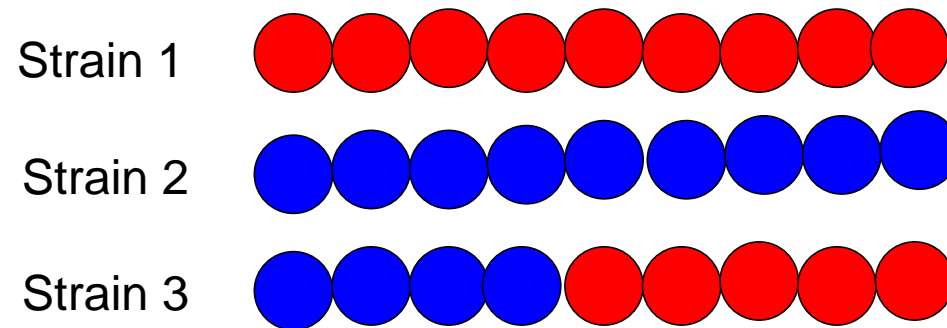


Maximum χ^2 method (J. M. Smith, 1992)



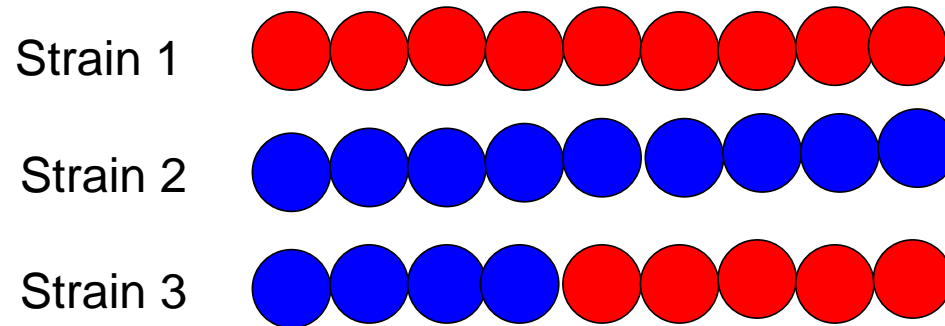
- Consider two sequences, N nucleotides long, with D polymorphic sites .

Maximum χ^2 method (J. M. Smith, 1992)



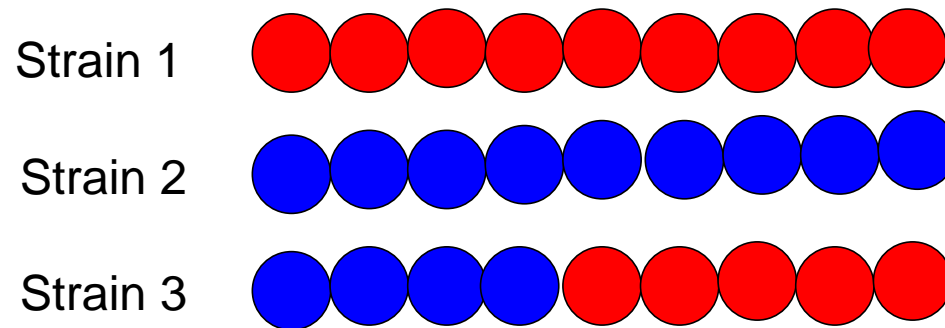
- Consider two sequences, N nucleotides long, with D polymorphic sites .
- Arbitrary cut at position t .
Polymorphic sites \rightarrow Left: x_1 , right: $x_2 = D - x_1$

Maximum χ^2 method (J. M. Smith, 1992)



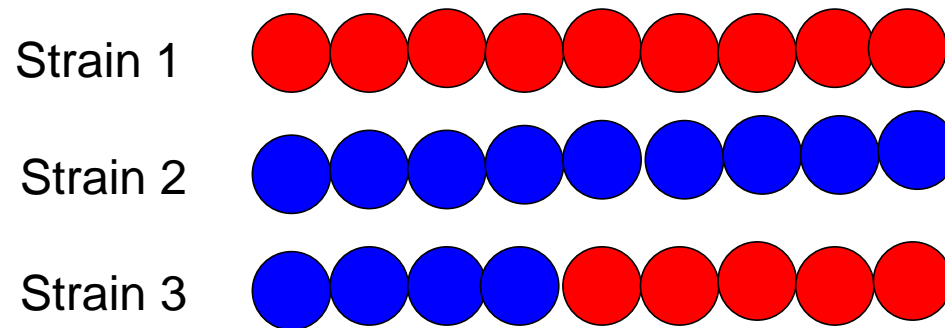
- Consider two sequences, N nucleotides long, with D polymorphic sites .
 - Arbitrary cut at position t .
Polymorphic sites \rightarrow Left: x_1 , right: $x_2 = D - x_1$
 - Expected number of polymorphic sites under a random distribution:
Left: $e_1 = \frac{D}{N}t$, right: $e_2 = \frac{D}{N}(N - t)$
-

Maximum χ^2 method (J. M. Smith, 1992)



- Consider two sequences, N nucleotides long, with D polymorphic sites .
 - Arbitrary cut at position t .
Polymorphic sites \rightarrow Left: x_1 , right: $x_2 = D - x_1$
 - Expected number of polymorphic sites under a random distribution:
Left: $e_1 = \frac{D}{N}t$, right: $e_2 = \frac{D}{N}(N - t)$
 - χ^2 statistic: $\chi^2 = \frac{(x_1 - e_1)^2}{e_1} + \frac{(x_2 - e_2)^2}{e_2}$
-

Maximum χ^2 method (J. M. Smith, 1992)



- Consider two sequences, N nucleotides long, with D polymorphic sites .
 - Arbitrary cut at position t .
Polymorphic sites \rightarrow Left: x_1 , right: $x_2 = D - x_1$
 - Expected number of polymorphic sites under a random distribution:
Left: $e_1 = \frac{D}{N}t$, right: $e_2 = \frac{D}{N}(N - t)$
 - χ^2 statistic: $\chi^2 = \frac{(x_1 - e_1)^2}{e_1} + \frac{(x_2 - e_2)^2}{e_2}$
 - Find cut t that **maximises χ^2** .
-

Maximum χ^2 method: Significance

Are the results **significant** ?

Maximum χ^2 method: Significance

Are the results **significant** ?

- **Asymptotic distribution** ($N \rightarrow \infty$) under the null hypothesis of no recombination: $\chi^2(1)$. But **unreliable for small N** .

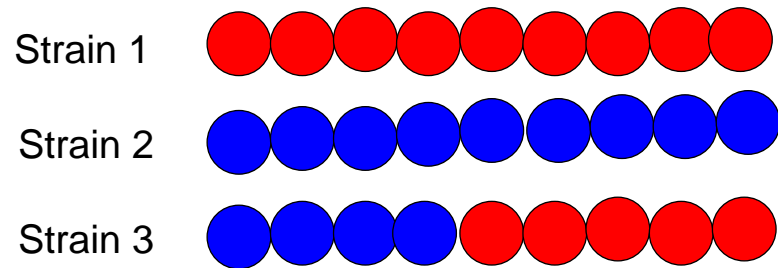
Maximum χ^2 method: Significance

Are the results **significant** ?

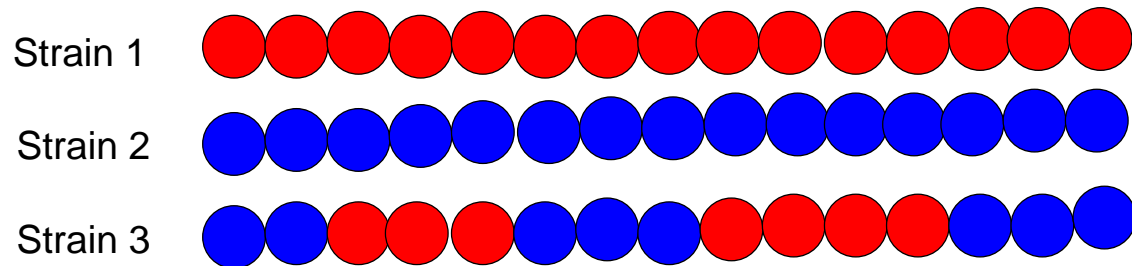
- **Asymptotic distribution** ($N \rightarrow \infty$) under the null hypothesis of no recombination: $\chi^2(1)$. But **unreliable for small N** .
 - **Permutation test:**
 - Permutation of columns, say M times.
 - For each randomised data set, find maximal χ^2 .
 - Empirical distribution under the null hypothesis.
 - If $|\{\chi_{rand}^2 > \chi^2\}| = m$, then $P = \frac{m}{M}$.
-

Maximum χ^2 : Shortcomings

Alignment must have a two-block structure:



Not directly applicable if recombinant regions lie in the middle of the alignment:



Split the data up into smaller subsets, and repeat the analysis many times.
But this is tedious.

Window methods

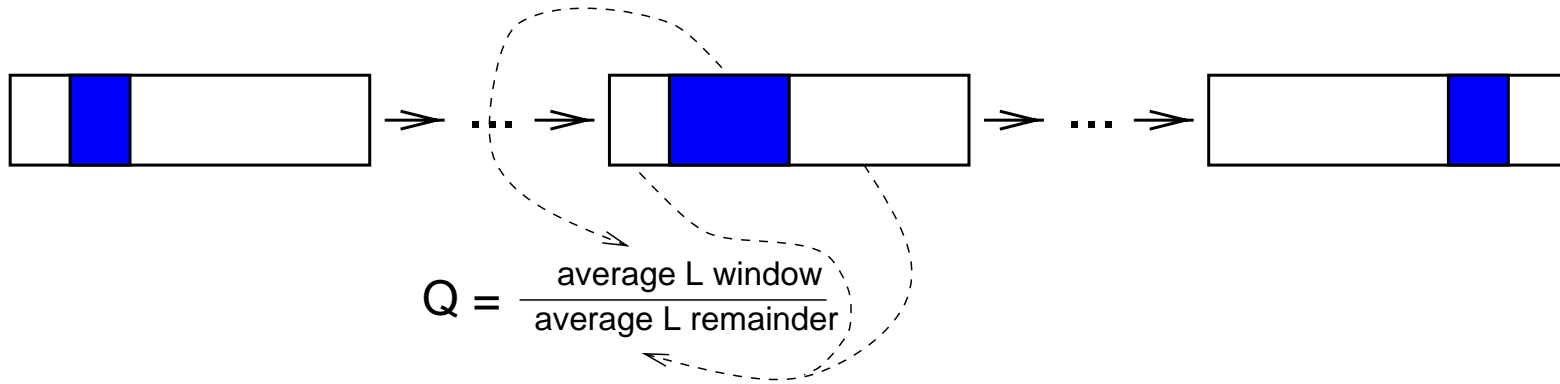
Window methods

- Slide a `window` across the alignment.

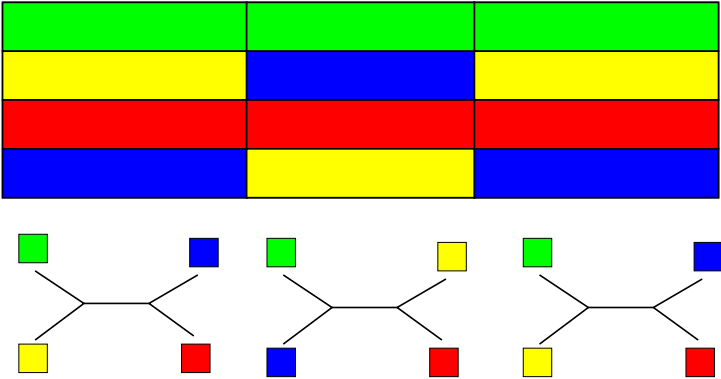
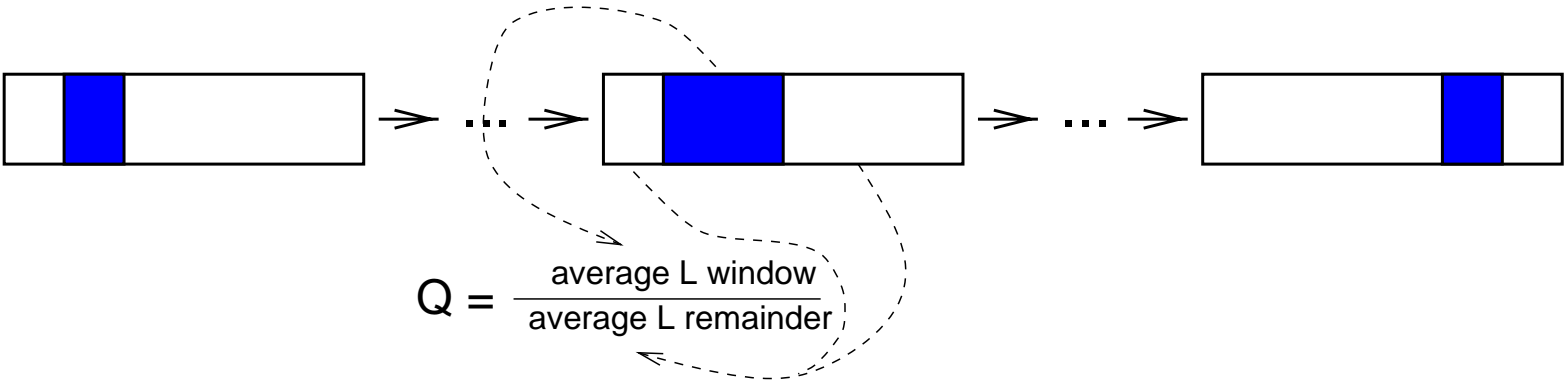
Window methods

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

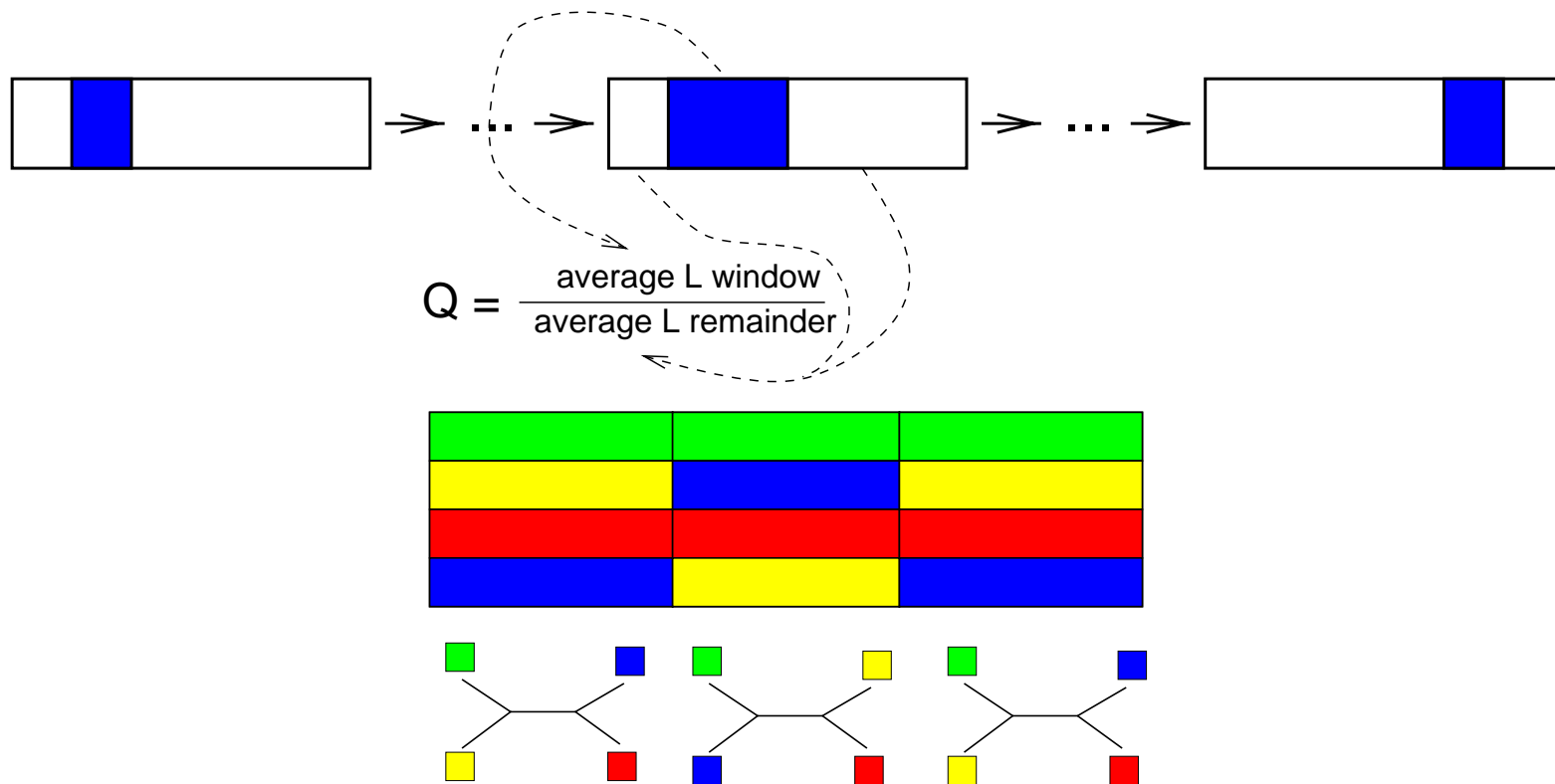
PLATO (Grassly & Holmes, 1997)



PLATO (Grassly & Holmes, 1997)



PLATO (Grassly & Holmes, 1997)



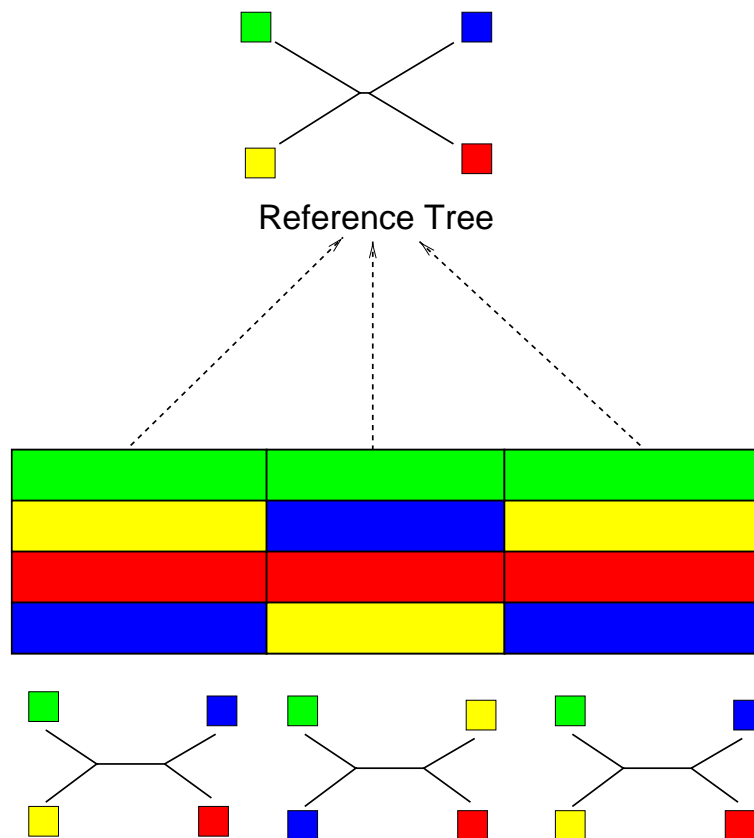
- Find regions with maximum Q values
- Test significance with parametric bootstrapping

Shortcoming of PLATO

- Need a reference tree
- Obtained with global maximum likelihood

Shortcoming of PLATO

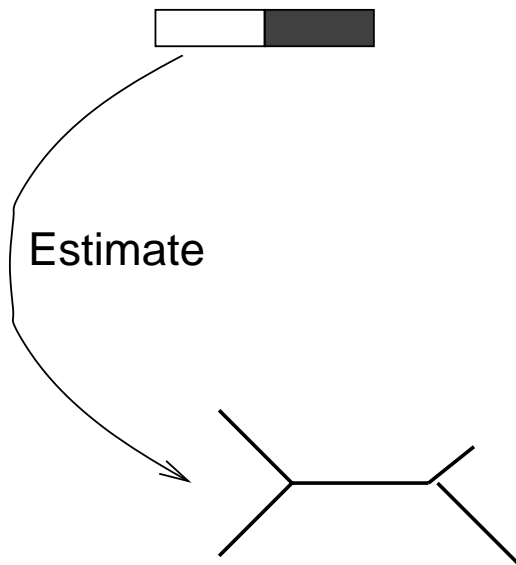
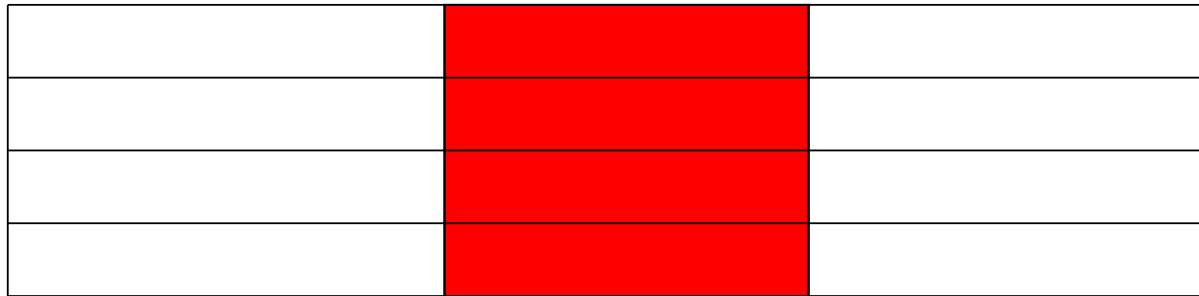
- Need a reference tree
- Obtained with global maximum likelihood



