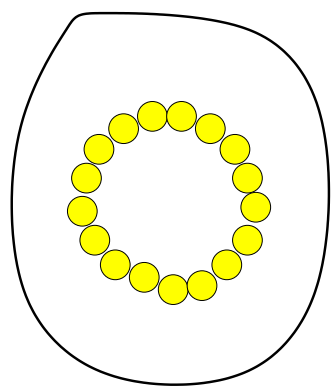

Detecting Recombination in DNA Sequence Alignments

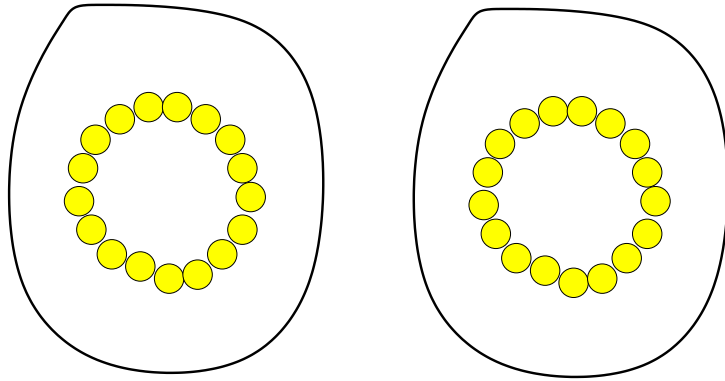
Dirk Husmeier

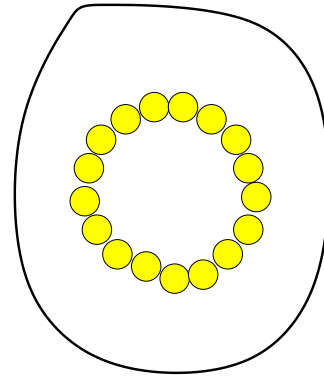
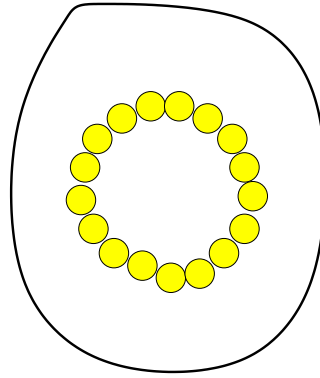
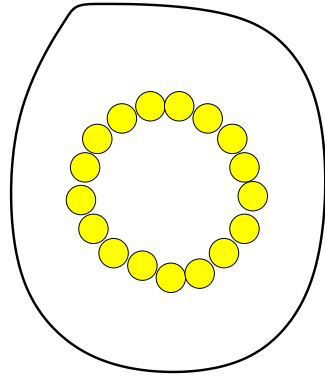
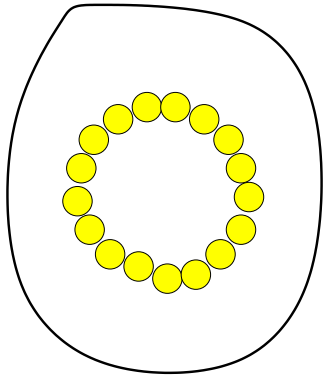
Biomathematics and Statistics Scotland
SCRI, Invergowrie, Dundee DD2 5DA, UK

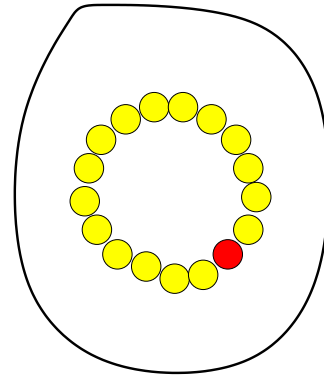
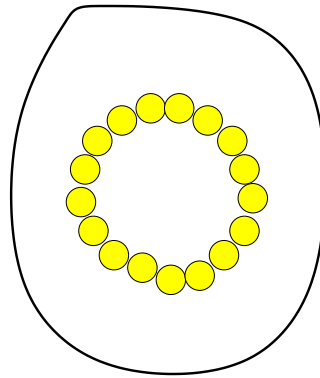
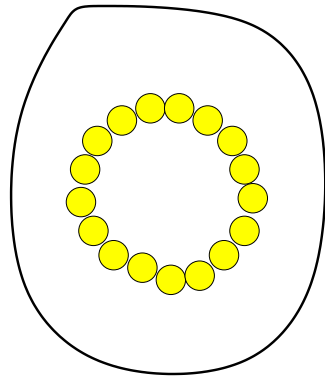
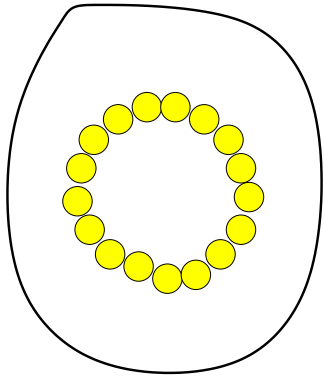
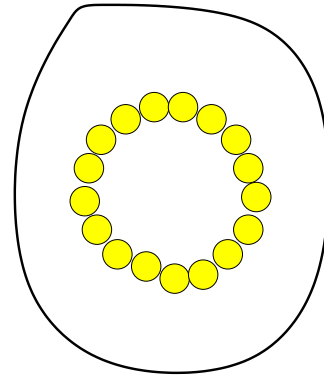
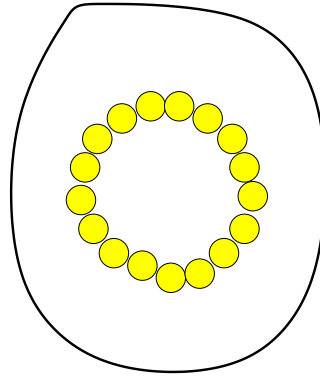
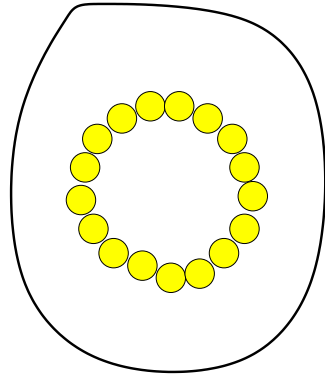
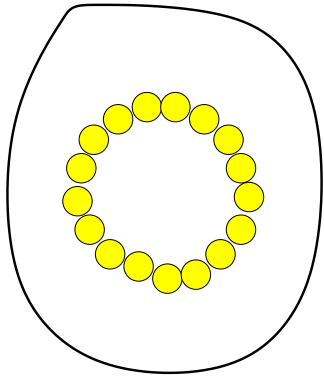
Email: dirk@bioinformatics.ac.uk

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

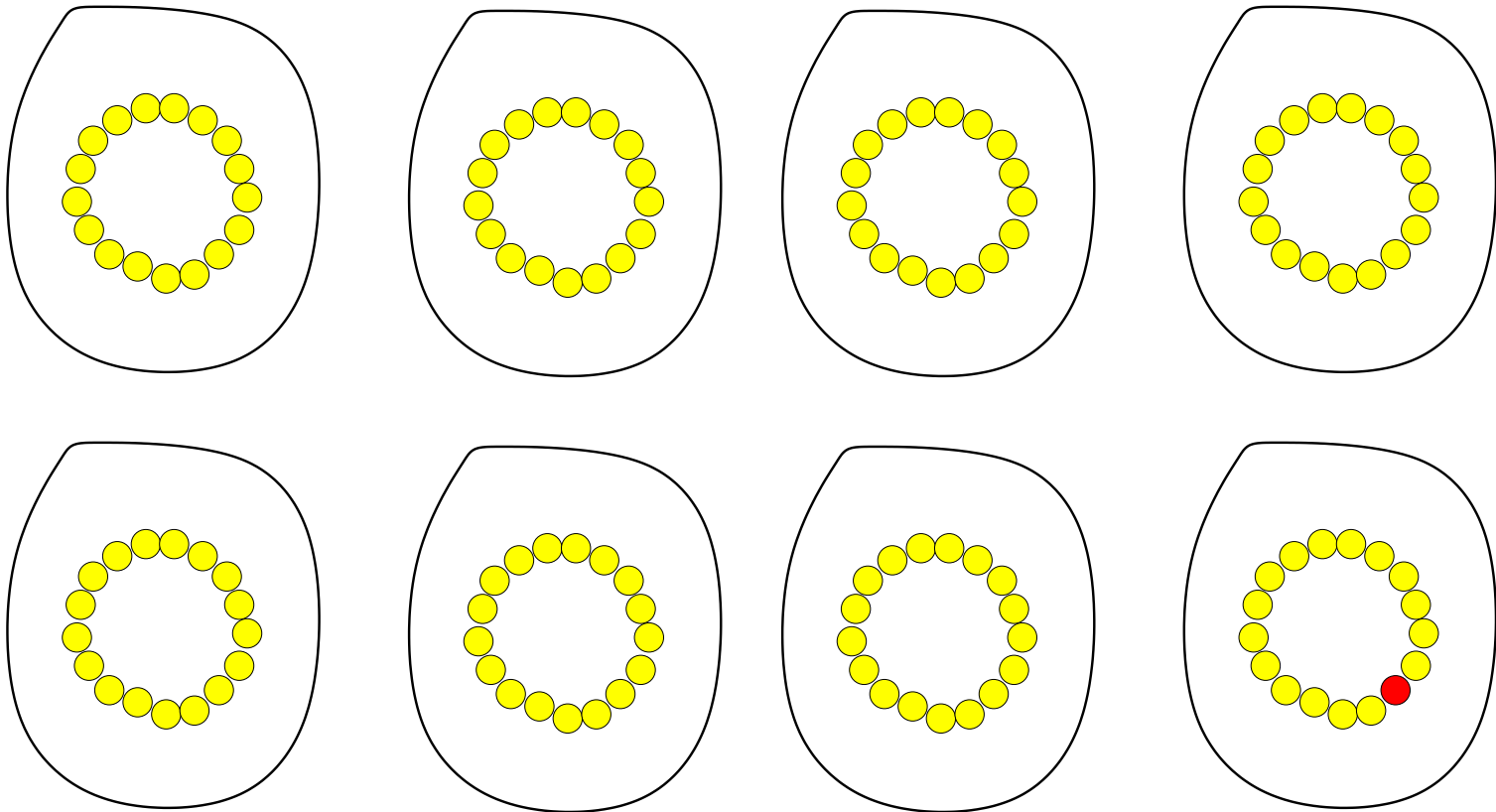


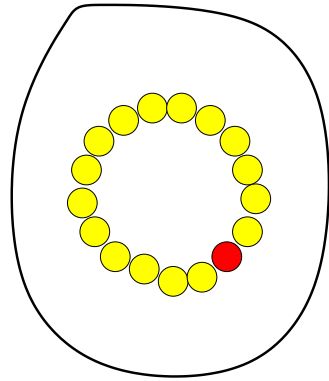






Apply antibiotics





Zhou, Spratt, 1992: *Neisseria*

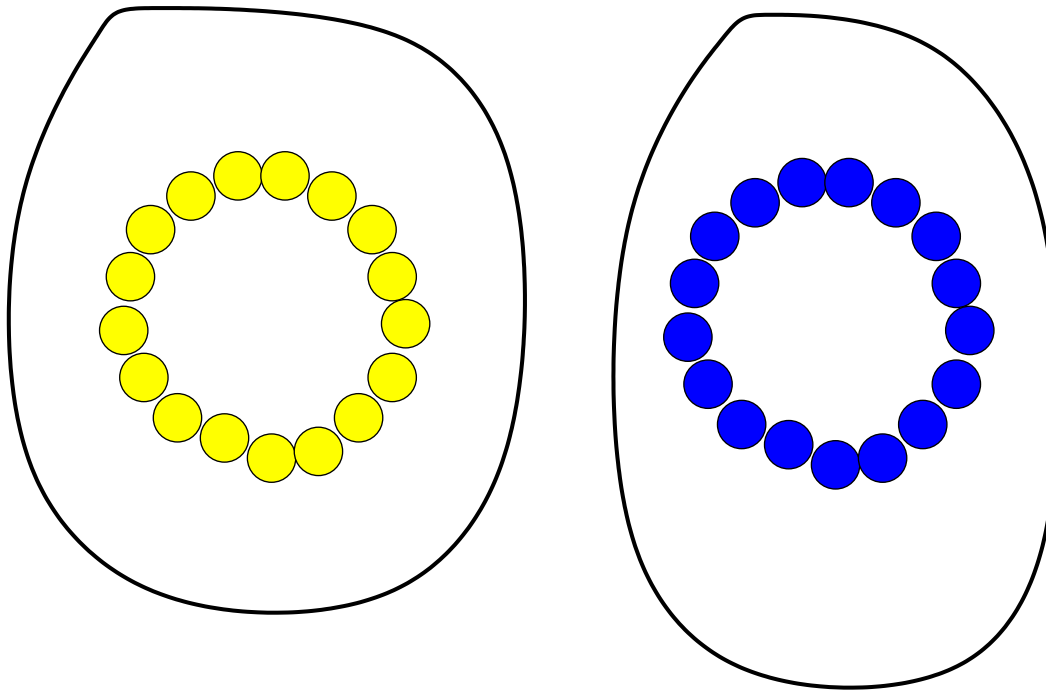
Zhou, Spratt, 1992: *Neisseria*

- *Neisseria.meningitidis*: Pathogenic, susceptible to penicillin

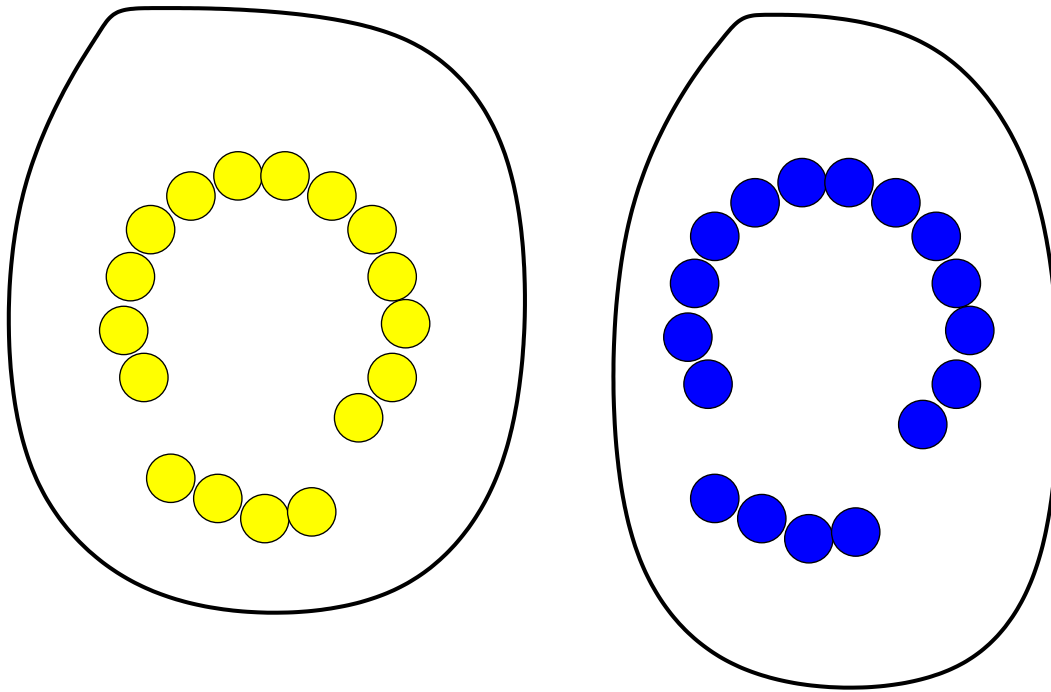
Zhou, Spratt, 1992: *Neisseria*

- *Neisseria.meningitidis*: Pathogenic, susceptible to penicillin
- *Neisseria.cinera*: Benign, resistant to penicillin

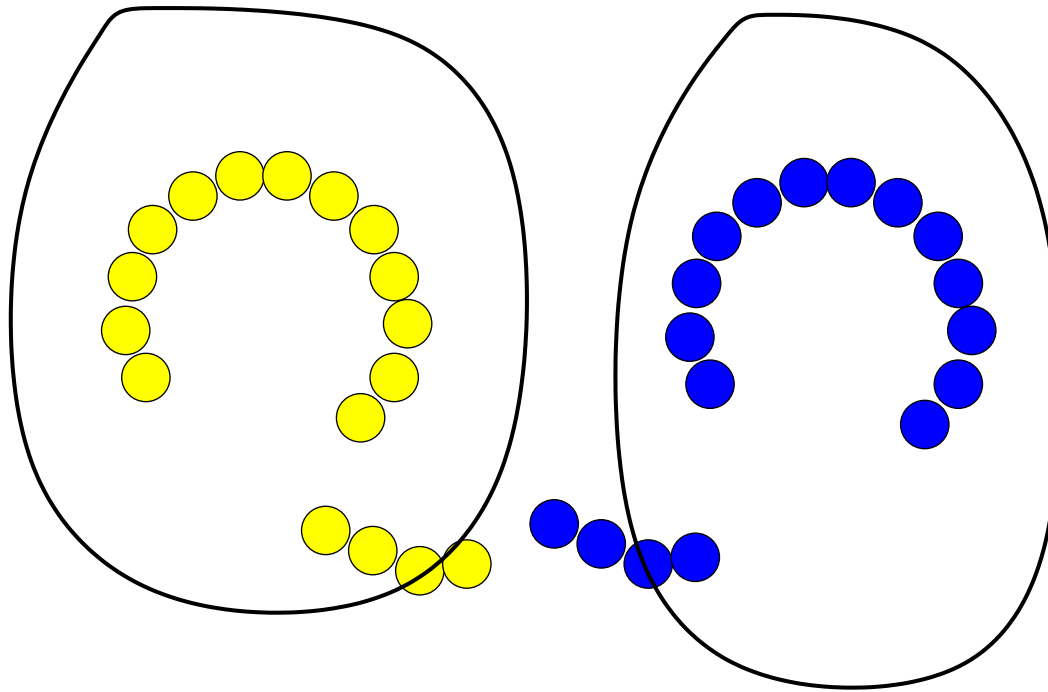
Recombination



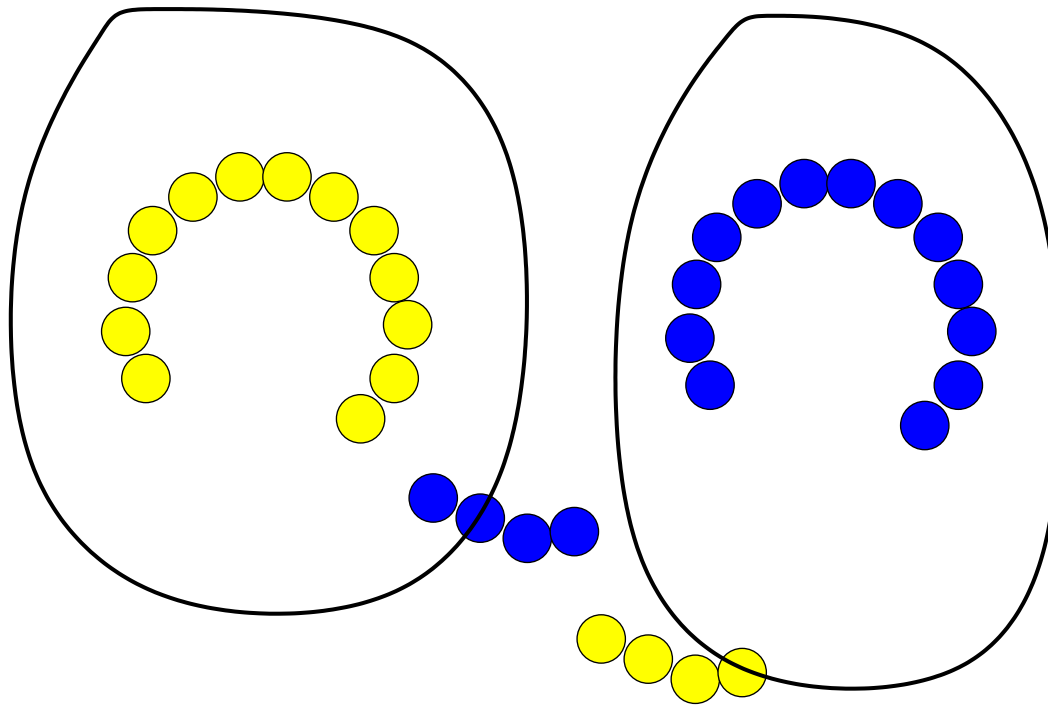
Recombination



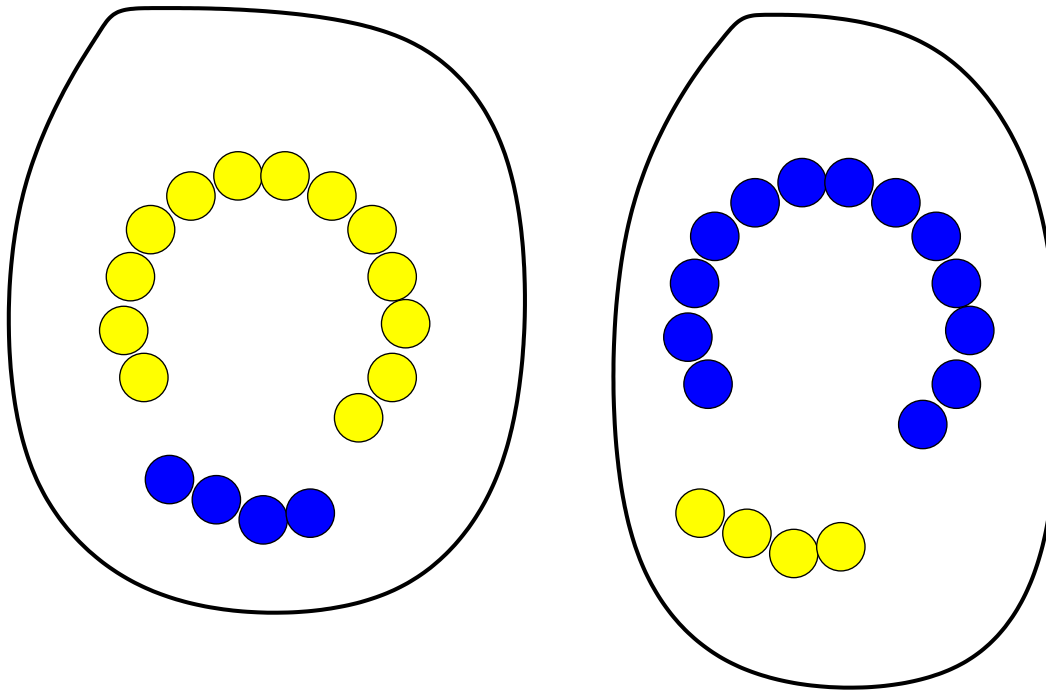
Recombination



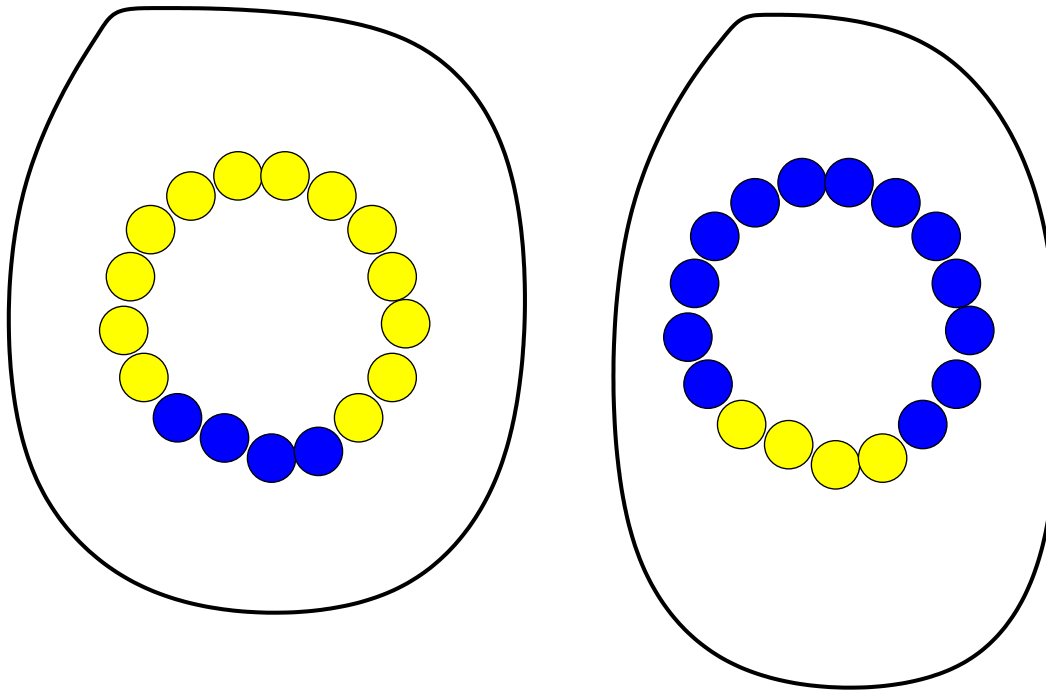
Recombination



Recombination



Recombination



1995

Robertson, Sharp, McCutchan, Hahn

Recombination in HIV-1

Nature 374, pp.124-126

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?

