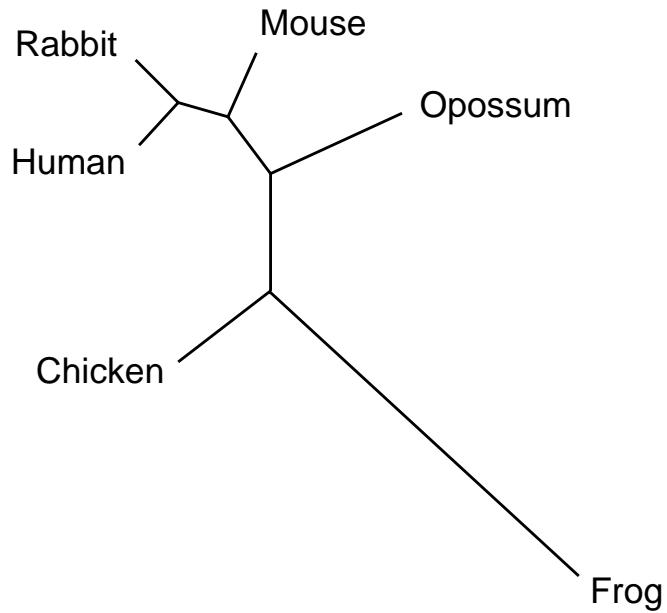

Probabilistic Divergence Measures for Detecting Interspecies Recombination

Dirk Husmeier & Frank Wright
Biomathematics and Statistics Scotland
at the Scottish Crop Research Institute
Invergowrie, Dundee DD2 5DA, UK
Email: dirk@bioass.ac.uk
<http://www.bioass.ac.uk/~dirk>

- Phylogenetics
- Sporadic recombination
- Statistical detection methods

Phylogenetics

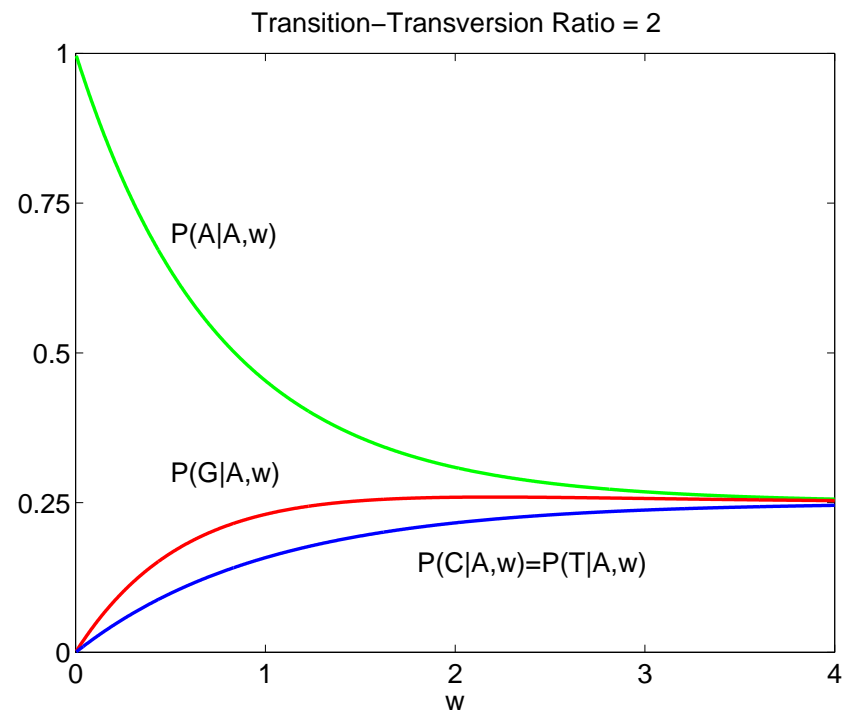
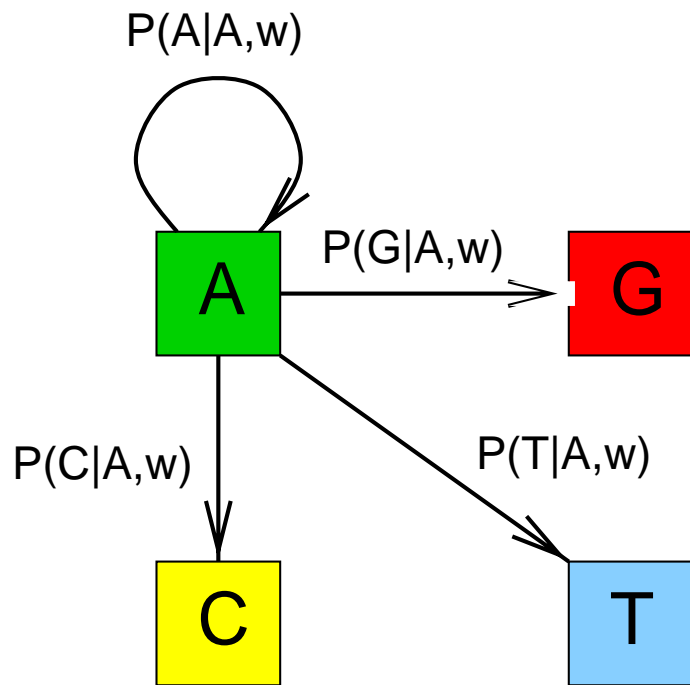
Frog	G	C	T	T	G	A	C	T	T	C	T	G	A	G	G	T	T
Chicken	G	C	G	T	A	A	C	T	T	C	A	C	A	T	G	A	T
Human	G	C	G	T	C	A	C	T	T	G	A	G	A	C	G	C	T
Rabbit	G	C	G	T	C	A	C	T	T	G	A	G	A	C	G	C	T
Mouse	G	C	G	T	C	A	C	T	T	G	A	C	A	G	G	C	T
Opossum	G	C	G	T	C	A	C	T	T	G	A	G	A	C	G	C	T



