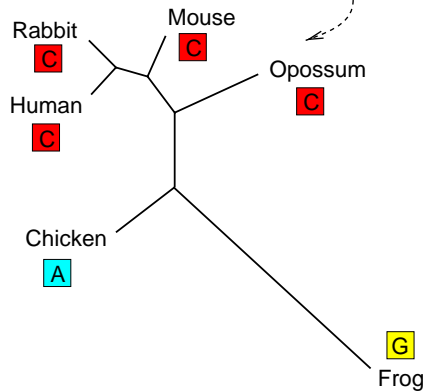

Approximate Bayesian Discrimination between Alternative Mosaic Structures of DNA Sequence Alignments

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>

To appear in **Bioinformatics**

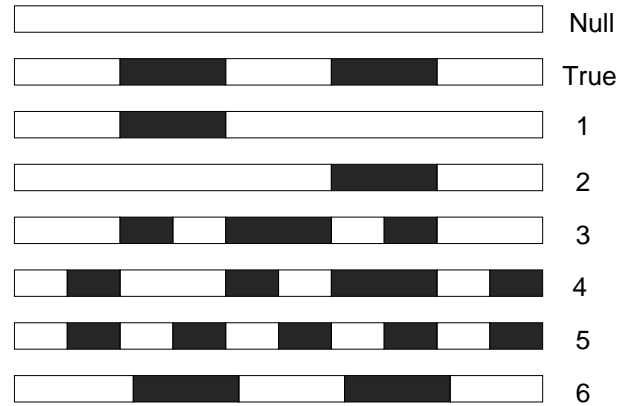
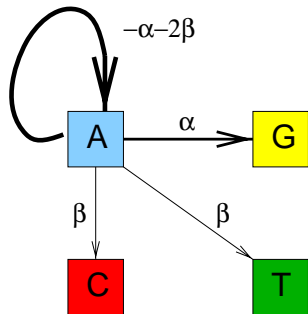
Introduction and Notation

	G	C	T	G	A	C	T	T	C	T	G	A	G	G	T	T	
Frog	G	C	G	T	A	A	C	T	T	C	A	C	A	T	G	A	T
Chicken	G	C	G	T	C	A	C	T	T	G	A	G	A	C	G	C	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



D Data = DNA Sequence Alignment

H Hypothesis = Mosaic Structure

ψ Tree Topology

w Branch Lengths

θ Nucleotide Substitution Parameters

$q = (w \ \psi \ \theta)$ Model Parameters

Bayesian Evidence: Background

$$P(\mathcal{D}|H) = \int P(\mathcal{D}|\mathbf{q}, H)P(\mathbf{q})d\mathbf{q}$$

$$P(\mathcal{D}|H) = \frac{1}{T} \sum_{t=1}^T P(\mathcal{D}|\mathbf{q}_t, H)$$

$$P(\mathcal{D}|\mathbf{q}, H)P(\mathbf{q}|H) = P(\mathbf{q}|\mathcal{D}, H)P(\mathcal{D}|H) \implies \frac{1}{P(\mathcal{D}|H)} = \int \frac{P(\mathbf{q}|\mathcal{D}, H)}{P(\mathcal{D}|\mathbf{q}, H)P(\mathbf{q}|H)}d\mathbf{q}$$

$$P(\mathcal{D}|H) = \left[\frac{1}{T} \sum_{t=1}^T \frac{1}{P(\mathcal{D}|\mathbf{q}_t, H)P(\mathbf{q}_t|H)} \right]^{-1}$$

Both estimators are consistent, but **not viable in practice** (diverging variance).

Bayesian Evidence: Decomposition

$$P(\mathbf{q}|\mathcal{D}, H)P(\mathcal{D}|H) = P(\mathcal{D}|\mathbf{q}, H)P(\mathbf{q}|H)$$

$$\ln P(\mathcal{D}|H) = \ln P(\mathcal{D}|\mathbf{q}, H) + \ln P(\mathbf{q}|H) - \ln P(\mathbf{q}|\mathcal{D}, H)$$

$$U(H) = \int P(\mathbf{q}|\mathcal{D}, H) [\ln P(\mathcal{D}|\mathbf{q}, H) + \ln P(\mathbf{q}|H)] d\mathbf{q}$$

$$S(H) = - \int P(\mathbf{q}|\mathcal{D}, H) \ln P(\mathbf{q}|\mathcal{D}, H) d\mathbf{q}$$

$$\ln P(\mathcal{D}|H) = U(H) + S(H)$$

$$U(H) = \frac{1}{T} \sum_{t=1}^T [\ln P(\mathcal{D}|\mathbf{q}_t, H) + \ln P(\mathbf{q}_t|H)]$$