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>

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.bioinformatics.ac.uk/~dirk>

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.bioinformatics.ac.uk/~dirk>

- Phylogenetics

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.bioinformatics.ac.uk/~dirk>

- Phylogenetics
- Hidden Markov models

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.bioinformatics.ac.uk/~dirk>

- Phylogenetics
- Hidden Markov models
- Parameter estimation: maximum likelihood, Bayesian methods

Detecting Recombination in DNA Sequence Alignments

Dirk Husmeier

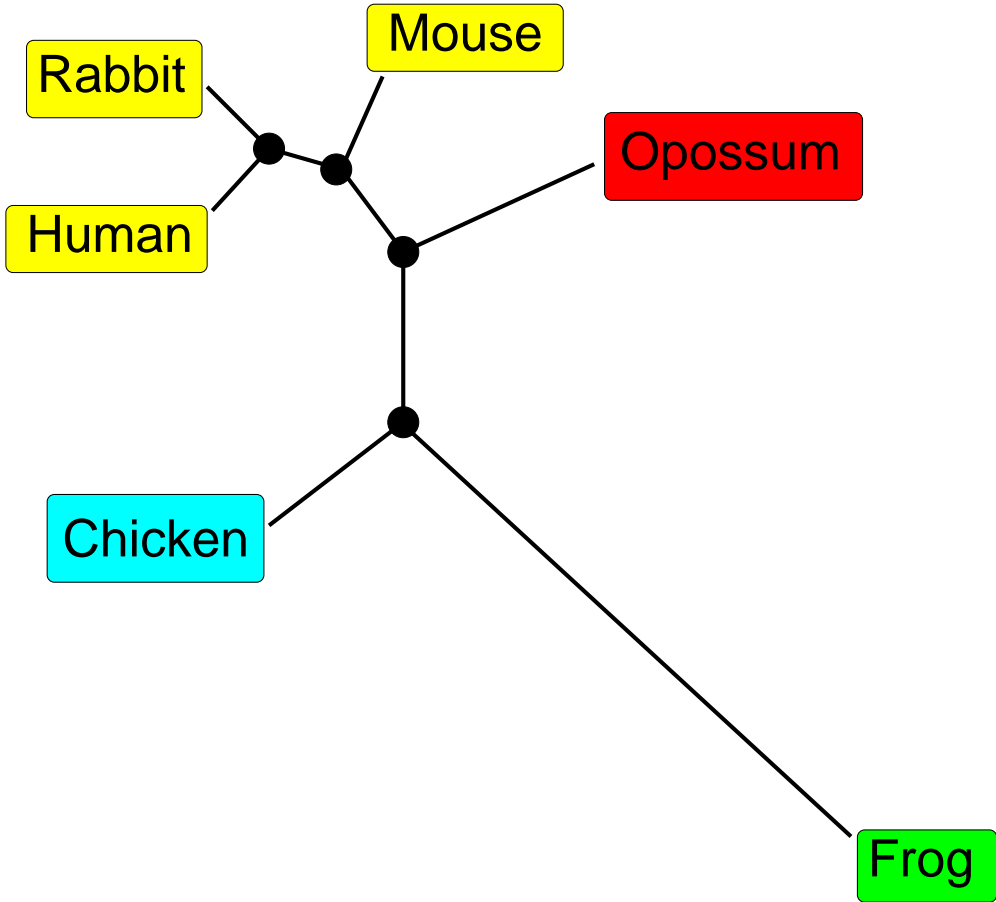
Biomathematics and Statistics Scotland
SCRI, Invergowrie, Dundee DD2 5DA, UK

Email: dirk@bioass.ac.uk

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

- Phylogenetics
- Hidden Markov models
- Parameter estimation: maximum likelihood, Bayesian methods
- Applications

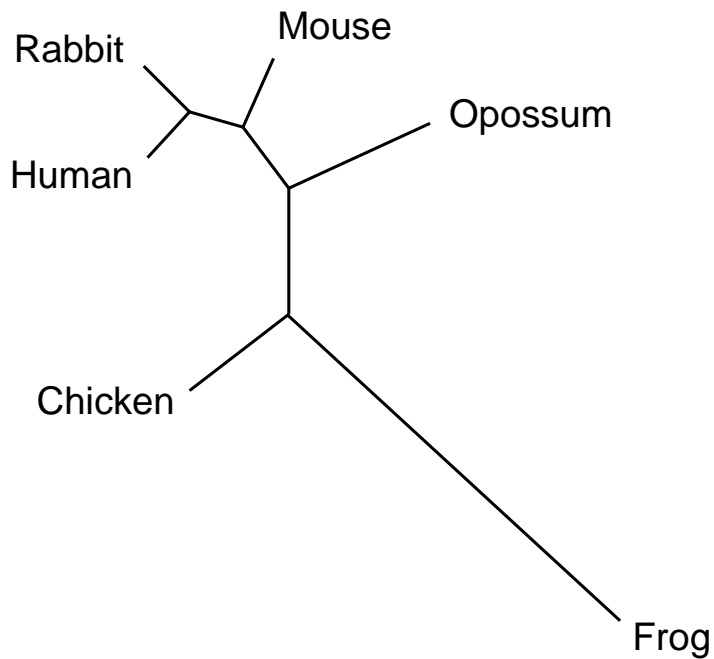
Phylogenetics



--> Topology
--> Branch lengths

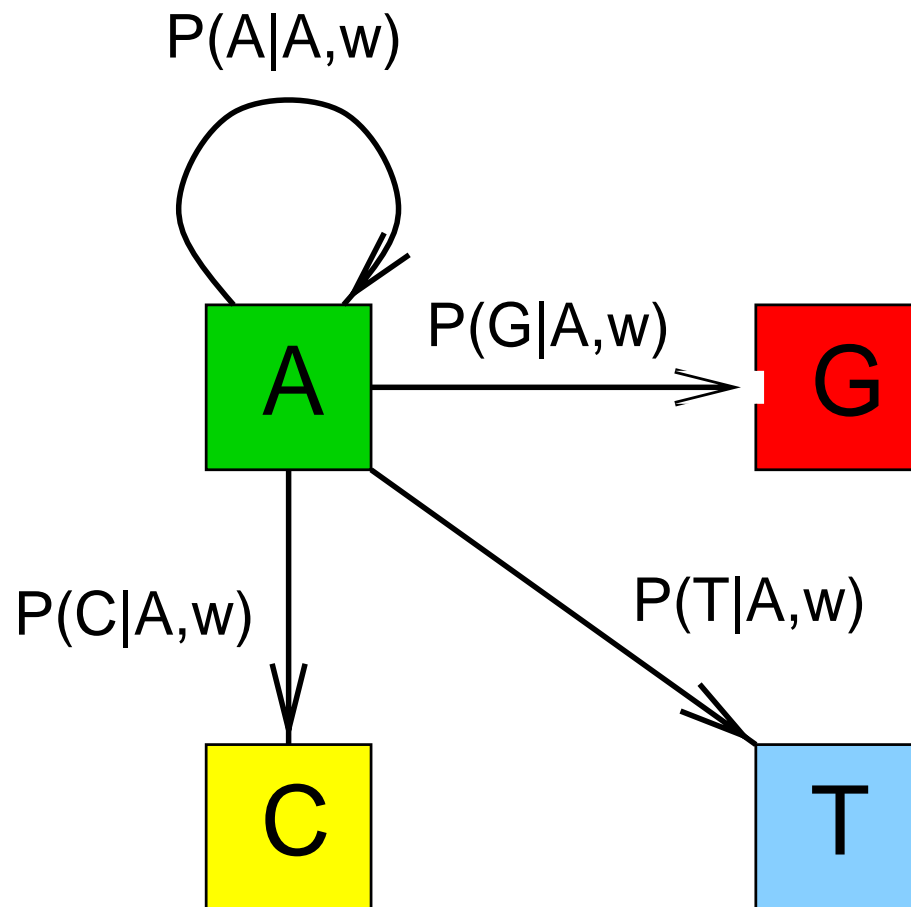
Phylogenetics

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

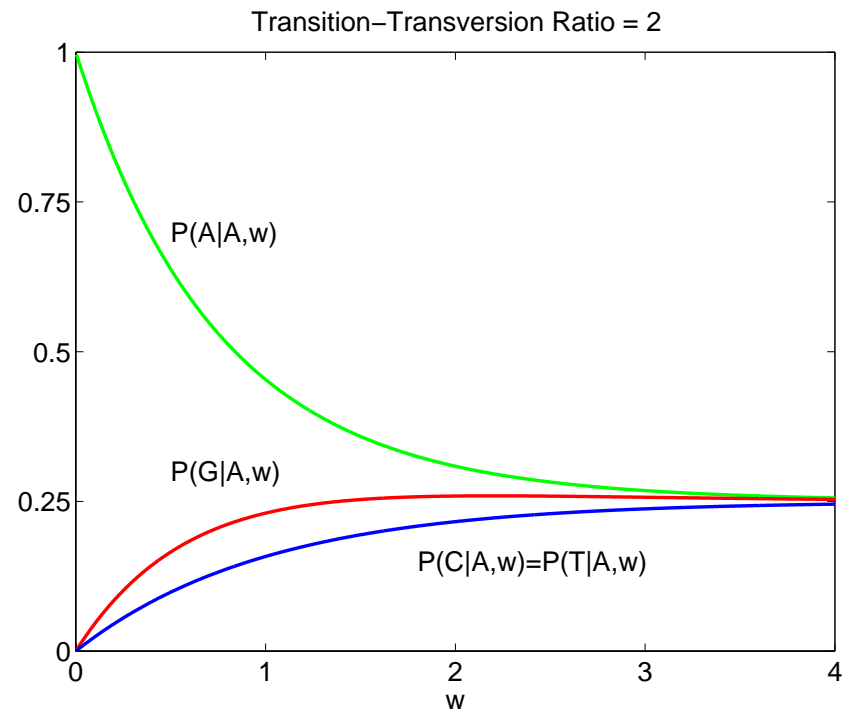
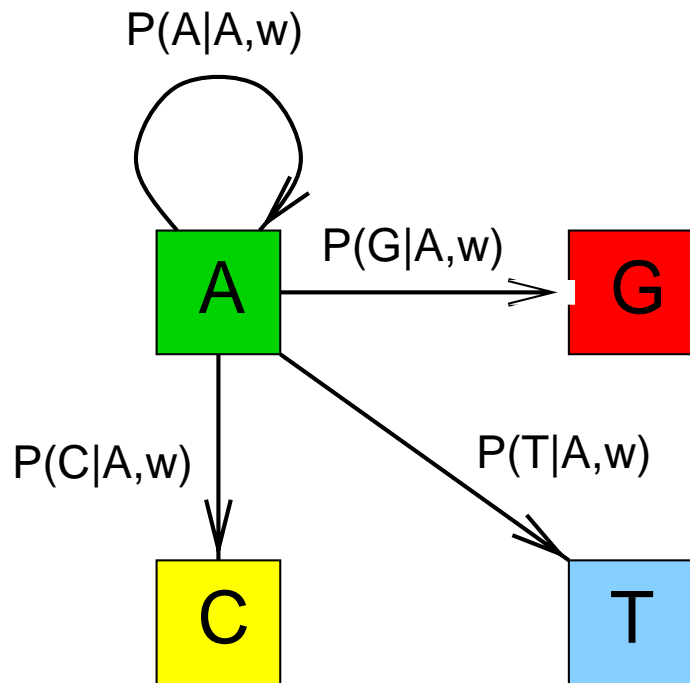


--> Topology
--> Branch lengths

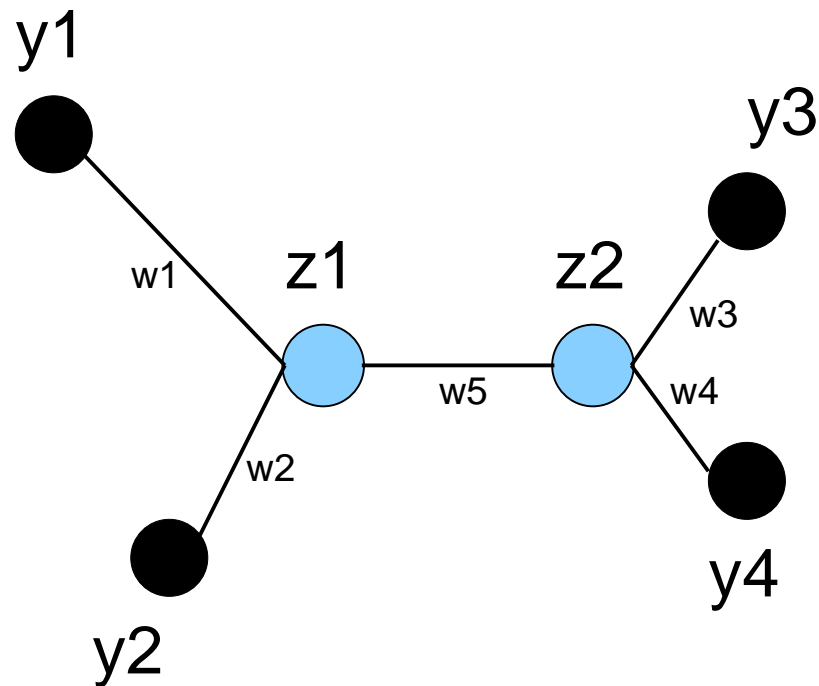
A probabilistic model of evolution



A probabilistic model of evolution



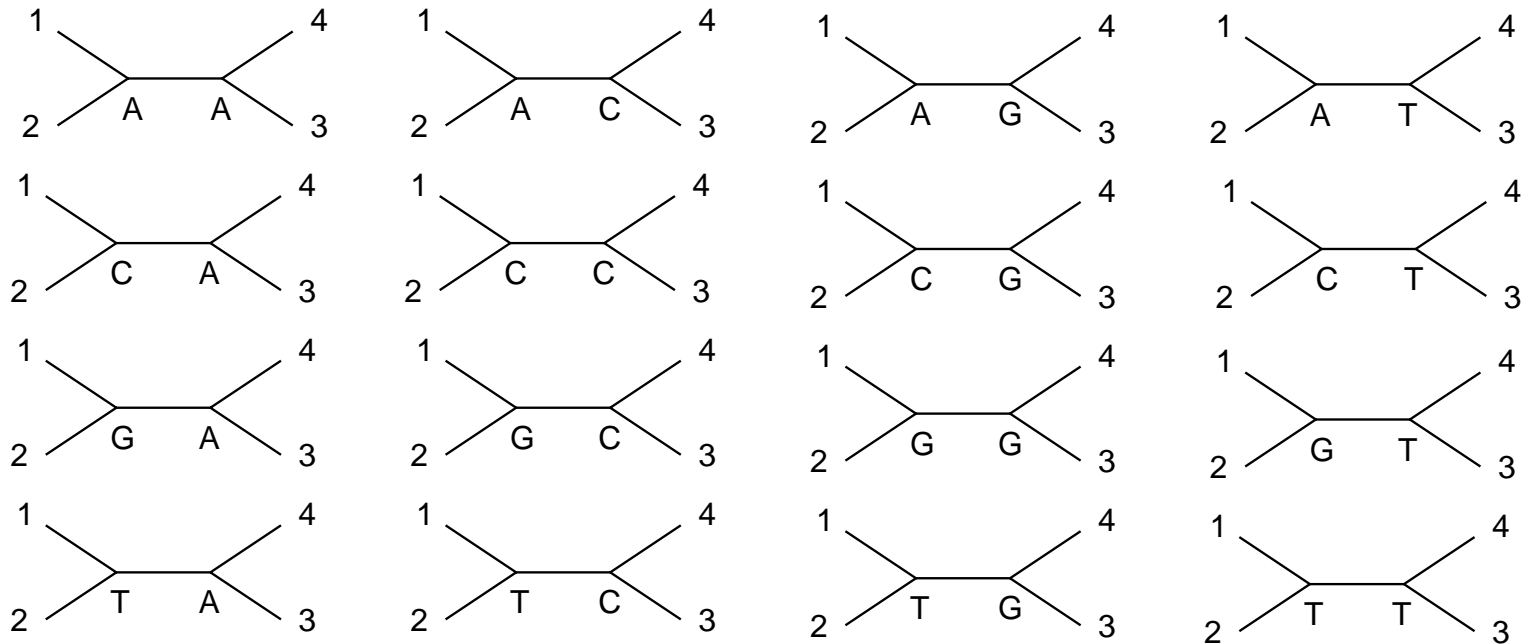
A probabilistic model of evolution



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

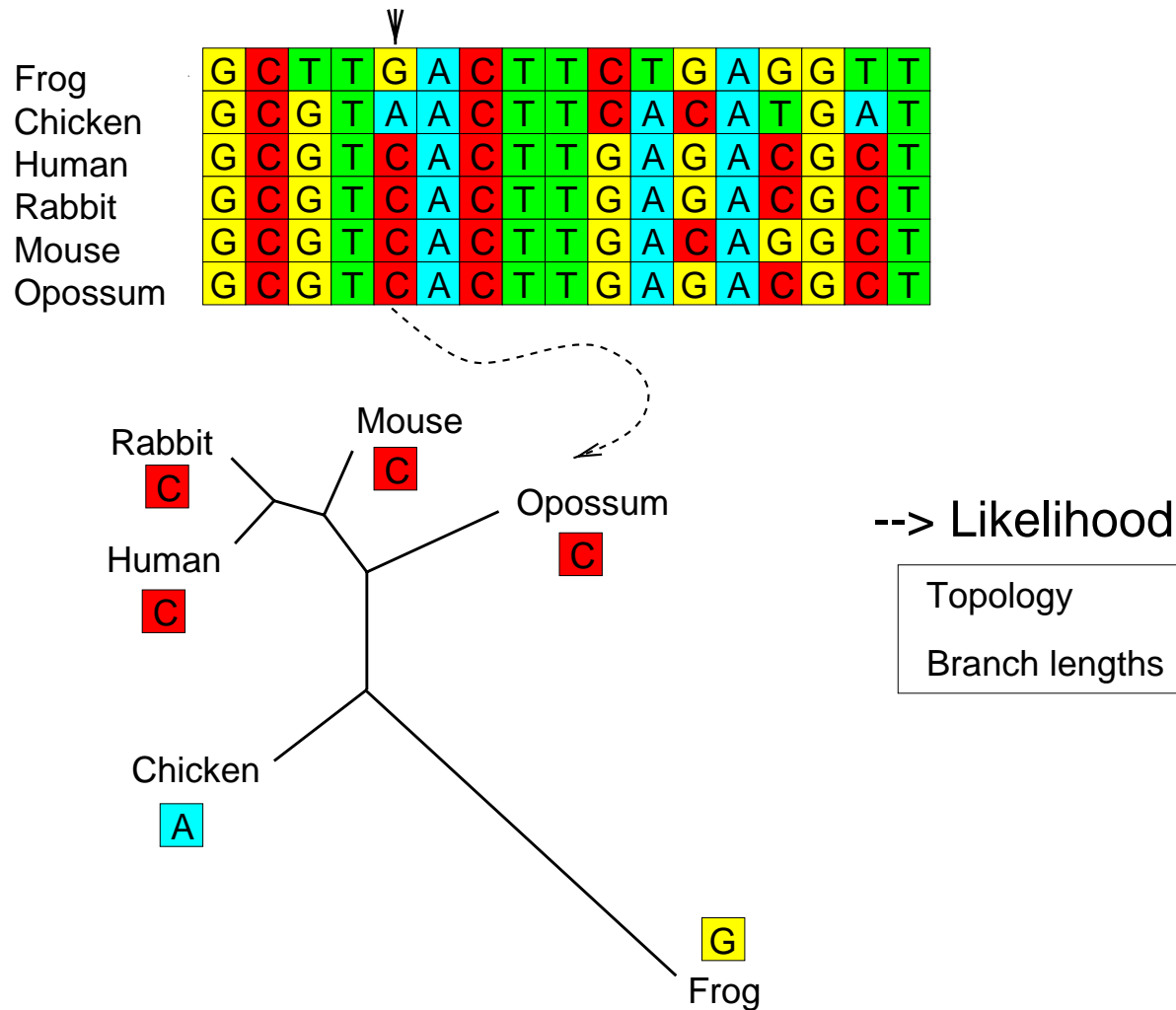
$$= P(y_1 | z_1, w_1) P(y_2 | z_1, w_2) P(z_2 | z_1, w_5) P(y_3 | z_2, w_3) P(y_4 | z_2, w_4)$$

Marginalisation

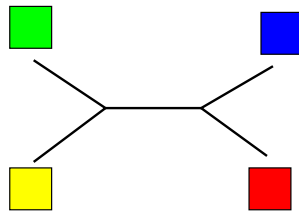
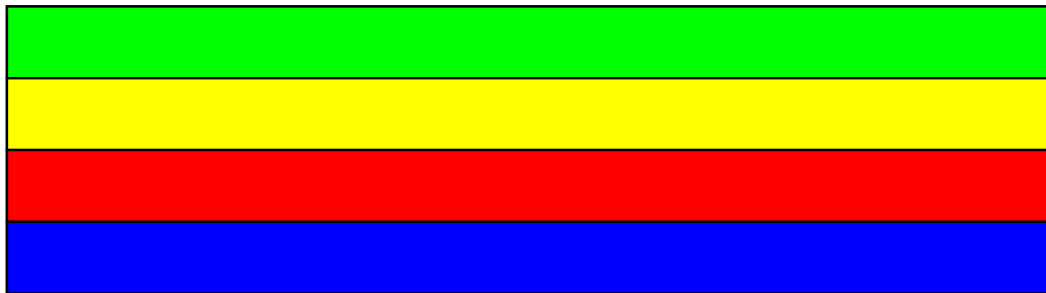
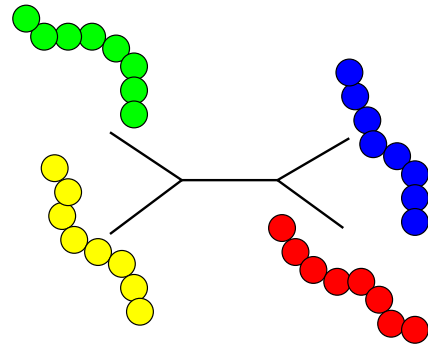


$$P(y_1, y_2, y_3, y_4 | \mathbf{w}) = \sum_{z_1} \sum_{z_2} P(y_1, y_2, y_3, y_4, z_1, z_2 | \mathbf{w})$$

