Detecting mosaic structures in DNA sequence alignments

Dirk Husmeier
Biomathematics and Statistics Scotland
Edinburgh, United Kingdom
Email: dirk@bioss.ac.uk
http://www.bioss.ac.uk/~dirk
Recombination
Recombination
Recombination
Recombination
Recombination
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?

Recombination
Recombination
Recombination
Recombination in HIV 1

Recombinant strain
ZR-VI 191
Various recombination detection methods

- Window methods
- Phylo-HMMs
- Phylo-FHMMs
- Window methods
- Phylo-HMMs
- Phylo-FHMMs
TOPAL (McGuire & Wright, 1997)
TOPAL (McGuire & Wright, 1997)

Estimate
TOPAL (McGuire & Wright, 1997)

Estimate

Measure goodness of fit
TOPAL (McGuire & Wright, 1997)

Measure goodness of fit

Difference DSS

Estimate
TOPAL (McGuire & Wright, 1997)

small
TOPAL (McGuire & Wright, 1997)

small

large
Detect significant peaks of the DSS signal.
Significance determined with parametric bootstrapping.
Parametric bootstrapping

Q → Data → H₀ → Data₁ → ̂Q₁
   |                        |          |          |          |          |
   |  Data₂ → ̂Q₂          |          |          |          |          |
   |                        |          |          |          |          |
   |  Data₃ → ̂Q₃          |          |          |          |          |
   |  ...                 |          |          |          |          |
   |  Dataᵦ → ̂Qᵦ          |          |          |          |          |

Q ←
Do not reject Ho: No recombination
Reject Ho: Recombination significant
Example: TOPAL, window size=200
Example: TOPAL, window size=100
- Window methods
- Phylo-HMMs
- Phylo-FHMMs
Detecting recombination with HMMs

- Husmeier, Wright (2001)
  Journal of Computational Biology 8

  Bioinformatics 18

  Molecular Biology and Evolution 20
Related work

- Felsenstein & Churchill (1996)
  Molecular Biology and Evolution 13

  Journal of Computational Biology 11
Hidden Markov models (HMMs)

- Probabilistic equivalent to RecPars.
- All parameters can be inferred from the data.
- No window needed.
- More precise location of the breakpoints.
- Can currently only deal with a small number of species.
Naive approach

\[ P(S_t|y_t) = \frac{P(y_t|S_t)P(S_t)}{P(y_t)} \]

\[ P(S_t) = Const \]

\[ P(S_t|y_t) = \frac{P(y_t|S_t)}{\sum_{S'_t} P(y_t|S'_t)} \]
Naive approach

\[
P(S_t|y_t) = \frac{P(y_t|S_t)P(S_t)}{P(y_t)}
\]

\[
P(S_t) = \text{Const}
\]

\[
P(S_t|y_t) = \frac{P(y_t|S_t)}{\sum_{S'_t} P(y_t|S'_t)}
\]

Unrealistic prior on \( S = (S_1, \ldots, S_N) \)
Naive approach

\[
P(S_t|y_t) = \frac{P(y_t|S_t)P(S_t)}{P(y_t)}
\]

\[
P(S_t) = \text{Const}
\]

\[
P(S_t|y_t) = \frac{P(y_t|S_t)}{\sum_{S'_t} P(y_t|S'_t)}
\]

Unrealistic prior on \( S = (S_1, \ldots, S_N) \)

Introduce first-order spatial correlations via a Markov model

\[
P(S) = \prod_t P(S_t|S_{t-1})P(S_1)
\]
State 1

State 2

State 3

s_1 \rightarrow s_2 \rightarrow \ldots \rightarrow s_t \rightarrow s_{(t+1)} \rightarrow \ldots \rightarrow s_N

y_1 \rightarrow y_2 \rightarrow \ldots \rightarrow y_t \rightarrow y_{(t+1)} \rightarrow \ldots \rightarrow y_N

GCATCGTTCTATTTTACCGGCTCCCCGA

GTGTCGCTCAAGATTGCCCATCGCGCGT

GTCGTGGTCTAGATTGCCATCGCGCGT

GTATCGCTCTAGTTTGCCCAGCTCCCGT

GTATCGCTCTAGTTTGCCCAGCTCCCGT
\[
P(\mathcal{D}, S) = \prod_{t=1}^{N} P(y_t|S_t) \prod_{t=2}^{N} P(S_t|S_{t-1}) P(S_1)
\]
Emission probabilities (vertical arrows)