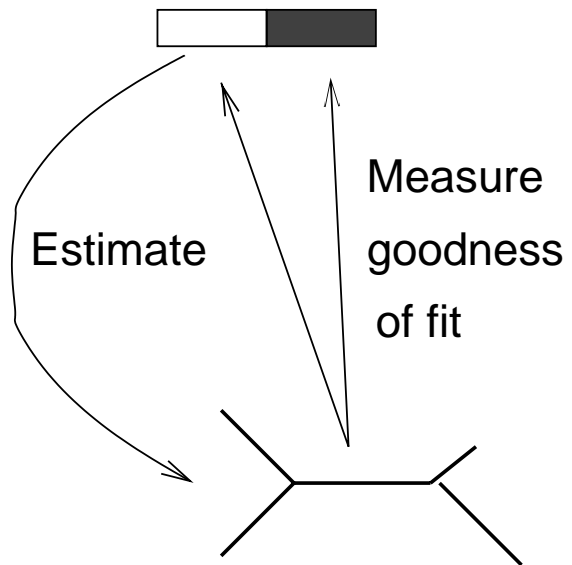
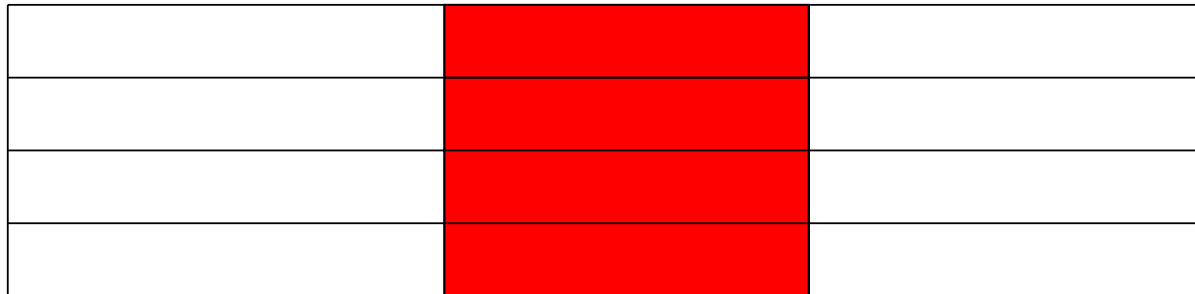
TOPAL (McGuire & Wright, 1997)



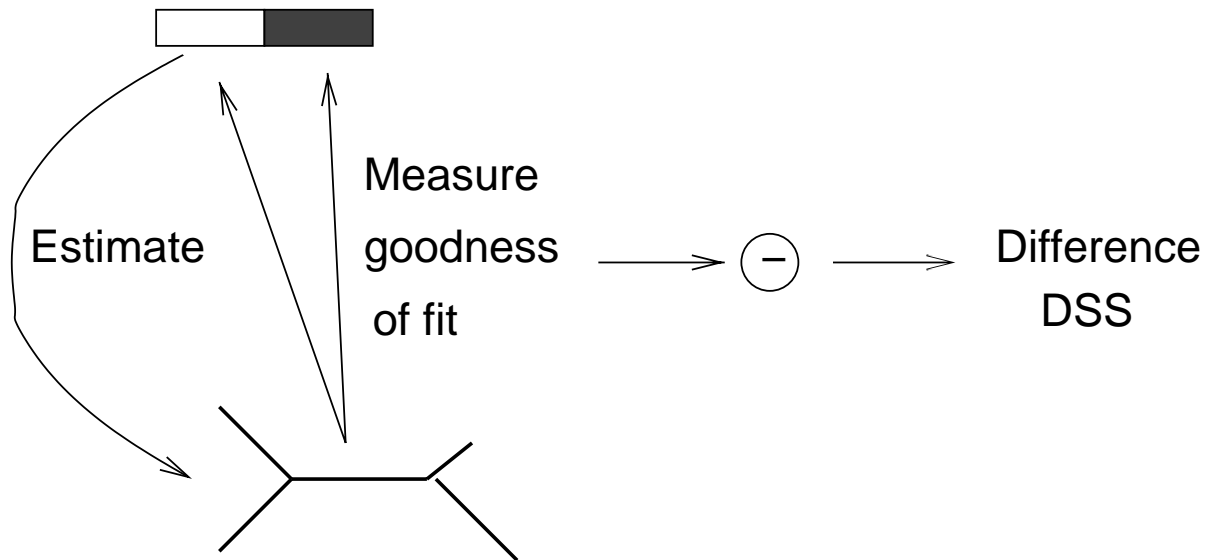
TOPAL (McGuire & Wright, 1997)



TOPAL (McGuire & Wright, 1997)



TOPAL (McGuire & Wright, 1997)



TOPAL (McGuire & Wright, 1997)



DSS small

TOPAL (McGuire & Wright, 1997)

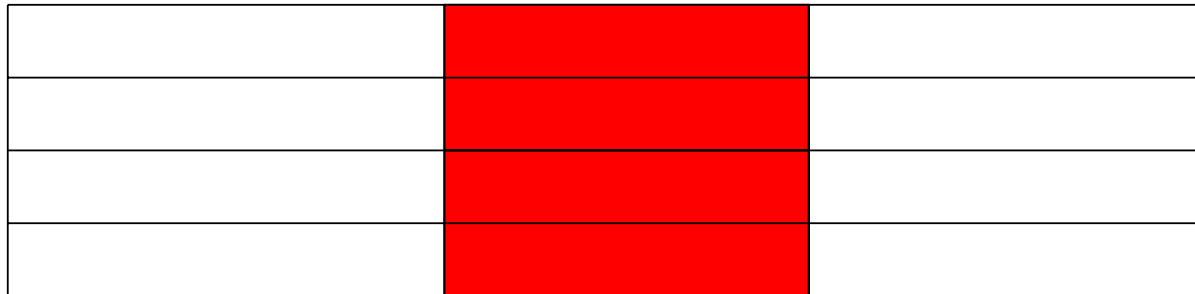


DSS small



DSS large

TOPAL (McGuire & Wright, 1997)



DSS small



DSS large

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

Problems with TOPAL

- Focuses on changes in the pairwise distances .

Problems with TOPAL

- Focuses on changes in the pairwise distances .

$$SS = \sum_i \sum_k (d_{ik} - \hat{d}_{ik})^2 \quad DSS = |SS_{left} - SS_{right}|$$

i, k	labels for taxa
\hat{d}_{ik}	fitted distances (Fitch or Neighbour Joining)
d_{ik}	true distances

Problems with TOPAL

- Focuses on changes in the pairwise distances .

$$SS = \sum_i \sum_k (d_{ik} - \hat{d}_{ik})^2 \quad DSS = |SS_{left} - SS_{right}|$$

i, k	labels for taxa
\hat{d}_{ik}	fitted distances (Fitch or Neighbour Joining)
d_{ik}	true distances

- Difficulties distinguishing between recombination and rate variation .

Problems with TOPAL

- Focuses on changes in the pairwise distances .

$$SS = \sum_i \sum_k (d_{ik} - \hat{d}_{ik})^2 \quad DSS = |SS_{left} - SS_{right}|$$

i, k	labels for taxa
\hat{d}_{ik}	fitted distances (Fitch or Neighbour Joining)
d_{ik}	true distances

- Difficulties distinguishing between recombination and rate variation .
 - Optimisation on the basis of a small data set .
-

Problems with TOPAL

- Focuses on changes in the pairwise distances .

$$SS = \sum_i \sum_k (d_{ik} - \hat{d}_{ik})^2 \quad DSS = |SS_{left} - SS_{right}|$$

i, k	labels for taxa
\hat{d}_{ik}	fitted distances (Fitch or Neighbour Joining)
d_{ik}	true distances

- Difficulties distinguishing between recombination and rate variation .
- Optimisation on the basis of a small data set .

Objective

Problems with TOPAL

- Focuses on changes in the pairwise distances .

$$SS = \sum_i \sum_k (d_{ik} - \hat{d}_{ik})^2 \quad DSS = |SS_{left} - SS_{right}|$$

i, k	labels for taxa
\hat{d}_{ik}	fitted distances (Fitch or Neighbour Joining)
d_{ik}	true distances

- Difficulties distinguishing between recombination and rate variation .
- Optimisation on the basis of a small data set .

Objective

- Focus on topology changes .
-

Problems with TOPAL

- Focuses on changes in the pairwise distances .

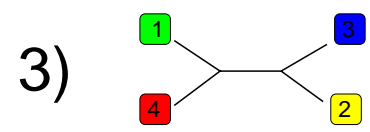
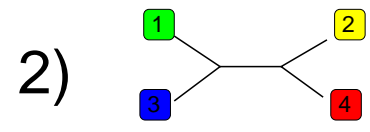
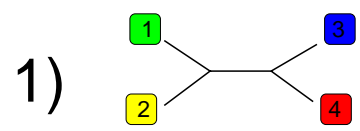
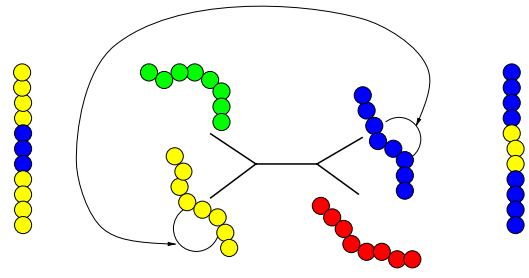
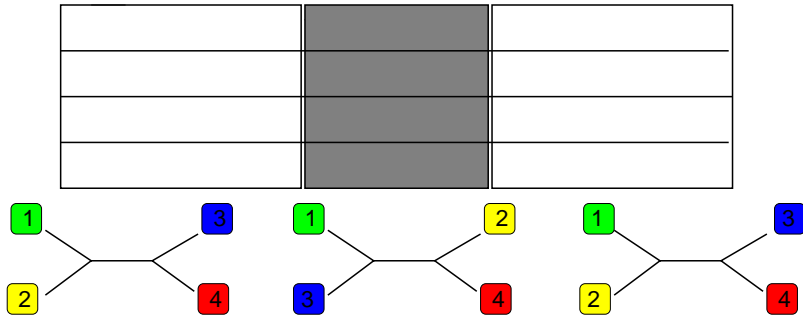
$$SS = \sum_i \sum_k (d_{ik} - \hat{d}_{ik})^2 \quad DSS = |SS_{left} - SS_{right}|$$

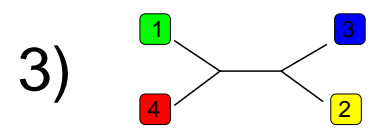
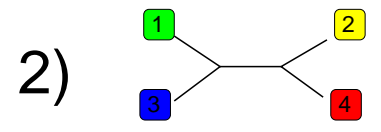
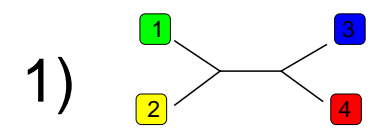
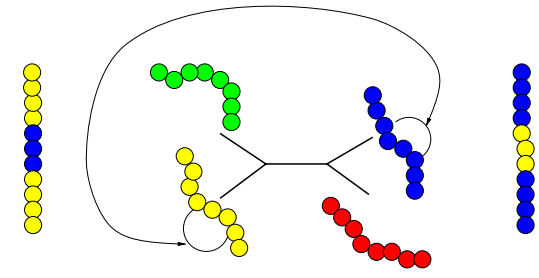
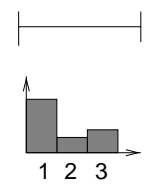
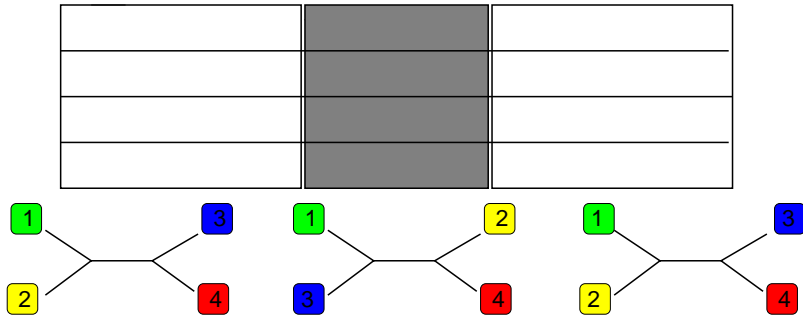
i, k	labels for taxa
\hat{d}_{ik}	fitted distances (Fitch or Neighbour Joining)
d_{ik}	true distances

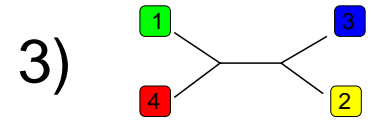
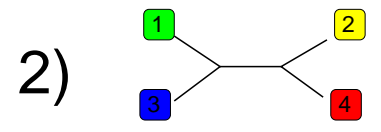
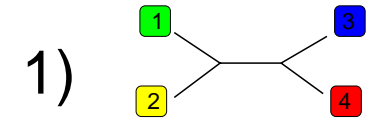
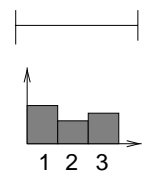
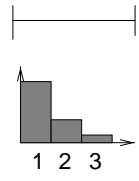
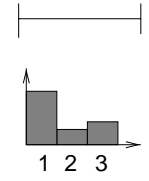
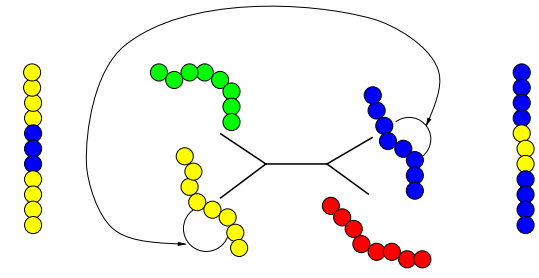
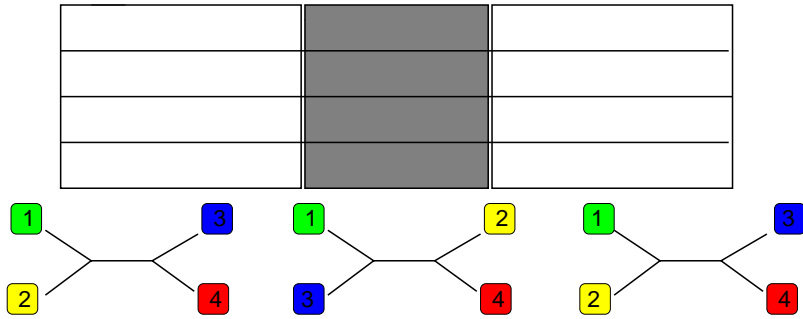
- Difficulties distinguishing between recombination and rate variation .
- Optimisation on the basis of a small data set .

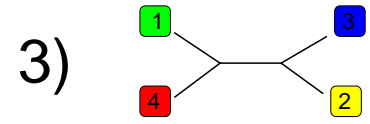
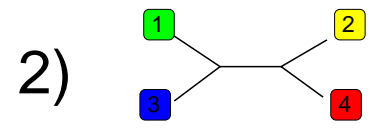
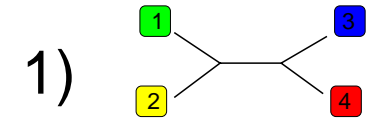
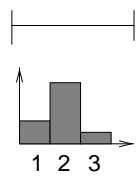
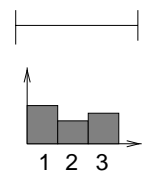
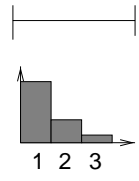
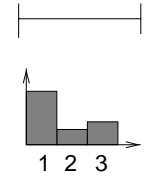
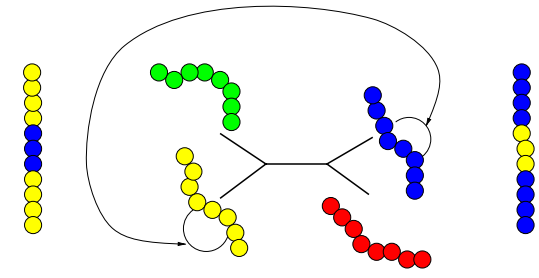
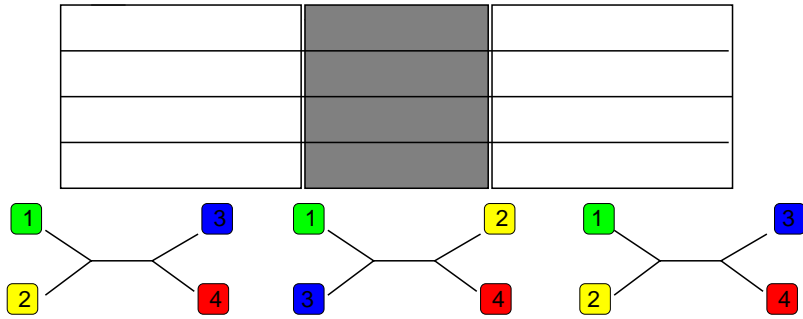
Objective

- Focus on topology changes .
 - Capture uncertainty .
-

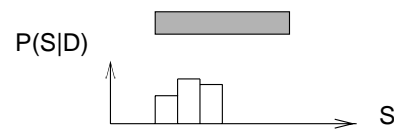
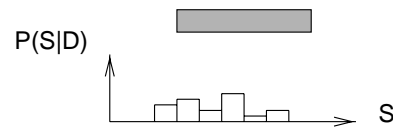
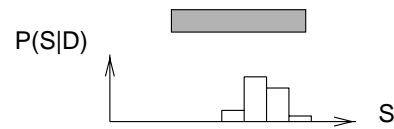
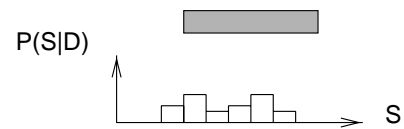
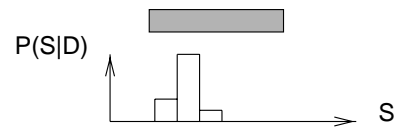
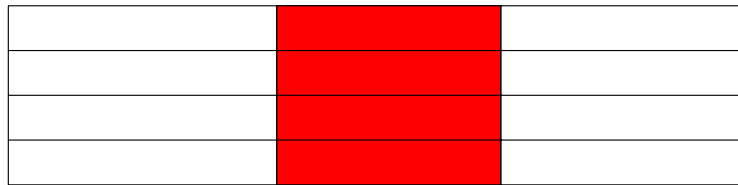




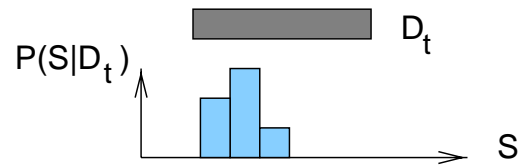




Detection of recombination with MCMC

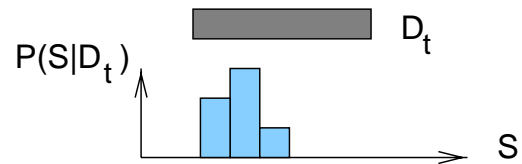
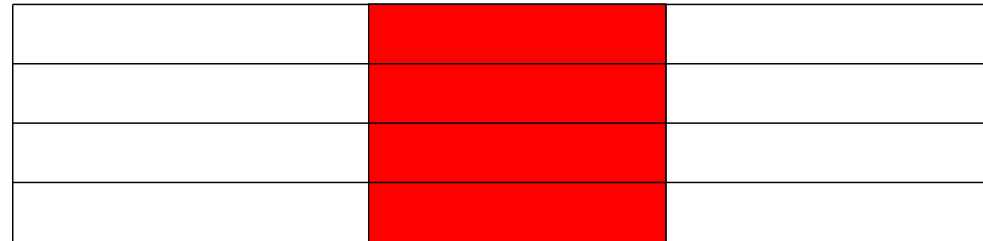


Marginal posterior distribution of tree topologies with MCMC



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

Marginal posterior distribution of tree topologies with MCMC



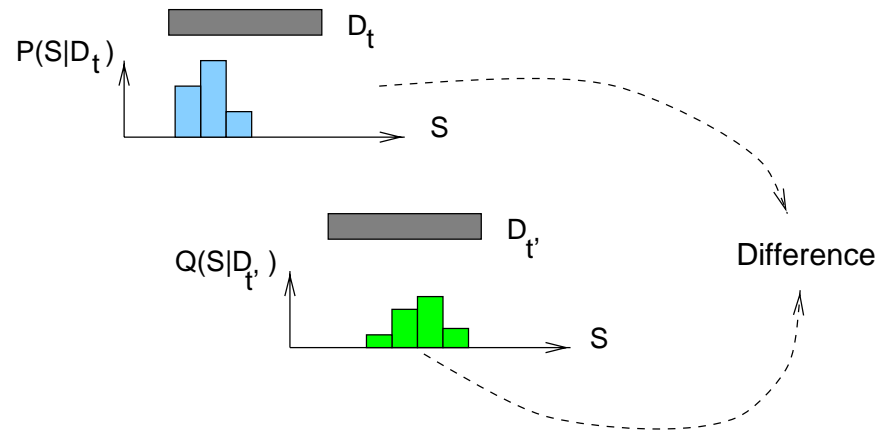
$$P(S|\mathcal{D}_t) = \int P(S, \mathbf{w}|\mathcal{D}_t) d\mathbf{w}$$

$$\text{MCMC} \longrightarrow \text{Sample : } \{S_{ti}, \mathbf{w}_{ti}\}_{i=1}^N$$

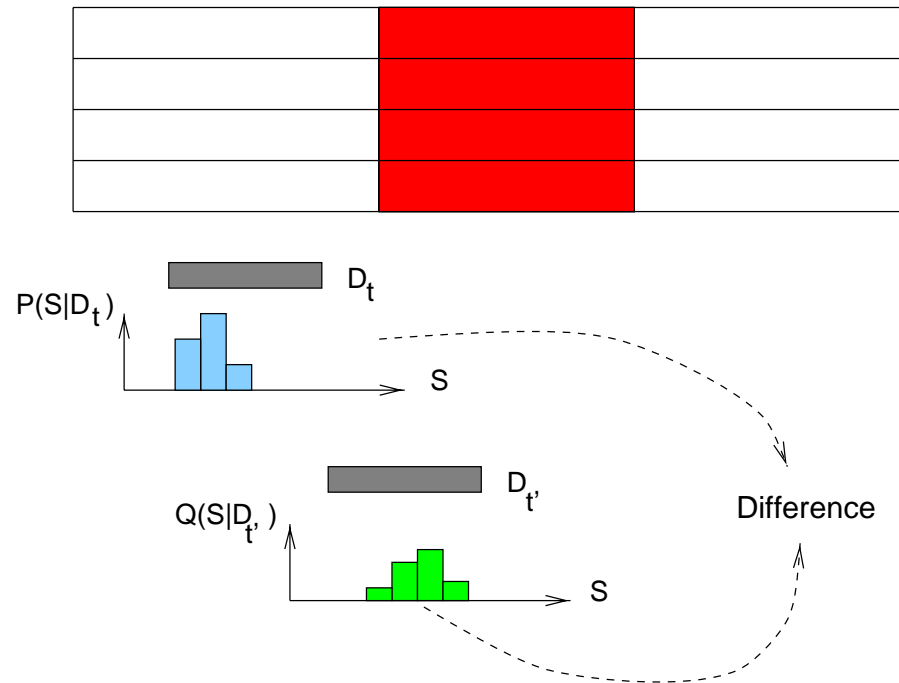
$$P(S, \mathbf{w}|\mathcal{D}_t) \approx \frac{1}{N} \sum_{i=1}^N \delta_{S, S_{ti}} \delta(\mathbf{w} - \mathbf{w}_{ti})$$

$$P(S|\mathcal{D}_t) = \frac{1}{N} \sum_{i=1}^N \delta_{S, S_{ti}} = \frac{N_S(t)}{N}$$