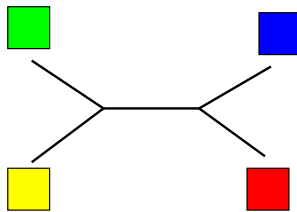
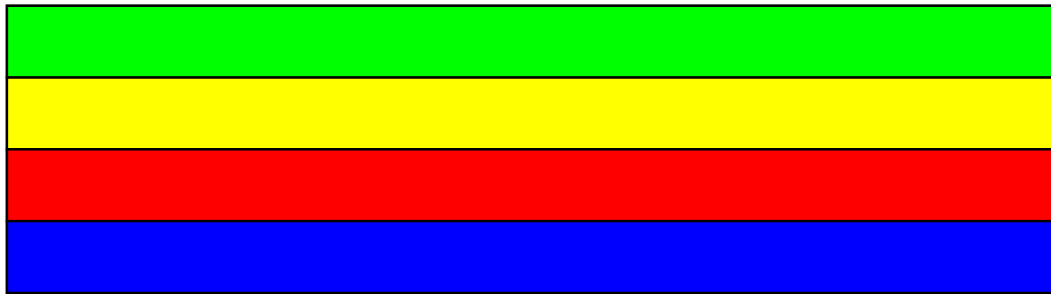
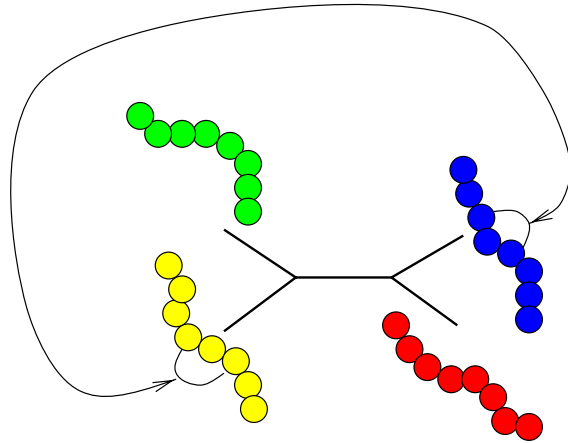
Statistical approach to phylogenetics



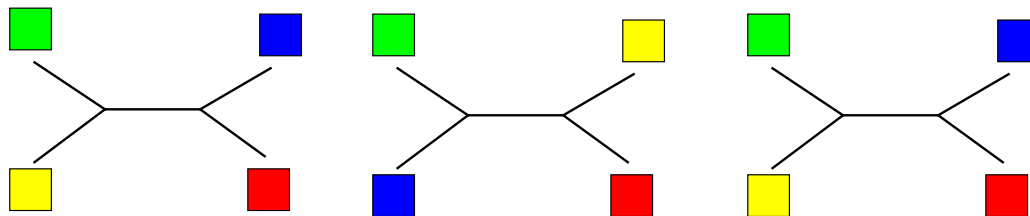
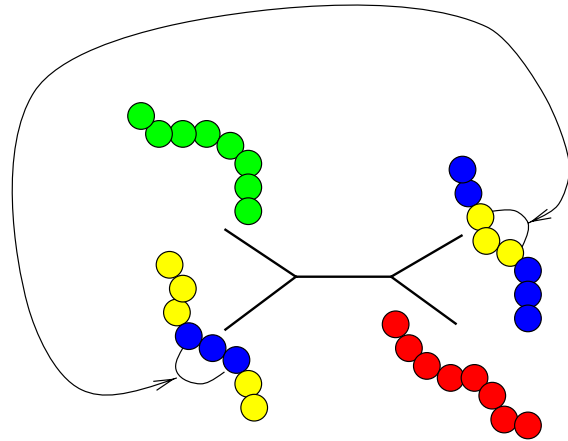
Recombination



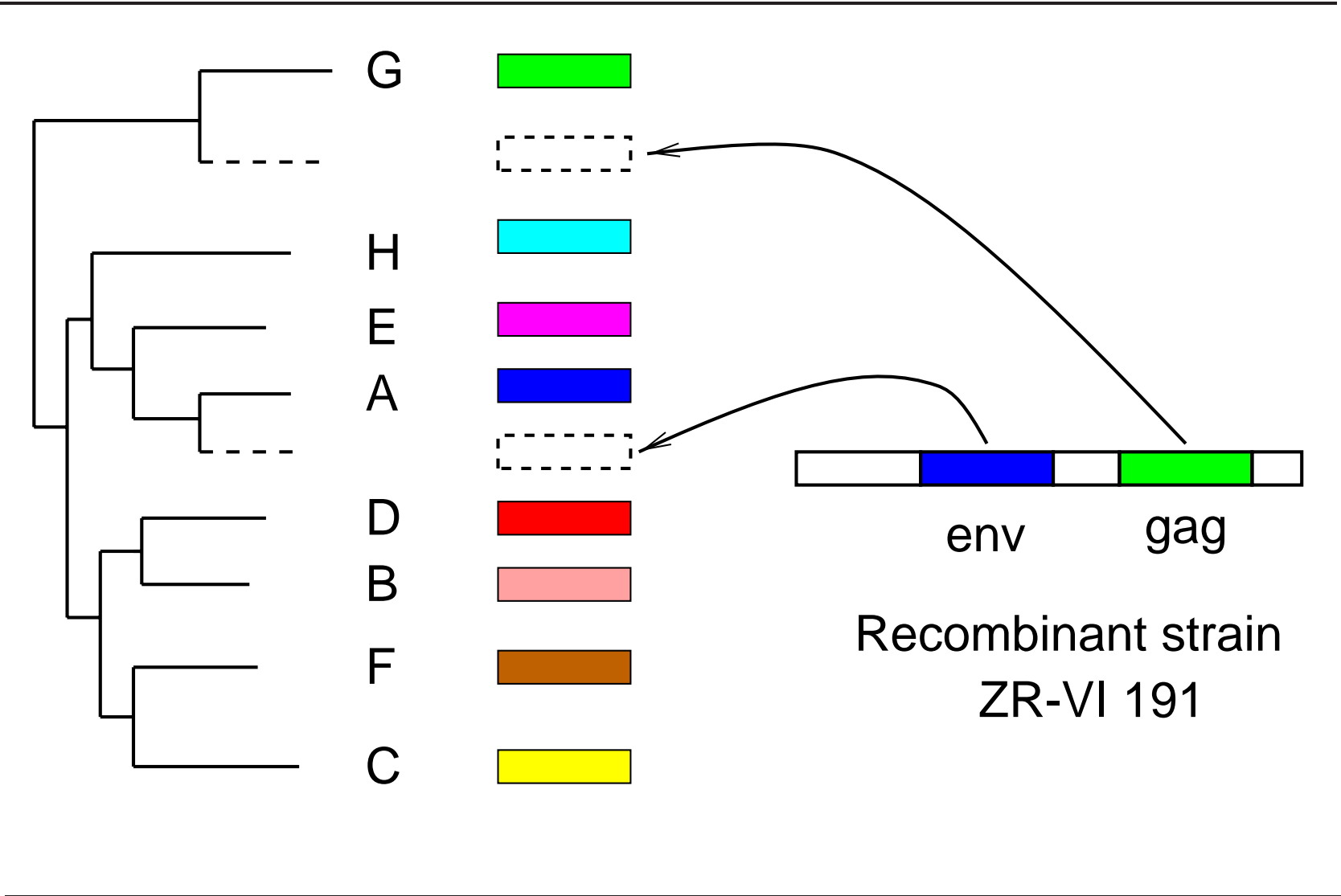
Recombination



Recombination



Recombination in HIV 1

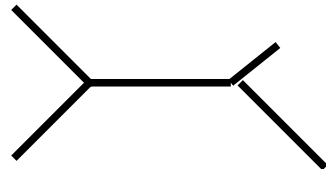
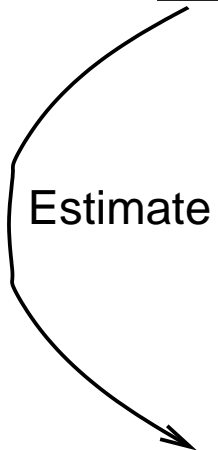
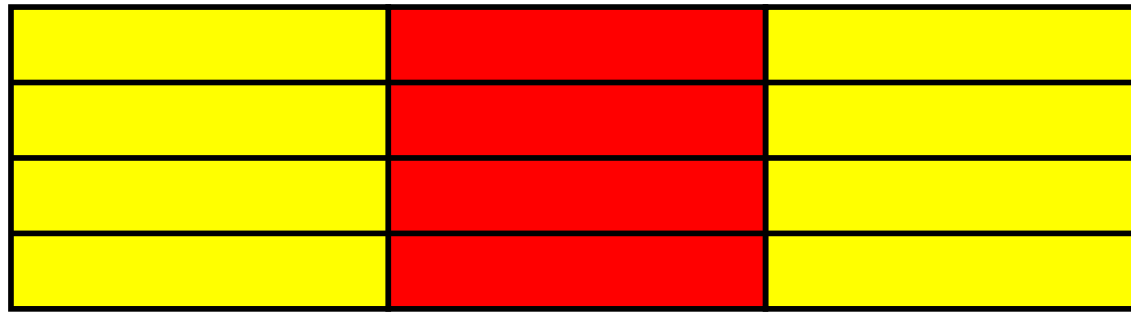


Detecting recombination with window methods

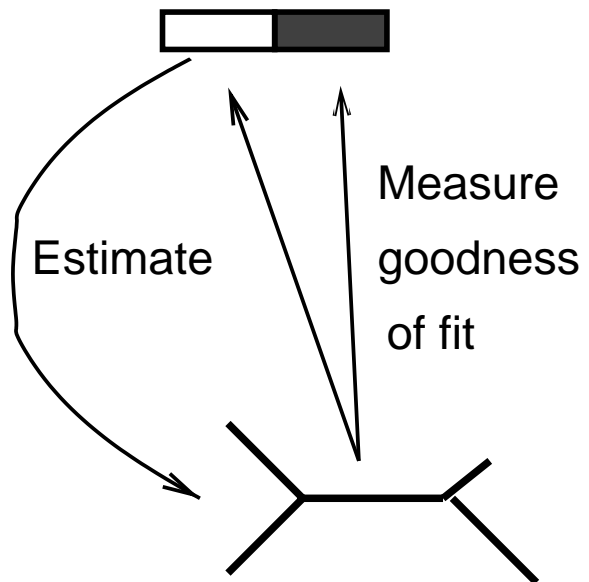
TOPAL (McGuire & Wright, 1997)



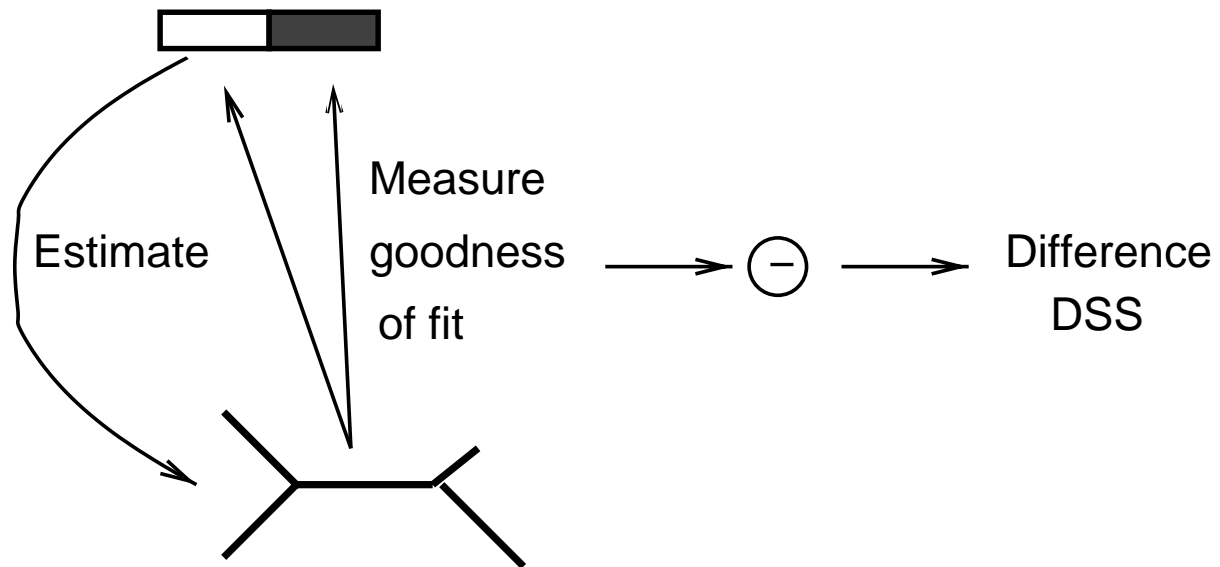
TOPAL (McGuire & Wright, 1997)



TOPAL (McGuire & Wright, 1997)



TOPAL (McGuire & Wright, 1997)



TOPAL (McGuire & Wright, 1997)



small

TOPAL (McGuire & Wright, 1997)



small



large

TOPAL (McGuire & Wright, 1997)



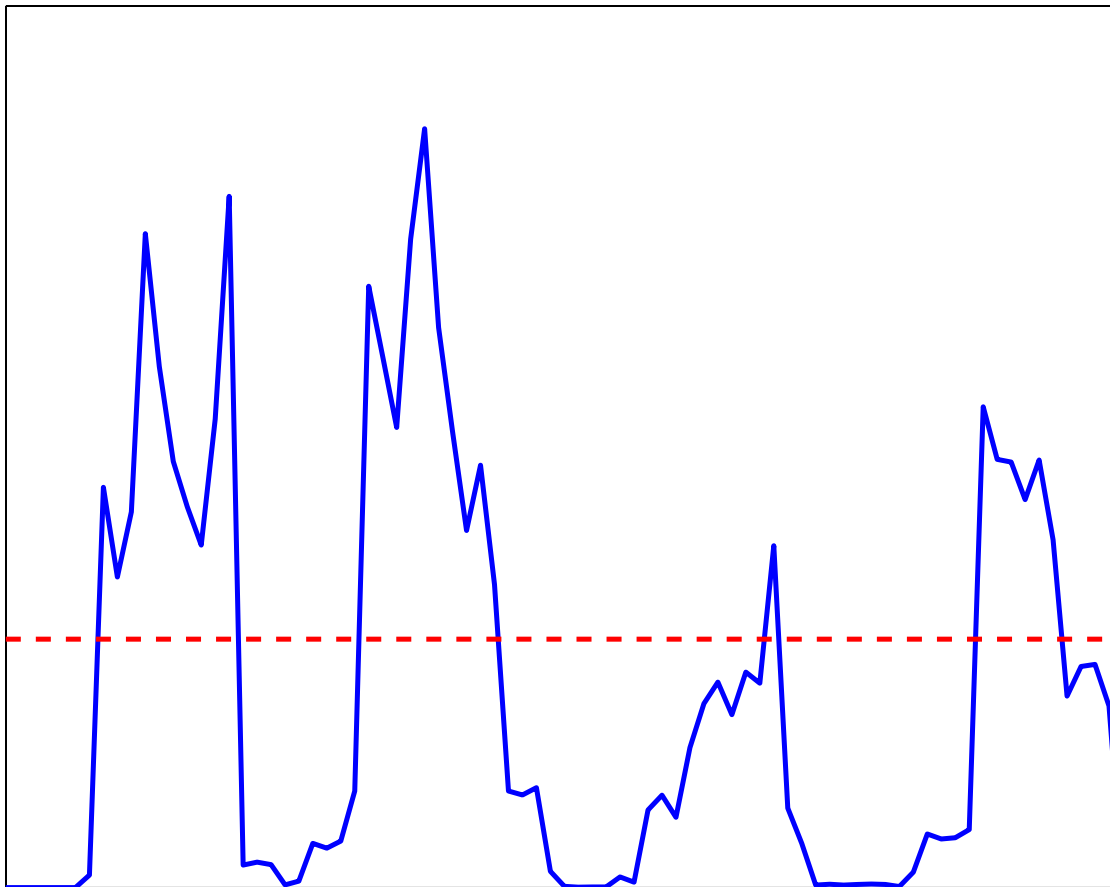
small



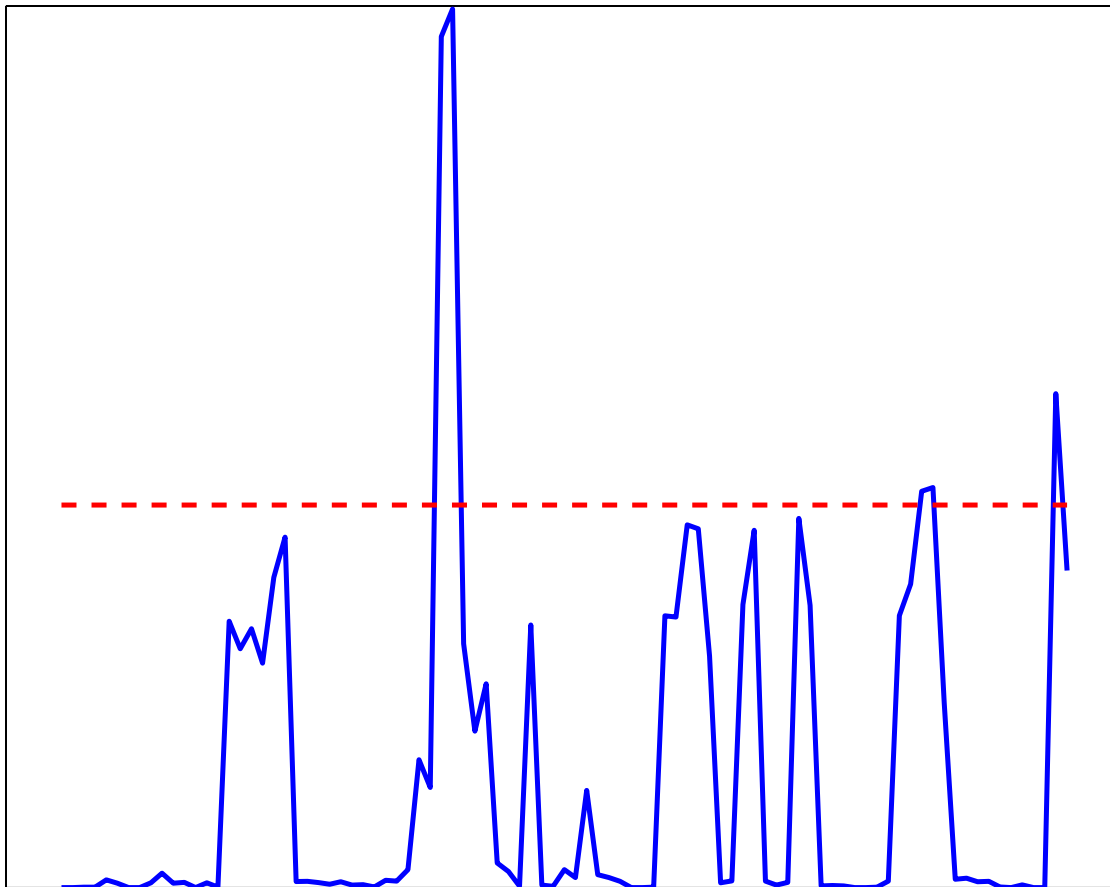
large

- 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.

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.
- 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.

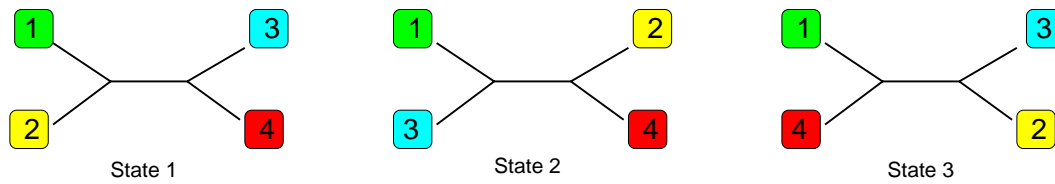
- Can currently only deal with a small number of species.

Hidden Markov models (HMMs)

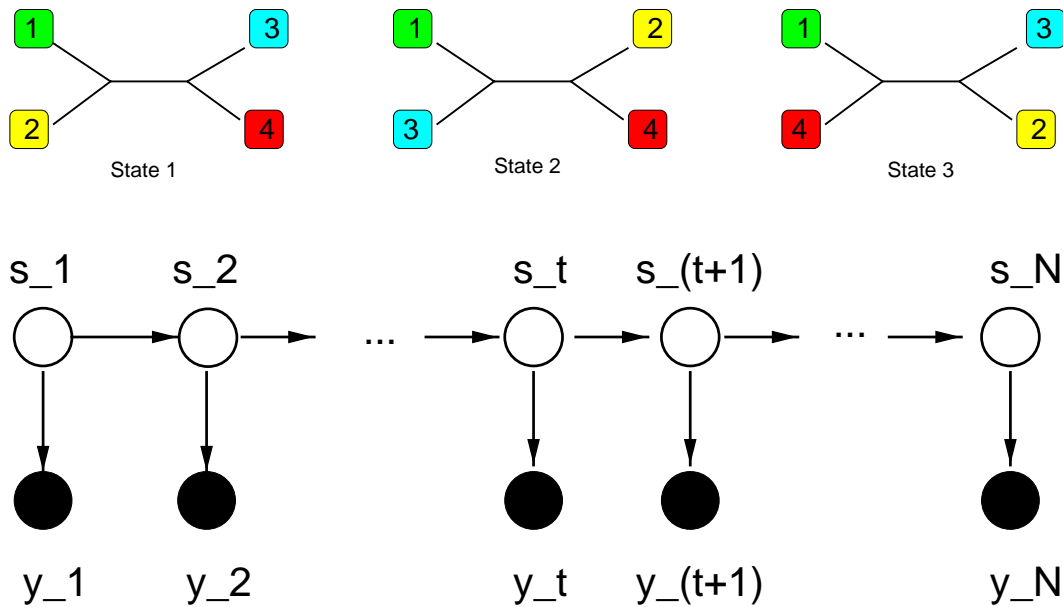
- No window needed.
 - More precise location of the breakpoints.
 - All parameters inferred from the data.

