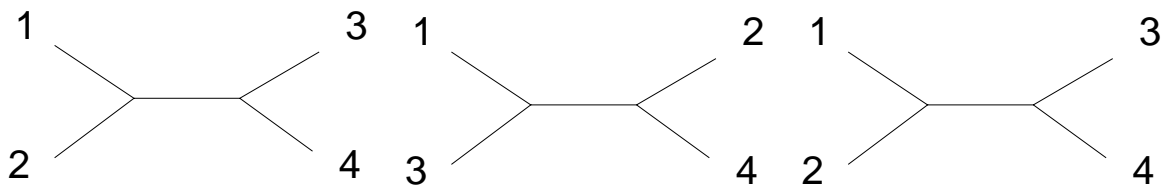
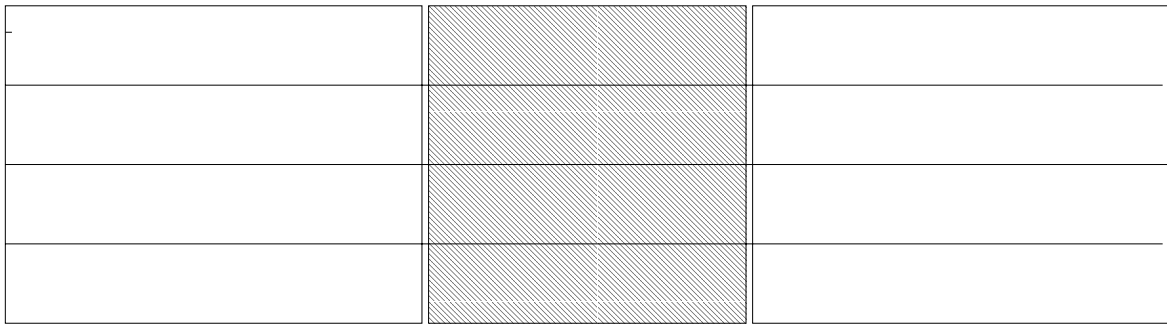
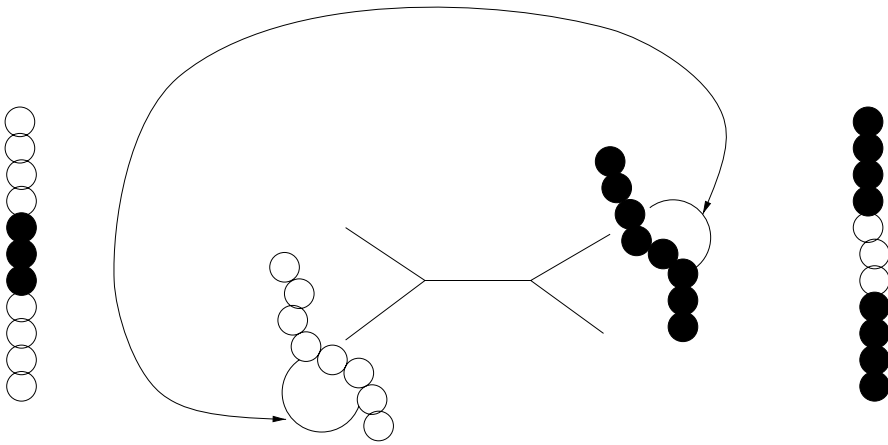

Detection of Recombination in Phylogenetic Data Sets with Hidden Markov Models

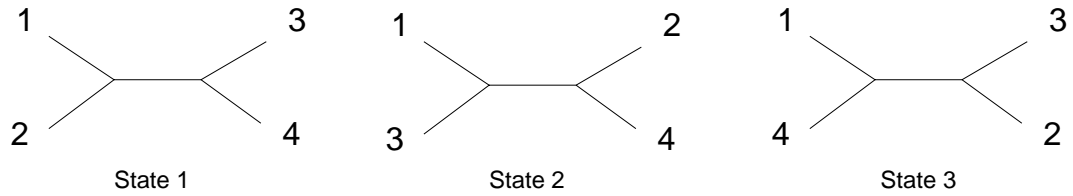
Dirk Husmeier and Frank Wright
Biomathematics and Statistics Scotland (BioSS)
SCRI, Invergowrie, Dundee DD2 5DA, UK
Email: dirk@bio.sari.ac.uk
Web: <http://www.bio.sari.ac.uk/~dirk>

