
Detecting mosaic structures in DNA sequence alignments with phylogenetic factorial hidden Markov models

Dirk Husmeier

Biomathematics and Statistics Scotland

Edinburgh, United Kingdom

Email: dirk@bioss.ac.uk

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

New Scientist, 8 October 2005

New Scientist, 8 October 2005

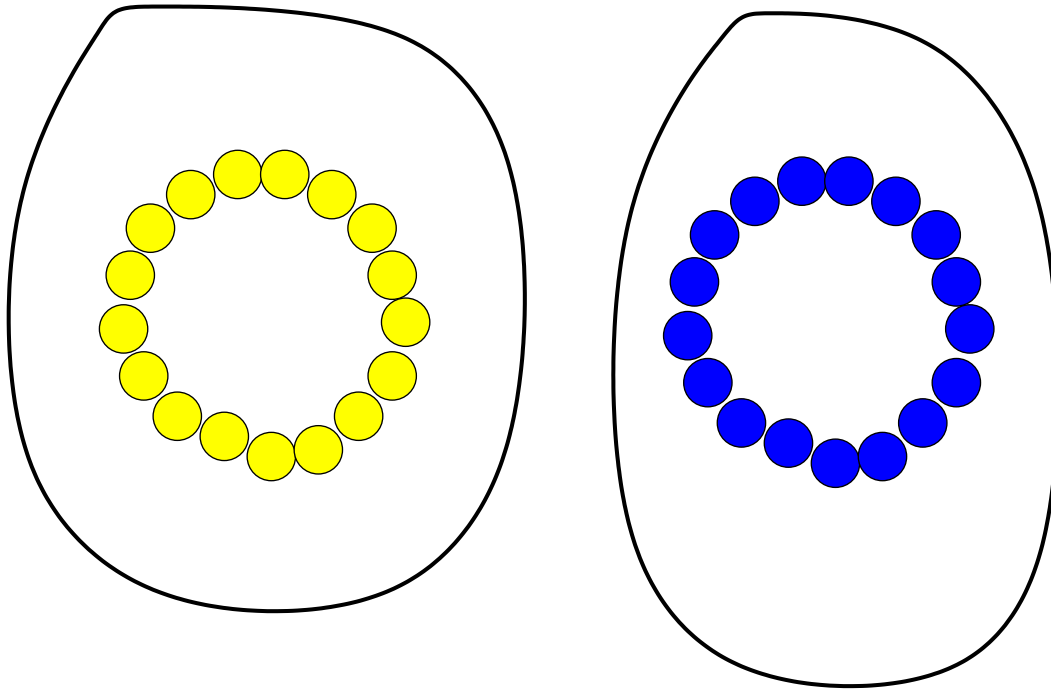
New **biological weapons** are being created right this minute - though not in any secret laboratory or military base. Out of reach of international legislation, there is a **genetic lending library of evil** in action, and while you may not realise it, **you are intimately involved**. It's been going on for millions of years, but we have only just begun to explore this sinister new territory.

Zhou & Spratt, 1992

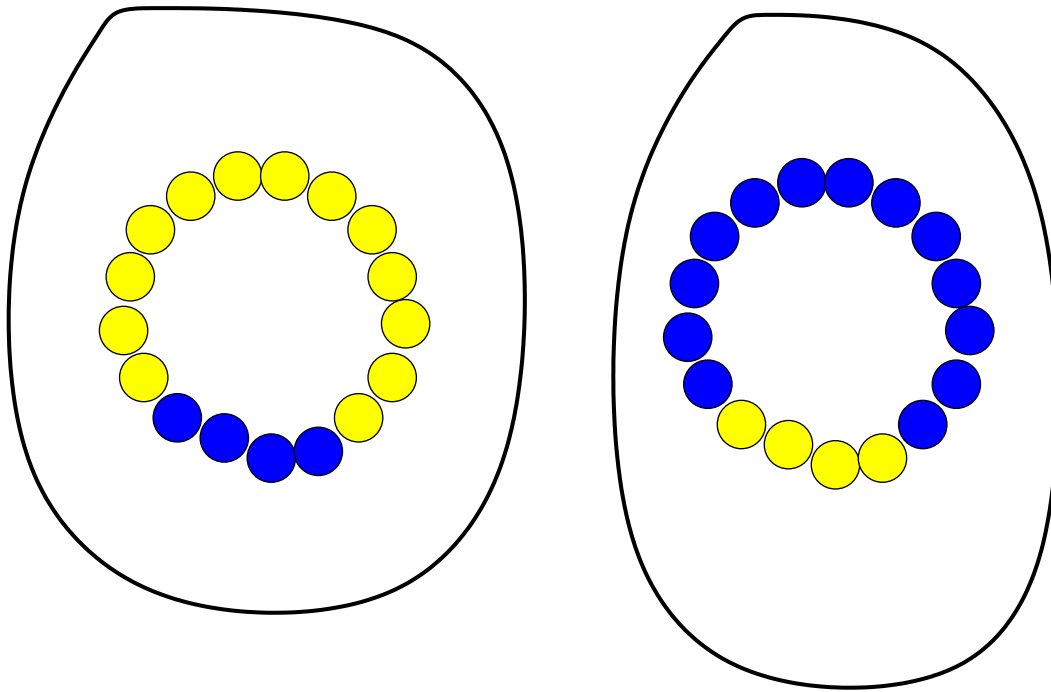
DNA alignment, 787 nucleotides (argF gene)

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

Recombination



Recombination



Zhou & Spratt, 1992

DNA alignment, 787 nucleotides (argF gene)

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

Pathogenicity and antibiotic resistance

1995

Robertson, Sharp, McCutchan, Hahn

Recombination in HIV-1

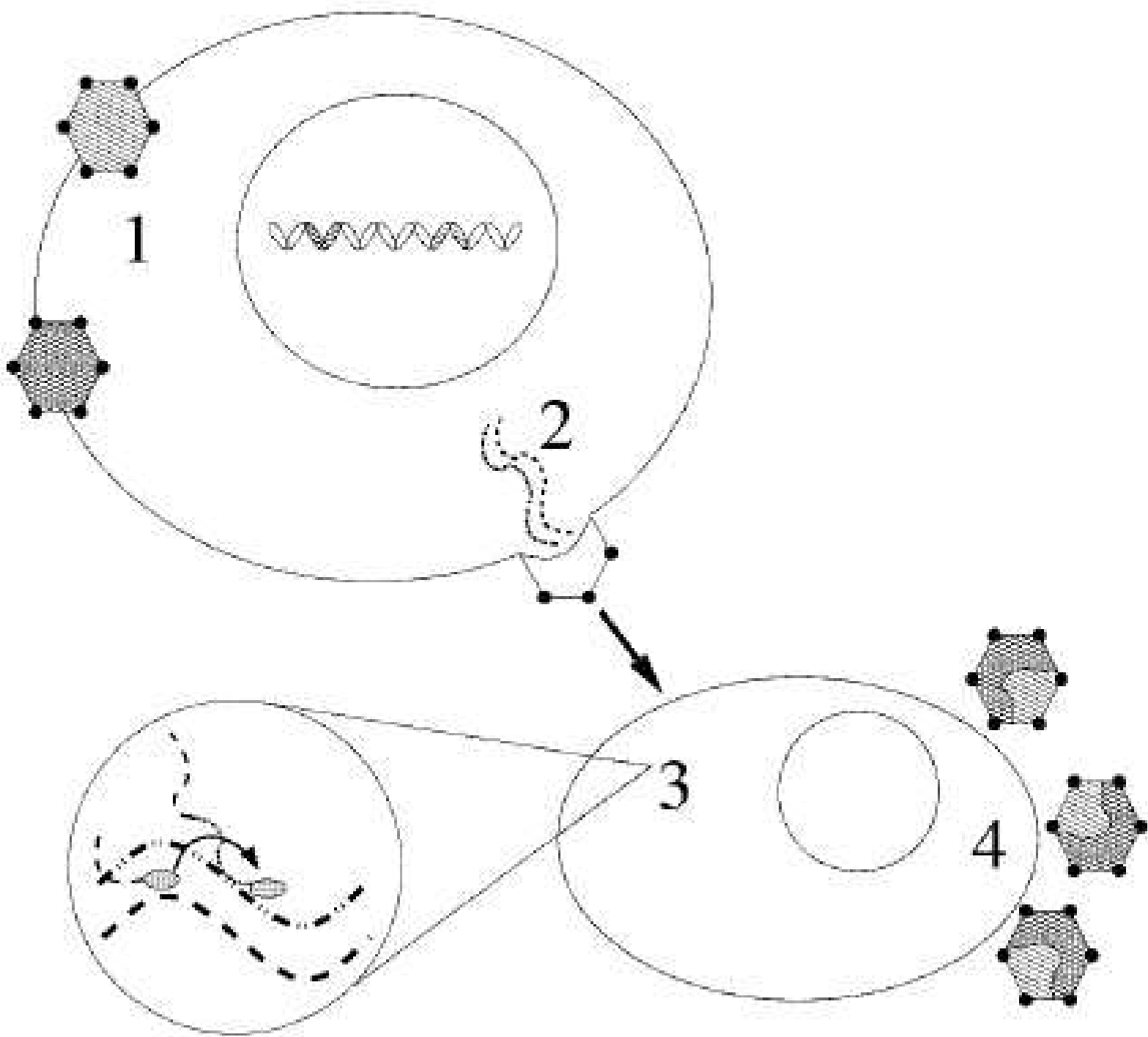
Nature 374, pp.124-126

1997

Dennis Blakeslee

Recombination in HIV: A fast track to resistance?

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



HIV-1 strain KAL 153

Caused epidemic outbreak of HIV-1 infection among intravenous drug users around Kaliningrad , Russia, in October 1996.

Rate of newly diagnosed seropositive individuals:
Less than 1/month → over 100/month

KAL-153: Recombination of subtypes A and B.

New Scientist, 8 October 2005

What is the **pathosphere**? It is the surprisingly vast and growing **gene pool** in which pathogens, the microbes that cause disease, **meet and mingle**. Scientists have long known that the so-called plastic genome of many pathogens allows them to readily swap genes, **transferring genetic material** in **information packages** called plasmids. Now it seems they have access to a much broader supply of genes than previously believed.

Systems Biology

Pal, Papp and Lercher (2005)

Horizontal gene transfer depends on gene content of the host
Bioinformatics 21, Suppl 2 (ECCB 05)

- Horizontal gene transfer is a major contributor to the evolution of bacterial genomes.
- The chance of acquiring a gene by horizontal gene transfer is up to six times higher if an **enzyme** that **catalyses** a **coupled metabolite** flux is already encoded in the host genome.

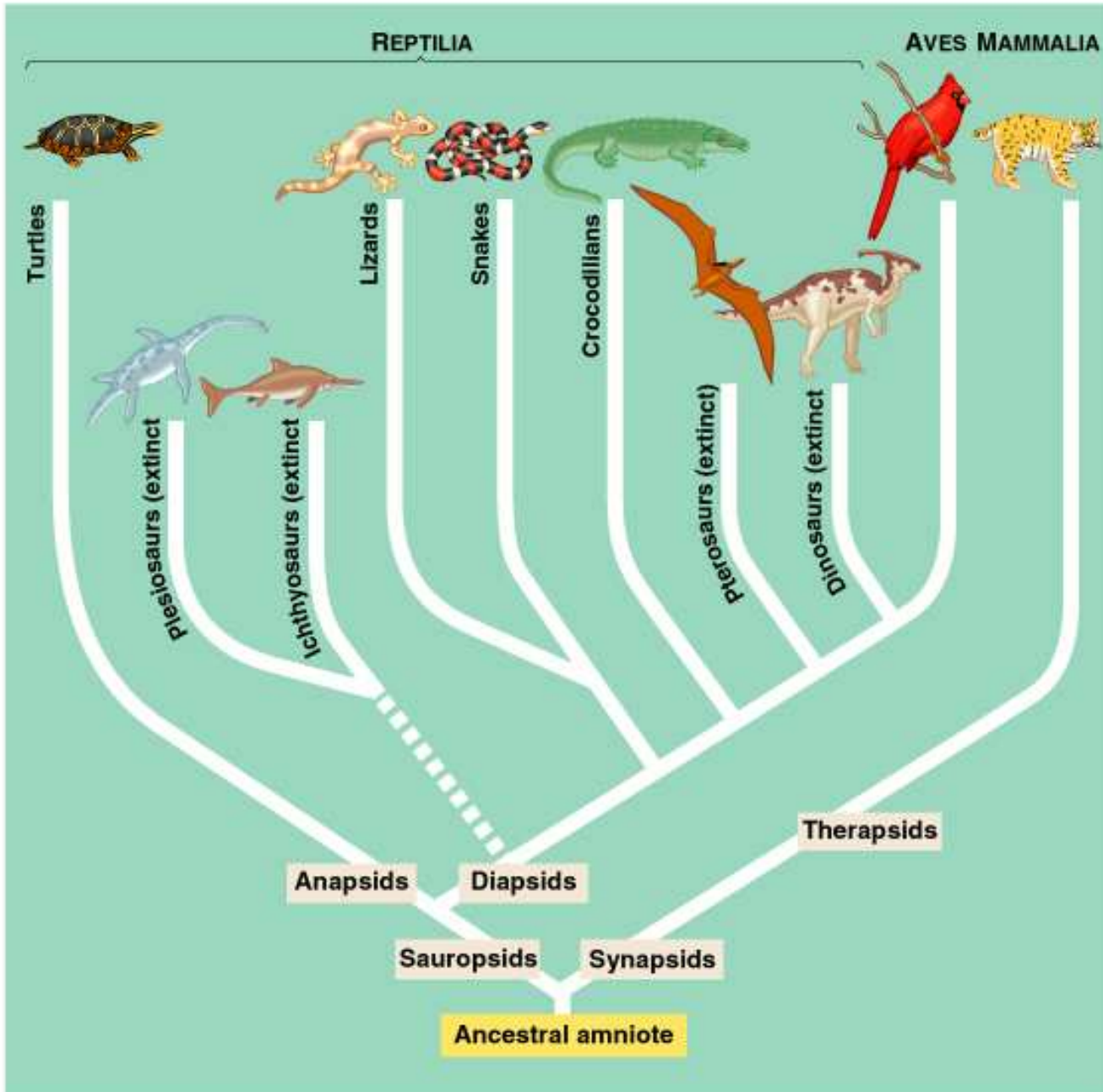
Overview

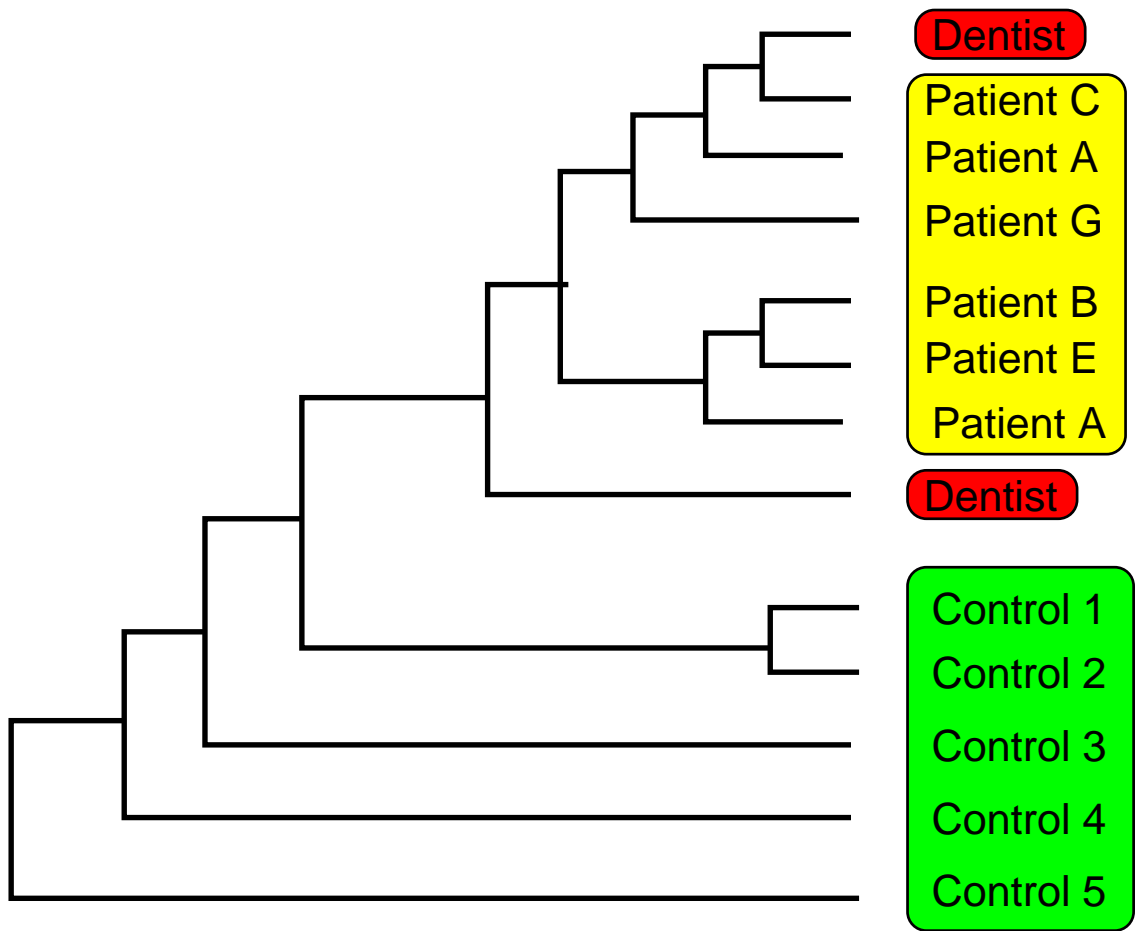
- Introduction: [Phylogenetics](#)
- [Detecting recombination: Phylogenetic HMMs](#)
Dirk Husmeier, Frank Wright and Grainne McGuire
2001-2003
- [Distinguishing between recombination and rate variation: Phylogenetic FHMMs](#)
Dirk Husmeier, 2005
- [Learning the number of genomic regions under selective pressure: Phylogenetic FHMMs trained with RJMCMC](#)
Wolfgang Lehrach and Dirk Husmeier, 2006

Overview

- **Introduction: Phylogenetics**

- Detecting recombination: Phylogenetic HMMs
Dirk Husmeier, Frank Wright and Grainne McGuire
2001-2003
- Distinguishing between recombination and rate variation:
Phylogenetic FHMMs
Dirk Husmeier, 2005
- Learning the number of genomic regions
under selective pressure:
Phylogenetic FHMMs trained with RJMCMC
Wolfgang Lehrach and Dirk Husmeier, 2006





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

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

Systems Biology

Predicting protein-protein interactions

Barker, Pagel (2005)

Predicting functional gene links from phylogenetic–statistical
analyses of whole genomes

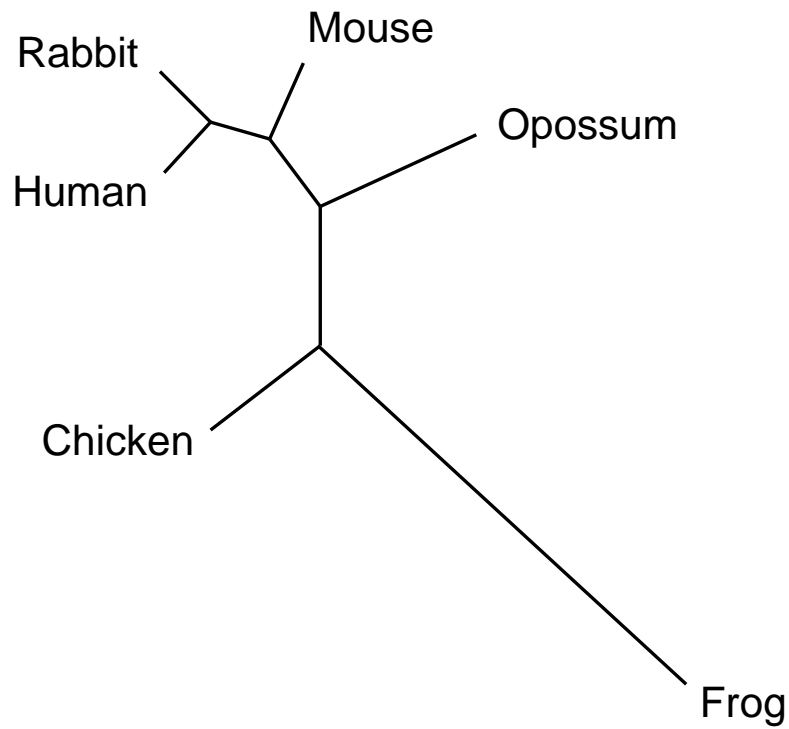
PLOS Computational Biology 1

Jothi, Kann, Przytycka (2005)

Predicting protein-protein interaction by searching
evolutionary tree automorphism space

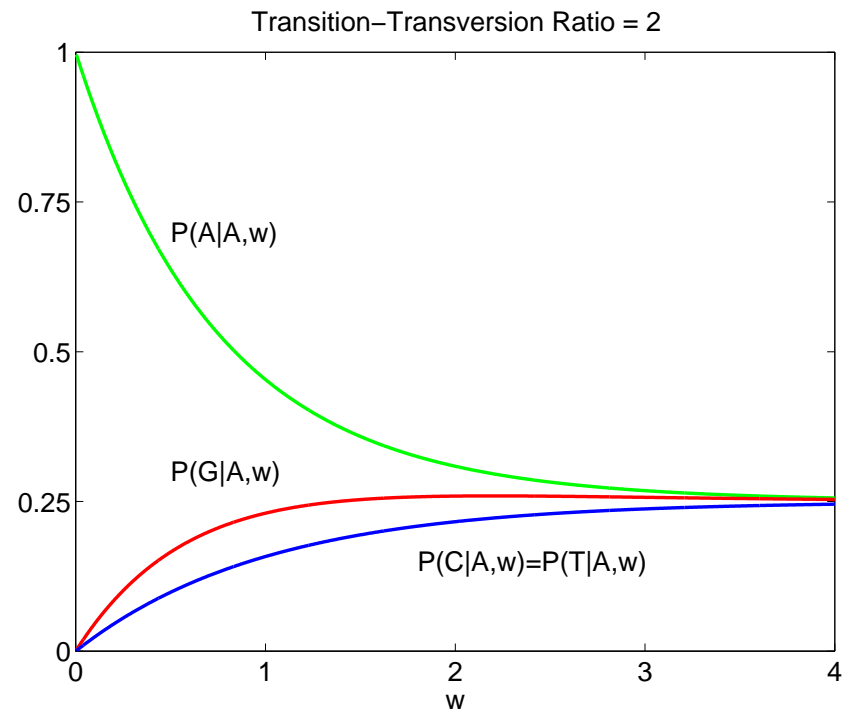
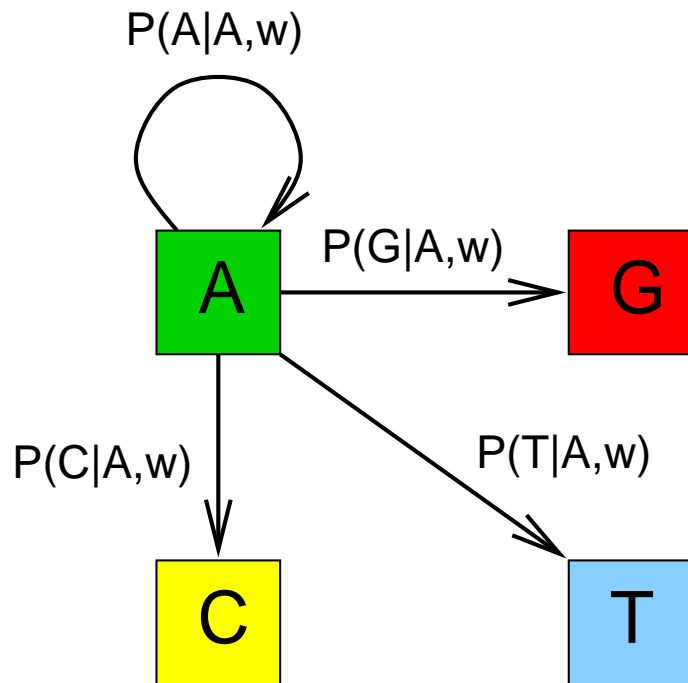
Bioinformatics 21, Suppl. 1 (ISMB 05)

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

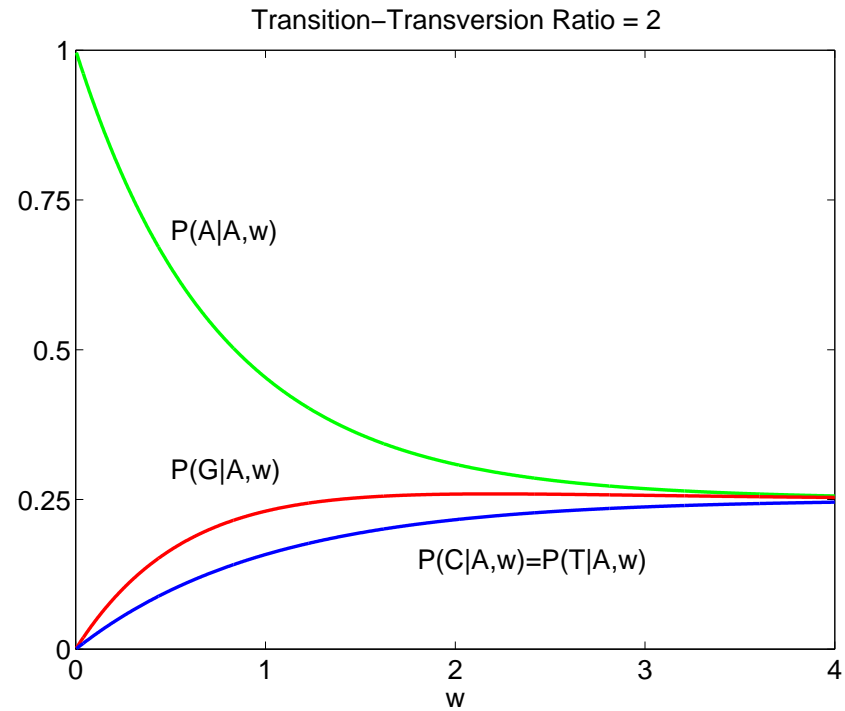
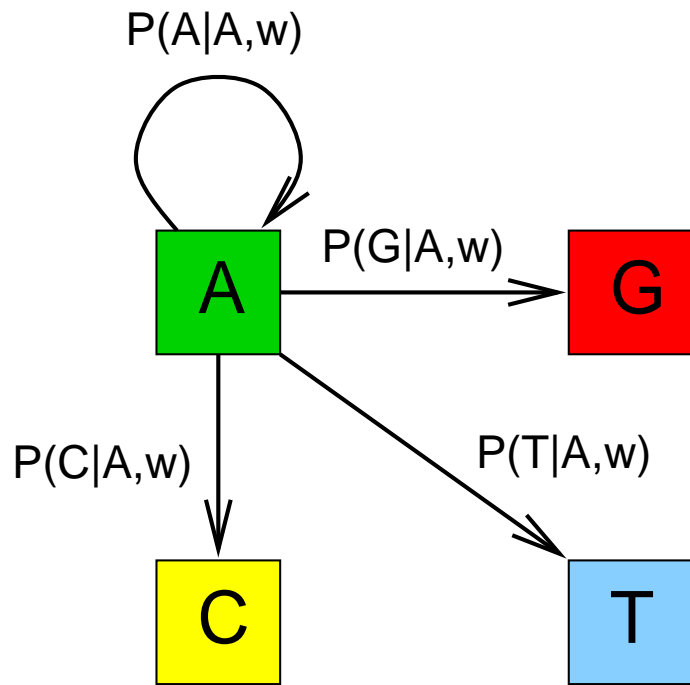


--> Topology
 --> Branch lengths

Mutation probabilities

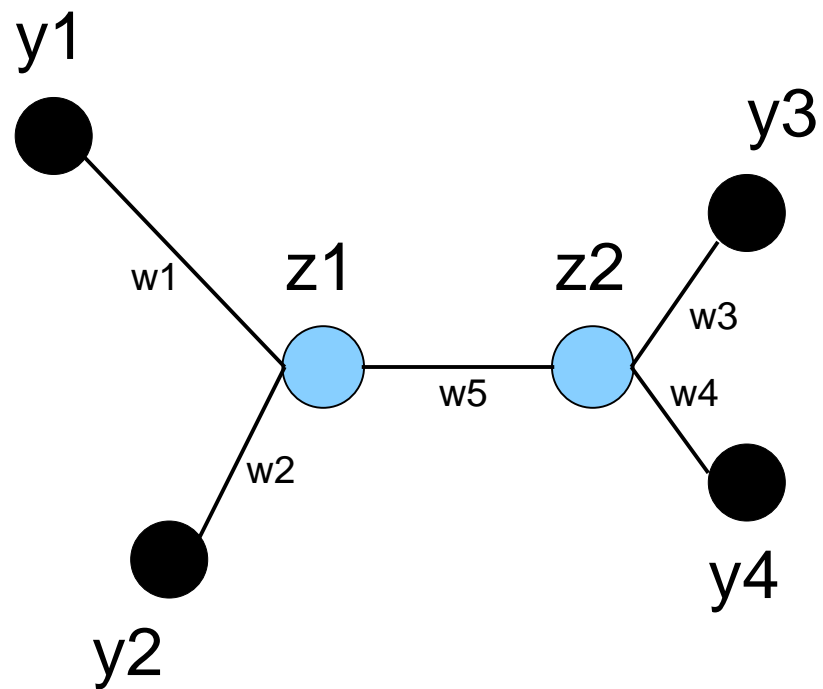


Mutation probabilities



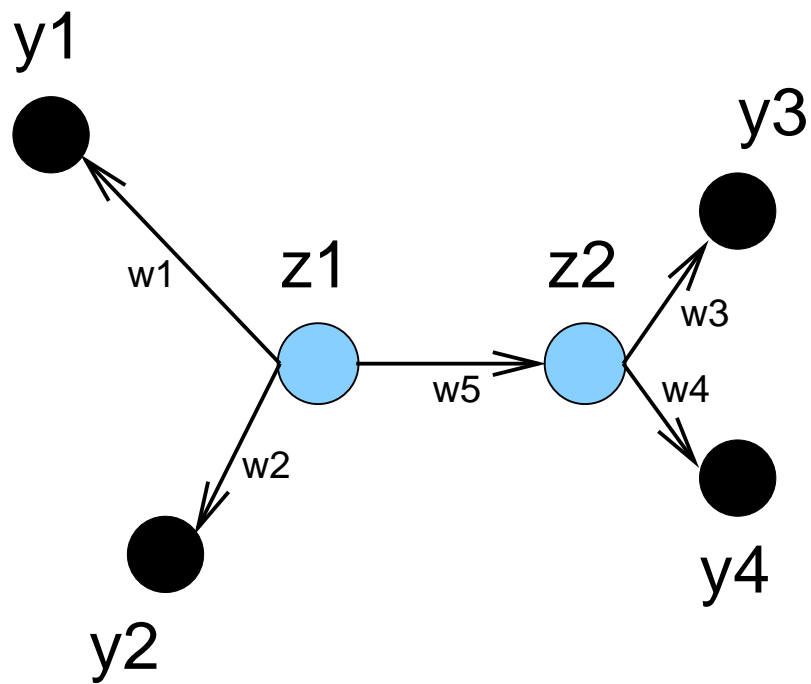
$w = \text{mutation rate} \times \text{time}$

Phylogenetic tree as an undirected graph



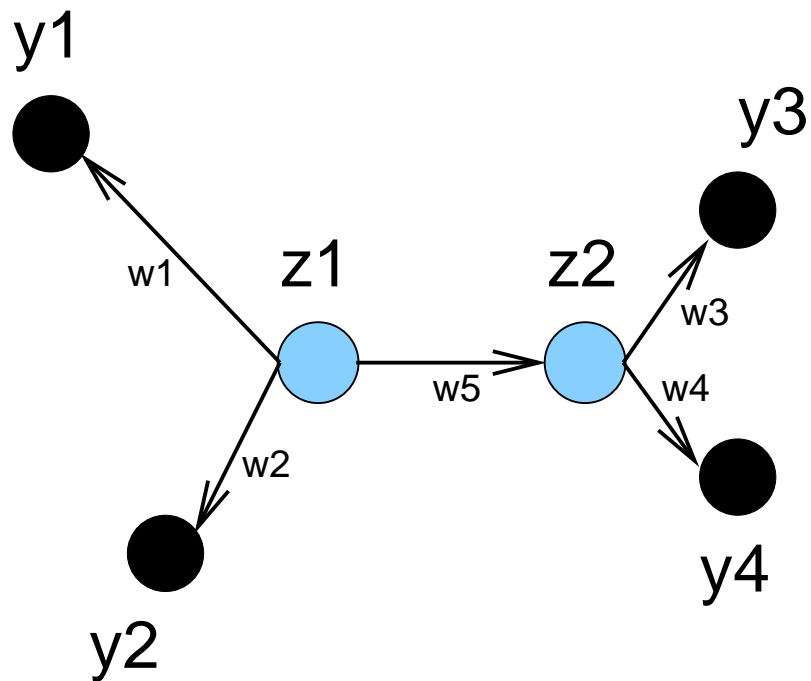
$$P(y_1, y_2, y_3, y_4, z_1, z_2 | \mathbf{w})$$

Phylogenetic tree as a directed graph



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

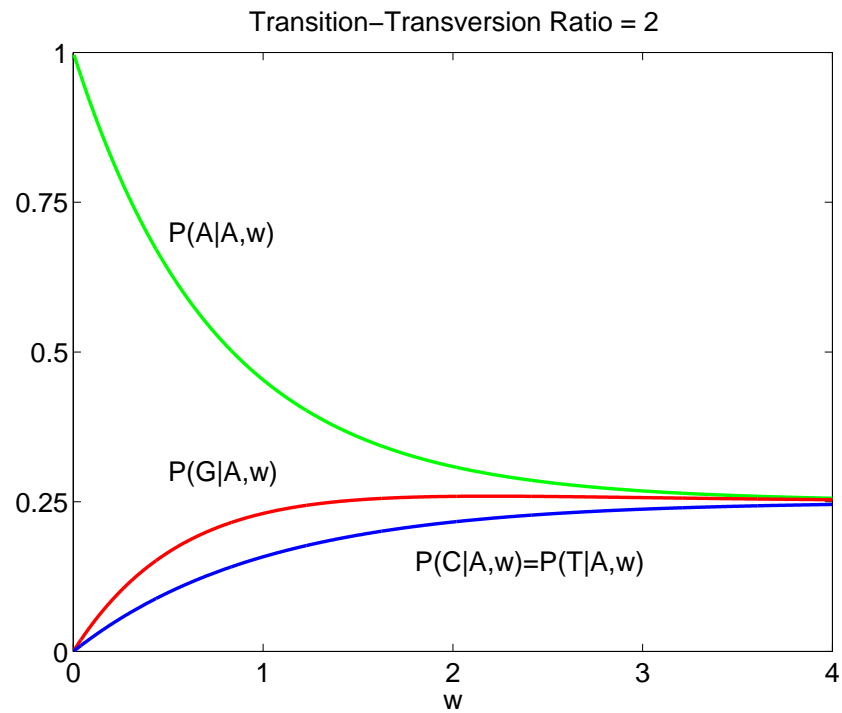
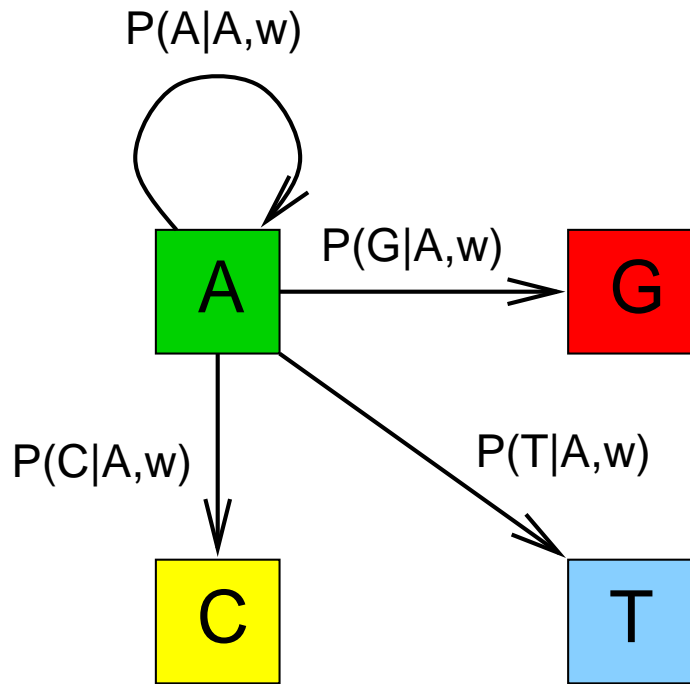
Phylogenetic tree as a directed graph



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

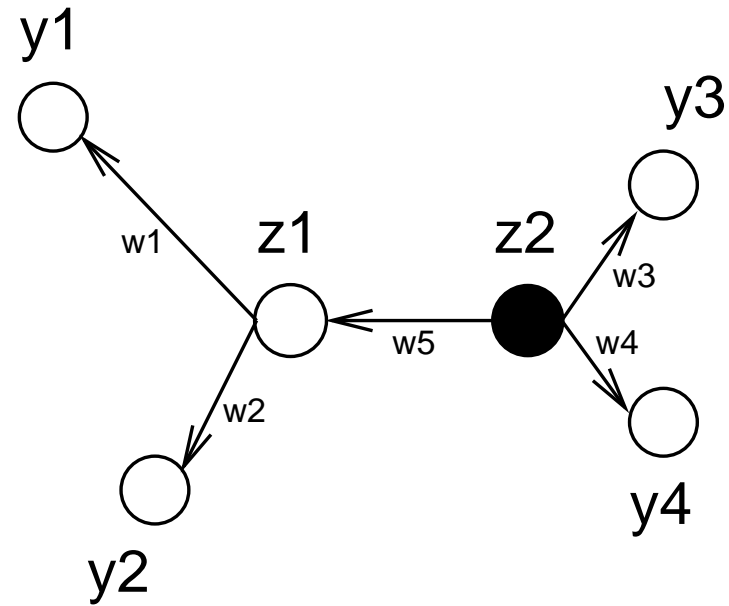
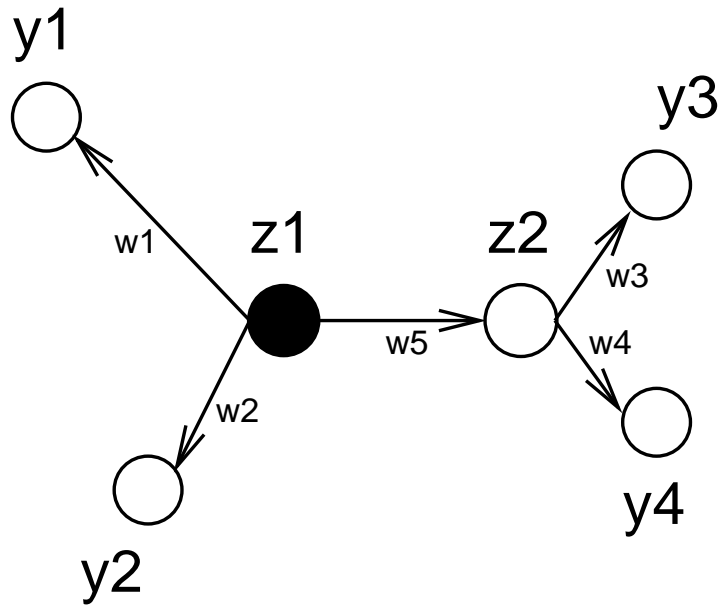
$$= P(y_1 | z_1, w_1) P(y_2 | z_1, w_2) P(z_2 | z_1, w_5) P(y_3 | z_2, w_3) P(y_4 | z_2, w_4) P(z_1)$$