Strain 1
G C T T G A C T T T C T G A G G T T
Strain 2
G C G T A 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 T A A C T T G A G A C G C T

\[ y_t \]

\[ S_t = 1 \]

\[ \begin{array}{c}
G 1 \\
A 2 \\
C 3 \\
T 4 \\
\end{array} \]

\[ S_t = 2 \]

\[ \begin{array}{c}
G 1 \\
A 2 \\
C 3 \\
T 4 \\
\end{array} \]

\[ S_t = 3 \]

\[ \begin{array}{c}
G 1 \\
A 2 \\
T 4 \\
C 3 \\
\end{array} \]

\[ \rightarrow P(y_t | S_t, w) \]

Topology \( S_t \)
Branch lengths \( w \)
Transition probabilities (horizontal arrows)

\[ \nu \]

\[ \frac{1}{2} (1 - \nu) \]

\[ \frac{1}{2} (1 - \nu) \]
HMM parameters

- $s_1, s_2, \ldots, s_N$
- $y_1, y_2, \ldots, y_N$

- $w$: Vector of branch lengths of all the trees
- $\nu$: Probability of not changing the tree topology
\[ P(S|\mathcal{D}) = P(S_1, S_2, \ldots, S_N|\mathcal{D}) \]

\[ P(S_t|\mathcal{D}) = \sum_{S_1, \ldots, S_{t-1}, S_{t+1}, \ldots, S_N} P(S|\mathcal{D}) \]
\[ P(S|D) = P(S_1, S_2, \ldots, S_N|D) \]

\[ P(S_t|D) = \sum_{S_1, \ldots, S_{t-1}, S_{t+1}, \ldots, S_N} P(S|D) \]

Four strains of Hepatitis B virus
HMM parameters

$w$ \rightarrow \text{Vector of branch lengths of all the trees}

$\nu$ \rightarrow \text{Probability of \textit{not} changing the tree topology}
Bayesian approach

Husmeier, McGuire (2002)
Bioinformatics 18, S345-S353

Molecular Biology and Evolution 20, 315-337
Bayesian approach

\[ P(S|\mathcal{D}) = \int P(S|\mathcal{D}, w, \nu) P(w, \nu|\mathcal{D}) dw d\nu \]
Bayesian approach

\[ P(S|D) = \int P(S|D, w, \nu)P(w, \nu|D)dw d\nu \]

Posterior \( P(w, \nu|D) \) ← Prior \( P(w, \nu) = \prod_i P(w_i)P(\nu) \)
Bayesian approach

\[
P(S|\mathcal{D}) = \int P(S|\mathcal{D}, w, \nu) P(w, \nu|\mathcal{D}) dw d\nu
\]

Posterior \( P(w, \nu|\mathcal{D}) \) ← Prior \( P(w, \nu) = \prod_i P(w_i) P(\nu) \)

\[
P(w_i) = \begin{cases} 
1/\Omega & \text{if } 0 \leq w_i \leq \Omega \\
0 & \text{otherwise}
\end{cases}
\]
Bayesian approach

\[ P(S|\mathcal{D}) = \int P(S|\mathcal{D}, w, \nu)P(w, \nu|\mathcal{D})dw d\nu \]

Posterior \( P(w, \nu|\mathcal{D}) \) ← Prior \( P(w, \nu) = \prod_i P(w_i)P(\nu) \)

\[
P(w_i) = \begin{cases} 
1/\Omega & \text{if } 0 \leq w_i \leq \Omega \\
0 & \text{otherwise}
\end{cases}
\]

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

Conjugate prior: Beta distribution.
Beta Prior, $\beta = 2, \mu = \alpha / (\alpha + \beta)$
Sampling from the posterior distribution

- **Sampling** from $P(S, w, \nu|\mathcal{D})$
Sampling from the posterior distribution

- Sampling from $P(S, w, \nu | D)$
- Gibbs sampling
$P(x,y)$
$P(x,y)$
$P(x,y)$

$P(x|y)$
$P(x,y)$
Sampling from the posterior distribution

- **Sampling** from $P(S, w, \nu | D)$
- **Gibbs-like** approach:
Sampling from the posterior distribution

- **Sampling** from $P(S, w, \nu|\mathcal{D})$
- **Gibbs-like** approach:
  - $S \sim P(S|w, \nu, \mathcal{D})$
Sampling from the posterior distribution

- **Sampling** from $P(S, w, \nu|D)$
- **Gibbs-like** approach:
  - $S \sim P(S|w, \nu, D)$
  - $w \sim P(w|S, \nu, D)$
Sampling from the posterior distribution

- **Sampling** from $P(S, w, \nu | D)$
- **Gibbs-like** approach:
  - $S \sim P(S| w, \nu, D)$
  - $w \sim P(w| S, \nu, D)$
  - $\nu \sim P(\nu| S, w, D)$
Sampling from the posterior distribution

- **Sampling** from $P(S, w, \nu|\mathcal{D})$
- **Gibbs-like** approach:
  - $S \sim P(S|w, \nu, \mathcal{D})$
  - $w \sim P(w|S, \nu, \mathcal{D})$
  - $\nu \sim P(\nu|S, w, \mathcal{D})$
- $\nu$: Sample from Beta distribution
Sampling from the posterior distribution