- Phylogeny
- Recombination
- Modelling Recombination with HMMs
- Maximum Likelihood (EM algorithm)

Recombination



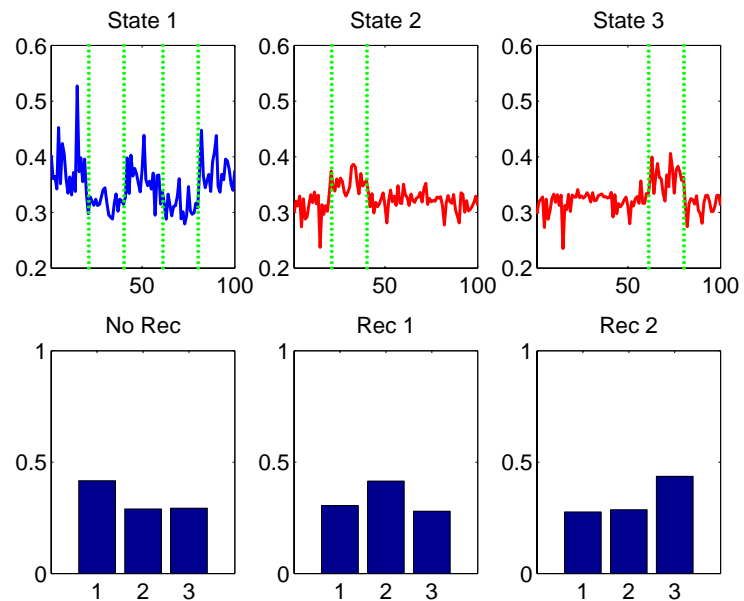
Naive approach



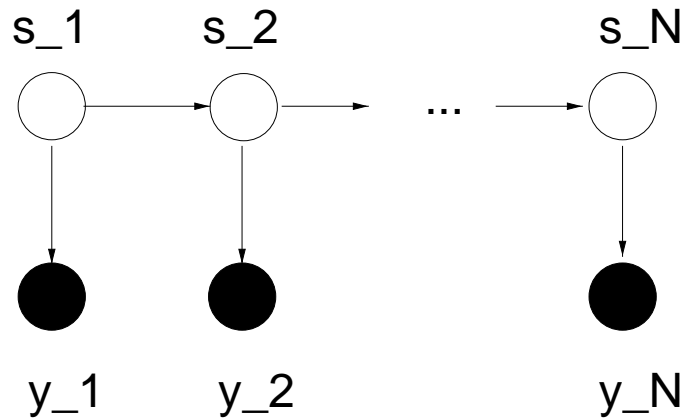
$$P(s_t | \mathbf{y}_t) = \frac{P(\mathbf{y}_t | s_t) P(s_t)}{P(\mathbf{y}_t)}$$

$$P(s_t) = \text{Const}$$

$$P(s_t | \mathbf{y}_t) = \frac{P(\mathbf{y}_t | s_t)}{\sum_{s'_t} P(\mathbf{y}_t | s'_t)}$$

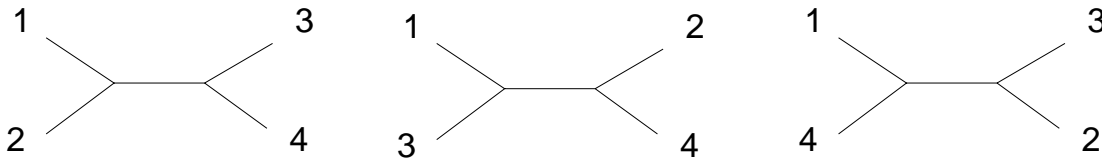


Modelling Recombination with HMMs I



$$P(\mathbf{y}_1, \dots, \mathbf{y}_N, s_1, \dots, s_N) = P(s_1) \prod_{t=2}^N P(s_t | s_{t-1}) \prod_{t=1}^N P(\mathbf{y}_t | s_t)$$