Mutation probabilities



branch length = mutation rate \times time

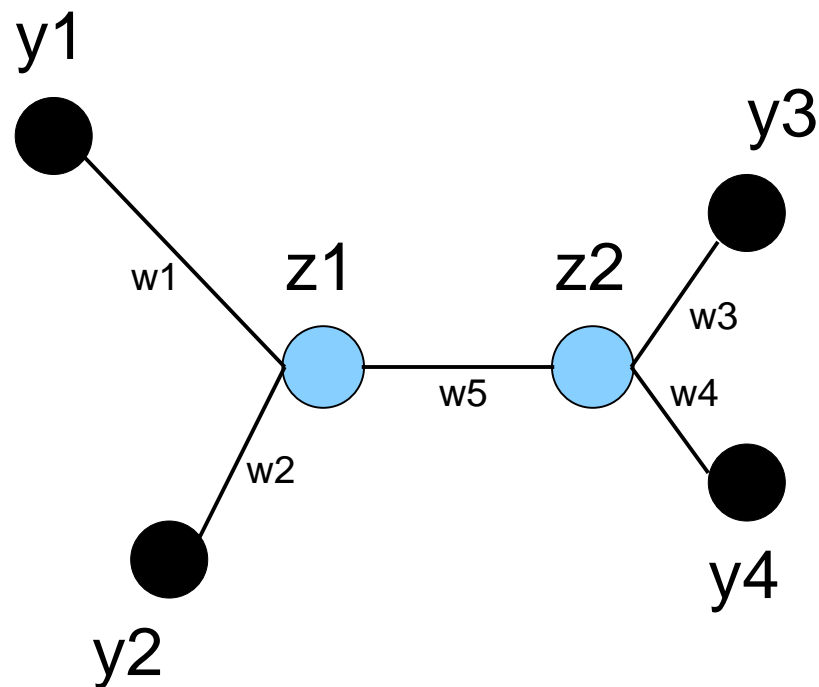
Different directed graphs



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

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

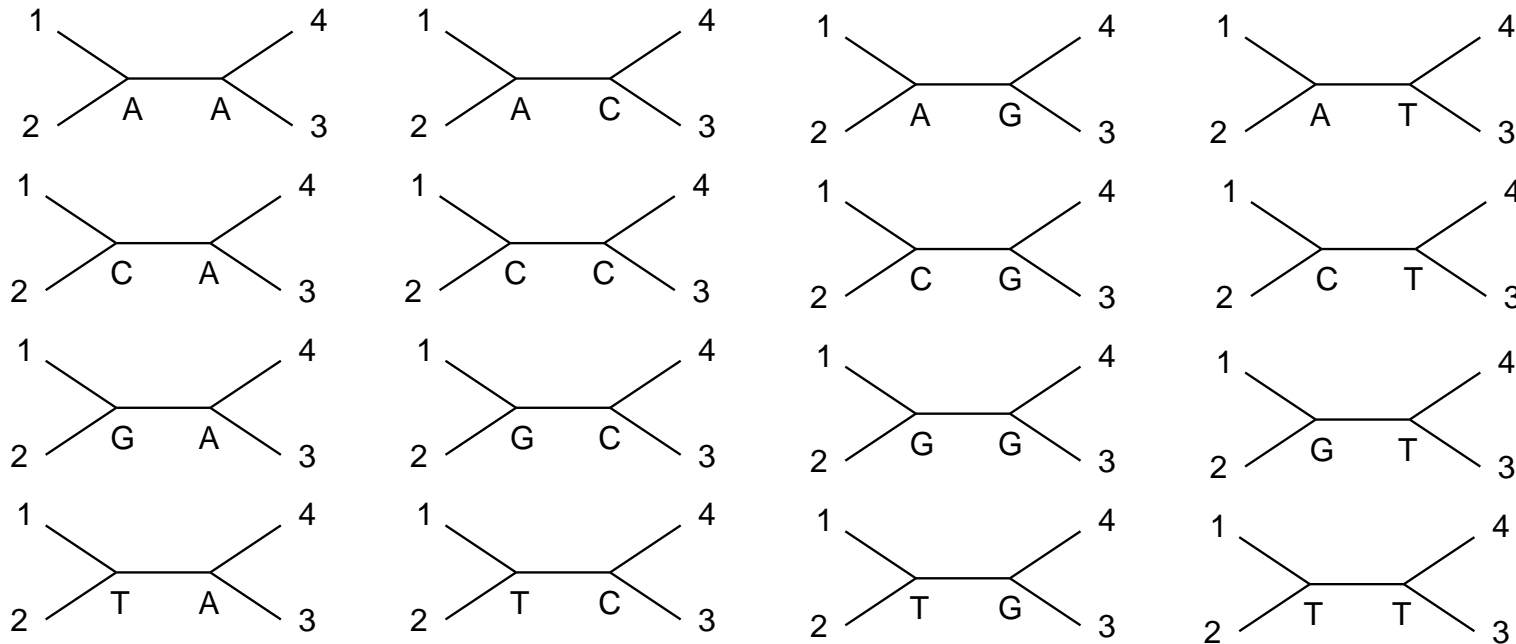
Expansion of the joint probability



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

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

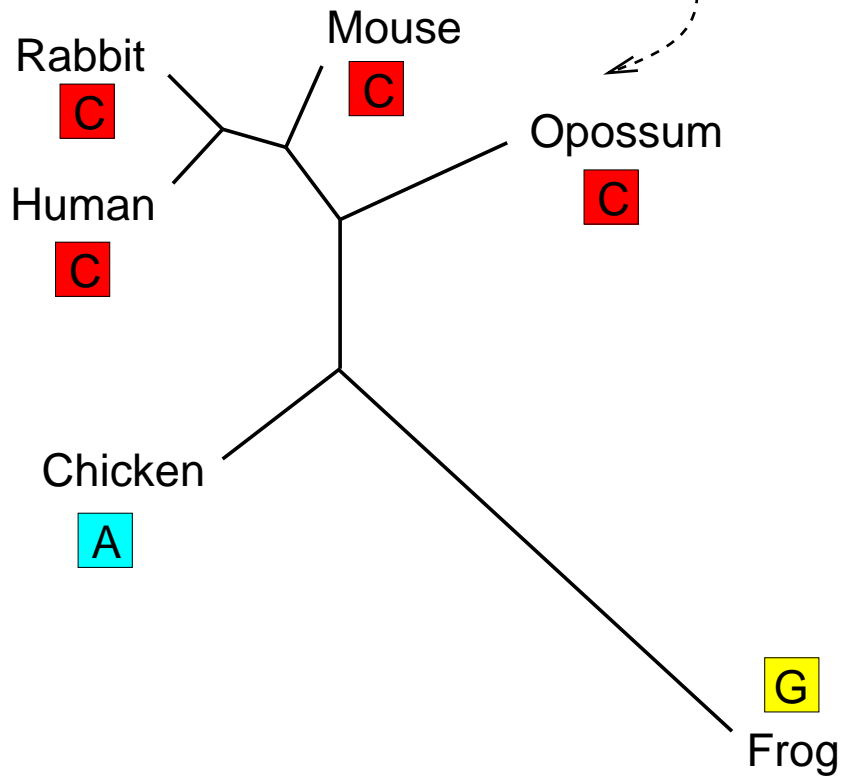
Marginalisation



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

∇

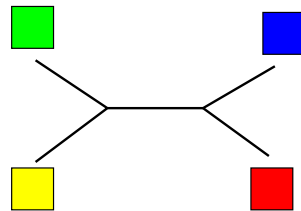
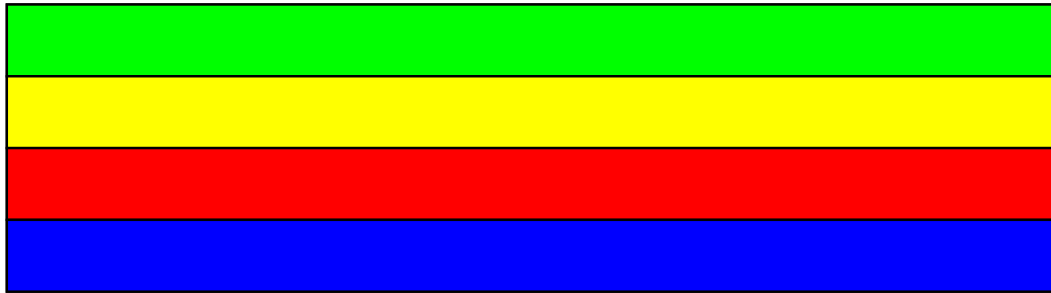
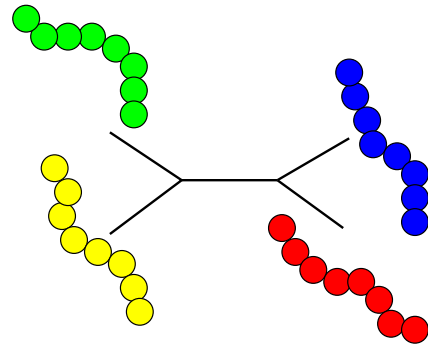
Frog	G	C	T	T	G	A	C	T	T	C	T	G	A	G	G	T	T
Chicken	G	C	G	T	A	A	C	T	T	C	A	C	A	T	G	A	T
Human	G	C	G	T	C	A	C	T	T	G	A	G	A	C	G	C	T
Rabbit	G	C	G	T	C	A	C	T	T	G	A	G	A	C	G	C	T
Mouse	G	C	G	T	C	A	C	T	T	G	A	C	A	G	G	C	T
Opossum	G	C	G	T	C	A	C	T	T	G	A	G	A	C	G	C	T



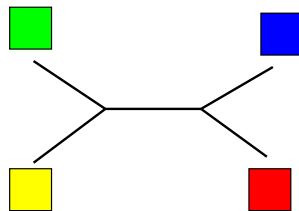
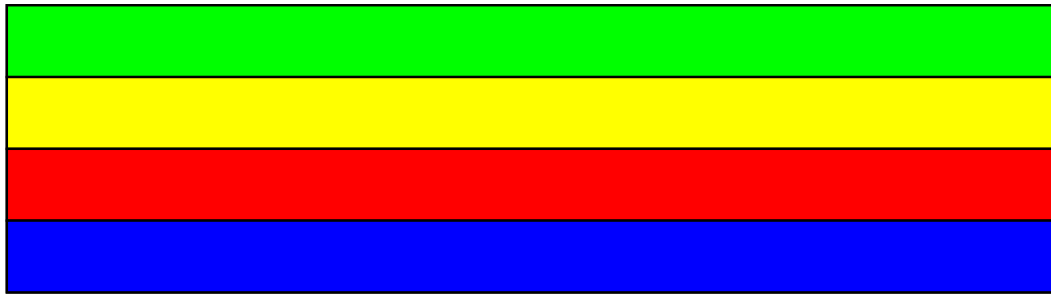
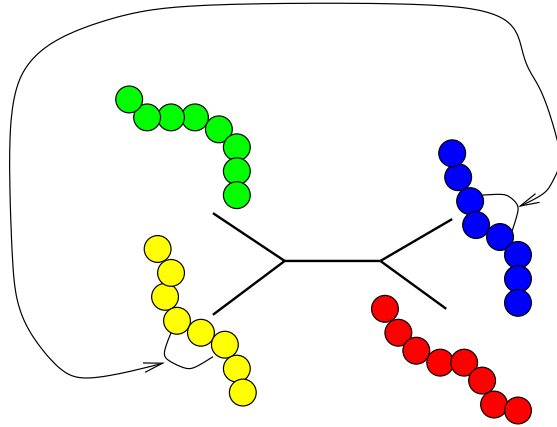
--> Likelihood

Topology
Branch lengths

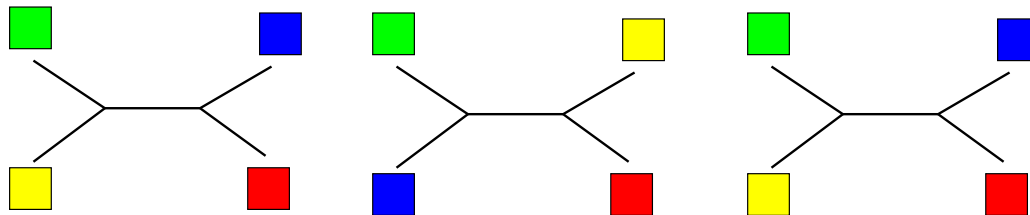
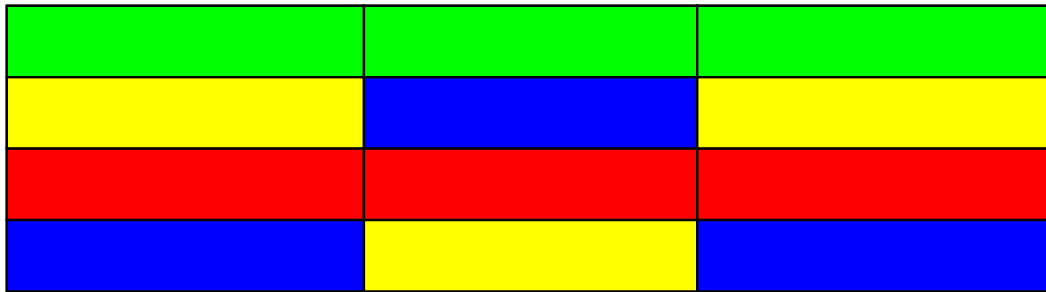
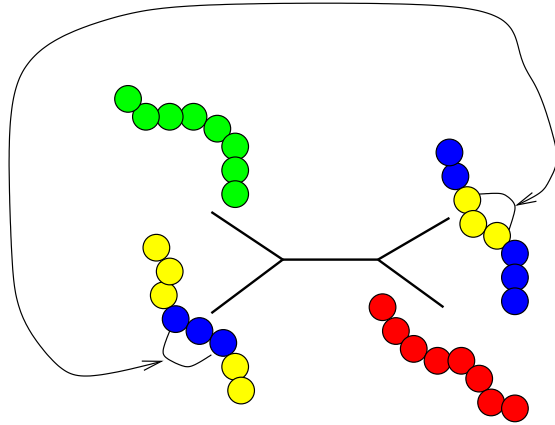
Recombination



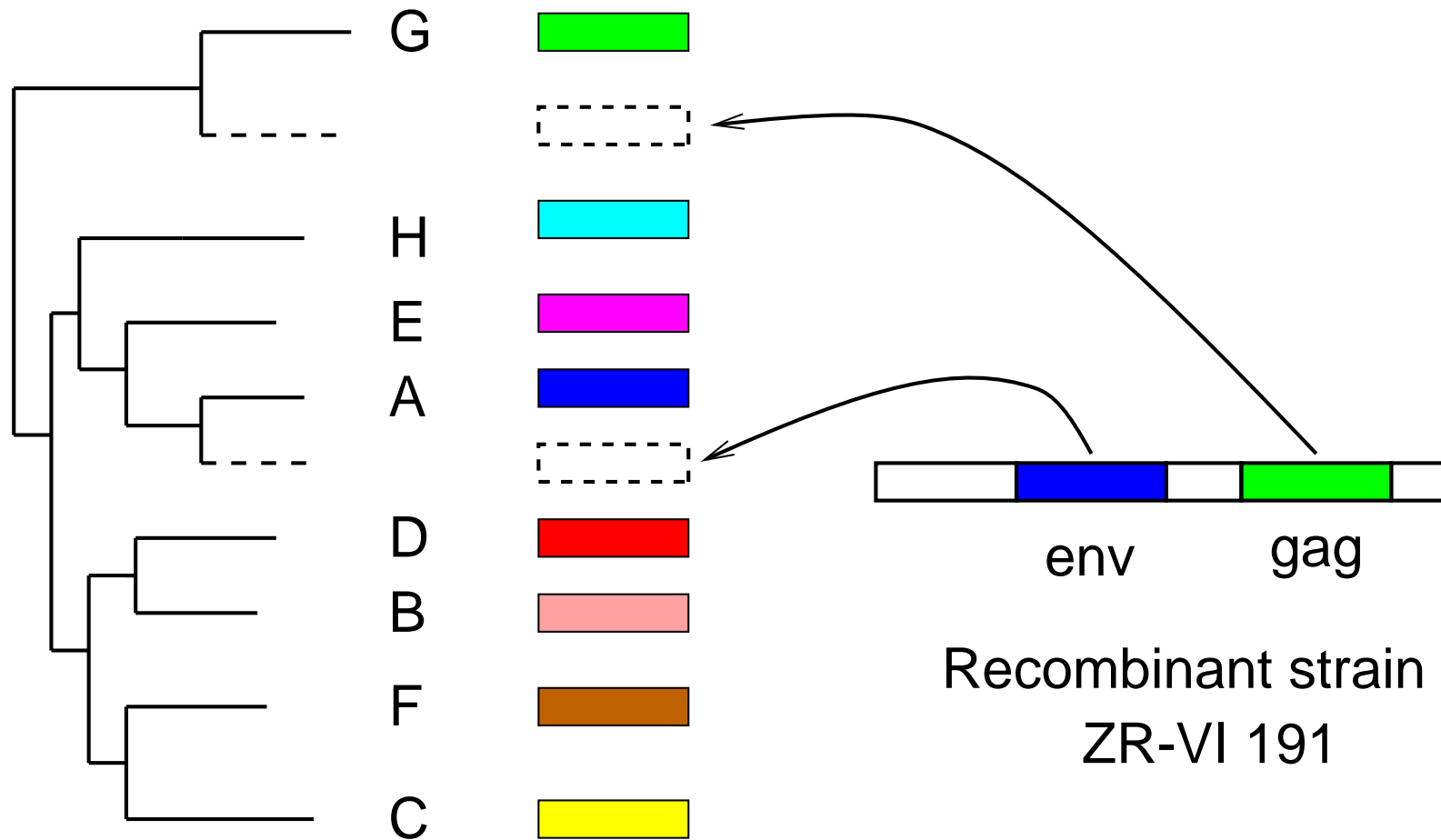
Recombination



Recombination



Recombination in HIV 1



Overview

- Introduction: Phylogenetics
- **Detecting recombination: Phylogenetic HMMs**
Dirk Husmeier, Frank Wright and Grainne McGuire
2001-2003
- Distinguishing between recombination and rate variation:
Phylogenetic FHMMs
Dirk Husmeier, 2005
- Learning the number of genomic regions
under selective pressure:
Phylogenetic FHMMs trained with RJMCMC
Wolfgang Lehrach and Dirk Husmeier, 2006

-
- Phylo-HMMs: Methodology
 - Phylo-HMMs: Applications
 - Phylo-HMMs: Limitations

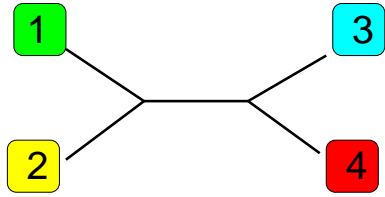
-
- **Phylo-HMMs: Methodology**
 - Phylo-HMMs: Applications
 - Phylo-HMMs: Limitations

Detecting recombination with Phylo-HMMs

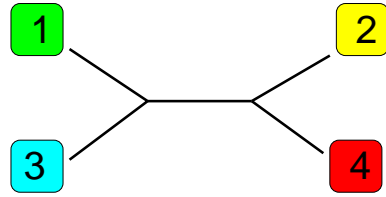
- Husmeier, Wright (2001)
Journal of Computational Biology 8
- Husmeier, McGuire (2002)
Bioinformatics 18
- Husmeier, McGuire (2003)
Molecular Biology and Evolution 20

Related work

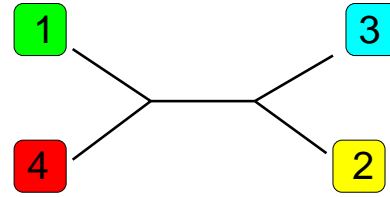
- Felsenstein & Churchill (1996)
Molecular Biology and Evolution 13
- Siepel & Haussler (2004)
Journal of Computational Biology 11



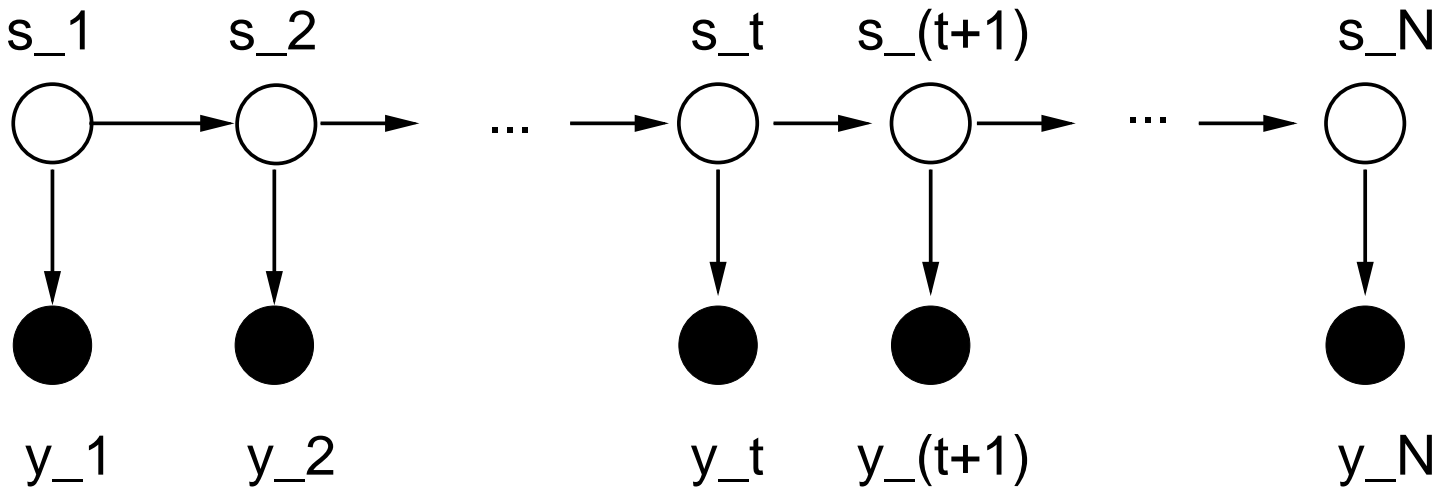
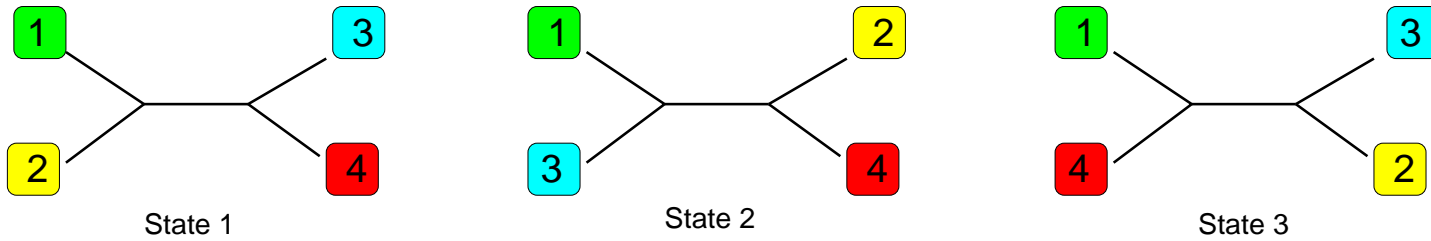
State 1



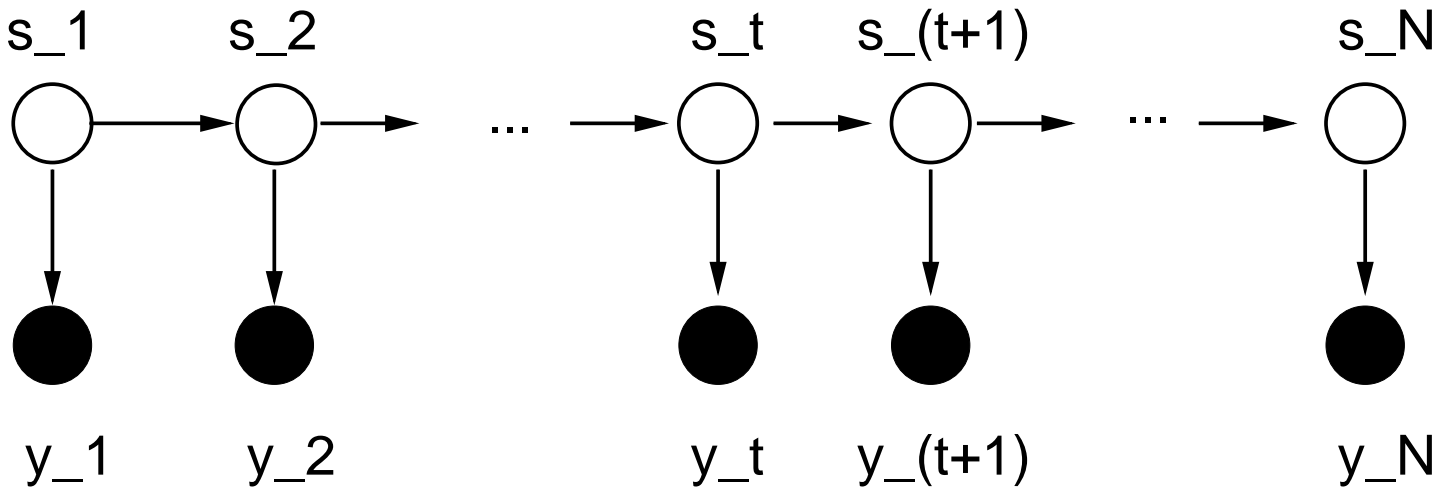
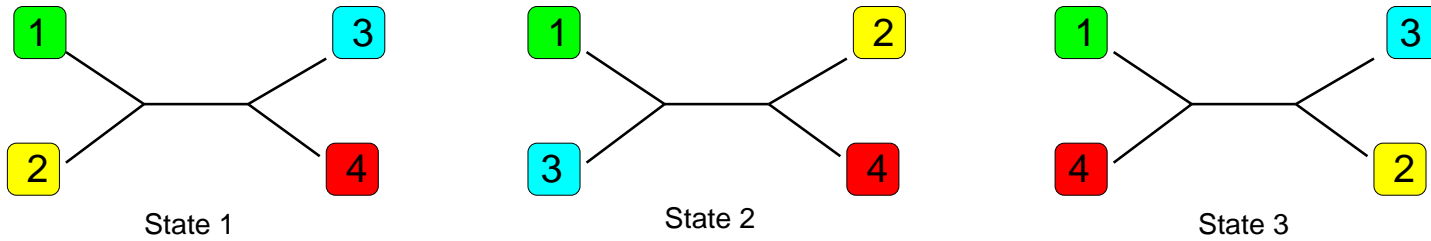
State 2



State 3



GCATCGTTCTATTTTACCGGCTCCCGA
 GTGTCGCTCAAGATTGCCATCGCGCGT
 GTCGTGGTCTAGATTGCCATCGCGCGT
 GTATCGCTCTAGTTTGCCAGCTCCCGT



GCATCGTTCTATTTTACCGGCTCCCGA
 GTGTCGCTCAAGATTGCCATCGCGCGT
 GTCGTGGTCTAGATTGCCATCGCGCGT
 GTATCGCTCTAGTTTGCCAGCTCCCGT

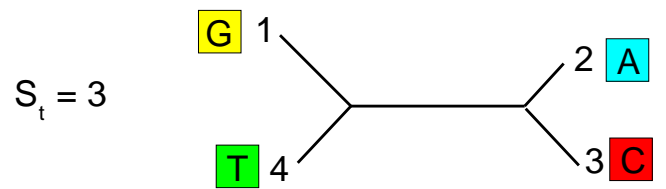
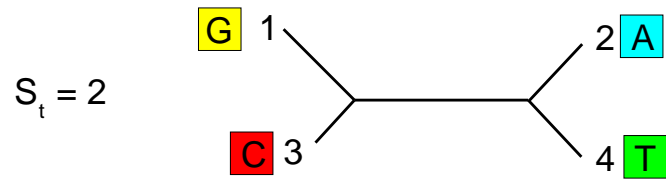
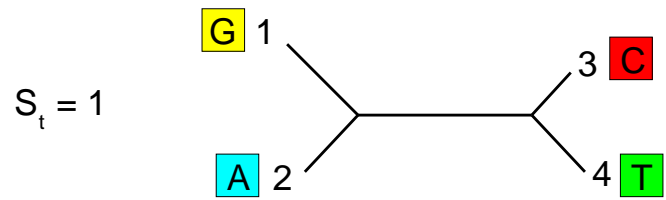
$$P(\mathcal{D}, \mathbf{S}) = \prod_{t=1}^N P(\mathbf{y}_t | S_t) \prod_{t=2}^N P(S_t | S_{t-1}) P(S_1)$$

Emission probabilities (vertical arrows)

↓

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

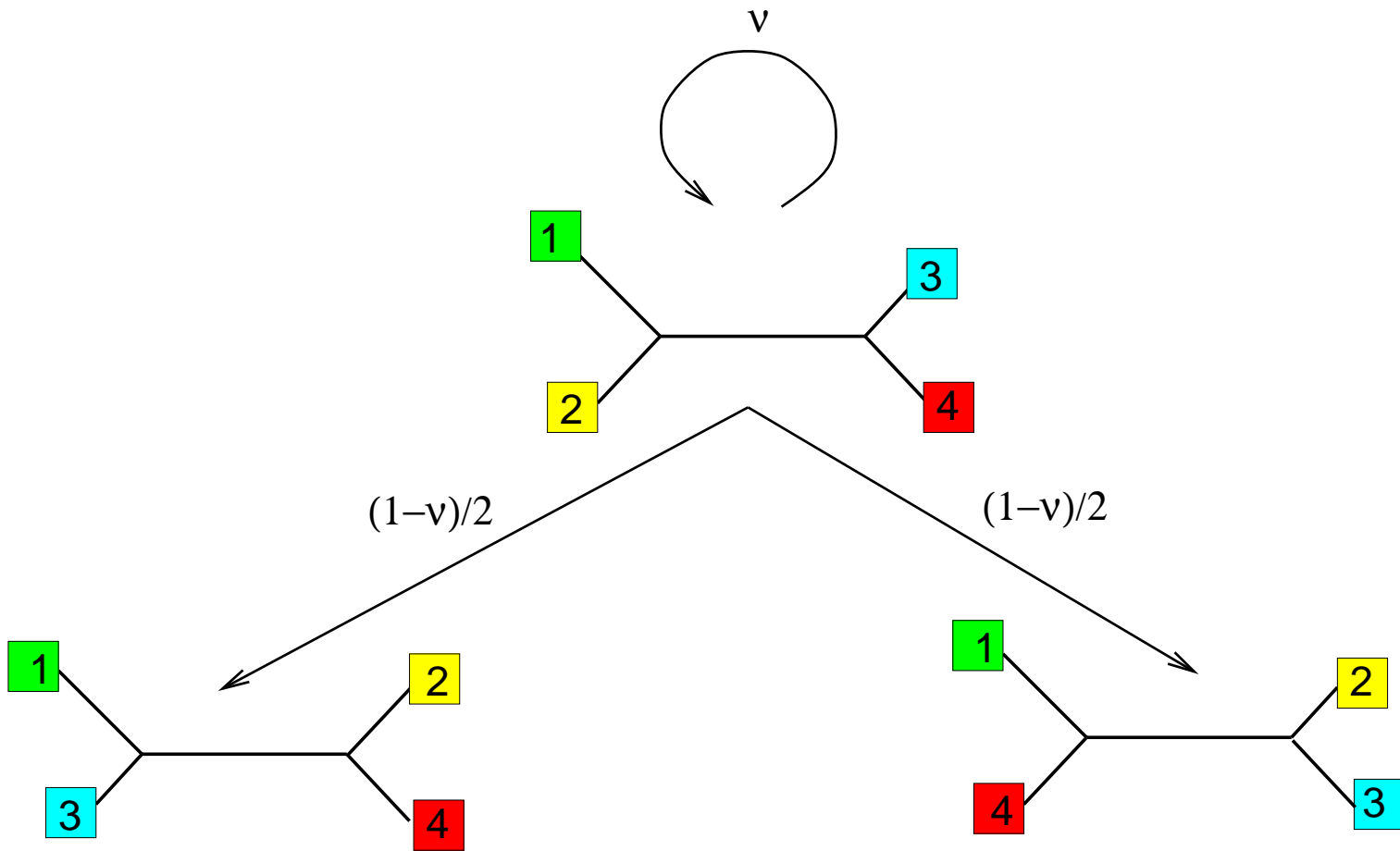
y_t



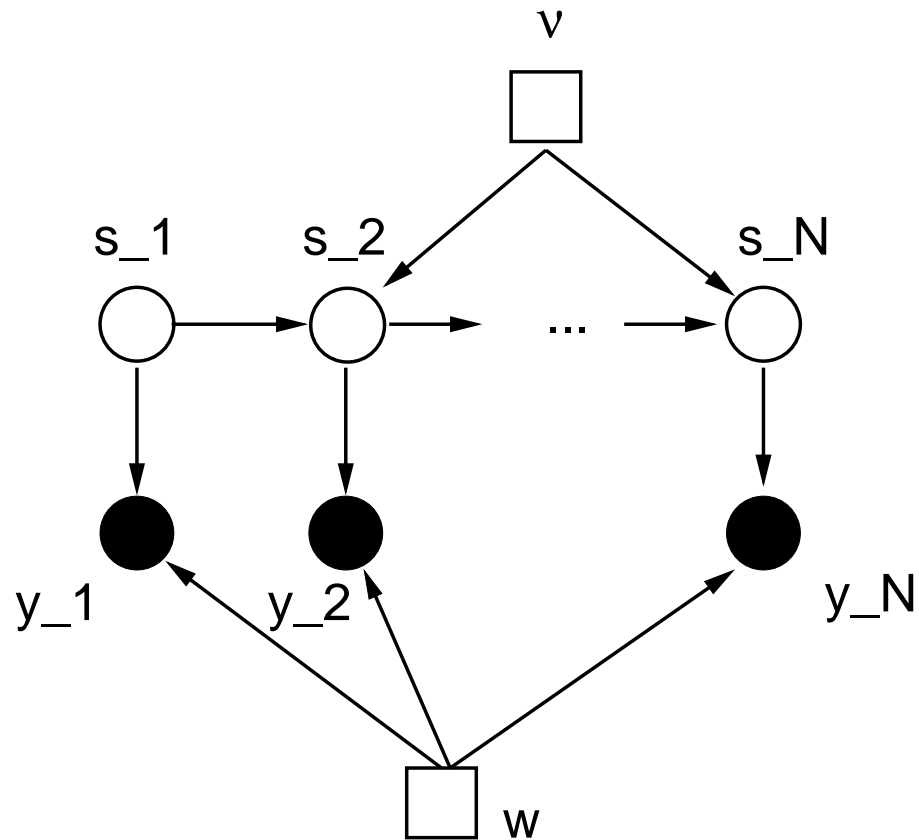
--> $P(y_t | S_t, w)$

Topology	S_t
Branch lengths	w

Transition probabilities (horizontal arrows)



HMM parameters



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

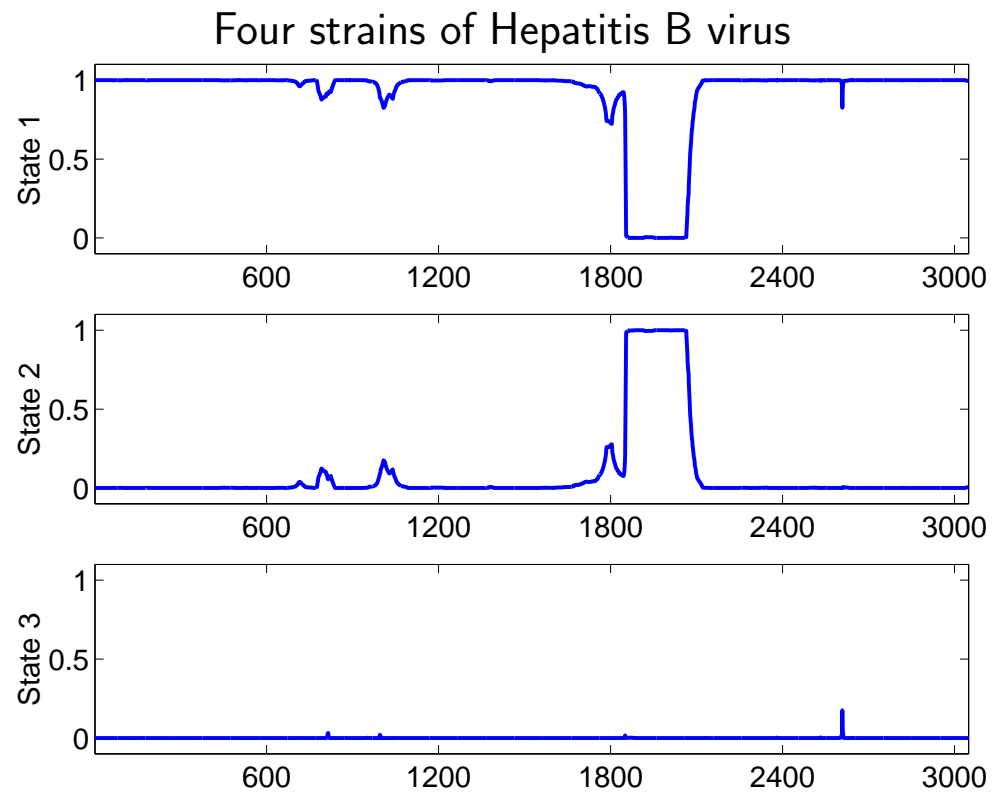
ν \longrightarrow Probability of *not* **changing** the tree **topology**

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

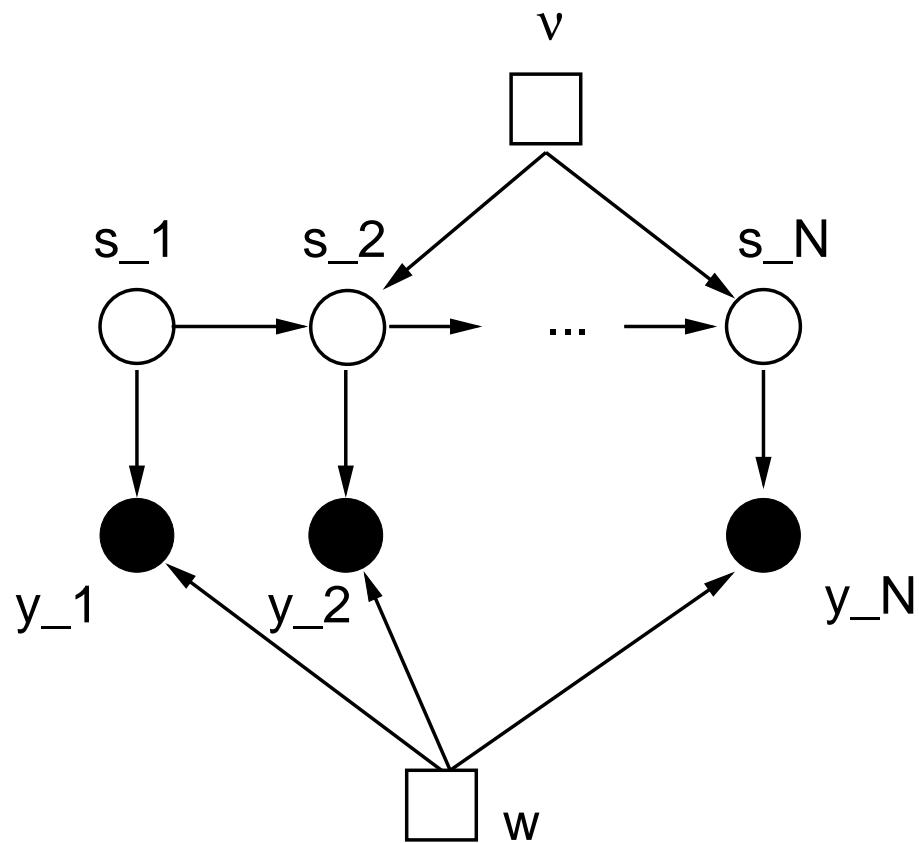
$$P(S_t|\mathcal{D}) = \sum_{S_1, \dots, S_{t-1}, S_{t+1}, \dots, S_N} P(\mathbf{S}|\mathcal{D})$$

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

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



HMM parameters



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

Bayesian approach

Husmeier, McGuire (2002)

Bioinformatics 18, S345-S353

Husmeier, McGuire (2003)

Molecular Biology and Evolution 20, 315-337

Bayesian approach

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

Bayesian approach

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

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

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

$$P(\mathbf{w}) = \left[\begin{array}{l} \text{Constant}(\Omega) \text{ if } 0 \leq w_i \leq \Omega \quad \forall i \\ 0 \text{ otherwise} \end{array} \right]$$

Bayesian approach

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

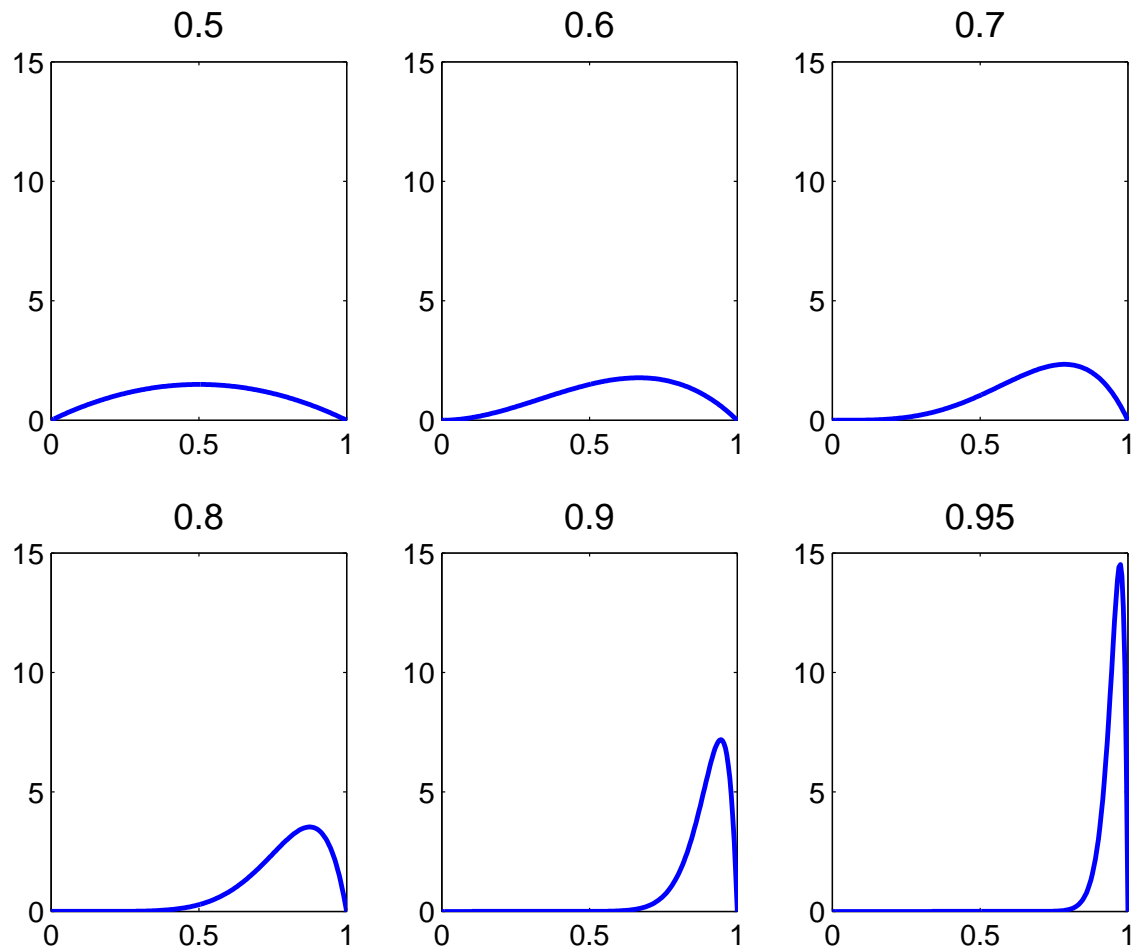
Posterior $P(\mathbf{w}, \nu|\mathcal{D}) \leftarrow$ Prior $P(\mathbf{w}, \nu) = P(\mathbf{w})P(\nu)$

$$P(\mathbf{w}) = \left[\begin{array}{l} \text{Constant}(\Omega) \text{ if } 0 \leq w_i \leq \Omega \quad \forall i \\ 0 \text{ otherwise} \end{array} \right]$$

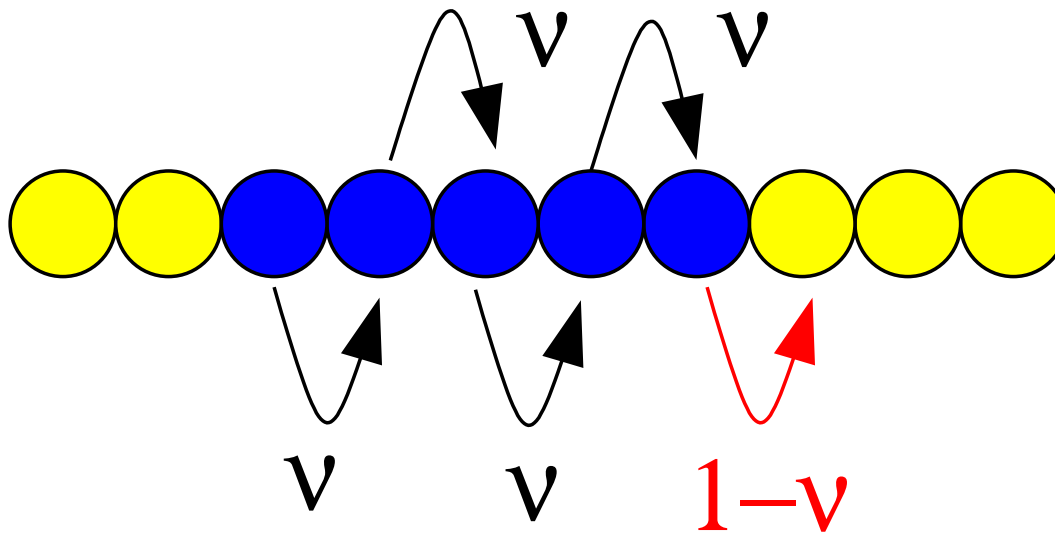
$$P(\nu) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \nu^{\alpha-1} (1-\nu)^{\beta-1}$$

Conjugate prior: Beta distribution.