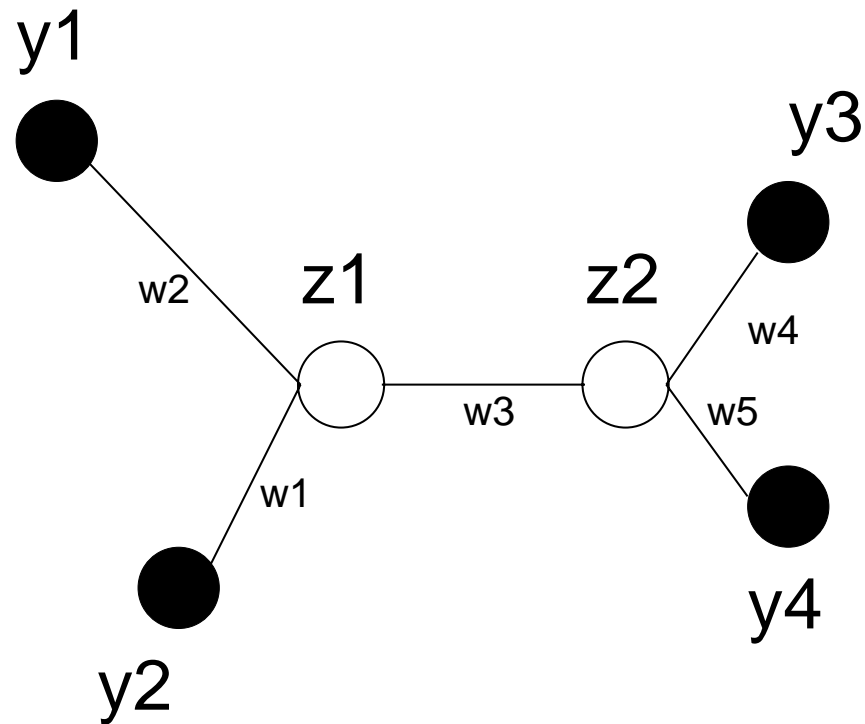
--> Topology

--> Branch lengths

A Probabilistic Model of Evolution 1



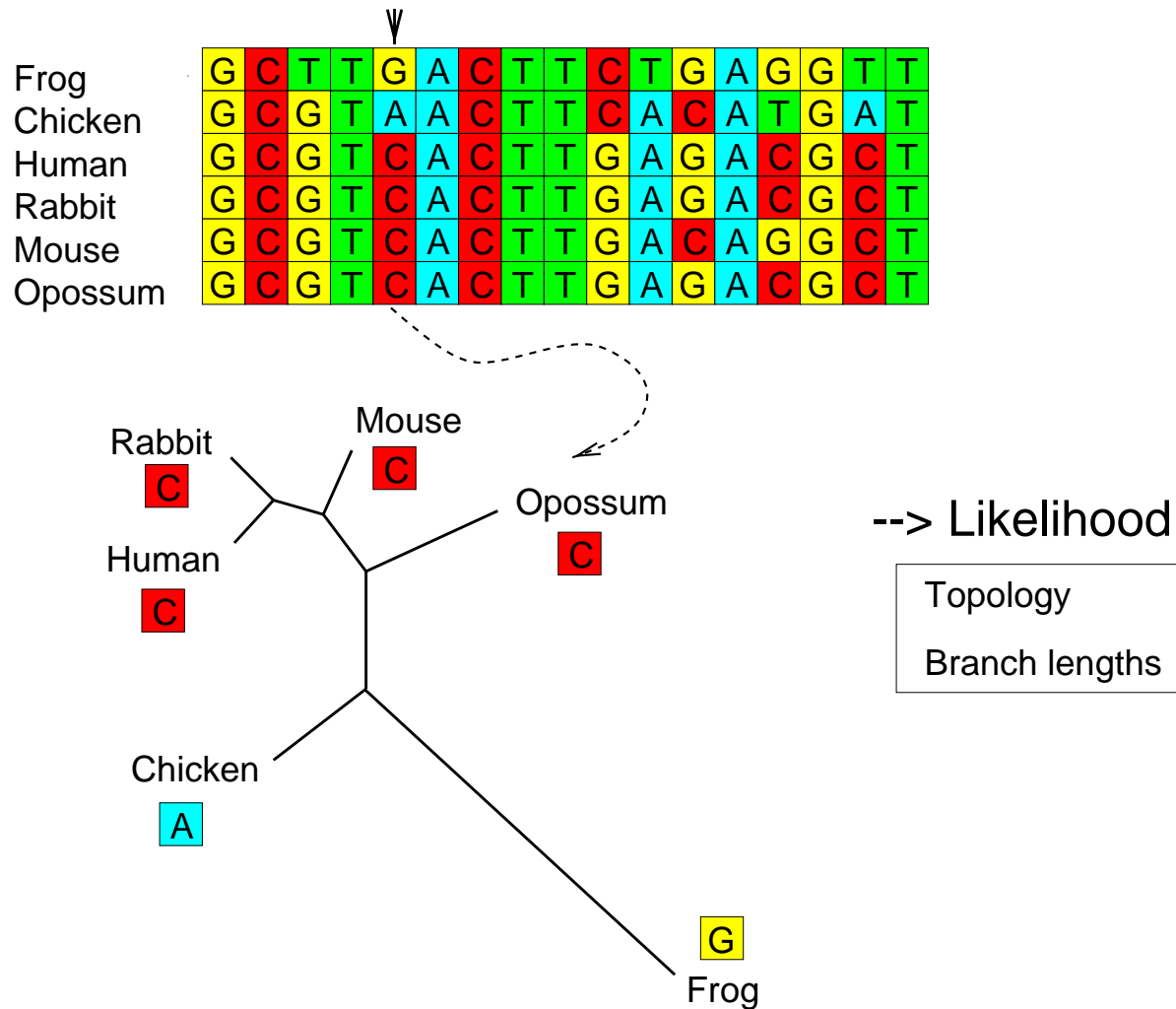
A Probabilistic Model of Evolution 2



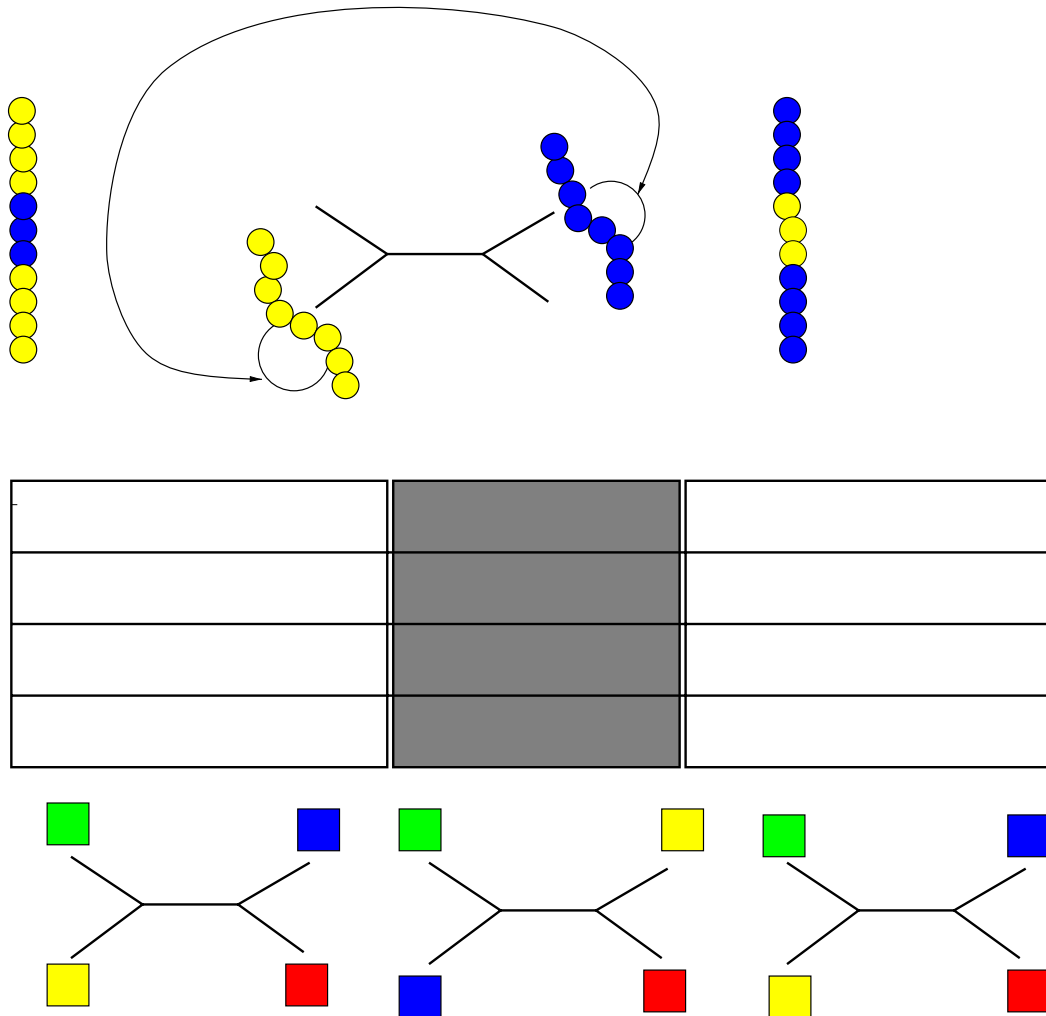
$$P(y_1, y_2, y_3, y_4, z_1, z_2 | \mathbf{w}) = P(y_1 | z_1, w_2) P(y_2 | z_1, w_1) P(z_2 | z_1, w_3) P(y_3 | z_2, w_4) P(y_4 | z_2, w_5)$$

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

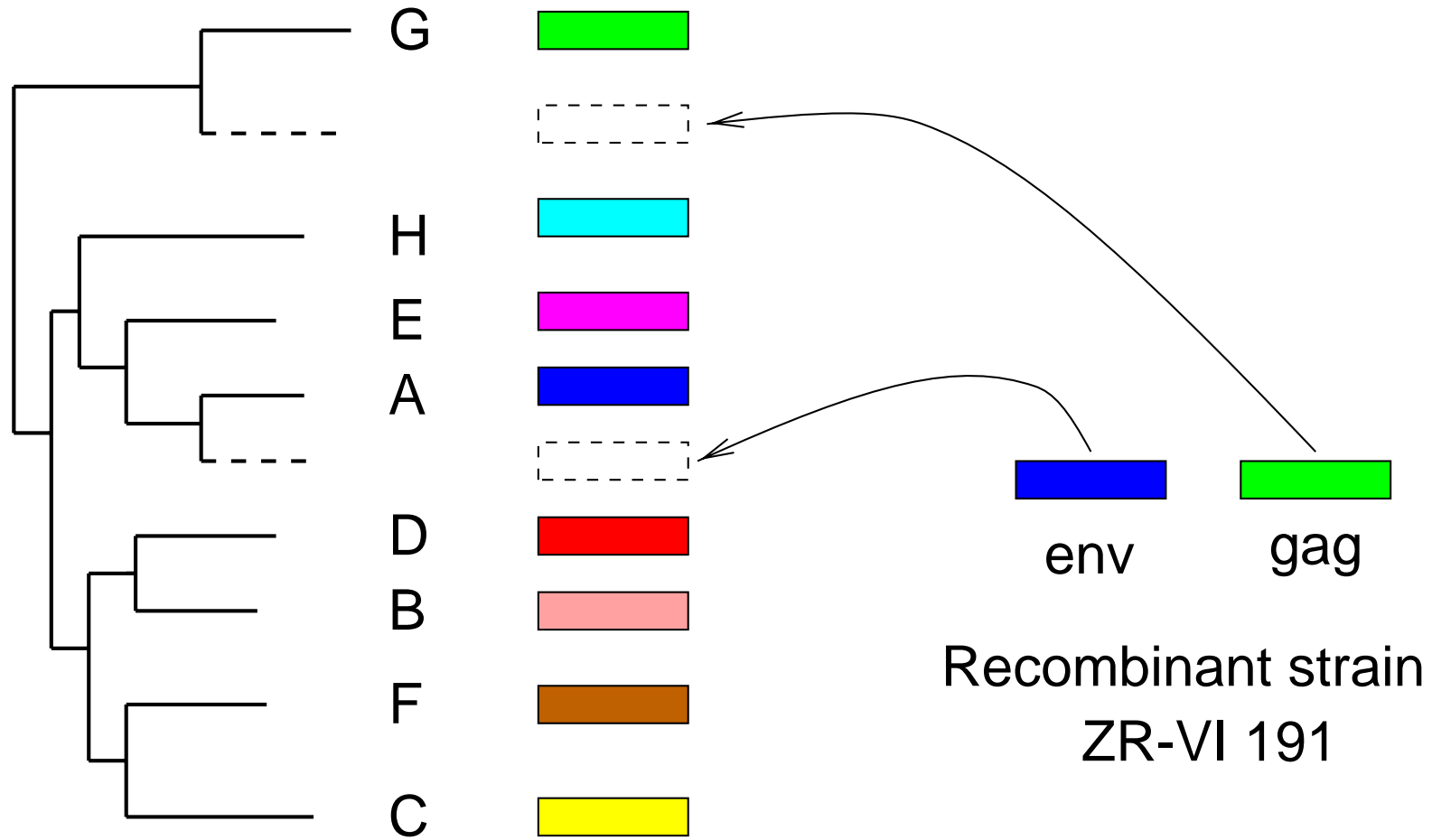
Statistical Approach to Phylogenetics



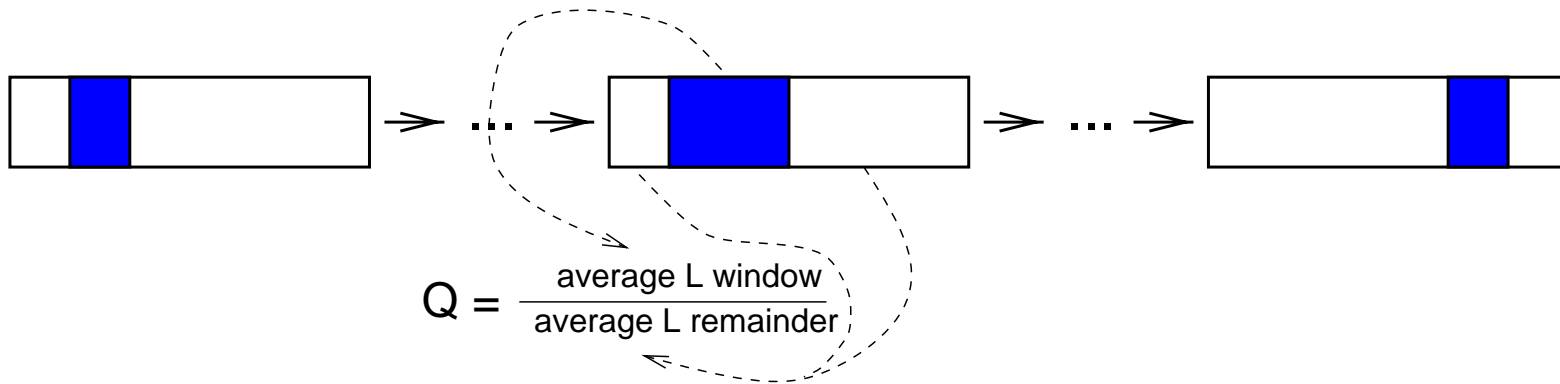
Recombination



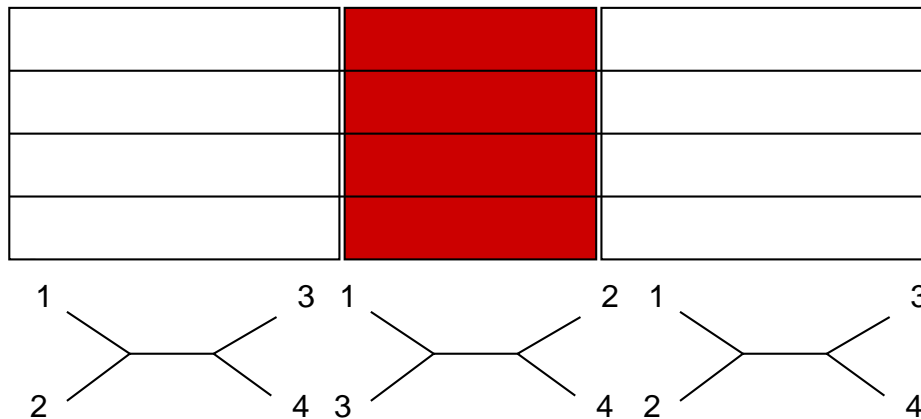
Recombination in HIV



PLATO (Grassly, Holmes)

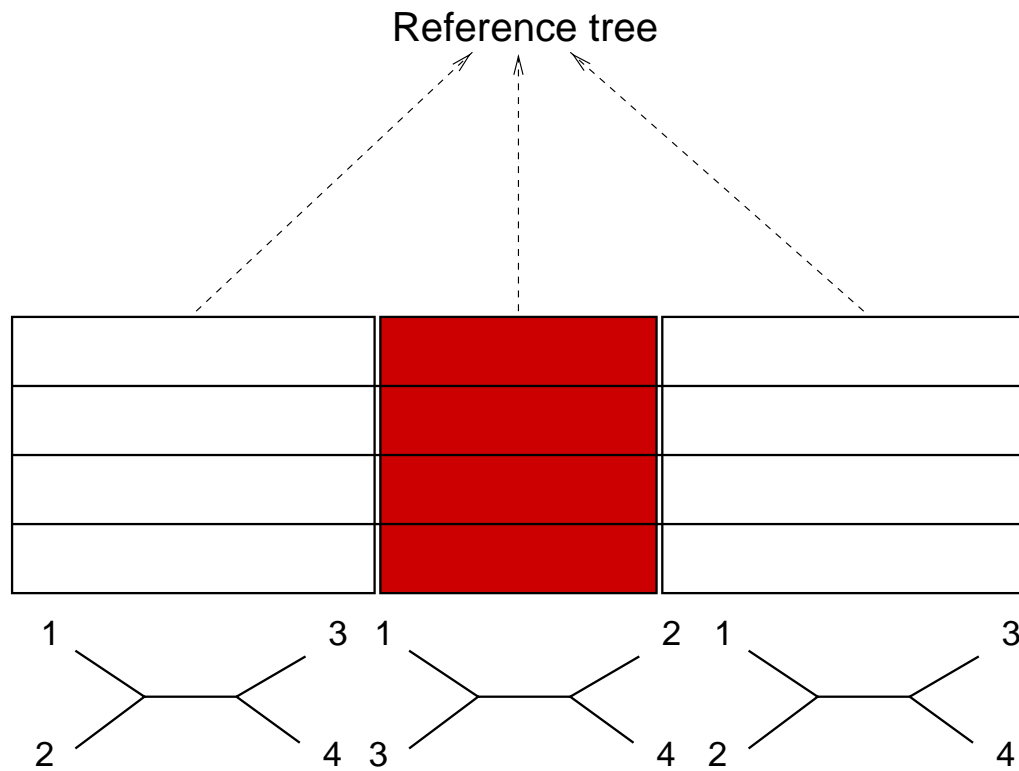


- Find maximum Q values
- Test significance with parametric bootstrapping

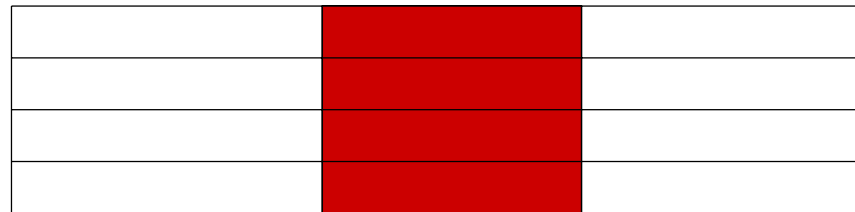


Shortcoming of PLATO

- Need a reference tree
- Obtained with global maximum likelihood



TOPAL (McGuire, Wright)