Problem: $S(H)$ is intractable (both analytically and numerically)

Parameter Replicas



\mathbf{q}_1

\mathbf{q}_2

\mathbf{q}_3

\mathbf{q}_4

\mathbf{q}_5

$$P(\mathbf{q}|\mathcal{D}) = P(\mathbf{q}_1, \dots, \mathbf{q}_K|\mathcal{D}) = \prod_{k=1}^K P(\mathbf{q}_k|\mathcal{D})$$

$$S = - \int P(\mathbf{q}|\mathcal{D}, H) \ln P(\mathbf{q}|\mathcal{D}, H) d\mathbf{q} = \sum_{k=1}^K S_k$$

$$S_k = - \int P(\mathbf{q}_k|\mathcal{D}, H) \ln P(\mathbf{q}_k|\mathcal{D}, H) d\mathbf{q}_k$$

Mean Field Approximation

$$P(\mathbf{q}_1, \dots, \mathbf{q}_K | \mathcal{D}, H) = \prod_{k=1}^K P(\mathbf{q}_k | \mathcal{D}, H)$$

$$P(\mathbf{q}_k | \mathcal{D}, H) = P(\psi_k | \mathcal{D}, H) P(\mathbf{w}_k | \mathcal{D}, H) P(\boldsymbol{\theta}_k | \mathcal{D}, H)$$

$$S(H) = \sum_{k=1}^K [S_{\psi_k}(H) + S_{\mathbf{w}_k}(H) + S_{\boldsymbol{\theta}_k}(H)]$$

$$S_{\psi_k}(H) = - \sum_{\psi_k} P(\psi_k | \mathcal{D}, H) \ln P(\psi_k | \mathcal{D}, H)$$

$$S_{\mathbf{w}_k}(H) = - \int P(\mathbf{w}_k | \mathcal{D}, H) \ln P(\mathbf{w}_k | \mathcal{D}, H) d\mathbf{w}_k$$

$$S_{\boldsymbol{\theta}_k}(H) = - \int P(\boldsymbol{\theta}_k | \mathcal{D}, H) \ln P(\boldsymbol{\theta}_k | \mathcal{D}, H) d\boldsymbol{\theta}_k$$

Laplace and BIC Approximations

$$P(\boldsymbol{\theta}_k | \mathcal{D}, H) \propto \exp \left[-\frac{1}{2} (\boldsymbol{\theta}_k - \hat{\boldsymbol{\theta}}_k)^\dagger \mathbf{C}_k^{-1} (\boldsymbol{\theta}_k - \hat{\boldsymbol{\theta}}_k) \right]$$
$$P(\mathbf{w}_k | \mathcal{D}, H) \propto \exp \left[-\frac{1}{2} (\mathbf{w}_k - \hat{\mathbf{w}}_k)^\dagger \mathbf{H}_k (\mathbf{w}_k - \hat{\mathbf{w}}_k) \right]$$

$$S_{\boldsymbol{\theta}_k}(H) = \frac{1}{2} \ln \det \mathbf{C}_k + c$$

$$S_{\mathbf{w}_k}(H) = -\frac{1}{2} \ln \det \mathbf{H}_k + c$$

$$= -\frac{1}{2} \sum_{i=1}^{\nu} \ln \varepsilon_k^i + c \quad \varepsilon_k^i \approx N_k \quad \forall i$$

$$\approx -\frac{\nu}{2} \ln N_k + c$$

$$\ln P(\mathcal{D} | H) \approx U(H) + \sum_{k=1}^K \left[S_{\psi_k}(H) + \frac{1}{2} \ln \det \mathbf{C}_k - \frac{\nu}{2} \ln N_k \right] + c$$