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

- $\nu$: Sample from Beta distribution
- $w$: Metropolis-Hastings
Sampling from the posterior distribution

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

- **Gibbs-like** approach:
  - $S \sim P(S | w, \nu, \mathcal{D})$
  - $w \sim P(w | S, \nu, \mathcal{D})$
  - $\nu \sim P(\nu | S, w, \mathcal{D})$

- $\nu$: Sample from Beta distribution

- $w$: Metropolis-Hastings

- $S$: Gibbs sampling
  $S_t \sim P(S_t | S_1, \ldots, S_{t-1}, S_{t+1}, \ldots, S_N, \mathcal{D}, w, \nu)$
Sampling from the posterior distribution

\[ P(S_t | S_1, \ldots, S_{t-1}, S_{t+1}, \ldots, S_N, D, w, \nu) \]
Sampling from the posterior distribution

\[ P(S_t|S_1, \ldots, S_{t-1}, S_{t+1}, \ldots, S_N, \mathcal{D}, w, \nu) = P(S_t|S_{t-1}, S_{t+1}, y_t, w, \nu) \]
Sampling from the posterior distribution

\[ P(S_t | S_1, \ldots, S_{t-1}, S_{t+1}, \ldots, S_N, D, w, \nu) = P(S_t | S_{t-1}, S_{t+1}, y_t, w, \nu) \]
\[ \propto P(S_{t+1} | S_t, \nu) P(S_t | S_{t-1}, \nu) P(y_t | S_t, w) \]
Gibbs-within-Gibbs scheme

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

J. R. Statist. Soc. B, 62, 57-75

Molecular Biology and Evolution 20, 315-337
Gibbs-within-Gibbs scheme

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

J. R. Statist. Soc. B, 62, 57-75

Molecular Biology and Evolution 20, 315-337

Slow mixing and convergence
Gibbs-within-Gibbs scheme

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

J. R. Statist. Soc. B, 62, 57-75

Molecular Biology and Evolution 20, 315-337

Slow mixing and convergence

Applied Statistics 49, 269–285
Simultaneous sampling of the states from $P(S|w, \nu, D)$

\[
P(S_t|S_{t+1}, \ldots, S_N, y_1, \ldots, y_N) \\
\propto P(S_t, S_{t+1}, \ldots, S_N, y_1, \ldots, y_N) \\
= P(y_{t+1}, \ldots, y_N, S_{t+1}, \ldots, S_N|S_t, y_1, \ldots, y_t)P(S_t, y_1, \ldots, y_t) \\
= P(y_{t+1}, \ldots, y_N, S_{t+1}, \ldots, S_N|S_t)\alpha_t(S_t) \\
= P(y_{t+1}, \ldots, y_N, S_{t+2}, \ldots, S_N|S_{t+1})P(S_{t+1}|S_t)\alpha_t(S_t) \\
\propto P(S_{t+1}|S_t)\alpha_t(S_t)
\]
Stochastic forward–backward algorithm

- Run the **forward algorithm** to obtain
  \[ \alpha_t(S_t) = P(S_t, y_1, \ldots, y_t) \]

- Sample \( S_N \) from
  \[ P(S_N = k | y_1, \ldots, y_N) = \frac{\alpha_N(S_N=k)}{\sum_i \alpha_N(S_N=i)} \]

- Sample the remaining states \( S_{N-1}, \ldots, S_1 \) recursively from
  \[ P(S_t = k | S_{t+1}, \ldots, S_N, y_1, \ldots, y_N) = \frac{P(S_{t+1} | S_t=k) \alpha_t(S_t=k)}{\sum_i P(S_{t+1} | S_t=i) \alpha_t(S_t=i)} \]
Hepatitis B Virus (Bollyky et al. 1995)

DNA alignment, 3049 nucleotides

1) HPBADW1  2) HPBADW2  3) HPBADWZCG  4) HPBADRC
\( P(S_t|\mathcal{D}) \): Marginal posterior probability

DNA alignment, 3049 nucleotides

1) HPBADW1  
2) HPBADW2  
3) HPBADWZCG  
4) HPBADRC
Gibbs-within-Gibbs sampling

<table>
<thead>
<tr>
<th>Burn-in:</th>
<th>$10^6$ Gibbs steps</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sampling:</td>
<td>$10^6$ Gibbs steps</td>
</tr>
<tr>
<td>Computational costs:</td>
<td>$&gt; 6$ hours (on SUN Ultra-60)</td>
</tr>
</tbody>
</table>
Gibbs-within-Gibbs sampling