DSS small



DSS large



DSS small

DSS large

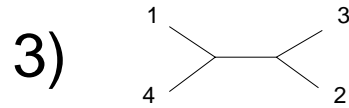
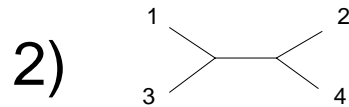
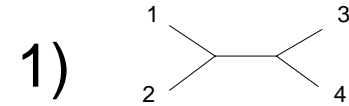
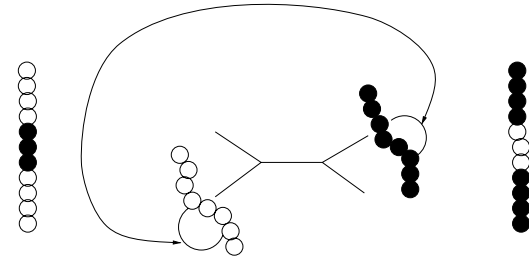
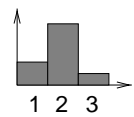
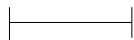
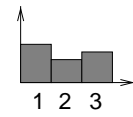
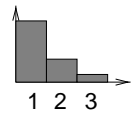
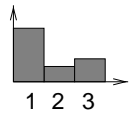
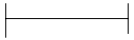
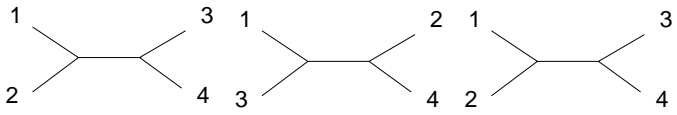
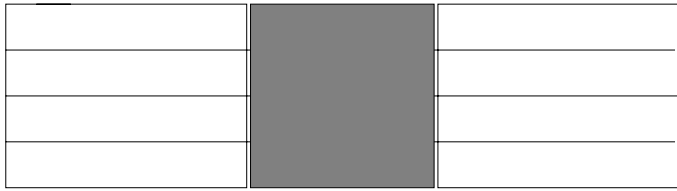


DSS small

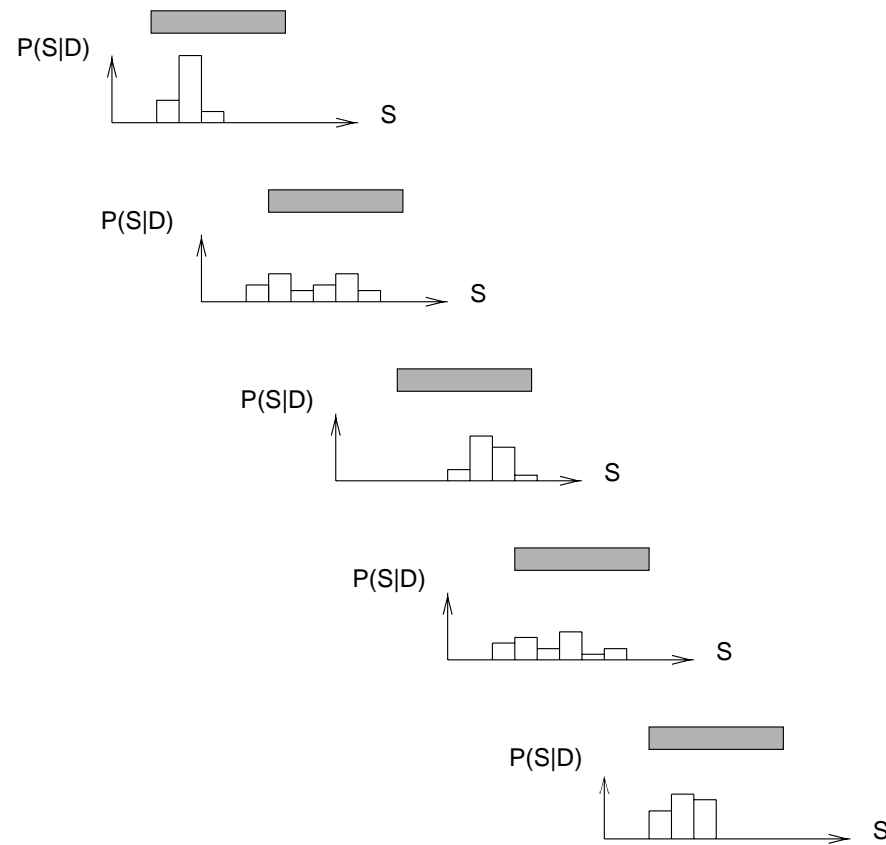
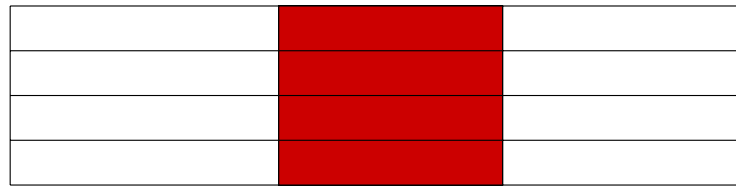


$$SoS = \sum_i \sum_k (d_{ik} - \hat{d}_{ik})^2 \quad DSS = |SoS_{left} - SoS_{right}|$$

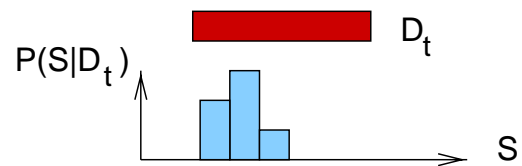
i, k	labels for taxa
\hat{d}_{ik}	fitted distances (Fitch or Neighbour Joining)
d_{ik}	true distances



Detection of Recombination with MCMC



Marginal Posterior Distribution over Tree Topologies with MCMC



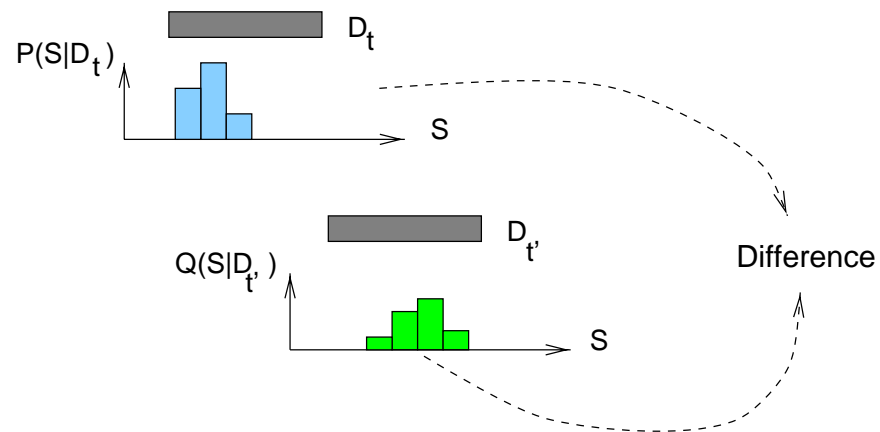
$$P(S|\mathbf{D}_t) := \int P(S, \mathbf{w}|\mathbf{D}_t) d\mathbf{w}$$

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

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

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

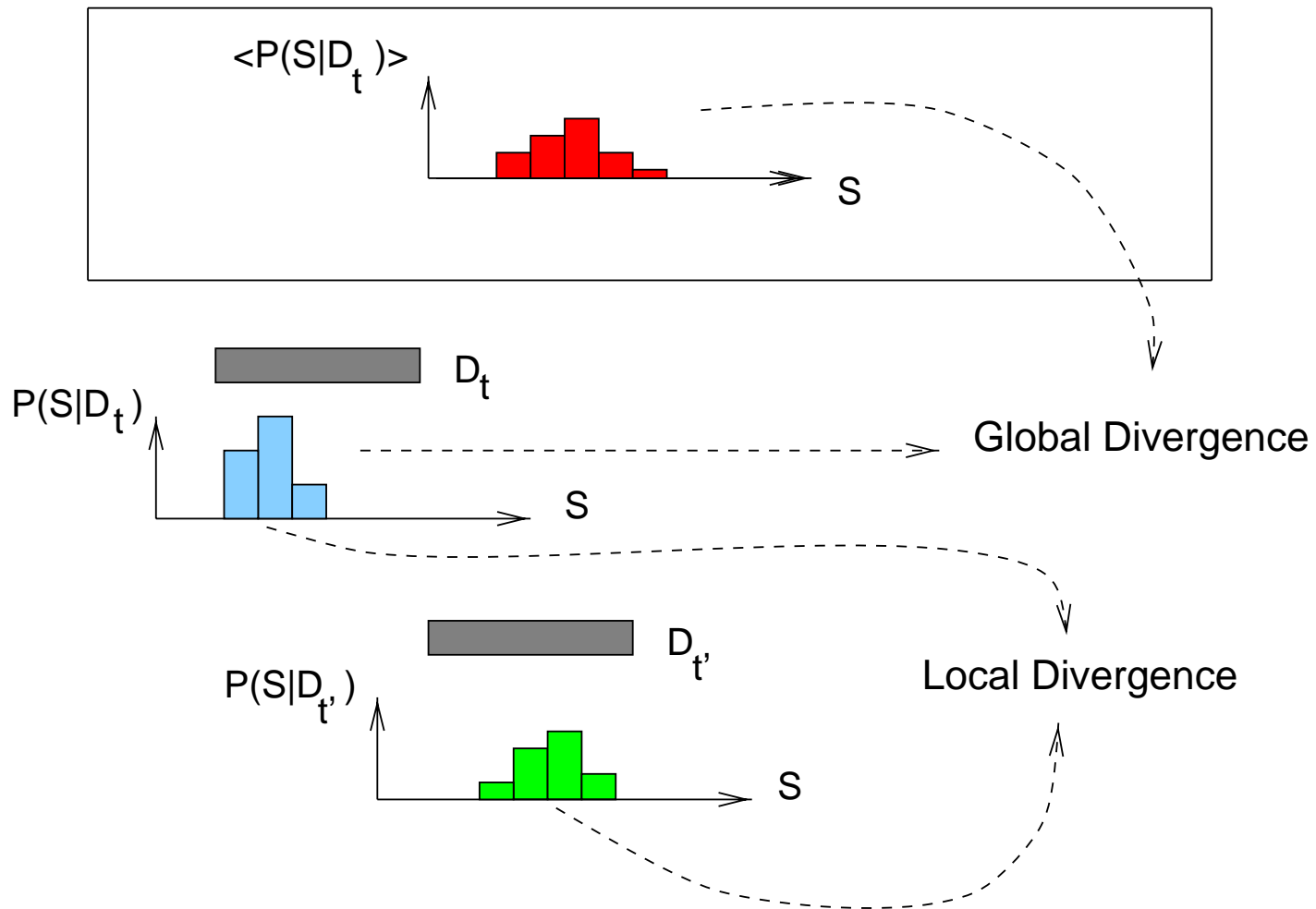
Divergence between Distributions



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

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

Local and Global Divergence Measures



Divergence Measures and Statistical Significance

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

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

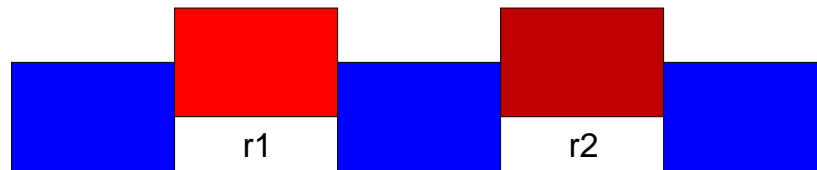
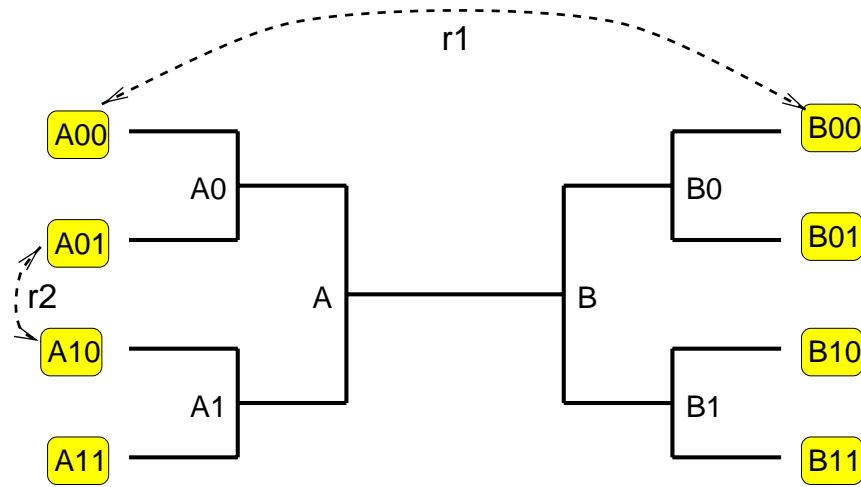
Divergence between the distributions over two adjacent windows, $P_S(t)$ and $P_S(t')$, where $\tilde{P}_S = \frac{P_S(t) + P_S(t')}{2}$ (Sibson):