Comparison: BIC Approximation (Schwarz 1978)

$$AIC = 2\hat{L}_H - 2\nu$$

$$BIC = 2\hat{L}_H - \nu \ln N$$

$$P(\mathcal{D}|H) = \int P(\mathcal{D}|\mathbf{q}, H)P(\mathbf{q}|H)d\mathbf{q} \approx P(\mathcal{D}|\hat{\mathbf{q}}, H)P(\hat{\mathbf{q}}|H)\sqrt{\frac{(2\pi)^\nu}{\det \mathbf{H}}}$$

$$\mathbf{H} = -\nabla_{\mathbf{q}}\nabla_{\mathbf{q}}^\dagger[\ln P(\mathcal{D}|\mathbf{q}, H) + \ln P(\mathbf{q}|H)]_{\mathbf{q}=\hat{\mathbf{q}}}$$

$$\ln P(\mathcal{D}|H) = \ln P(\mathcal{D}|\hat{\mathbf{q}}, H) + \ln P(\hat{\mathbf{q}}|H) - \frac{1}{2} \ln \det \mathbf{H} + \frac{\nu}{2} \ln(2\pi)$$

$$\approx \hat{L}_H - \frac{1}{2} \ln \det \mathbf{H}$$

$$= \hat{L}_H - \frac{1}{2} \sum_{i=1}^{\nu} \ln \varepsilon_i, \quad \varepsilon_i \approx N \quad \forall i$$

$$\approx \hat{L}_H - \frac{\nu}{2} \ln N$$

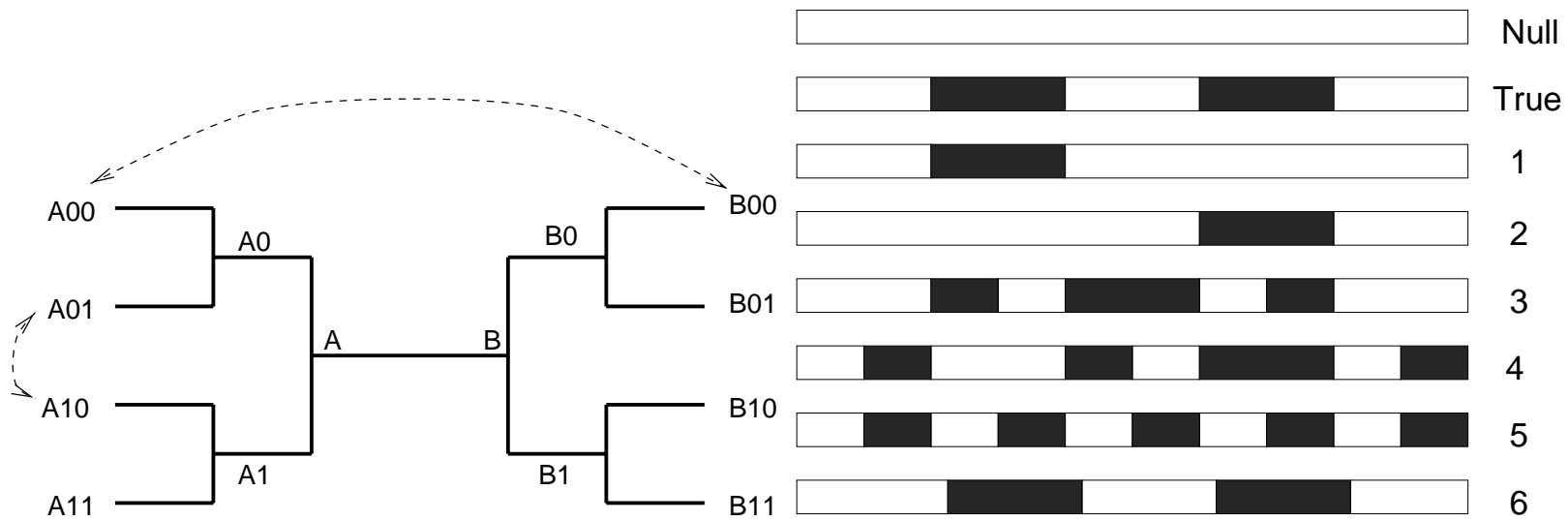
Improvement on BIC Approximation

$$\text{BIC} : \ln P(\mathcal{D}|H) = \hat{L}_H - \frac{\nu}{2} \ln N$$

$$\text{NEW} : \ln P(\mathcal{D}|H) = U(H) + \sum_{k=1}^K \left[S_{\psi_k}(H) + \frac{1}{2} \ln \det \mathbf{C}_k - \frac{\nu}{2} \ln N_k \right] + c$$