 - Can currently only deal with a small number of species.
 - Current software: Only 4 taxa.
-

Modelling recombination with HMMs

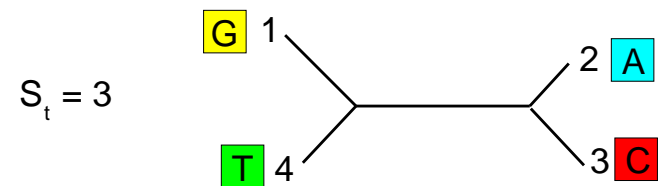
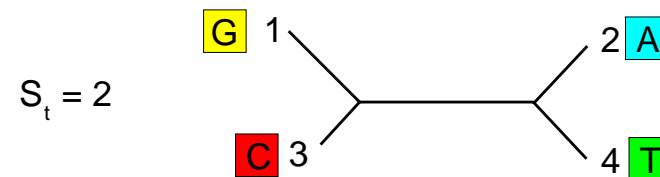
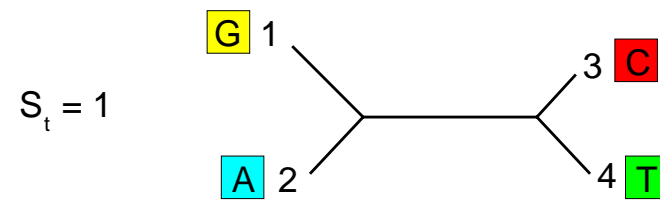
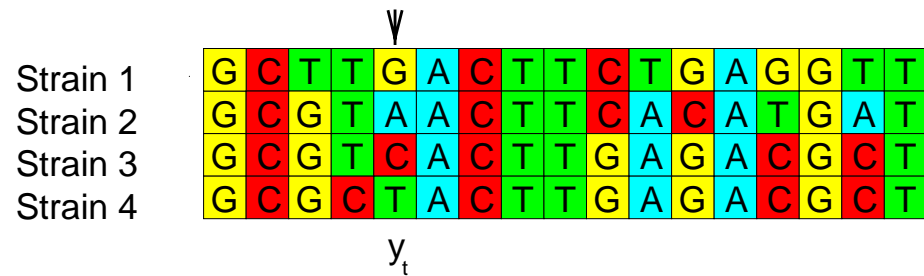


Modelling recombination with HMMs



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

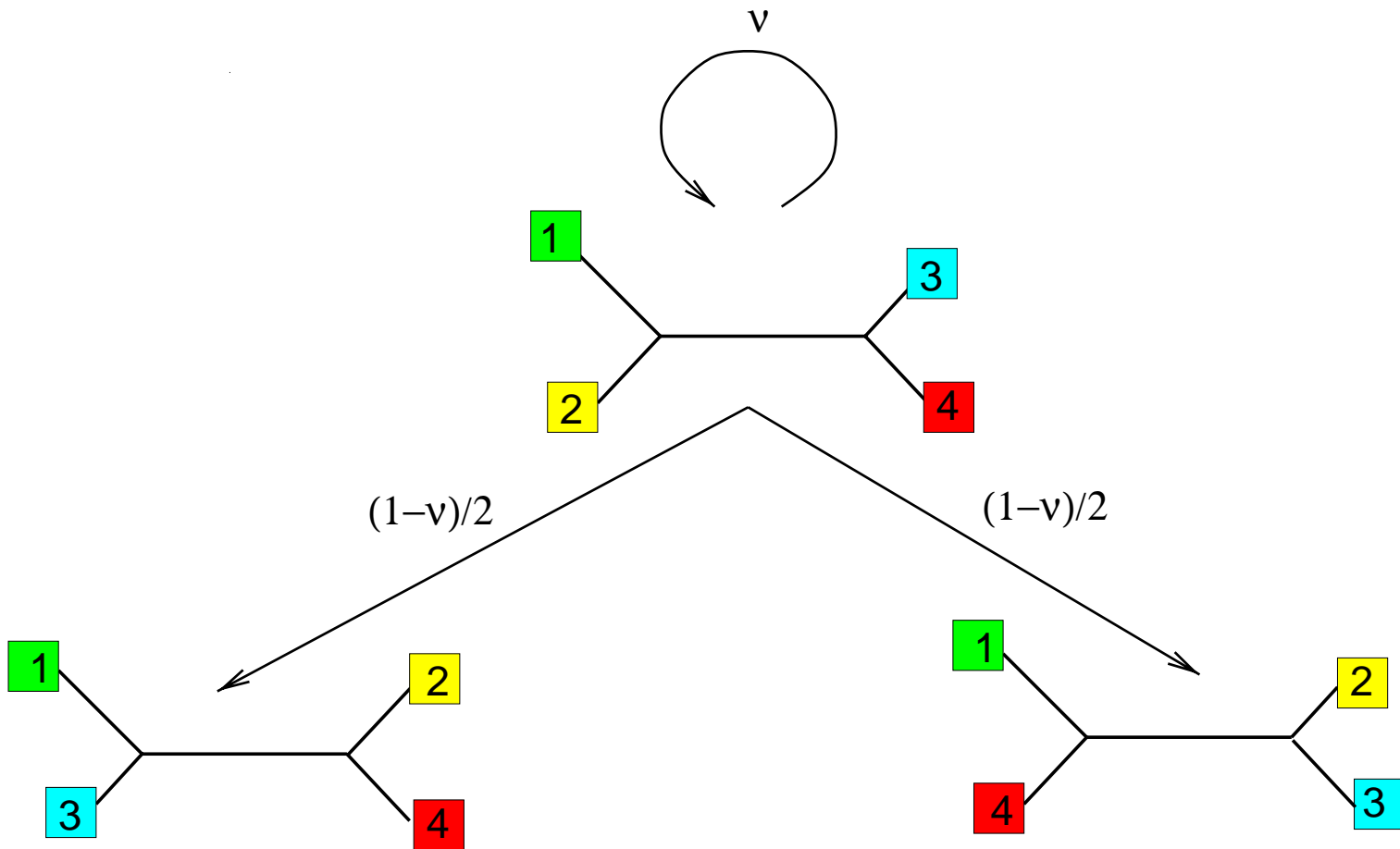
Emission probabilities (vertical arrows)



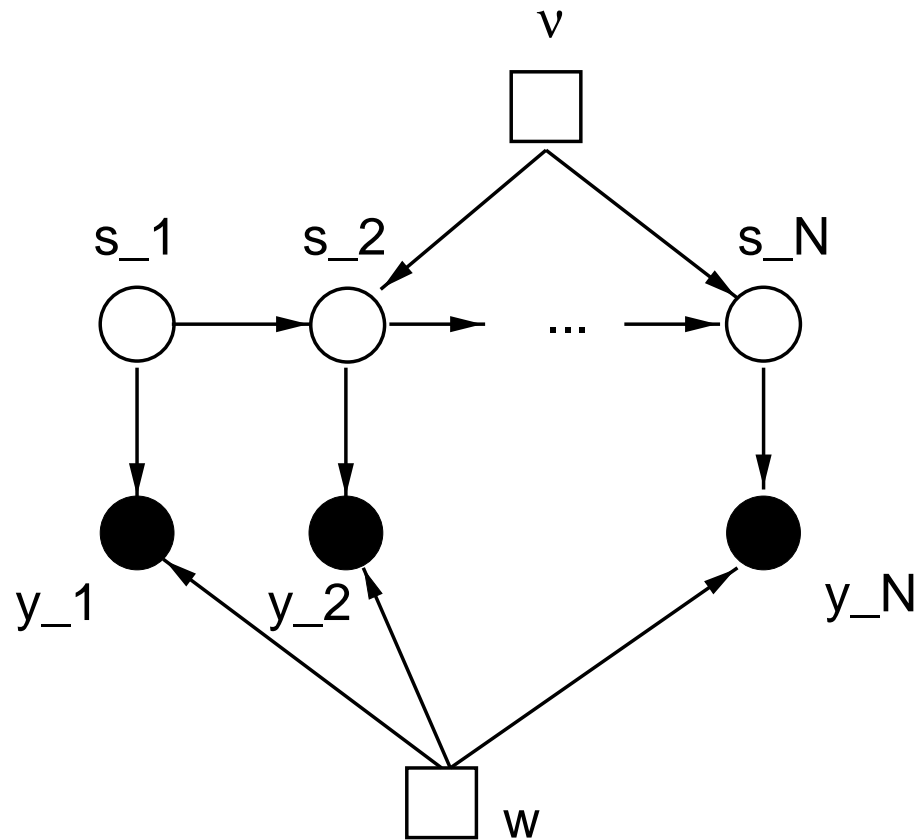
--> $P(y_t | S_t, w)$

Topology	S_t
Branch lengths	w

Transition probabilities (horizontal arrows)



HMM parameters



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

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

Parameter estimation

Parameter estimation

- **Maximum likelihood (EM algorithm)**
Husmeier, Wright (2001)
Journal of Computational Biology 8

Parameter estimation

- **Maximum likelihood (EM algorithm)**

Husmeier, Wright (2001)

Journal of Computational Biology 8

- **Bayesian approach**

Husmeier, McGuire (2002)

Current work

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

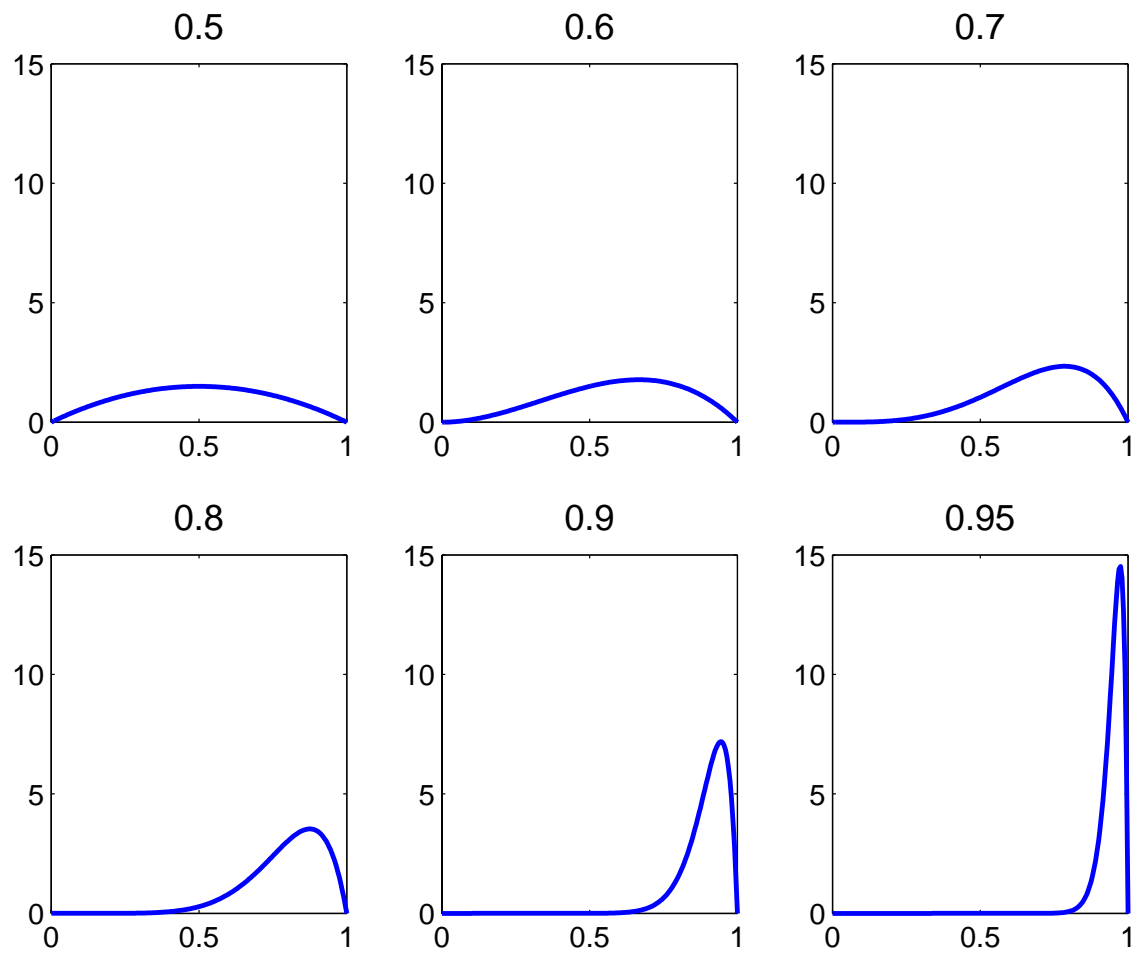
Bayesian approach

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

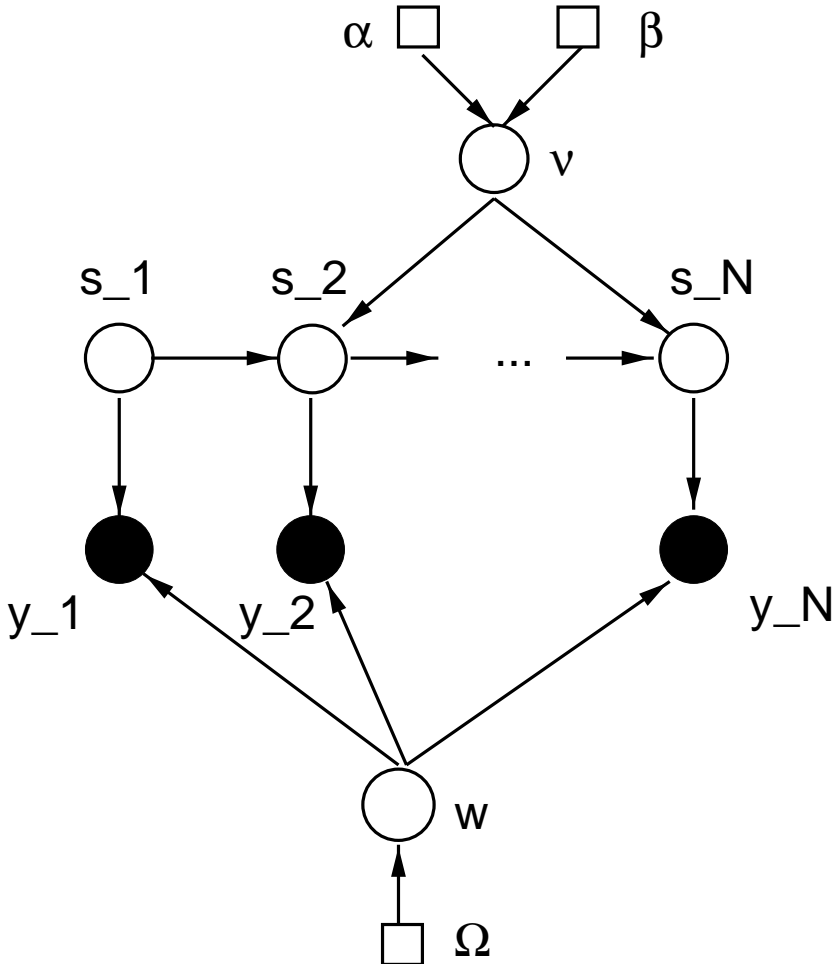
Bayesian approach

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



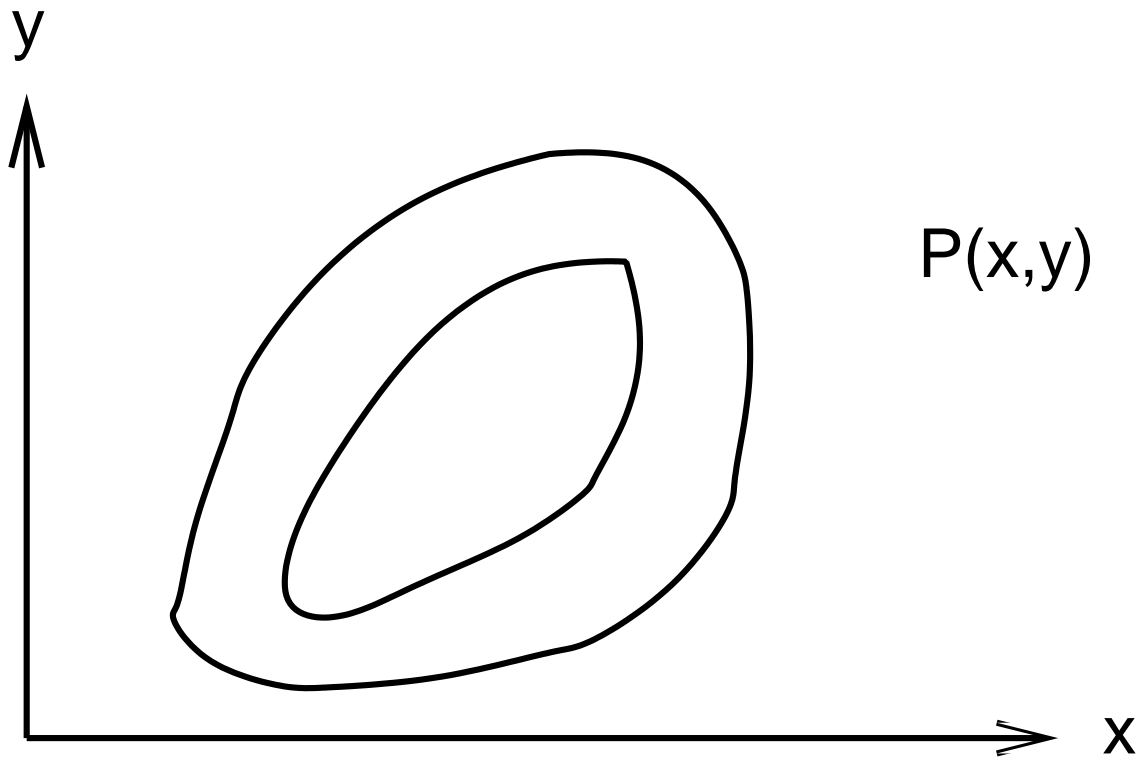
Sampling from the posterior distribution

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

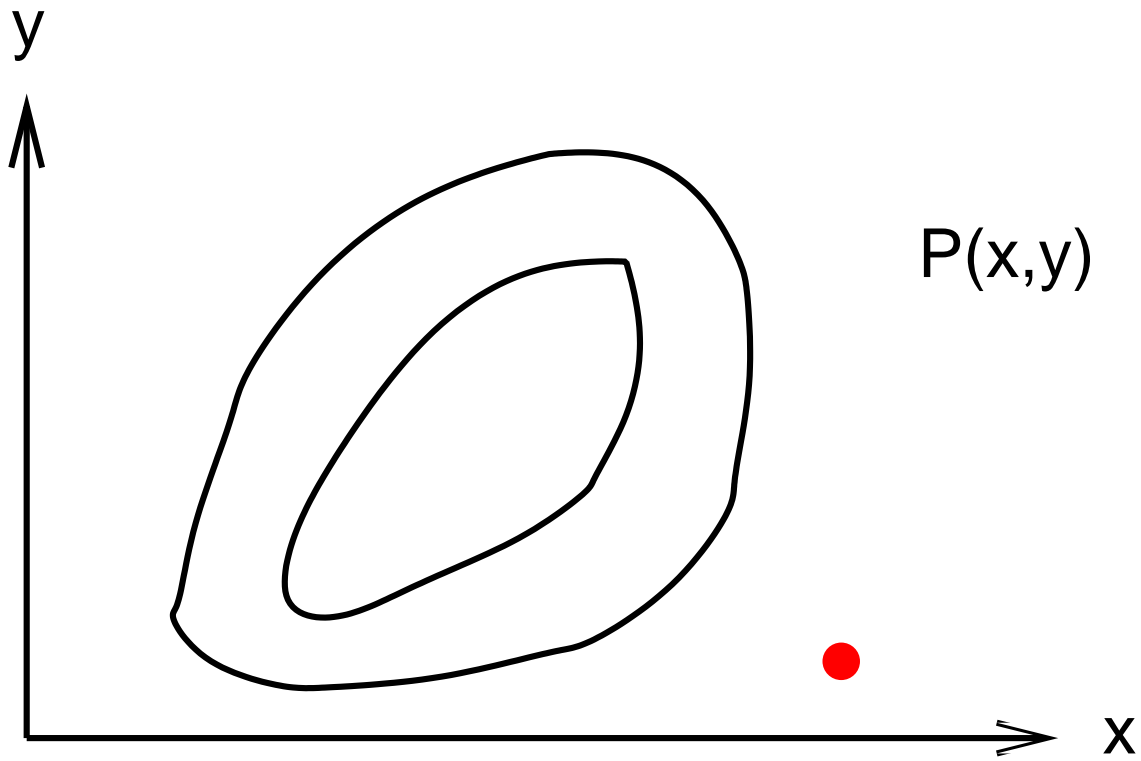
Sampling from the posterior distribution