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

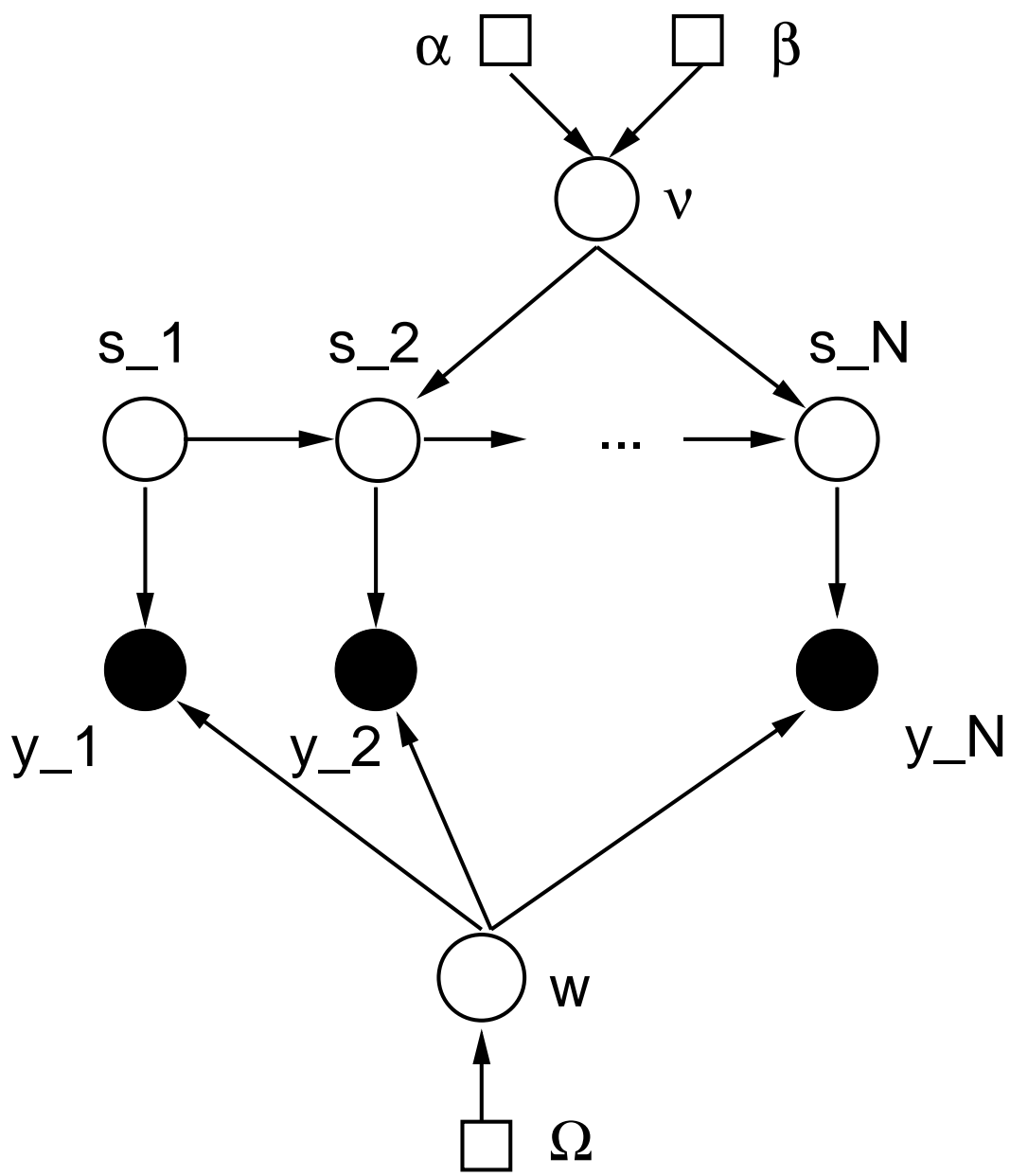


Average segment length



Segment length n . Probability: $P(n) = \nu^{n-1}(1 - \nu)$

$$\begin{aligned}\langle n \rangle &= \sum nP(n) = (1 - \nu) \sum n\nu^{n-1} = (1 - \nu) \frac{d}{d\nu} \sum \nu^n \\ &= (1 - \nu) \frac{d}{d\nu} \frac{1}{1 - \nu} = \frac{1}{1 - \nu}\end{aligned}$$



Numerical integration with MCMC

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

MCMC \longrightarrow Sample : $\{\mathbf{S}_i, \mathbf{w}_i, \nu_i\}_{i=1}^N$

$$P(\mathbf{S}, \mathbf{w}, \nu | \mathcal{D}) \approx \frac{1}{N} \sum_{i=1}^N \delta_{\mathbf{S}, \mathbf{S}_i} \delta(\mathbf{w} - \mathbf{w}_i) \delta(\nu - \nu_i)$$

$$P(S_t = k | \mathcal{D}) = \frac{1}{N} \sum_{i=1}^N \delta_{k, S_{t,i}} = \frac{N_{S_t=k}}{N}$$

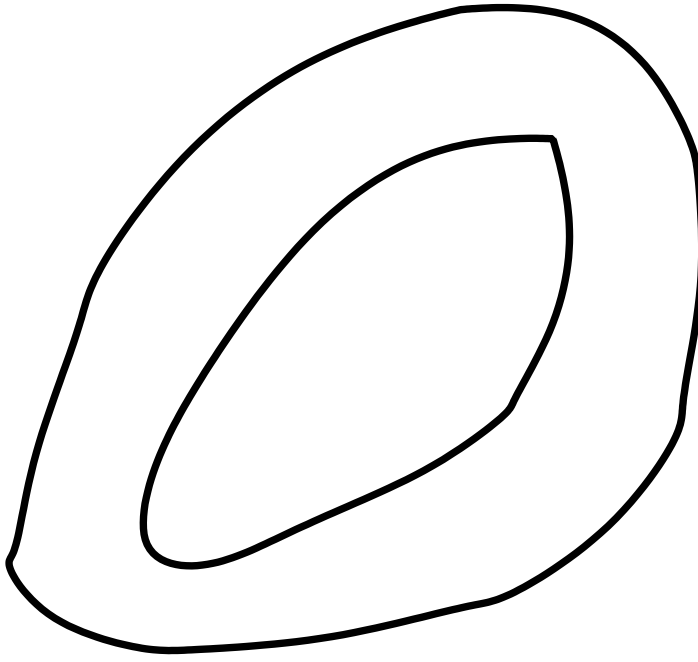
Sampling from the posterior distribution

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

Sampling from the posterior distribution

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

y

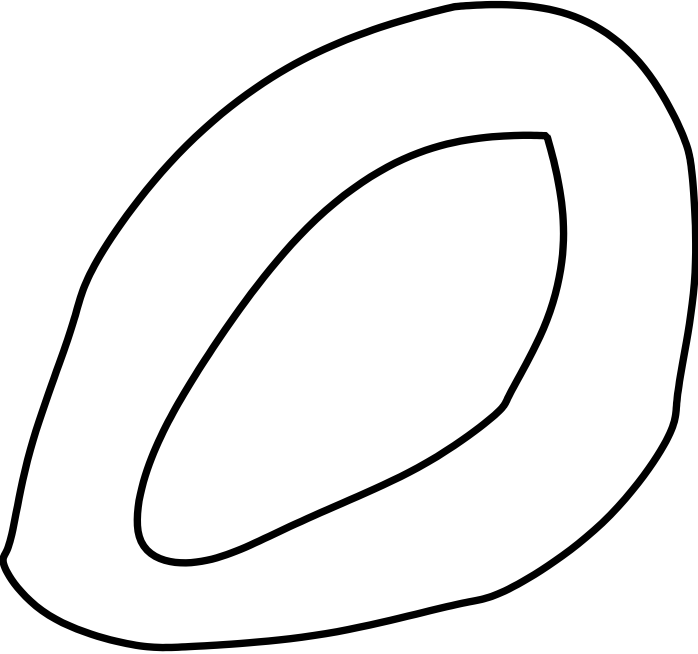


$P(x,y)$

x



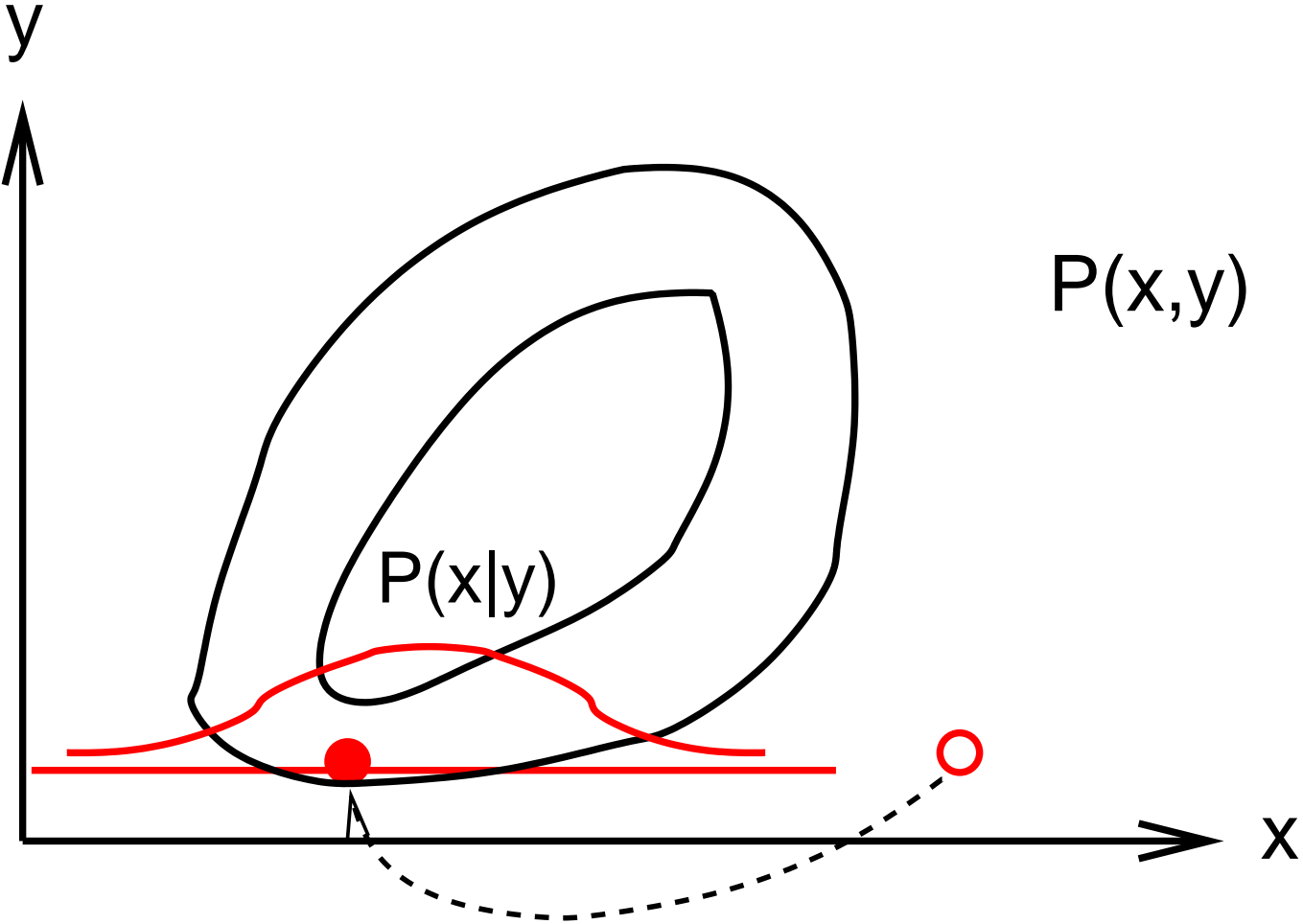
y

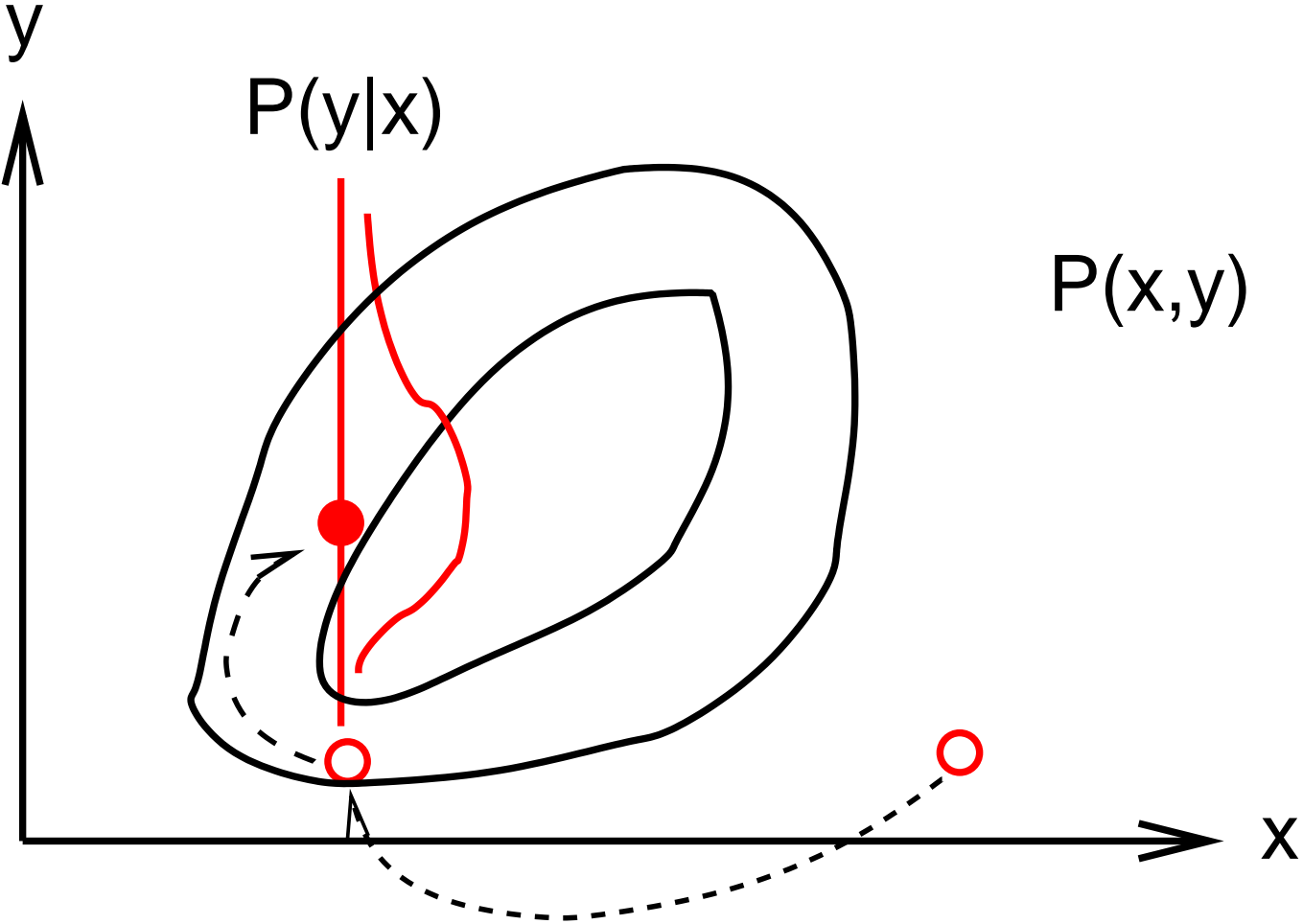


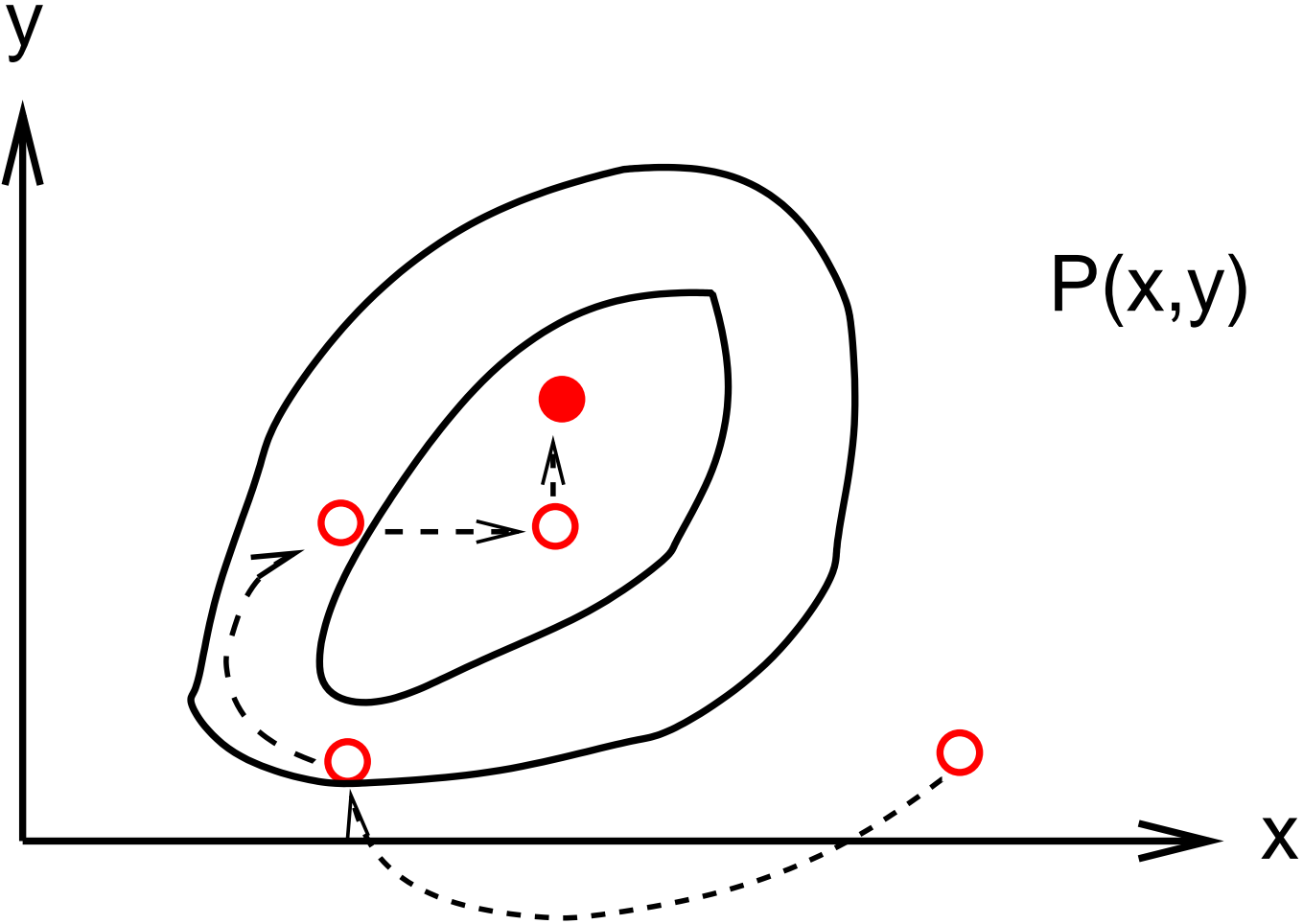
$P(x,y)$



x







Sampling from the posterior distribution

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

Sampling from the posterior distribution

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

Sampling from the posterior distribution

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

Sampling from the posterior distribution

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

Sampling from the posterior distribution

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

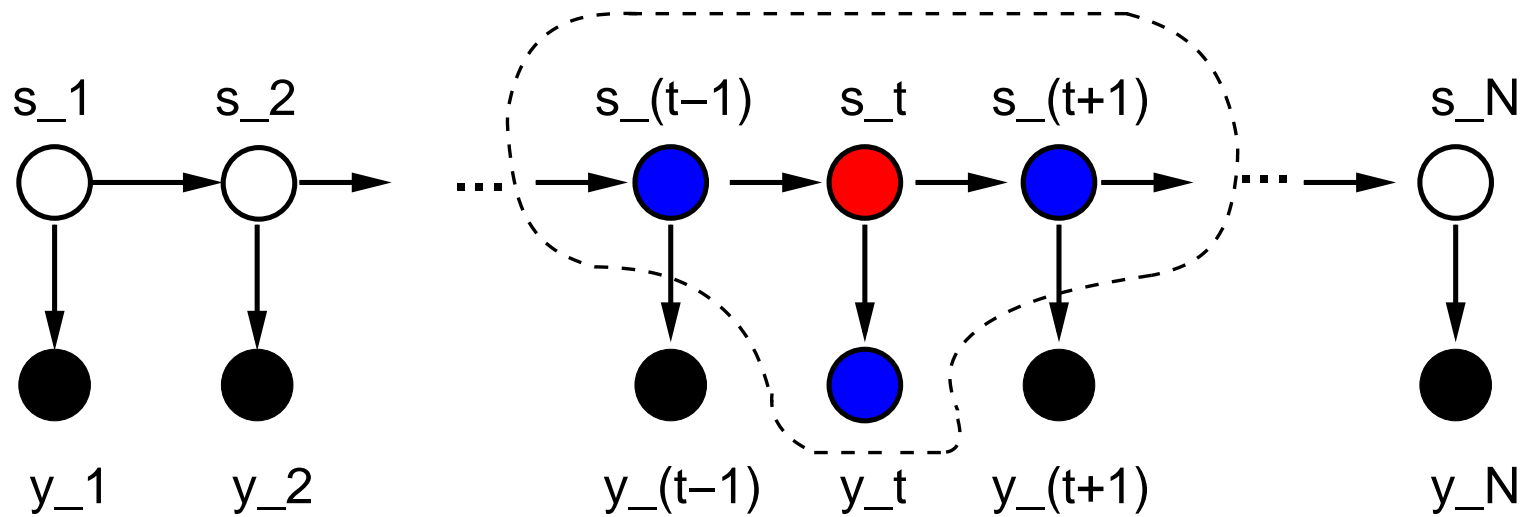
Sampling from the posterior distribution

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

Sampling from the posterior distribution

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



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

Gibbs-within-Gibbs scheme

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

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

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

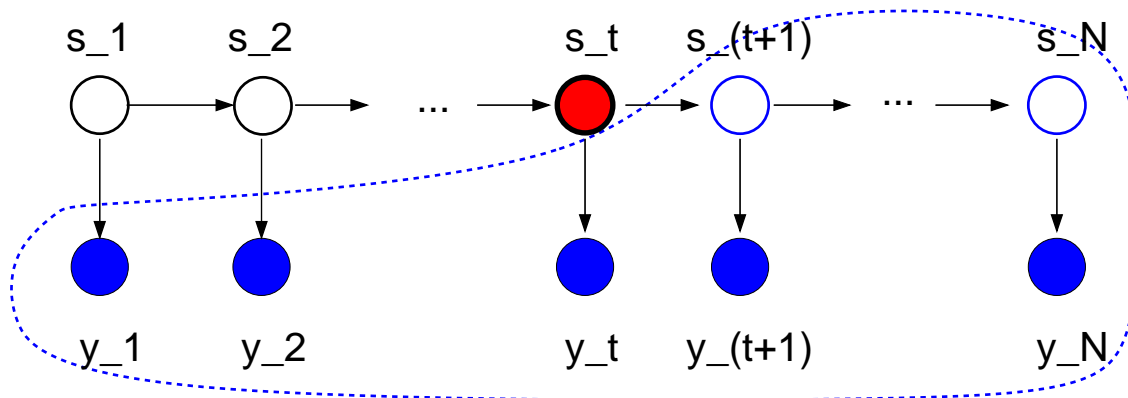
Slow mixing and convergence

Stochastic forward-backward algorithm

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

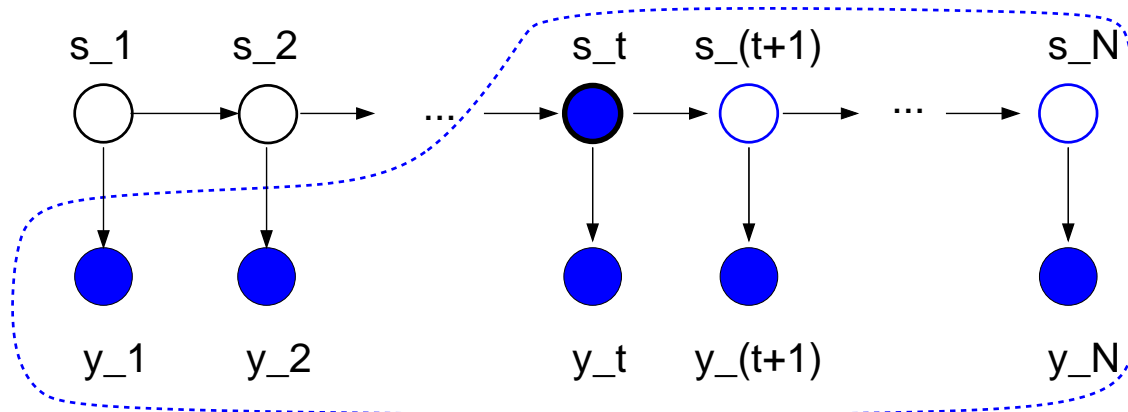
Simultaneous sampling of the states from $P(\mathbf{S}|\mathbf{w}, \nu, \mathcal{D})$

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



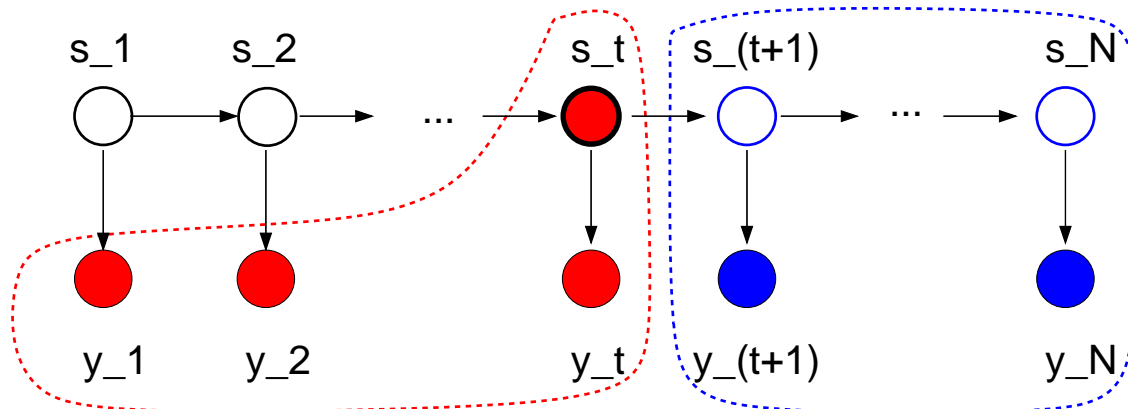
Simultaneous sampling of the states from $P(\mathbf{S}|\mathbf{w}, \nu, \mathcal{D})$

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



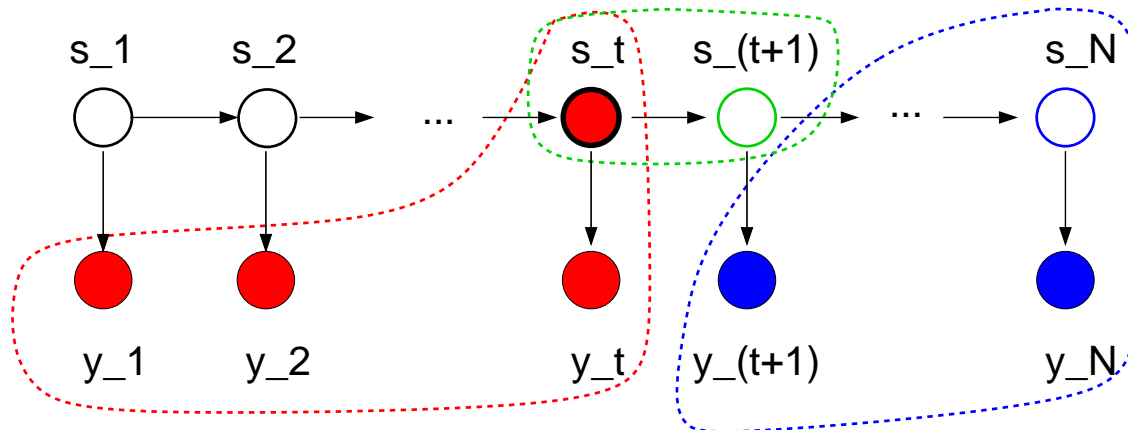
Simultaneous sampling of the states from $P(\mathbf{S}|\mathbf{w}, \nu, \mathcal{D})$

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



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

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



HMMs: The forward algorithm

$$P(y_1, \dots, y_N) = \sum_{S_1, \dots, S_N} P(y_1, \dots, y_N, S_1, \dots, S_N)$$

$$P(y_1, \dots, y_N) = \sum_{S_N} \alpha_N(S_N)$$

$$\alpha_n(S_n) = P(y_1, \dots, y_n, S_n)$$

$$= \sum_{S_1} \dots \sum_{S_{n-1}} P(y_1, \dots, y_n, S_1, \dots, S_{n-1}, S_n)$$

$$= \sum_{S_1} \dots \sum_{S_{n-1}} \prod_{t=1}^n P(y_t | S_t) P(S_t | S_{t-1})$$

$$= \sum_{S_1} \dots \sum_{S_{n-1}} P(y_n | S_n) P(S_n | S_{n-1}) \prod_{t=1}^{n-1} P(y_t | S_t) P(S_t | S_{t-1})$$

$$= P(y_n | S_n) \sum_{S_{n-1}} P(S_n | S_{n-1}) \sum_{S_1} \dots \sum_{S_{n-2}} \prod_{t=1}^{n-1} P(y_t | S_t) P(S_t | S_{t-1})$$

$$= P(y_n | S_n) \sum_{S_{n-1}} P(S_n | S_{n-1}) \alpha_{n-1}(S_{n-1})$$

Stochastic forward–backward algorithm

- Run the **forward algorithm** to obtain

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

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

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

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

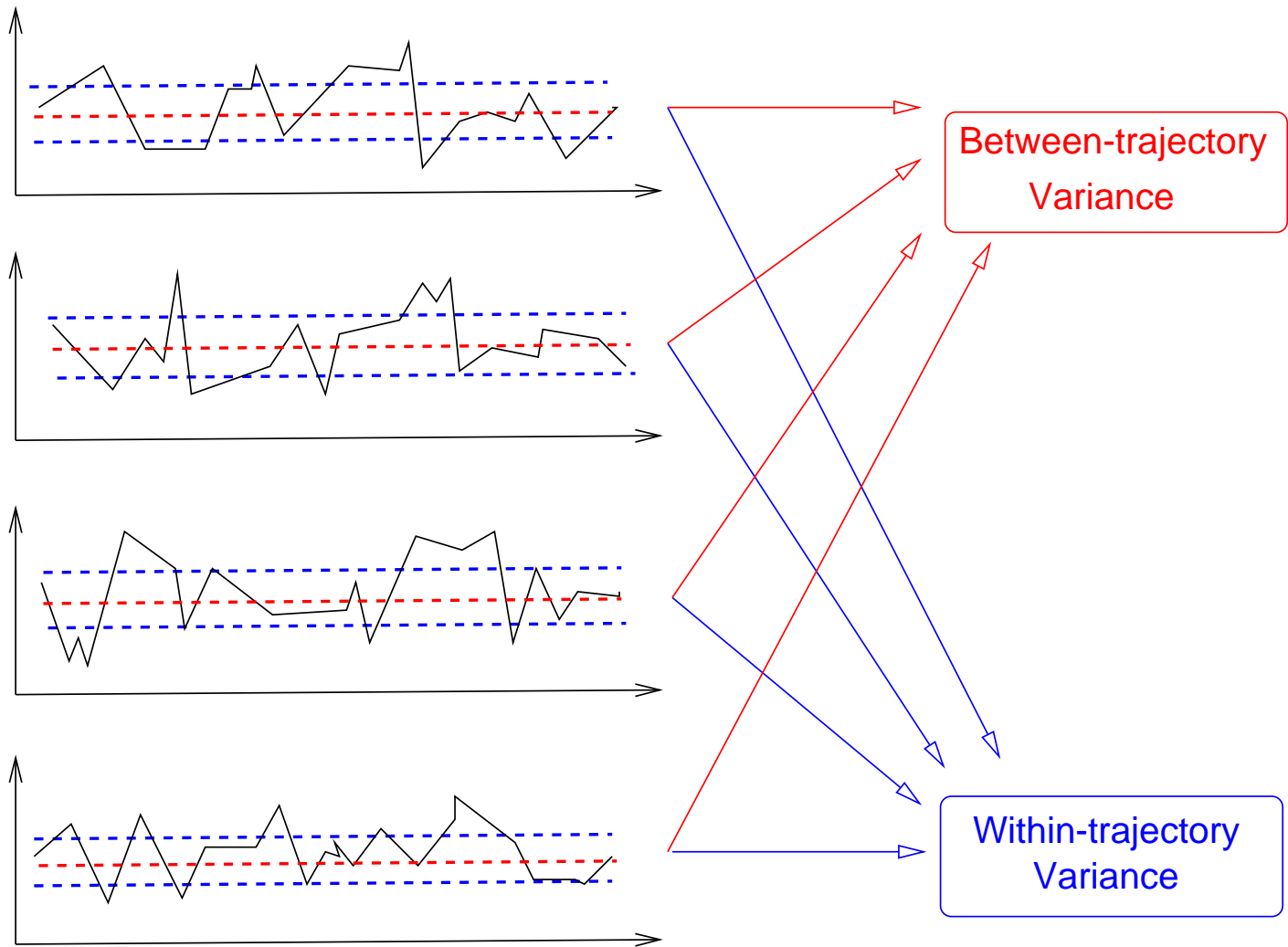
Sampling from the posterior distribution

- Sampling from $P(\mathbf{S}, \mathbf{w}, \nu | \mathcal{D})$
- Gibbs sampling :
 - $\mathbf{S} \sim P(\mathbf{S} | \mathbf{w}, \nu, \mathcal{D})$
 - $\mathbf{w} \sim P(\mathbf{w} | \mathbf{S}, \nu, \mathcal{D})$
 - $\nu \sim P(\nu | \mathbf{S}, \mathbf{w}, \mathcal{D})$
- ν : Sample from Beta distribution
- \mathbf{w} : Metropolis-Hastings within Gibbs
- \mathbf{S} : Stochastic forward-backward algorithm
 - $\mathbf{S} \sim P(\mathbf{S} | \mathbf{w}, \nu, \mathcal{D})$

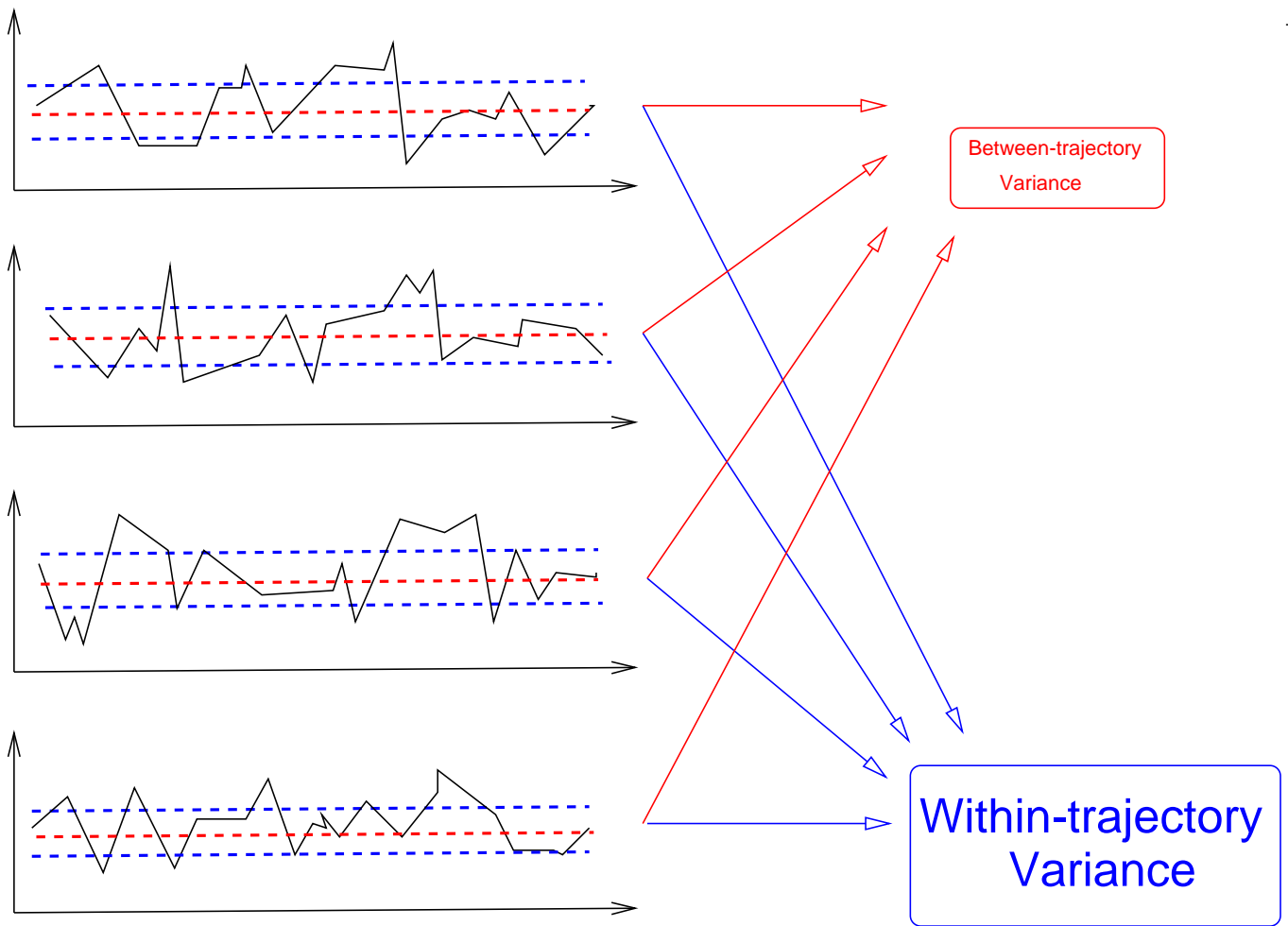
MCMC convergence diagnostics

Gelman and Rubin (1992)

MCMC convergence diagnostic



MCMC convergence diagnostic



Convergence diagnostic of Gelman and Rubin (1992)

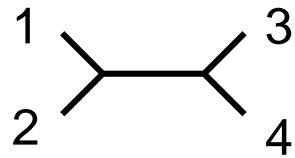
Potential scale reduction factor (PSRF):

Factor by which the estimated scale of the posterior distribution will shrink as $T \rightarrow \infty$

$$PSRF \simeq \frac{\text{Within-trajectory variance} + \text{Between-trajectory variance}}{\text{Within-trajectory variance}}$$

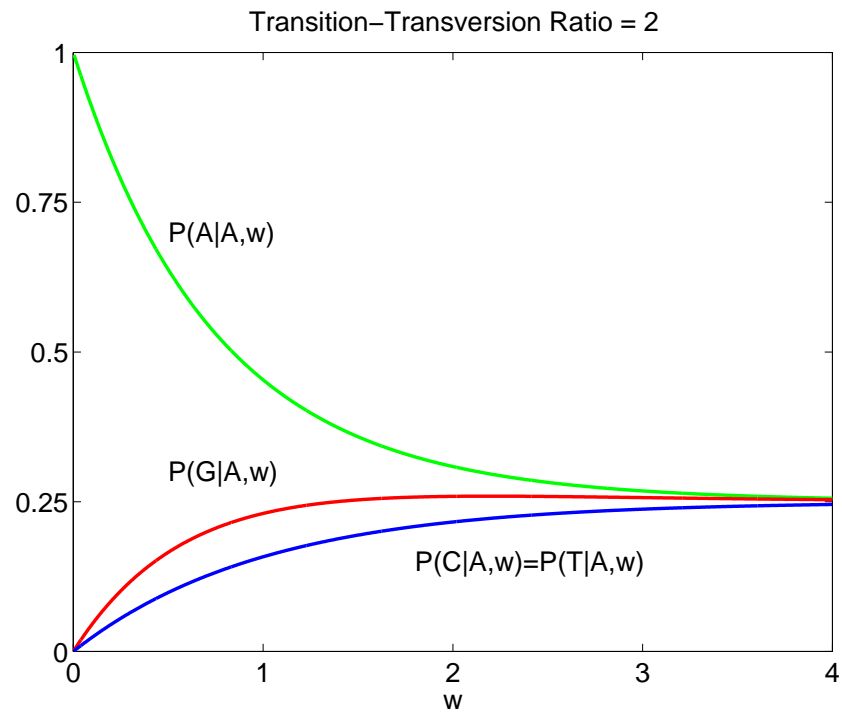
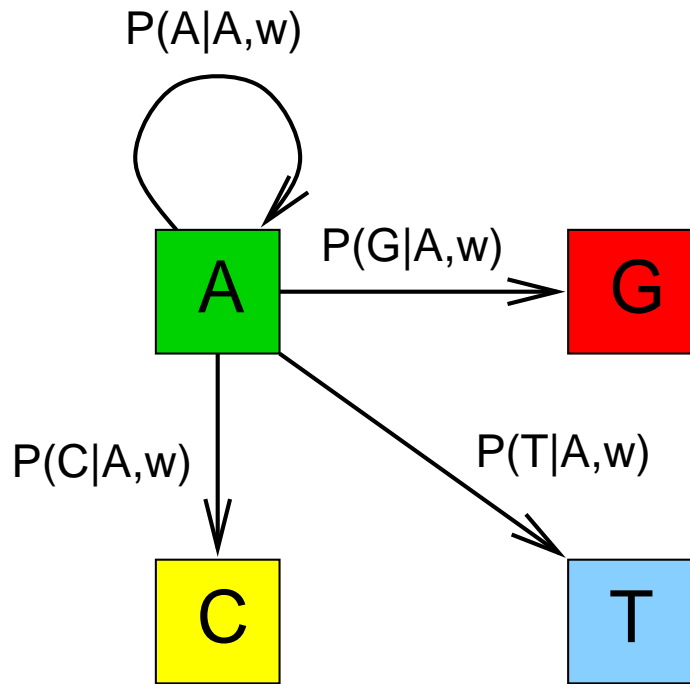
-
- Phylo-HMMs: Methodology
 - **Phylo-HMMs: Applications**
 - Phylo-HMMs: Limitations

Synthetic example



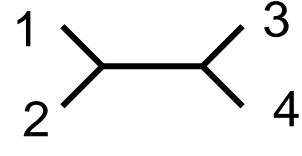
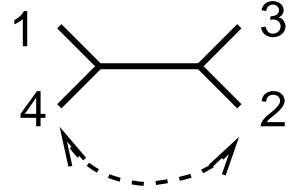
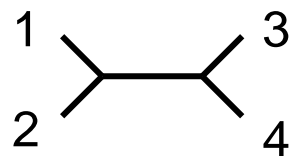
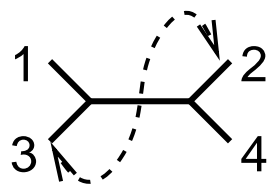
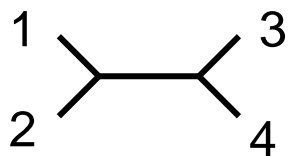
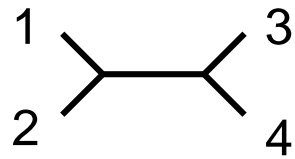
- Model of nucleotide substitution: Kimura 2-parameter, $\tau = 2$.
- Alignment of length $N = 1000$ nucleotides.