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

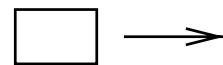
Null hypotheses: $P_S(t) = \bar{P}_S$ and $P_S(t) = P_S(t')$

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

Simulation Experiment A

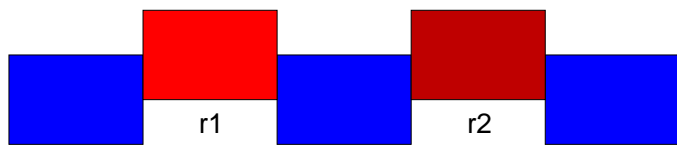
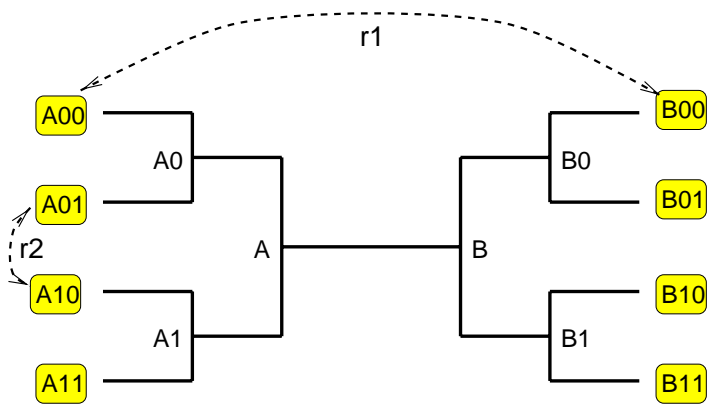


5000 nucleotides

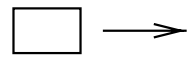


window size = 500 nucleotides

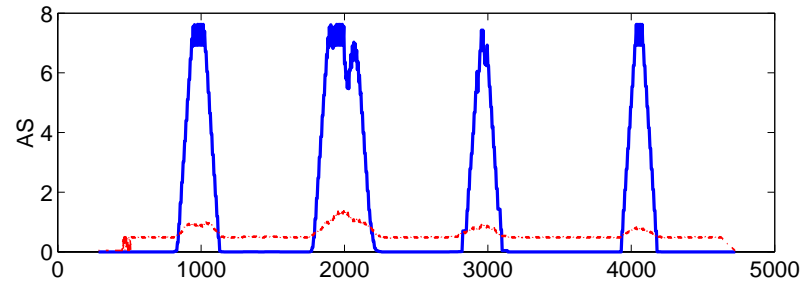
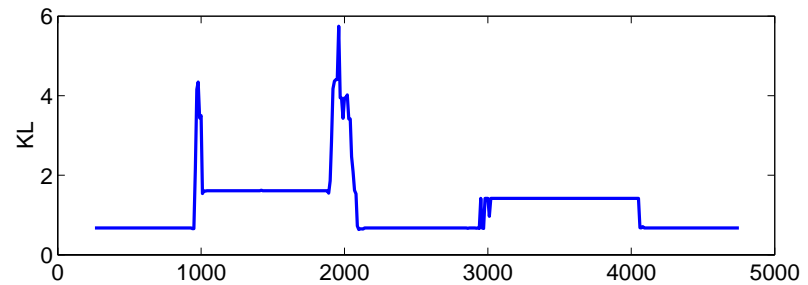
Simulation Experiment A



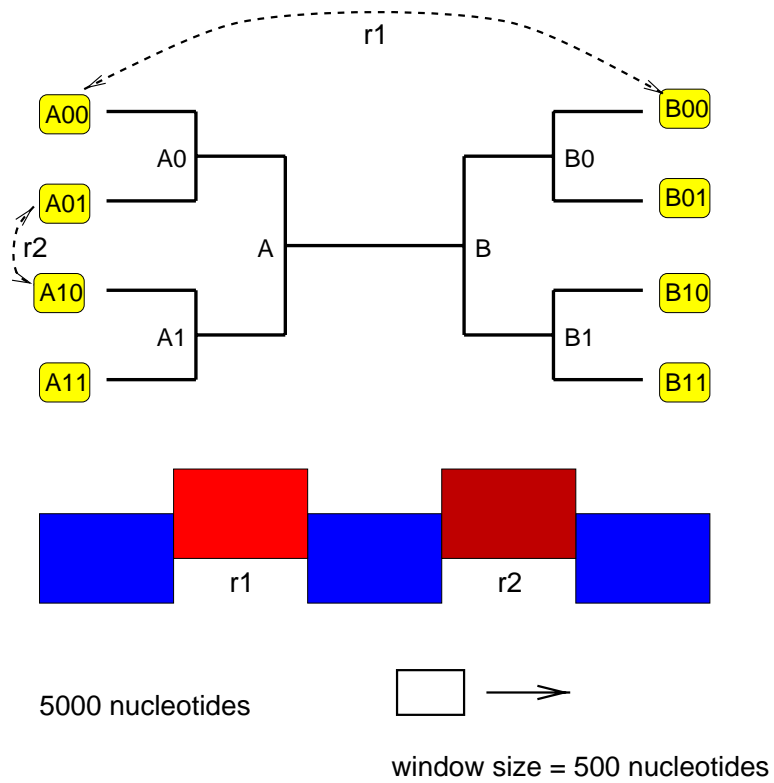
5000 nucleotides



window size = 500 nucleotides

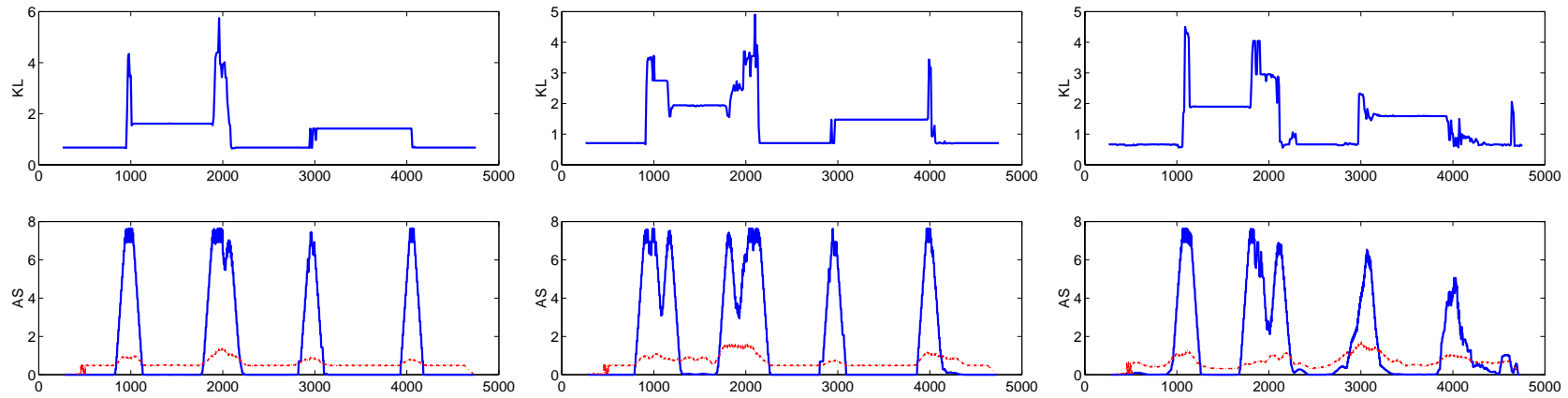


Simulation Experiment A

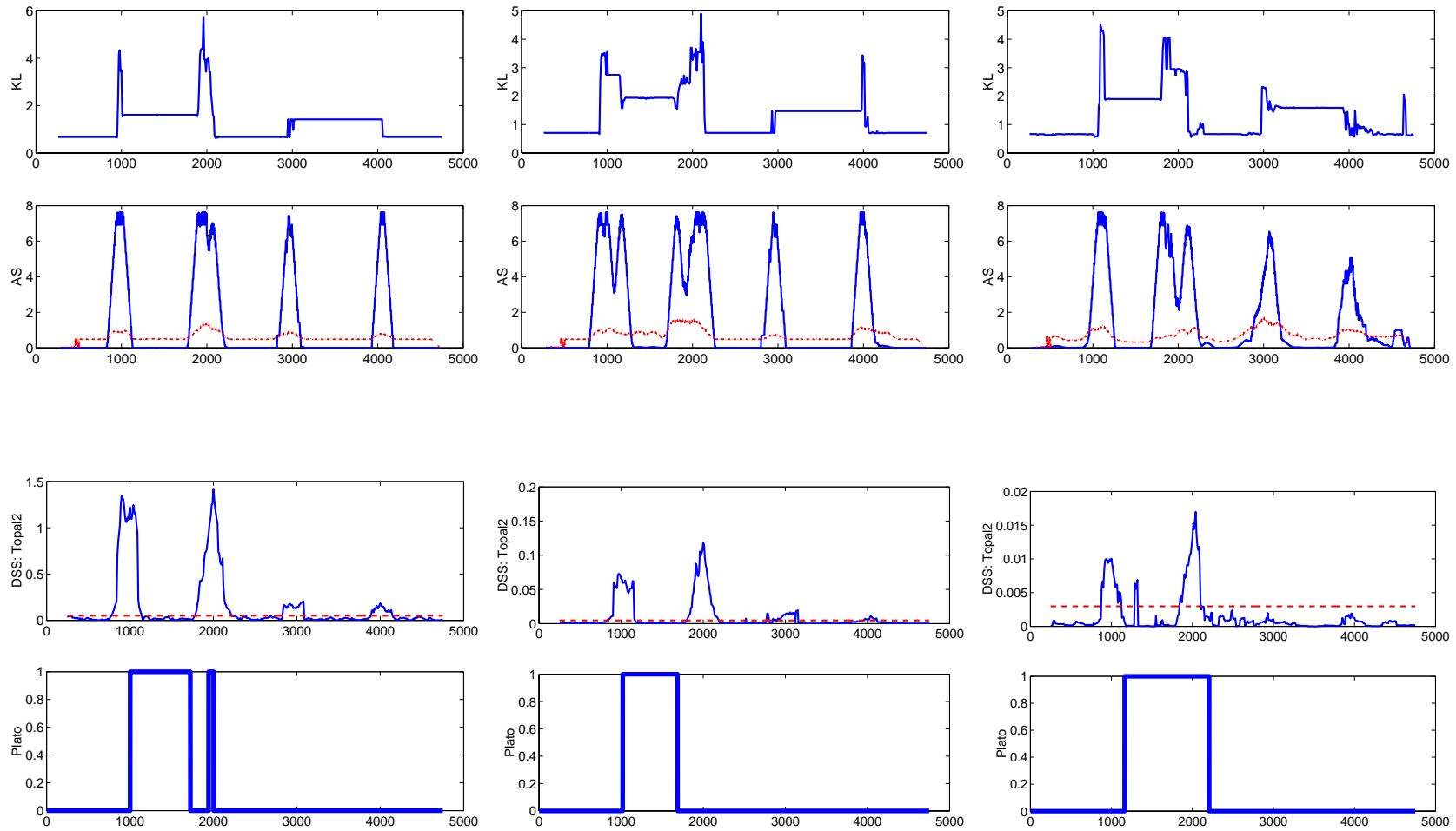


MCMC Global			
MCMC Local			
TOPAL			
PLATO			

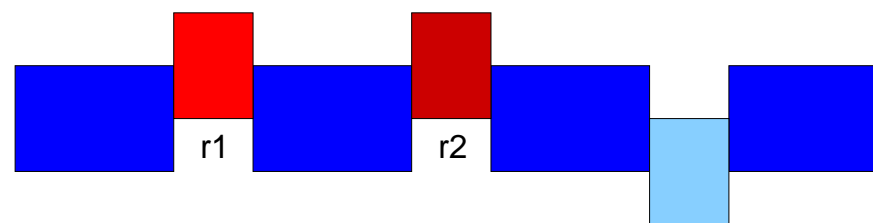
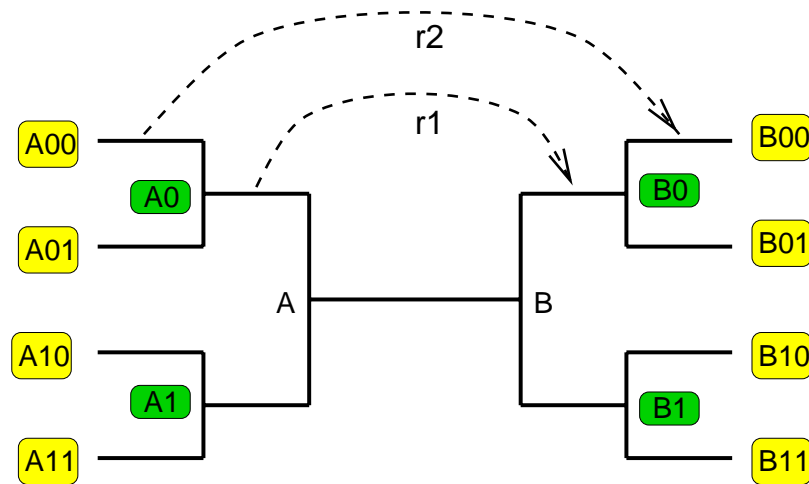
Results - Simulation Experiment A