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

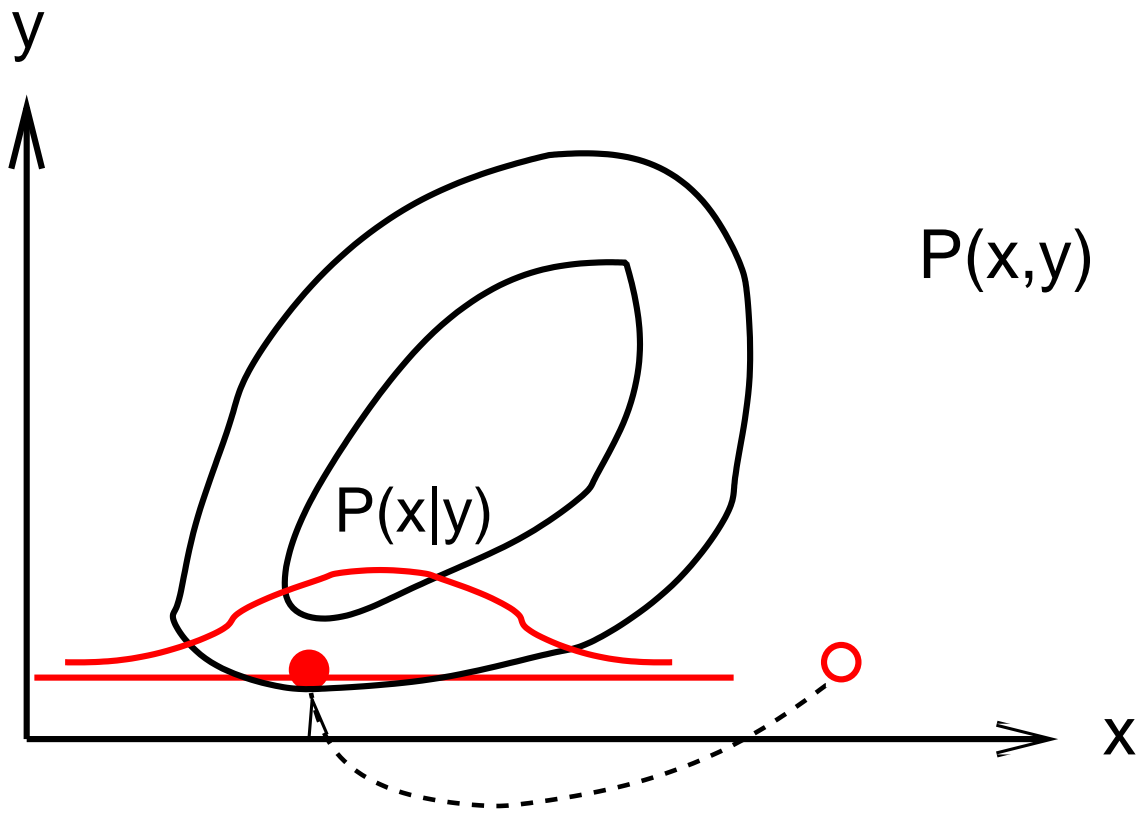
Gibbs sampling



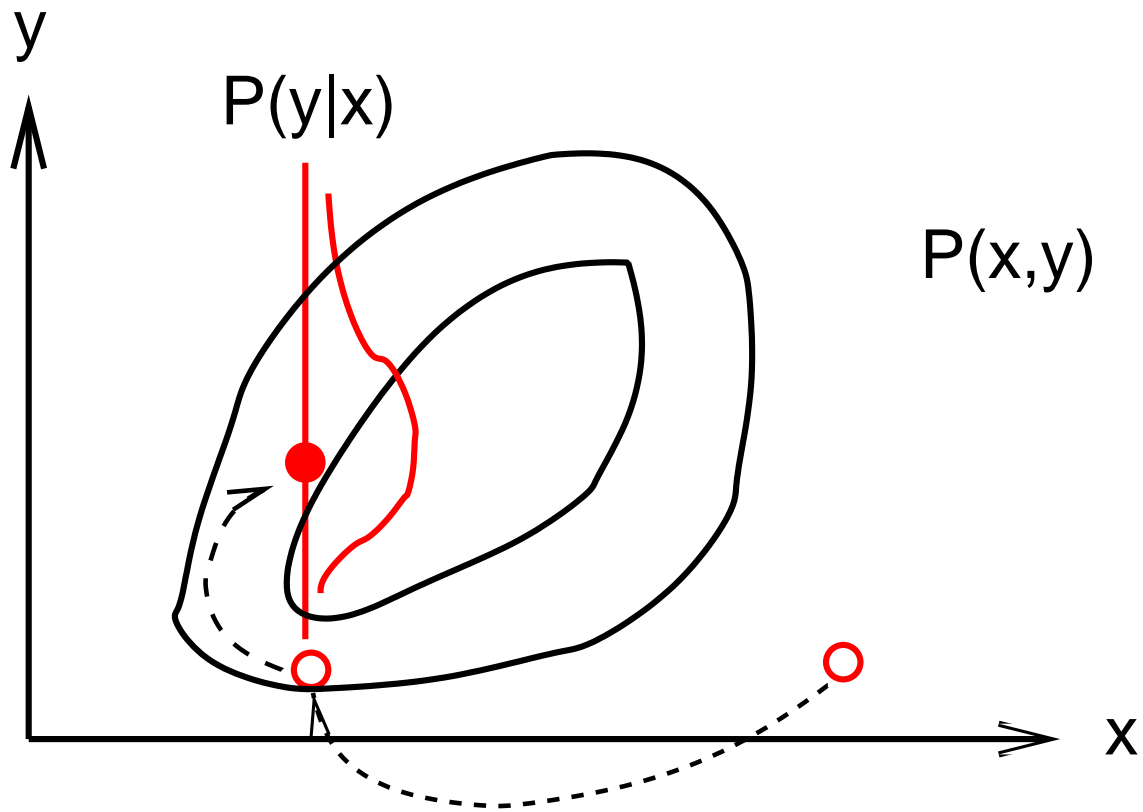
Gibbs sampling



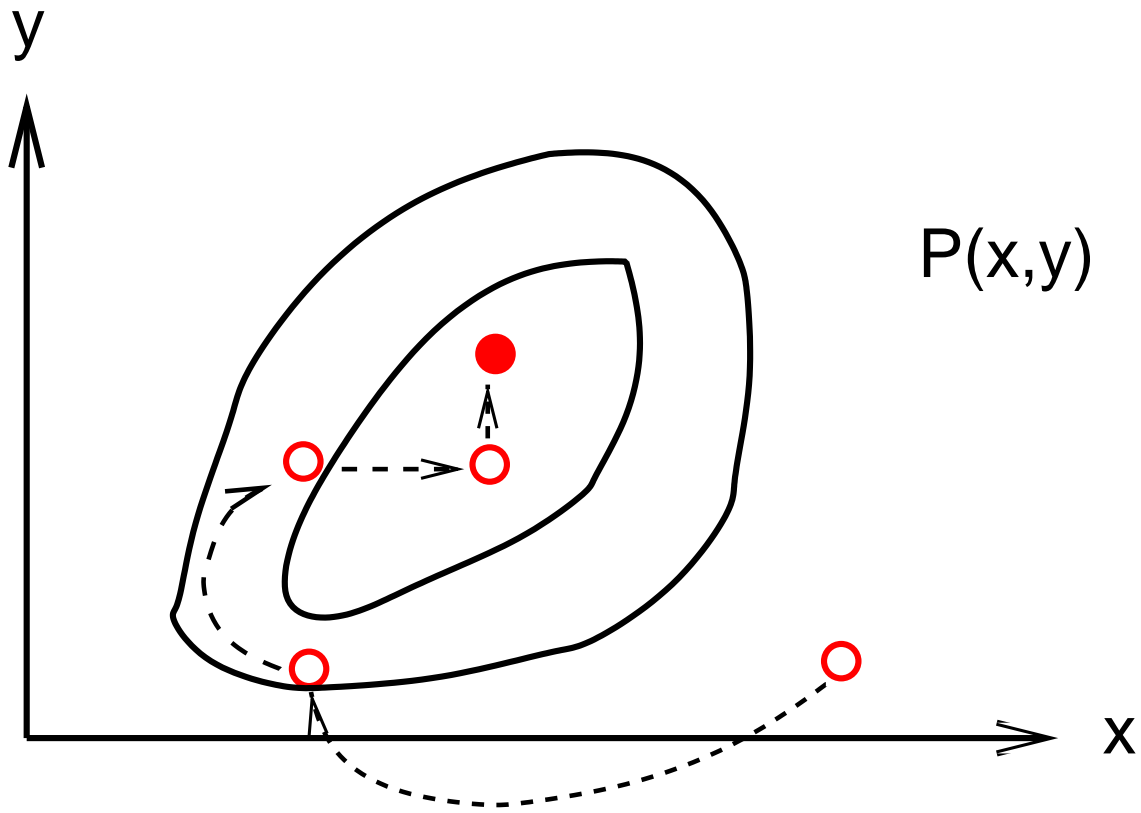
Gibbs sampling



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

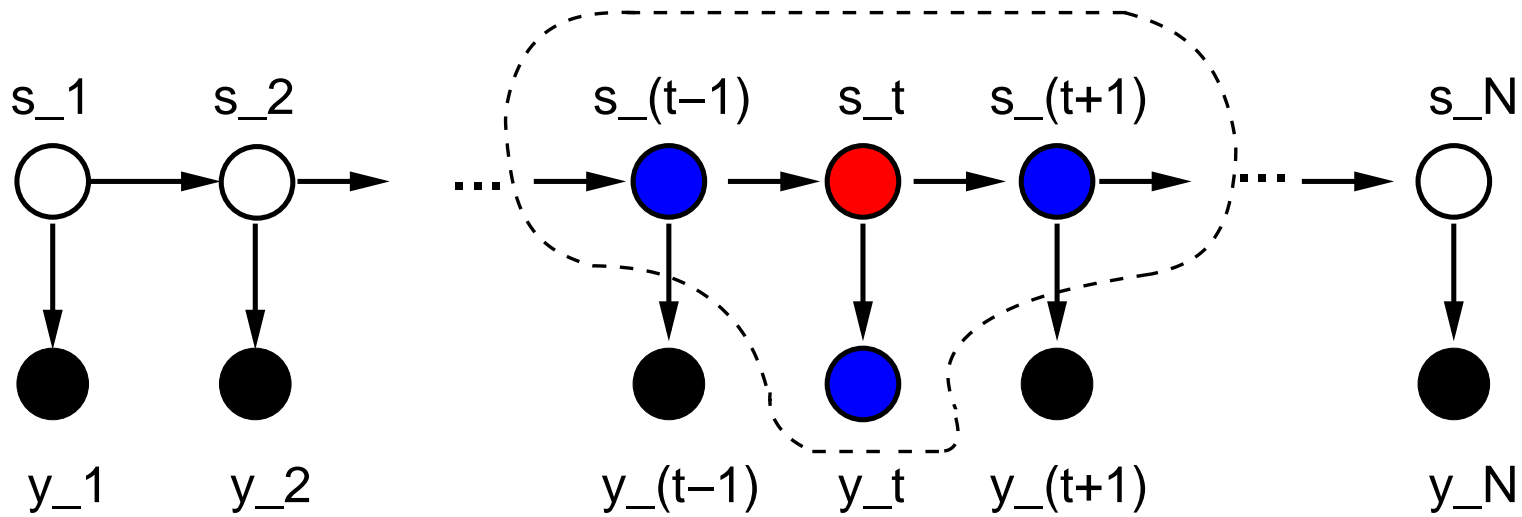
Sampling from the posterior distribution

- 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

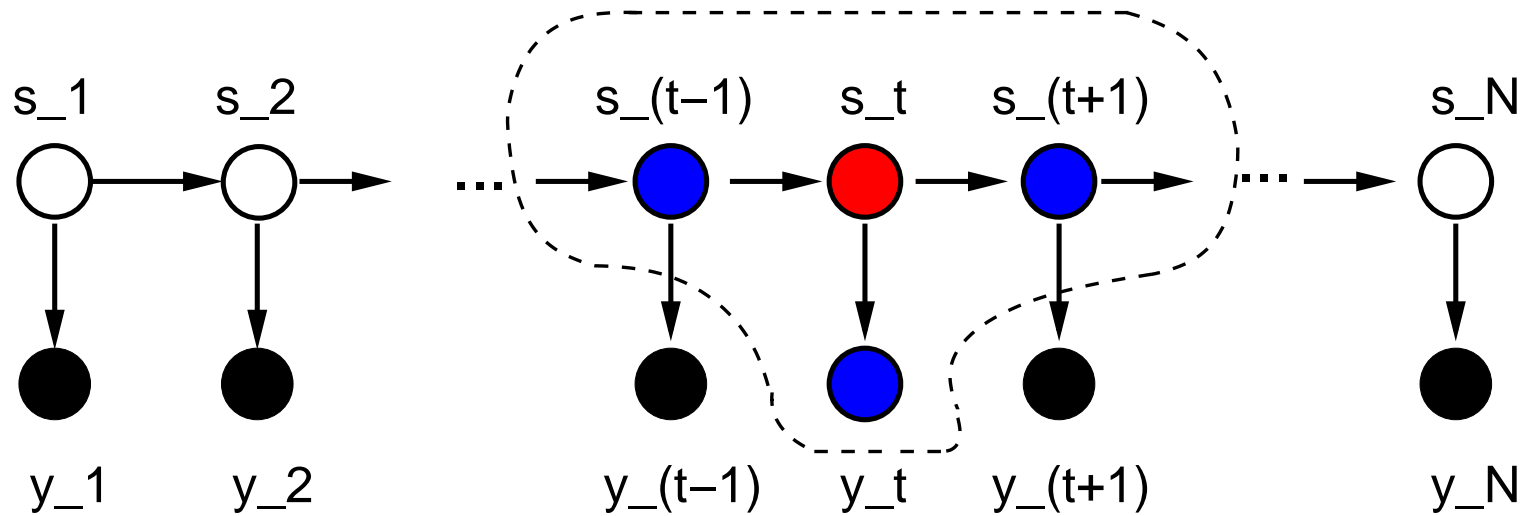
- Sampling from $P(\mathbf{S}, \mathbf{w}, \nu | \mathcal{D})$
 - Gibbs-like approach:
 - $\mathbf{S} \sim P(\mathbf{S} | \mathbf{w}, \nu, \mathcal{D})$
 - $\mathbf{w} \sim P(\mathbf{w} | \mathbf{S}, \nu, \mathcal{D})$
 - $\nu \sim P(\nu | \mathbf{S}, \mathbf{w}, \mathcal{D})$
 - ν : Sample from Beta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



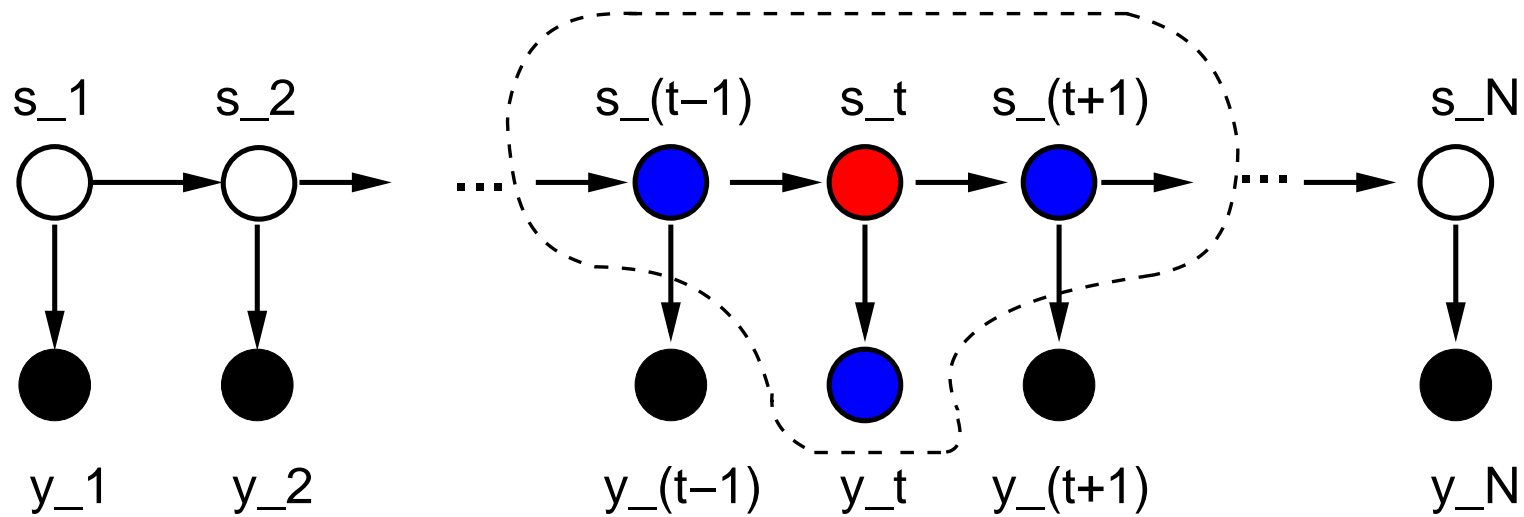
$$P(S_t | S_1, \dots, S_{t-1}, S_{t+1}, \dots, S_N, \mathcal{D}, \mathbf{w}, \nu)$$

Sampling from the posterior distribution



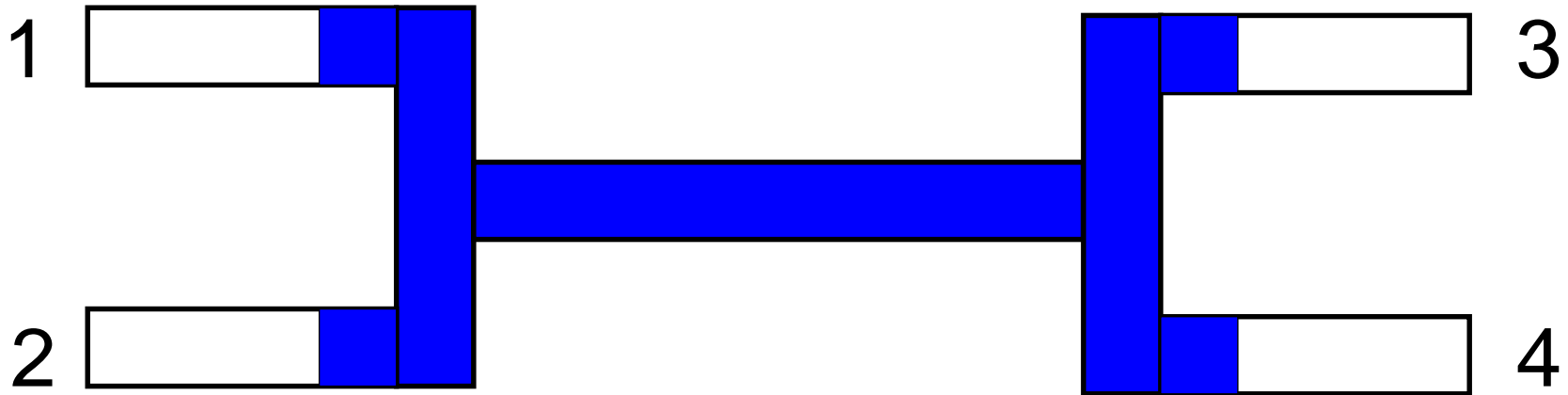
$$\begin{aligned} P(S_t | S_1, \dots, S_{t-1}, S_{t+1}, \dots, S_N, \mathcal{D}, \mathbf{w}, \nu) \\ = P(S_t | S_{t-1}, S_{t+1}, y_t, \mathbf{w}, \nu) \end{aligned}$$

Sampling from the posterior distribution

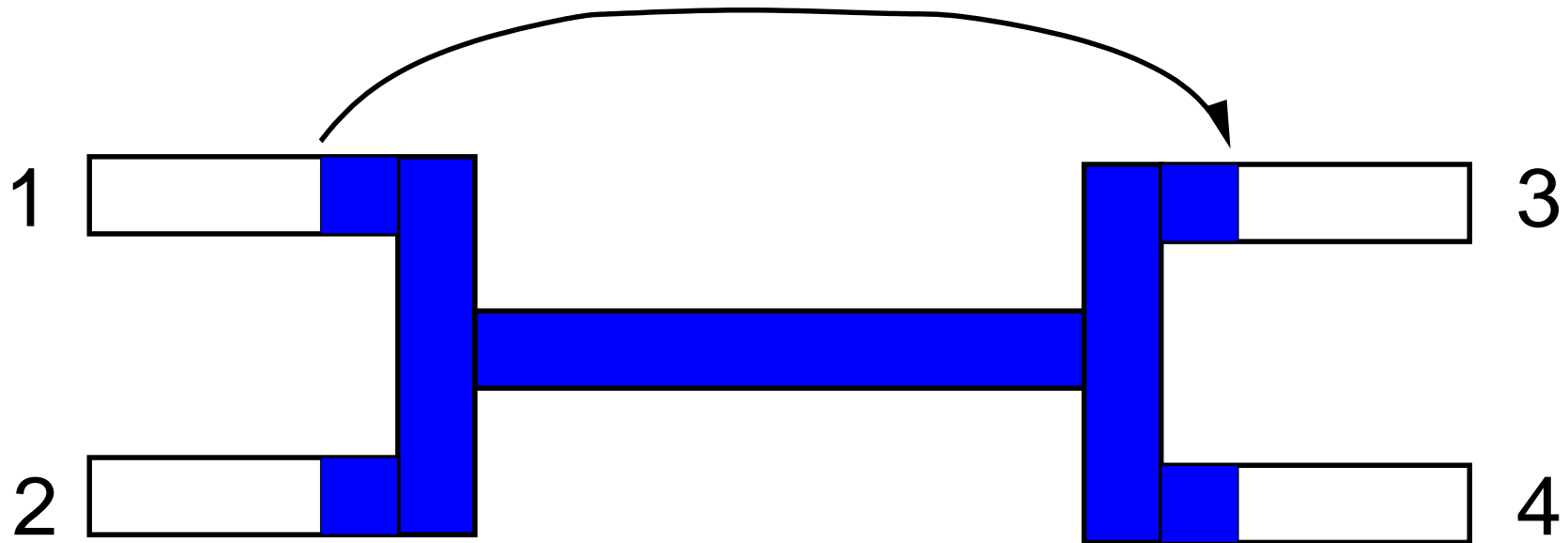


$$\begin{aligned} P(\mathbf{S}_t | S_1, \dots, S_{t-1}, S_{t+1}, \dots, S_N, \mathcal{D}, \mathbf{w}, \nu) \\ &= P(\mathbf{S}_t | S_{t-1}, S_{t+1}, \mathbf{y}_t, \mathbf{w}, \nu) \\ &\propto P(S_{t+1} | S_t, \nu) P(S_t | S_{t-1}, \nu) P(\mathbf{y}_t | S_t, \mathbf{w}) \end{aligned}$$