Divergence between distributions



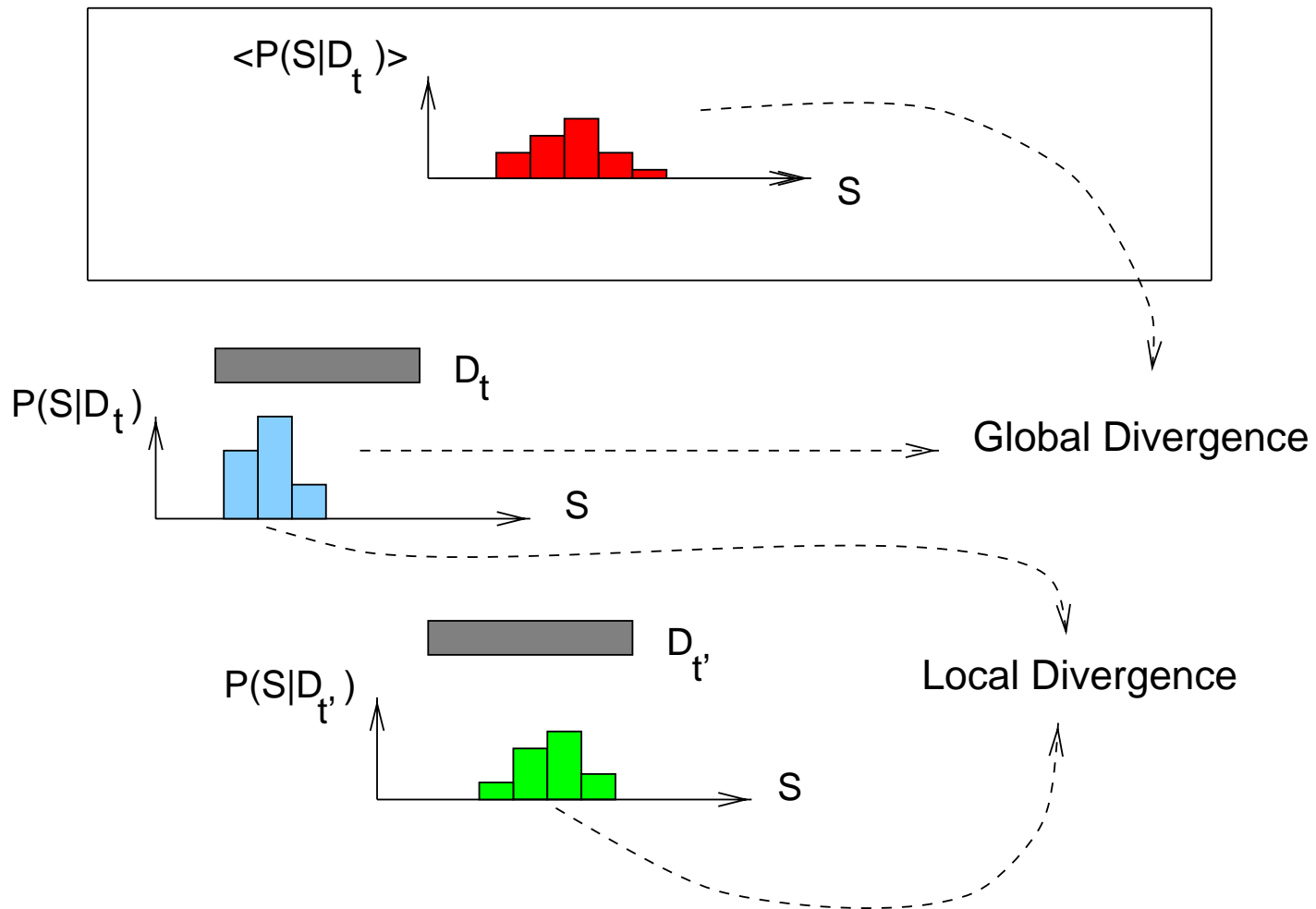
Divergence between distributions



Divergence measure in probability space: **Kullback-Leibler divergence**

$$KL(P, Q) = \sum_S P_S \ln \left(\frac{P_S}{Q_S} \right)$$

Local and global divergence measures



Divergence measures and statistical significance

Global divergence between the distribution over the window, $P_S(t)$, and the average distribution, $\bar{P} = \frac{1}{W} \sum_{t=1}^W P_S(t)$:

$$d[P_S(t), \bar{P}] = \sum_S P_S(t) \ln \left(\frac{P_S(t)}{\bar{P}_S} \right)$$

Divergence measures and statistical significance

Global divergence between the distribution over the window, $P_S(t)$, and the average distribution, $\bar{P} = \frac{1}{W} \sum_{t=1}^W P_S(t)$:

$$d[P_S(t), \bar{P}] = \sum_S P_S(t) \ln \left(\frac{P_S(t)}{\bar{P}_S} \right)$$

Local divergence between the distributions over two adjacent windows, $P_S(t)$ and $P_S(t')$, where $\tilde{P}_S = \frac{P_S(t) + P_S(t')}{2}$ (Sibson):

$$d[P_S(t), P_S(t')] = \frac{1}{2} \sum_S \left[P_S(t) \ln \left(\frac{P_S(t)}{\tilde{P}_S} \right) + P_S(t') \ln \left(\frac{P_S(t')}{\tilde{P}_S} \right) \right]$$

Divergence measures and statistical significance

Global divergence between the distribution over the window, $P_S(t)$, and the average distribution, $\bar{P} = \frac{1}{W} \sum_{t=1}^W P_S(t)$:

$$d[P_S(t), \bar{P}] = \sum_S P_S(t) \ln \left(\frac{P_S(t)}{\bar{P}_S} \right)$$

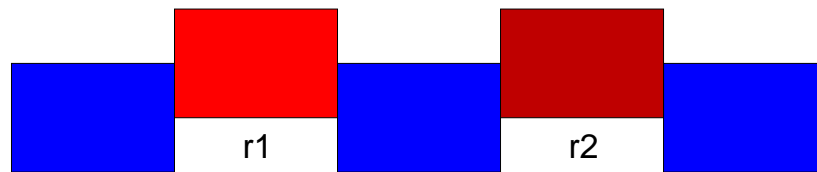
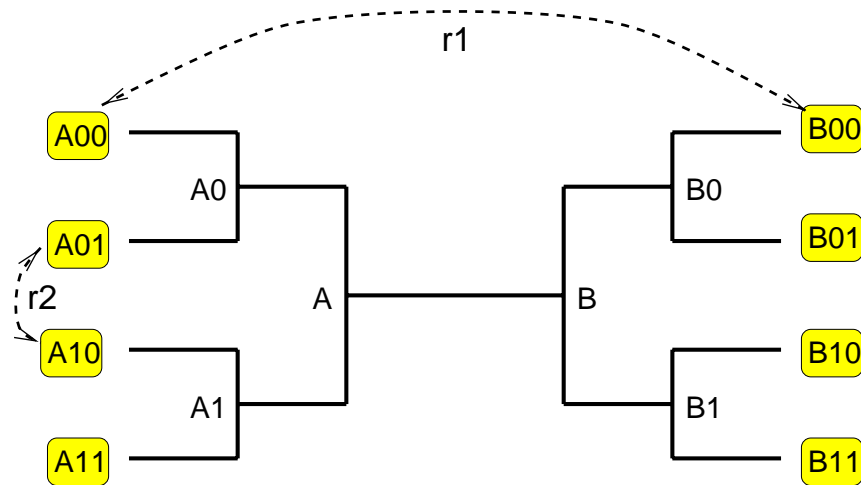
Local divergence between the distributions over two adjacent windows, $P_S(t)$ and $P_S(t')$, where $\tilde{P}_S = \frac{P_S(t) + P_S(t')}{2}$ (Sibson):

$$d[P_S(t), P_S(t')] = \frac{1}{2} \sum_S \left[P_S(t) \ln \left(\frac{P_S(t)}{\tilde{P}_S} \right) + P_S(t') \ln \left(\frac{P_S(t')}{\tilde{P}_S} \right) \right]$$

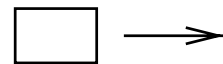
Null hypotheses: $P_S(t) = \bar{P}_S$ and $P_S(t) = P_S(t')$

$$\begin{aligned} 2Nd[P_S(t), \bar{P}] &\rightarrow \chi^2(\nu - 1), & \nu &= |\text{Support}(\bar{P})| \\ 2Nd[P_S(t), P_S(t')] &\rightarrow \chi^2(\tilde{\nu} - 1), & \tilde{\nu} &= |\text{Support}(\tilde{P})| \end{aligned}$$

Simulation experiment A

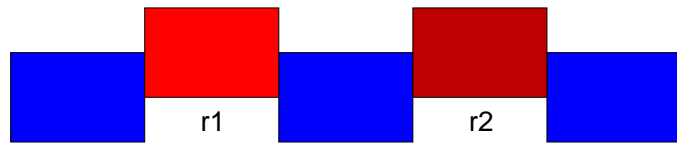
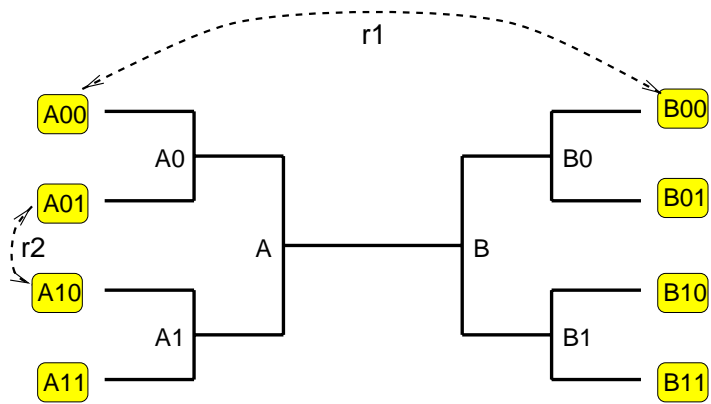


5000 nucleotides

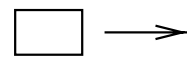


window size = 500 nucleotides

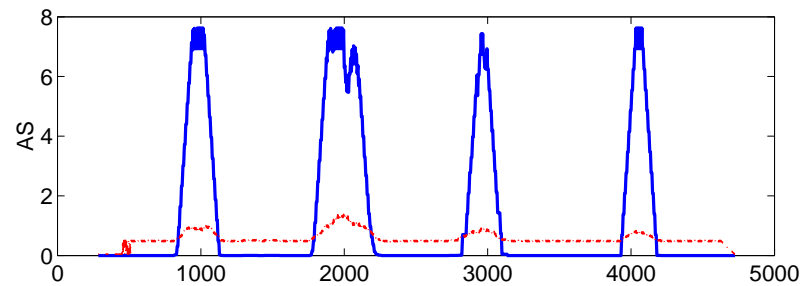
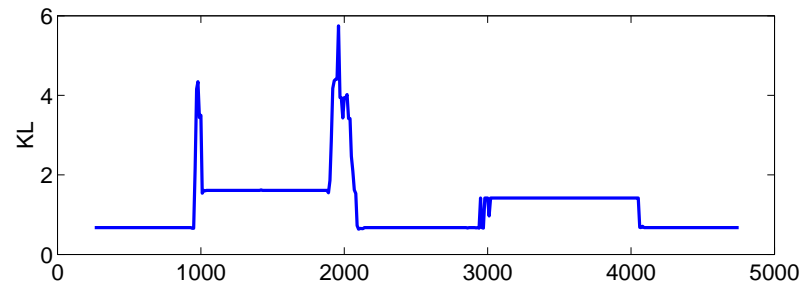
Simulation experiment A



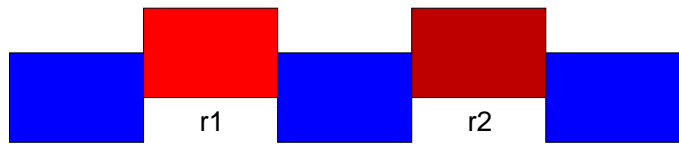
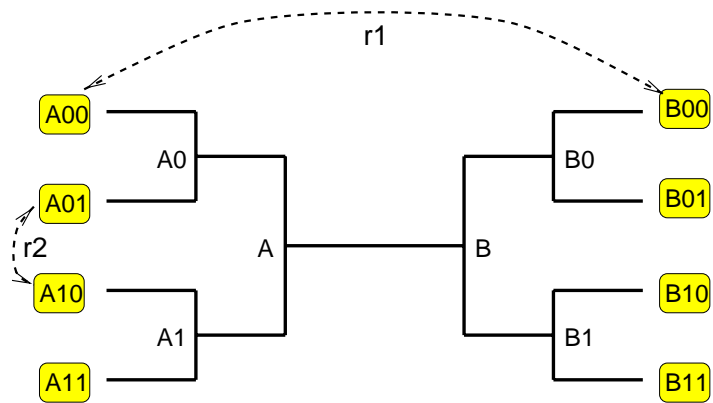
5000 nucleotides



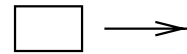
window size = 500 nucleotides



Simulation experiment A



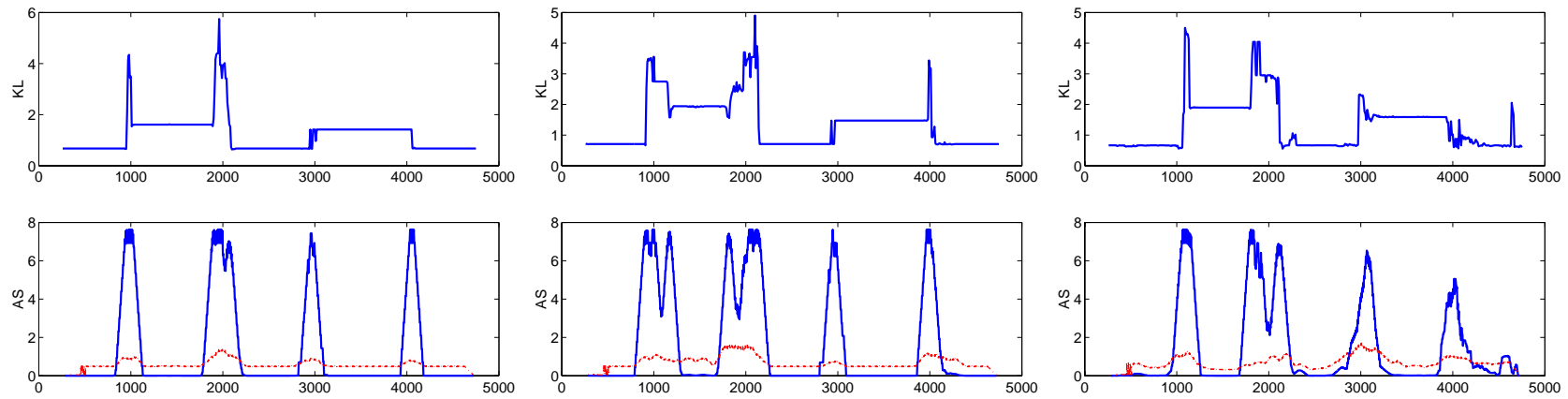
5000 nucleotides



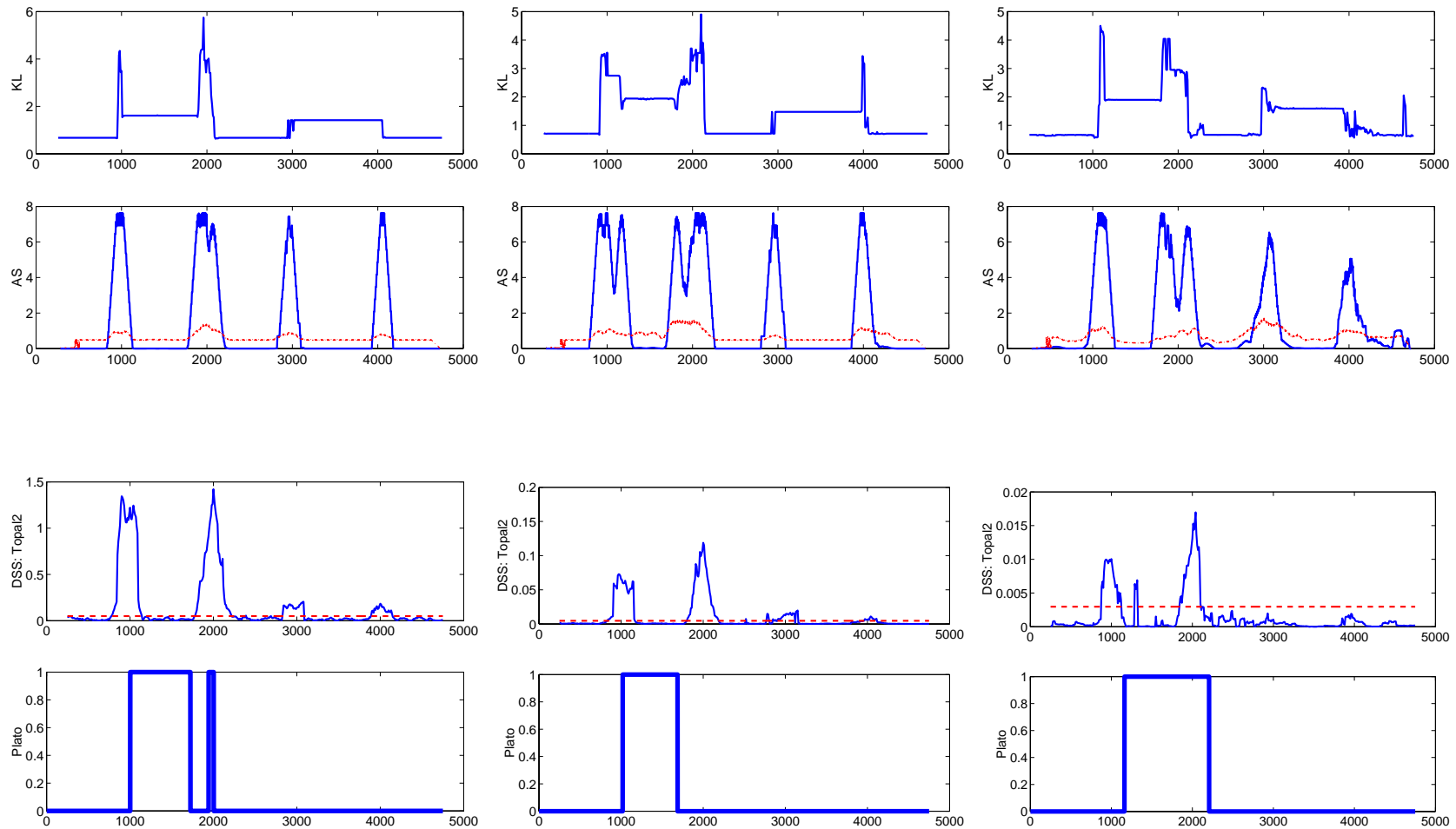
window size = 500 nucleotides

MCMC Global			
MCMC Local			
TOPAL			
PLATO			

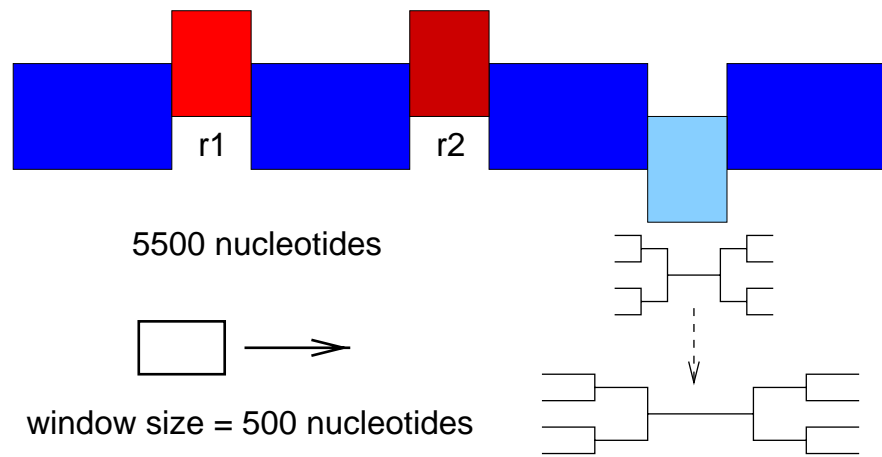
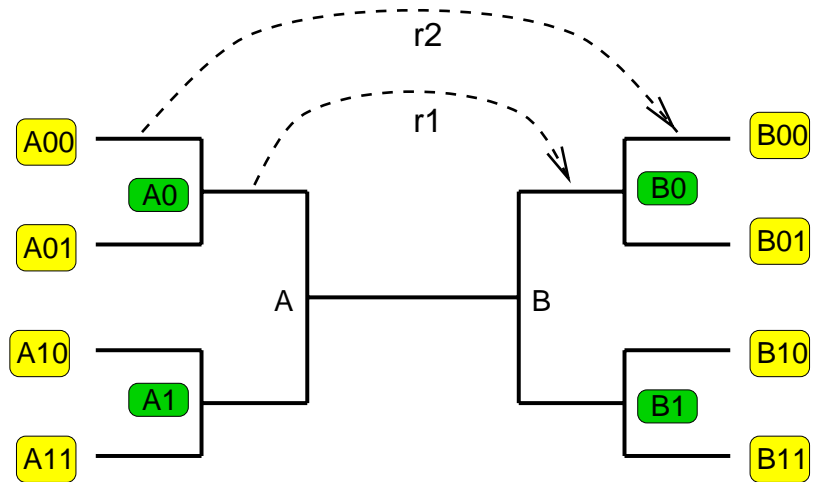
Results - Simulation experiment A