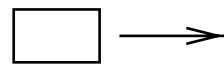
Results - Simulation Experiment A



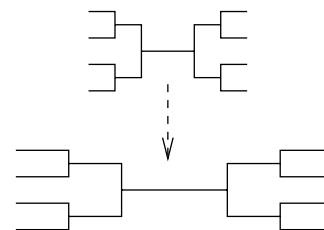
Simulation Experiment B



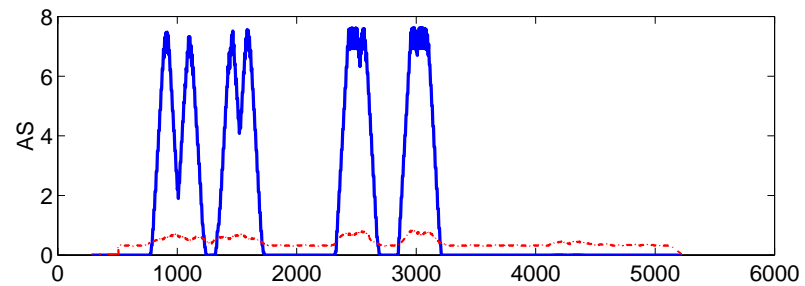
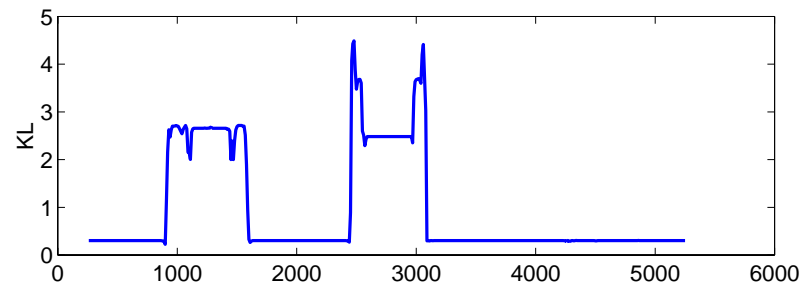
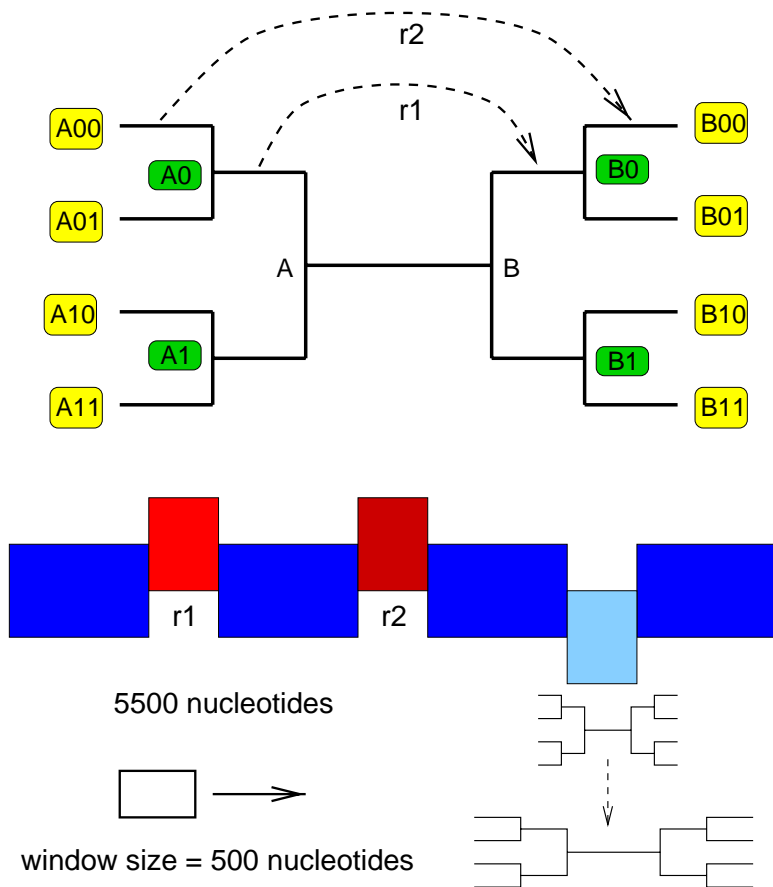
5500 nucleotides



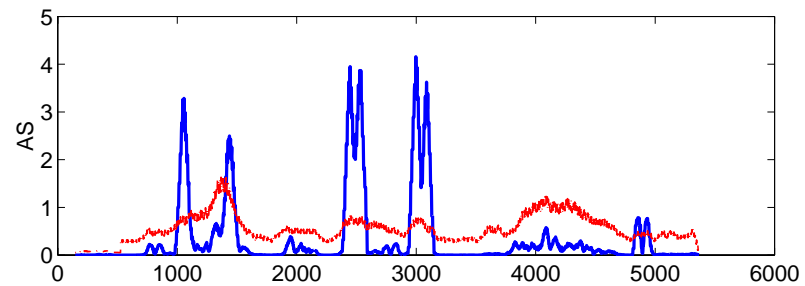
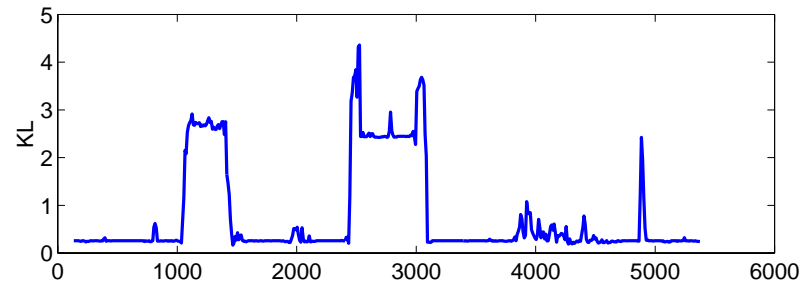
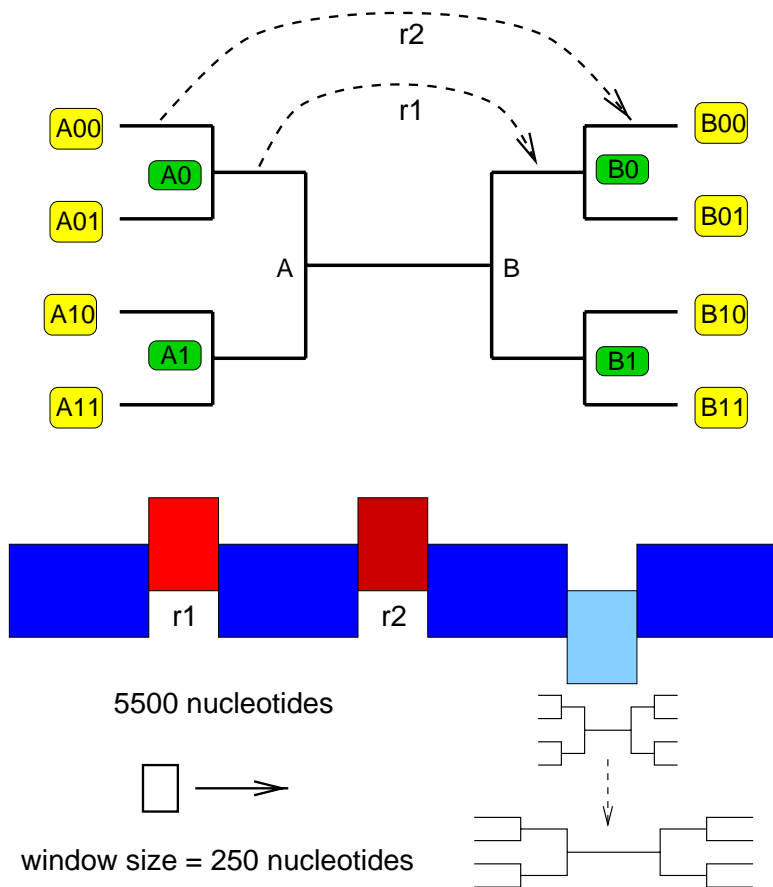
window size = 500 nucleotides