Mutation probabilities

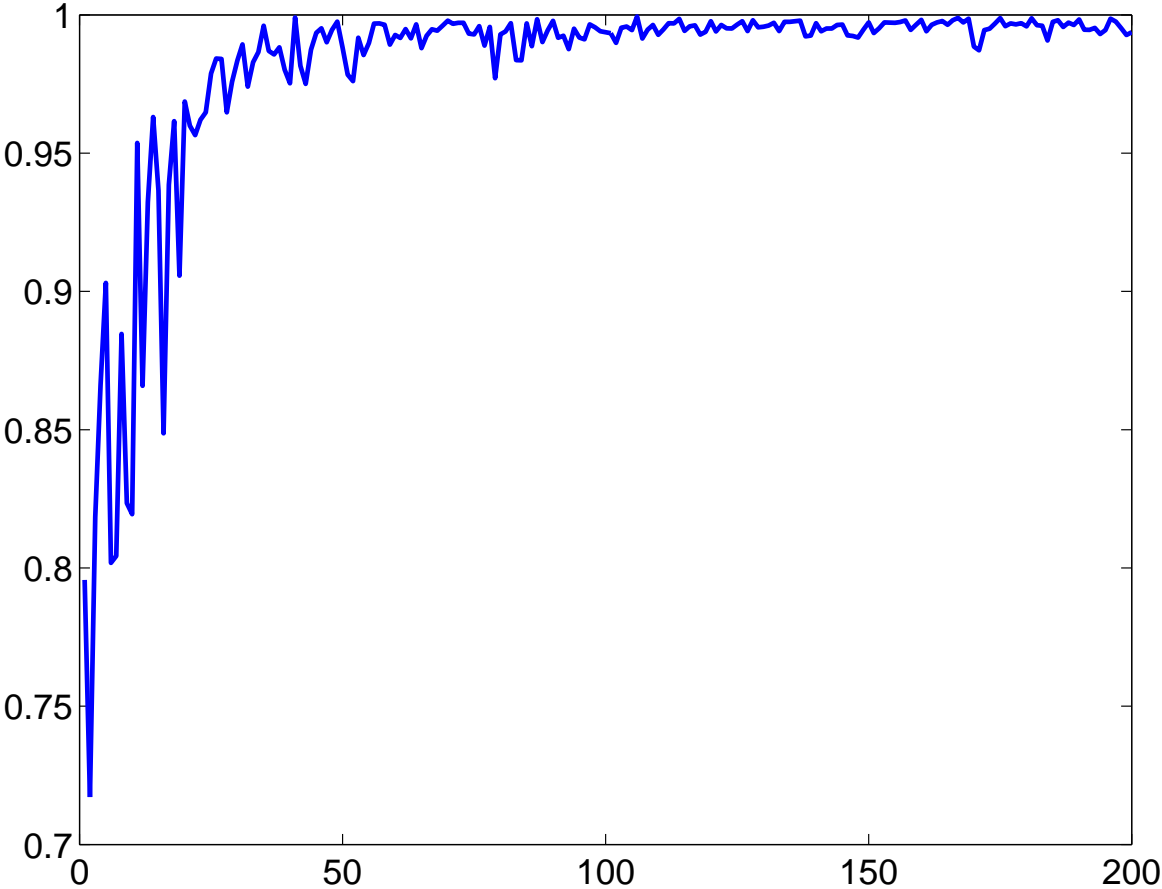


branch length = mutation rate \times time

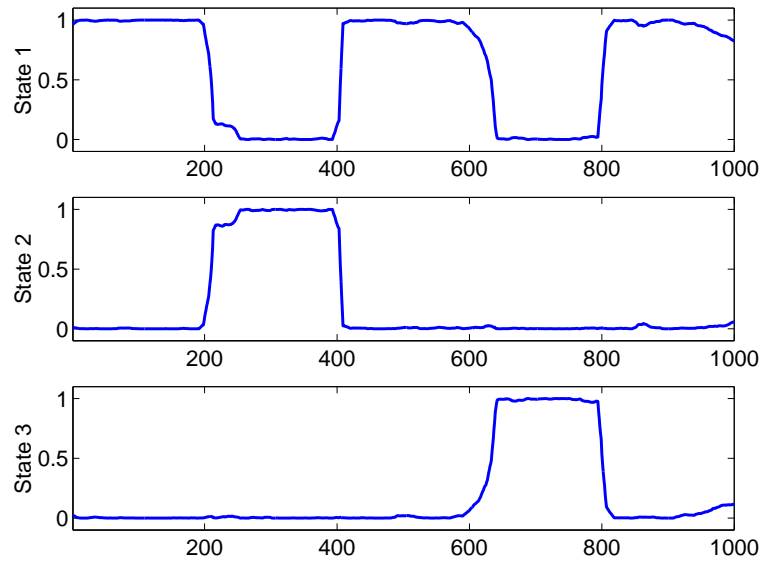
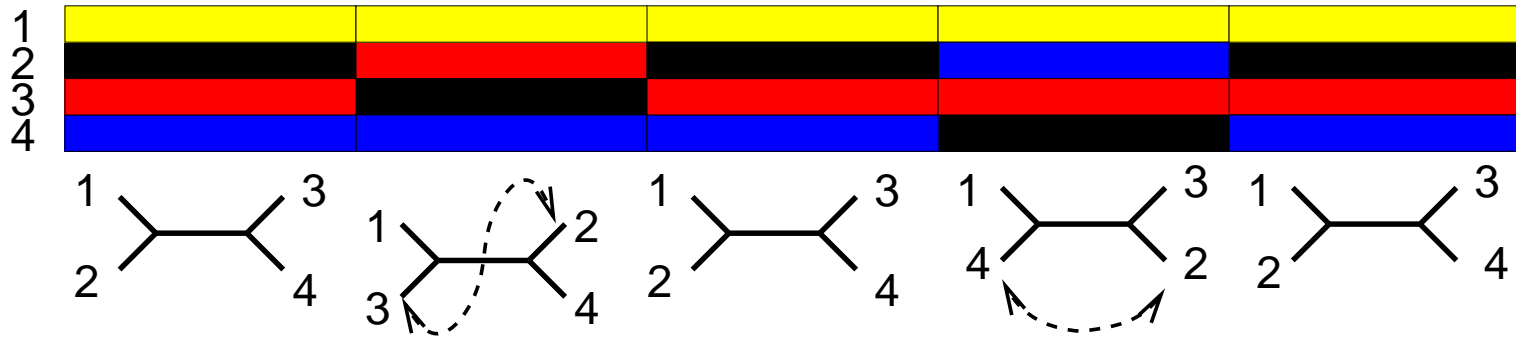
Synthetic example



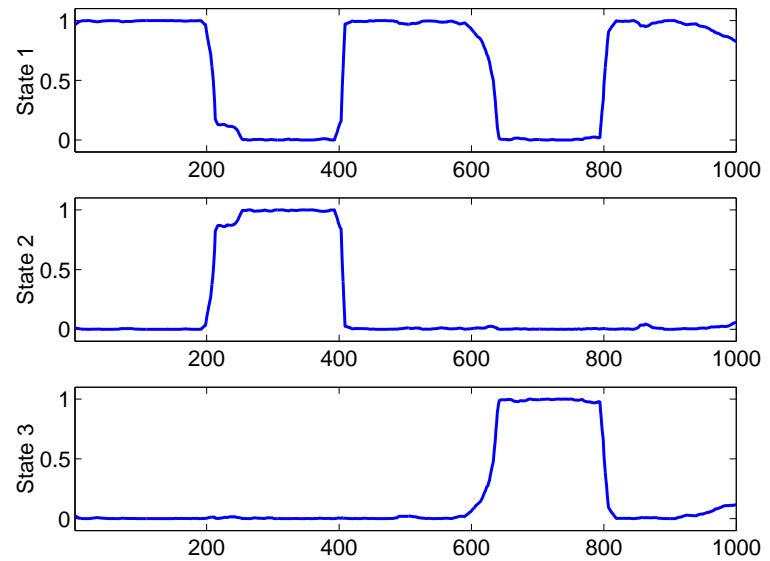
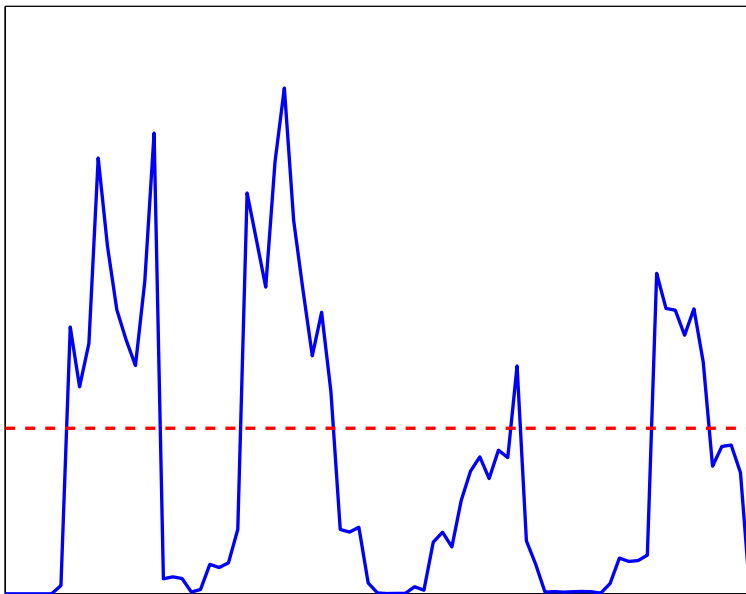
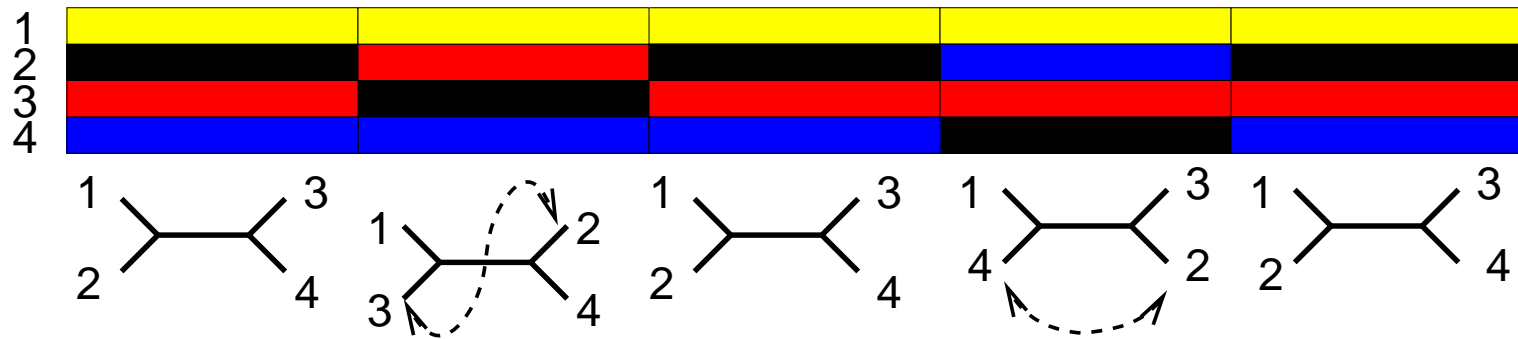
Trace plot of ν



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



Phylo-HMM vs. Topal, window size=200



Hepatitis B Virus (Bollyky et al. 1995)

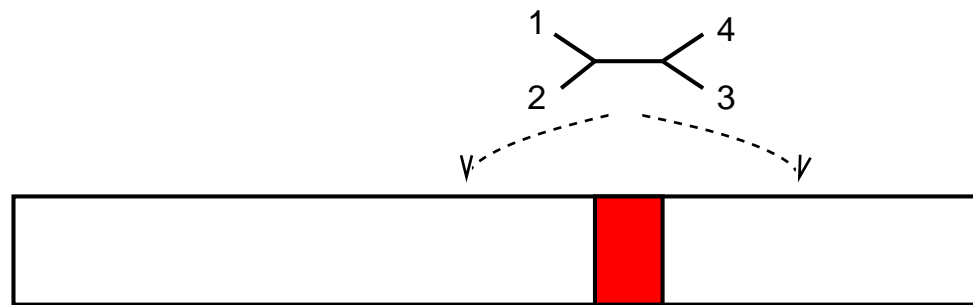
DNA alignment, 3049 nucleotides

1) HPBADW1

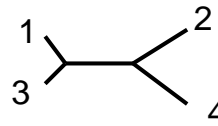
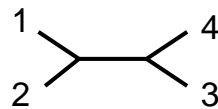
2) HPBADW2

3) HPBADWZCG

4) HPBADRC



State 1



State 2

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

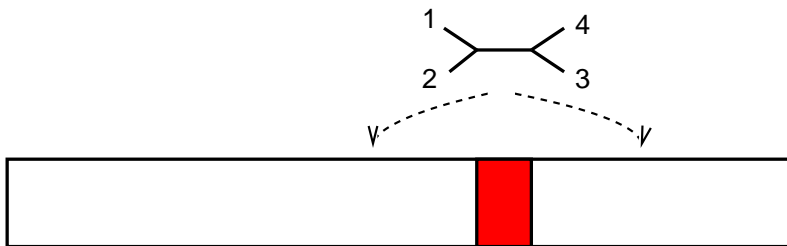
DNA alignment, 3049 nucleotides

1) HPBADW1

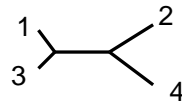
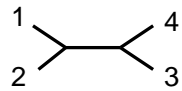
2) HPBADW2

3) HPBADWZCG

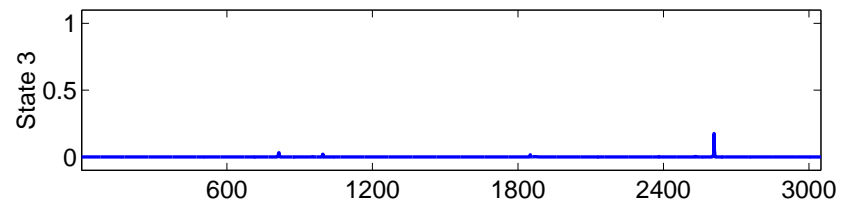
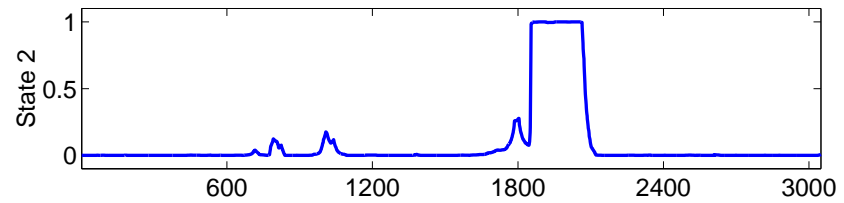
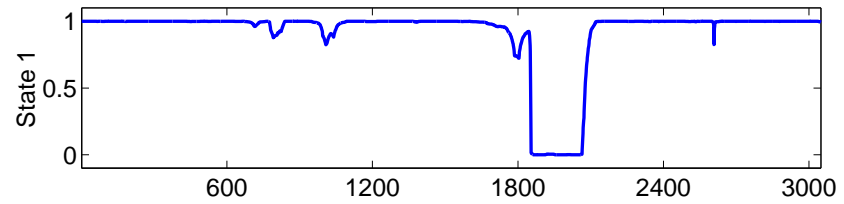
4) HPBADRC



State 1



State 2



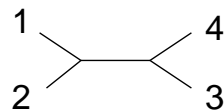
Gene conversion in maize (Moniz de Sa, Drouin, 1996)

Actin genes: DNA alignment of 1008 nucleotides

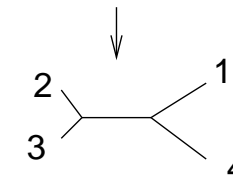
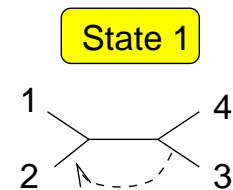
- 1) Maz56 3) Maz63
- 2) Maz63 4) Maz89

875 bases

133 bases



State 1

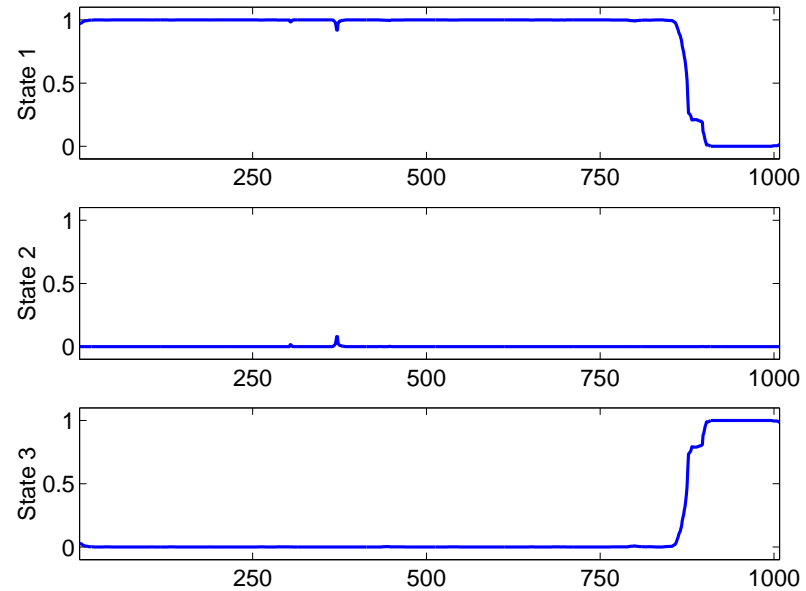
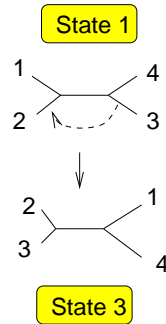
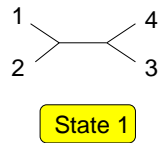


State 3

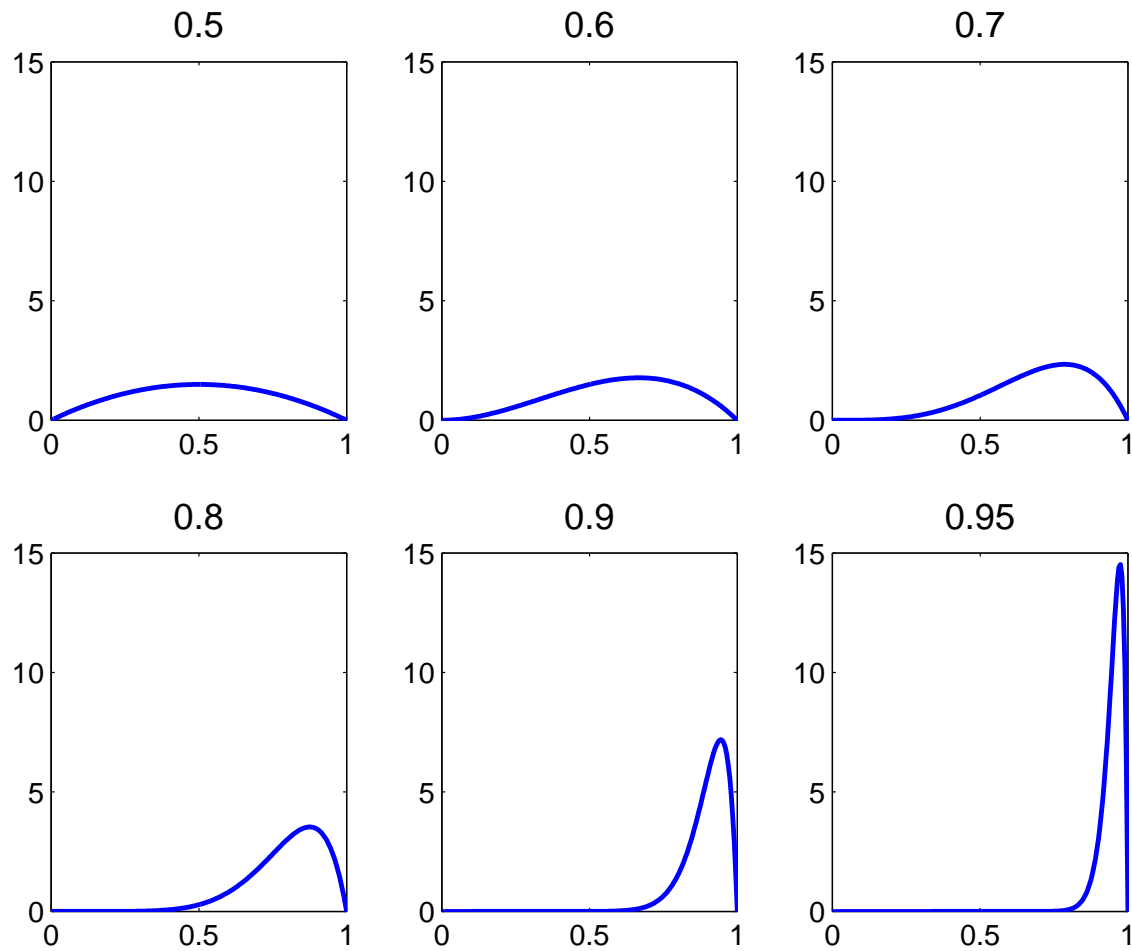
$P(S_t|\mathcal{D})$: Marginal posterior probability

Actin genes: DNA alignment of 1008 nucleotides

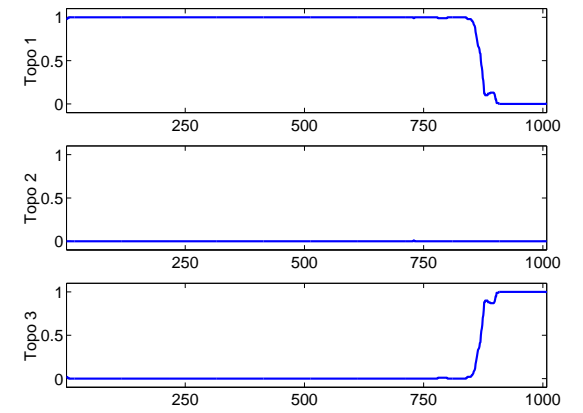
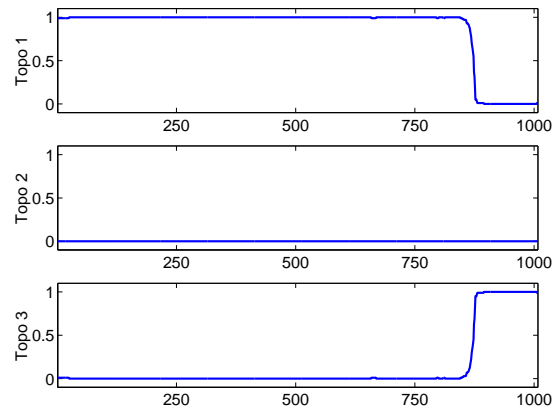
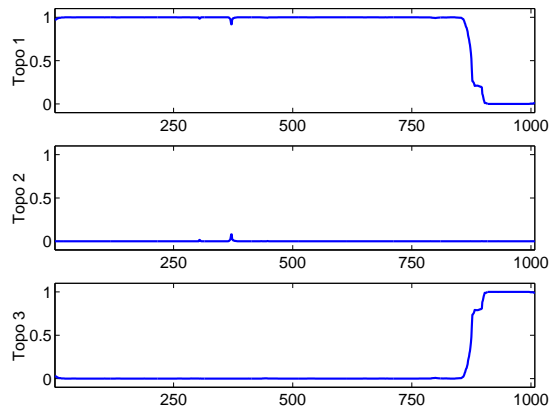
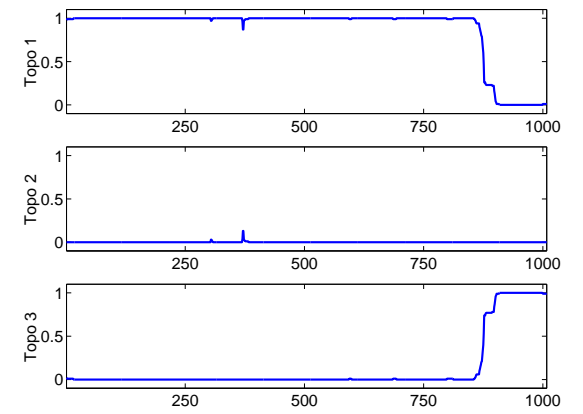
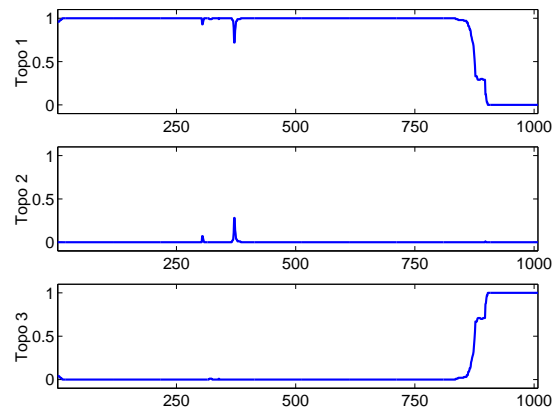
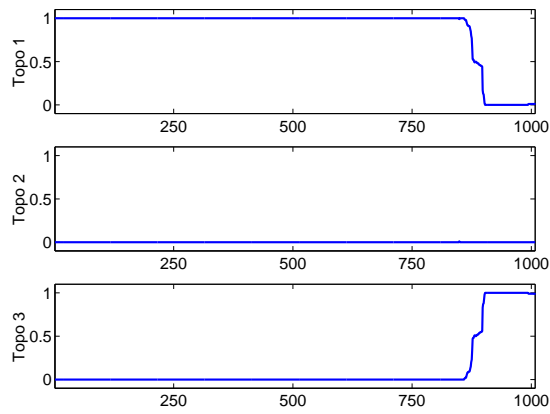
- 1) Maz56
- 2) Maz63
- 3) Maz63
- 4) Maz89



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



Dependence on the prior and the initialization



-
- Phylo-HMMs: Methodology
 - Phylo-HMMs: Applications
 - **Phylo-HMMs: Limitations**

Neisseria (Zhou & Spratt, 1992)

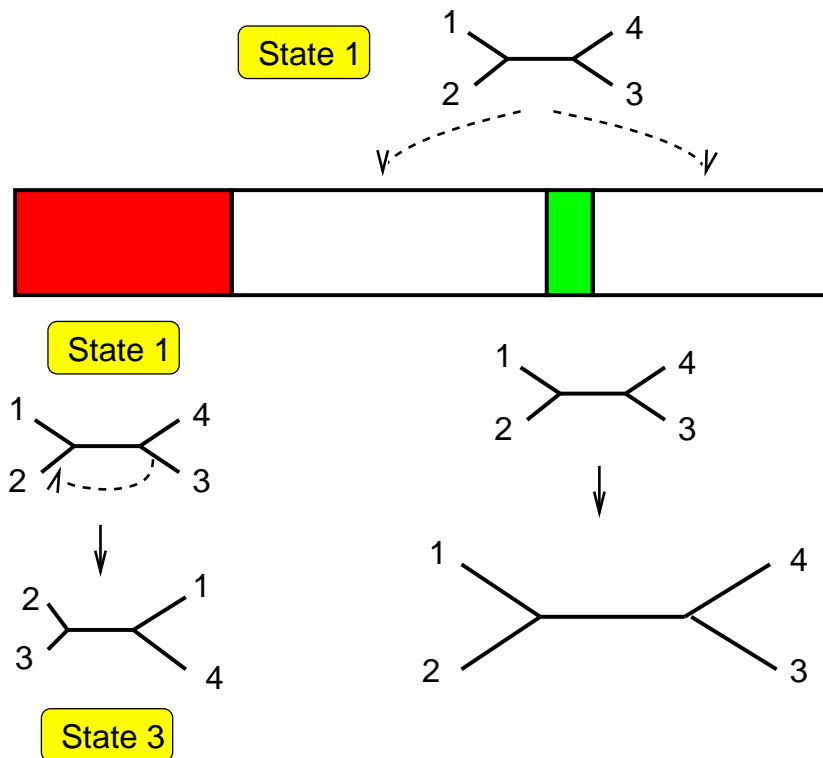
DNA alignment, 787 nucleotides (argF gene)

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

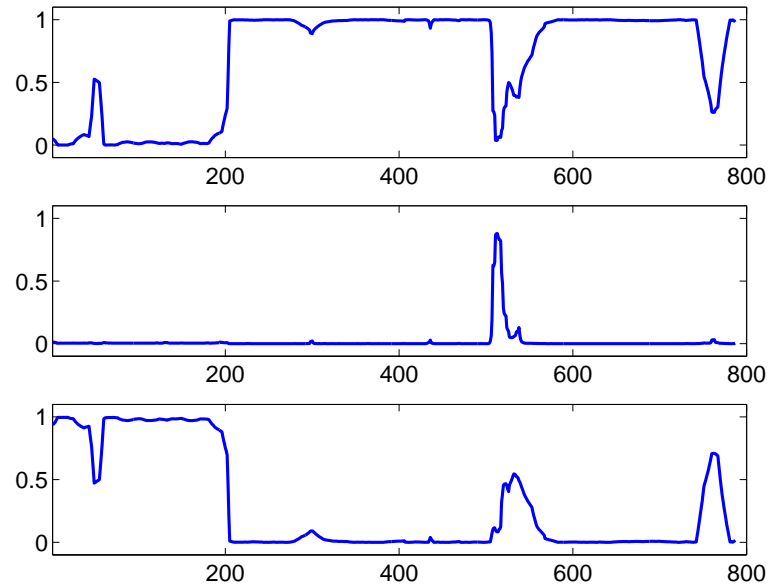
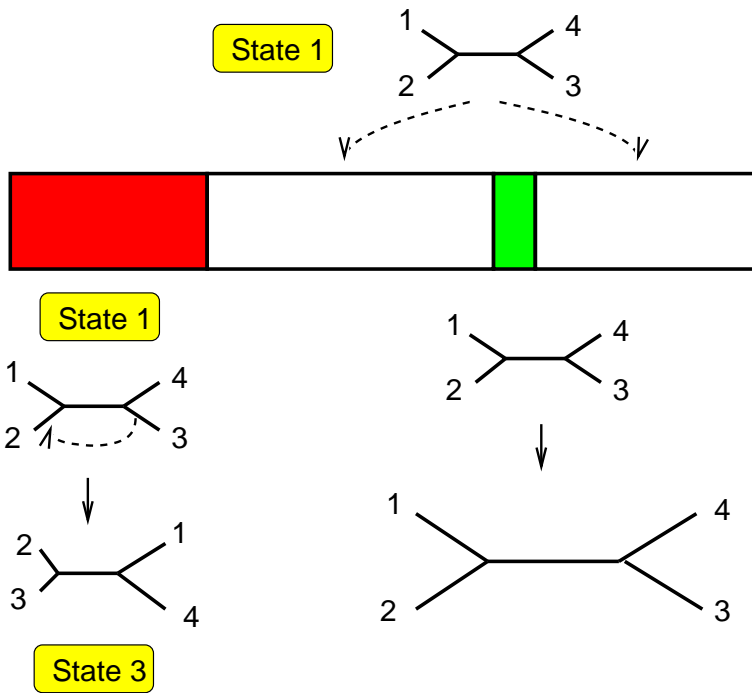
Neisseria (Zhou & Spratt, 1992)

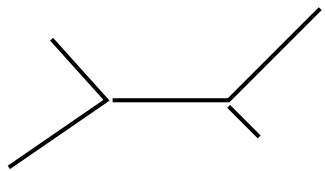
DNA alignment, 787 nucleotides (argF gene)

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

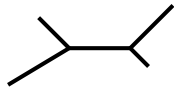


$P(S_t|\mathcal{D})$: Marginal posterior probability





$$w = \alpha t$$

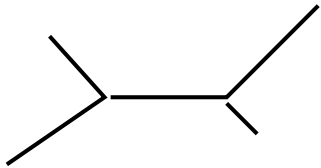


$$\alpha \rightarrow r^- \alpha$$

negative selective pressure

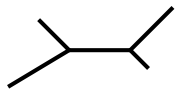
$$w \rightarrow r^- w$$

$$0 < r^- < 1$$



$$w = \alpha t$$

reference ("neutral") state

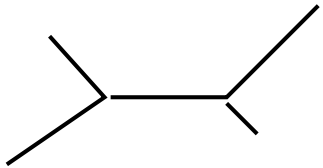


$$\alpha \rightarrow r^- \alpha$$

negative selective pressure

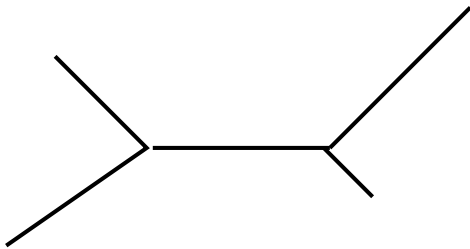
$$w \rightarrow r^- w$$

$$0 < r^- < 1$$



$$w = \alpha t$$

reference ("neutral") state



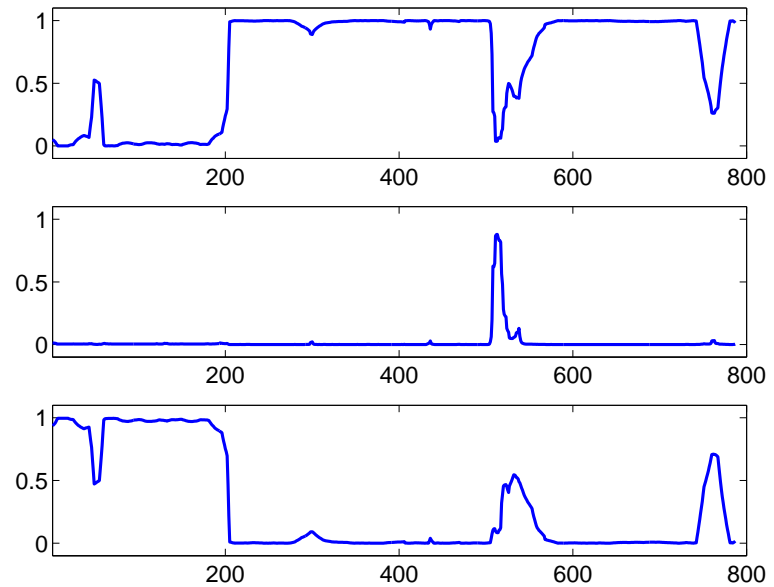
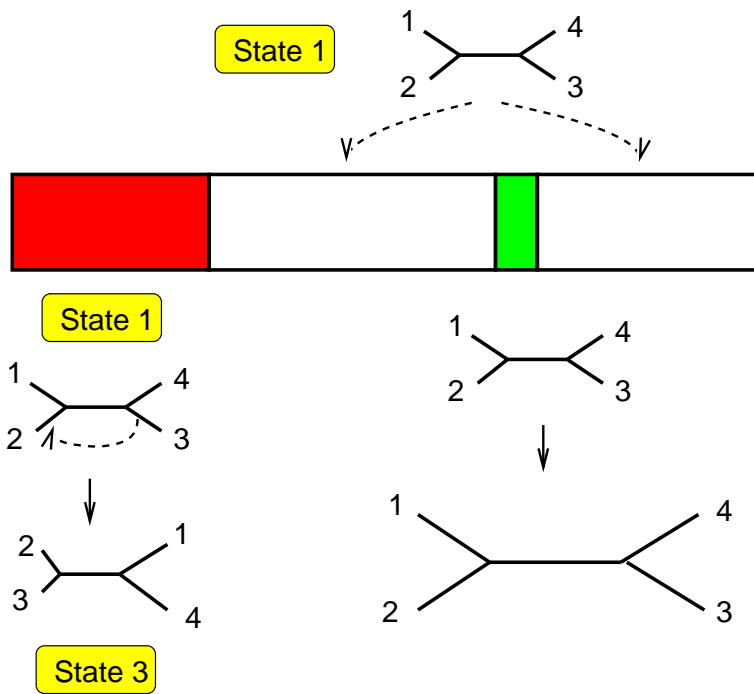
$$\alpha \rightarrow r^+ \alpha$$

positive selective pressure

$$w \rightarrow r^+ w$$

$$r^+ > 1$$

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



Problem:

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

Challenge

Distinguish between
recombination
and
rate heterogeneity

Overview

- Introduction: Phylogenetics
- Detecting recombination: Phylogenetic HMMs
Dirk Husmeier, Frank Wright and Grainne McGuire
2001-2003
- **Distinguishing between recombination and rate variation:**
Phylogenetic FHMMs
Dirk Husmeier, 2005
- Learning the number of genomic regions under selective pressure:
Phylogenetic FHMMs trained with RJMCMC
Wolfgang Lehrach and Dirk Husmeier, 2006

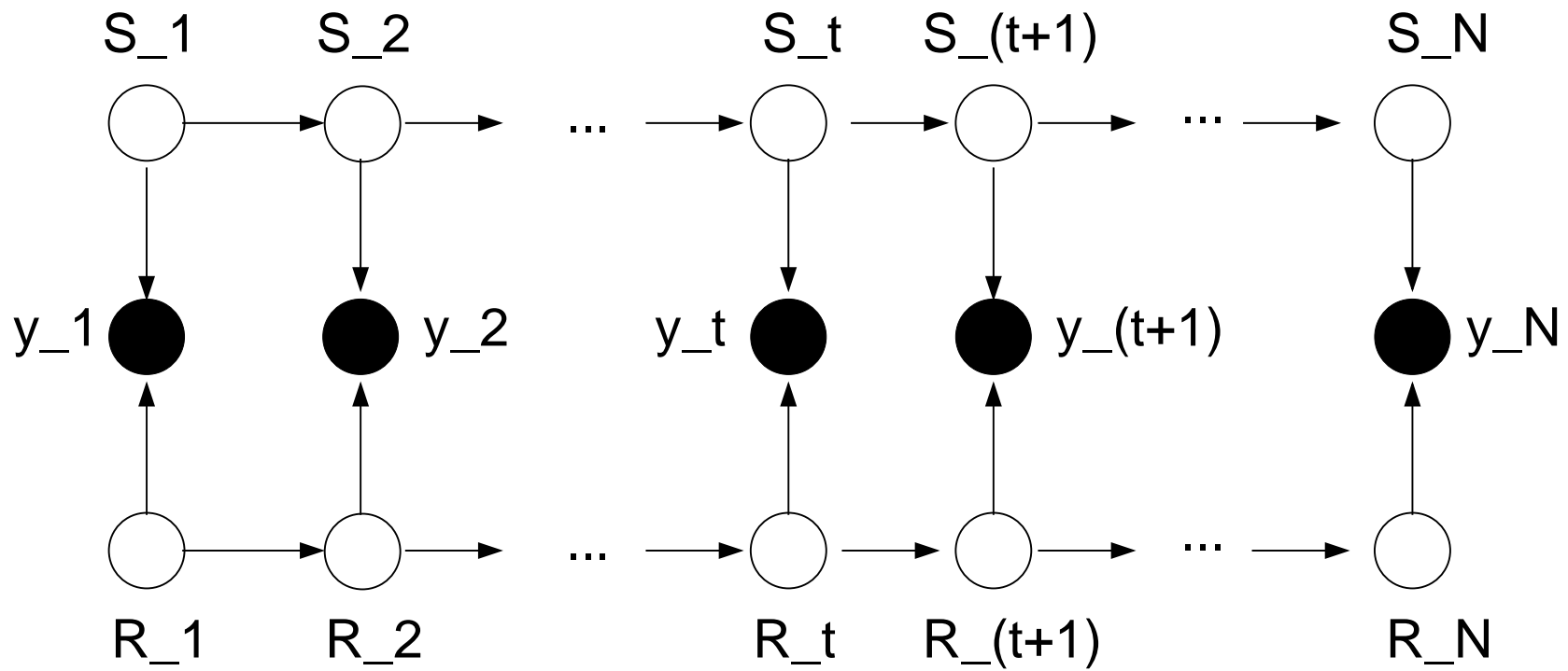
-
- Phylo-FHMMs: Methodology
 - Phylo-FHMMs: Applications

-
- **Phylo-FHMMs: Methodology**
 - Phylo-FHMMs: Applications

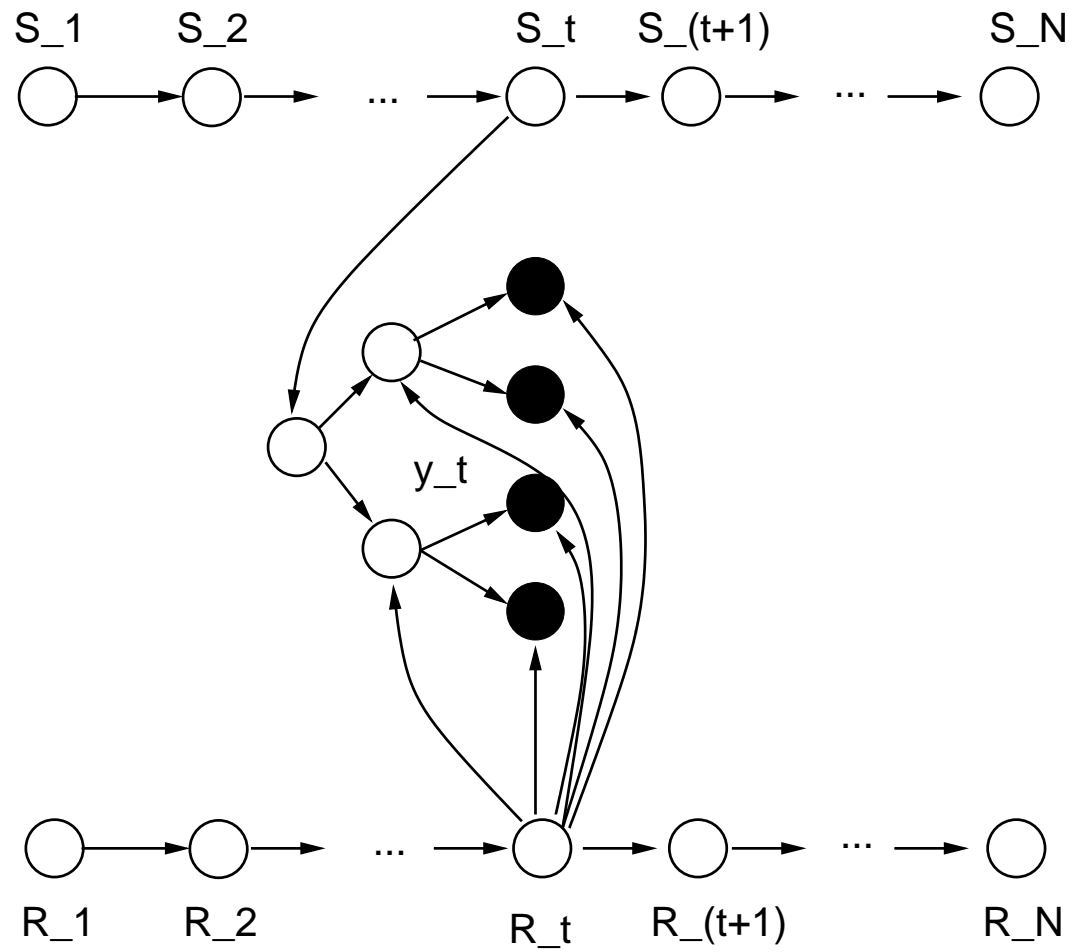
Distinguishing between
recombination
and
rate variation
with
factorial hidden Markov models (FHMMs)

Husmeier (2005)
Bioinformatics 21, Suppl. 2 (ECCB 05)

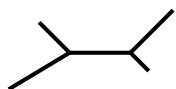
Factorial hidden Markov model (FHMM)



Phylo-FHMM



Rate states

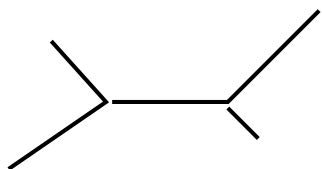


$$R = R^-$$

negative selective pressure

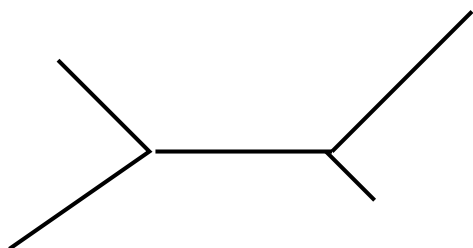
$$w \rightarrow r^- w$$

$$0 < r^- < 1$$



$$R = R^0$$

reference ("neutral") state



$$R = R^+$$

positive selective pressure

$$w \rightarrow r^+ w$$

$$r^+ > 1$$

Parameters

- Topology state sequences:

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

- Rate state sequences:

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

- Rate variation parameters:

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

- Branch lengths:

$$\mathbf{w}$$

- Transition probability parameters:

$$\nu_S, \nu_R$$

Sampling from the posterior distribution

- Sampling from

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

Sampling from the posterior distribution

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

Sampling from the posterior distribution

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

Sampling from the posterior distribution

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

Sampling from the posterior distribution

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

Sampling from the posterior distribution

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

Integrating out the branch lengths w

Integrating out the branch lengths \mathbf{w}

Suchard et al. (2003), JASA 98, 427–437

$$P(\mathbf{w}) = \prod_i P(w_i) = \frac{1}{r} \exp\left(-\frac{w_i}{r}\right)$$

Integrating out the branch lengths w

Suchard et al. (2003), JASA 98, 427–437

$$P(\mathbf{w}) = \prod_i P(w_i) = \frac{1}{r} \exp\left(-\frac{w_i}{r}\right)$$

Probability of nucleotide A mutating to B along a branch of length w :

$$P(Y|X, w) = \pi_y + A \exp(-Bw) + C \exp(-Dw)$$

A, B, C, D determined by the eigensystem of the rate matrix;
 π_y equilibrium frequency of nucleotide Y.

