
Detecting Sporadic Recombination in DNA Sequence Alignments

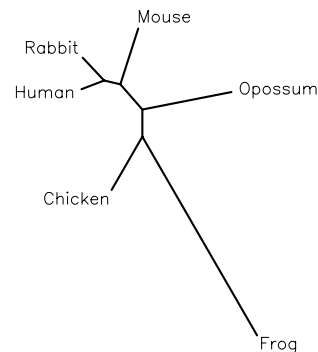
Dirk Husmeier & Frank Wright
Biomathematics and Statistics Scotland
at the Scottish Crop Research Institute
Invergowrie, Dundee DD2 5DA, UK
Email: dirk@bioss.ac.uk
<http://www.bioss.ac.uk/~dirk>

Probabilistic Approach to Phylogeny

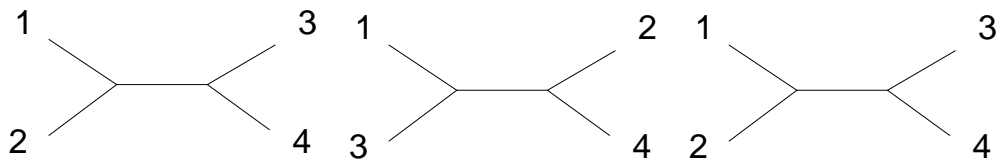
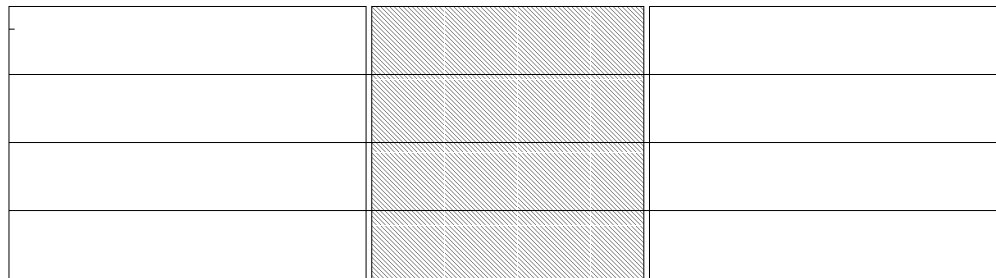
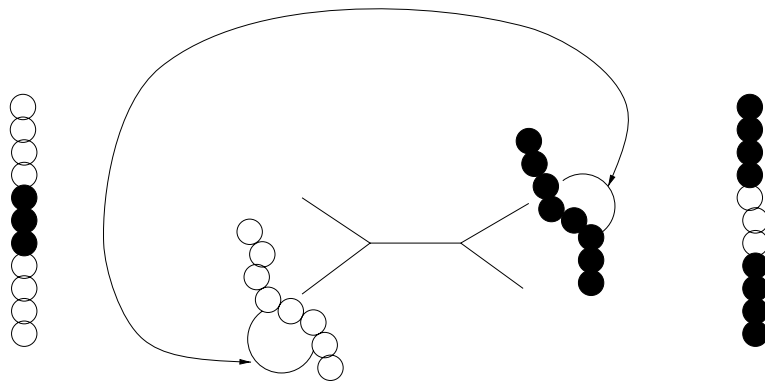
Frog	GT C GCGGGTCAAAC TTTCCGTCTCGCG
Chicken	AG C ATCGTTCTATTTTACCGGCTCCCG
Human	TG T ATCGCTCAAGATTGCCATCGCGCG
Rabbit	TG T GTCGCTCAAGATTGCCATCGCGCG
Mouse	TG T CGTGGTCTAGATTGCCATCGCGCG
Opossum	TG T ATCGCTCTAGTTTGCCAGCTCCCG

$$\mathbf{D} = (\mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_N) \quad P(\mathbf{D}|\mathbf{w}, S) = \prod_{t=1}^N P(\mathbf{y}_t|\mathbf{w}, S)$$

Optimisation of the topology S and the branch lengths \mathbf{w} by maximum likelihood .