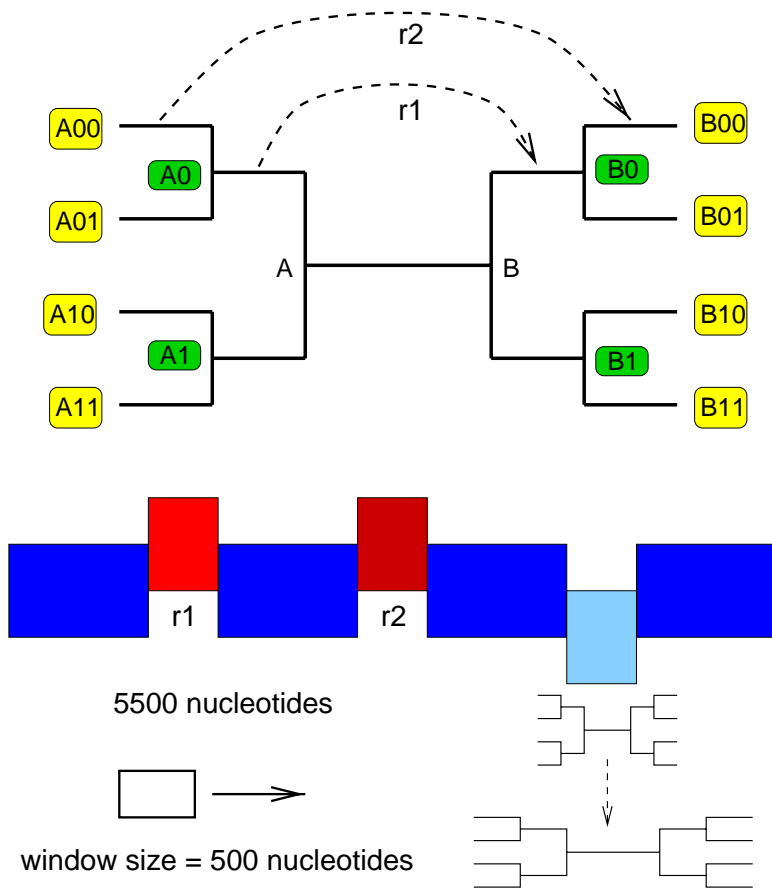
Simulation Experiment B



Simulation Experiment B: Smaller Window



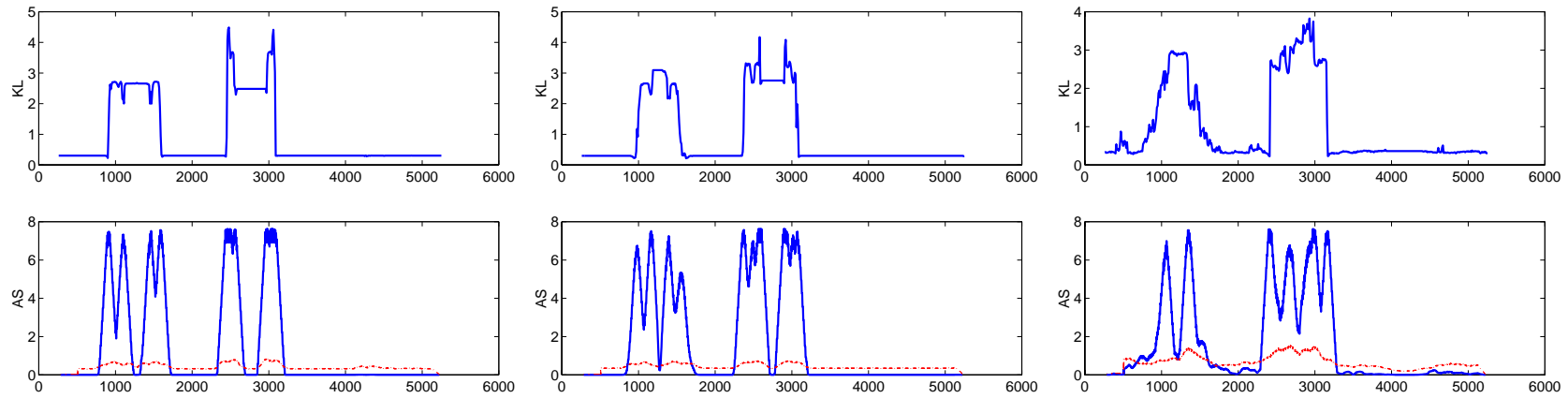
Simulation Experiment B



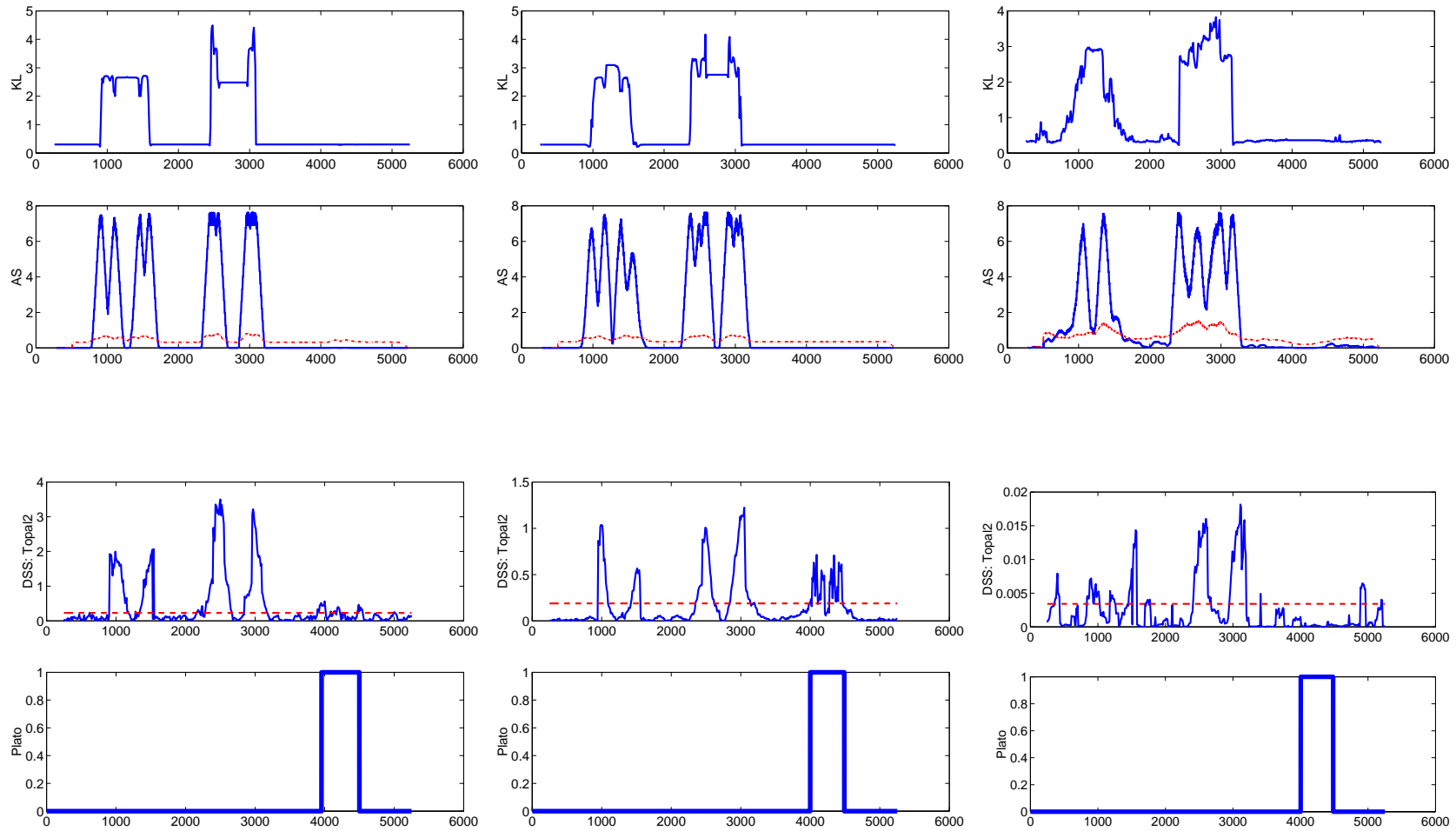
The table compares four methods across three different tree topologies. The topologies are represented by three small tree diagrams above the columns. The methods are listed in the rows. All cells in the table are shaded light red, indicating that all methods performed similarly or were applicable in all cases.

MCMC Global			
MCMC Local			
TOPAL			
PLATO			

Results - Simulation Experiment B



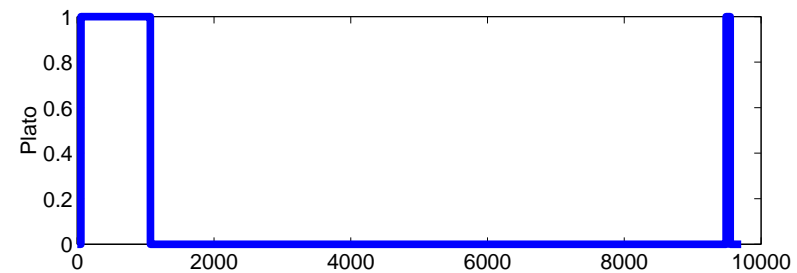
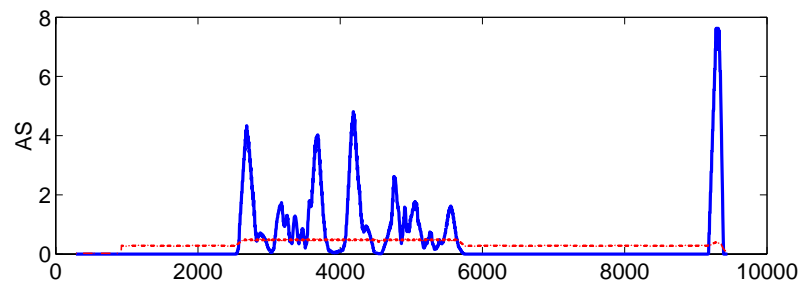
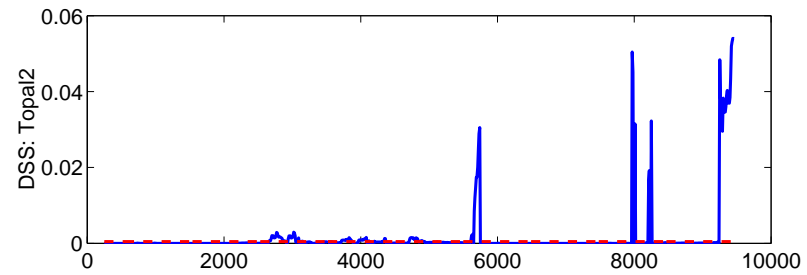
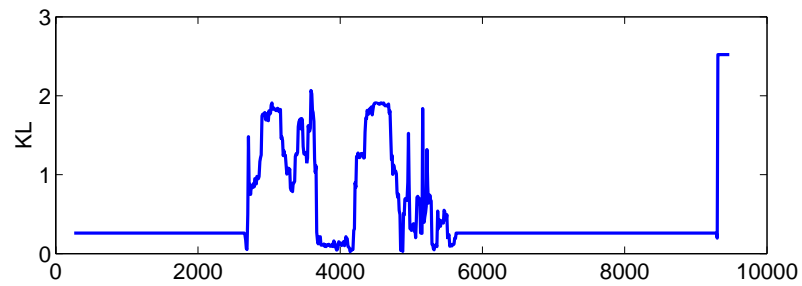
Results - Simulation Experiment B



Potato Virus Y

Four strains, 9700 bases, window size= 500 bases.