Integrating out the branch lengths w

Suchard et al. (2003), JASA 98, 427–437

$$P(\mathbf{w}) = \prod_i P(w_i) = \frac{1}{r} \exp\left(-\frac{w_i}{r}\right)$$

Probability of nucleotide A mutating to B along a branch of length w :

$$P(Y|X, w) = \pi_y + A \exp(-Bw) + C \exp(-Dw)$$

A, B, C, D determined by the eigensystem of the rate matrix;
 π_y equilibrium frequency of nucleotide Y.

$$\begin{aligned} P(Y|X, r) &= \int P(Y|X, w)P(w|r)dw \\ &= \pi_y + \frac{A}{r} \int \exp\left(-\left[B + \frac{1}{r}\right]w\right) dw + \frac{C}{r} \int \exp\left(-\left[D + \frac{1}{r}\right]w\right) dw \\ &= \pi_y + \frac{A}{1 + Br} + \frac{C}{1 + Dr} \end{aligned}$$

Sampling from the posterior distribution

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

FHMM: rate factors r

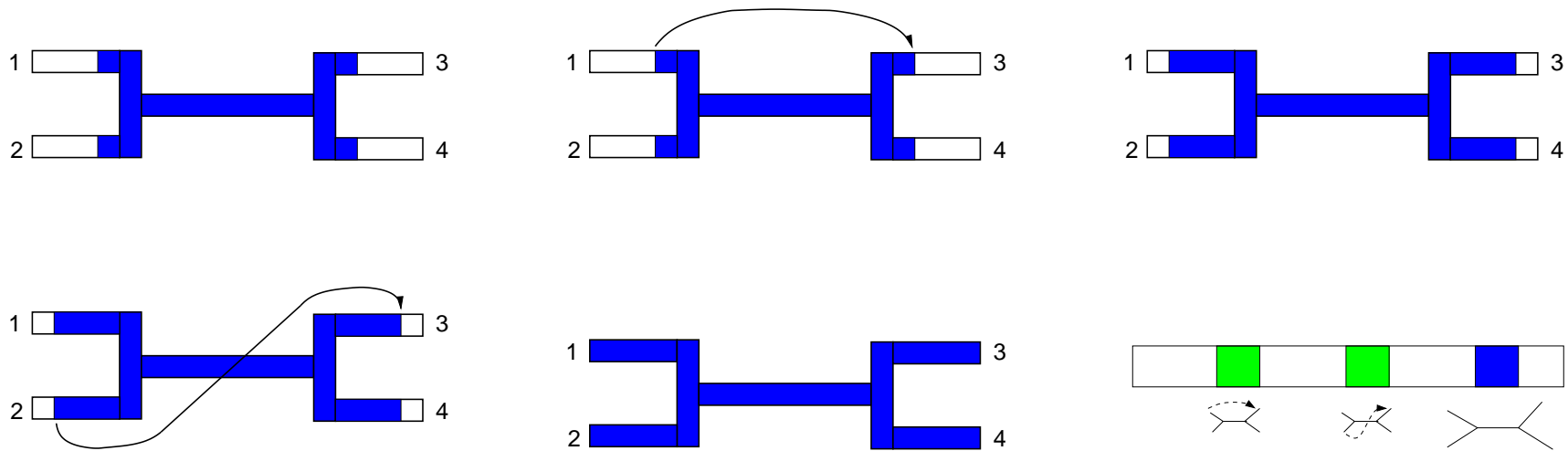
- 10 rate states
- Fixed rate factors
- Between $r = 0.001$ and $r = 100$ approximately uniform on a log scale

-
- Phylo-FHMMs: Methodology
 - **Phylo-FHMMs: Applications**

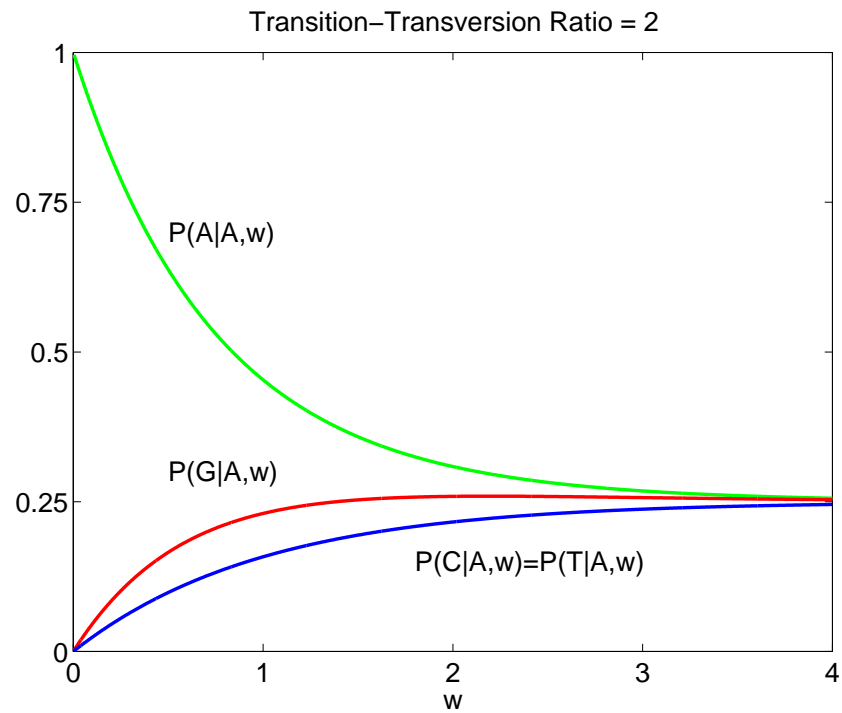
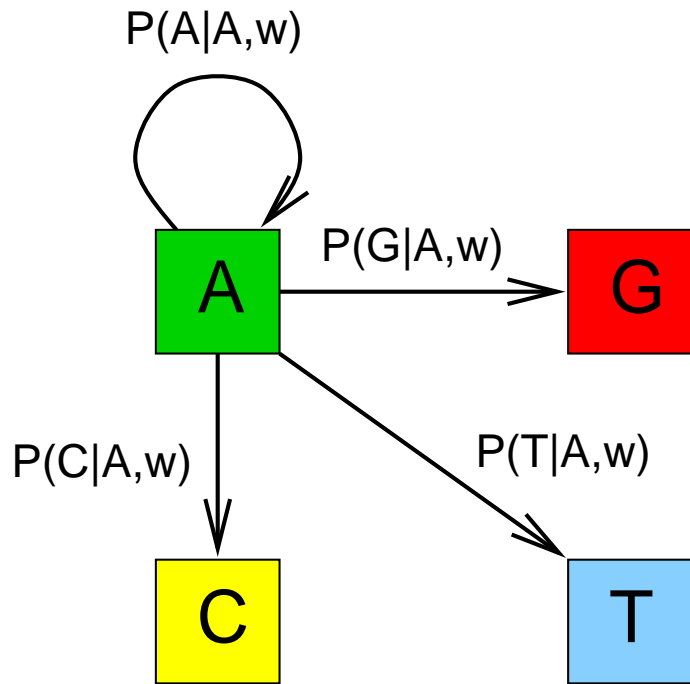
-
- Phylo-FHMMs: Methodology
 - **Phylo-FHMMs: Applications**
 - Synthetic data
 - Neisseria
 - HIV-1

-
- Phylo-FHMMs: Methodology
 - **Phylo-FHMMs: Applications**
 - **Synthetic data**
 - Neisseria
 - HIV-1

Synthetic simulation study

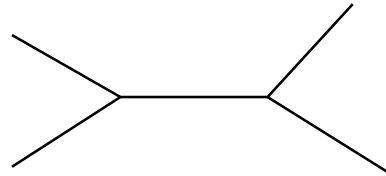
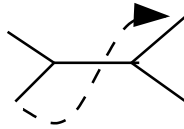
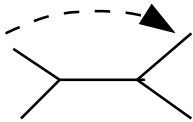


Mutation probabilities



branch length = mutation rate \times time

Synthetic simulation study



Synthetic simulation

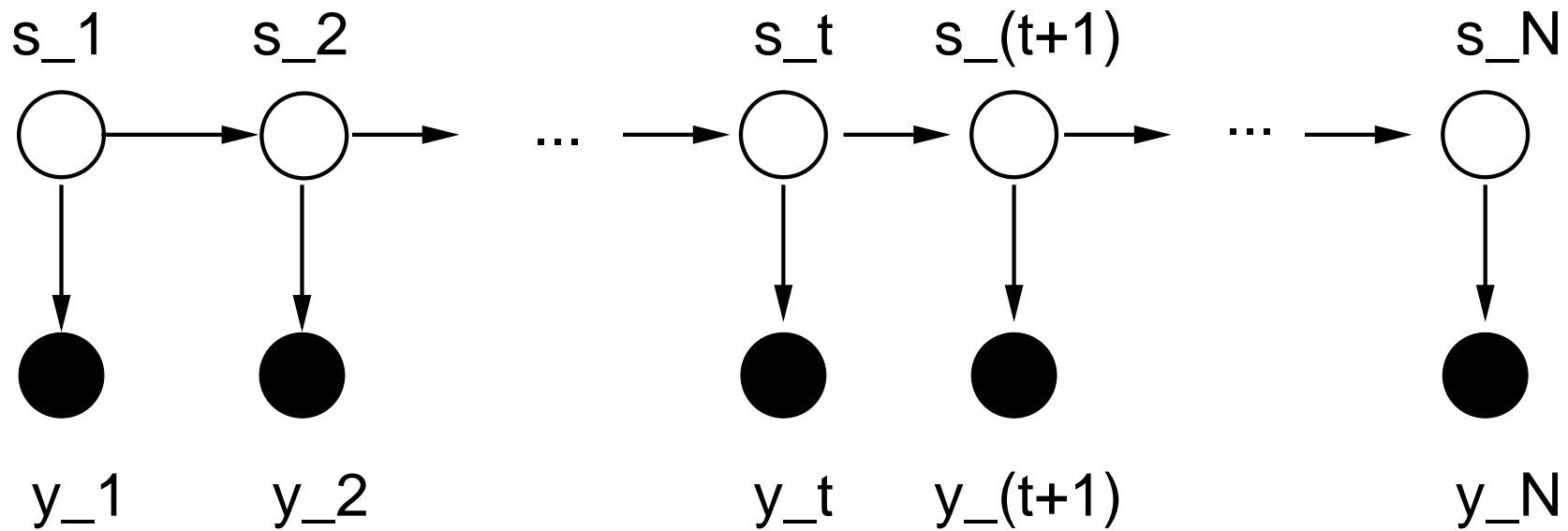
Two simulations:

- Long branch lengths: $\overline{w}_i = 0.1$
- Short branch lengths: $\overline{w}_i = 0.01$

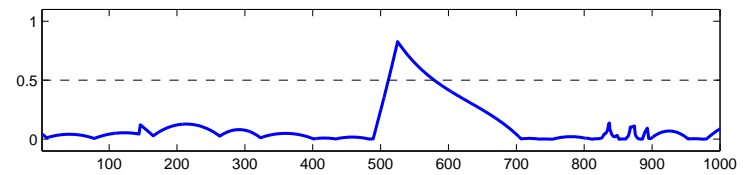
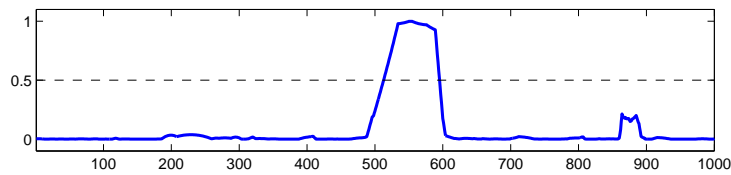
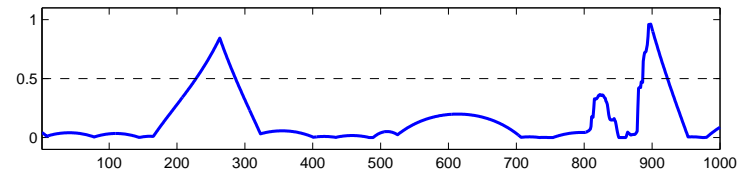
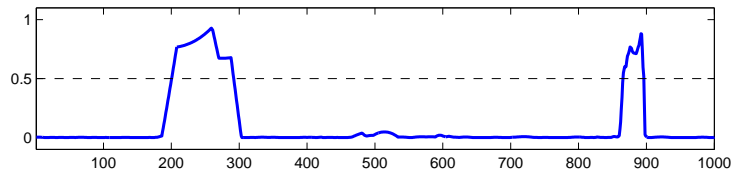
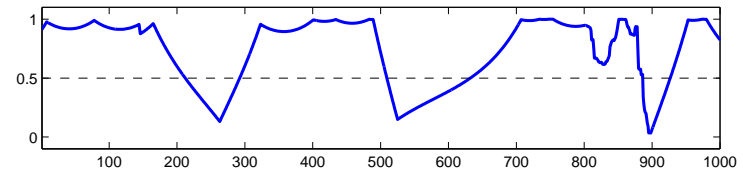
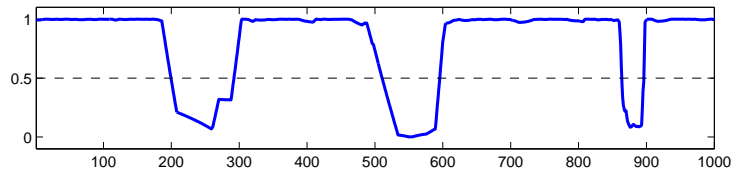
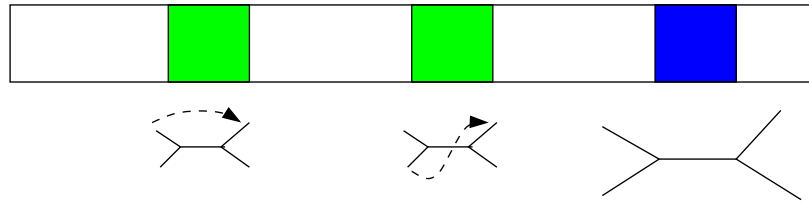
Differently diverged region:

$$w_i \rightarrow 10 \times w_i$$

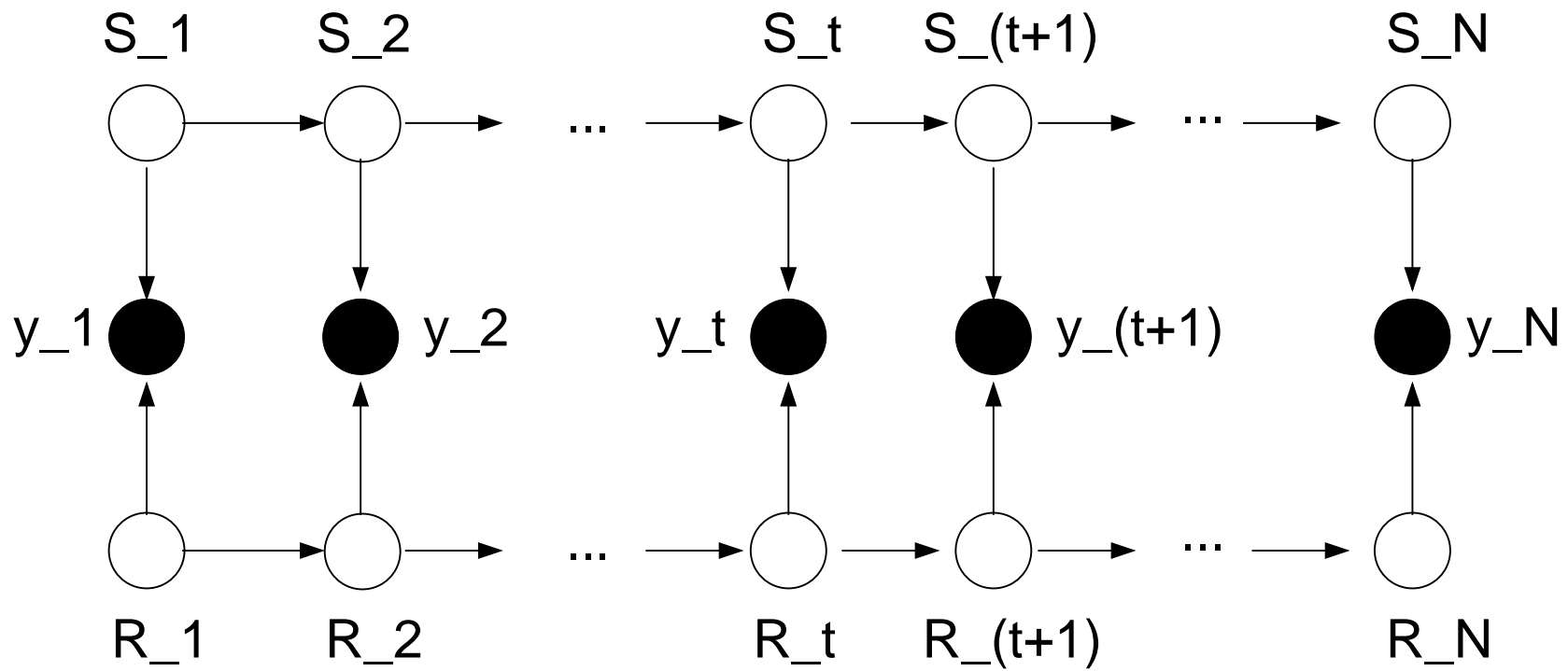
Hidden Markov model (Phylo-HMM)



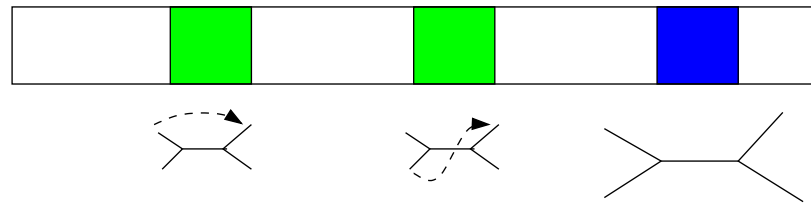
HMM. Left: long branches. Right: short branches



Factorial hidden Markov model (Phylo-FHMM)



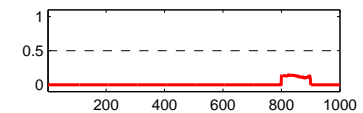
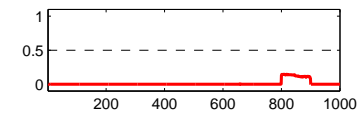
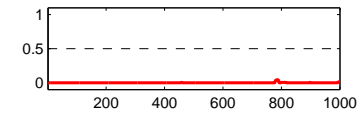
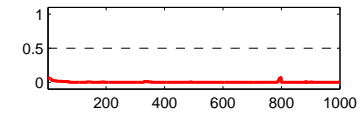
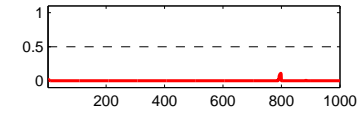
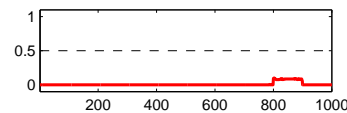
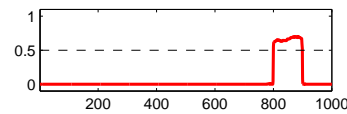
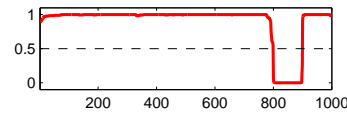
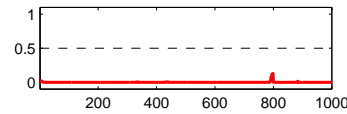
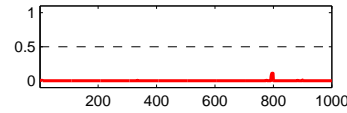
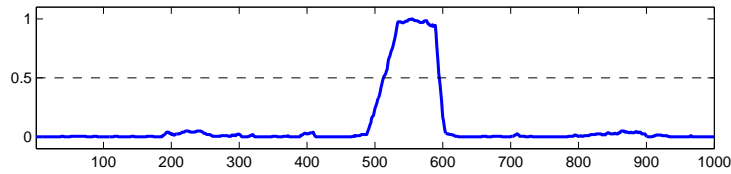
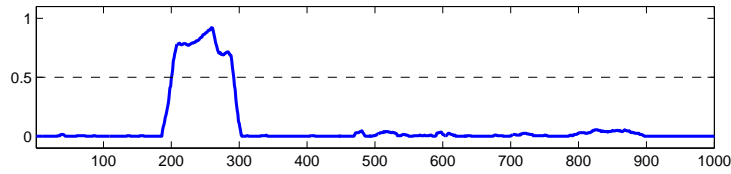
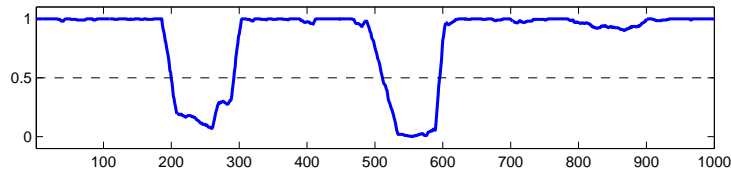
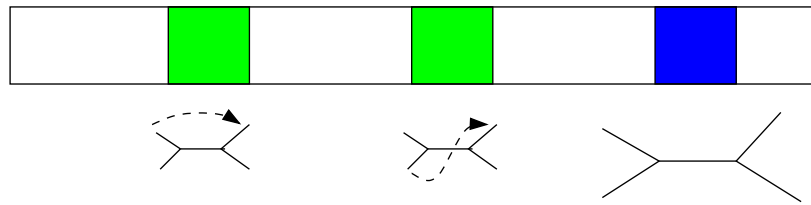
FHMM, long branch lengths



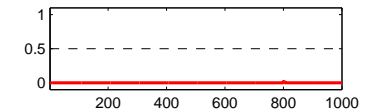
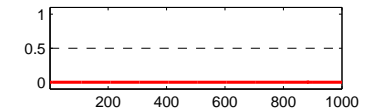
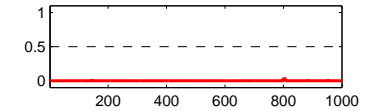
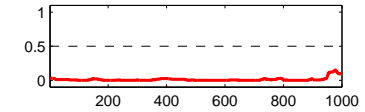
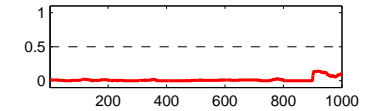
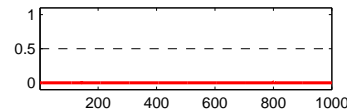
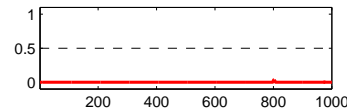
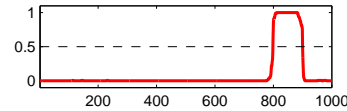
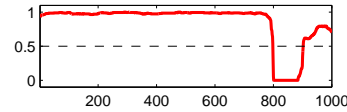
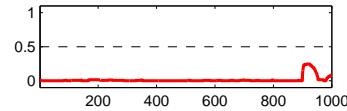
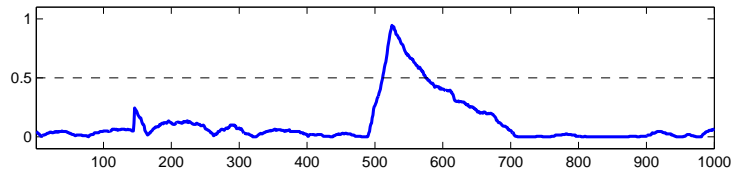
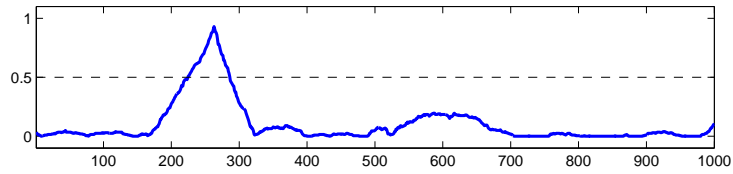
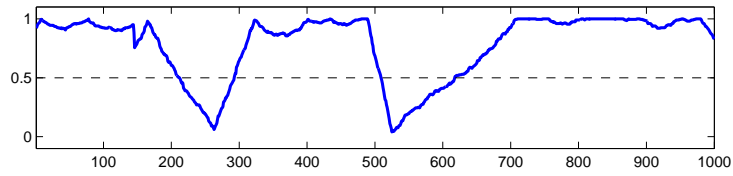
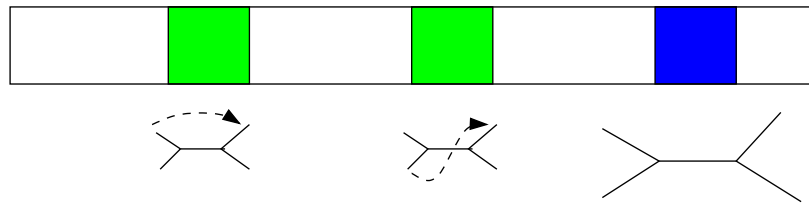
$P(S_t = \text{topo1} D)$
$P(S_t = \text{topo2} D)$
$P(S_t = \text{topo3} D)$

$P(r_t = 0.001 D)$	$P(r_t = 0.003 D)$
$P(r_t = 0.01 D)$	$P(r_t = 0.03 D)$
$P(r_t = 0.1 D)$	$P(r_t = 0.3 D)$
$P(r_t = 1 D)$	$P(r_t = 3 D)$
$P(r_t = 10 D)$	$P(r_t = 100 D)$

FHMM, long branch lengths



FHMM, short branch lengths

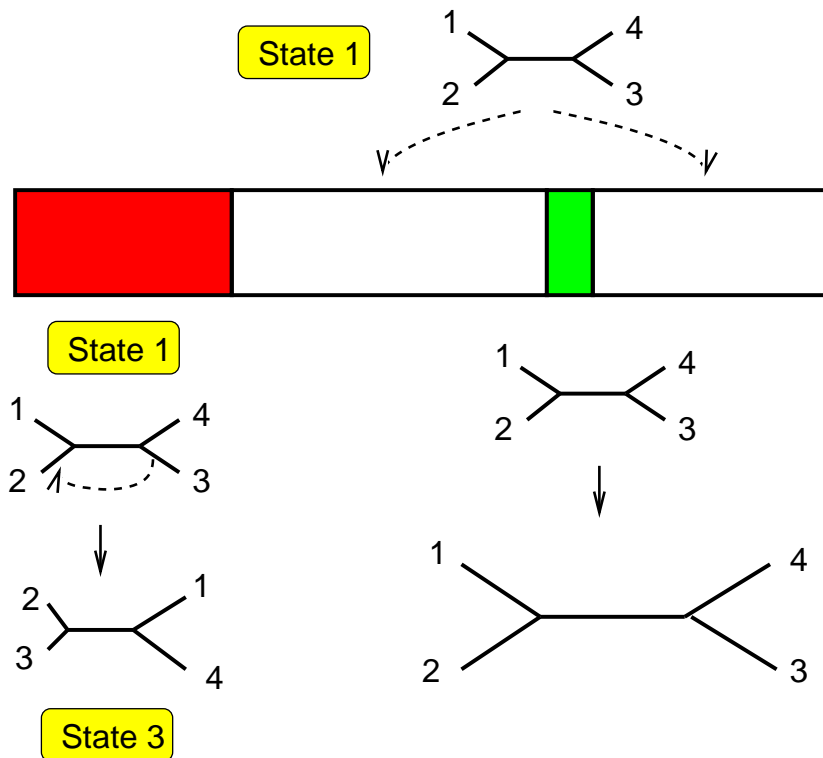


-
- Phylo-FHMMs: Methodology
 - **Phylo-FHMMs: Applications**
 - Synthetic data
 - **Neisseria**
 - HIV-1

Neisseria (Zhou & Spratt, 1992)

DNA alignment, 787 nucleotides (argF gene)

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

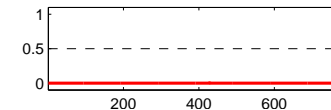
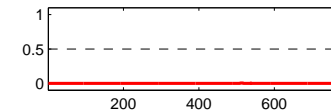
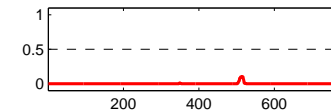
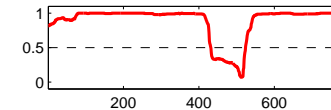
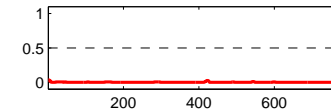
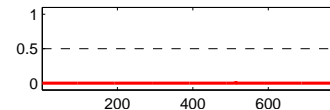
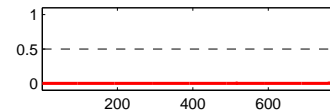
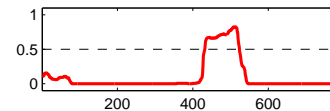
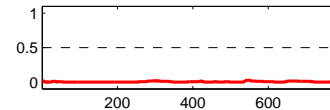
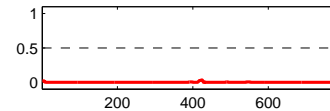
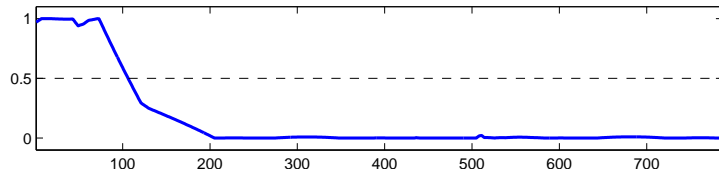
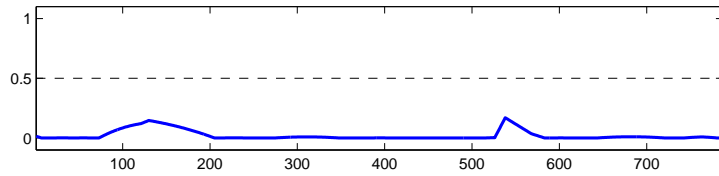
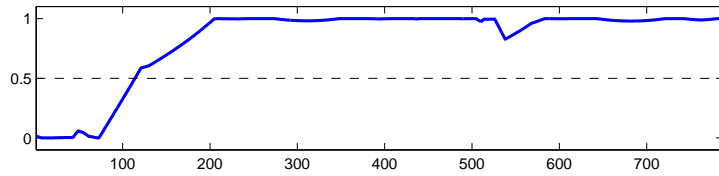


FHMM, Neisseria

$P(S_t = \text{topo1} D)$
$P(S_t = \text{topo2} D)$
$P(S_t = \text{topo3} D)$

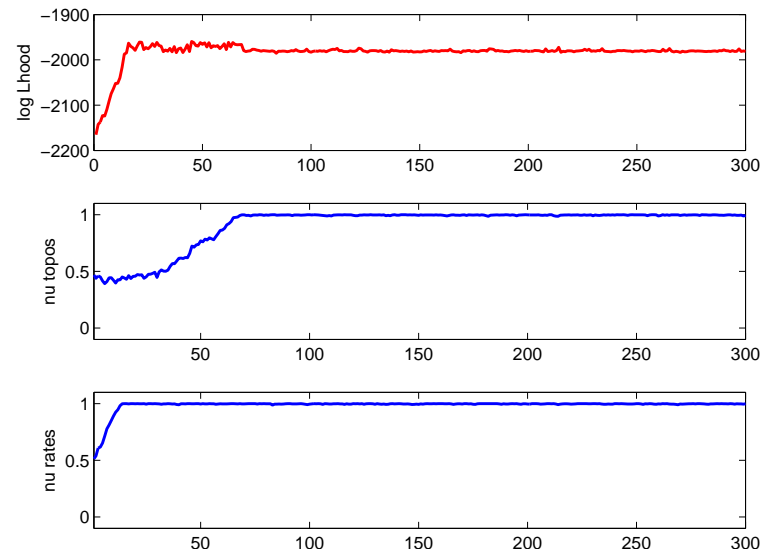
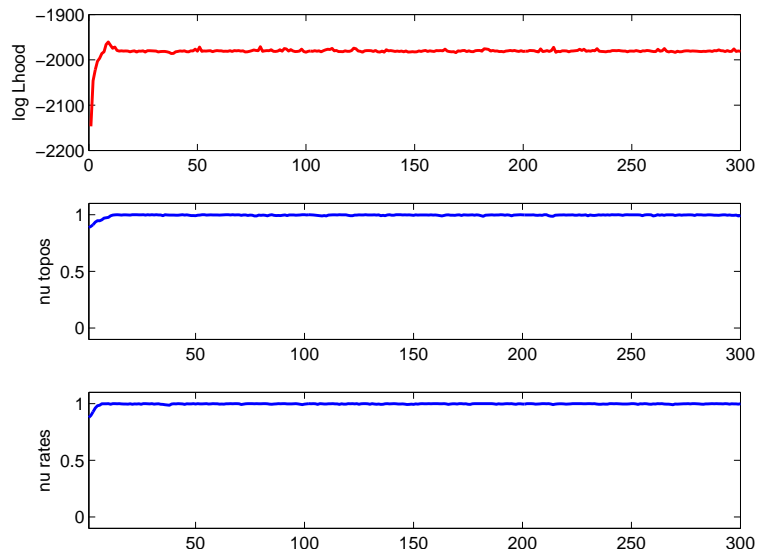
$P(r_t = 0.001 D)$	$P(r_t = 0.003 D)$
$P(r_t = 0.01 D)$	$P(r_t = 0.03 D)$
$P(r_t = 0.1 D)$	$P(r_t = 0.3 D)$
$P(r_t = 1 D)$	$P(r_t = 3 D)$
$P(r_t = 10 D)$	$P(r_t = 100 D)$

FHMM, Neisseria



FHMM, Neisseria

Initialisation:	$\nu_S = \nu_R = 0.9$	$\nu_S = \nu_R = 0.5$
Posterior average ν_S :	0.997 ± 0.003	0.996 ± 0.003
Posterior average ν_R :	0.998 ± 0.002	0.998 ± 0.002



-
- Phylo-FHMMs: Methodology
 - **Phylo-FHMMs: Applications**
 - Synthetic data
 - Neisseria
 - **HIV-1**

HIV-1 (KAL 153)

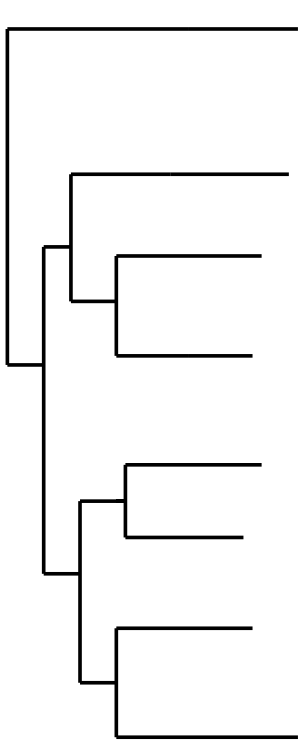
KAL-153









Caused epidemic outbreak of HIV-1 infection among intravenous drug users around Kaliningrad , Russia, in October 1996.

Rate of newly diagnosed seropositive individuals:

Less than 1/month → over 100/month

Complete genome (>8500 nucleotides) aligned with consensus sequences of subtypes A, B, and F.