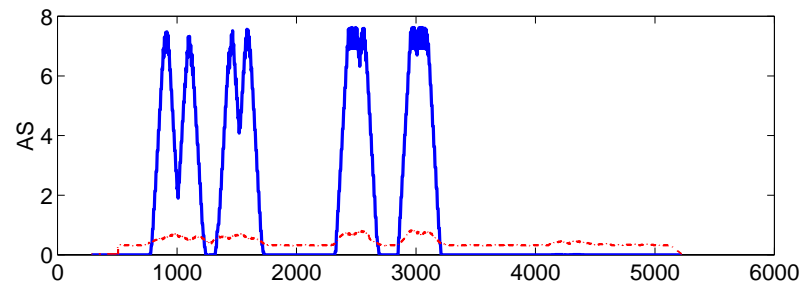
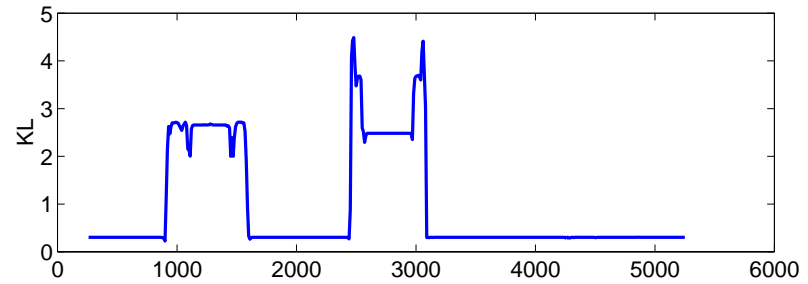
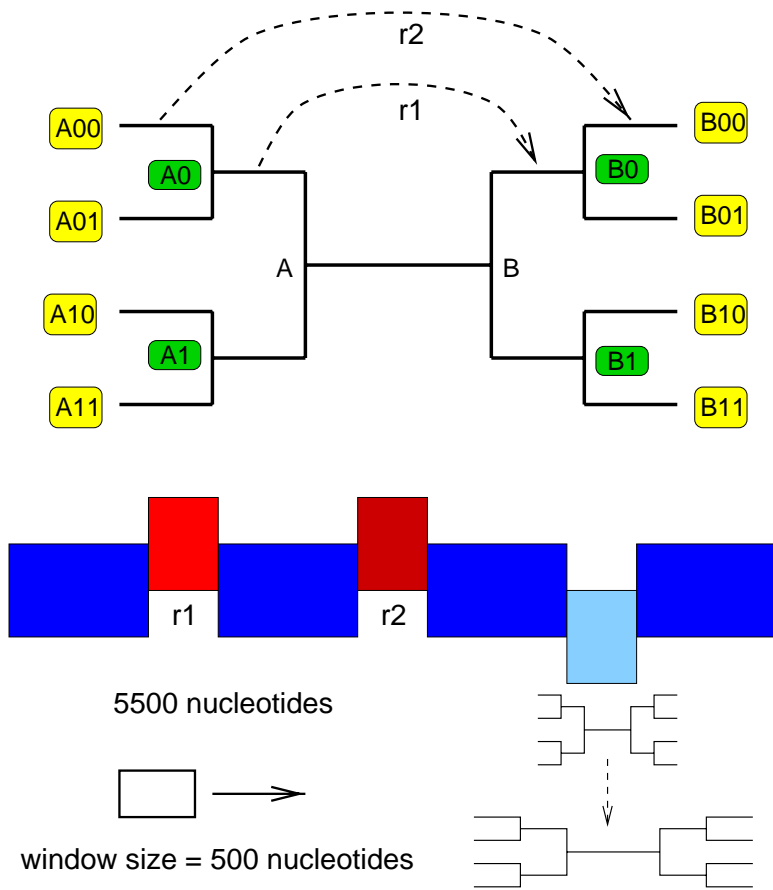
Results - Simulation experiment A



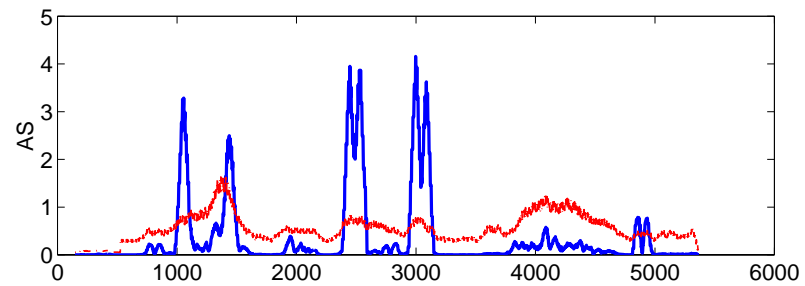
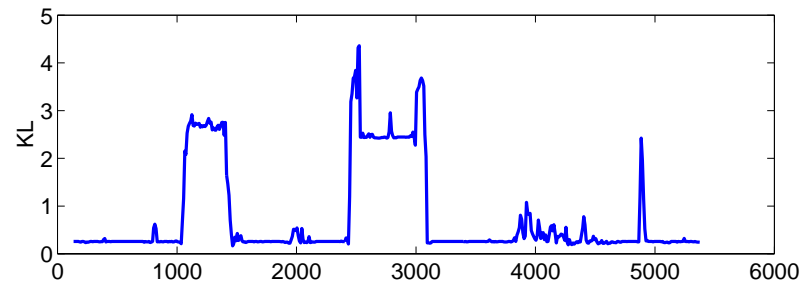
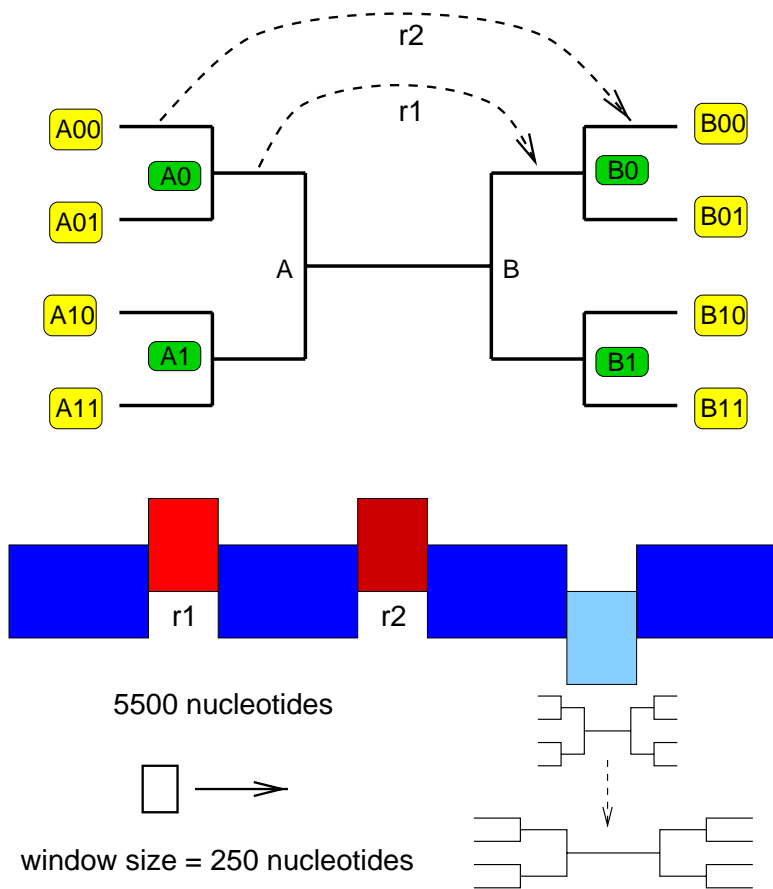
Simulation experiment B



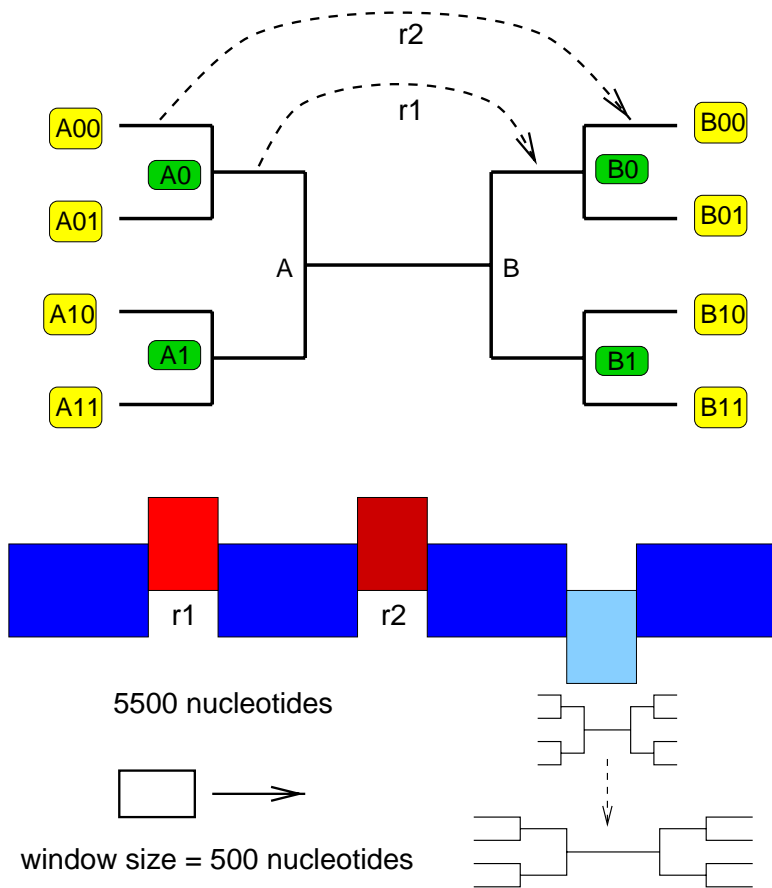
Simulation experiment B



Simulation experiment B: smaller window



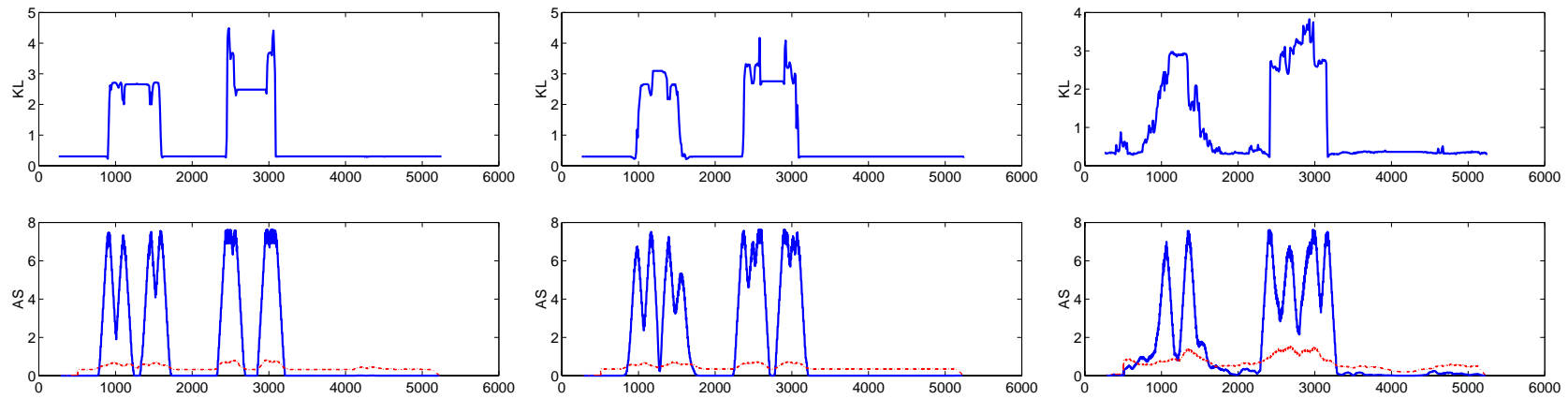
Simulation experiment B



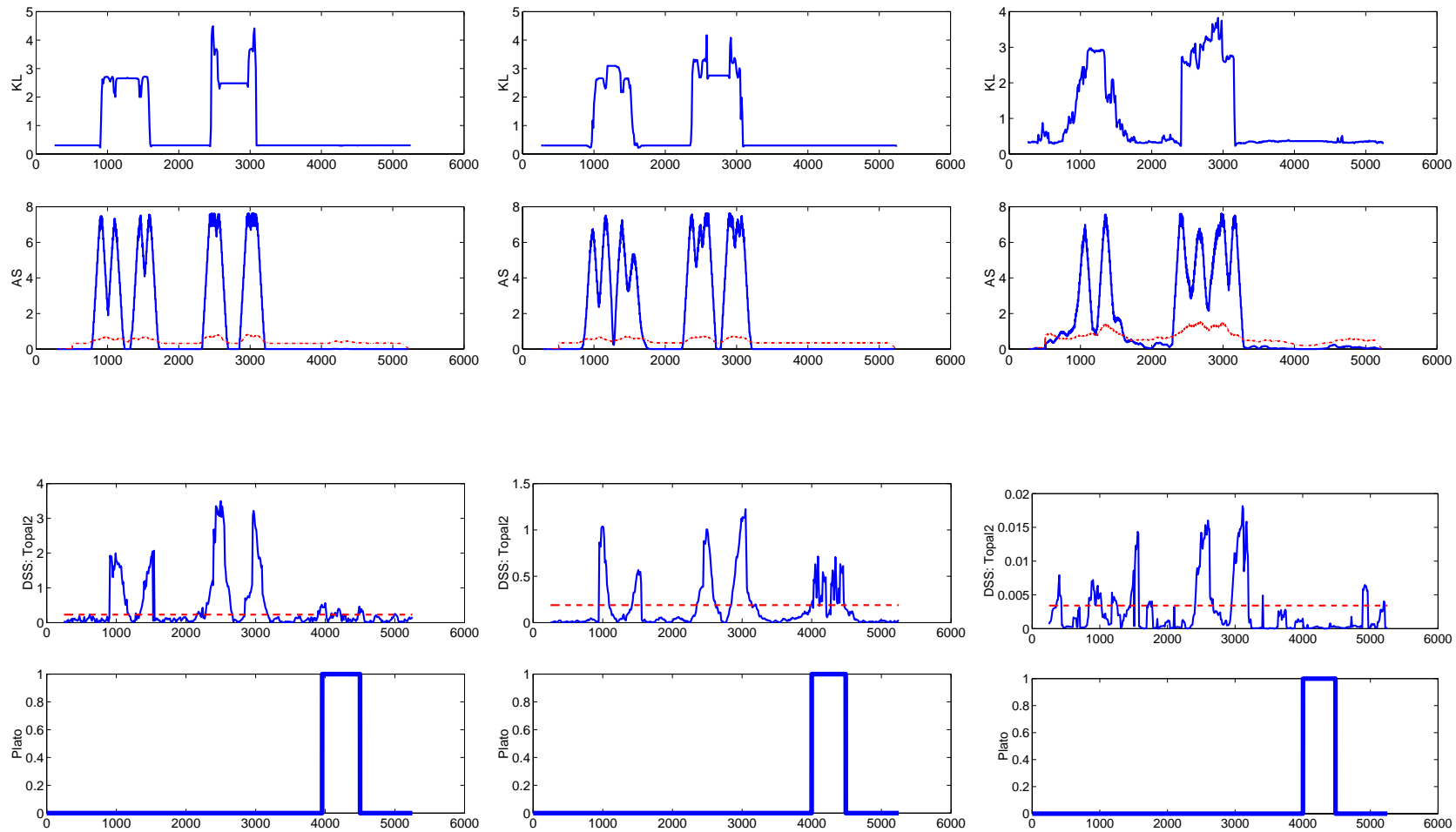
The table compares three methods: MCMC Global, MCMC Local, and PLATO. The columns represent different tree topologies, as indicated by the three tree diagrams above the table. The cells in the table are shaded pink, indicating that all methods are applicable to all topologies.

MCMC Global			
MCMC Local			
TOPAL			
PLATO			

Results - Simulation experiment B



Results - Simulation experiment B



Hepatitis B virus

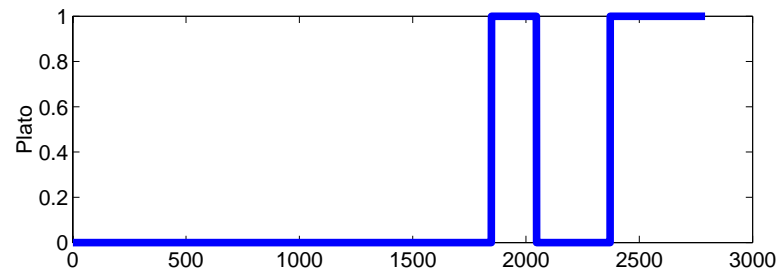
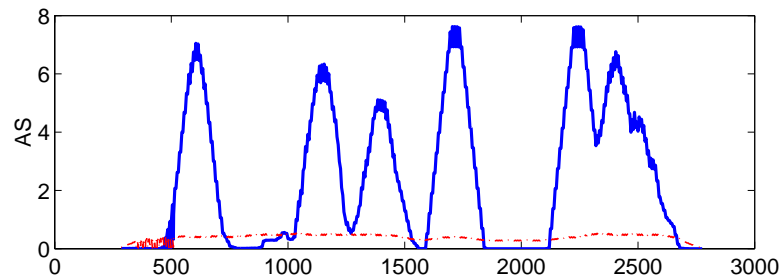
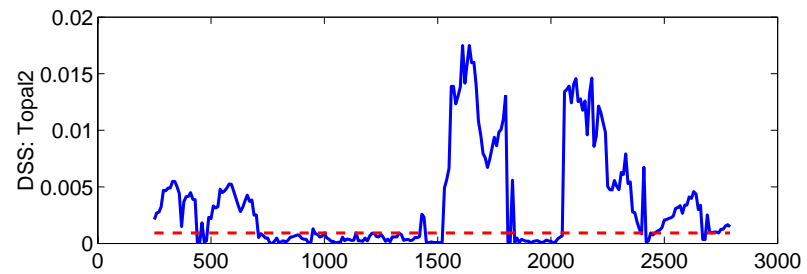
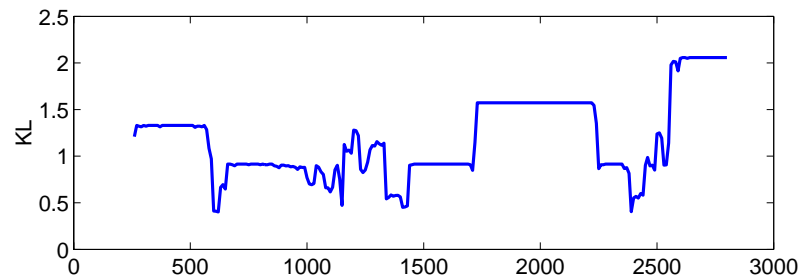
Five strains, 3050 bases, window size= 500 bases.

MCMC, global	TOPAL
MCMC, local	PLATO

Hepatitis B virus

Five strains, 3050 bases, window size= 500 bases.

MCMC, global	TOPAL
MCMC, local	PLATO



Conclusions

- MCMC window method:

Conclusions

- MCMC window method:
 - Improvement on PLATO : Local distributions.

Conclusions

- MCMC window method:
 - Improvement on PLATO : Local distributions.
 - Improvement on TOPAL : Focuses on topology changes and captures uncertainty .

Conclusions

- MCMC window method:
 - Improvement on PLATO : Local distributions.
 - Improvement on TOPAL : Focuses on topology changes and captures uncertainty .
- Catch 22:

Conclusions

- MCMC window method:
 - Improvement on PLATO : Local distributions.
 - Improvement on TOPAL : Focuses on topology changes and captures uncertainty .
- Catch 22:
 - Informative posterior distributions \longrightarrow large window .

Conclusions

- MCMC window method:
 - Improvement on PLATO : Local distributions.
 - Improvement on TOPAL : Focuses on topology changes and captures uncertainty .
- Catch 22:
 - Informative posterior distributions \longrightarrow large window .
 - Good spatial resolution \longrightarrow small window .

Hidden Markov models (HMMs)

Hidden Markov models (HMMs)

- No window needed.