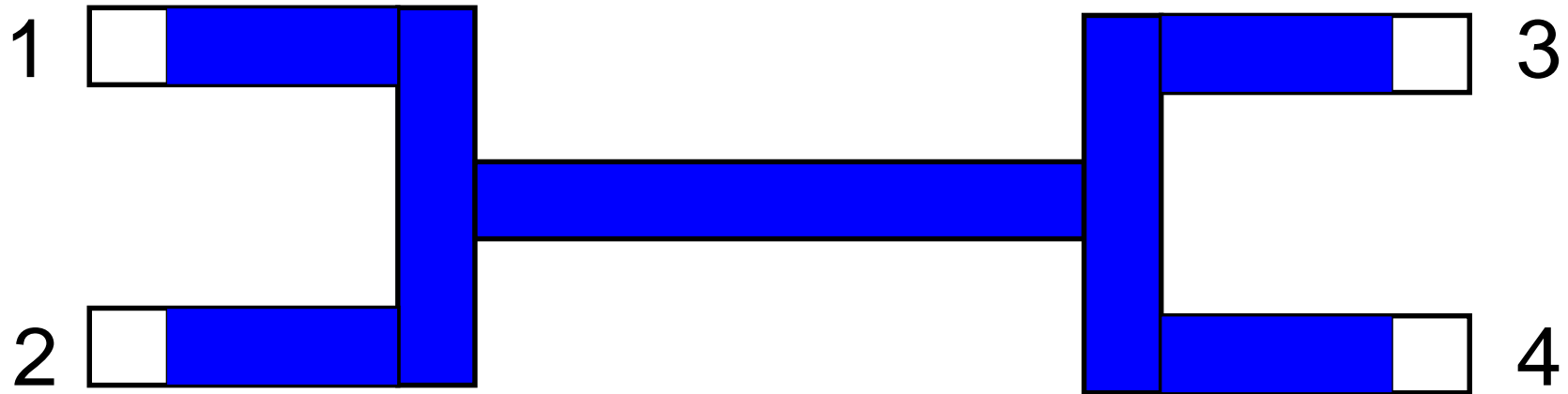
Simulation of recombination



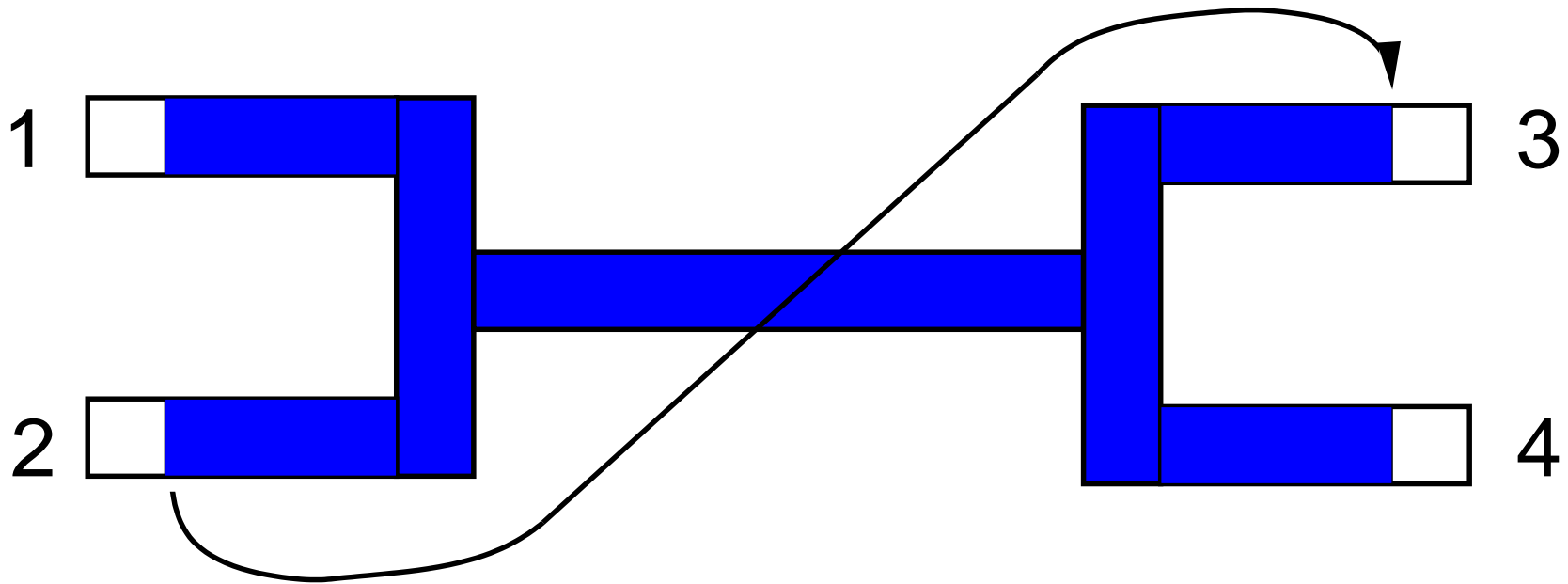
Simulation of recombination



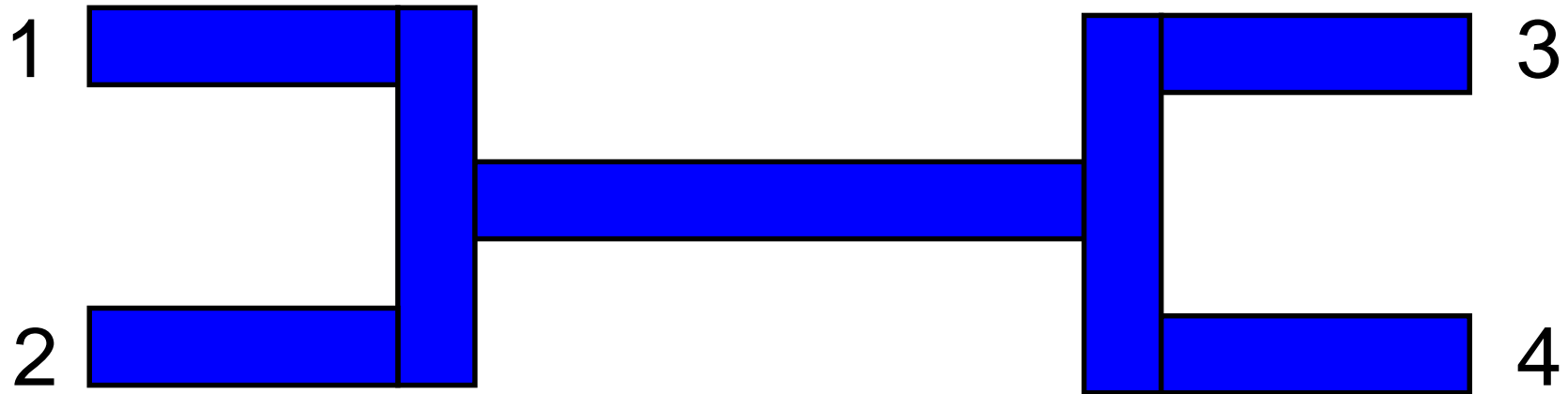
Simulation of recombination



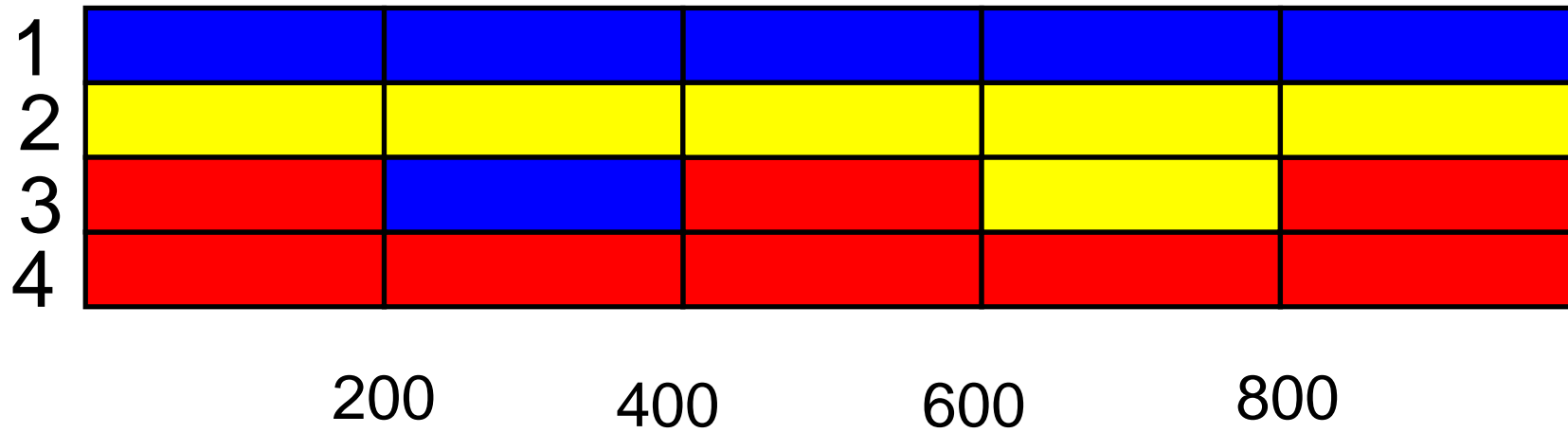
Simulation of recombination



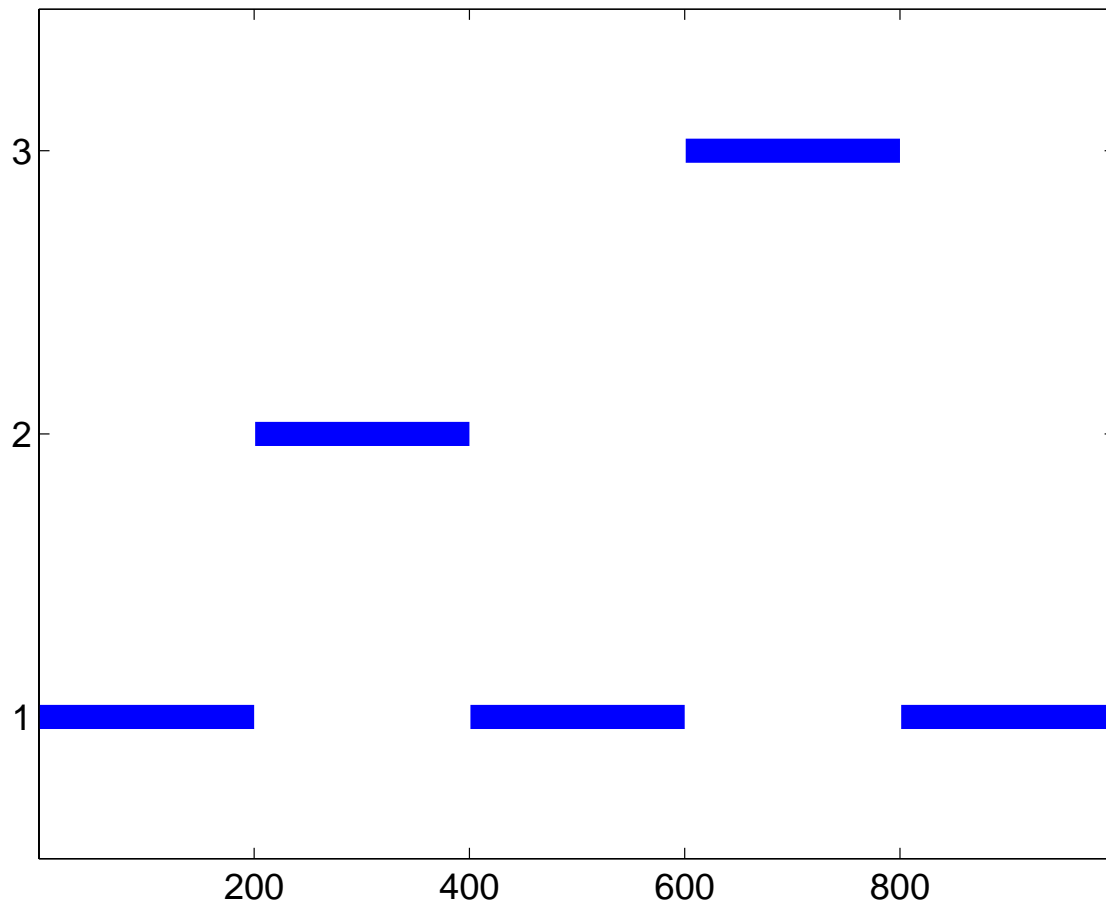
Simulation of recombination



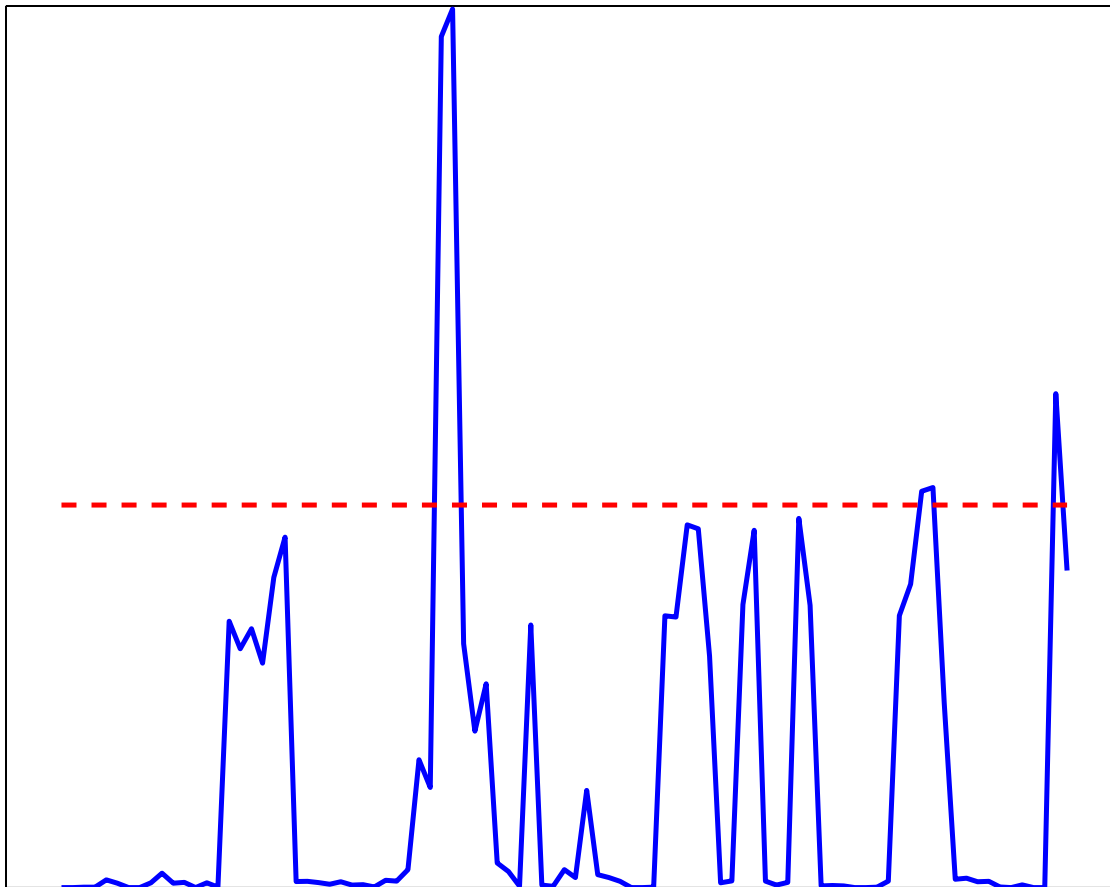
Simulation of recombination



True mosaic structure



Example: TOPAL, window size=100

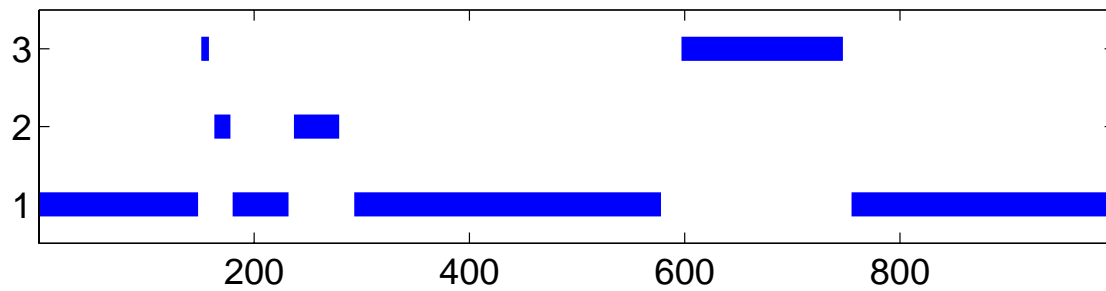
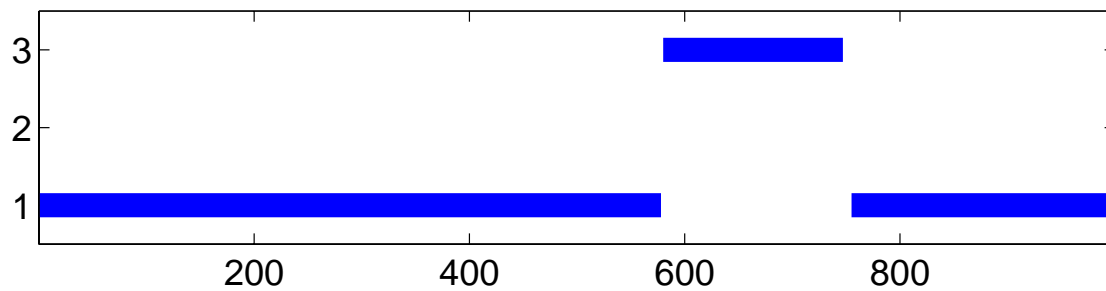
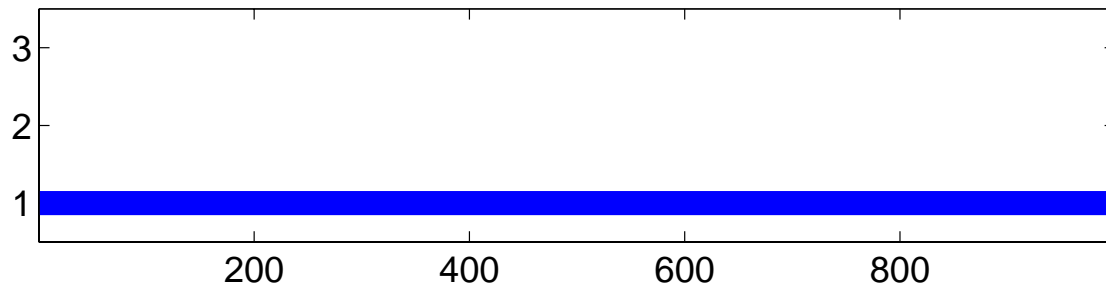


Prediction with RECPARS

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

Vertical axis :

$$P(S_t|\mathcal{D})$$

Horizontal axis :

Site in the DNA sequence alignment, t

Prediction with HMM-Bayes

Vertical axis :

$$P(S_t|\mathcal{D})$$

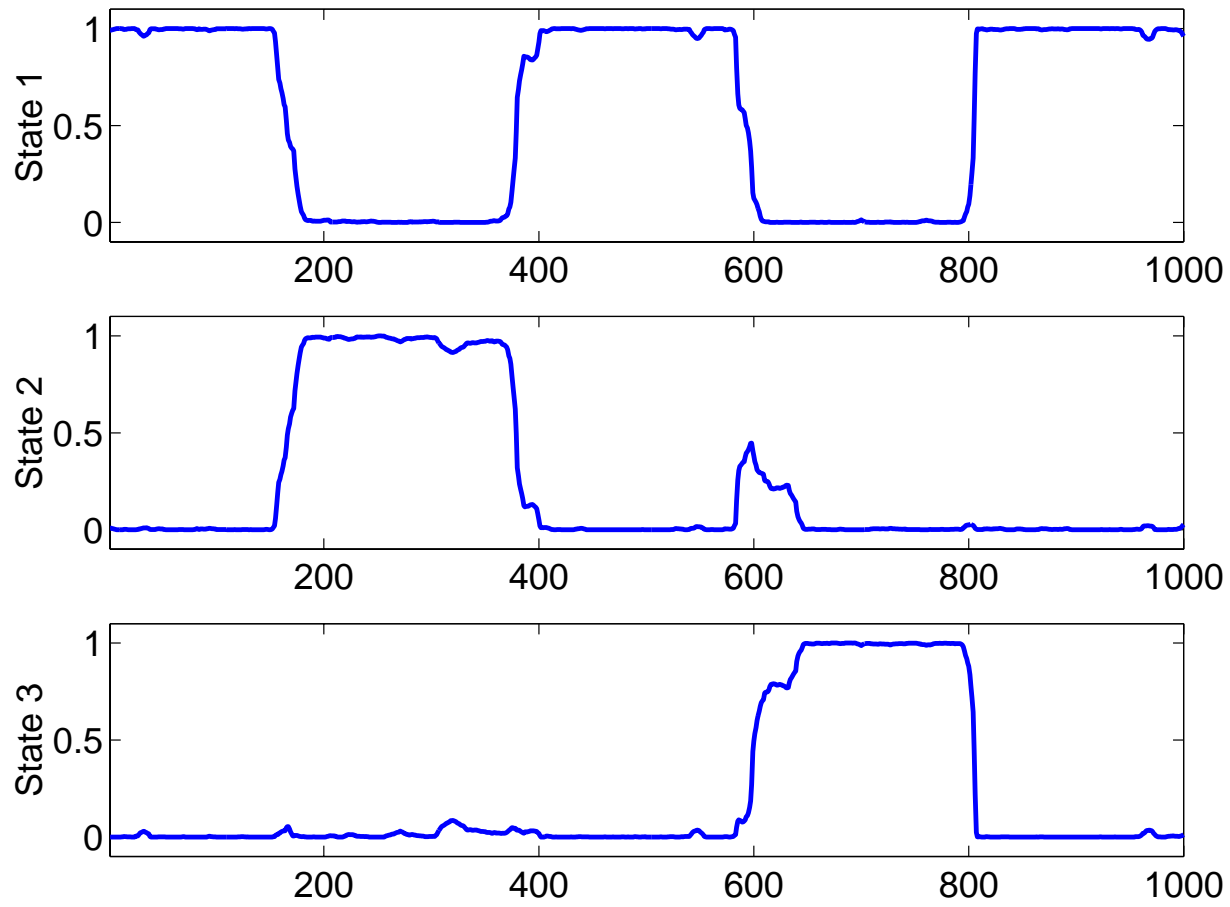
Horizontal axis :

Site in the DNA sequence alignment, t

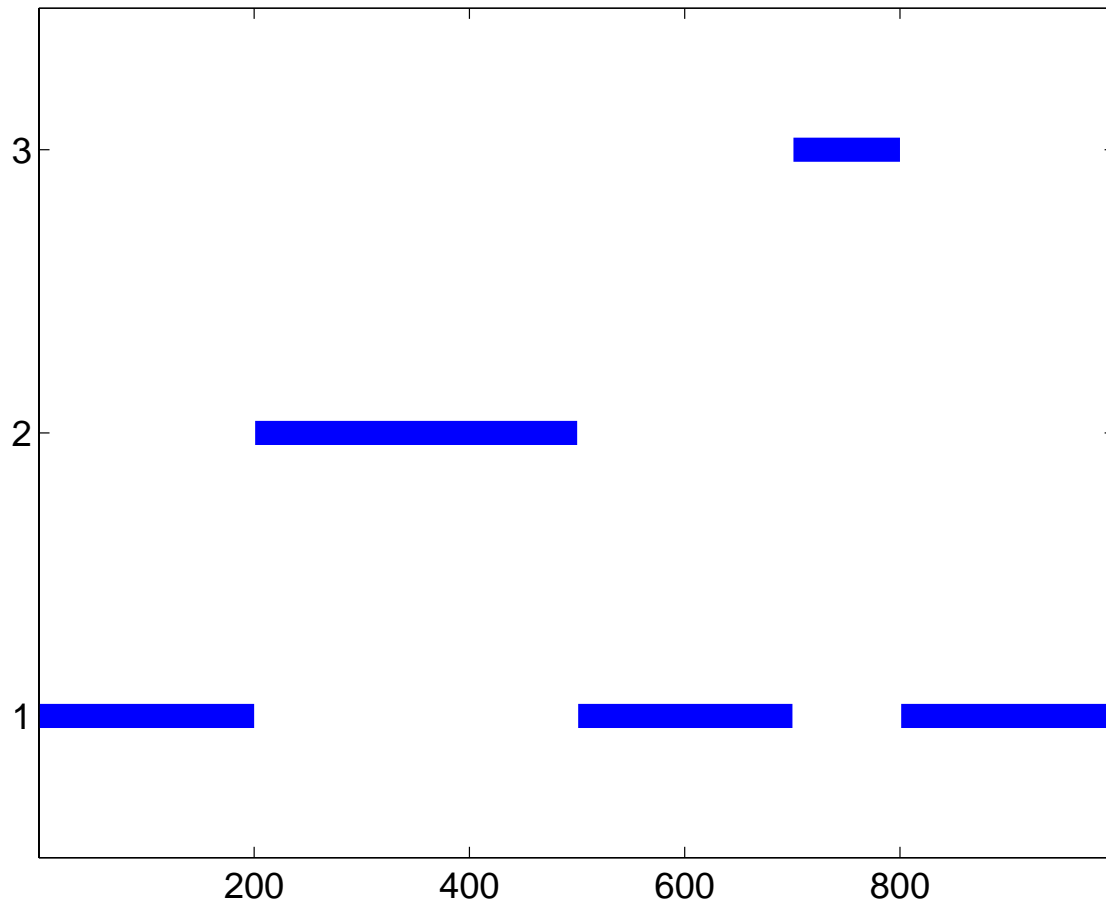
Three graphs :

$$P(S_t = 1|\mathcal{D}), P(S_t = 2|\mathcal{D}), P(S_t = 3|\mathcal{D})$$

Prediction with HMM-Bayes



True mosaic structure

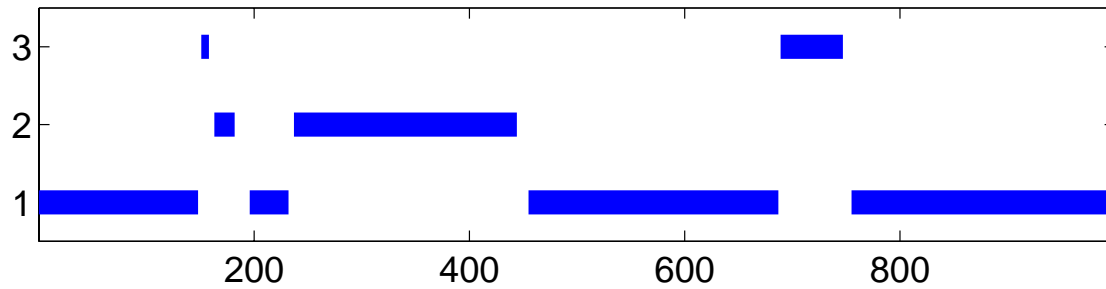
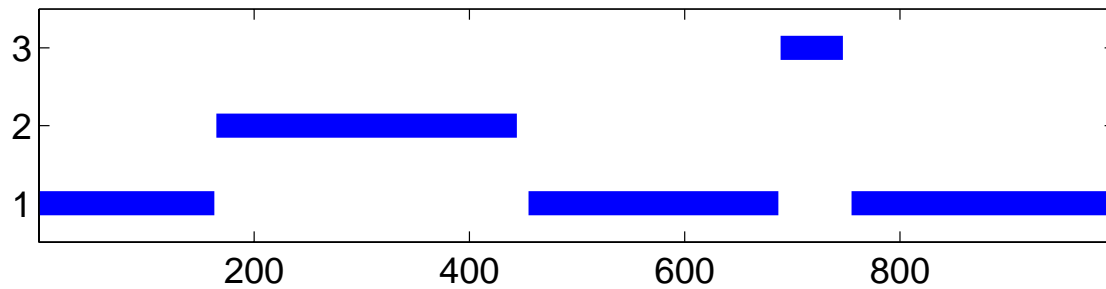
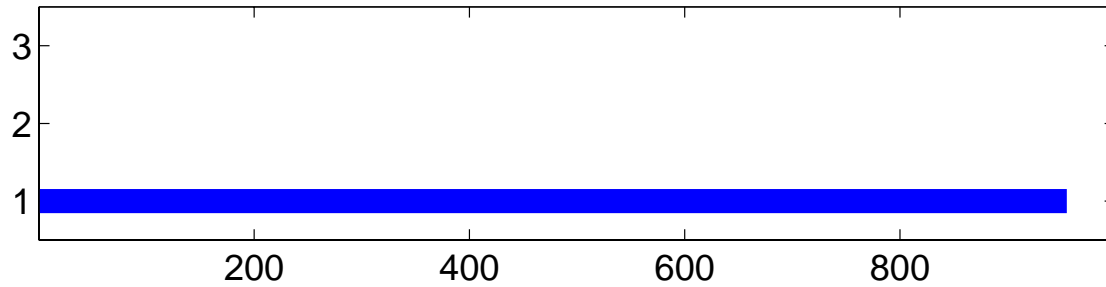