States s_t :



Transition Probabilities: $P(s_t | s_{t-1}) = \nu \delta_{s_t, s_{t-1}} + (1 - \delta_{s_t, s_{t-1}}) \frac{1 - \nu}{K - 1}$

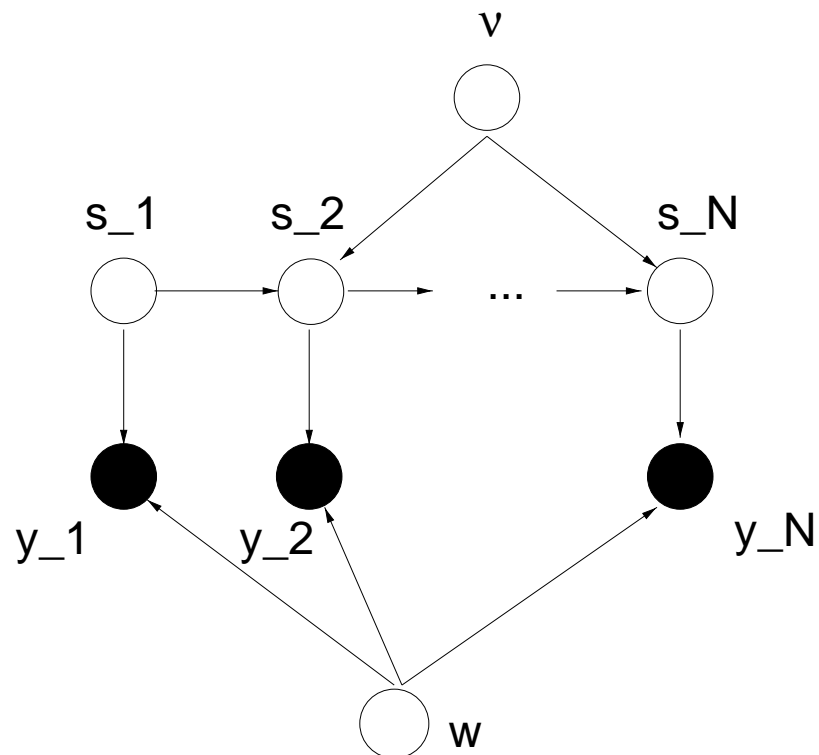
State Sequence: $\mathbf{s} = (s_1, s_2, \dots, s_N)$

Mode of $P(\mathbf{s} | \mathbf{D}) \longrightarrow$ **Viterbi Path**

Modelling Recombination with HMMs II

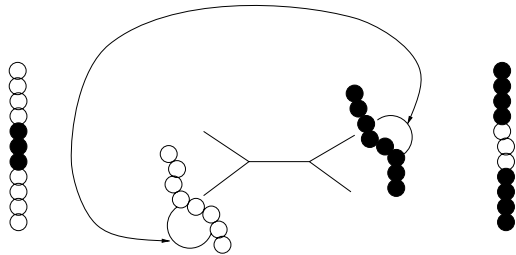
Transition Probabilities: $P(s_t | s_{t-1}, \nu)$