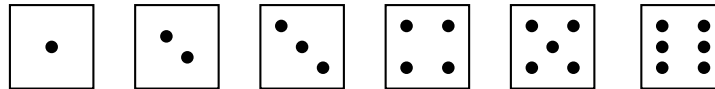
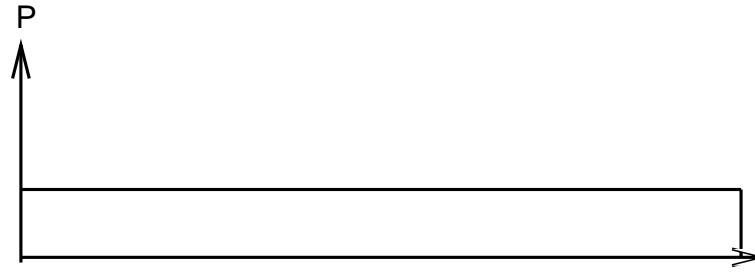
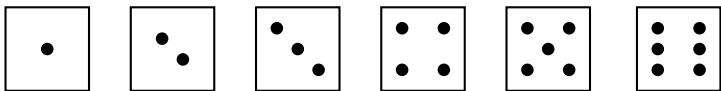
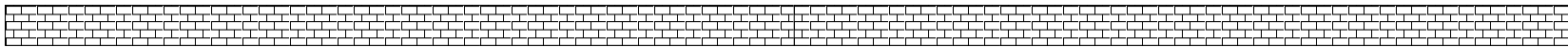
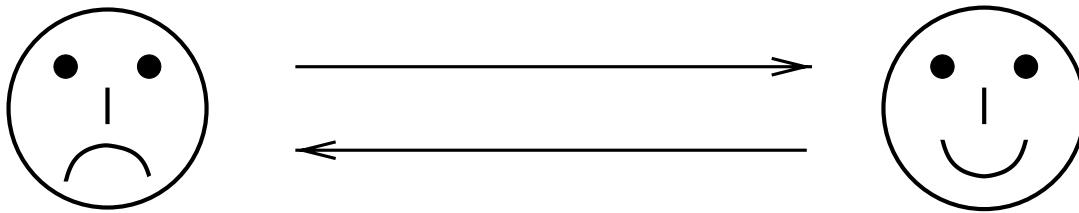
Hidden Markov models (HMMs)

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

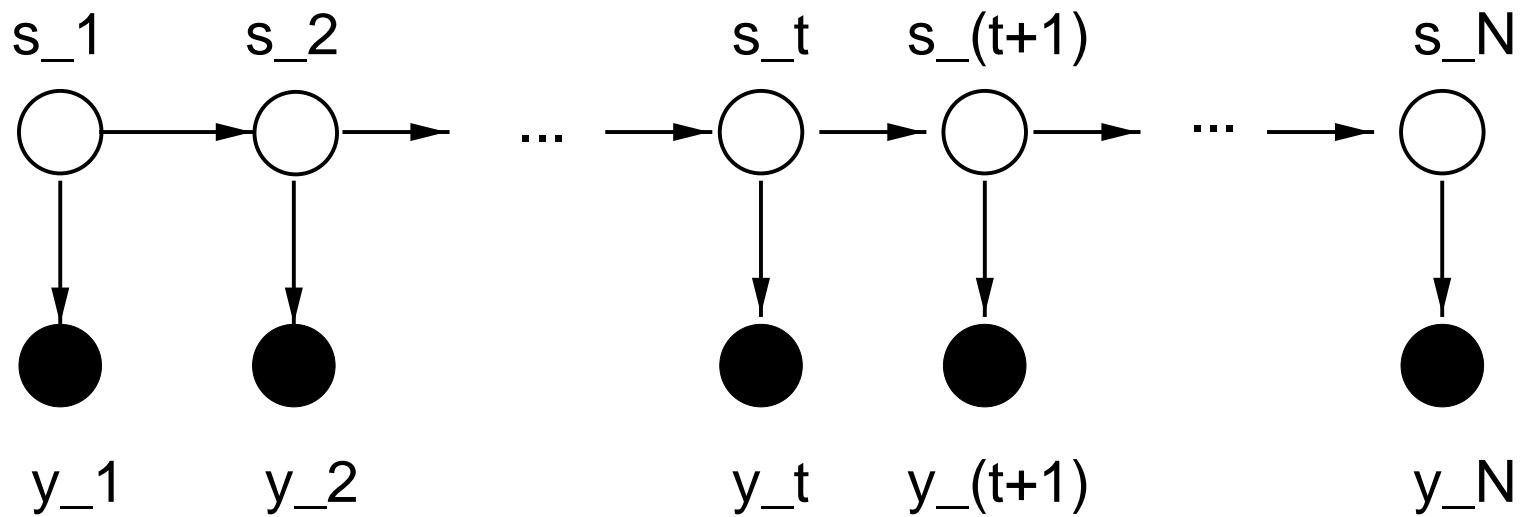
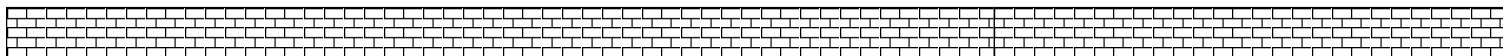
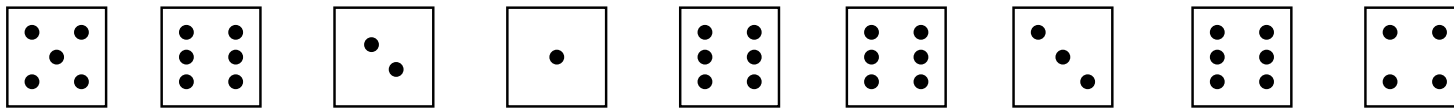
Hidden Markov models (HMMs)

- No window needed.
 - More precise location of the breakpoints.
 - Can only deal with a small number of species (4 or 5).
-

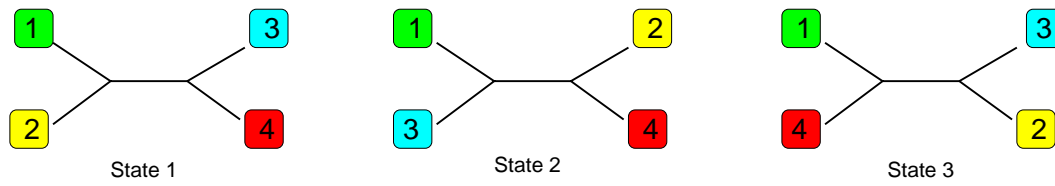
Introduction to HMMs



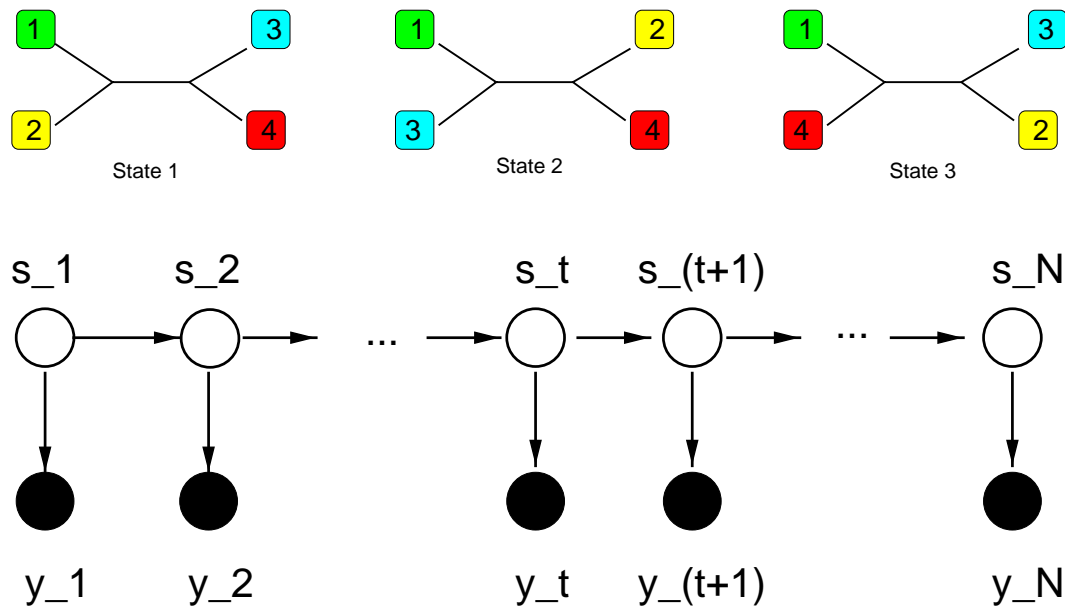
Introduction to HMMs



Modelling recombination with HMMs

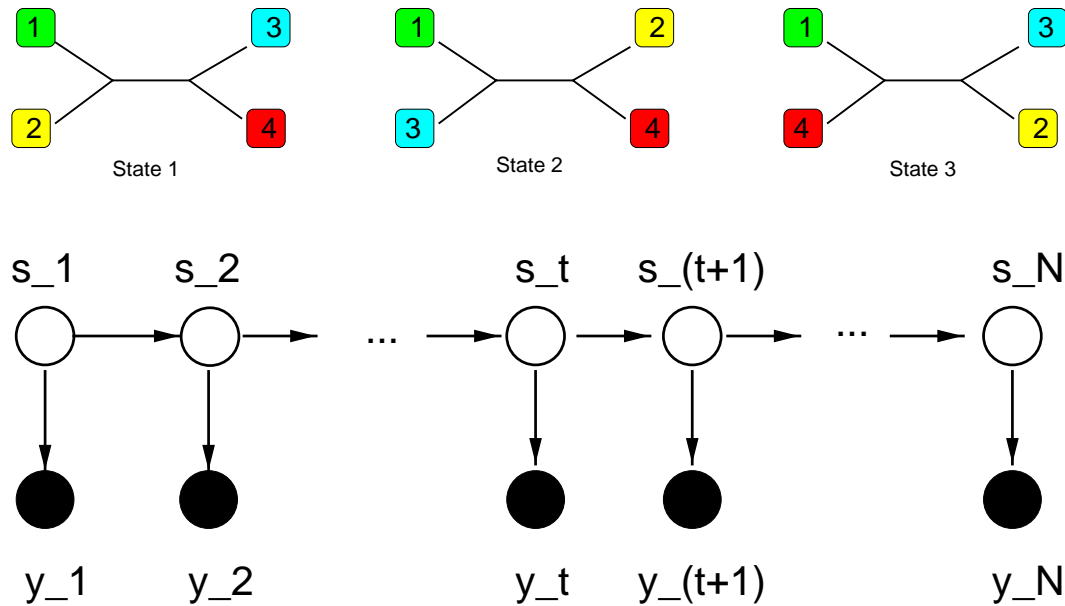


Modelling recombination with HMMs



AG C ATCGTTCTATTTTACCGGCTCCCG
TG T GTCGCTCAAGATTGCCATCGCGCG
TG T CGTGGTCTAGATTGCCATCGCGCG
TG T ATCGCTCTAGTTTGCCAGCTCCCG

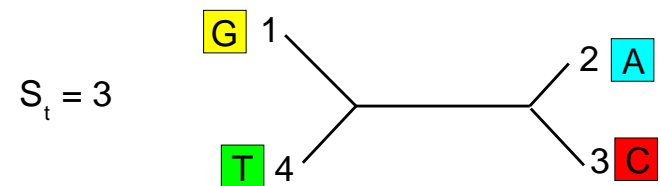
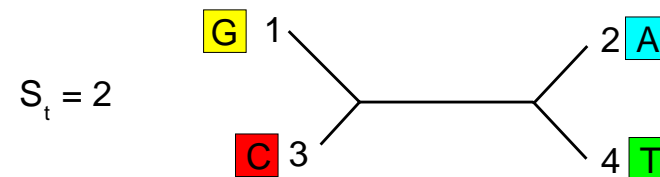
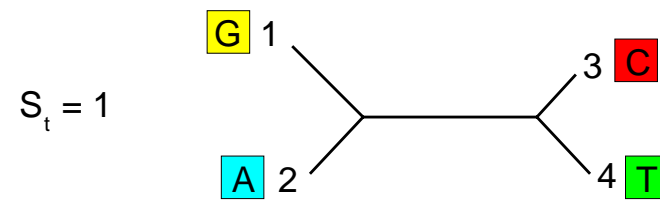
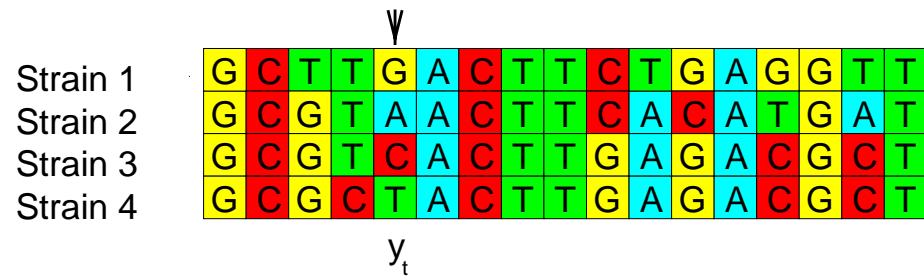
Modelling recombination with HMMs



AG C ATCGTTCTATTTTACCGGCTCCCG
 TG T GTCGCTCAAGATTGCCATCGCGCG
 TG T CGTGGTCTAGATTGCCATCGCGCG
 TG T ATCGCTCTAGTTTGCCAGCTCCCG

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

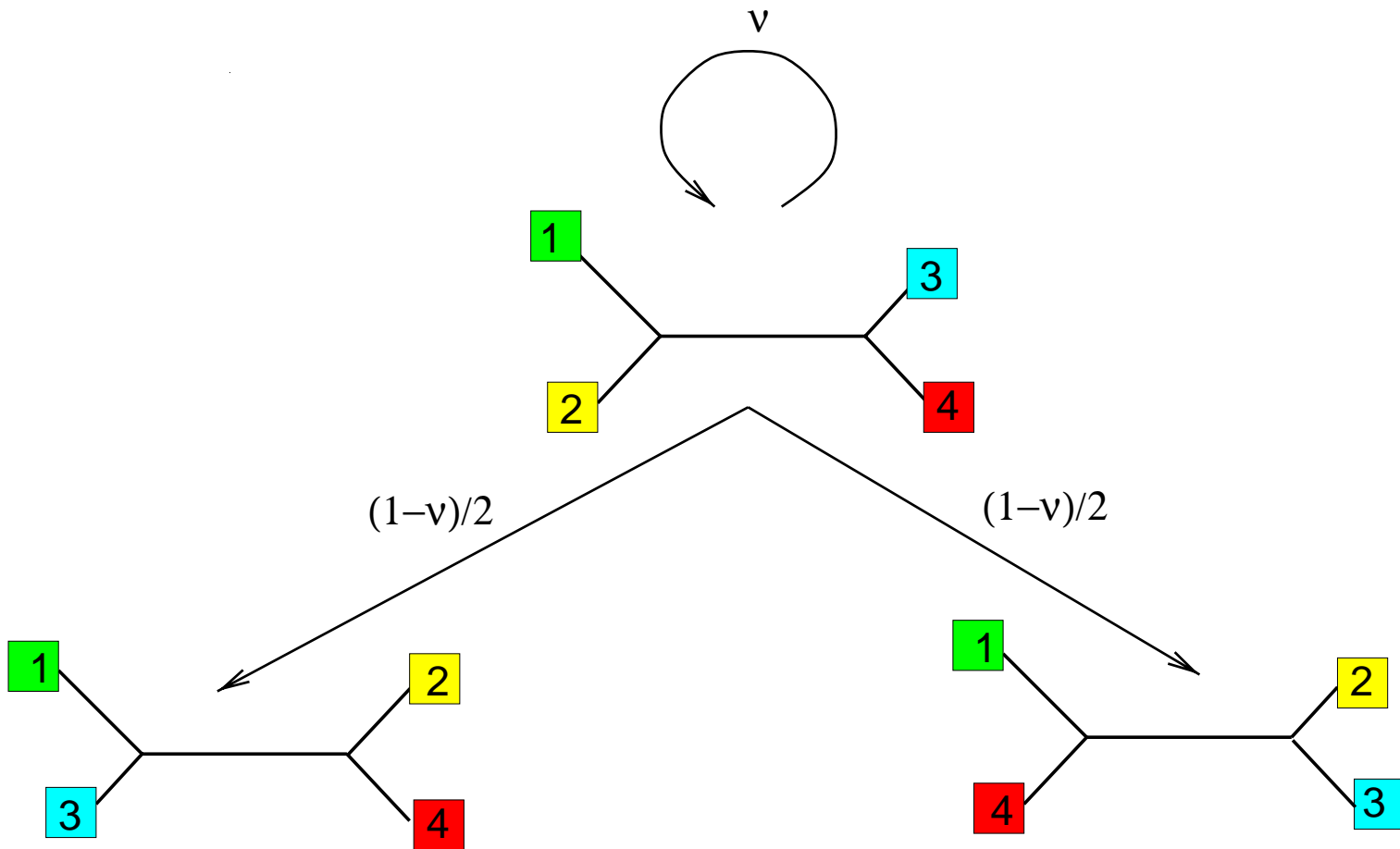
Emission probabilities (vertical arrows)



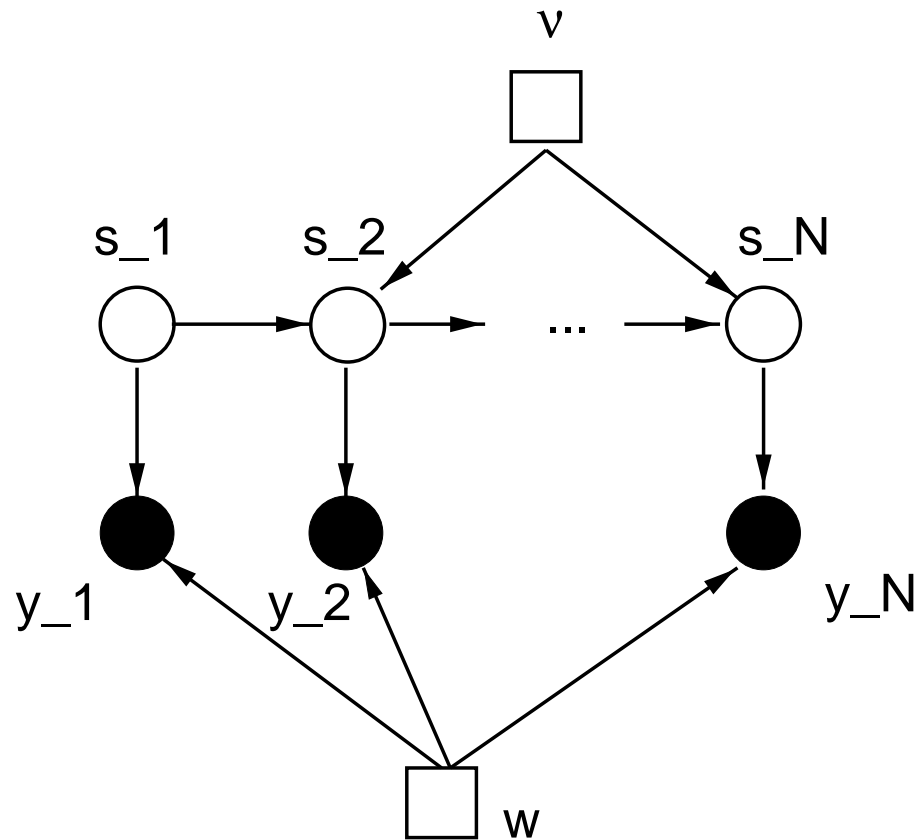
--> $P(y_t | S_t, w)$

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**

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**

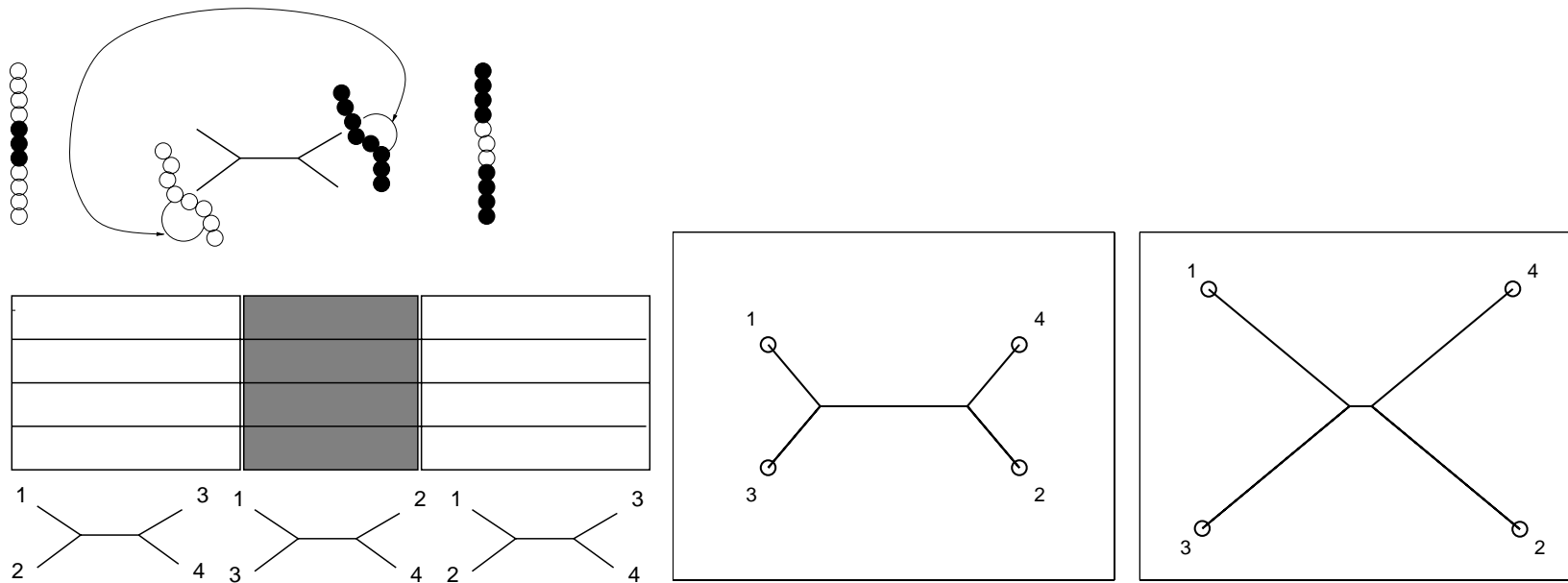
Husmeier, McGuire (2001)
Current work

1) Heuristic method

Optimise the branch lengths w_S for each tree topology S *separately* from the whole alignment.

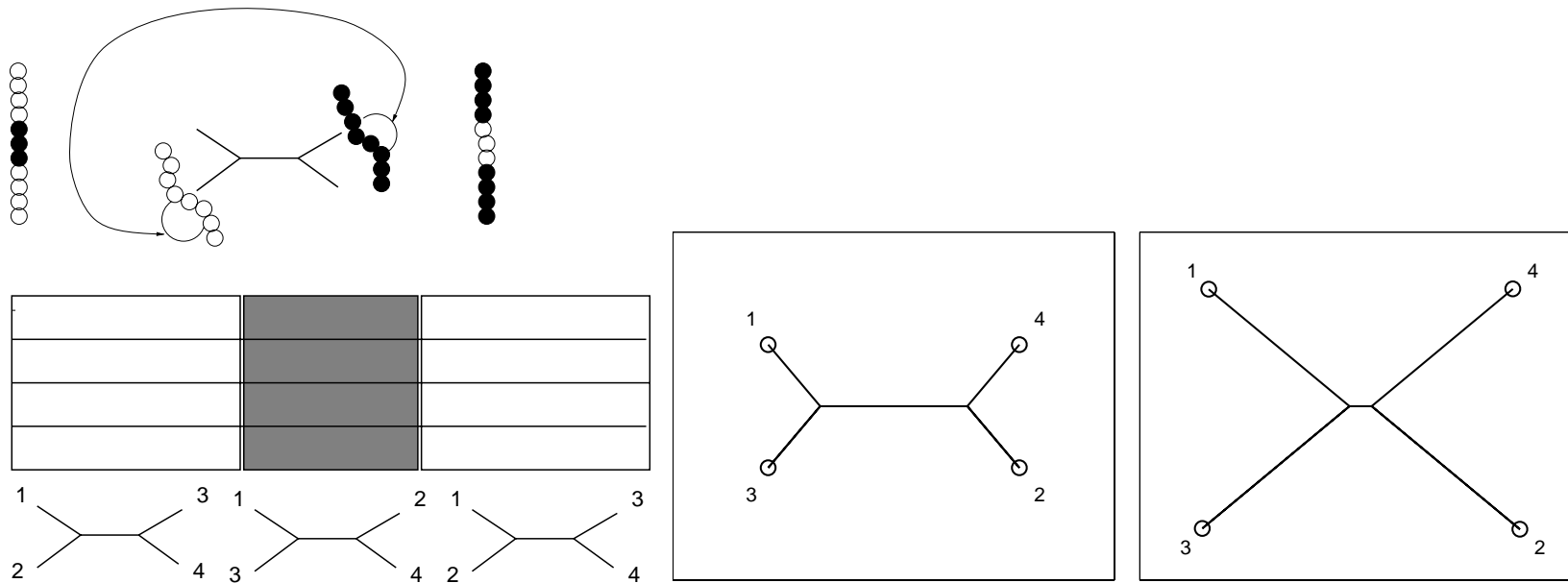
1) Heuristic method

Optimise the branch lengths w_S for each tree topology S separately from the whole alignment.