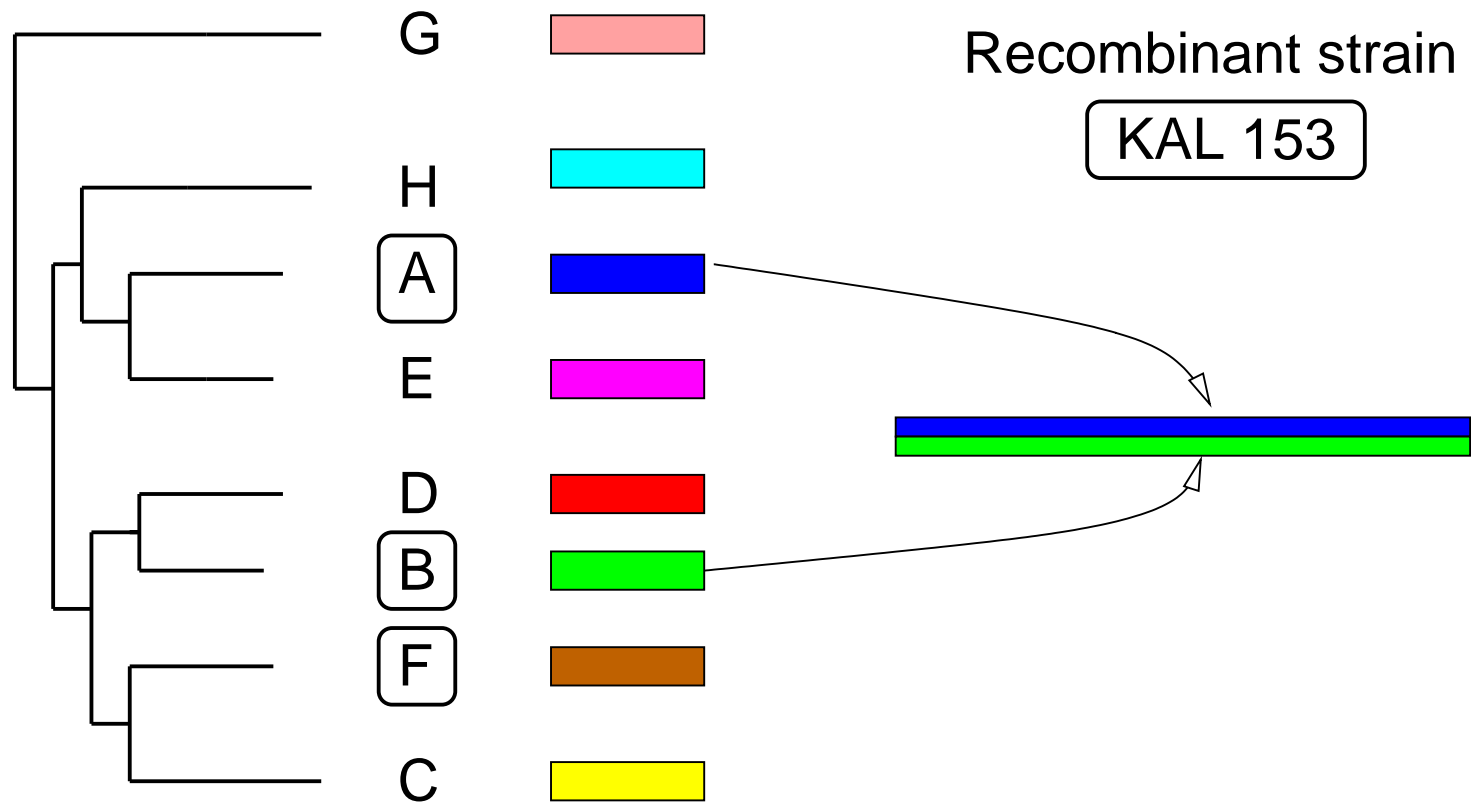


- G 
- H 
- A 
- E 
- D 
- B 
- F 
- C 

Recombinant strain
KAL 153



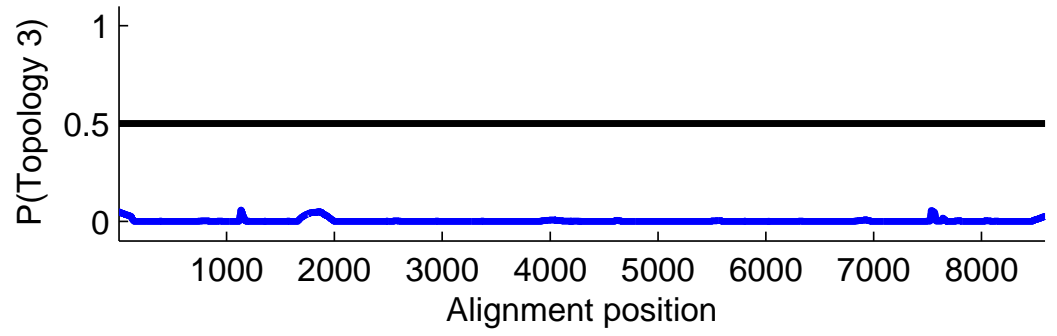
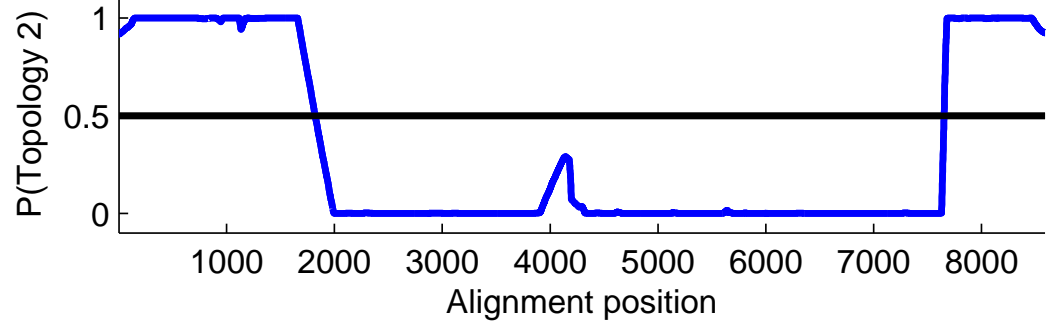
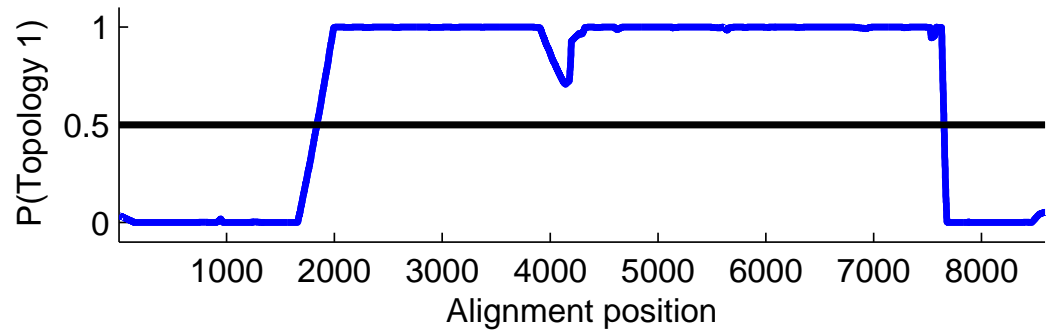
Recombination in HIV 1



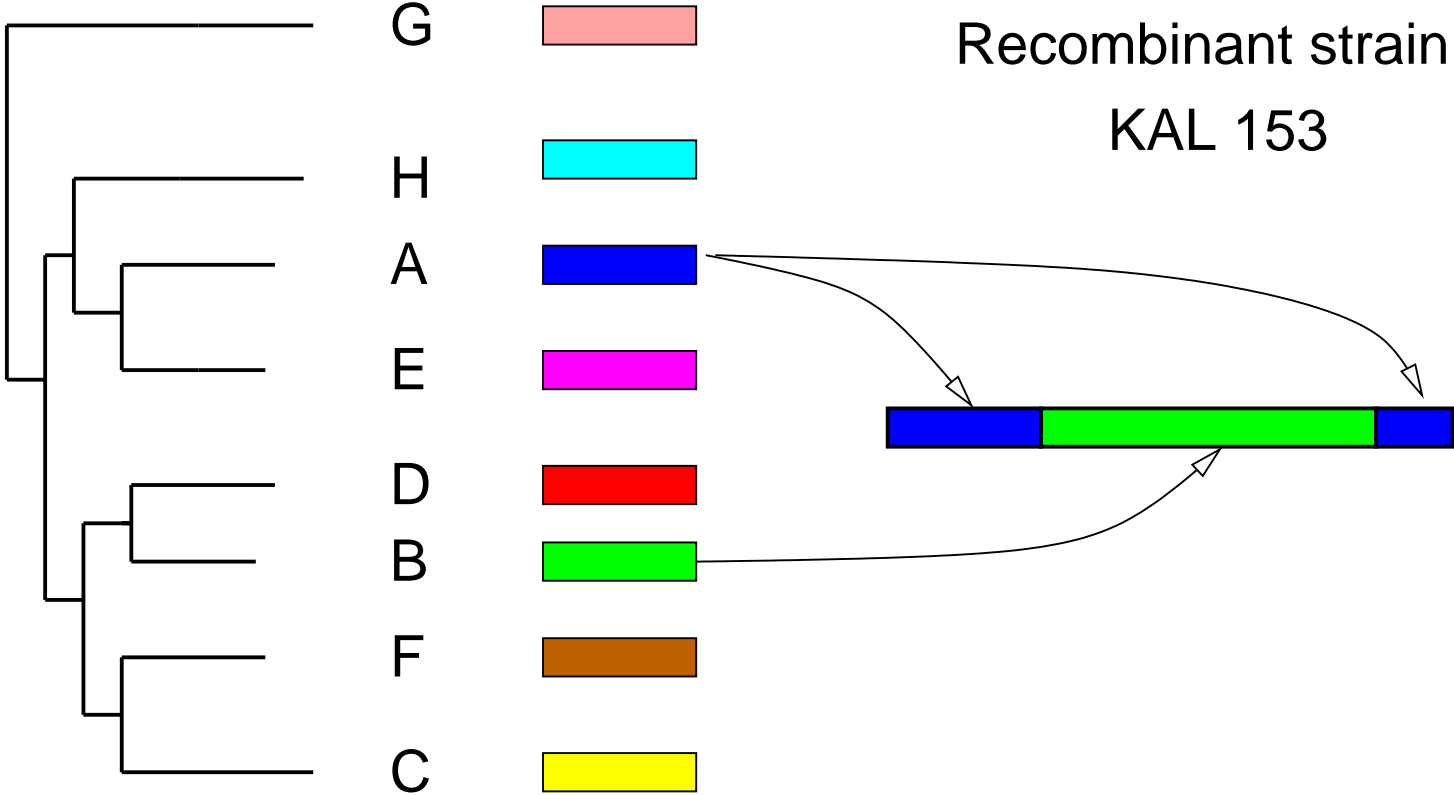
Topo 1: B - KAL 153

Topo 2: A - KAL 153

Topo 3: F - KAL 153



Recombination in KAL 153



Suchard, Weiss, Dorman and Sinsheimer (2003)

Journal of the American Statistical Association 98, 427–437

Bayesian multiple change-point model

(Topo,r, Q)



(Topo1,r1, Q1)



(Topo2, r2, Q2)



(Topo1,r1, Q1)



(Topo2, r2, Q2)



(Topo3,r3, Q3)



Minin, Dorman, Fang and Suchard (2005)

Bioinformatics 21, 3034–3042

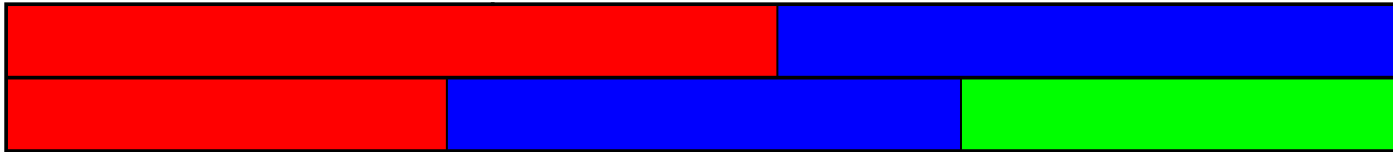
Bayesian dual multiple change-point model

(Topo,r, Q)



Topo1

Topo2



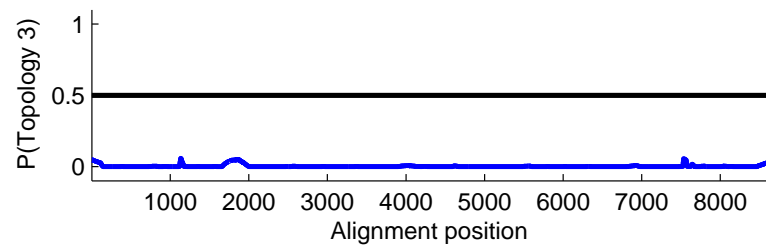
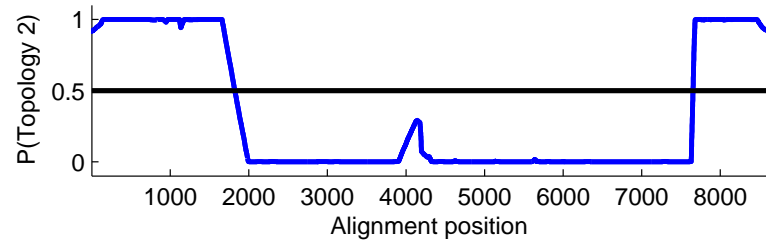
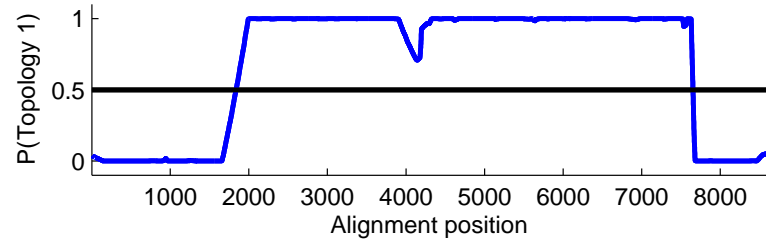
(r1, Q1)

(r2, Q2)

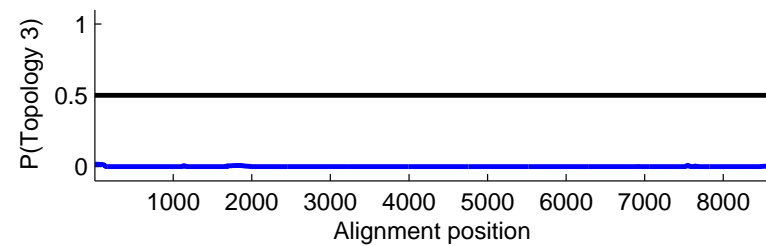
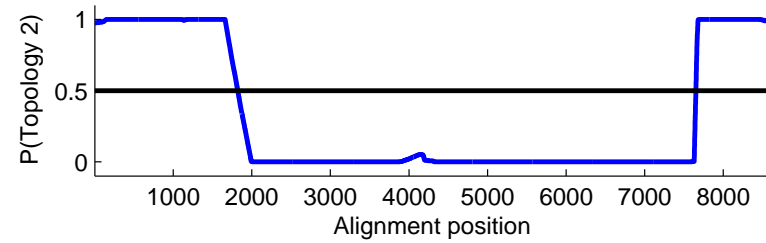
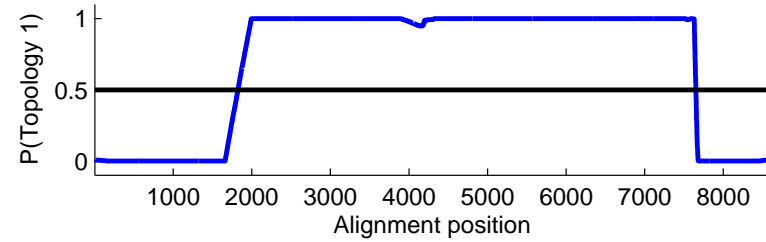
(r3, Q3)



Comparison of predictions



Phyl-FHMM



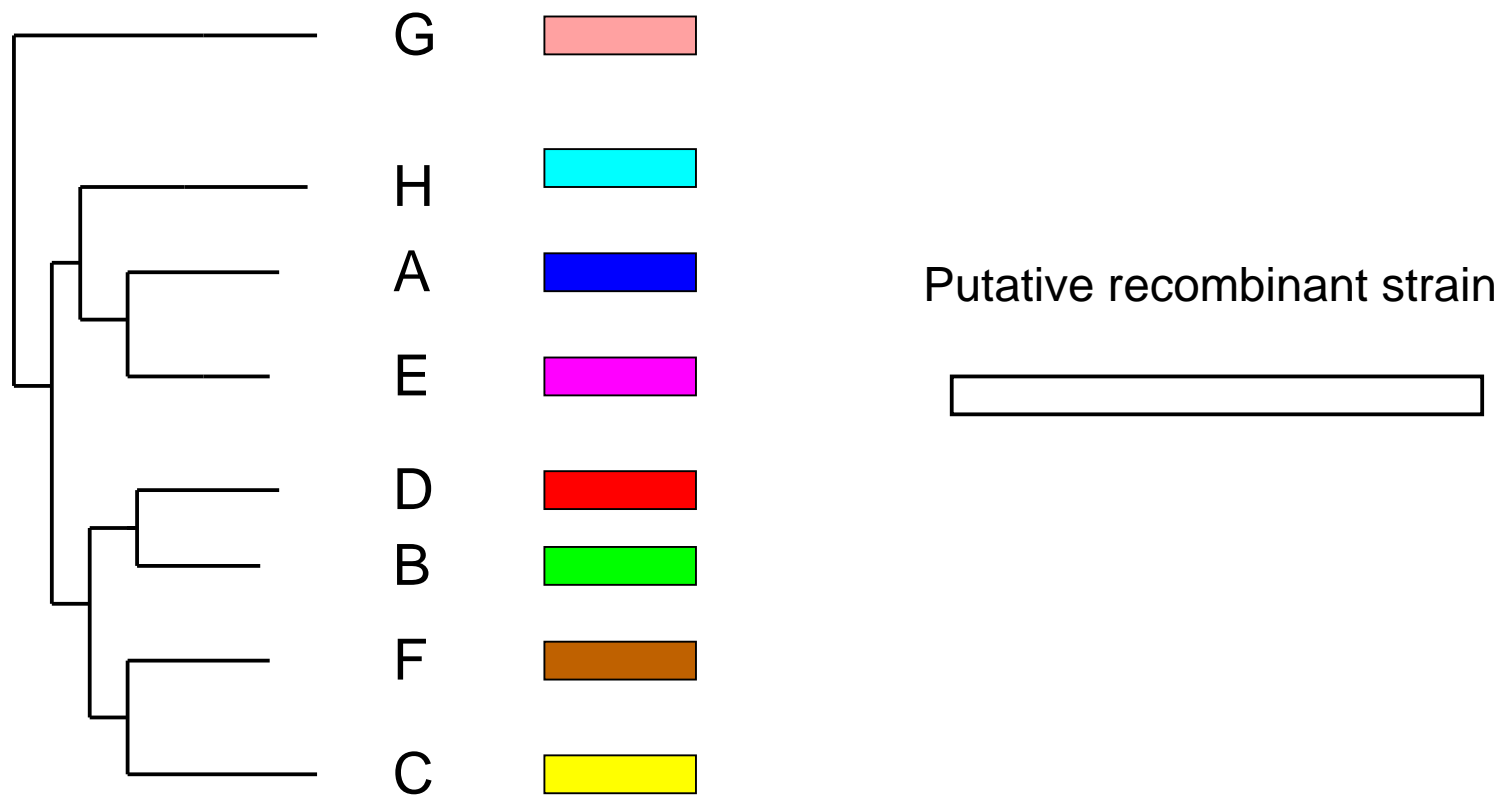
Dual change-point model

Extension to more than 4 sequences

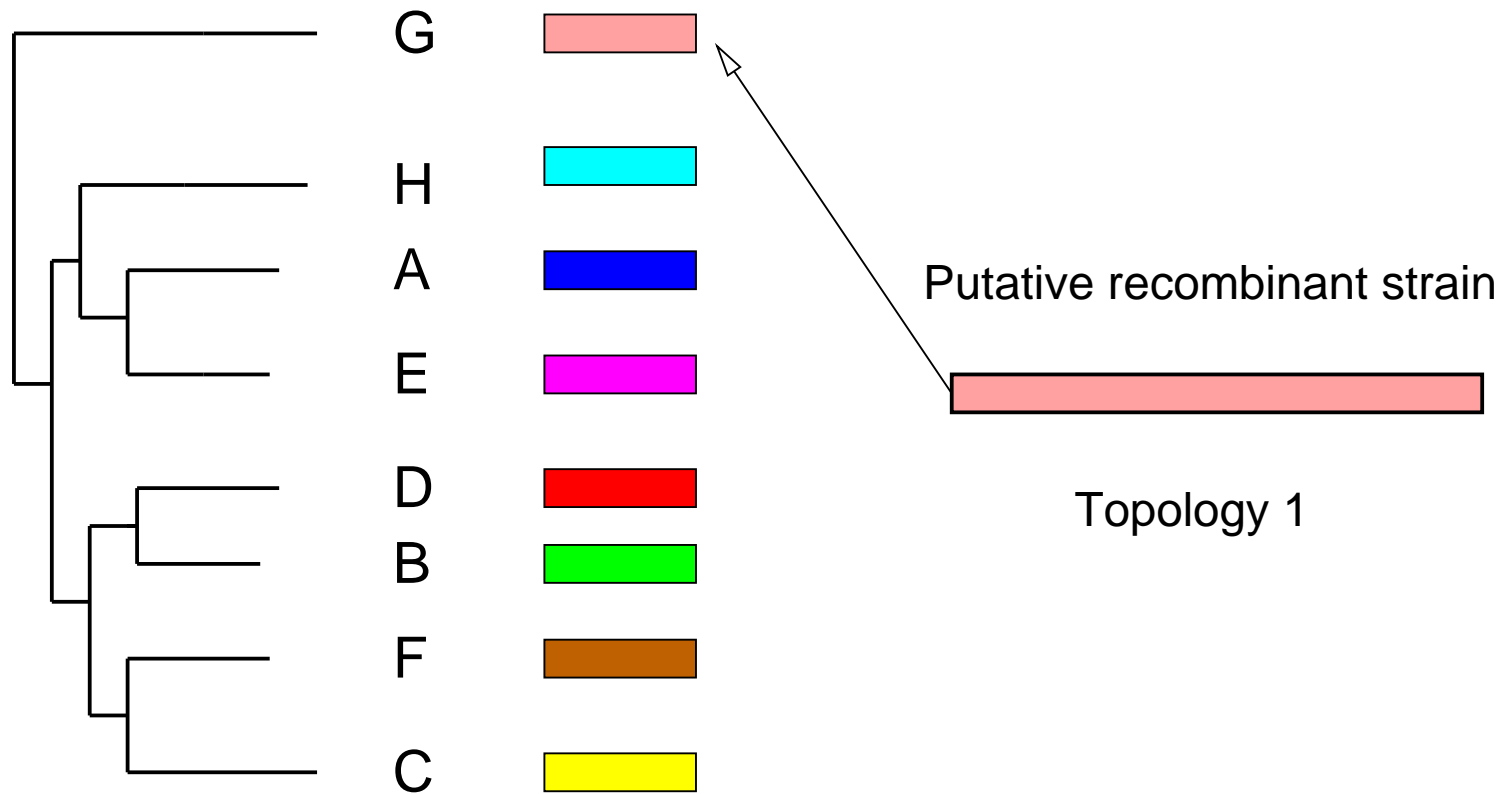
Alex Mantzaris

MSc project
BioSS and Edinburgh University
Summer 2006

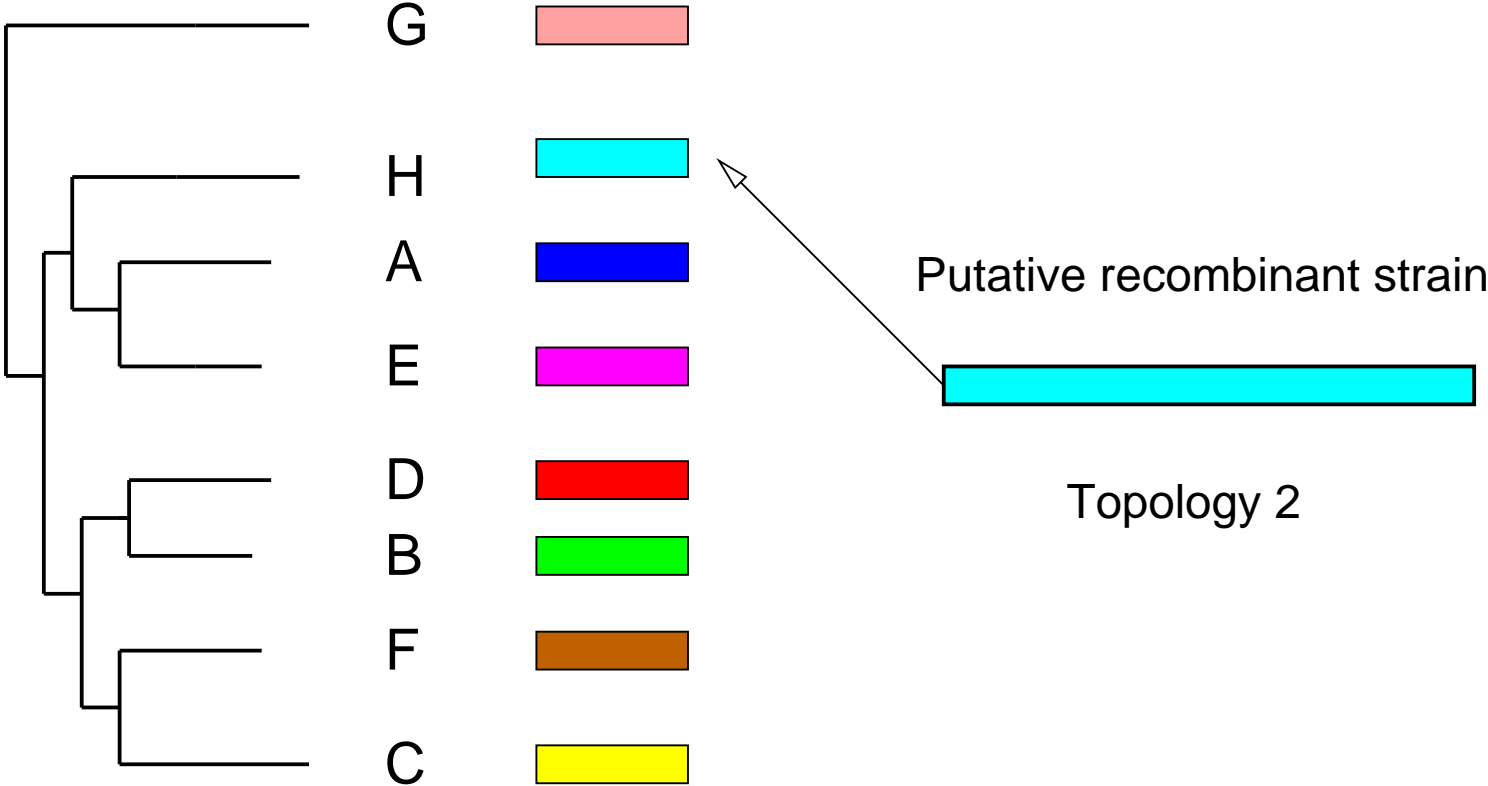
Extend method beyond 4 sequences



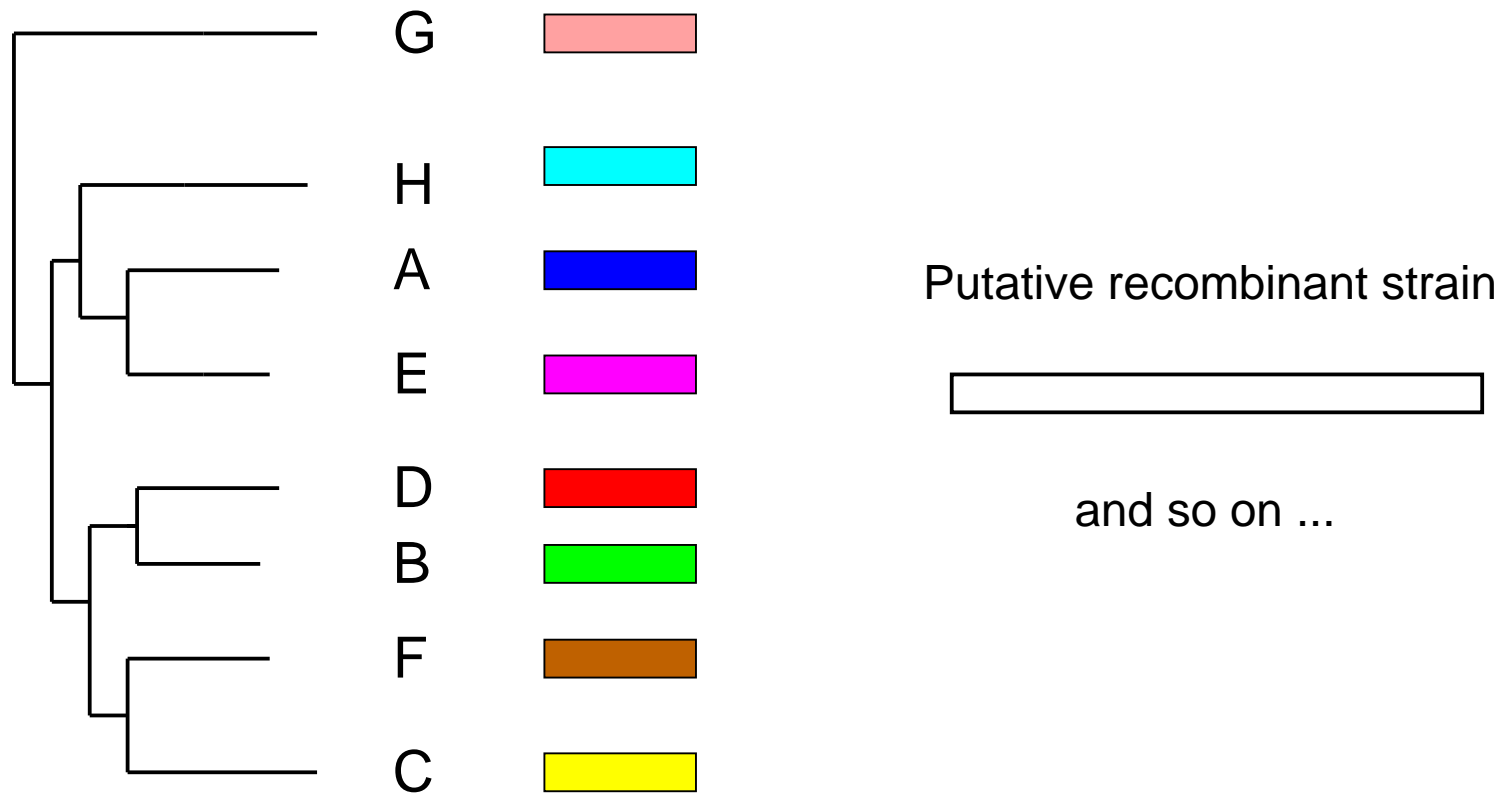
Extend method beyond 4 sequences



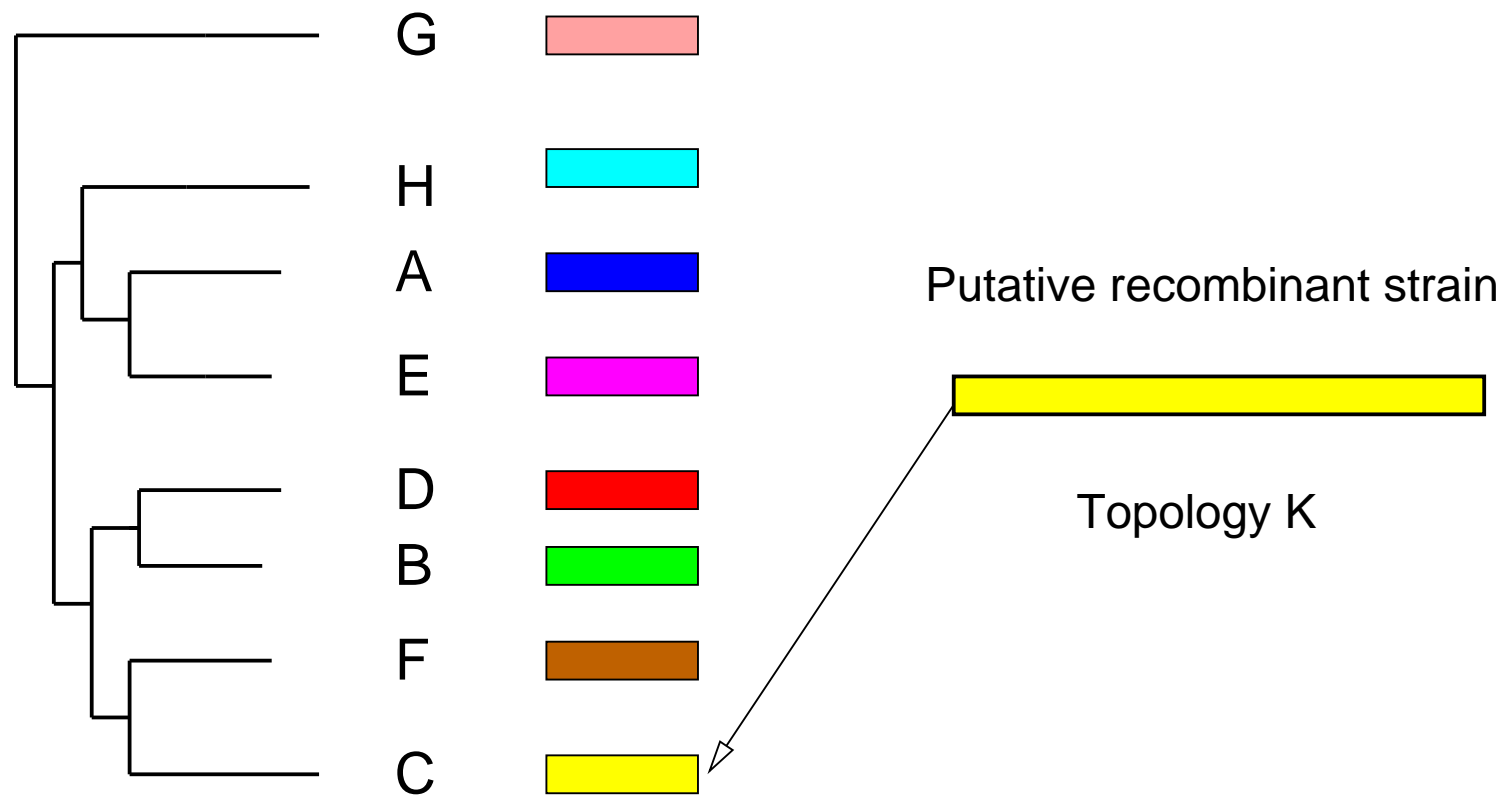
Extend method beyond 4 sequences



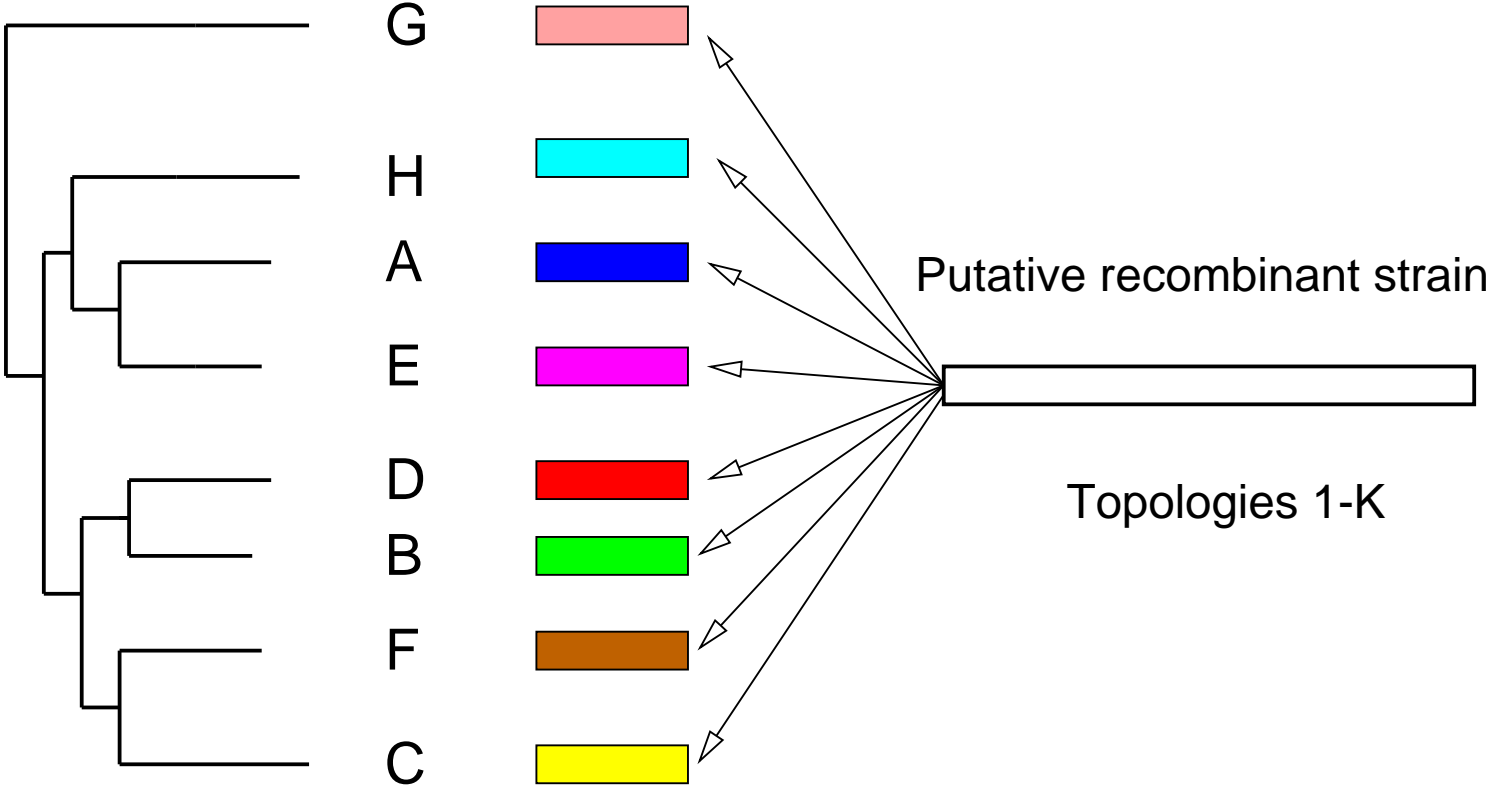
Extend method beyond 4 sequences



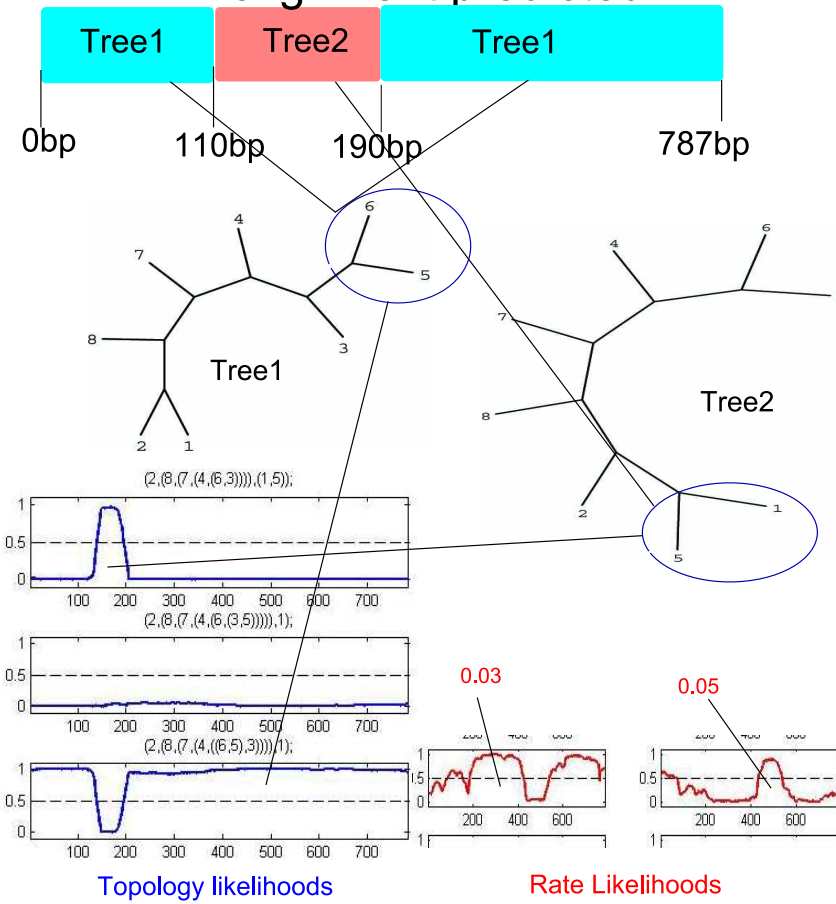
Extend method beyond 4 sequences



Extend method beyond 4 sequences



DNA alignment predicted



Application to 8 strains of Neisseria

Overview

- Introduction: Phylogenetics
- Detecting recombination: Phylogenetic HMMs
Dirk Husmeier, Frank Wright and Grainne McGuire
2001-2003
- Distinguishing between recombination and rate variation:
Phylogenetic FHMMs
Dirk Husmeier, 2005
- **Learning the number of genomic regions
under selective pressure:
Phylogenetic FHMMs trained with RJMCMC
Wolfgang Lehrach and Dirk Husmeier, 2006**

Phylogenetic FHMMs trained with RJMCMC

Wolfgang Lehrach

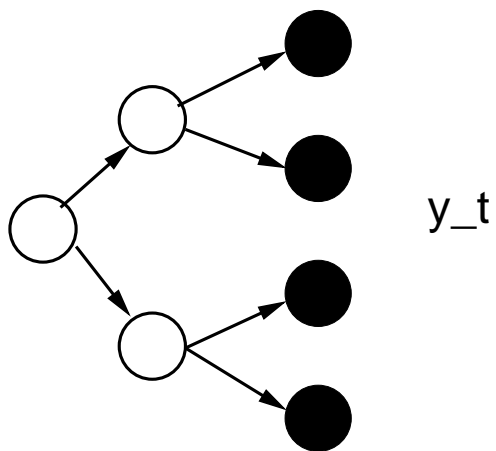
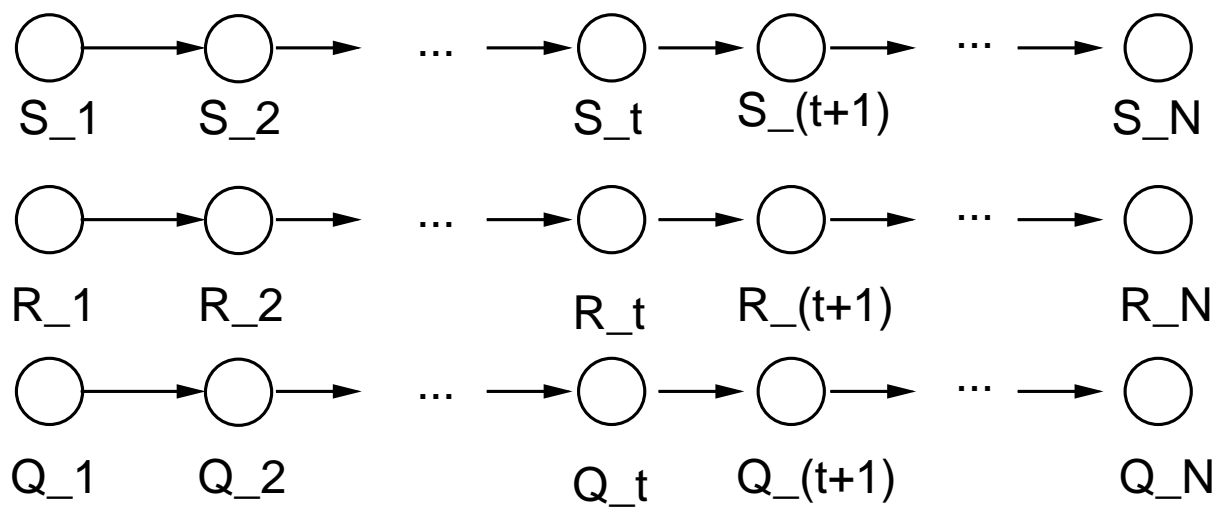
PhD project
BioSS and Edinburgh University
2006

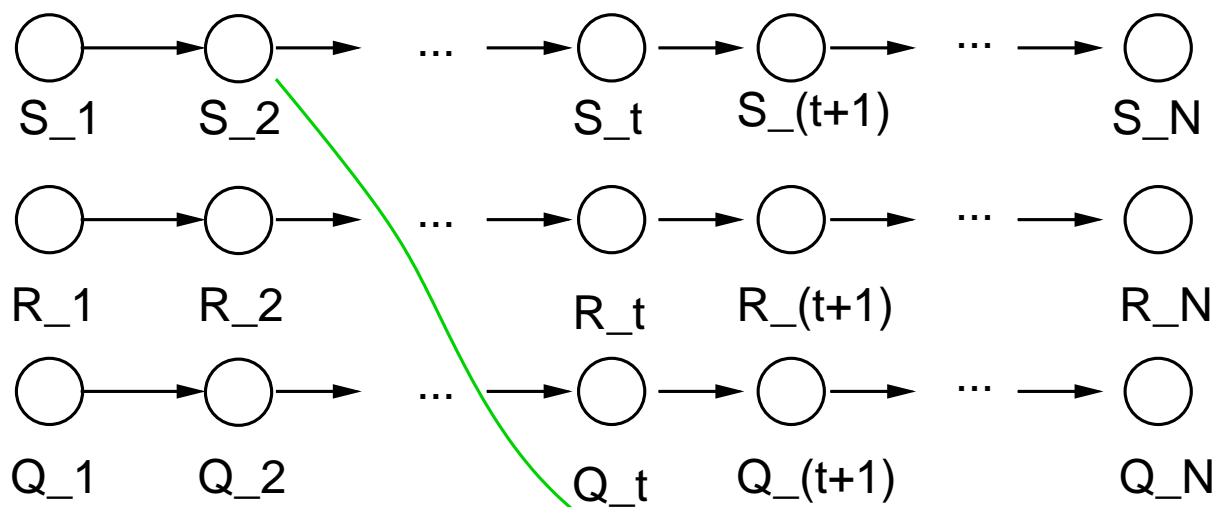
Work in progress

- Decoupling rate variation from nucleotide substitution processes
- Adaptable rate factors \mathbf{r}
- Adaptable dimension of \mathbf{R} with RJMCMC