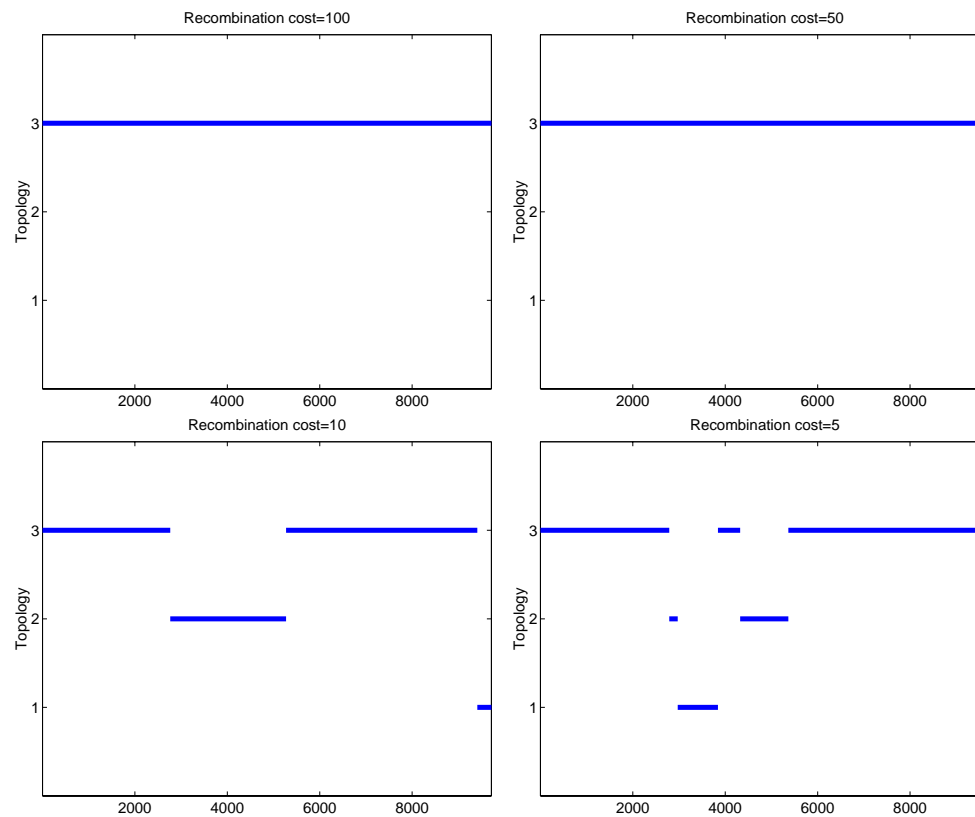
MCMC, global	TOPAL
MCMC, local	PLATO



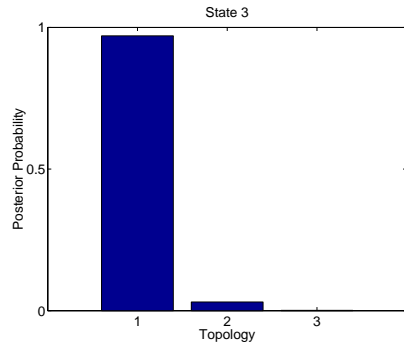
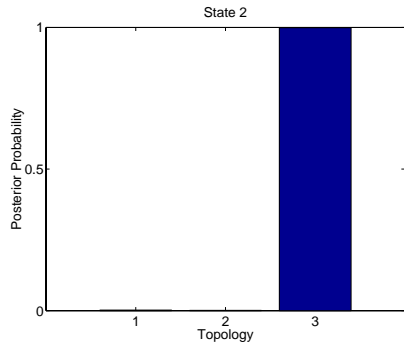
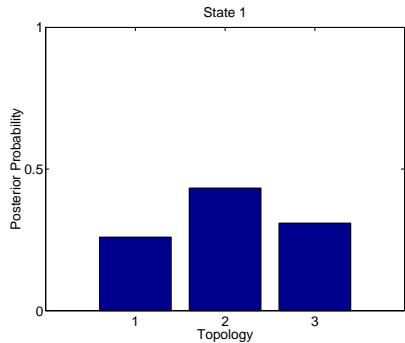
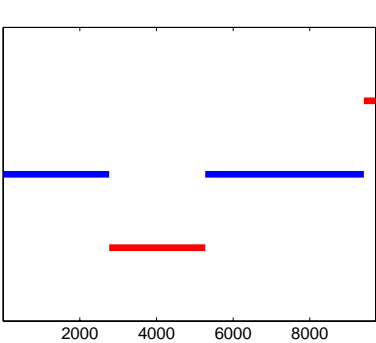
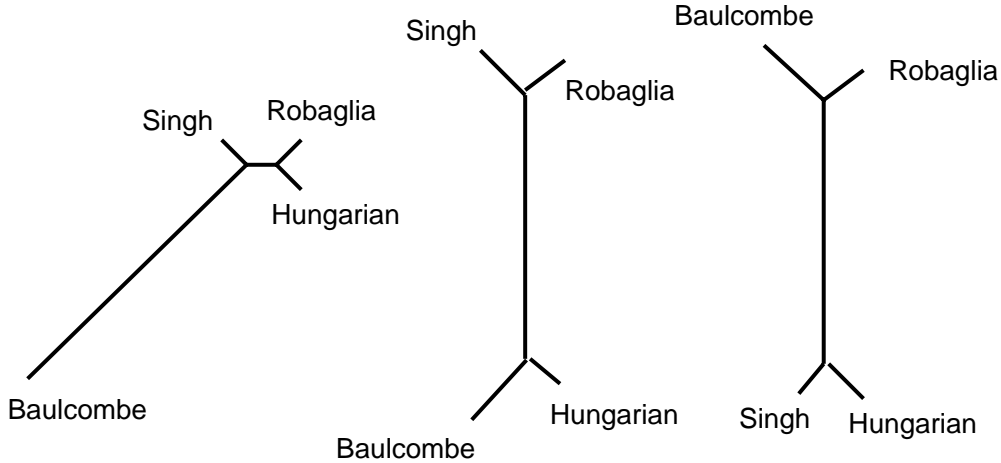
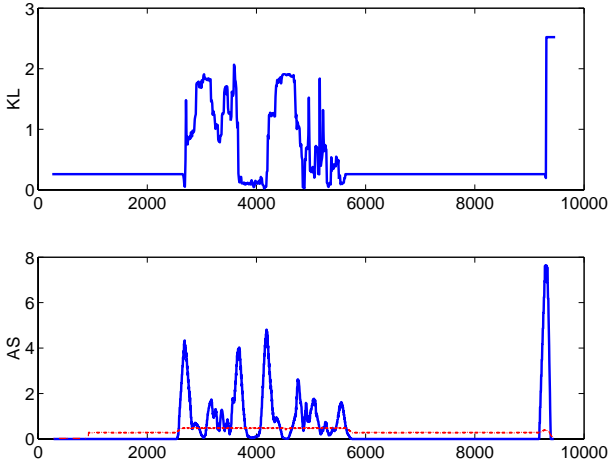
Potato Virus Y: RecPars (J. Hein)

Transition cost= 2, transversion cost= 5, recombination cost:

100	50
10	5



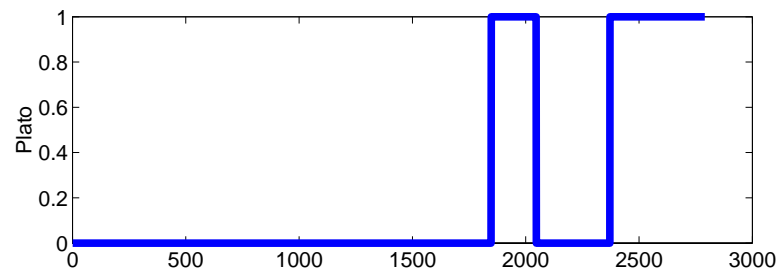
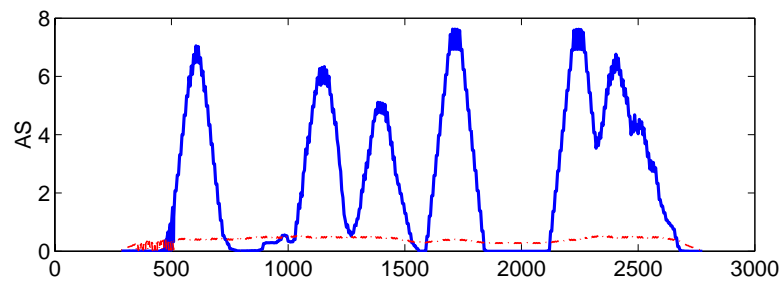
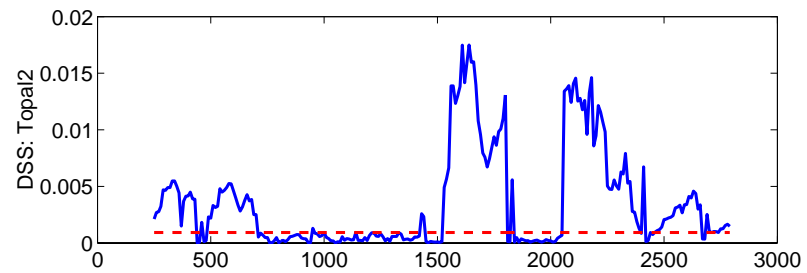
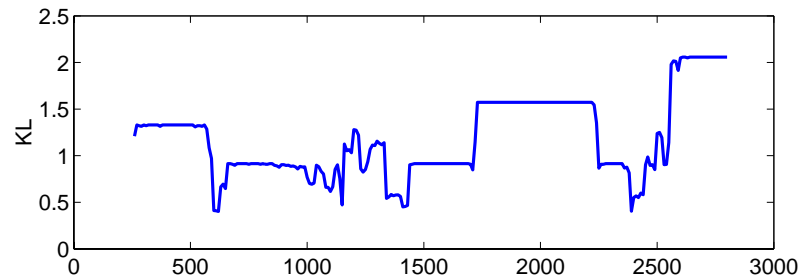
Potato Virus Y



Hepatitis B Virus

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

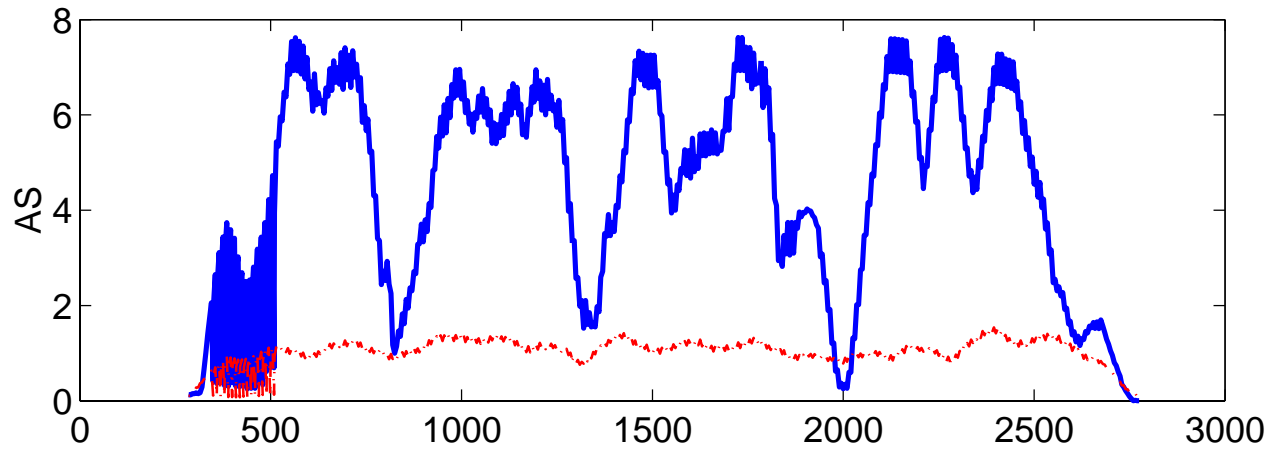
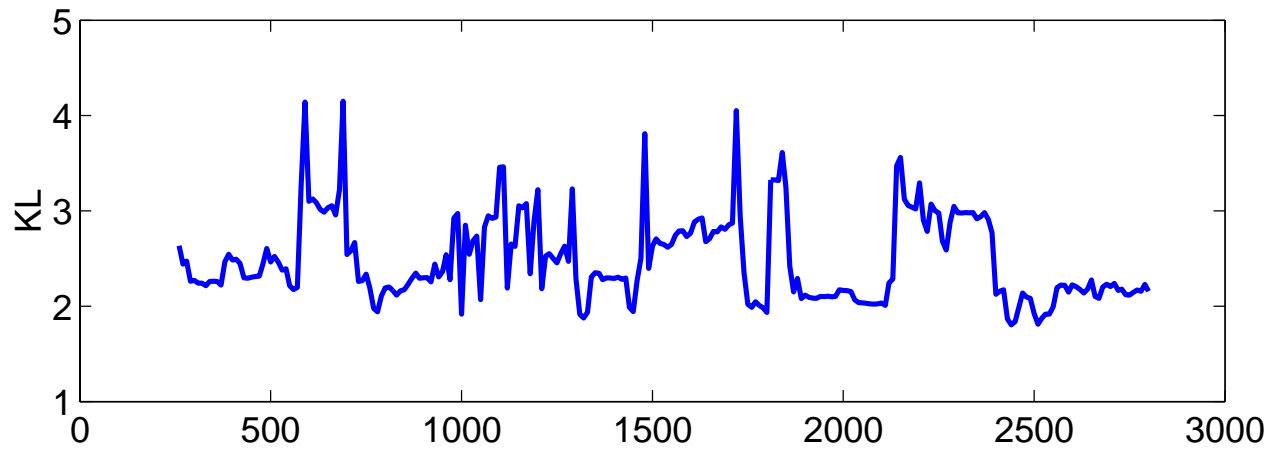
MCMC, global	TOPAL
MCMC, local	PLATO



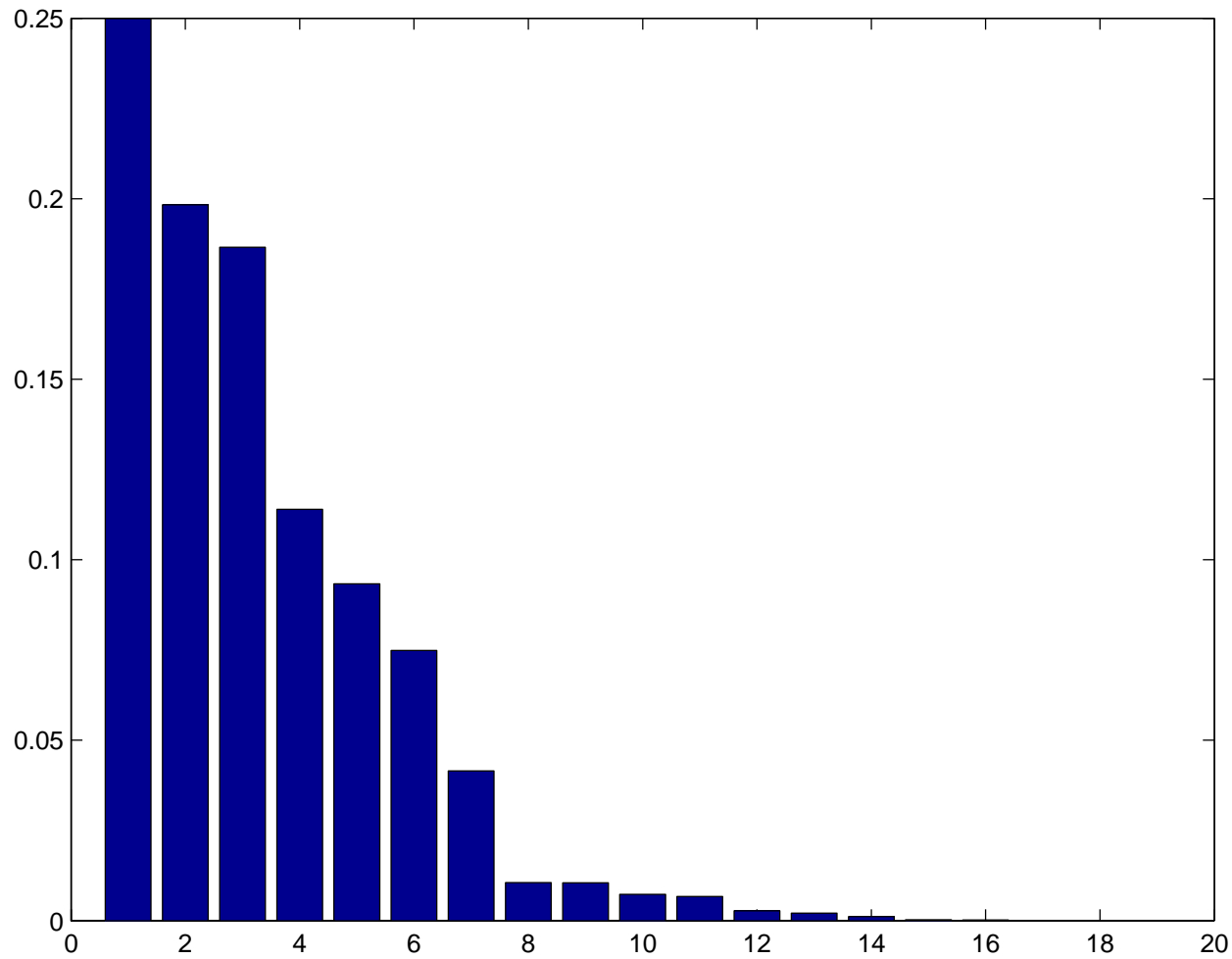
Conclusions

- **Sliding window:** marginal posterior distribution over tree topologies , conditional on the selected subset of the alignment.
- **Global divergence measure:** Kullback-Leibler divergence between a local distribution and the global distribution.
- **Local divergence measure:** Modified Kullback-Leibler divergence between adjacent local distributions.
- **Comparison** with TOPAL and PLATO on several synthetic benchmark problems.
- **Distinguishes** between recombination and rate variation .
- **Detects all** recombination events.
- **Hepatitis B virus:** New method detects breakpoints predicted with TOPAL plus two additional breakpoints.