Emission Probabilities: $P(y_t | s_t, \mathbf{w})$

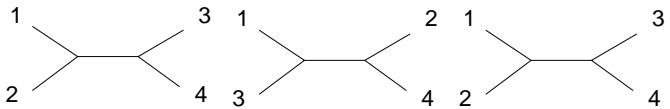


Maximum Likelihood: $\operatorname{argmax} P(\mathbf{D} | \mathbf{w}, \nu)$

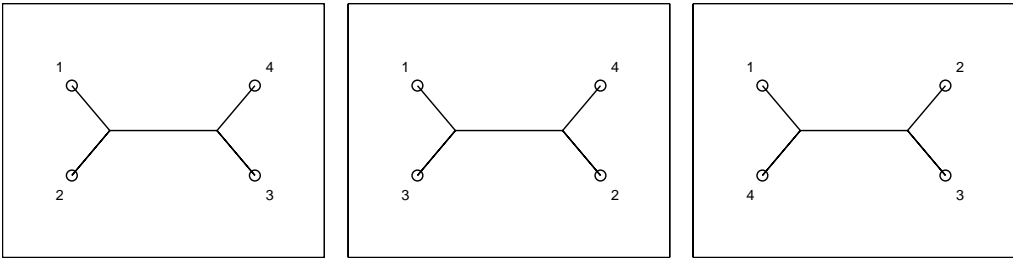
Wrong: Maximum Likelihood for Each Tree Separately



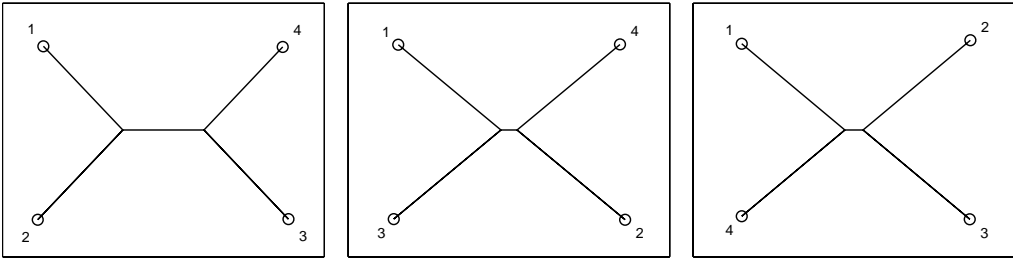
| | | |
|--|--|--|
| | | |
| | | |
| | | |
| | | |



True trees



Predicted trees



Joint Maximum Likelihood with the EM Algorithm

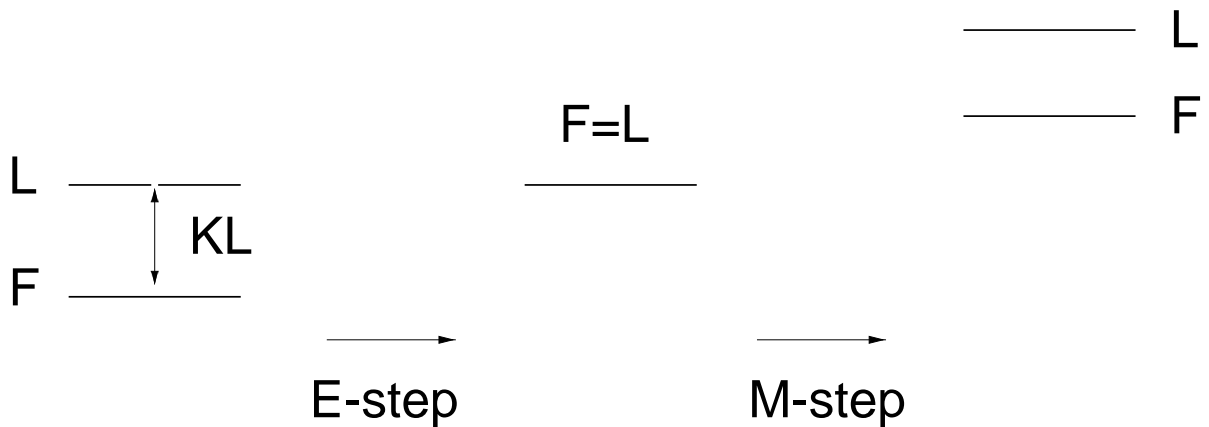
$$\begin{aligned}
 L(\mathbf{w}, \nu) &= \ln P(\mathbf{D}|\mathbf{w}, \nu) = \ln \sum_{\mathbf{s}} P(\mathbf{D}, \mathbf{s}|\mathbf{w}, \nu) \\
 &= \ln \sum_{\mathbf{s}} \frac{P(\mathbf{D}, \mathbf{s}|\mathbf{w}, \nu)}{Q(\mathbf{s})} Q(\mathbf{s}) \geq \sum_{\mathbf{s}} Q(\mathbf{s}) \ln \frac{P(\mathbf{D}, \mathbf{s}|\mathbf{w}, \nu)}{Q(\mathbf{s})}
 \end{aligned}$$