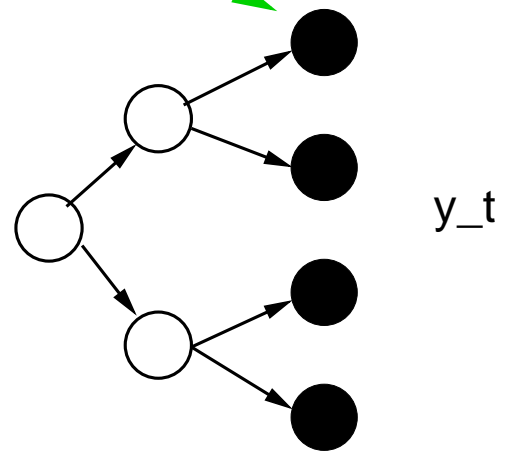
Work in progress

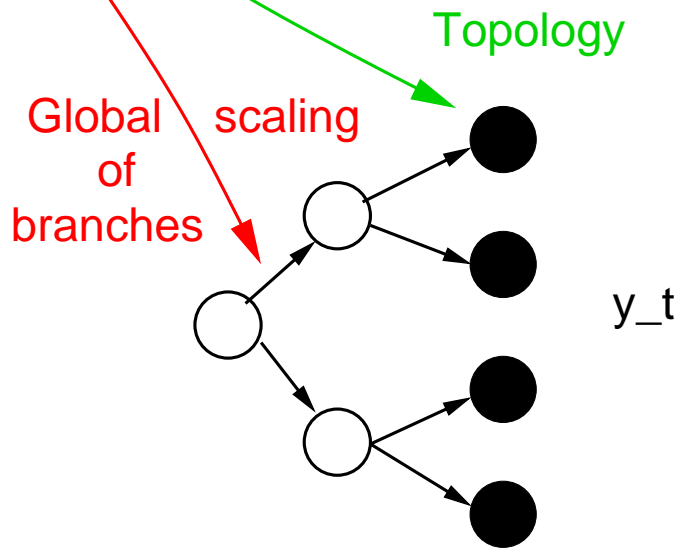
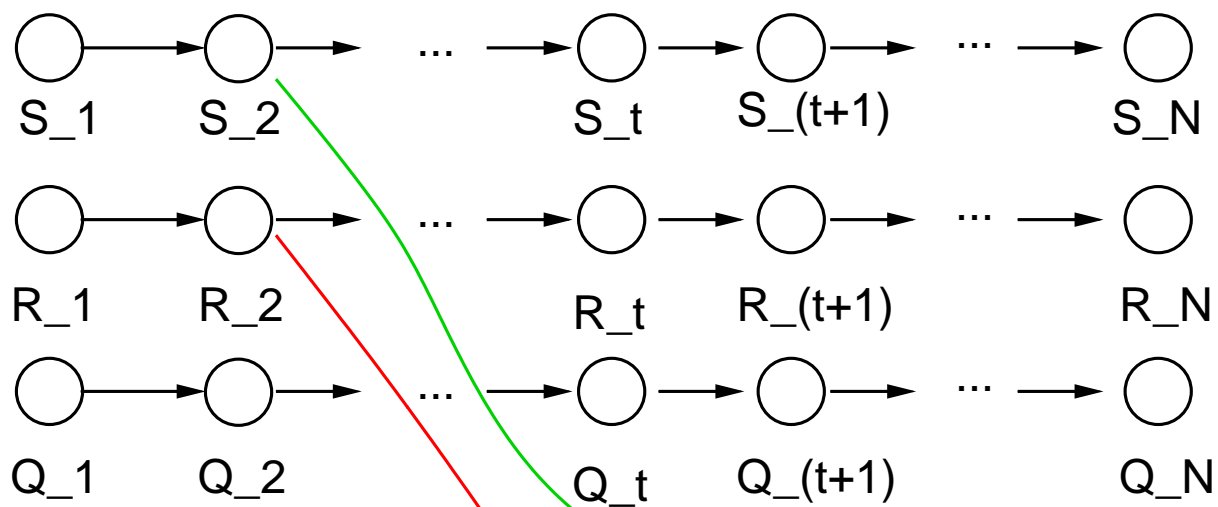
- Decoupling rate variation from nucleotide substitution processes
- Adaptable rate factors \mathbf{r}
- Adaptable dimension of \mathbf{R} with RJMCMC

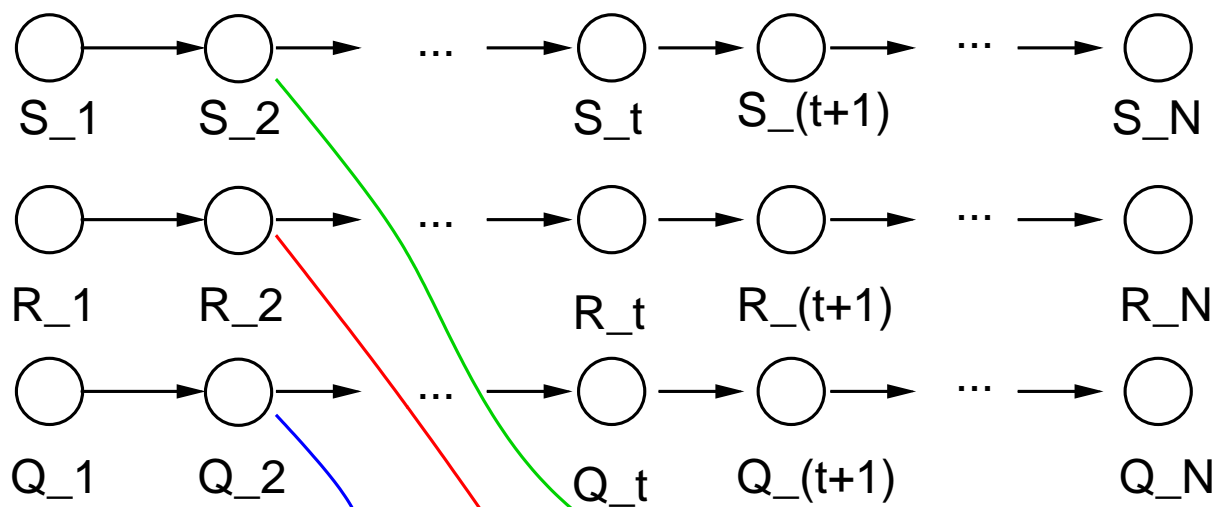




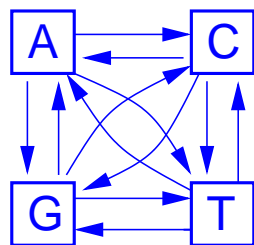
Topology





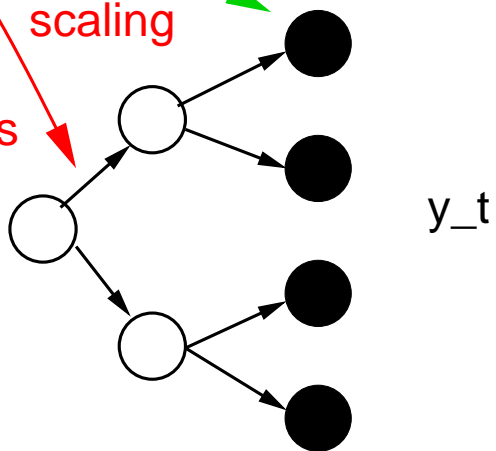


Nucleotide substitution model

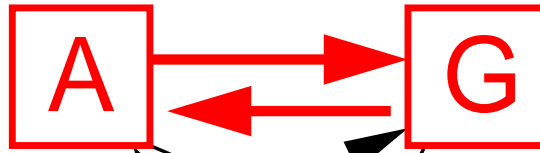


Global scaling of branches

Topology



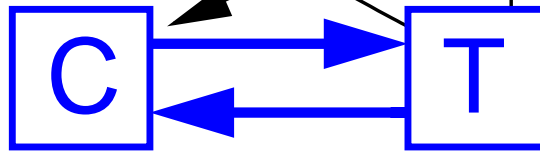
Purines



Transitions

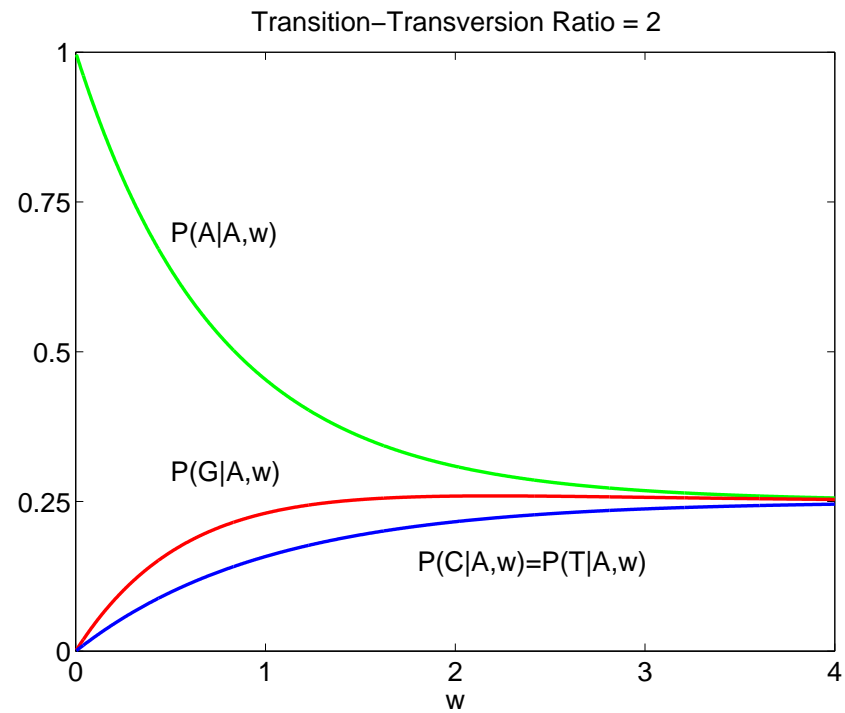
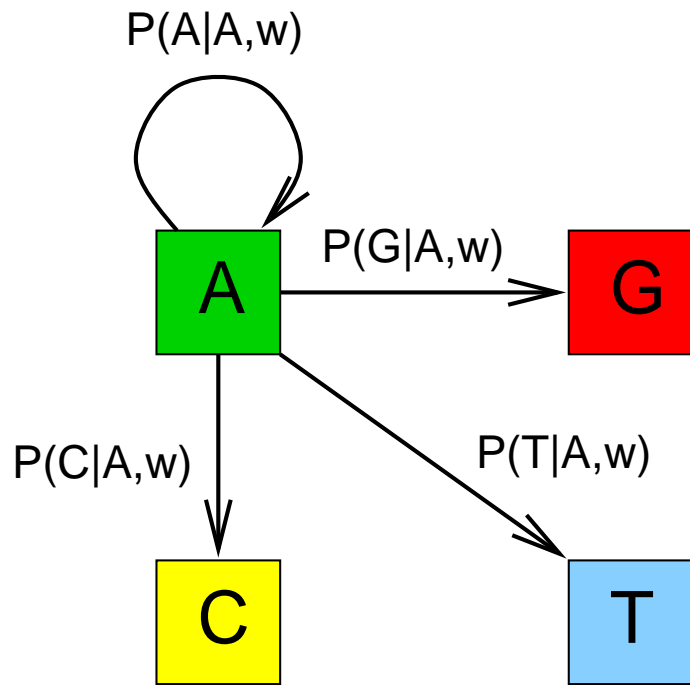
Transversions

Pyrimidines



Transitions

Mutation probabilities



Branch length = mutation rate \times time

Parameters

- Topology state sequences:

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

- Rate state sequences:

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

- Nucleotide substitution model state sequences:

$$\mathbf{Q} = (Q_1, \dots, Q_N)$$

- Transition probability parameters:

$$\nu_S, \nu_R, \nu_Q$$

Sampling from the posterior distribution

- Sampling from

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

- Gibbs sampling

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

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

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

- ...

Sampling from the posterior distribution

- Sampling from

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

- Gibbs sampling

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

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

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

- ...

- $\mathbf{S}, \mathbf{R}, \mathbf{Q}$: Stochastic forward–backward algorithm

- ν_S, ν_R, ν_Q : Sample from Beta distribution

Work in progress

- Decoupling rate variation from nucleotide substitution processes
- **Adaptable rate factors \mathbf{r}**
- Adaptable dimension of \mathbf{R} with RJMCMC

Adaptation of the rate hyperparameters

$$\mathbf{r} \sim \mathbf{P}(\mathbf{r}|\mathbf{R}, \mathbf{S}, \nu_{\mathbf{S}}, \nu_{\mathbf{R}}, \mathcal{D})$$

Bayes' rule:

$$P(\mathbf{r}|\mathbf{R}, \mathbf{S}, \nu_{\mathbf{S}}, \nu_{\mathbf{R}}, \mathcal{D}) \propto \mathbf{P}(\mathcal{D}|\mathbf{r}, \mathbf{R}, \mathbf{S}, \nu_{\mathbf{S}}, \nu_{\mathbf{R}})\mathbf{P}(\mathbf{r}|\mathbf{R}, \mathbf{S}, \nu_{\mathbf{S}}, \nu_{\mathbf{R}})$$

How to choose the prior: $P(\mathbf{r}|\mathbf{R}, \mathbf{S}, \nu_{\mathbf{S}}, \nu_{\mathbf{R}}) = \mathbf{P}(\mathbf{r})$?

Adaptation of the rate hyperparameters

$$\mathbf{r} \sim \mathbf{P}(\mathbf{r}|\mathbf{R}, \mathbf{S}, \nu_{\mathbf{S}}, \nu_{\mathbf{R}}, \mathcal{D})$$

Bayes' rule:

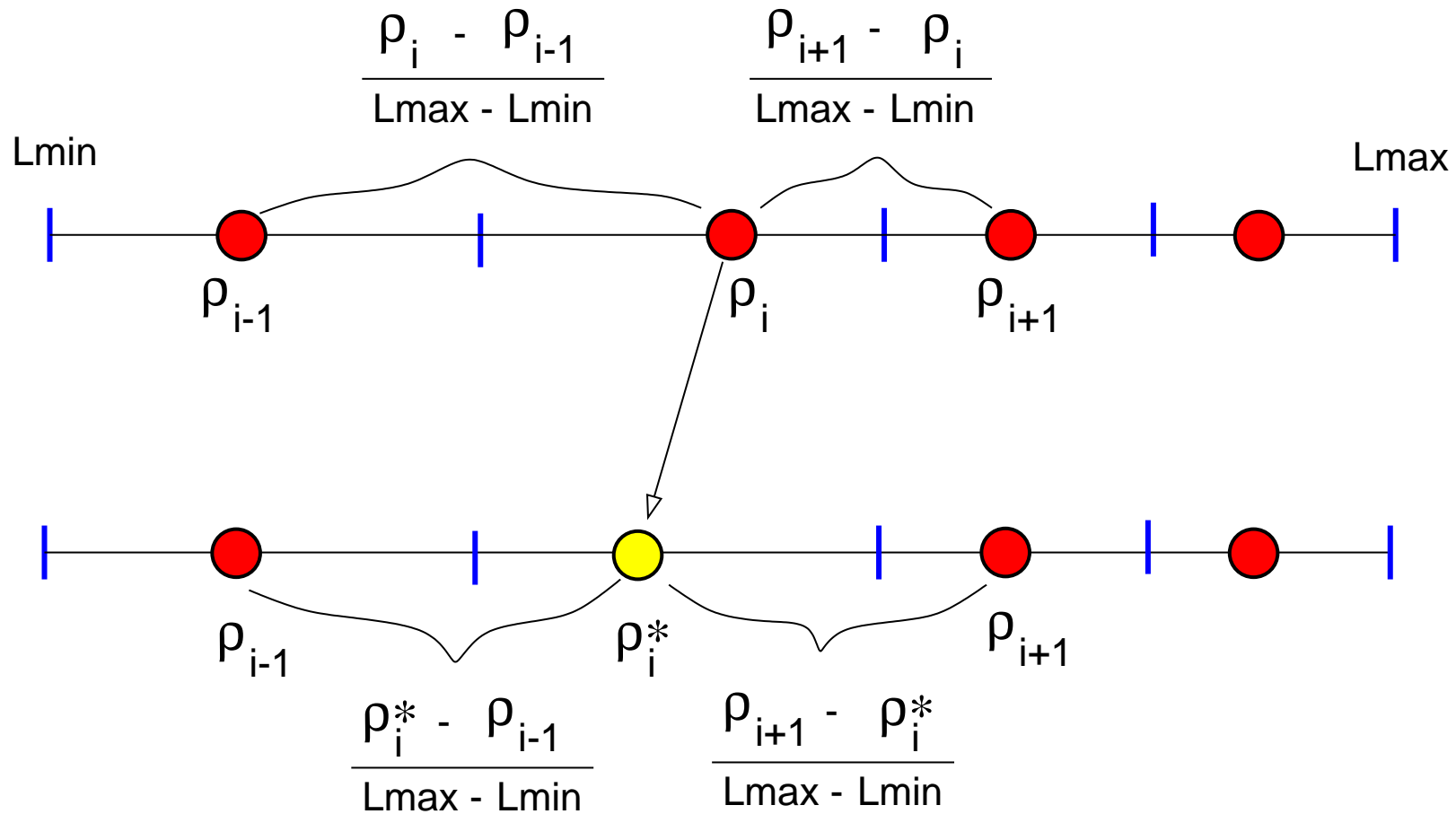
$$P(\mathbf{r}|\mathbf{R}, \mathbf{S}, \nu_{\mathbf{S}}, \nu_{\mathbf{R}}, \mathcal{D}) \propto \mathbf{P}(\mathcal{D}|\mathbf{r}, \mathbf{R}, \mathbf{S}, \nu_{\mathbf{S}}, \nu_{\mathbf{R}})\mathbf{P}(\mathbf{r}|\mathbf{R}, \mathbf{S}, \nu_{\mathbf{S}}, \nu_{\mathbf{R}})$$

How to choose the prior: $P(\mathbf{r}|\mathbf{R}, \mathbf{S}, \nu_{\mathbf{S}}, \nu_{\mathbf{R}}) = \mathbf{P}(\mathbf{r})$?

- Log scale
- Uniform distribution
- Equi-distant spacing

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

Rate adaptation: even-numbered order statistics



Adaptation of the rate hyperparameters

Prior ratio for a relocation move:

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

Adaptation of the rate hyperparameters

Prior ratio for a relocation move:

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

Proposal move :

Propose ρ'_i from a uniform distribution over the interval $[\rho_{i-1}, \rho_{i+1}]$:

$$Q(\rho'_i) = Q(\rho_i) = \frac{1}{\rho_{i+1} - \rho_{i-1}}$$

Metropolis-Hastings within Gibbs

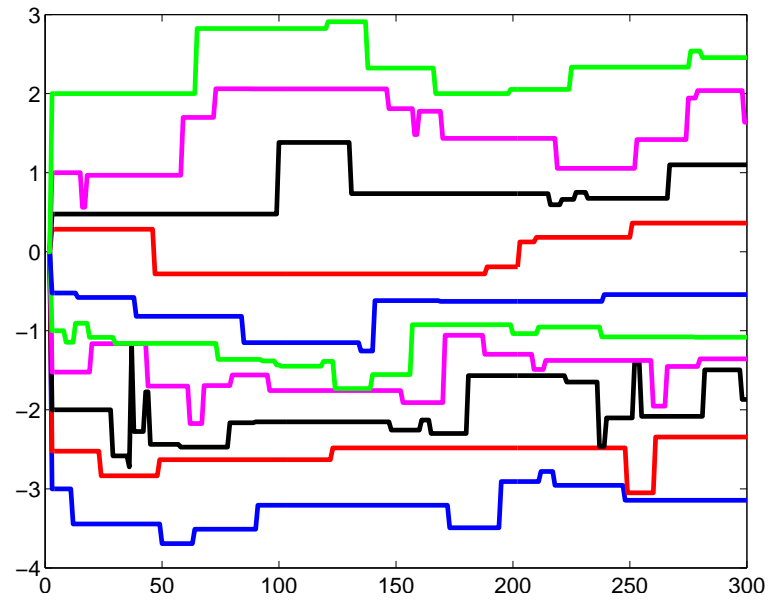
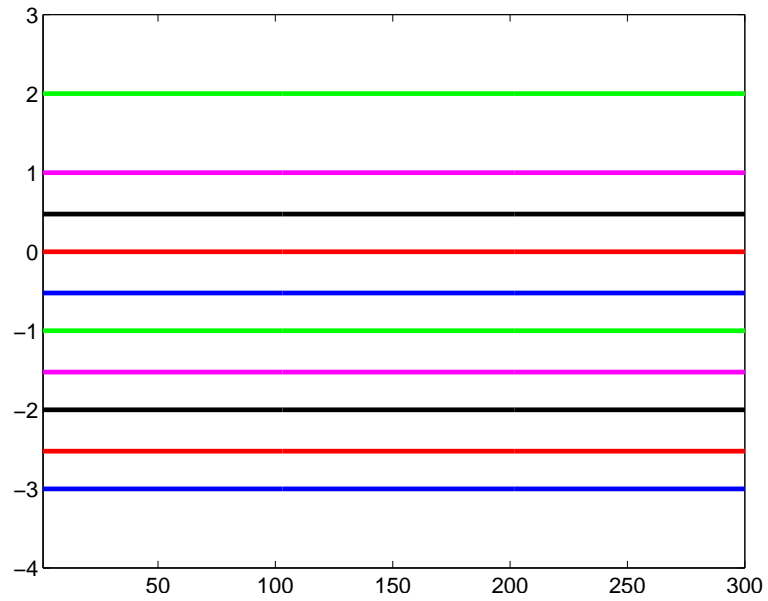
Acceptance probability for ρ'_i : $\min\{1, A\}$

$A =$ likelihood ratio \times prior ratio \times inverse proposal probability ratio

$$= \frac{P(\mathcal{D}|\boldsymbol{\rho}', \mathbf{R}, \mathbf{S}, \nu_S, \nu_R)}{P(\mathcal{D}|\boldsymbol{\rho}, \mathbf{R}, \mathbf{S}, \nu_S, \nu_R)} \times \frac{P(\boldsymbol{\rho}')}{P(\boldsymbol{\rho})} \times \frac{Q(\boldsymbol{\rho})}{Q(\boldsymbol{\rho}')}$$

$$= \frac{P(\mathcal{D}|\boldsymbol{\rho}', \mathbf{R}, \mathbf{S}, \nu_S, \nu_R)}{P(\mathcal{D}|\boldsymbol{\rho}, \mathbf{R}, \mathbf{S}, \nu_S, \nu_R)} \times \frac{(\rho_{i+1} - \rho_i)(\rho_i - \rho_{i-1})}{(\rho_{i+1} - \rho'_i)(\rho'_i - \rho_{i-1})}$$

Rate adaptation for Neisseria



Work in progress

- Decoupling rate variation from nucleotide substitution processes
- Adaptable rate factors \mathbf{r}
- Adaptable dimension of \mathbf{R} with RJMCMC

RJMCMC

Green (1995)
Biometrika 82, 711-732

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

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

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

RJMCMC

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

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

RJMCMC

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

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

Acceptance probability=

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

Prior probabilities

Prior probabilities

Number of rate states k : Truncated Poisson distribution

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

Prior probabilities

Number of rate states k : Truncated Poisson distribution

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

Rate factors

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

Prior probabilities

Number of rate states k : Truncated Poisson distribution

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

Rate factors

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

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

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

Proposal probabilities

Proposal probabilities

Birth move:

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

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

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

Death move:

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

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

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

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

Jacobian

Bijection: $(\boldsymbol{\rho}, u) \longleftrightarrow \tilde{\boldsymbol{\rho}} : \tilde{\boldsymbol{\rho}} = \mathbf{f}(\boldsymbol{\rho}, u)$

$\dim(\boldsymbol{\rho}) = n$; $\dim(u) = 1$; $\dim(\tilde{\boldsymbol{\rho}}) = n + 1$

$$\det(\mathbf{J}) = \begin{vmatrix} \frac{\partial \tilde{\rho}_1}{\partial \rho_1} & \frac{\partial \tilde{\rho}_1}{\partial \rho_2} & \cdots & \frac{\partial \tilde{\rho}_1}{\partial \rho_n} & \frac{\partial \tilde{\rho}_1}{\partial u} \\ \frac{\partial \tilde{\rho}_2}{\partial \rho_1} & \frac{\partial \tilde{\rho}_2}{\partial \rho_2} & \cdots & \frac{\partial \tilde{\rho}_2}{\partial \rho_n} & \frac{\partial \tilde{\rho}_2}{\partial u} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \frac{\partial \tilde{\rho}_n}{\partial \rho_1} & \frac{\partial \tilde{\rho}_n}{\partial \rho_2} & \cdots & \frac{\partial \tilde{\rho}_n}{\partial \rho_n} & \frac{\partial \tilde{\rho}_n}{\partial u} \\ \frac{\partial \tilde{\rho}_{n+1}}{\partial \rho_1} & \frac{\partial \tilde{\rho}_{n+1}}{\partial \rho_2} & \cdots & \frac{\partial \tilde{\rho}_{n+1}}{\partial \rho_n} & \frac{\partial \tilde{\rho}_{n+1}}{\partial u} \end{vmatrix}$$

Jacobian

Bijection: $(\boldsymbol{\rho}, u) \longleftrightarrow \tilde{\boldsymbol{\rho}} : \tilde{\boldsymbol{\rho}} = (\boldsymbol{\rho}, u)$

$\dim(\boldsymbol{\rho}) = n$; $\dim(u) = 1$; $\dim(\tilde{\boldsymbol{\rho}}) = n + 1$

$$\det(\mathbf{J}) = \begin{vmatrix} \frac{\partial \rho_1}{\partial \rho_1} & \frac{\partial \rho_1}{\partial \rho_2} & \cdots & \frac{\partial \rho_1}{\partial \rho_n} & \frac{\partial \rho_1}{\partial u} \\ \frac{\partial \rho_2}{\partial \rho_1} & \frac{\partial \rho_2}{\partial \rho_2} & \cdots & \frac{\partial \rho_2}{\partial \rho_n} & \frac{\partial \rho_2}{\partial u} \\ \frac{\partial \rho_3}{\partial \rho_1} & \frac{\partial \rho_3}{\partial \rho_2} & \cdots & \frac{\partial \rho_3}{\partial \rho_n} & \frac{\partial \rho_3}{\partial u} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \frac{\partial \rho_n}{\partial \rho_1} & \frac{\partial \rho_n}{\partial \rho_2} & \cdots & \frac{\partial \rho_n}{\partial \rho_n} & \frac{\partial \rho_n}{\partial u} \\ \frac{\partial u}{\partial \rho_1} & \frac{\partial u}{\partial \rho_2} & \cdots & \frac{\partial u}{\partial \rho_n} & \frac{\partial u}{\partial u} \end{vmatrix}$$

Jacobian

Bijection: $(\boldsymbol{\rho}, u) \longleftrightarrow \tilde{\boldsymbol{\rho}} : \tilde{\boldsymbol{\rho}} = (\boldsymbol{\rho}, u)$

$\dim(\boldsymbol{\rho}) = n$; $\dim(u) = 1$; $\dim(\tilde{\boldsymbol{\rho}}) = n + 1$

$$\det(\mathbf{J}) = \begin{vmatrix} 1 & 0 & \dots & 0 & 0 \\ 0 & 1 & \dots & 0 & 0 \\ \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & 1 & 0 \\ 0 & 0 & \dots & 0 & 1 \end{vmatrix} = 1$$

Acceptance probability for a birth move

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

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

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

$$\text{Jacobian} = 1$$

Acceptance probability of a birth move

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

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

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

$$\text{Jacobian} = 1$$

Acceptance probability of a birth move:

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

Acceptance probability of a birth move

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

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

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

$$\text{Jacobian} = 1$$

Acceptance probability of a birth move:

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

Acceptance probability of a birth move

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

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

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

Jacobian = 1

Acceptance probability of a birth move:

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

Acceptance probability of a birth move

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

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

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

$$\text{Jacobian} = 1$$

Acceptance probability of a birth move:

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

Acceptance probability of a birth move

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

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

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

$$\text{Jacobian} = 1$$

Acceptance probability of a birth move:

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

What happened to the ratio of factorials?

In [Boys, Henderson \(2001\)](#), *Comp. Sci. and Statist.*, 33, 35-49, there is an extra term:

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

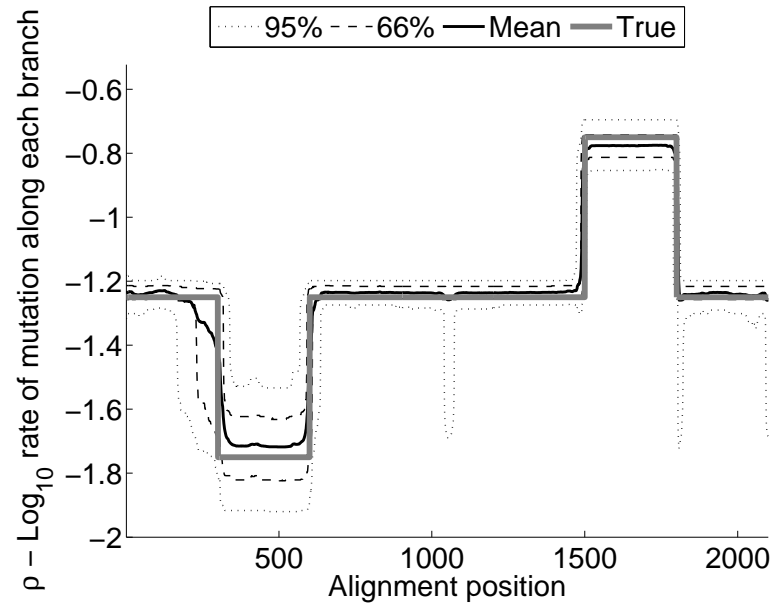
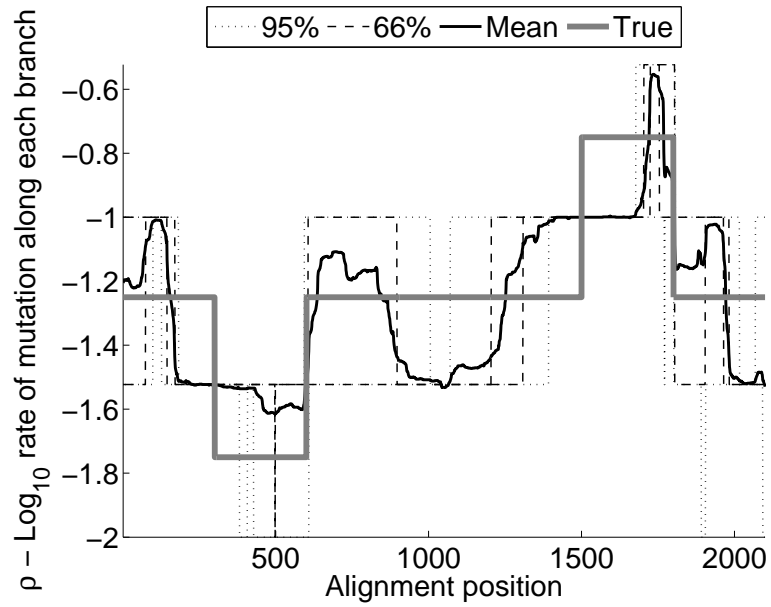
Note that this term is already included in

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

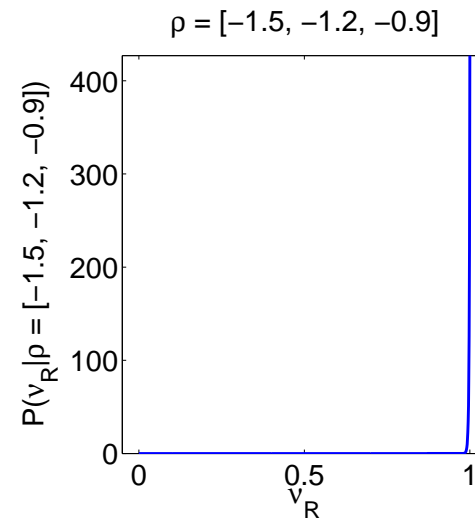
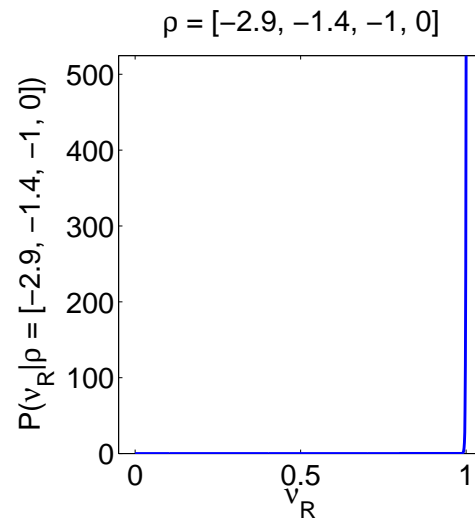
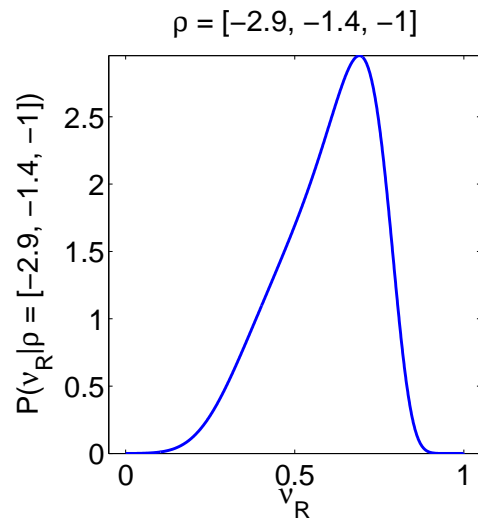
from [Green \(1995\)](#), *Biometrika* 82, 711-732

Results

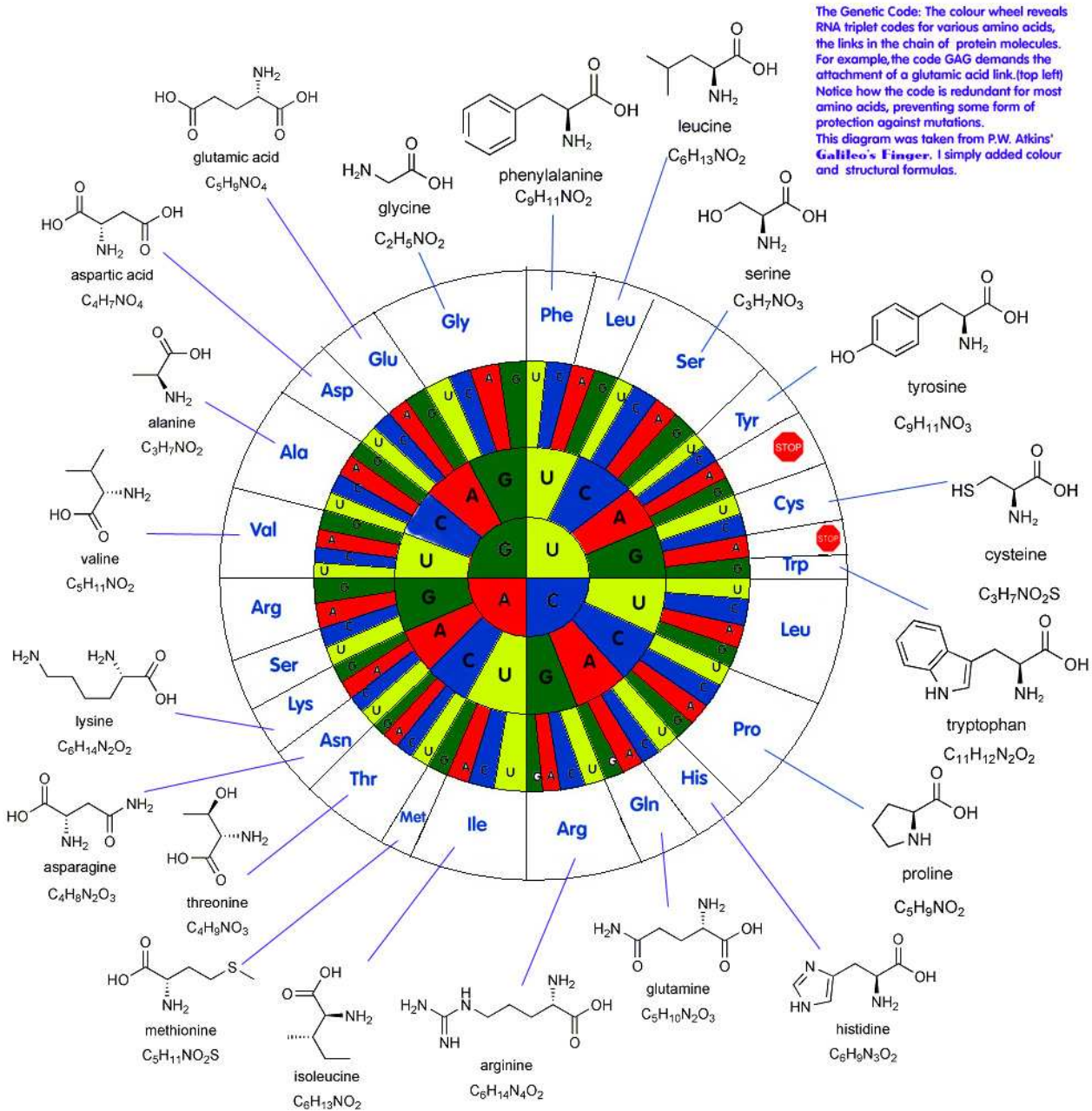
Synthetic data: MCMC (left) versus RJMCMC (right)



Neisseria: posterior distribution of ν_R : bi-modal

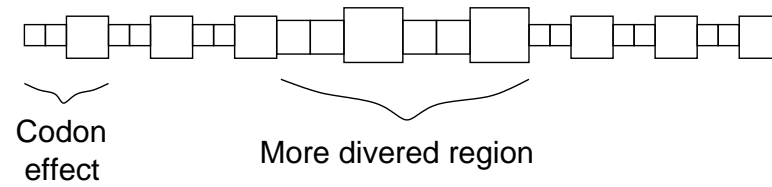
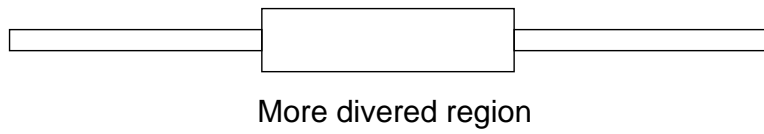


- Why is the distribution bi-modal?
- Why does bi-modality occur with RJMCMC, but not necessarily with fixed-dimension MCMC?

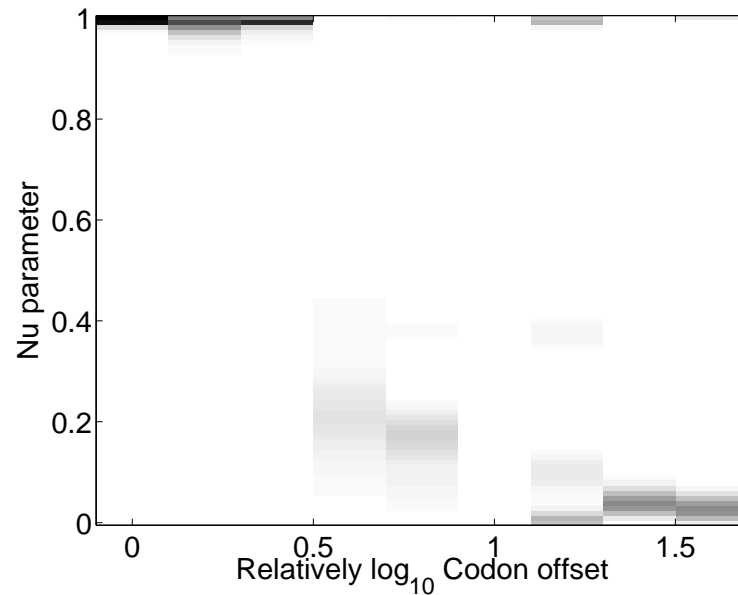
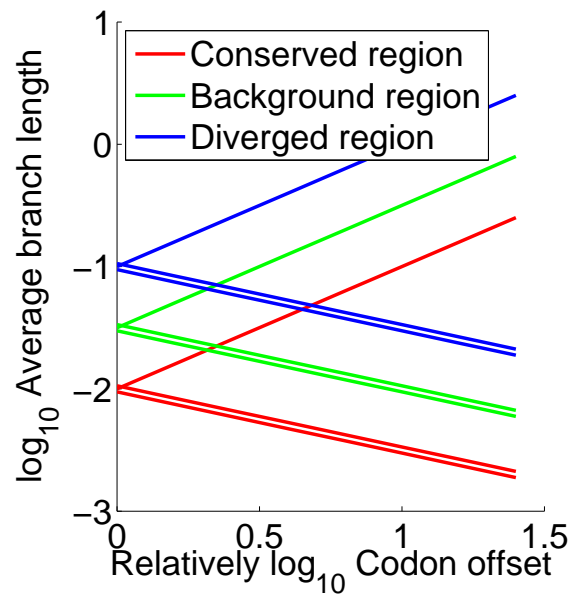
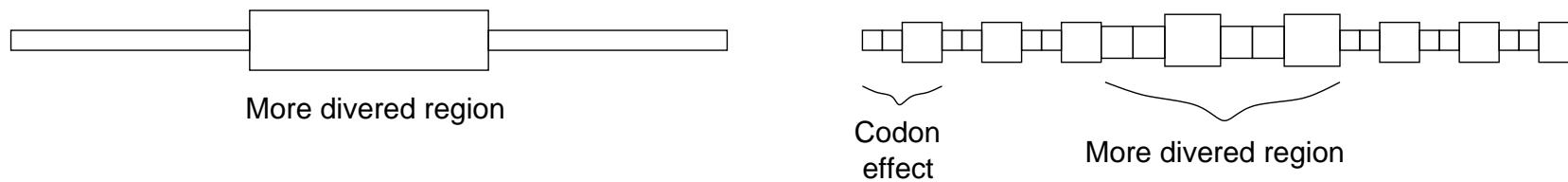


The Genetic Code: The colour wheel reveals RNA triplet codes for various amino acids, the links in the chain of protein molecules. For example, the code GAG demands the attachment of a glutamic acid link. (top left) Notice how the code is redundant for most amino acids, preventing some form of protection against mutations. This diagram was taken from P.W. Atkins' *Galileo's Finger*. I simply added colour and structural formulas.

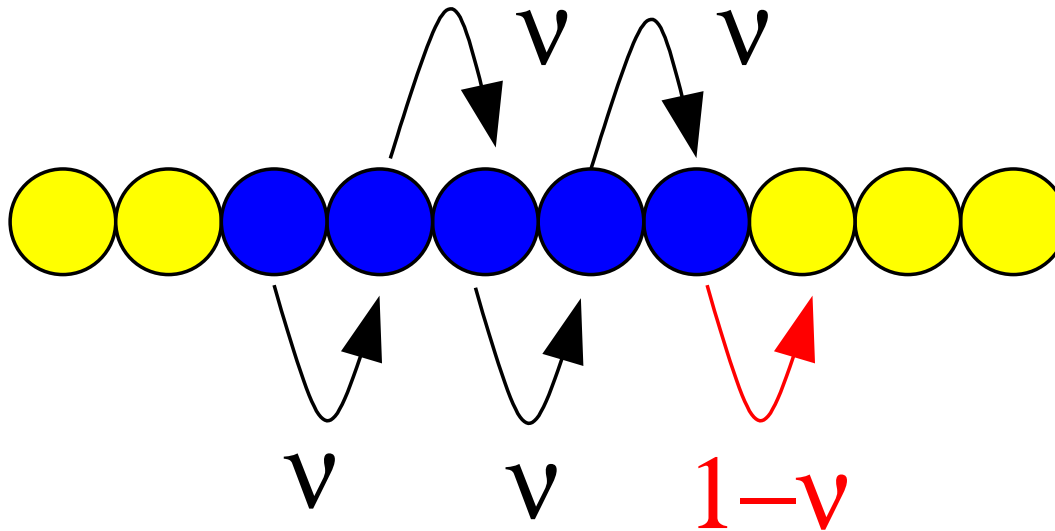
Why can the distribution of ν_R become bi-modal ?