Prediction with RECPARS

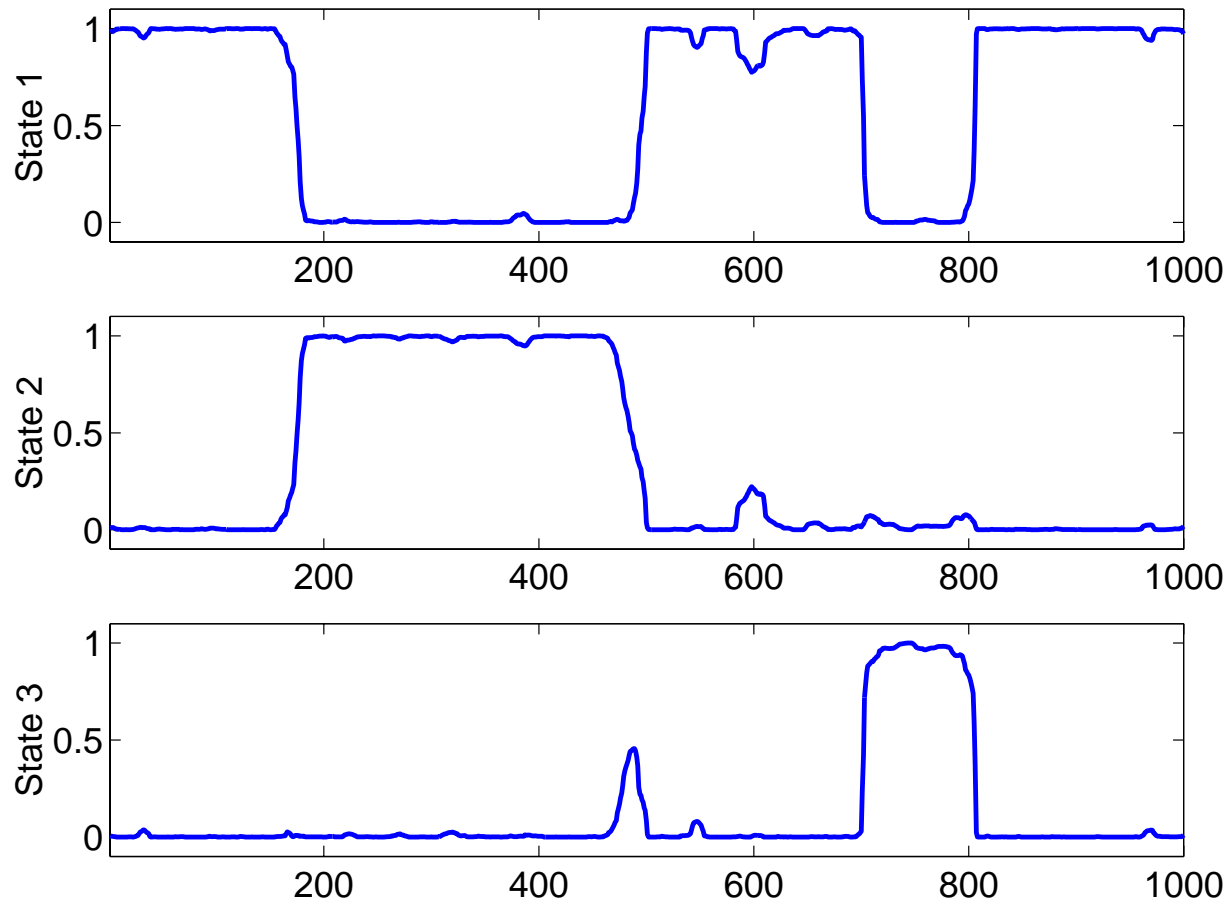
Top: $C_{recomb}/C_{mut} = 10.0$

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

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

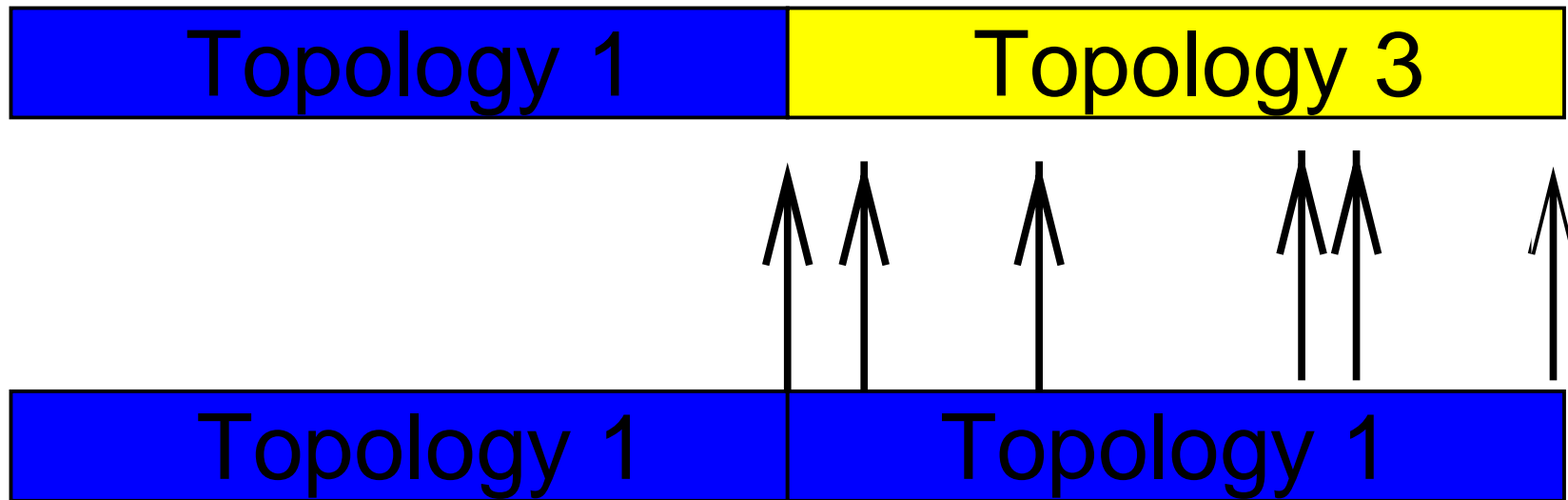


Prediction with HMM-Bayes

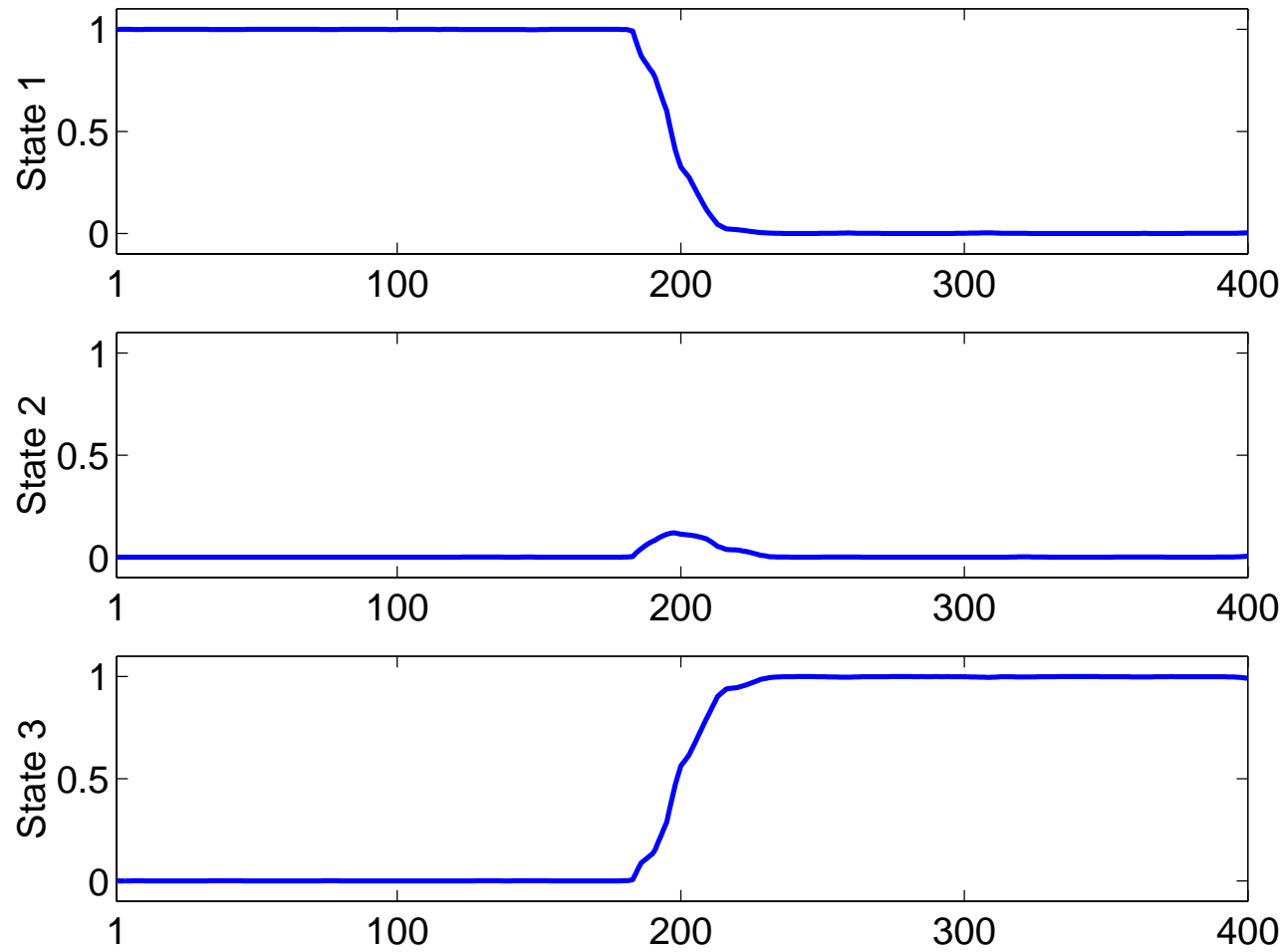


Comparison between HMM-ML and HMM-Bayes

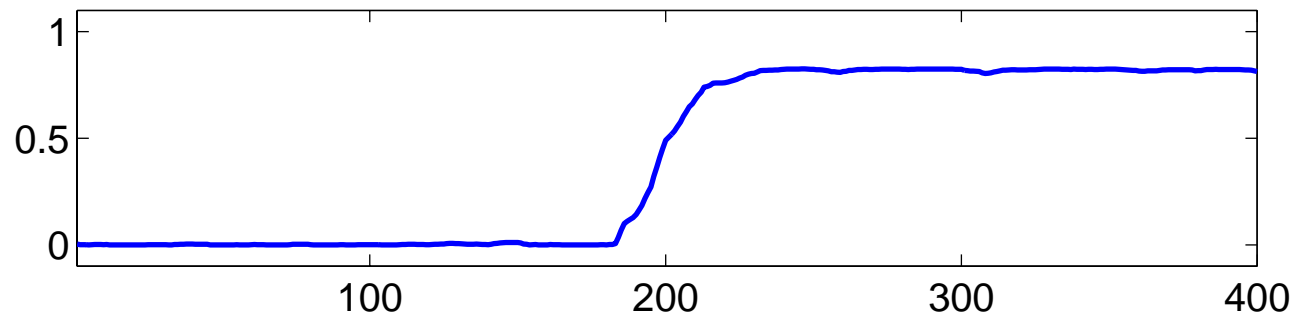
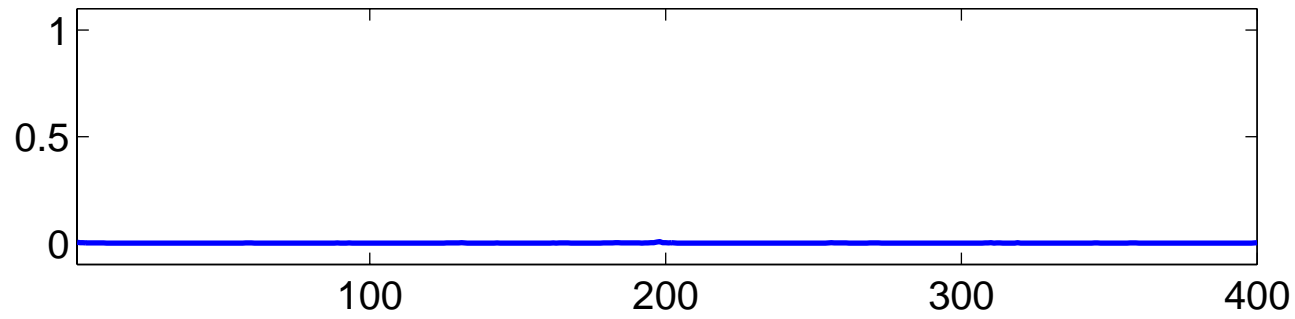
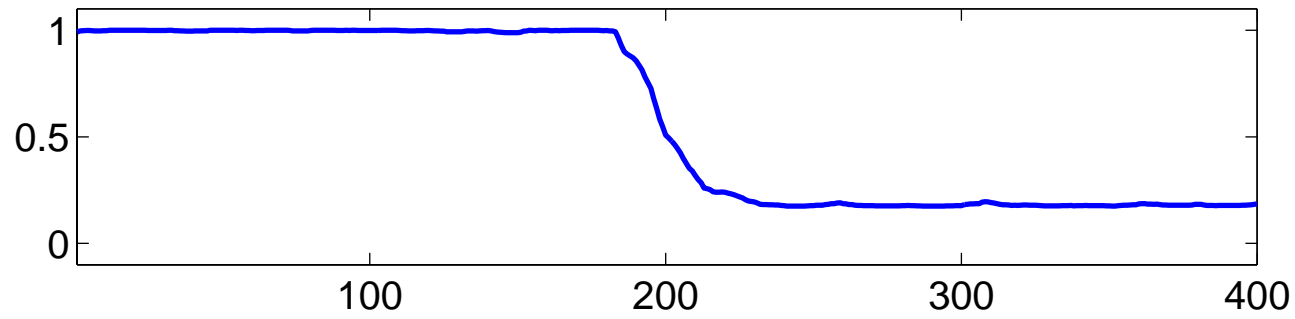
Comparison between HMM-ML and HMM-Bayes



HMM-ML



HMM-Bayes



Hepatitis B Virus (Bollyky et al. 1995)

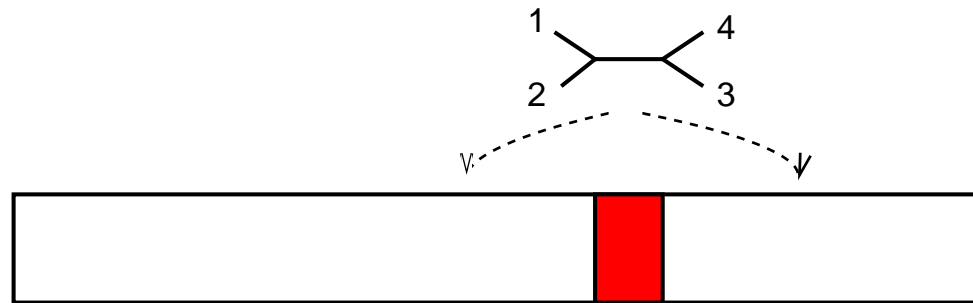
DNA alignment, 3049 nucleotides

1) HPBADW1

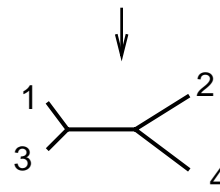
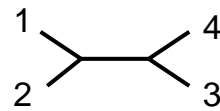
2) HPBADW2

3) HPBADWZCG

4) HPBADRC



State 1



State 2

TOPAL, window size = 100

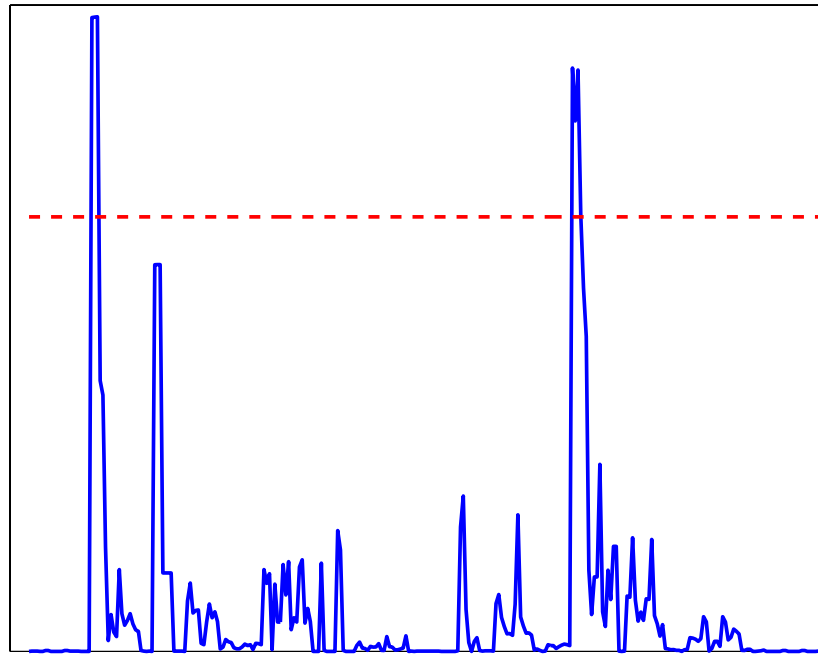
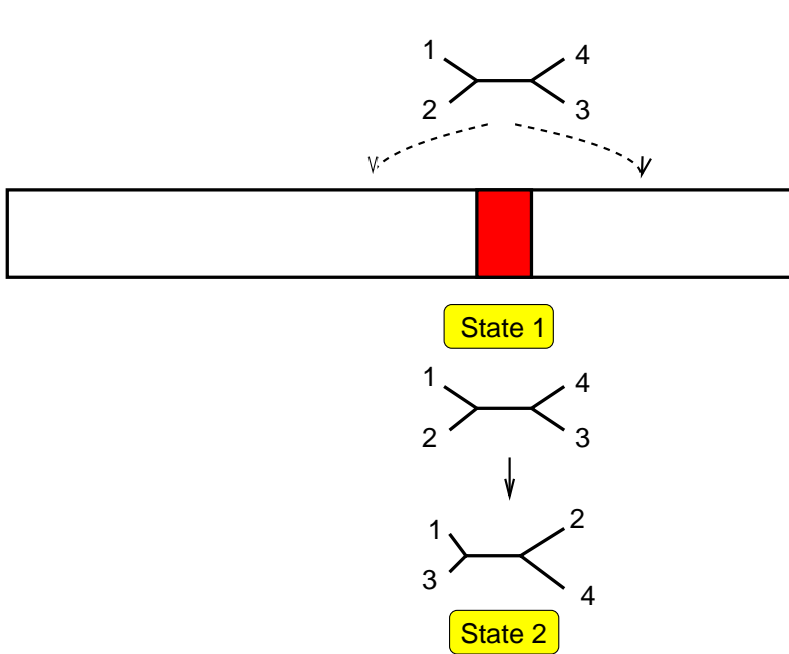
DNA alignment, 3049 nucleotides

1) HPBADW1

2) HPBADW2

3) HPBADWZCG

4) HPBADRC



TOPAL, window size = 200

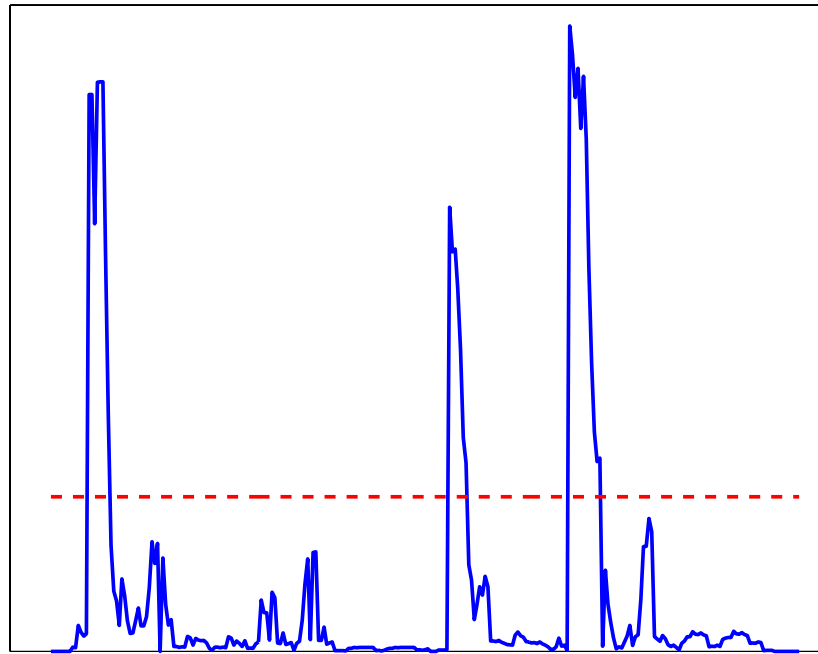
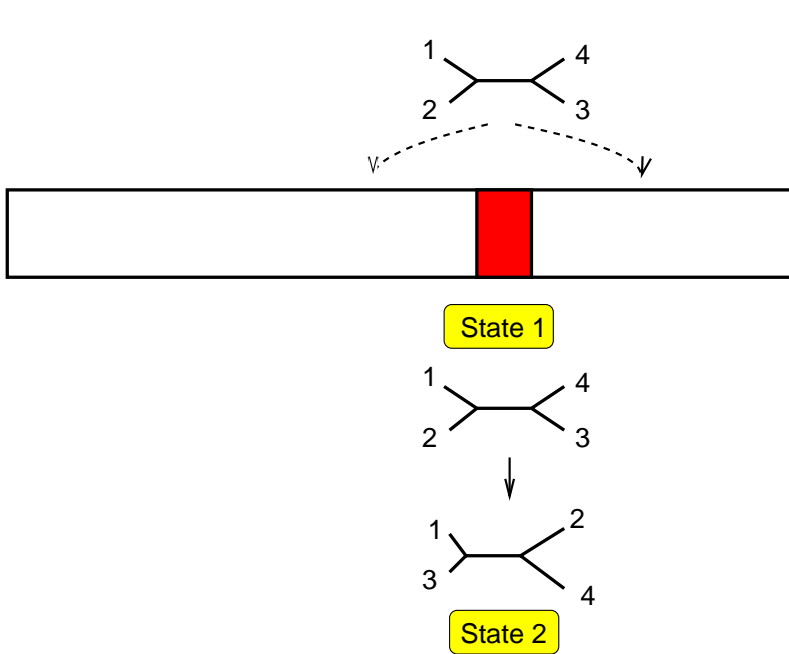
DNA alignment, 3049 nucleotides

1) HPBADW1

2) HPBADW2

3) HPBADWZCG

4) HPBADRC



RECPARS

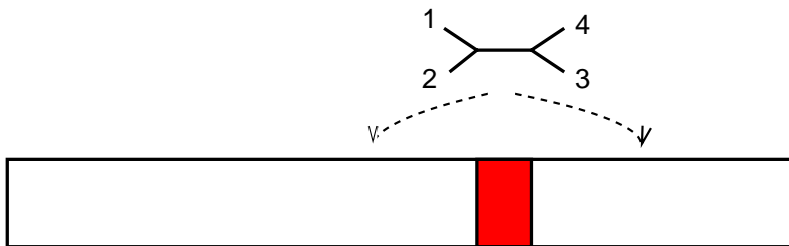
DNA alignment, 3049 nucleotides

1) HPBADW1

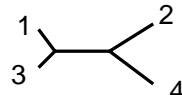
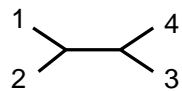
2) HPBADW2

3) HPBADWZCG

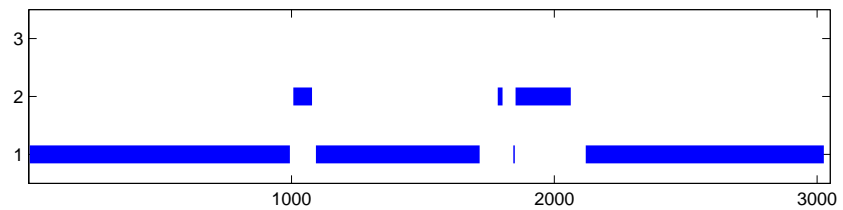
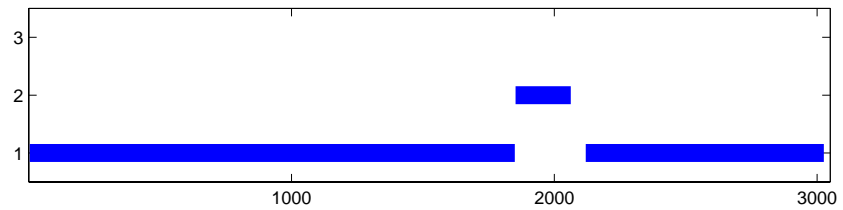
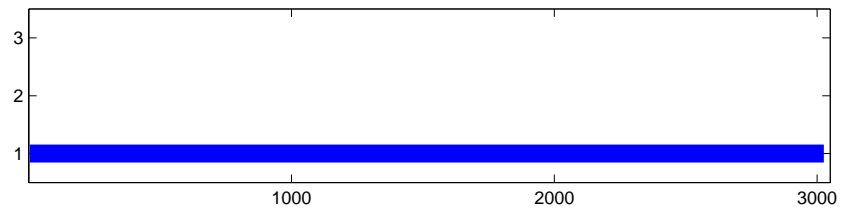
4) HPBADRC