Recombination

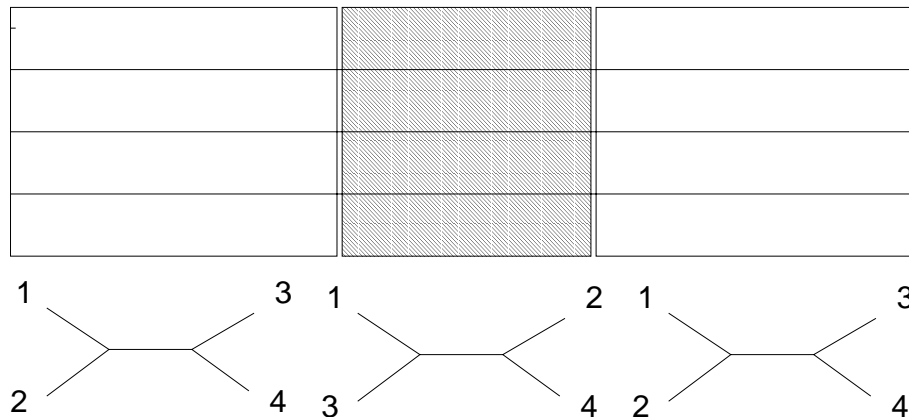


PLATO (Grassly, Holmes)

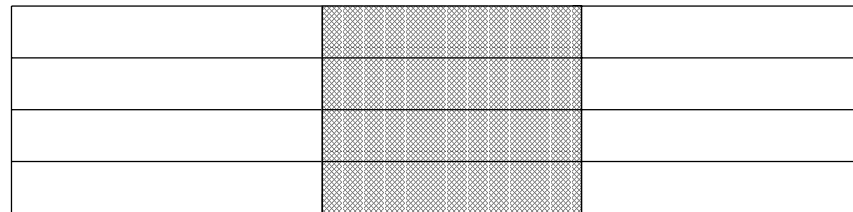
- Obtain a **global** phylogenetic tree, $\hat{\psi} = (\hat{\mathbf{w}}, \hat{S})$, e.g. by maximum likelihood on the whole sequence alignment **D**.
- Compute the **site likelihoods**, $L_t = \ln P(\mathbf{y}_t | \hat{\psi})$; $t = 1, \dots, N$.
- Look for **subsets** of significantly low likelihood. Compute the statistic

$$Q = \frac{\sum_{t=bW}^{(b+1)W-1} L_t}{W} / \frac{\sum_{t=1}^{bW-1} L_t + \sum_{t=(b+1)W}^N L_t}{N - W}$$

for **all positions** and **varying window size** ($5 \leq W \leq N/2$).



TOPAL (McGuire, Wright)



DSS small

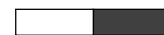


DSS large



DSS small

DSS large



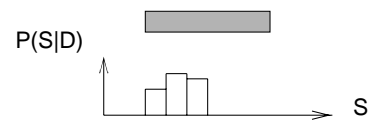
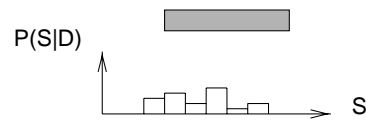
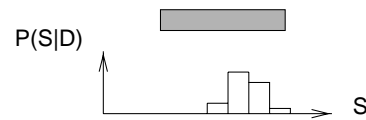
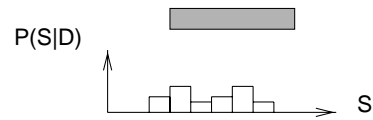
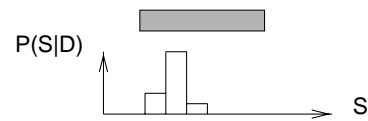
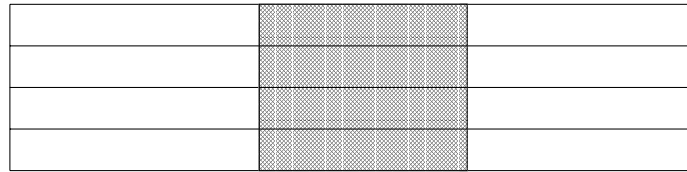
DSS small