1) Heuristic method

Optimise the branch lengths w_S for each tree topology S separately from the whole alignment.



No optimisation of ν .

2) Maximum likelihood

Likelihood:

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

2) 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
 $\mathbf{S} = (S_1, S_2, \dots, S_N)$.

2) 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
 $\longrightarrow K^N$ state sequences.

2) 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.

$$\begin{aligned} \ln P(\mathcal{D}|\mathbf{w}, \nu) &= \ln \sum_{\mathbf{S}} P(\mathcal{D}, \mathbf{S}|\mathbf{w}, \nu) = \ln \sum_{\mathbf{S}} \frac{P(\mathcal{D}, \mathbf{S}|\mathbf{w}, \nu)}{Q(\mathbf{S})} Q(\mathbf{S}) \\ &\geq \sum_{\mathbf{S}} Q(\mathbf{S}) \ln \frac{P(\mathcal{D}, \mathbf{S}|\mathbf{w}, \nu)}{Q(\mathbf{S})} \end{aligned}$$

2) Maximum likelihood: EM algorithm

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

2) Maximum likelihood: EM algorithm

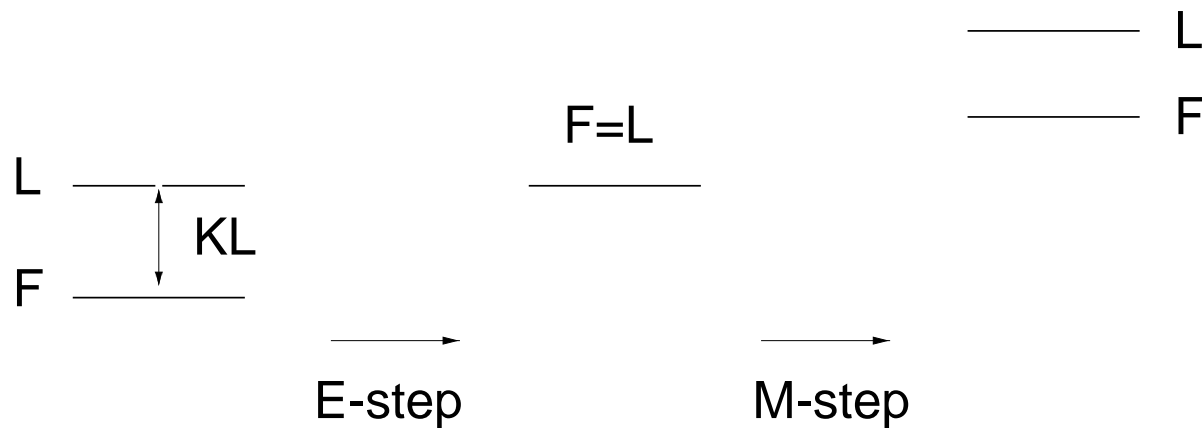
$$F(\mathbf{w}, \nu) = \sum_{\mathbf{S}} Q(\mathbf{S}) \ln \frac{P(\mathcal{D}, \mathbf{S} | \mathbf{w}, \nu)}{Q(\mathbf{S})} = \sum_{\mathbf{S}} Q(\mathbf{S}) \ln \frac{P(\mathbf{S} | \mathcal{D}, \mathbf{w}, \nu)}{Q(\mathbf{S})} + \ln P(\mathcal{D} | \mathbf{w}, \nu)$$

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

2) Maximum likelihood: EM algorithm

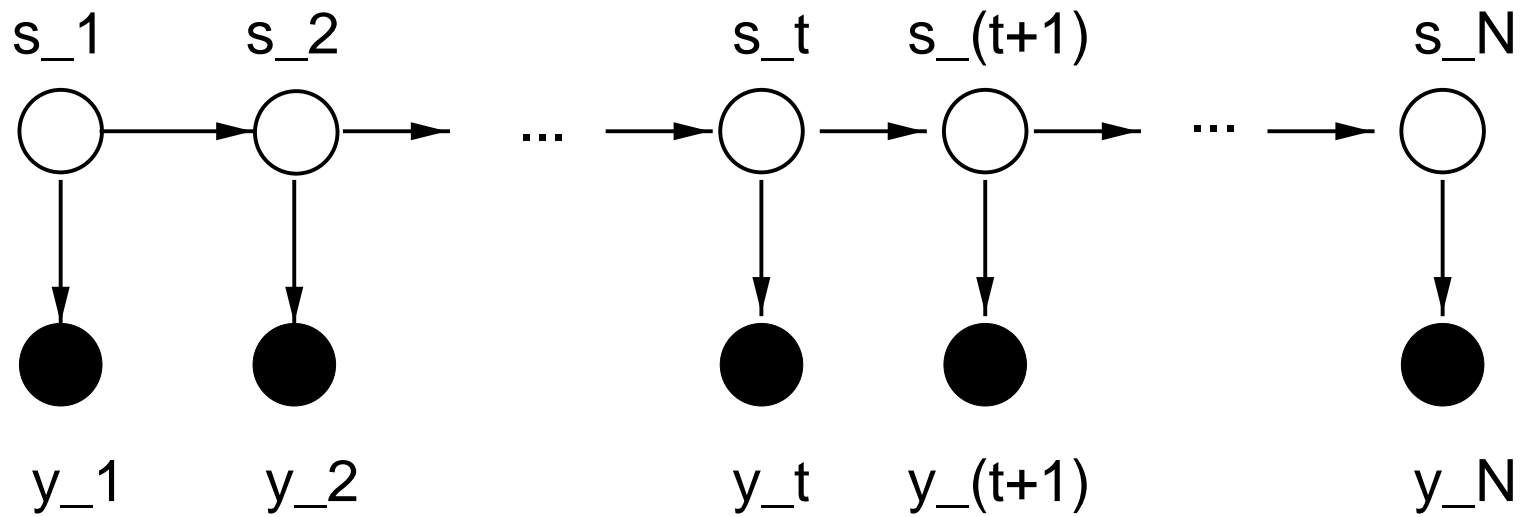
$$F(\mathbf{w}, \nu) = \sum_{\mathbf{S}} Q(\mathbf{S}) \ln \frac{P(\mathcal{D}, \mathbf{S} | \mathbf{w}, \nu)}{Q(\mathbf{S})} = \sum_{\mathbf{S}} Q(\mathbf{S}) \ln \frac{P(\mathbf{S} | \mathcal{D}, \mathbf{w}, \nu)}{Q(\mathbf{S})} + \ln P(\mathcal{D} | \mathbf{w}, \nu)$$

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



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)

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

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 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})

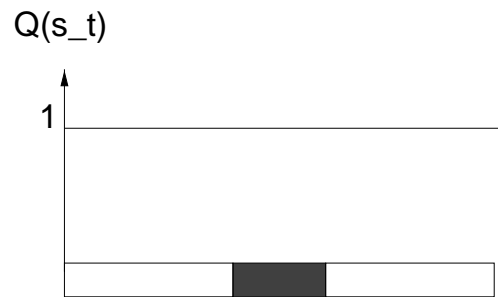
$$P(\mathcal{D}, \mathbf{S} | \mathbf{w}, \nu) = \prod_{t=1}^N P(\mathbf{y}_t | S_t, \mathbf{w}) \prod_{t=2}^N P(S_t | S_{t-1}, \nu) P(S_1)$$

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

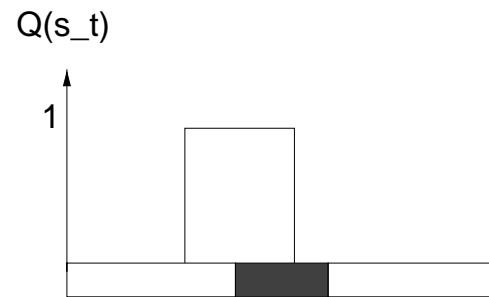
$$F(\mathbf{w}, \nu) = \sum_{\mathbf{S}} Q(\mathbf{S}) \sum_{t=1}^N \ln P(\mathbf{y}_t | S_t, \mathbf{w}) + C$$

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

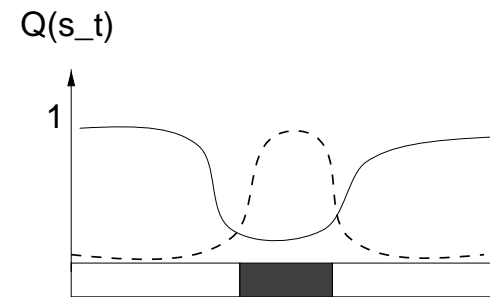
EM weighting scheme



Standard

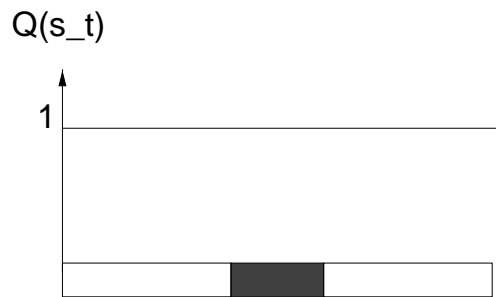


Window

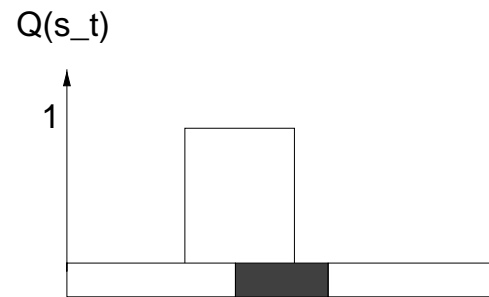


EM

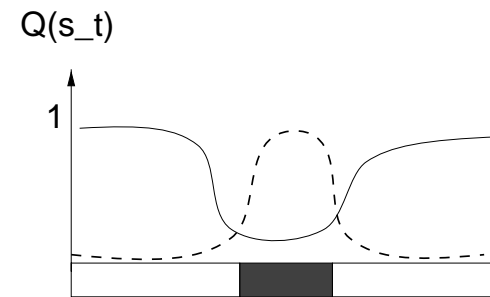
EM weighting scheme



Standard



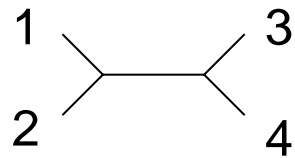
Window



EM

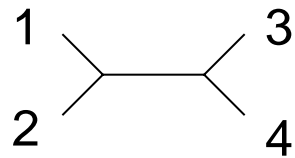
$Q(S_t) \longrightarrow Q(S_t | \mathcal{D}, \mathbf{w}, \nu) \longrightarrow$ Forward-backward algorithm

Synthetic example



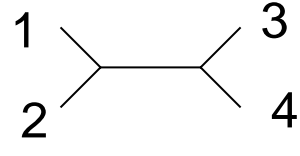
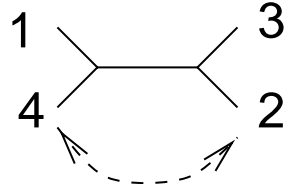
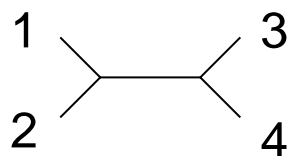
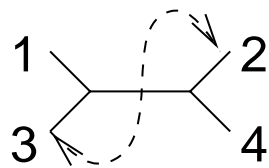
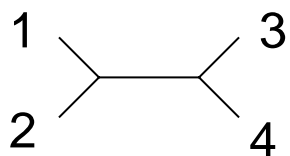
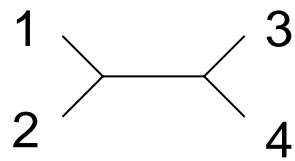
- Model of nucleotide substitution: Kimura 2-parameter, $\tau = 2$.
- Alignment of length $N = 1000$ nucleotides.

Synthetic example

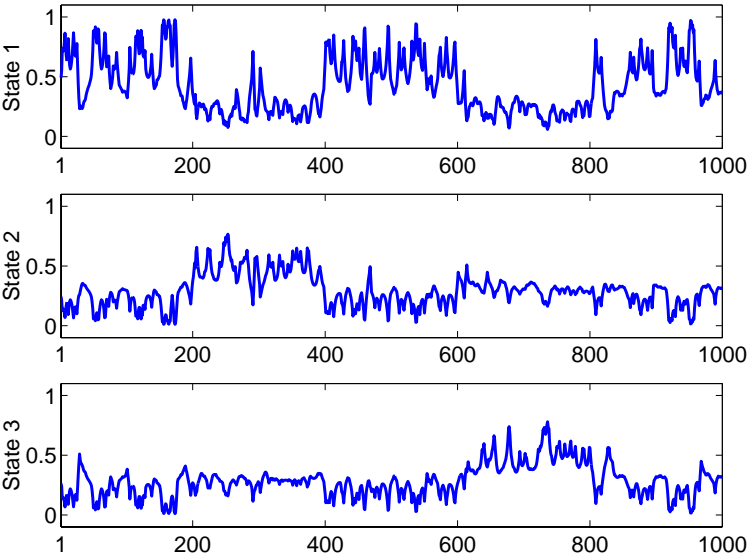
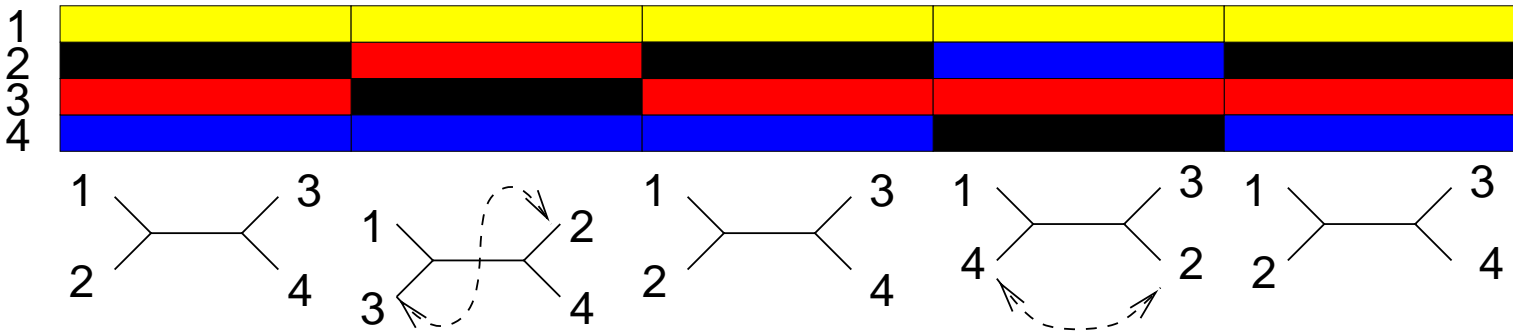


- Model of nucleotide substitution: Kimura 2-parameter, $\tau = 2$.
- Alignment of length $N = 1000$ nucleotides.
- Two recombinant regions of length 200 nucleotides.

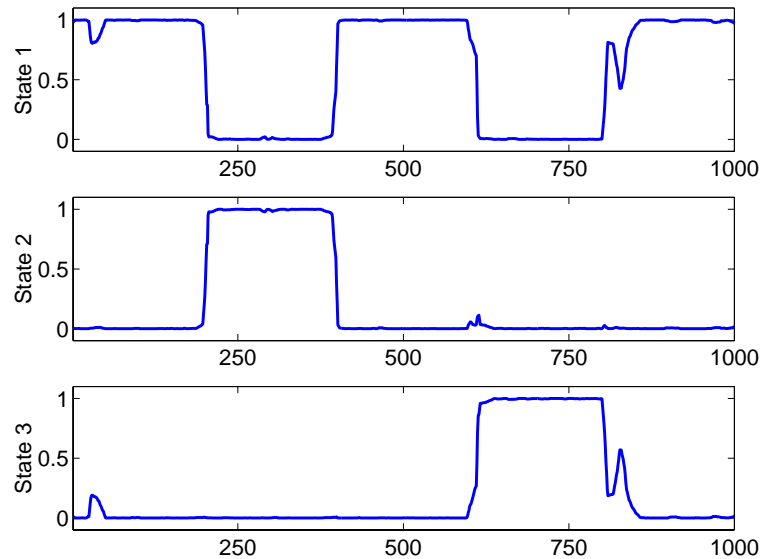
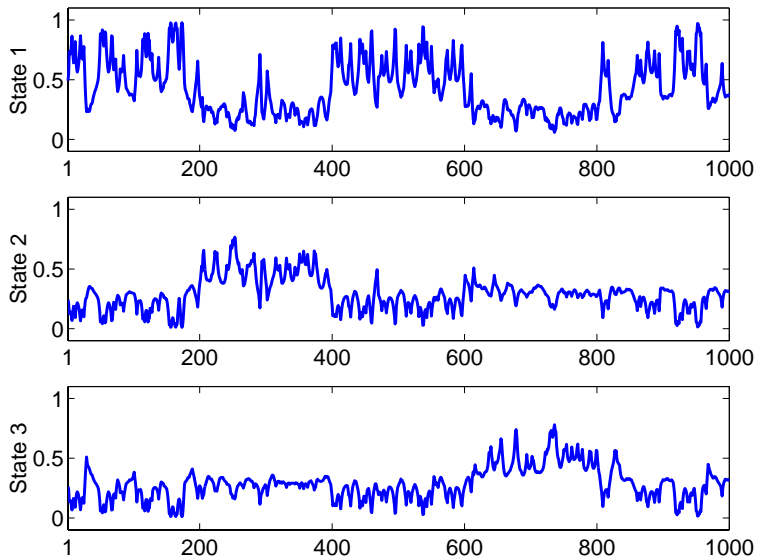
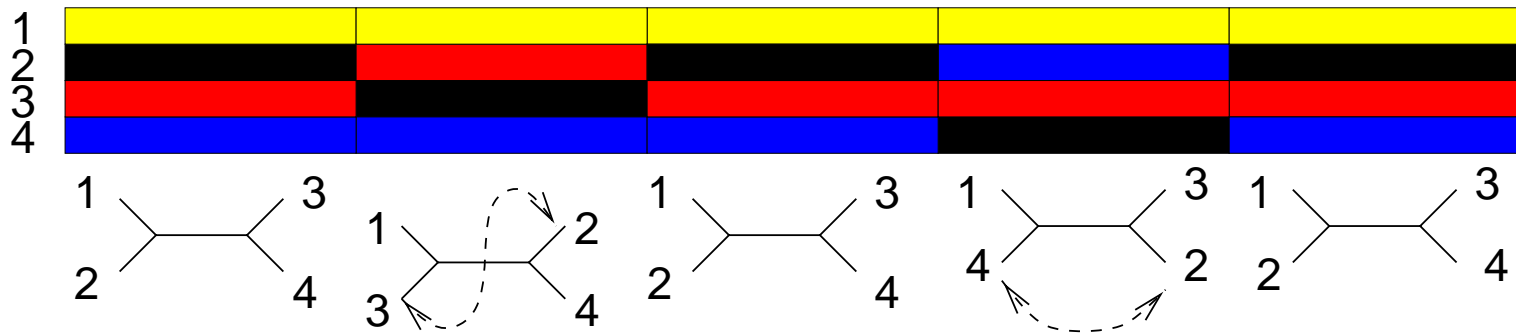
Synthetic example



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



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



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

Actin genes : DNA alignment of 1008 nucleotides

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

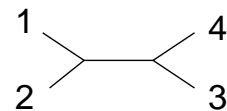
Gene conversion in maize (Moniz de Sa, Drouin, 1996)

Actin genes : DNA alignment of 1008 nucleotides

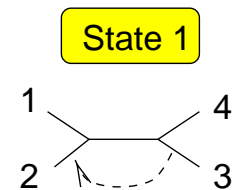
- 1) Maz56
- 2) Maz63
- 3) Maz63
- 4) Maz89

875 bases

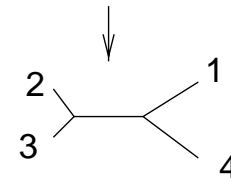
133 bases



State 1



State 1

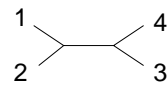


State 3

$P(S_t|\mathcal{D})$, heuristic method, $\nu = 0.95$

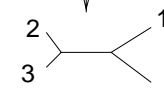
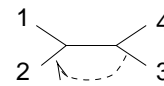
875 bases

133 bases

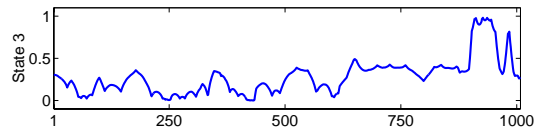
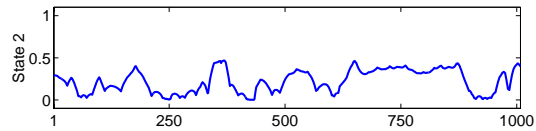
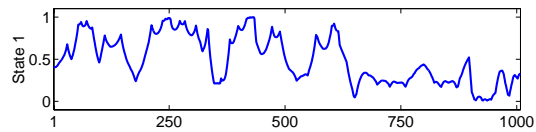


State 1

State 1



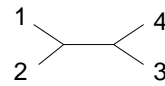
State 3



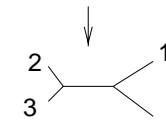
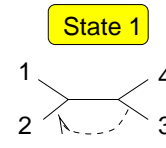
$P(S_t|\mathcal{D})$, heuristic method, $\nu = 0.997$