$$F(\mathbf{w}, \nu) = \sum_{\mathbf{s}} Q(\mathbf{s}) \ln \frac{P(\mathbf{D}, \mathbf{s}|\mathbf{w}, \nu)}{Q(\mathbf{s})} = \sum_{\mathbf{s}} Q(\mathbf{s}) \ln \frac{P(\mathbf{s}|\mathbf{D}, \mathbf{w}, \nu)}{Q(\mathbf{s})} + \ln P(\mathbf{D}|\mathbf{w}, \nu)$$

$$L(\mathbf{w}, \nu) = F(\mathbf{w}, \nu) + KL[Q, P]$$

E-step $\longrightarrow Q(\mathbf{s}) = P(\mathbf{s}|\mathbf{D}, \mathbf{w}, \nu)$

M-step \longrightarrow Maximise $F(\mathbf{w}, \nu)$



Application of the EM Algorithm to Recombination

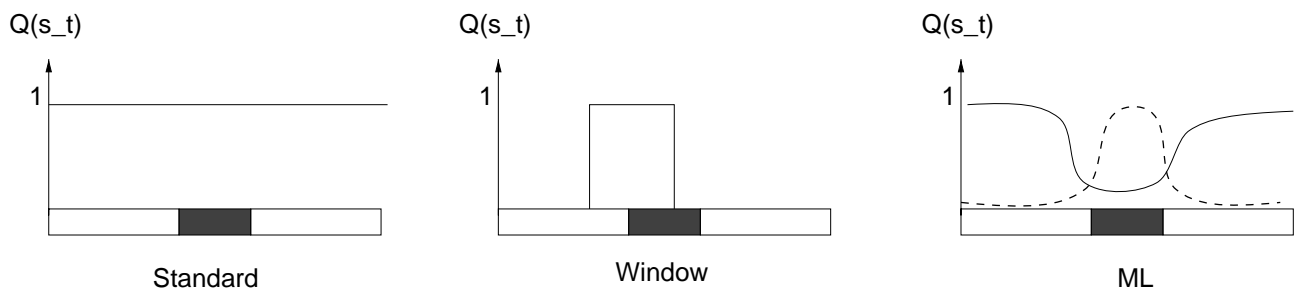
E-step: $Q(\mathbf{s}) = P(\mathbf{s}|\mathbf{D}, \mathbf{w}, \nu)$

→ Forward-backward algorithm for HMMs

M-step: Maximise $F(\mathbf{w}, \nu) = \sum_{\mathbf{s}} Q(\mathbf{s}) \ln P(\mathbf{D}, \mathbf{s}|\mathbf{w}, \nu)$

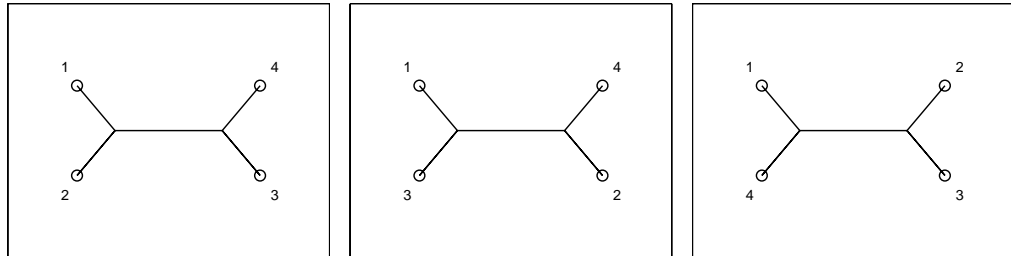
Recall:

$$P(\mathbf{D}, \mathbf{s}|\mathbf{w}, \nu) = P(s_1) \prod_{t=2}^N P(s_t|s_{t-1}, \nu) \prod_{t=1}^N P(y_t|s_t, \mathbf{w})$$
$$\Rightarrow F(\mathbf{w}, \nu) = \sum_{t=1}^N \sum_{s_t} Q(s_t) \ln P(y_t|s_t, \mathbf{w})$$