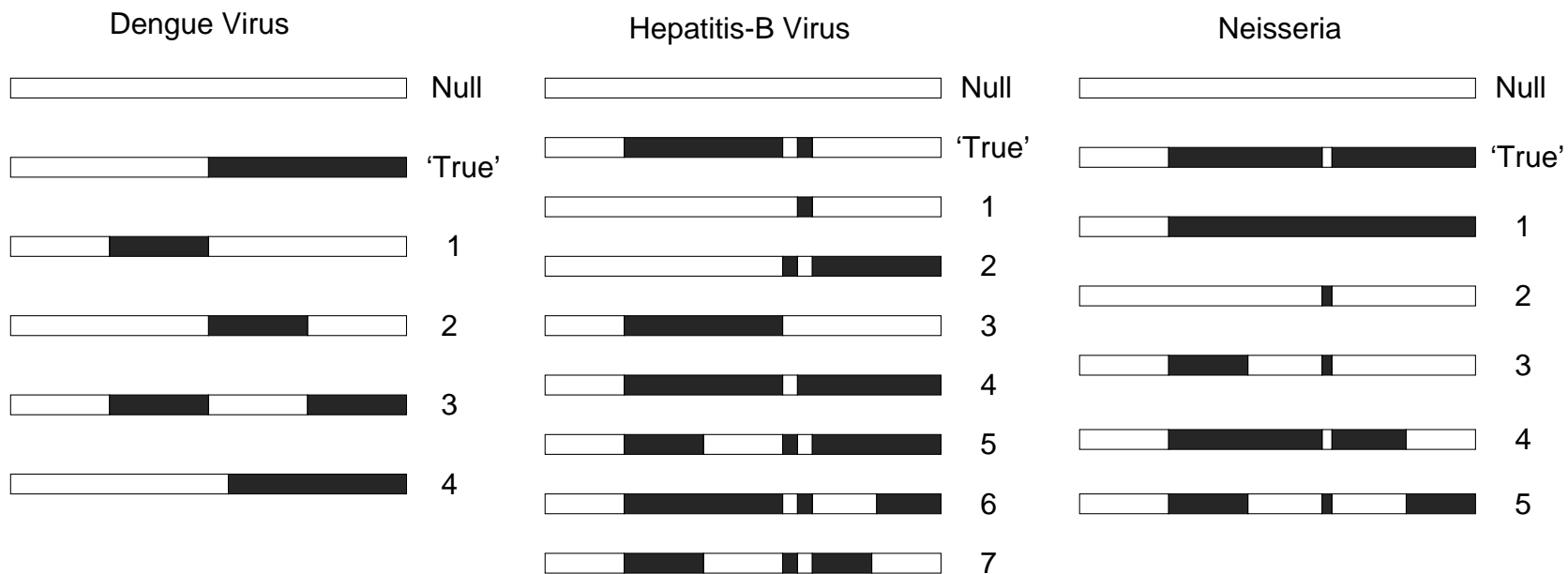
- BIC better than AIC (Byron J. Y. Morgan)
- NEW better than BIC
 - MCMC to compute U rather than ML to obtain \hat{L}_H .
 - Constraints on the eigenvalues only partially imposed.

Synthetic Problem



Total length of the alignment: 5000 nucleotides

Real-World DNA Sequence Alignments



Dengue virus: 7 taxa, 2295 nucleotides
Hepatitis B virus: 10 taxa, 3200 nucleotides
Neisseria: 8 taxa, 787 nucleotides

$$U(H) - U(H_0)$$

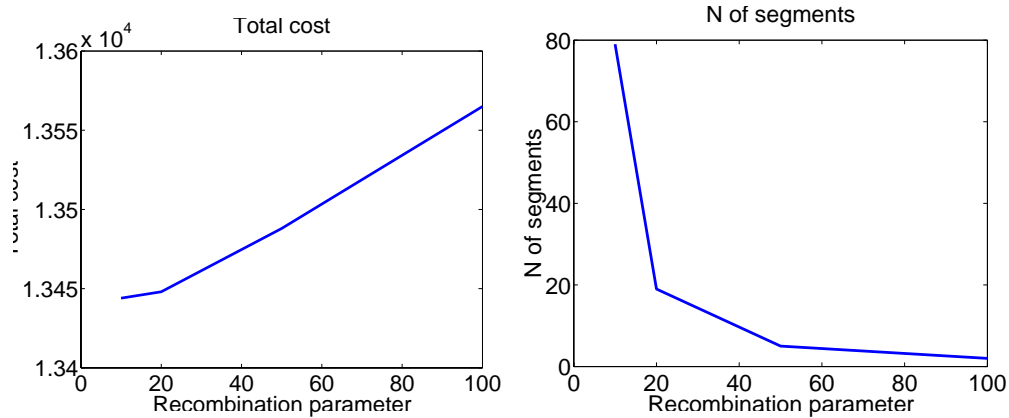
Data	True	1	2	3	4	5	6	7
Synthetic, $w = 0.1$	1506	1057	500	1500	1503	1496	1453	
	1245	1055	259	1242	1247	1244	1212	
Synthetic, $w = 0.075$	1296	907	425	1302	1295	1301	1271	
	1157	974	260	1156	1162	1162	1134	
Synthetic, $w = 0.05$	1135	833	349	1128	1138	1130	1102	
	1040	888	230	1046	1039	1045	1026	
Synthetic, $w = 0.025$	796	596	255	799	796	799	793	
	658	578	125	661	653	657	651	
Synthetic, $w = 0.01$	341	259	87	338	345	343	327	
	350	300	73	351	349	350	349	
Neisseria	141.0	120.2	72.3	141.3	149.6	149.9		
	141.7	119.9	73.0	142.0	150.9	151.2		
Hepatitis-B virus	369	167	288	168	293	396	383	410
	371	166	290	168	294	399	385	413
Dengue virus	83.4	91.3	88.4	96.2	71.6			
	84.2	92.5	89.3	97.6	68.6			