875 bases

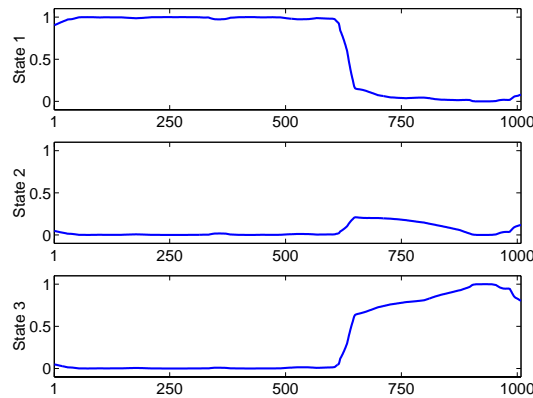
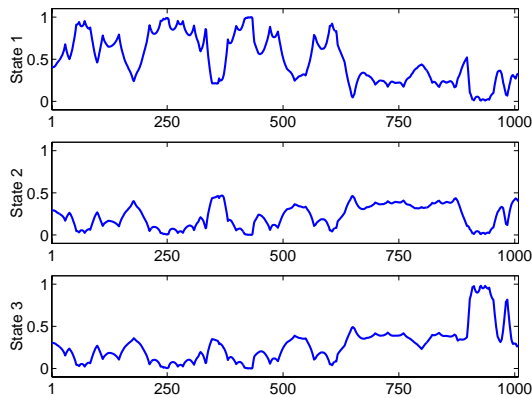
133 bases



State 1



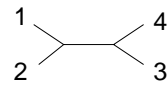
State 3



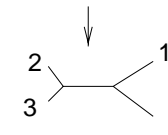
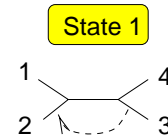
$P(S_t|\mathcal{D})$, maximum likelihood

875 bases

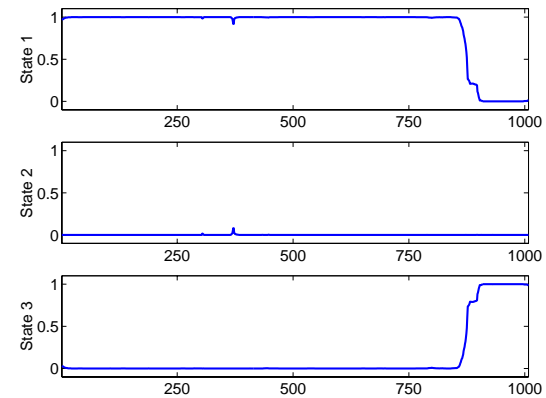
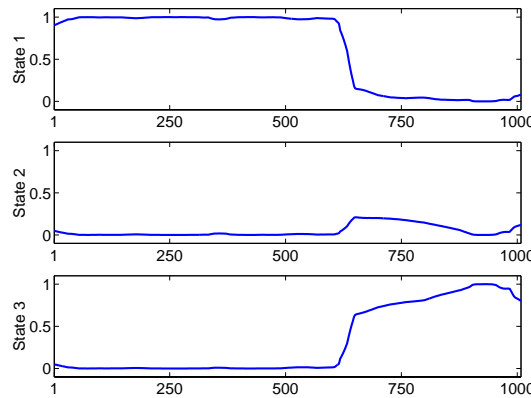
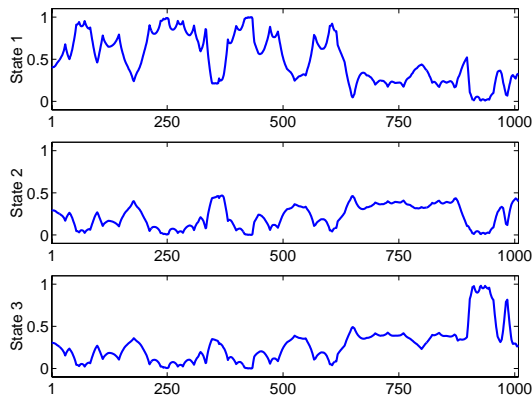
133 bases



State 1



State 3



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 .

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** $P(\mathbf{w}, \nu) = \prod_i P(w_i)P(\nu)$

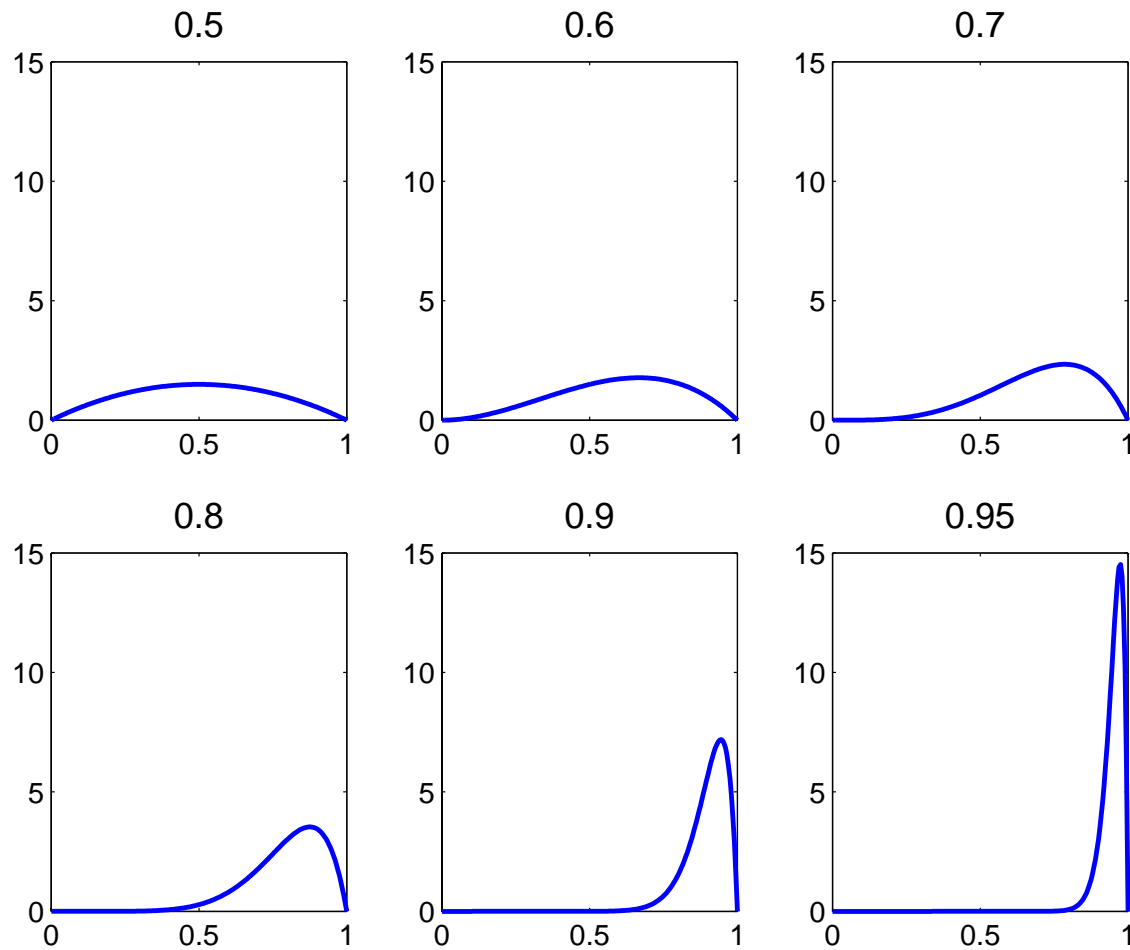
Bayesian approach

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

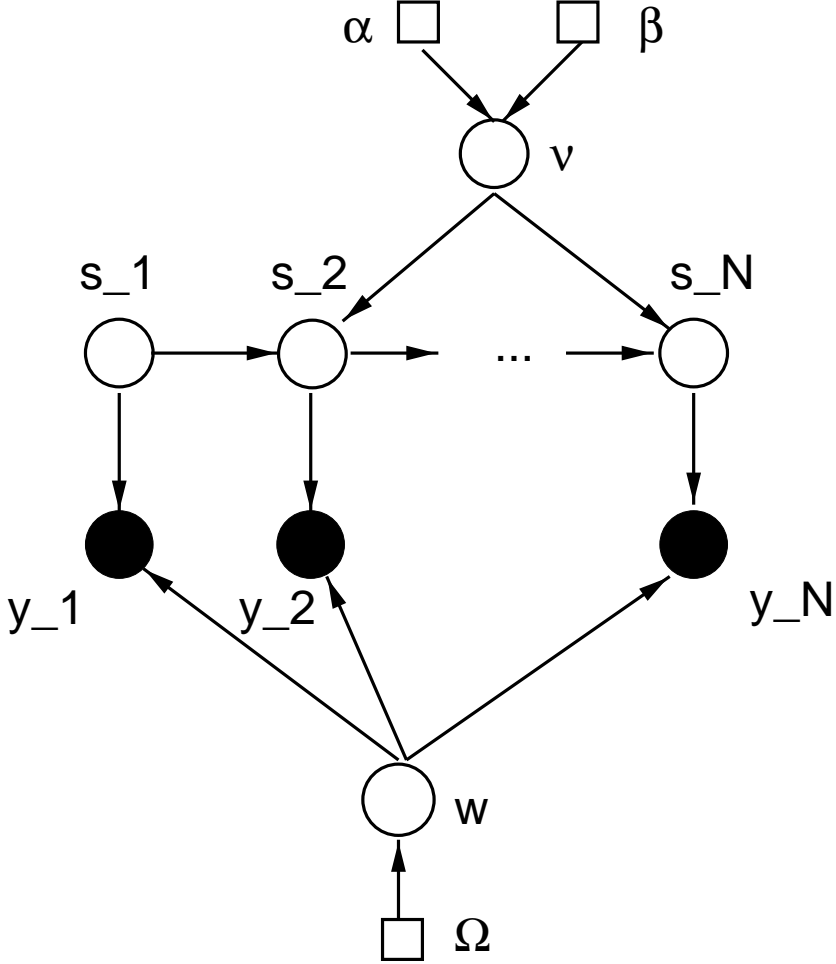
Bayesian approach

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

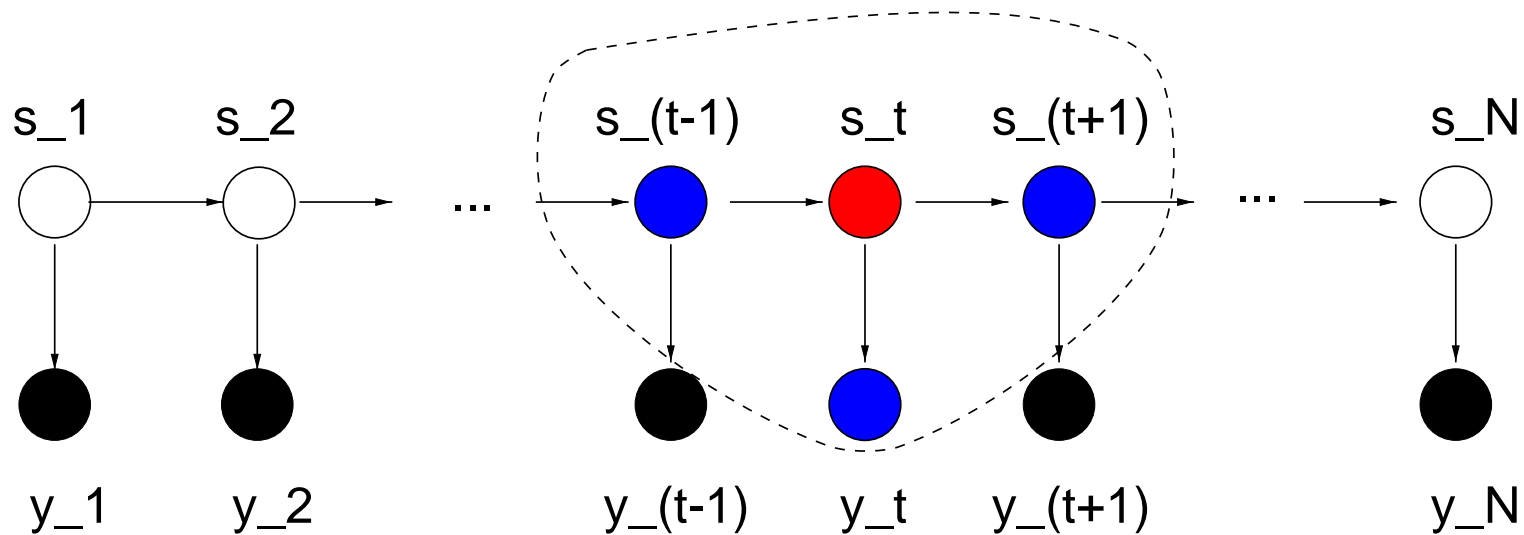
Sampling from the posterior distribution

- 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

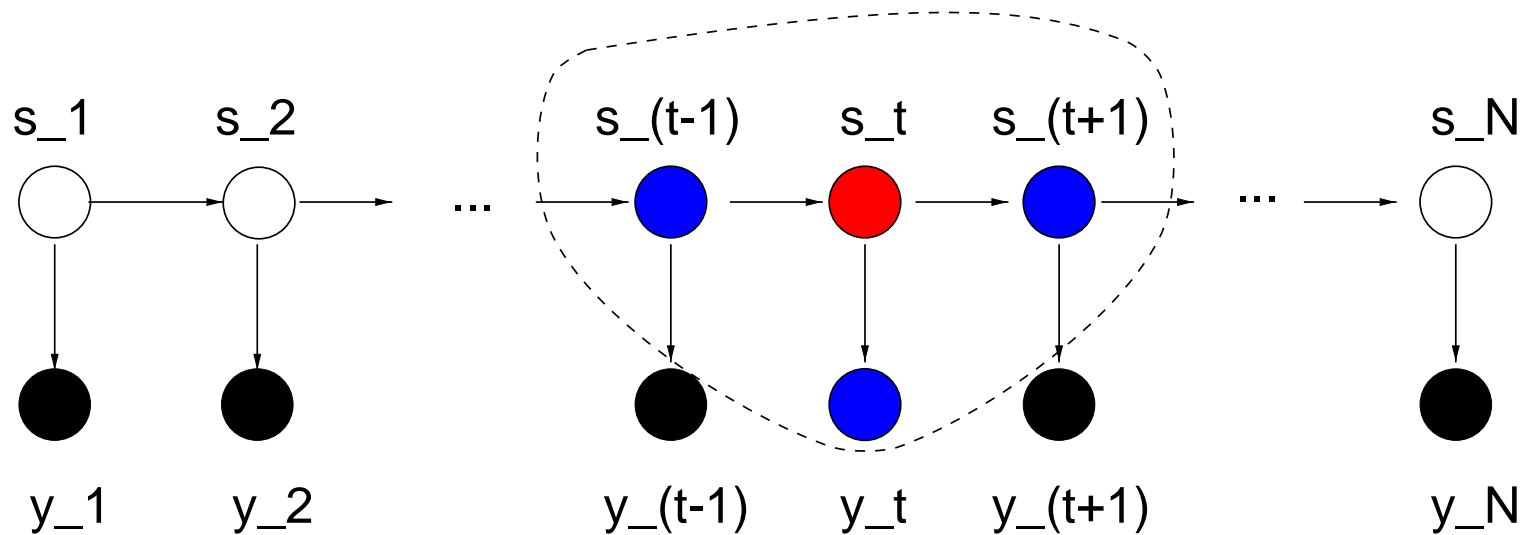
- Sampling from $P(\mathbf{S}, \mathbf{w}, \nu | \mathcal{D})$
 - Gibbs-like approach:
 - $\mathbf{S} \sim P(\mathbf{S} | \mathbf{w}, \nu, \mathcal{D})$
 - $\mathbf{w} \sim P(\mathbf{w} | \mathbf{S}, \nu, \mathcal{D})$
 - $\nu \sim P(\nu | \mathbf{S}, \mathbf{w}, \mathcal{D})$
 - ν : Sample from Beta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



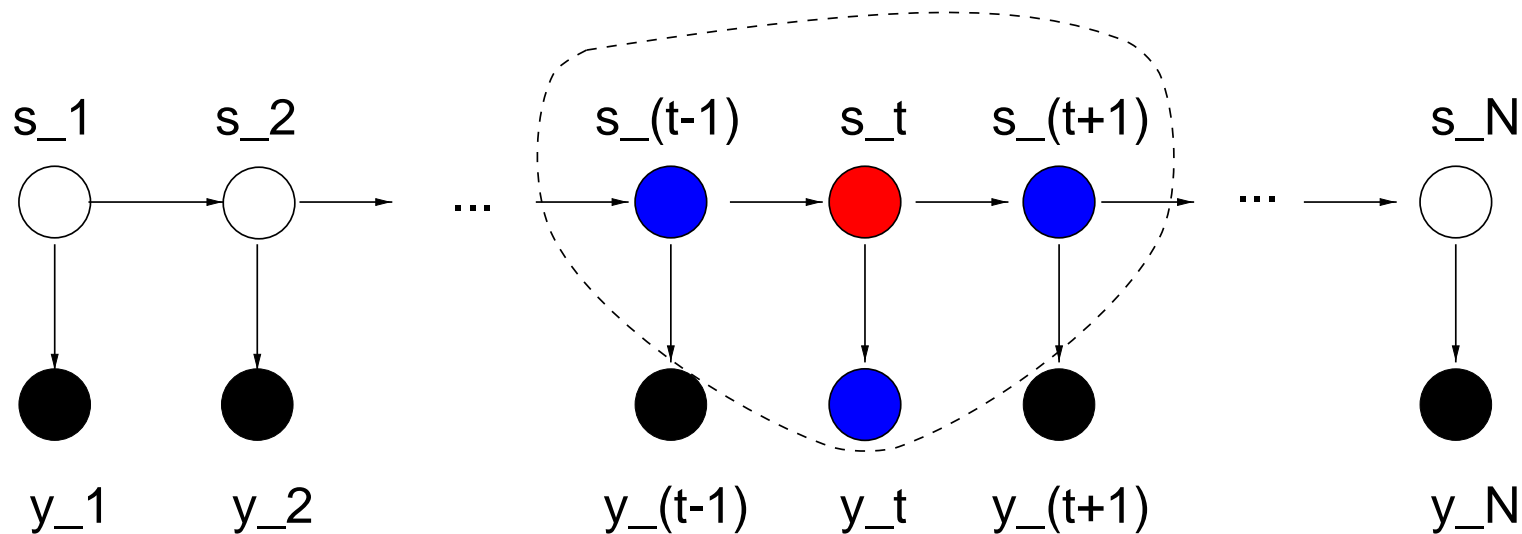
$$P(S_t | S_1, \dots, S_{t-1}, S_{t+1}, \dots, S_N, \mathcal{D}, \mathbf{w}, \nu)$$

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



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

Neisseria (Zhou & Spratt, 1992)

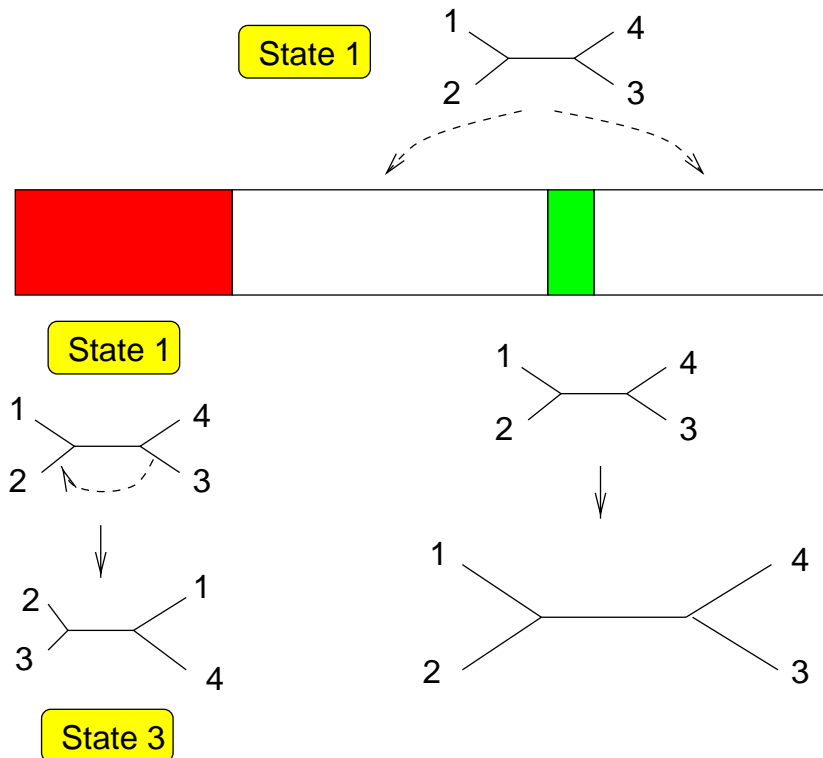
DNA alignment, 787 nucleotides (argF gene)

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

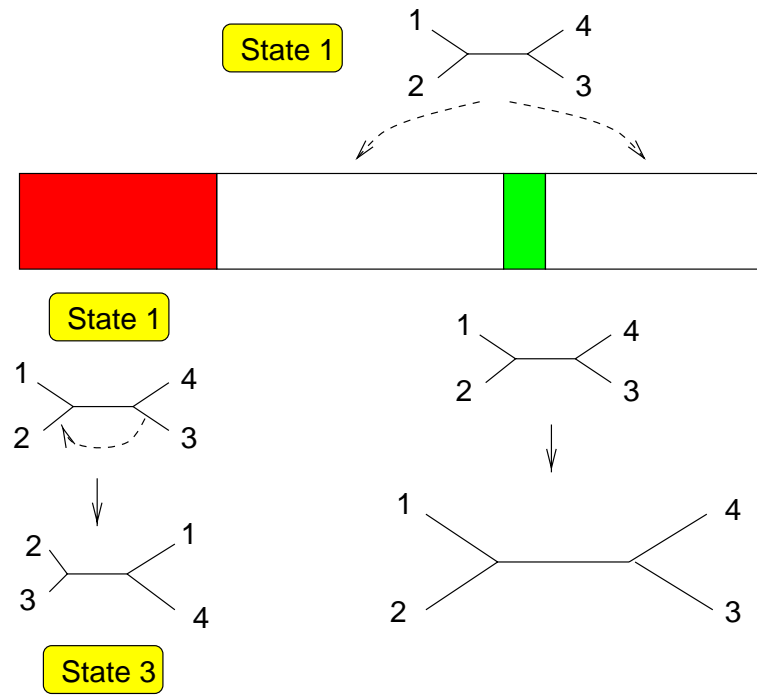
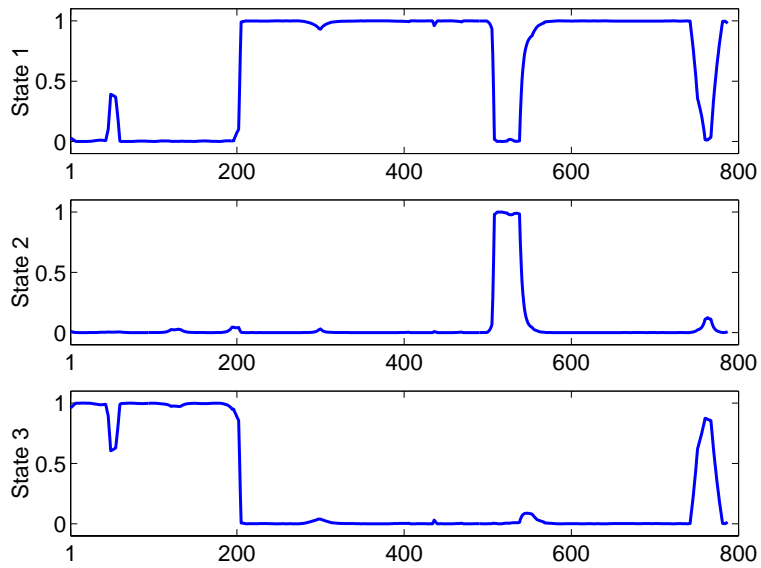
Neisseria (Zhou & Spratt, 1992)