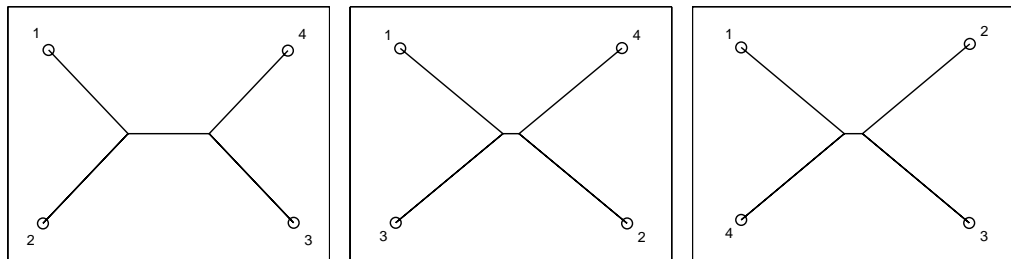


True and Predicted Trees

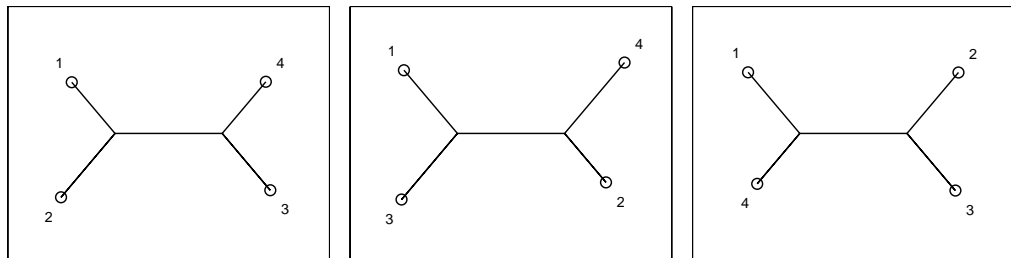
True trees



Maximum likelihood for each tree separately



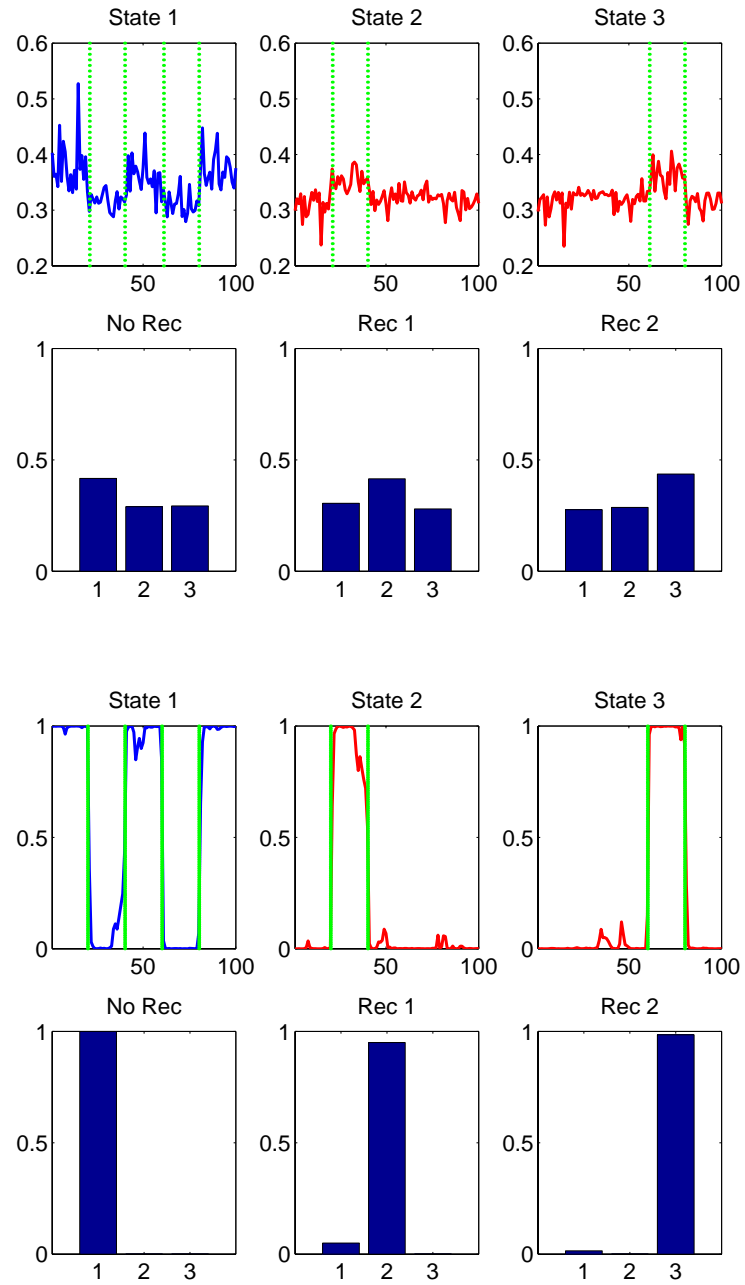
Joint maximum likelihood with EM



Root-mean-square deviation between the branch lengths:

| Training method | State 1 | State 2 | State 3 |
|-----------------|---------|---------|---------|
| Separate | 0.13 | 0.23 | 0.22 |
| EM | 0.04 | 0.06 | 0.04 |