Evidence: $\ln P(\mathcal{D}|H) - \ln P(\mathcal{D}|H_0)$






Data	Polymorph	True	1	2	3	4	5	6	7
Synthetic, $w = 0.1$	75%	1323	962	407	1236	1206	1119	1271	
		1058	959	166	979	946	868	1029	
Synthetic, $w = 0.075$	65%	1119	814	333	1048	1009	938	1093	
		979	883	168	903	872	796	955	
Synthetic, $w = 0.05$	51%	965	745	262	887	858	780	933	
		865	798	140	802	758	695	852	
Synthetic, $w = 0.025$	30%	641	517	175	583	550	492	639	
		500	496	42	443	404	347	495	
Synthetic, $w = 0.01$	13%	209	191	18	160	138	89	192	
		215	233	5	173	148	106	221	
Neisseria	30%	81.9	70.4	36.0	65.5	75.2	58.9		
		84.0	70.1	38.2	67.6	78.4	62.0		
Hepatitis-B virus	22%	247	104	198	94	192	242	232	227
		249	102	200	94	193	245	234	230
Dengue virus	12%	55.1	40.5	36.7	22.1	43.6			
		56.1	41.8	37.3	23.0	40.4			

RecPars (Hein) applied to the Synthetic Data

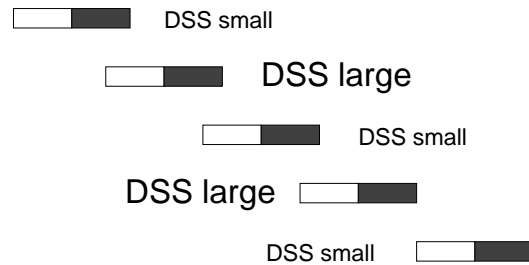
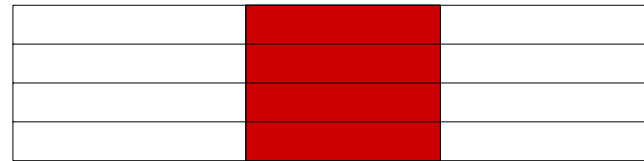
- Find most parsimonious history.
- Parameters: transition cost, transversion cost, recombination cost.
- Prior selection required, no optimisation possible.



RecPars

		Relative Evidence
	Null	0
	True	500
	RecPars 100	485
	RecPars 50	488
	RecPars 10	457

TOPAL (McGuire, Wright) applied to Neisseria



$$SoS_1 = \sum_i \sum_k (d_{ik} - \hat{d}_{ik})^2 \quad SoS_2 = \sum_i \sum_k \left(\frac{d_{ik} - \hat{d}_{ik}}{d_{ik}} \right)^2$$

TOPAL

		Relative Evidence
	Null	0
	'True'	81.9
	'Spurious'	51.7

Conclusions

- Bayesian model selection: Bayes factor $P(\mathcal{D}|H_\alpha)/P(\mathcal{D}|H_\beta)$
- Requires marginalisation: $P(\mathcal{D}|H) = \int P(\mathcal{D}|\mathbf{q}, H)P(\mathbf{q}|H)d\mathbf{q}$
- Analytically and numerically intractable.
- Decomposition into energy and entropy terms.
- Approximation for the entropy: partial mean field , Laplace method , BIC .
- Test data: 10 synthetic alignments, 3 real-world alignments.
- Approximate evidence almost consistently selects the true segmentation .