DNA alignment, 787 nucleotides (argF gene)

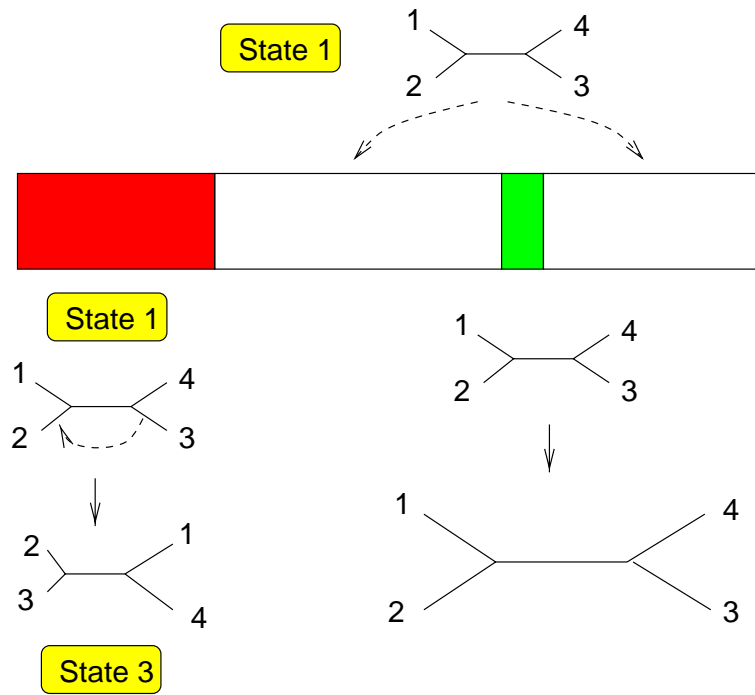
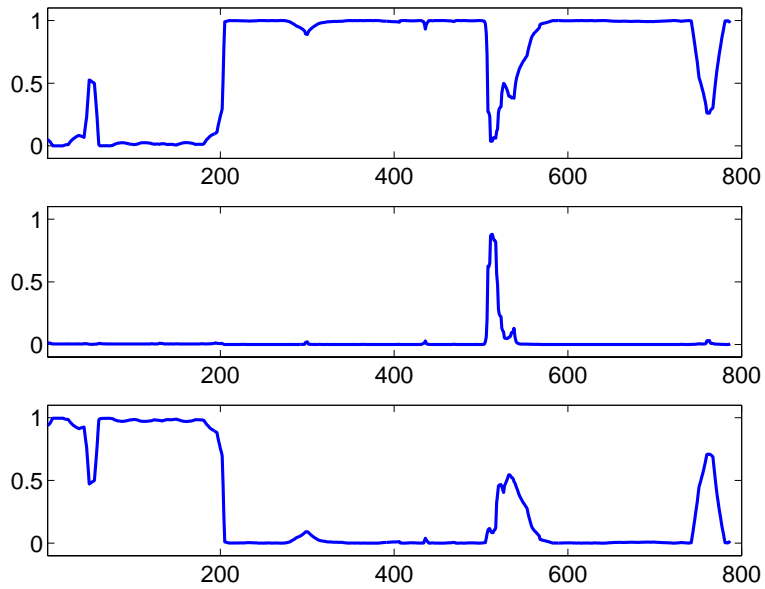
- 1) Neisseria *gonorrhoeae*
- 2) Neisseria *meningitidis*
- 3) Neisseria *cinerea*
- 4) Neisseria *mucosa*



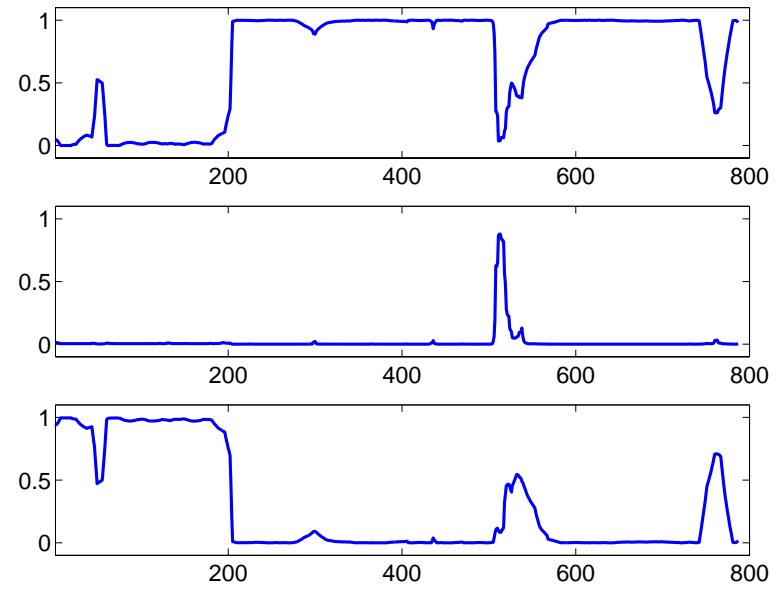
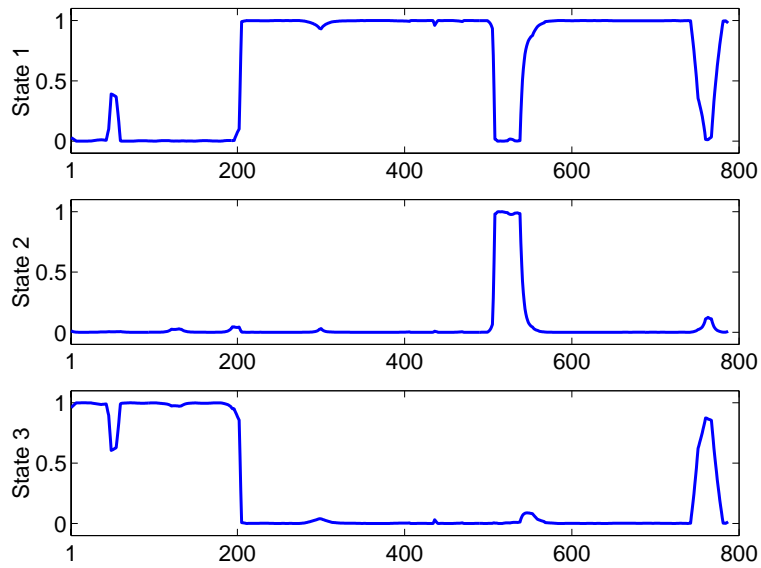
Predition of $P(S_t|\mathcal{D})$ with ML



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



Prediction of $P(S_t|\mathcal{D})$: Comparison between ML and Bayes



Hepatitis B Virus (Bollyky et al. 1995)

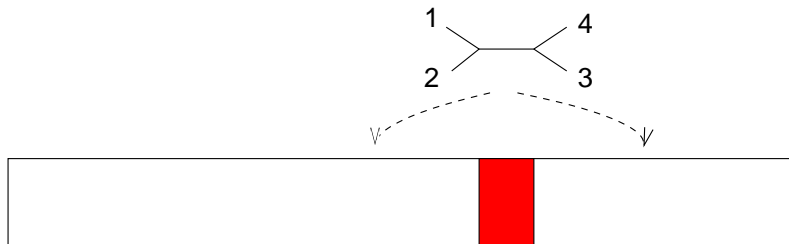
DNA alignment, 3049 nucleotides

1) HPBADW1

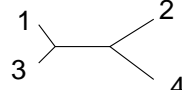
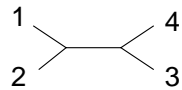
2) HPBADW2

3) HPBADWZCG

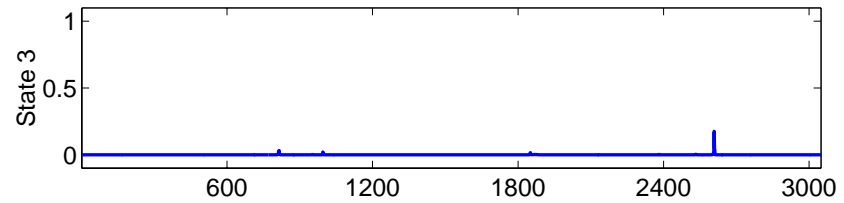
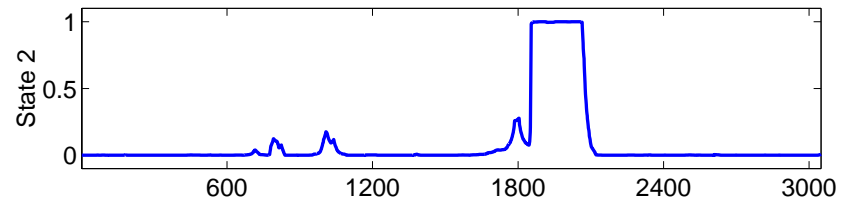
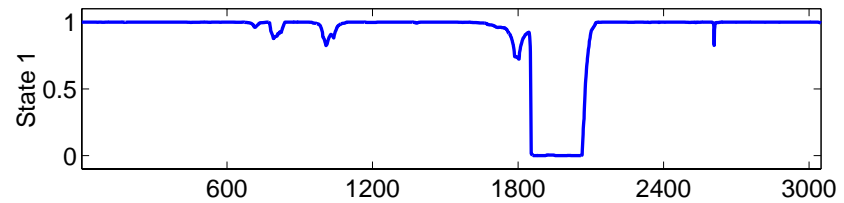
4) HPBADRC



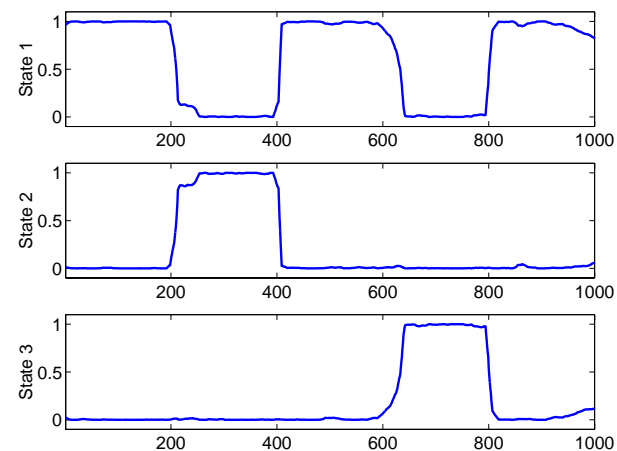
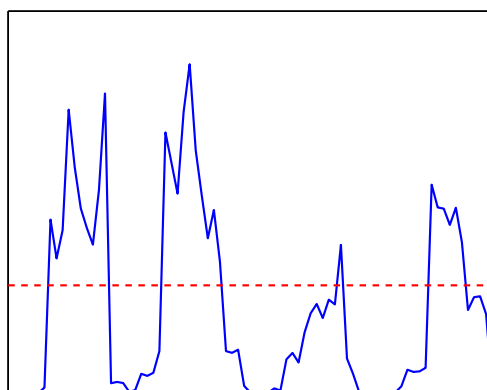
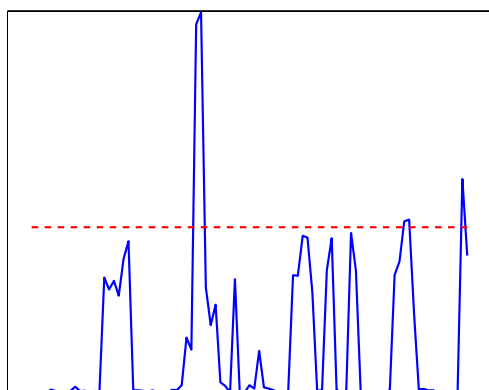
State 1



State 2

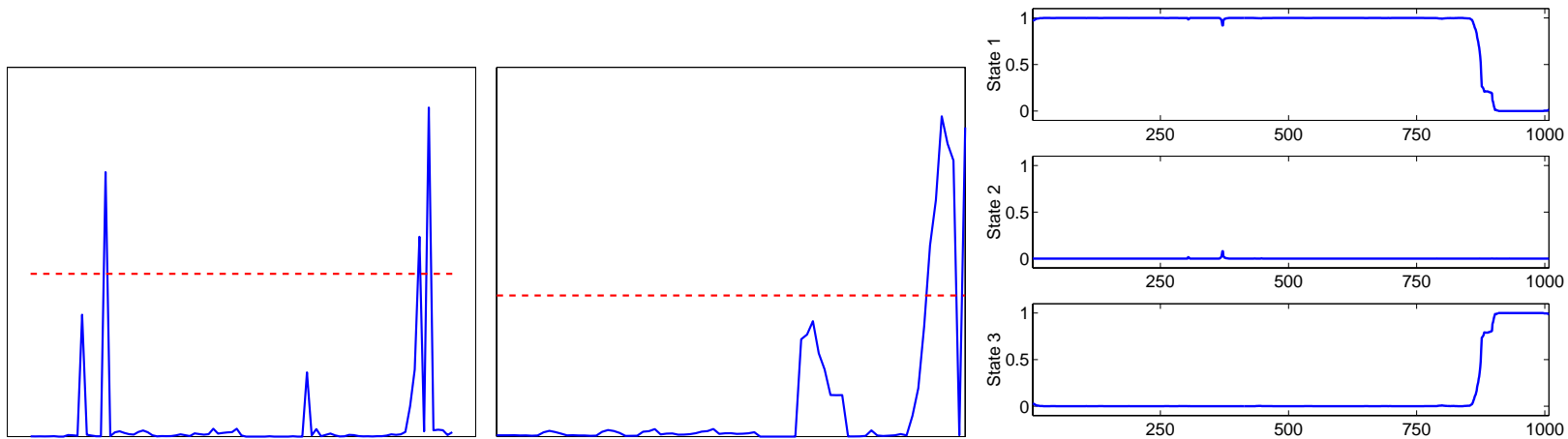


Comparison with TOPAL: Synthetic alignment



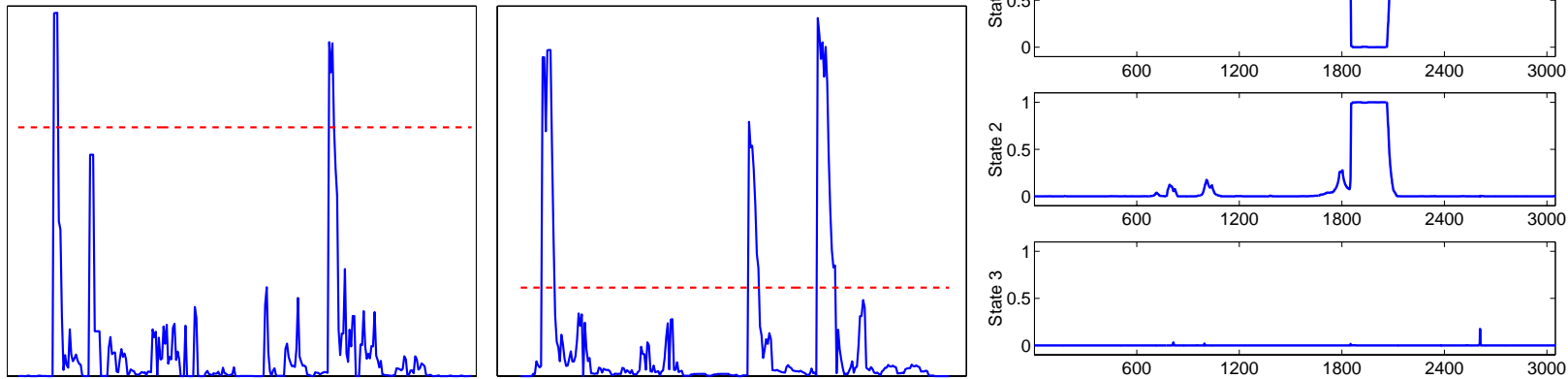
Left: TOPAL, window size=100
Middle: TOPAL, window size=200
Right: Bayesian HMM

Comparison with TOPAL: Maize



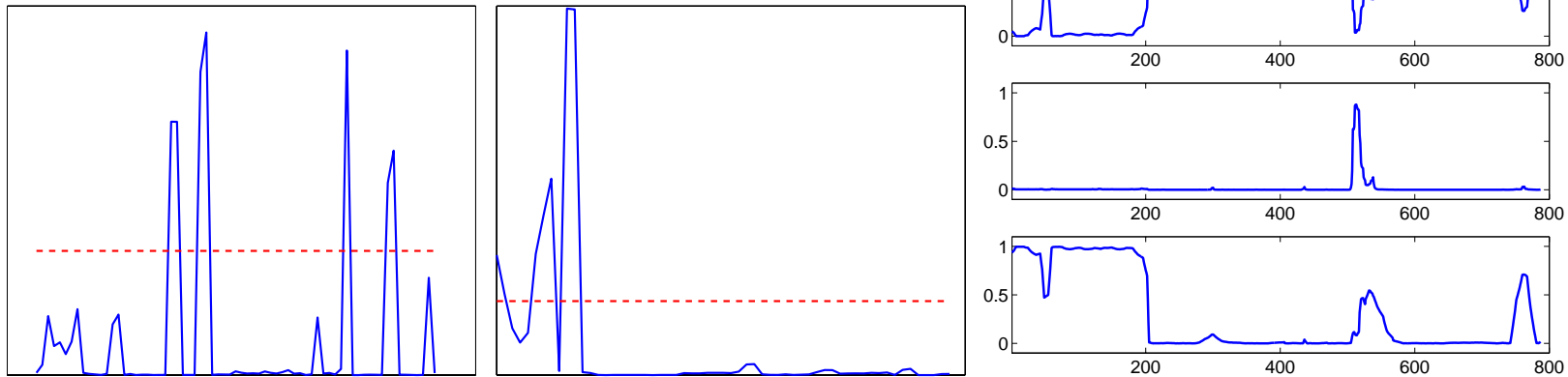
Left: TOPAL, window size=100
Middle: TOPAL, window size=200
Right: Bayesian HMM

Comparison with TOPAL: Hepatitis B



Left: TOPAL, window size=100
Middle: TOPAL, window size=200
Right: Bayesian HMM

Comparison with TOPAL: Neisseria



Left: TOPAL, window size=100

Middle: TOPAL, window size=200

Right: Bayesian HMM

Conclusion and future work

Conclusion and future work

- No window.

Conclusion and future work

- No window.
- Precise location of breakpoints.

Conclusion and future work

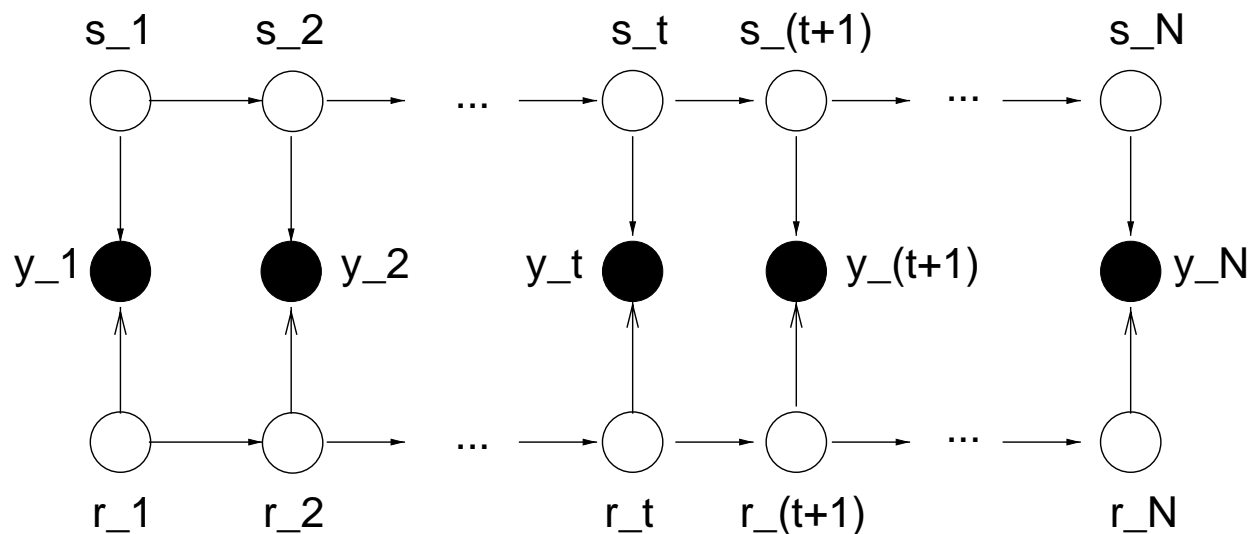
- No window.
- Precise location of breakpoints.
- Limited in the number of different tree topologies.

Conclusion and future work

- No window.
- Precise location of breakpoints.
- Limited in the number of different tree topologies.
- Problem: rate heterogeneity.

Conclusion and future work

- No window.
- Precise location of breakpoints.
- Limited in the number of different tree topologies.
- **Problem: rate heterogeneity.**
- Future work: factorial HMM



Acknowledgements

Collaboration

Frank Wright
Gráinne McGuire

Funding

BBSRC
SEERAD