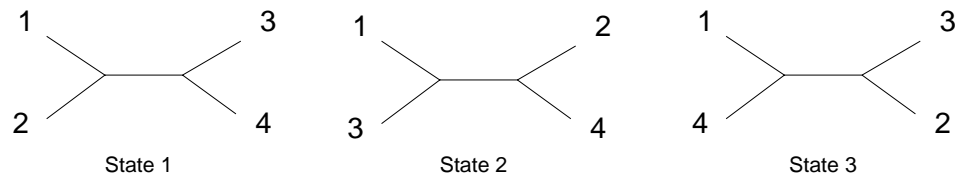
Comparison: Naive Approach versus HMM: $P(s_t|y_t)$



Real-world Data: Neisseria

- 1) Neisseria gonorrhoeae
- 2) Neisseria meningitidis

- 3) Neisseria cinerea
- 4) Neisseria mucosa

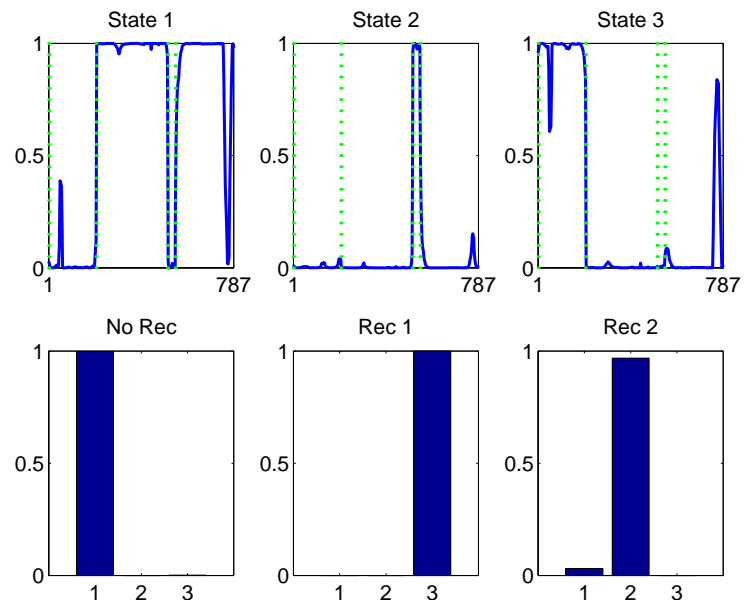


DNA alignment, 787 nucleotides (argF gene)

Dominant topology: State 1

Two anomalous (more diverged) regions:

[1-202] State 3 (Zhou, Spratt), [507-538] (unknown).



Neisseria: Two predictions