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

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

Detection of Recombination with MCMC



Detection of Recombination with MCMC (→ BAMBE)

Marginal posterior distribution over tree topologies

$$P_k(t) := P(k|\mathbf{D}_t) = \int P(k, \mathbf{w}|\mathbf{D}_t) d\mathbf{w}$$
$$P(k, \mathbf{w}|\mathbf{D}_t) \approx \frac{1}{N} \sum_{i=1}^N \delta_{k, k_{ti}} \delta(\mathbf{w} - \mathbf{w}_{ti})$$
$$P_k(t) = \frac{1}{N} \sum_{i=1}^N \delta_{k, k_{ti}} = \frac{N_k(t)}{N}$$

Entropy

$$H(t) = - \sum_k P_k(t) \ln P_k(t) \quad 0 \leq H(t) \leq \ln K$$

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

$$KL(P, Q) = \sum_k P_k \ln \left(\frac{P_k}{Q_k} \right)$$

Divergence measures and statistical significance

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

$$d[P_k(t), \bar{P}] = \sum_k P_k(t) \ln \left(\frac{P_k(t)}{\bar{P}_k} \right)$$

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

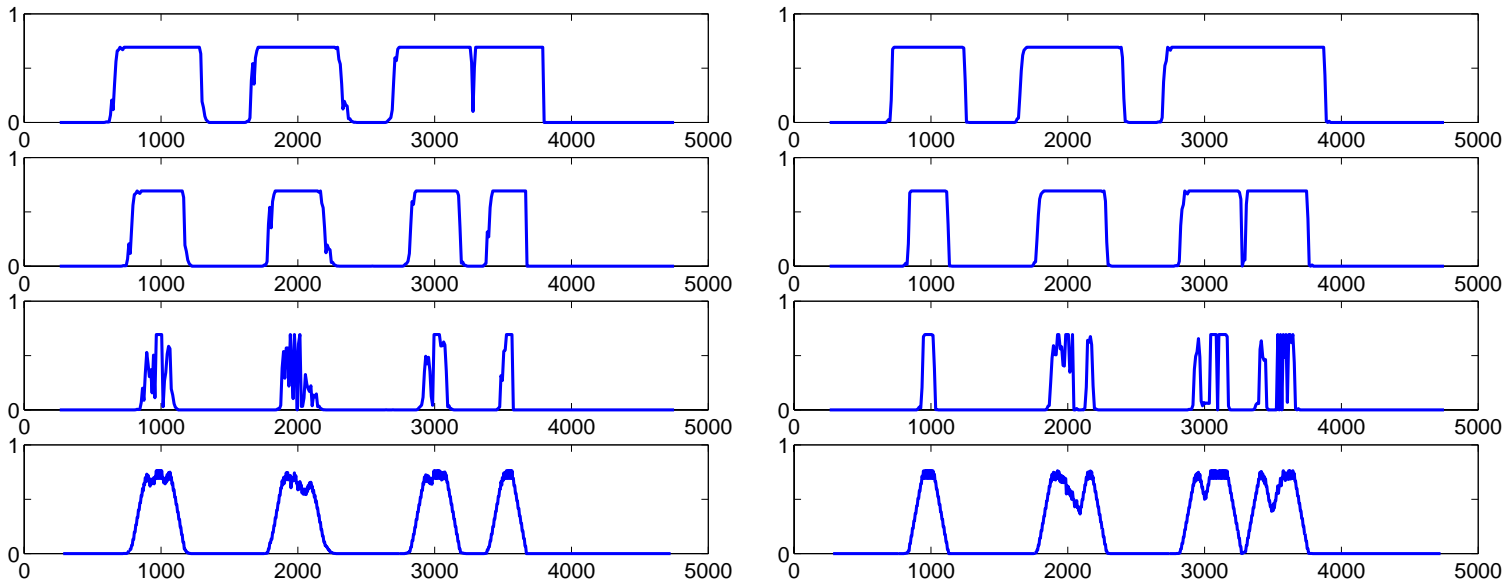
$$d[P_k(t), P_k(t')] = \frac{1}{2} \sum_k \left[P_k(t) \ln \left(\frac{P_k(t)}{\tilde{P}_k} \right) + P_k(t') \ln \left(\frac{P_k(t')}{\tilde{P}_k} \right) \right]$$