State 1



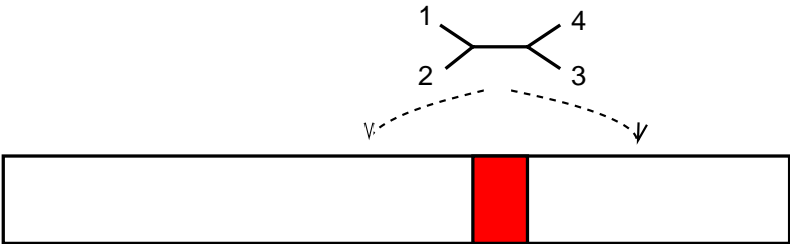
State 2



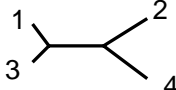
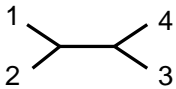
HMM-Bayes

DNA alignment, 3049 nucleotides

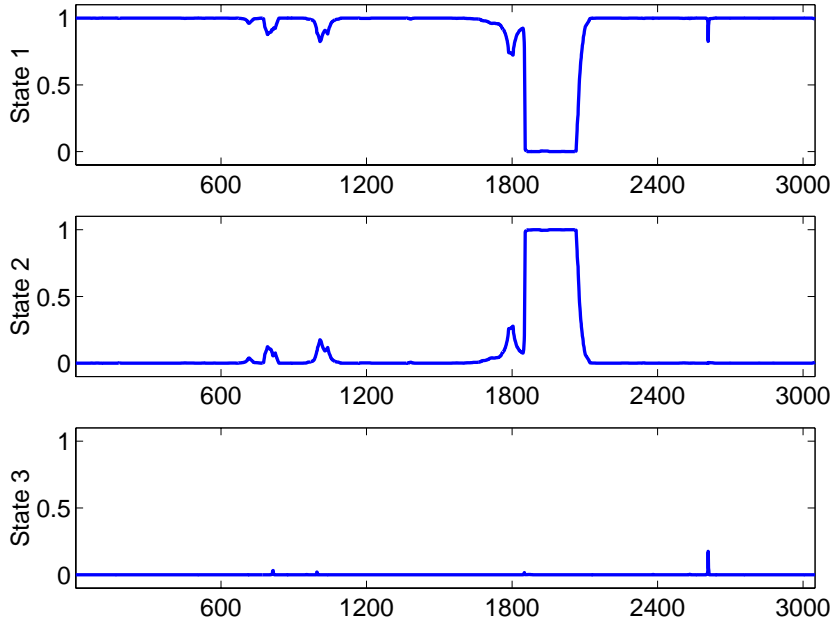
- 1) HPBADW1
- 2) HPBADW2
- 3) HPBADWZCG
- 4) HPBADRC



State 1



State 2



Neisseria (Zhou & Spratt, 1992)

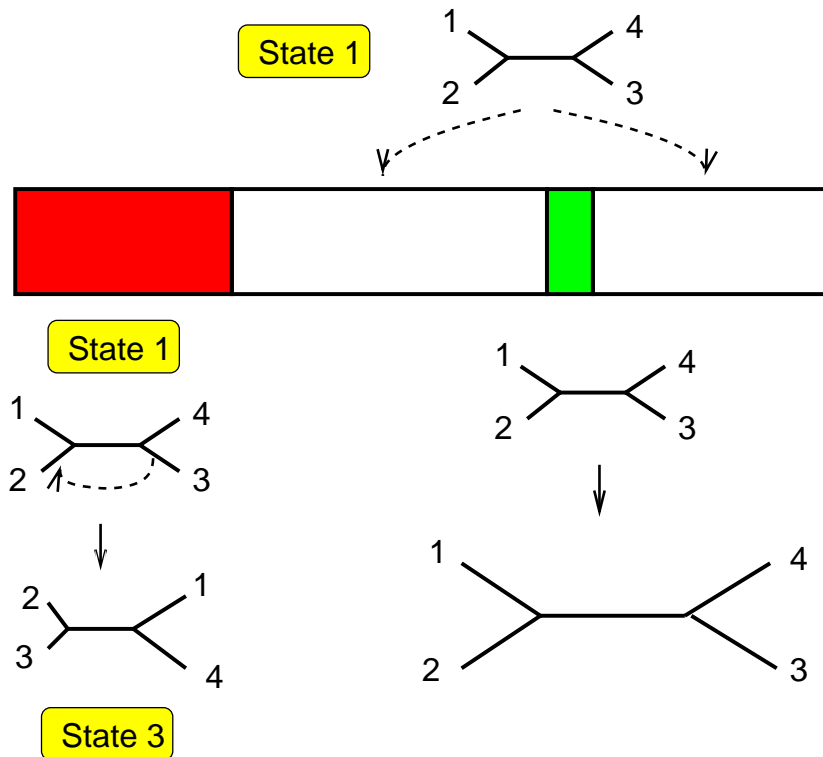
DNA alignment, 787 nucleotides (argF gene)

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

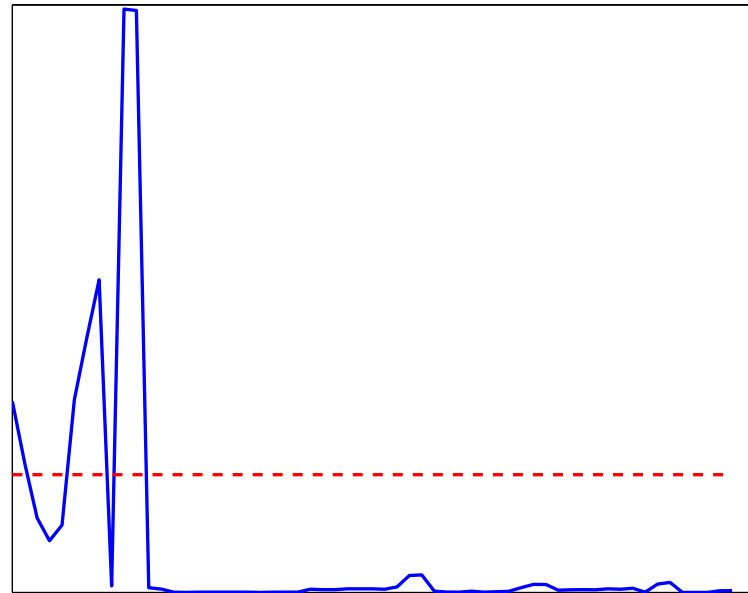
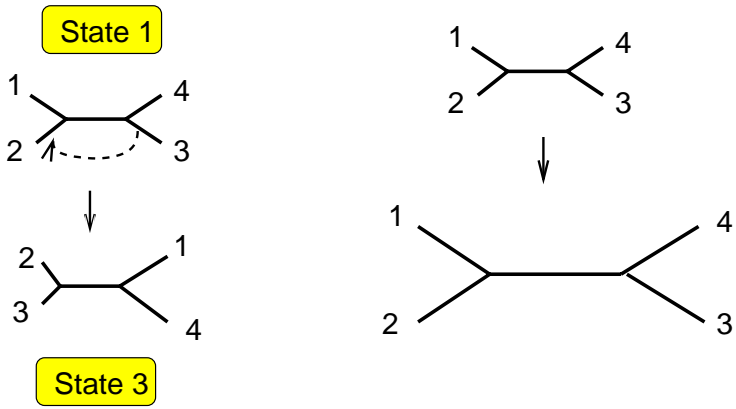
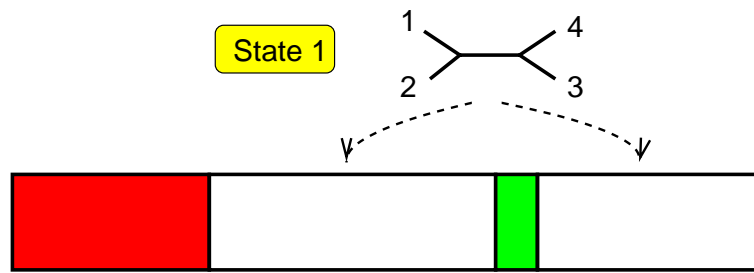
Neisseria (Zhou & Spratt, 1992)

DNA alignment, 787 nucleotides (argF gene)

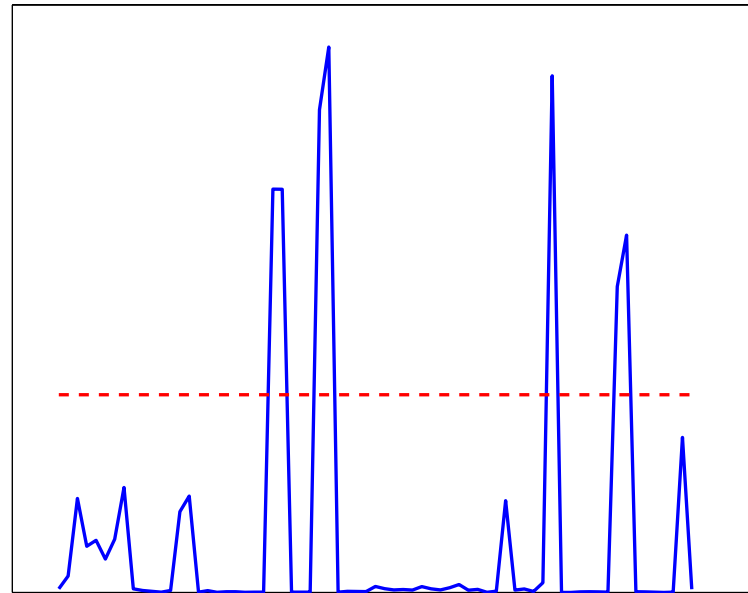
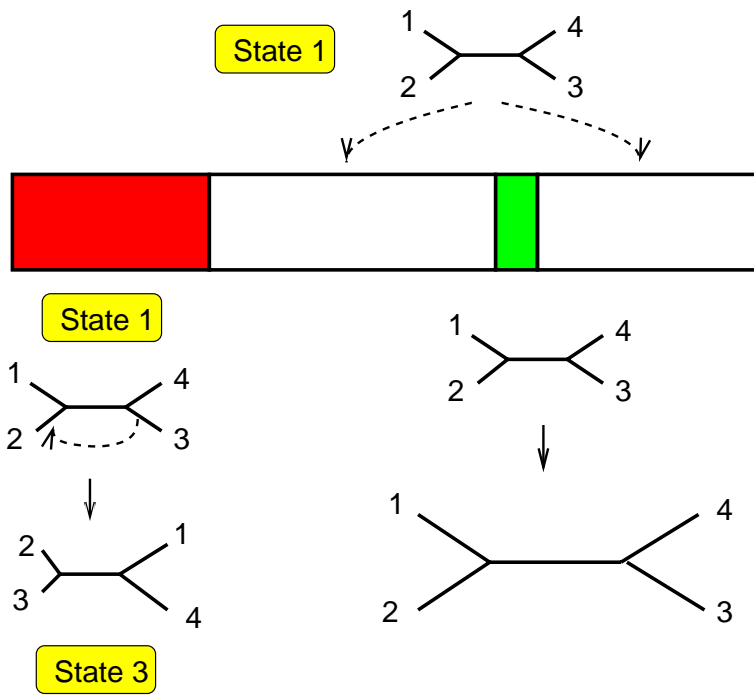
- 1) *Neisseria gonorrhoeae*
- 2) *Neisseria meningitidis*
- 3) *Neisseria cinerea*
- 4) *Neisseria mucosa*



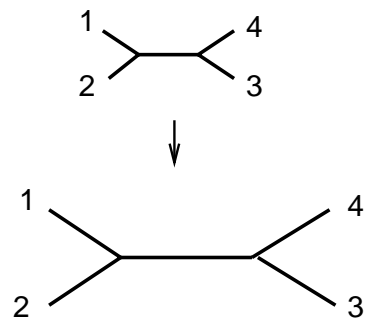
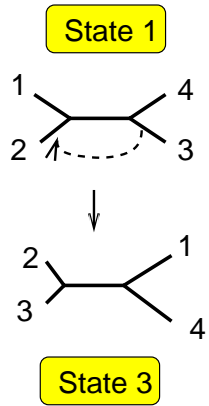
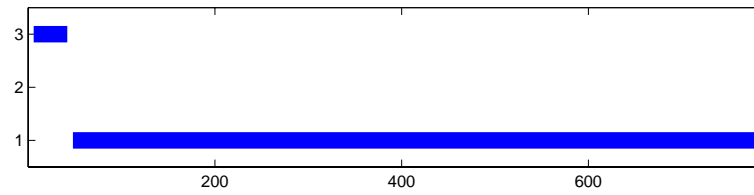
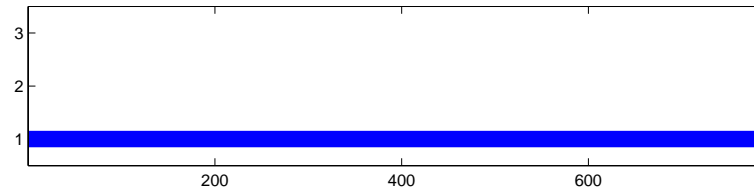
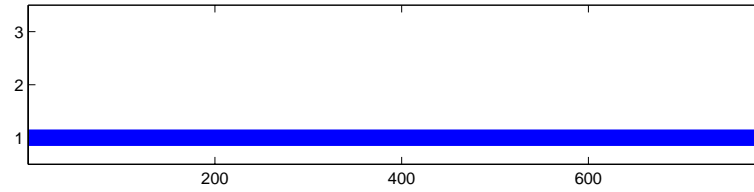
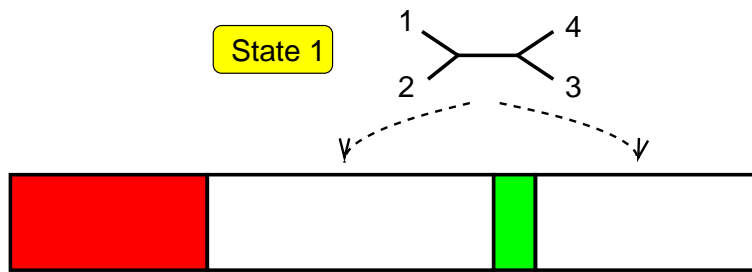
Topal, window size 200



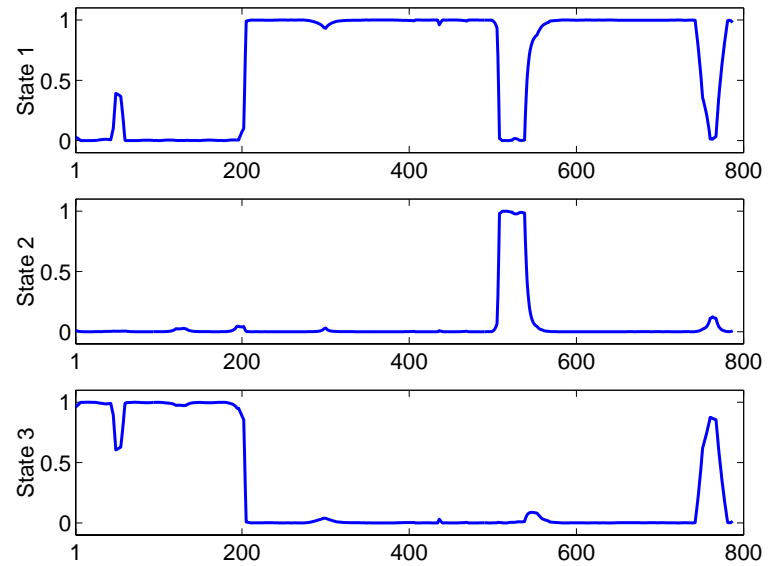
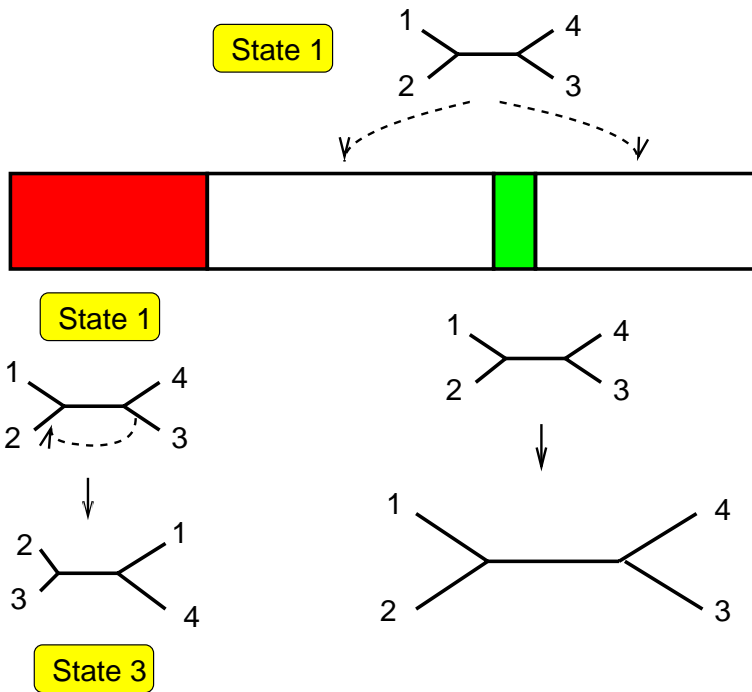
Topal, window size 100



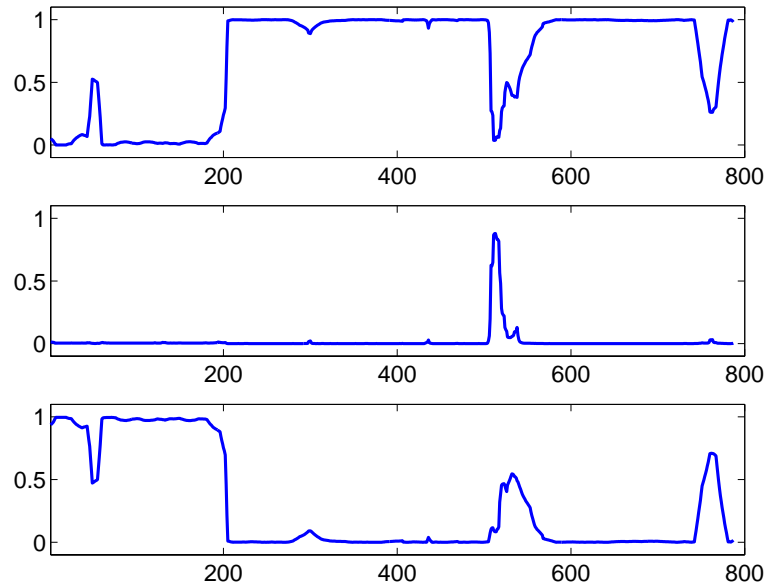
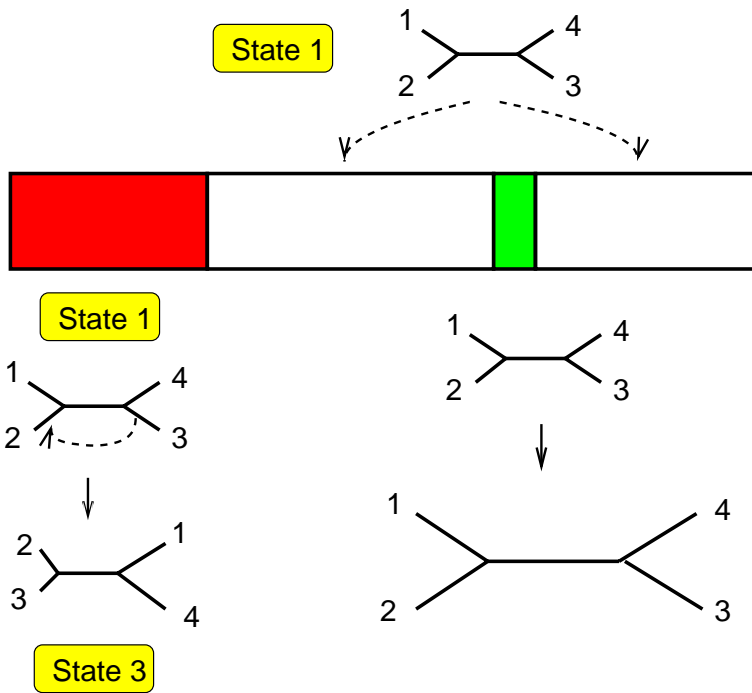
RECPARS



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



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



Conclusion and future work

Conclusion and future work

- Parameters inferred from the data

Conclusion and future work

- Parameters inferred from the data
- Precise location of breakpoints.

Conclusion and future work

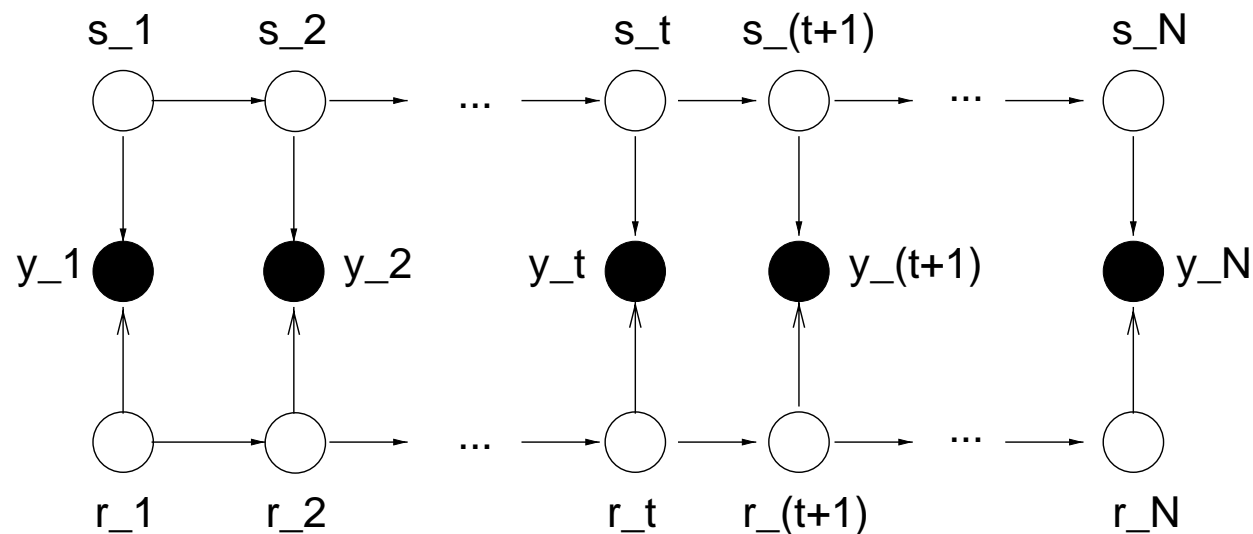
- Parameters inferred from the data
- Precise location of breakpoints.
- Limited in the number of different tree topologies.

Conclusion and future work

- Parameters inferred from the data
 - Precise location of breakpoints.
 - Limited in the number of different tree topologies.
 - Problem: rate heterogeneity.
-

Conclusion and future work

- Parameters inferred from the data
- 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