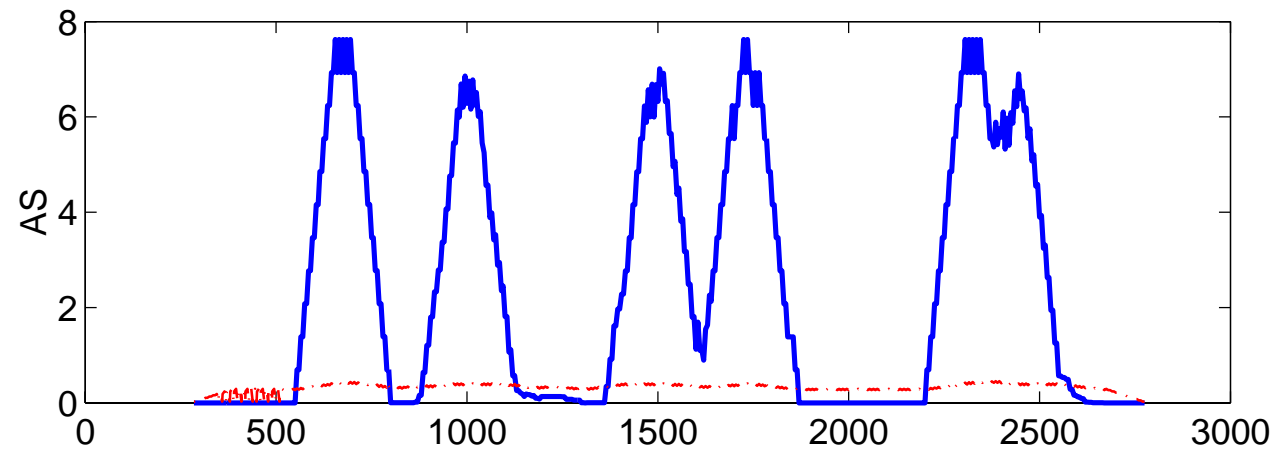
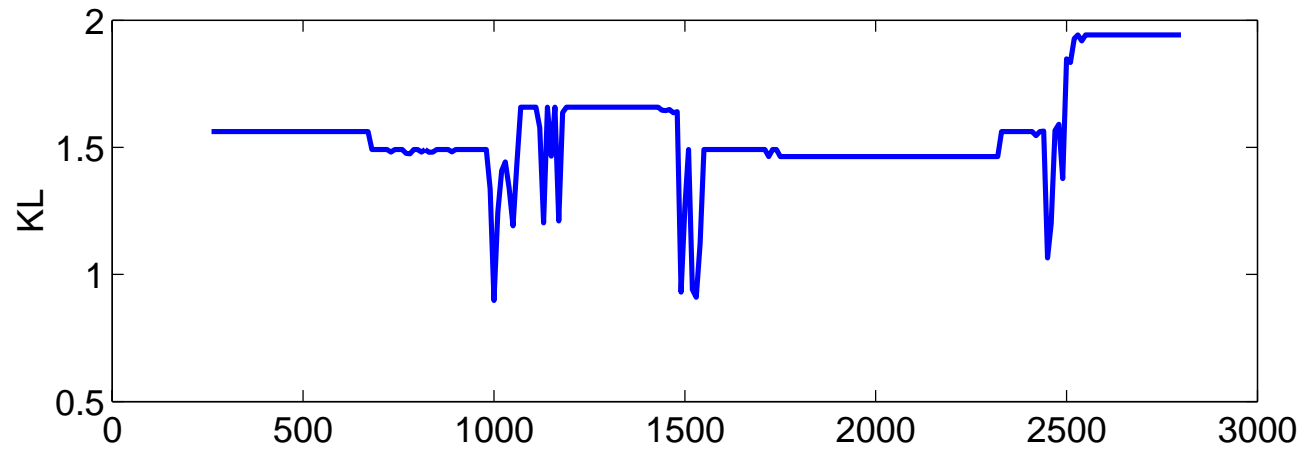
Hepatitis B Virus, 10 Strains



Hepatitis B Virus: Spectrum

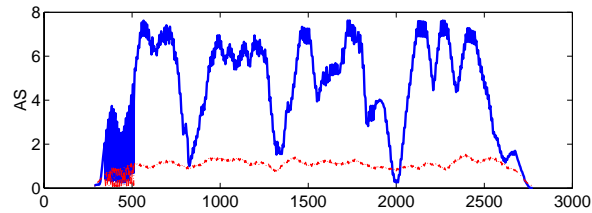
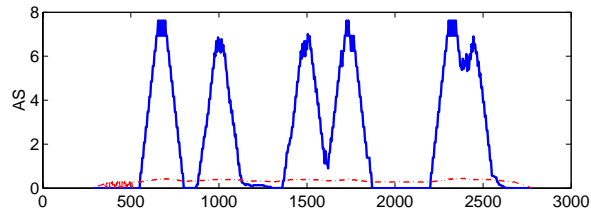
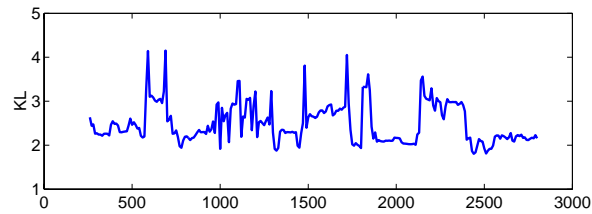
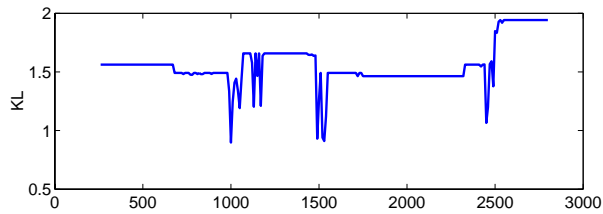
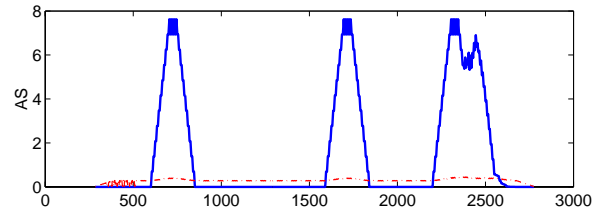
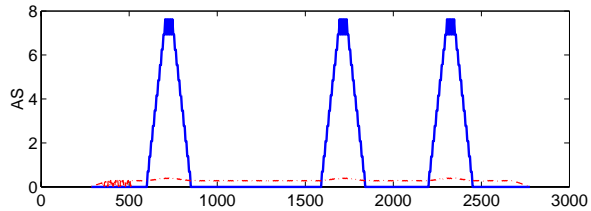
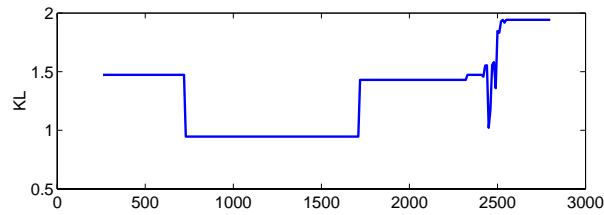
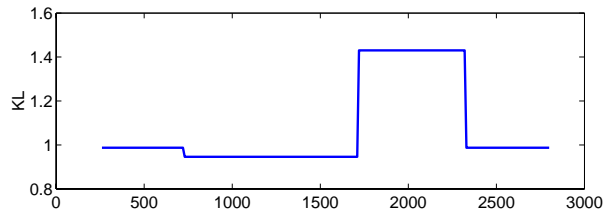


Hepatitis B Virus: Pruning, $K = 5$



Hepatitis B Virus, Pruning:

$K = 3$	$K = 4$
$K = 5$	$K = \infty$

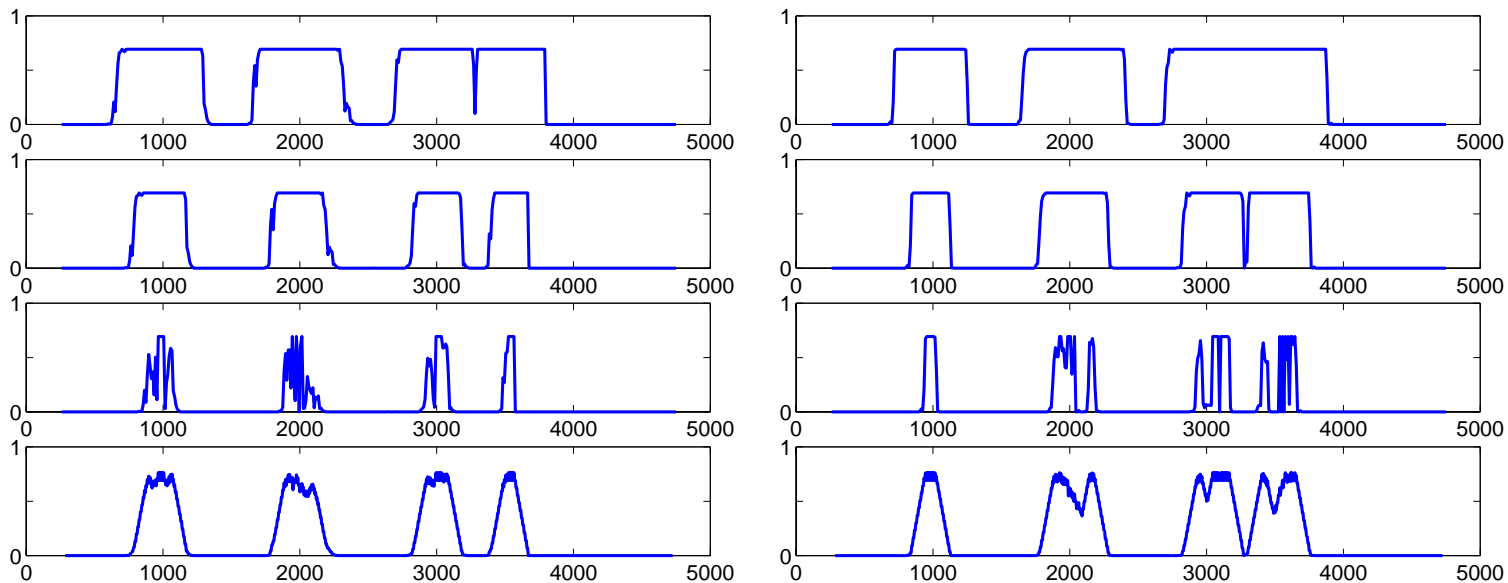


Average Divergence Measure

- Divergence measure $d[P(t), P(t + \Delta t)]$

- How to choose Δt ?

- Average over different degrees of overlap:
$$\bar{d} = \frac{1}{M} \sum_{m=1}^M d[P(t), P(t + m\Delta t)]$$



From top to bottom: 0%, 50%, 90% overlap, averaging between 50% and 90%