Null hypotheses: $P_k(t) = \bar{P}_k$ and $P_k(t) = P_k(t')$

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

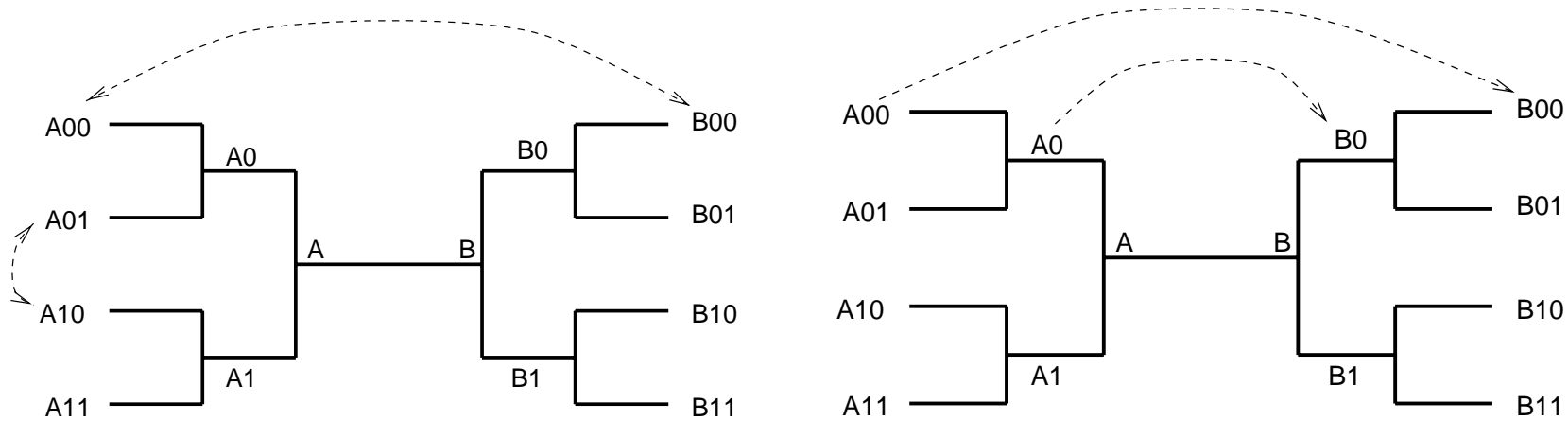
Average Divergence Measure