Why can the distribution of ν_R become bi-modal ?



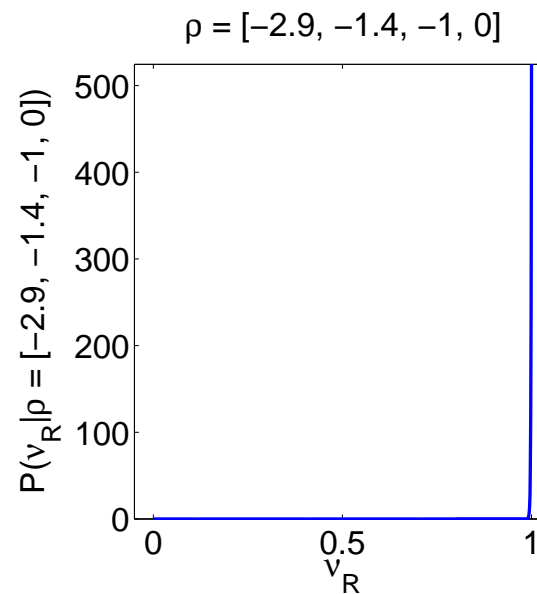
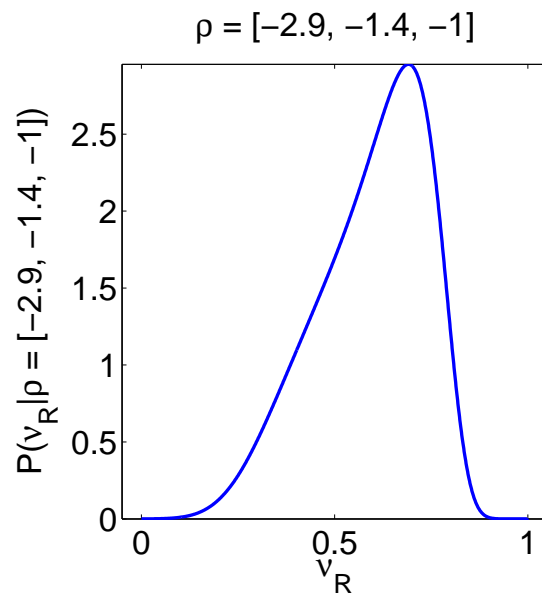
Average segment length



Segment length n . Probability: $P(n) = \nu^{n-1}(1 - \nu)$

$$\begin{aligned}\langle n \rangle &= \sum nP(n) = (1 - \nu) \sum n\nu^{n-1} = (1 - \nu) \frac{d}{d\nu} \sum \nu^n \\ &= (1 - \nu) \frac{d}{d\nu} \frac{1}{1-\nu} = \frac{1}{1-\nu}\end{aligned}$$

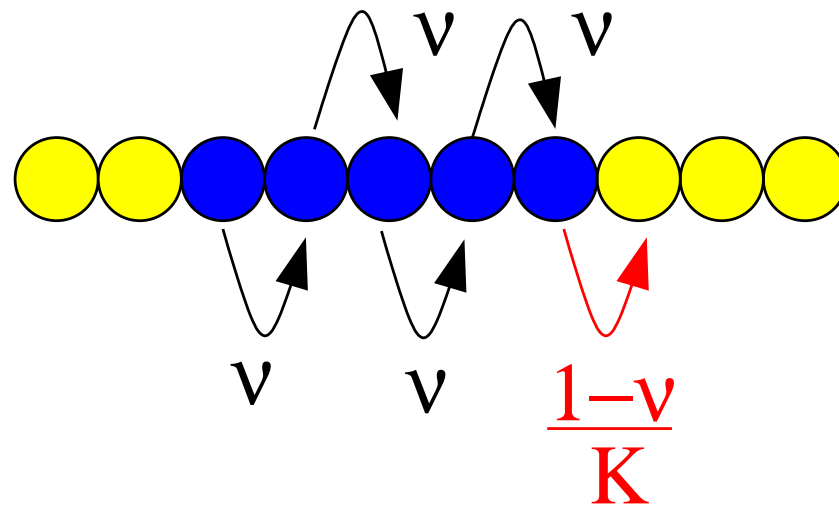
Neisseria: posterior distribution of ν_R : bi-modal



- Why is the distribution bi-modal?
- Why does bi-modality occur with RJMCMC, but not necessarily with fixed-dimension MCMC?

Why is the bimodality repressed with fixed-dimension MCMC when including extreme rate states ?

$$P(\nu|\mathcal{D}) \propto P(\mathcal{D}|\nu) = \sum_{\mathbf{R}} P(\mathcal{D}, \mathbf{R}|\nu) = \sum_{\mathbf{R}} P(\mathbf{y}_t|R_t)P(R_t|R_{t-1}, \nu)$$



Extreme state R^* : $P(\mathbf{y}_t|R^*) = 0$
 $P(R_t \neq R_{t-1}|R_{t-1}, \nu) = \frac{1-\nu}{K}$
 $P(\mathcal{D}|\nu)$ suppressed for low values of ν .

Alternative explanation

$$P(\nu|\mathcal{D}) \propto P(\mathcal{D}|\nu) = \sum_{\mathbf{R}} P(\mathcal{D}, \mathbf{R}|\nu) = \sum_{\mathbf{R}} P(\mathcal{D}|\mathbf{R})P(\mathbf{R}|\nu)$$

Simple Monte Carlo: Propose state sequences $\{\mathbf{R}_i\}$ from the prior $P(\mathbf{R}|\nu)$.

$$P(\mathcal{D}|\nu) = \frac{1}{N} \sum_{i=1}^N P(\mathcal{D}|\mathbf{R}_i)$$

When \mathbf{R}_i contains a **transition** into an **extreme state**, then $P(\mathcal{D}|\mathbf{R}_i) \approx 0$.

This is **more likely** to happen as ν gets **smaller**.

Consequently, $P(\mathcal{D}|\nu)$ gets small for small values of ν .

Alternative explanation

$$P(\nu|\mathcal{D}) \propto P(\mathcal{D}|\nu) = \sum_{\mathbf{R}} P(\mathcal{D}, \mathbf{R}|\nu) = \sum_{\mathbf{R}} P(\mathcal{D}|\mathbf{R})P(\mathbf{R}|\nu)$$

Simple Monte Carlo: Propose state sequences $\{\mathbf{R}_i\}$ from the prior $P(\mathbf{R}|\nu)$.

$$P(\mathcal{D}|\nu) = \frac{1}{N} \sum_{i=1}^N P(\mathcal{D}|\mathbf{R}_i)$$

When \mathbf{R}_i contains a **transition** into an **extreme state**, then $P(\mathcal{D}|\mathbf{R}_i) \approx 0$.

This is **more likely** to happen as ν gets smaller.

Consequently, $P(\mathcal{D}|\nu)$ gets small for small values of ν .

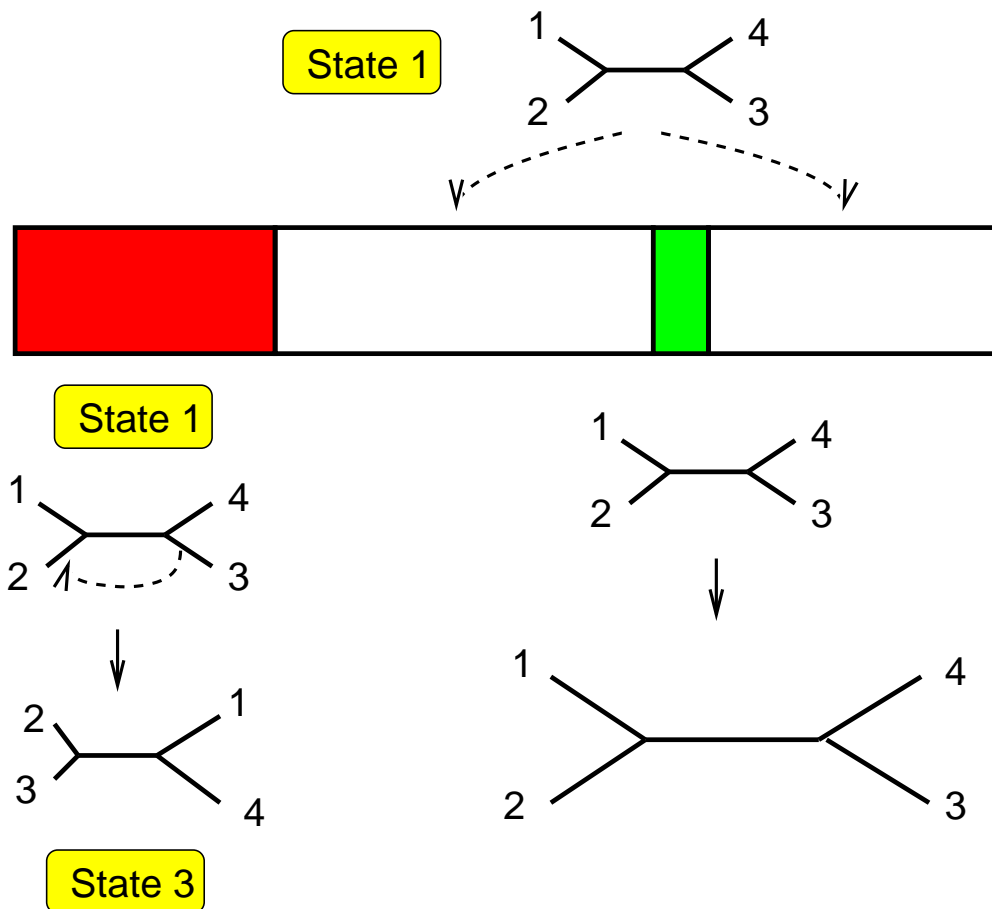
RJMCMC → death of extreme states → no pressure to keep ν high

We need a more informative prior on ν .

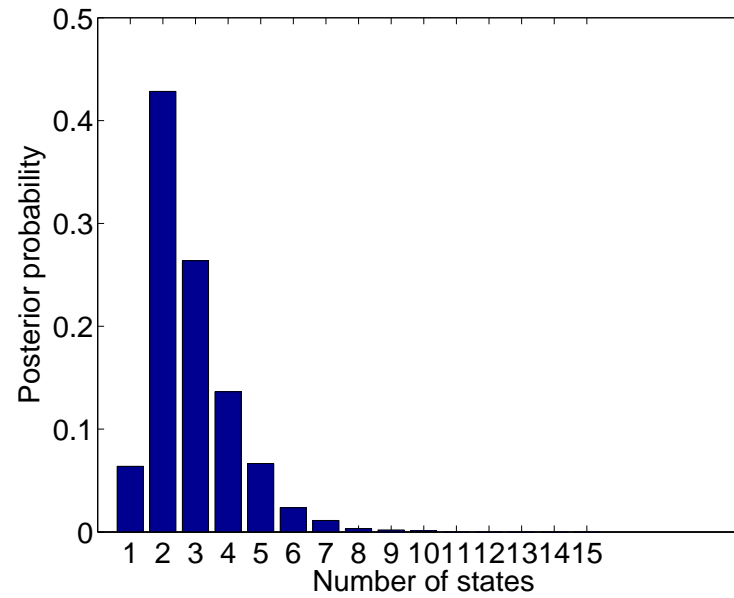
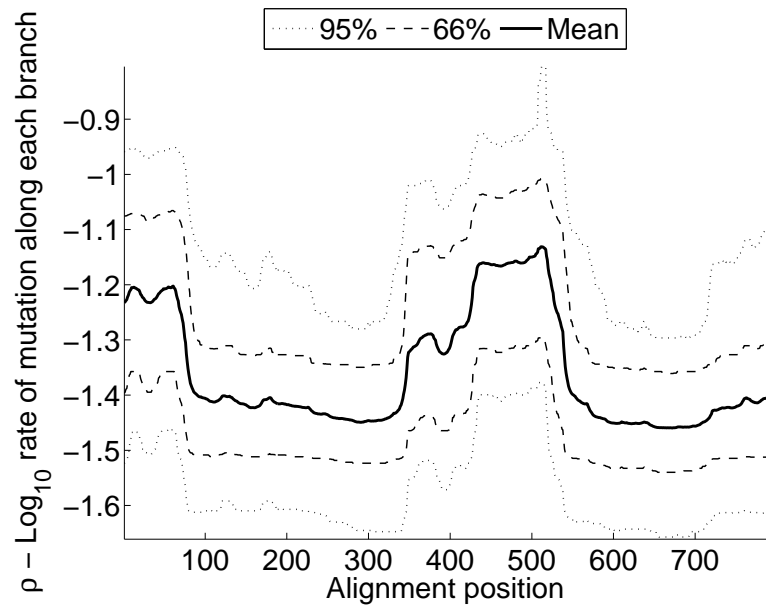
$$P(\nu) = 0 \text{ if } \nu < \nu_{max}$$

$$\nu_{max} = 0.99$$

Application to Neisseria

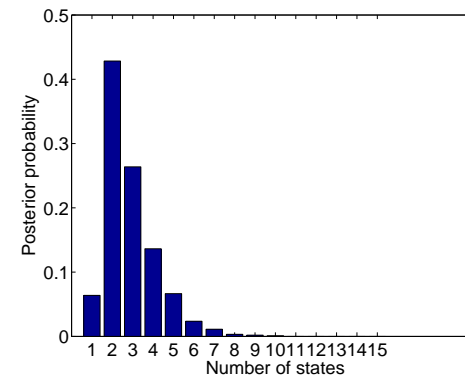
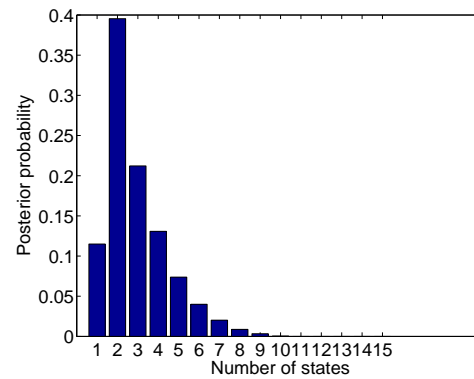
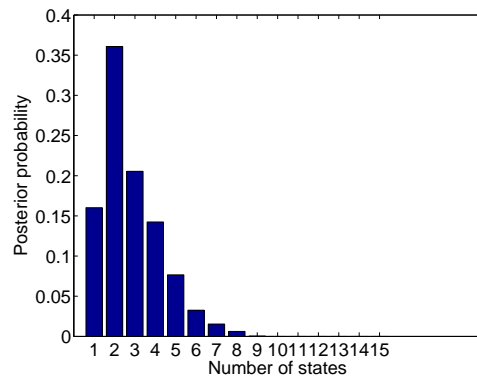
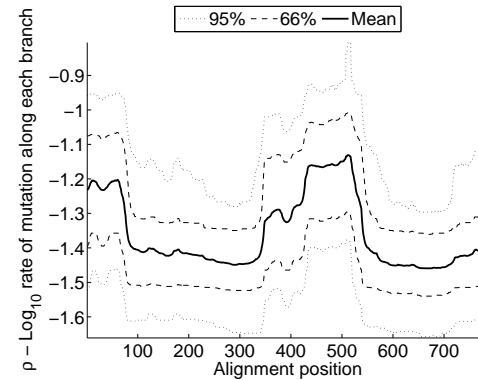
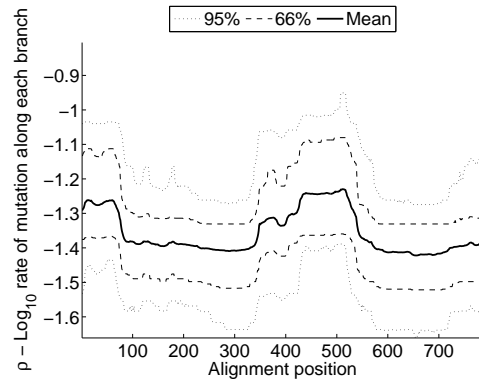
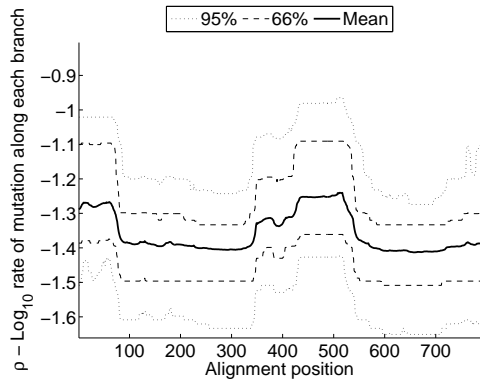


Application to Neisseria

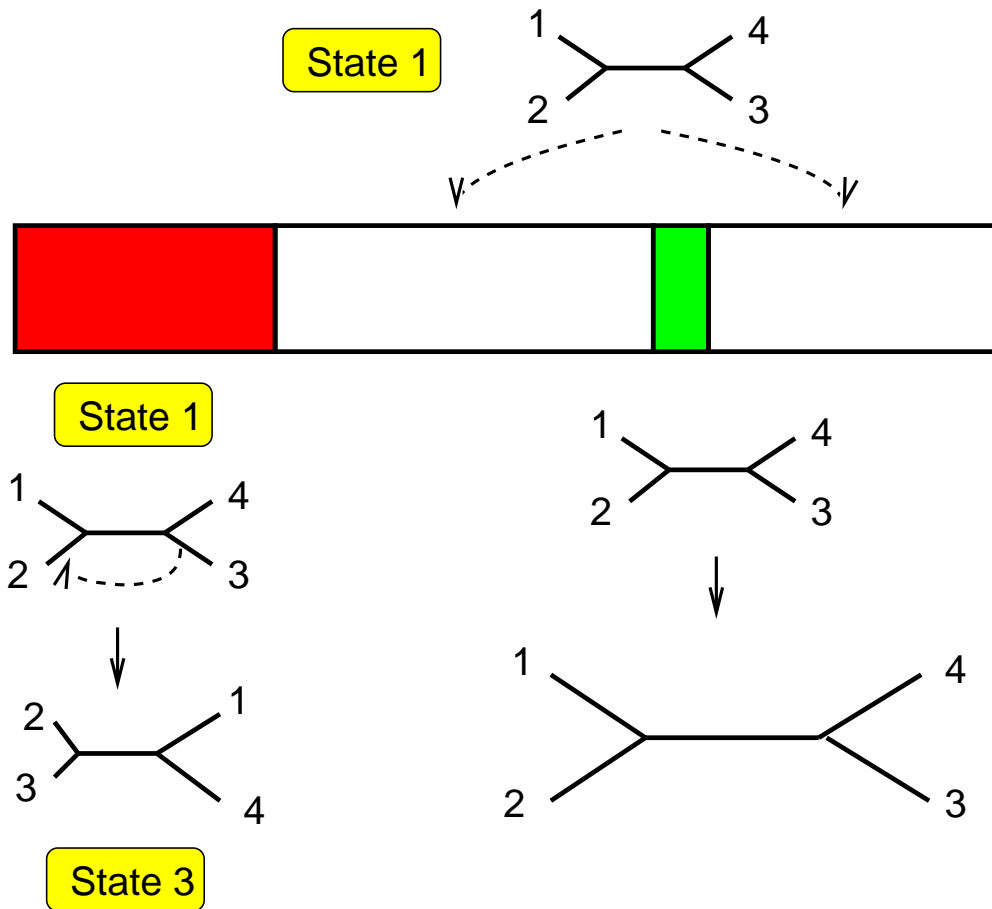


Neisseria: Dependence on the prior

Even-order statistics, uniform, Gaussian



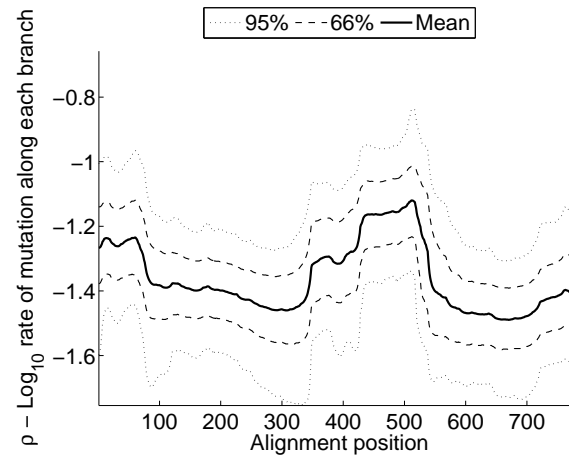
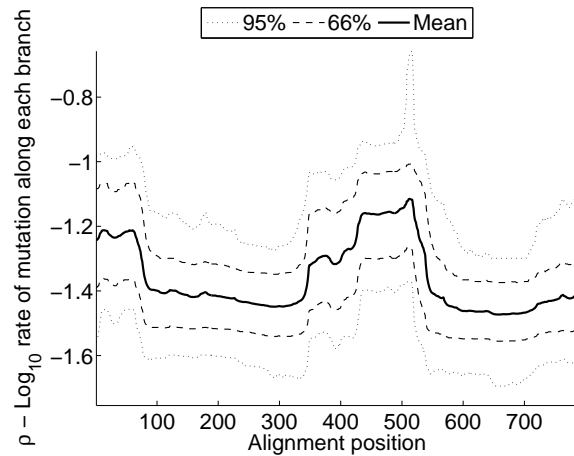
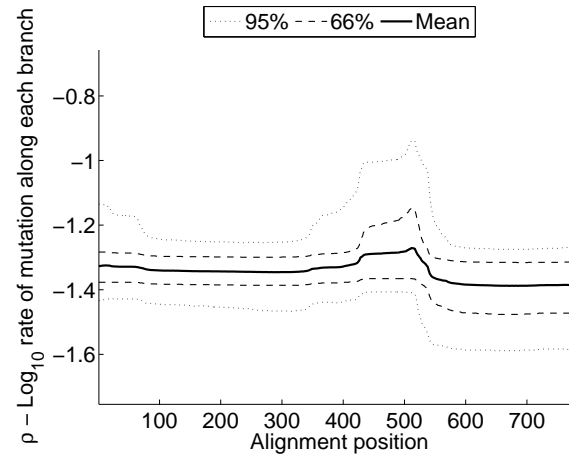
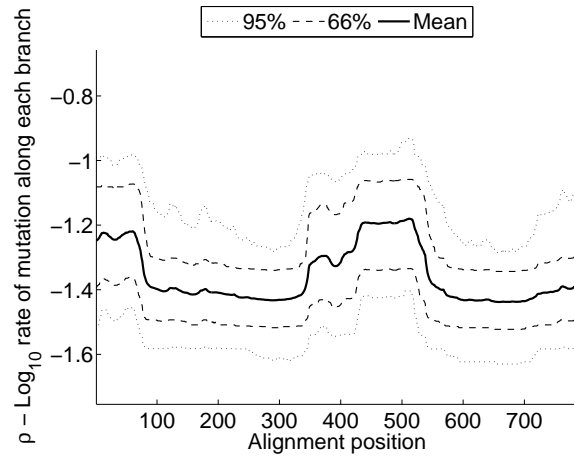
Application to Neisseria



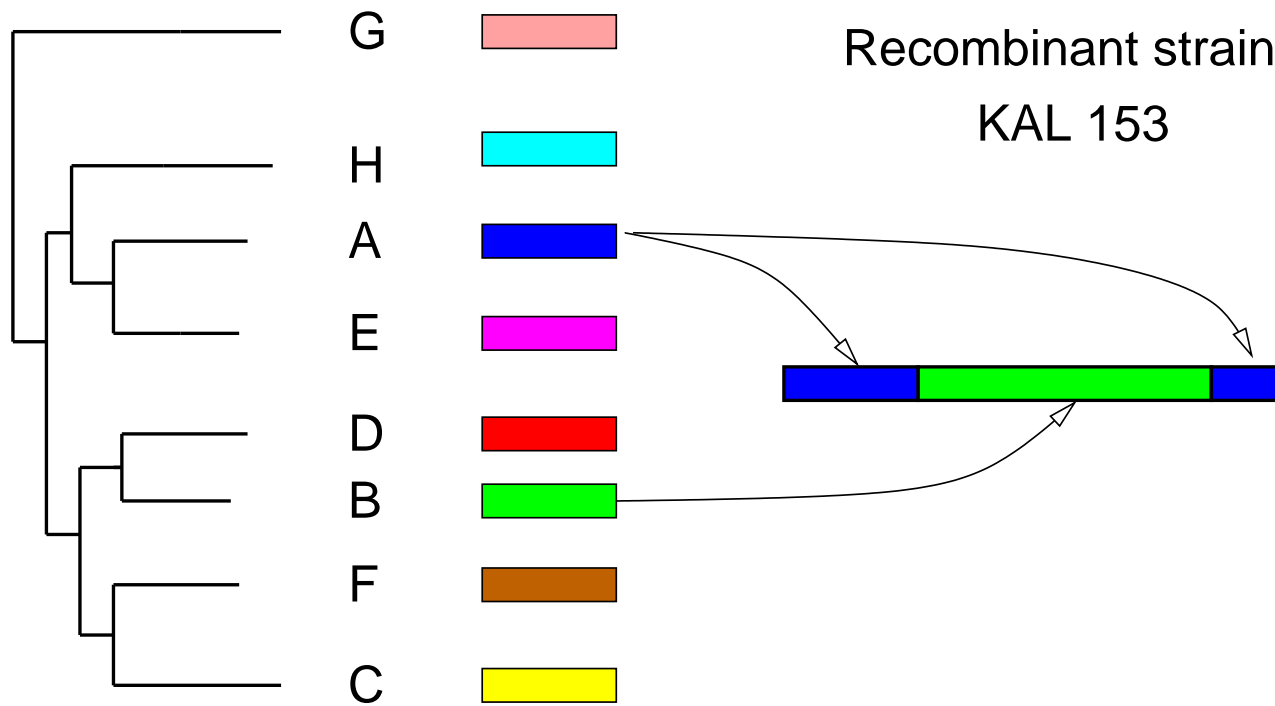
Comparison with the dual breakpoint model

Left: Phylo-FHMM Right: Dual breakpoint model

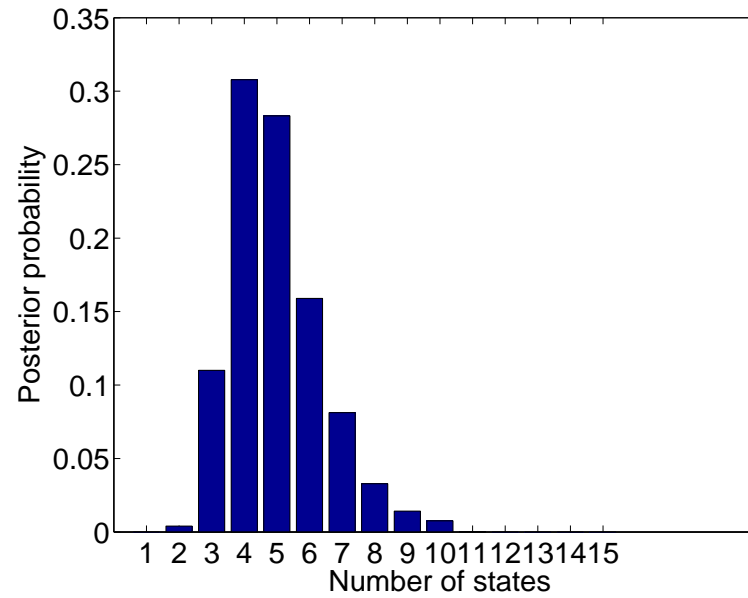
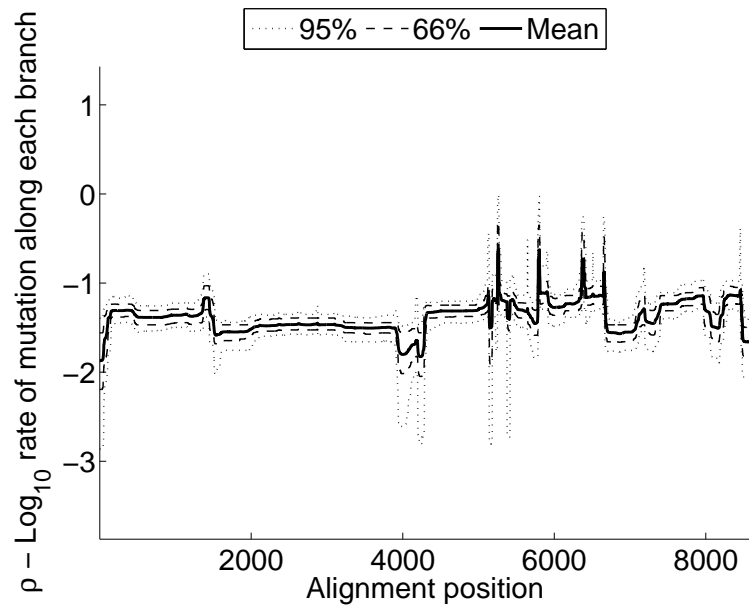
Top: Prior mean=1 Bottom: Prior mean= 5 states, 9 regions



Application to HIV-1

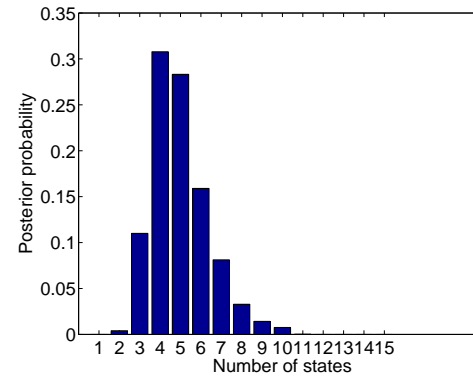
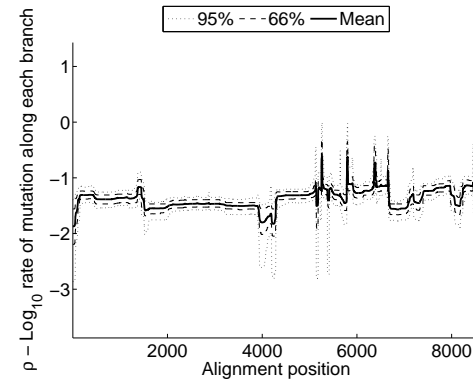
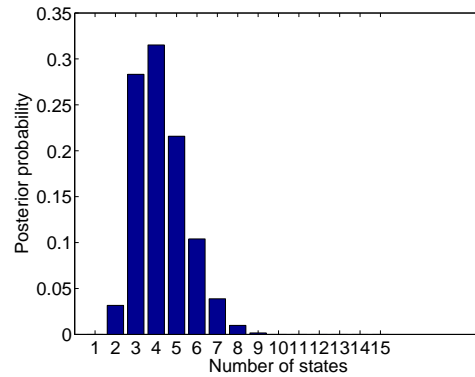
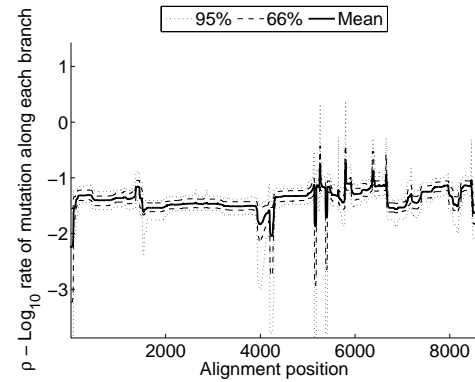
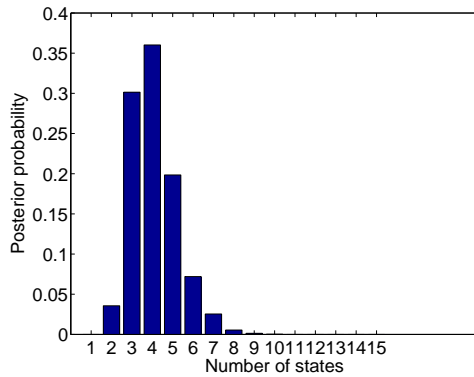
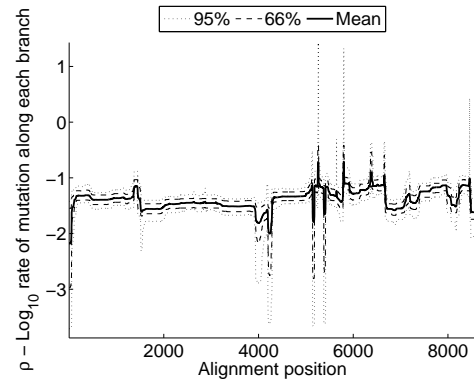


Application to HIV-1



Application to HIV-1: Dependence on the prior

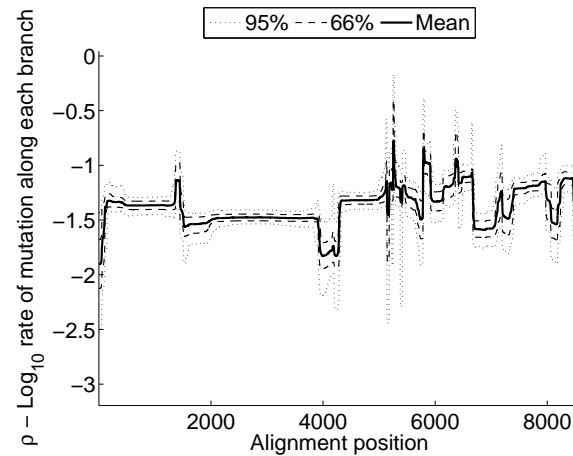
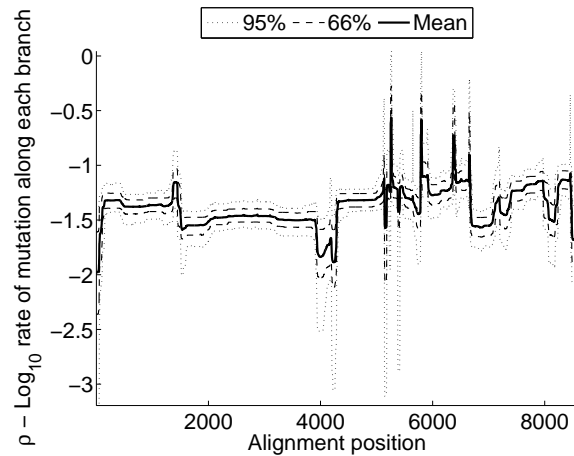
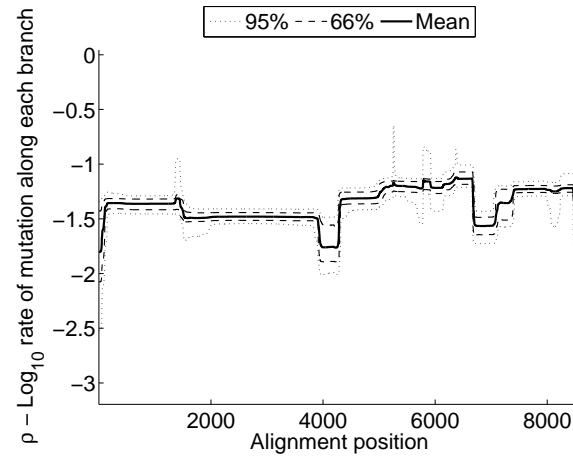
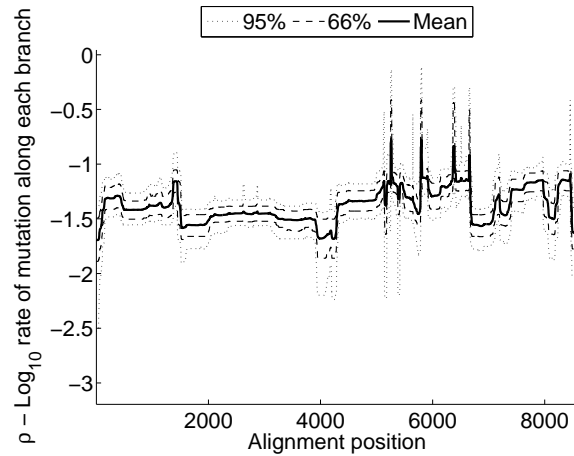
Even-order statistics, uniform, Gaussian



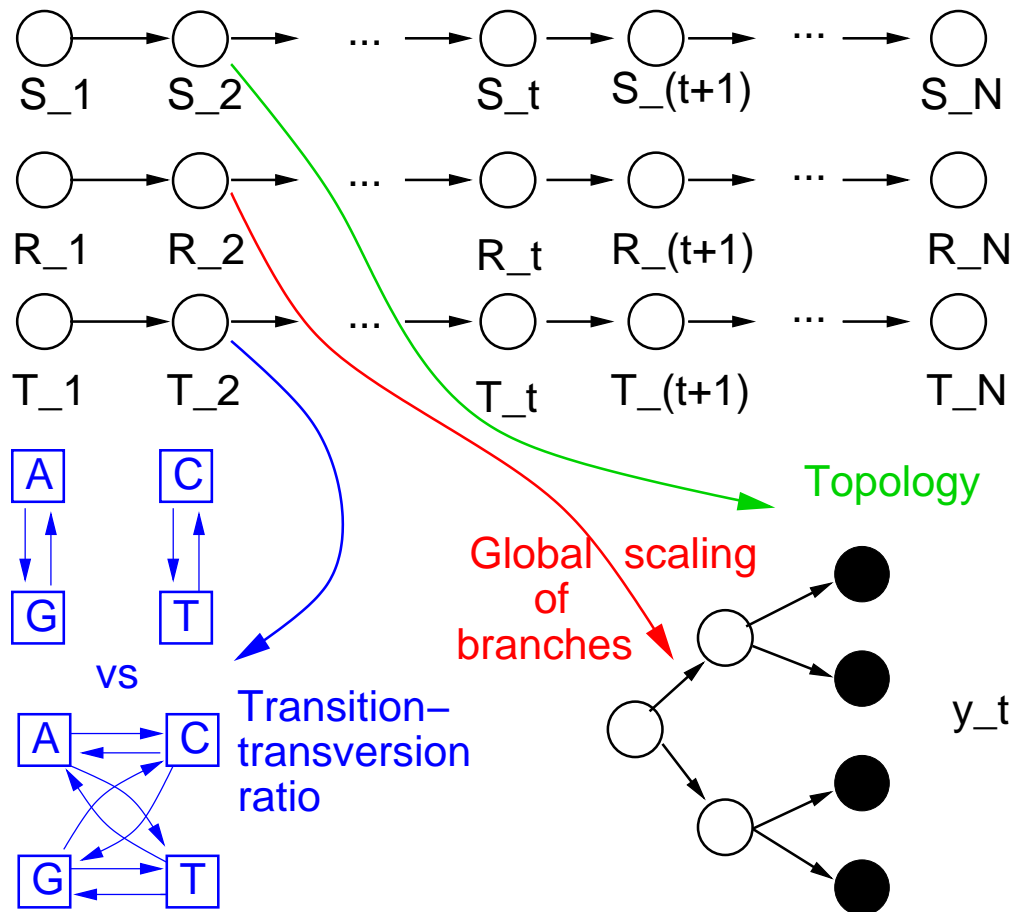
Comparison with the dual breakpoint model

Left: Phylo-FHMM Right: Dual breakpoint model

Top: Prior mean=1 Bottom: Prior mean= 5 states, 9 regions

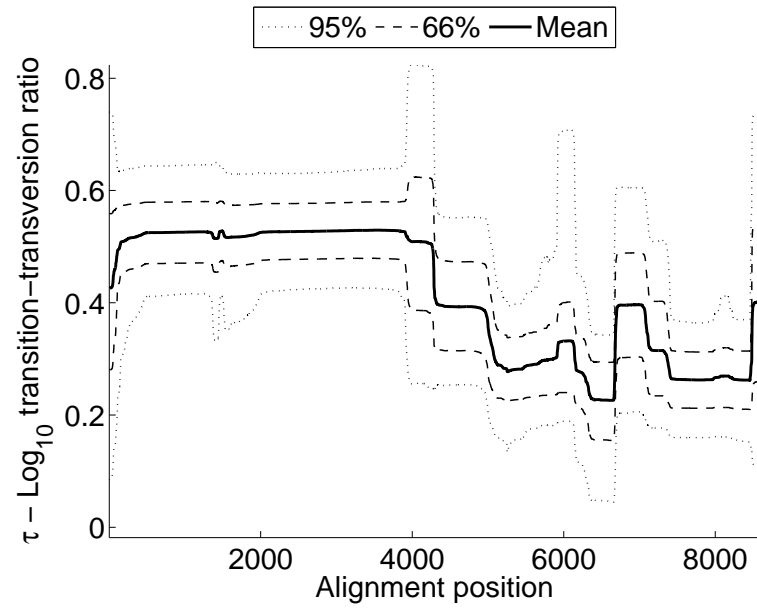
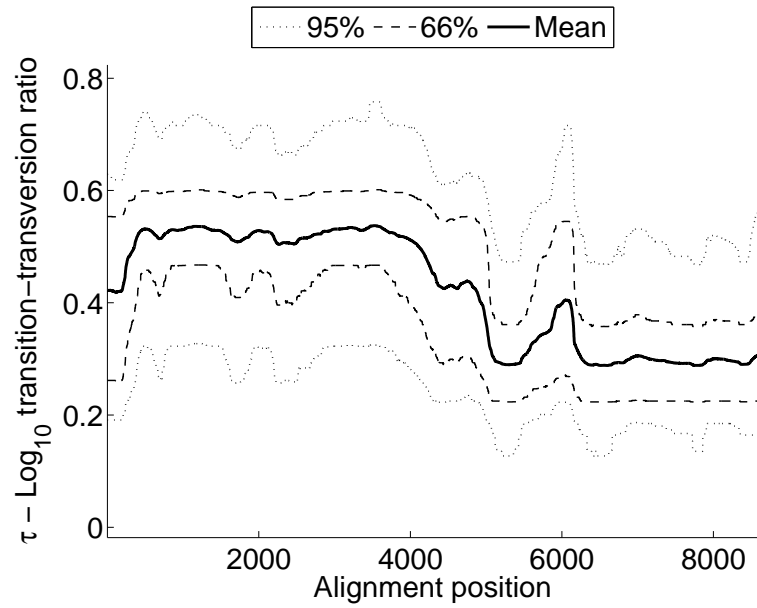


Transition-transversion ratio in HIV-1



Transition-transversion ratio in HIV-1

Left: Phylo-FHMM Right: Dual breakpoint model



Summary

- Detecting recombination: [Phylogenetic HMMs](#)
Dirk Husmeier, Frank Wright and Grainne McGuire
2000-2003
- Distinguishing between recombination and rate variation:
[Phylogenetic FHMMs](#)
Dirk Husmeier, 2005
- Learning the number of genomic regions
under selective pressure:
[Phylogenetic FHMMs trained with RJMCMC](#)
Wolfgang Lehrach and Dirk Husmeier, 2006

Future work

Predicting protein-protein interactions

Nimrod, Glaser, Steinberg, Ben-Tal and Pupko (2005)

In silico identification of functional
regions in proteins

Bioinformatics 21, Suppl. 1 (ISMB 05)