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



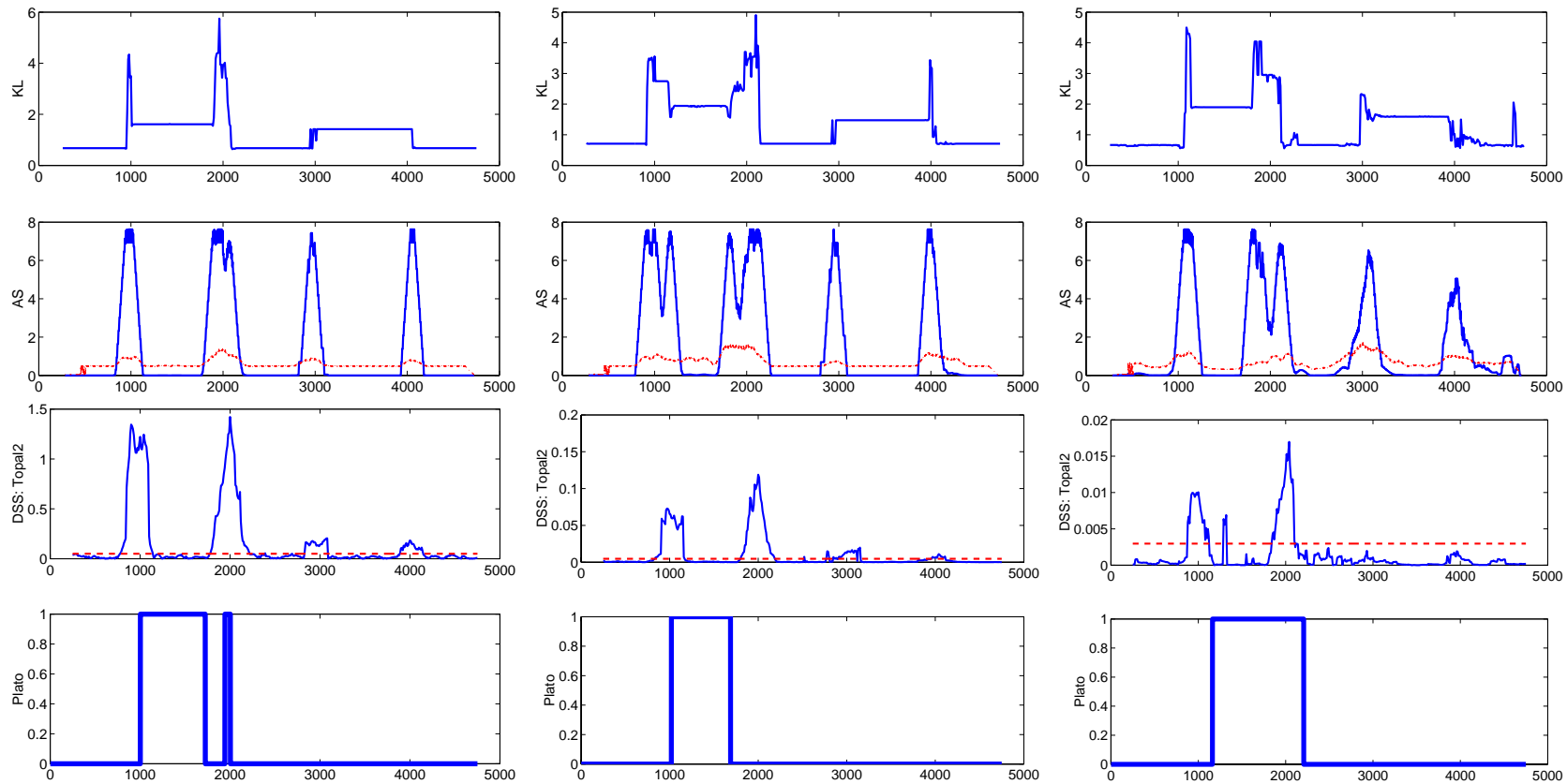
From top to bottom: 0%, 50%, 90% overlap, averaging between 50% and 90%

Simulation Experiments



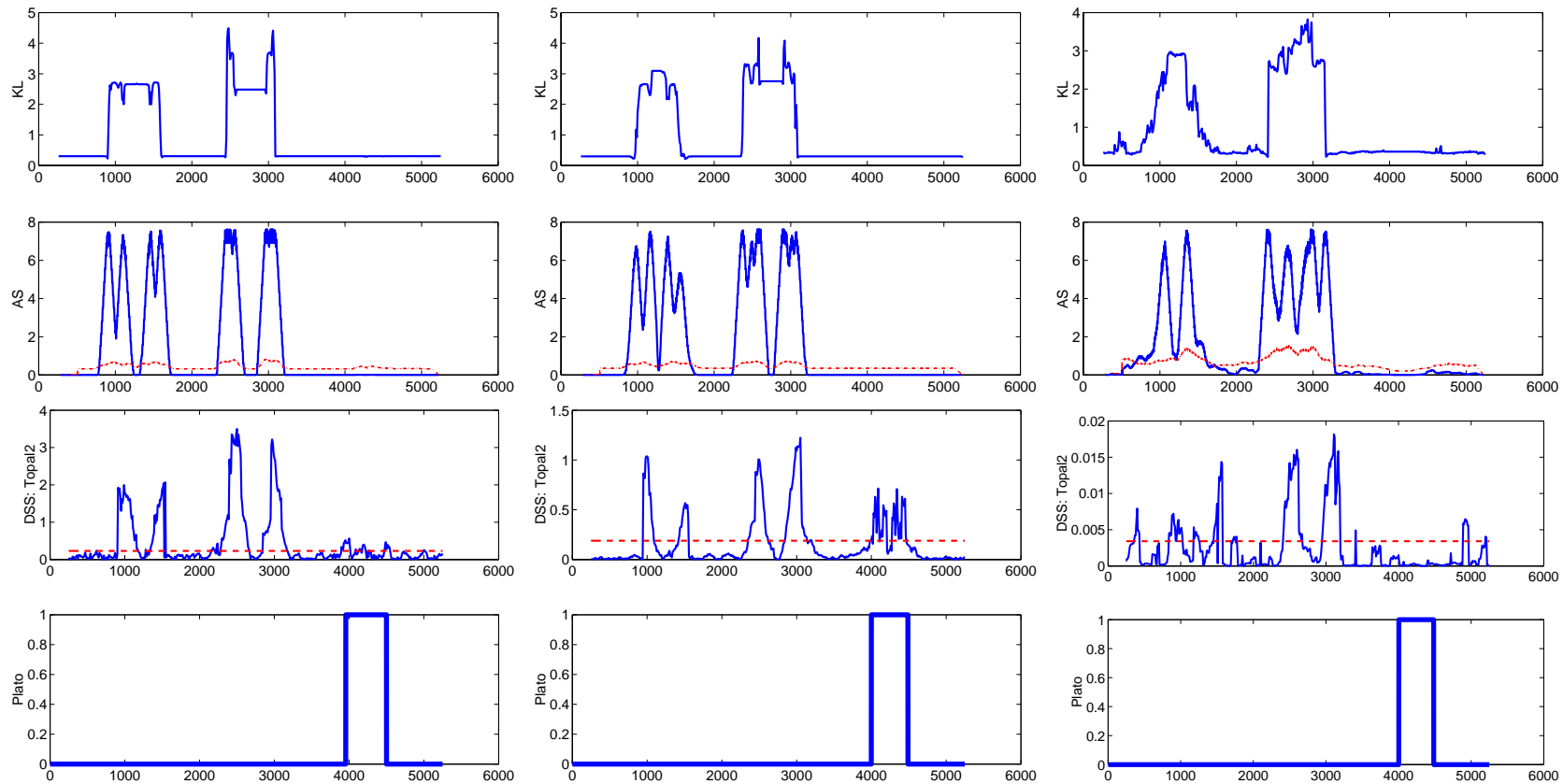
Exp	Recombination			Rate Variation	Basic Rate
	Recent, close	Recent, distant	Ancient		
A.1	1000-2000	3000-4000	–	–	0.1
A.2	1000-2000	3000-4000	–	–	0.025
A.3	1000-2000	3000-4000	–	–	0.01
B.2	–	2500-3000	1000-1500	4000-4500	0.1
B.3	–	2500-3000	1000-1500	4000-4500	0.05
B.4	–	2500-3000	1000-1500	4000-4500	0.01

Results - Simulation Experiments A



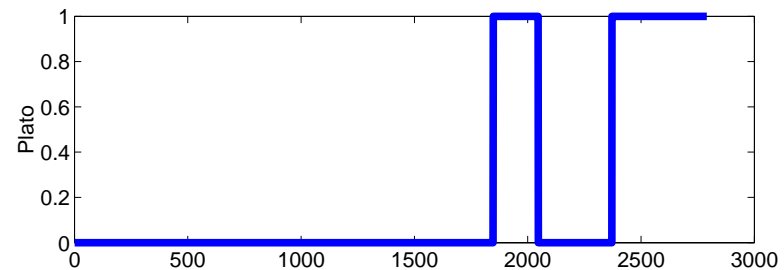
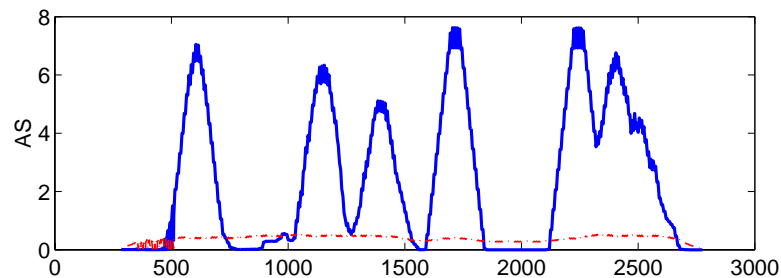
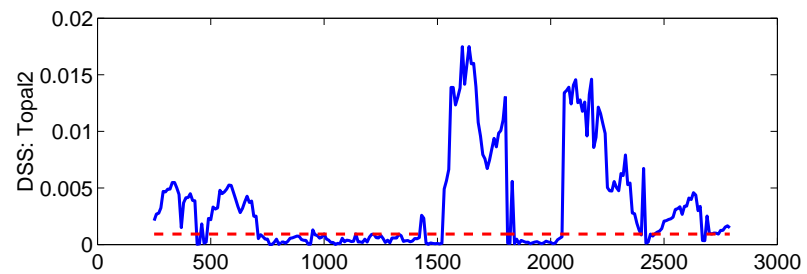
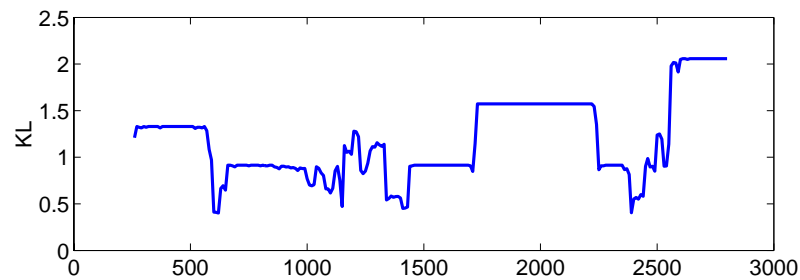
- **Top to bottom:** Kullback-Leibler divergence, average Simpson divergence, TOPAL, PLATO
- **From left to right:** $w = 0.1$, $w = 0.025$, $w = 0.01$

Results - Simulation Experiments B



- **Top to bottom:** Kullback-Leibler divergence, average Sibson divergence, TOPAL, PLATO
- **From left to right:** $w = 0.1$, $w = 0.05$, $w = 0.01$

Hepatitis B Virus



- **Data:** Five strains of Hepatitis B virus.
- **Left:** Kullback-Leibler divergence (**top**) and average Simpson divergence (**bottom**).
- **Right:** TOPAL (**top**) and PLATO (**bottom**).

Conclusions

- Sliding window: **marginal posterior distribution over tree topologies** , conditional on the selected subset of the alignment.
- Global divergence measure: **Kullback-Leibler divergence** between the local and the global distributions.
- Local divergence measure: **Averaged Sibson divergence** .
- 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.