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Burn-in:</strong></td>
<td>$10^6$ Gibbs steps</td>
</tr>
<tr>
<td><strong>Sampling:</strong></td>
<td>$10^6$ Gibbs steps</td>
</tr>
<tr>
<td><strong>Computational costs:</strong></td>
<td>$&gt; 6$ hours (on SUN Ultra-60)</td>
</tr>
</tbody>
</table>

By how much can we reduce the computational costs with the modified forward-backward algorithm?
Gibbs-within-Gibbs sampling

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Burn-in:</strong></td>
<td>$10^6$ Gibbs steps</td>
</tr>
<tr>
<td><strong>Sampling:</strong></td>
<td>$10^6$ Gibbs steps</td>
</tr>
<tr>
<td><strong>Computational costs:</strong></td>
<td>&gt; 6 hours (on SUN Ultra-60)</td>
</tr>
</tbody>
</table>

By how much can we reduce the computational costs with the modified forward-backward algorithm?

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Burn-in:</strong></td>
<td>$10^4$ Gibbs steps</td>
</tr>
<tr>
<td><strong>Sampling:</strong></td>
<td>$10^4$ Gibbs steps</td>
</tr>
<tr>
<td><strong>Computational costs:</strong></td>
<td>a few minutes (on SUN Ultra-60)</td>
</tr>
</tbody>
</table>

Adriano Werhli
Comparison Gibbs-within-Gibbs versus stochastic forward-backward algorithm
$P(x, y)$

$P(y| x)$

$P(x, y)$
Comparison Gibbs-within-Gibbs versus stochastic forward-backward algorithm
Comparison Gibbs-within-Gibbs versus stochastic forward-backward algorithm
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
Simulation of recombination
Simulation of recombination
Simulation of recombination
Simulation of recombination
Simulation of recombination
Synthetic simulation study

400 600 800 200
True mosaic structure
$P(S_t|\mathcal{D})$: Marginal posterior probability
Trace plots of the log likelihood (left) and $\nu$ (right)
Phylo-HMM vs. naive method
Phylo-HMM vs. Topal, window size=200
Phylo-HMM vs. Topal, window size=100
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

---

- **State 1**: 875 bases
- **State 2**: 133 bases

Graphs showing the posterior probability distribution for States 1, 2, and 3 from 0 to 1000.
Beta Prior, $\beta = 2$, $\mu = \alpha / (\alpha + \beta)$
Dependence on the prior and the initialization
Neisseria (Zhou & Spratt, 1992)

DNA alignment, 787 nucleotides (argF gene)

1) Neisseria gonorrhoeae 3) Neisseria cinerea
2) Neisseria meningitidis 4) Neisseria mucosa
DNA alignment, 787 nucleotides (argF gene)

1) Neisseria gonorrhoeae  
2) Neisseria meningitidis  
3) Neisseria cinerea  
4) Neisseria mucosa
$P(S_t|\mathcal{D})$: Marginal posterior probability
\[ w = \alpha t \]
\[ \alpha \rightarrow r^- \alpha \]
\[ w \rightarrow r^- w \]
\[ 0 < r^- < 1 \]

\[ w = \alpha t \]

reference ("neutral") state
negative selective pressure

\[ \alpha \rightarrow r^- \alpha \]

\[ w \rightarrow r^- w \]

\[ 0 < r^- < 1 \]

reference ("neutral") state

\[ w = \alpha t \]

positive selective pressure

\[ \alpha \rightarrow r^+ \alpha \]

\[ w \rightarrow r^+ w \]

\[ r^+ > 1 \]
Problem:
Model cannot distinguish between recombination and rate variation.
Challenge

Distinguish between recombination and rate heterogeneity
- Window methods
- Phylo-HMMs
- Phylo-FHMMs
Distinguishing between recombination and rate variation with factorial hidden Markov models (FHMMs)
Factorial hidden Markov model (FHMM)
Phylo-FHMM

S_1 \rightarrow S_2 \rightarrow \ldots \rightarrow S_t \rightarrow S_{(t+1)} \rightarrow \ldots \rightarrow S_N

R_1 \rightarrow R_2 \rightarrow \ldots \rightarrow R_t \rightarrow R_{(t+1)} \rightarrow \ldots \rightarrow R_N

y_t
Rate states

$R = R^-$  
$w \rightarrow r^- w$
$0 < r^- < 1$

$R = R^0$  
reference ("neutral") state

$R = R^+$  
positive selective pressure
$w \rightarrow r^+ w$
$r^+ > 1$
Parameters

- Topology state sequences:
  \( S = (S_1, \ldots, S_N) \)

- Rate state sequences:
  \( R = (R_1, \ldots, R_N) \)

- Rate variation parameters:
  \( r = (r_1, \ldots, r_N) \)

- Branch lengths:
  \( w \)

- Transition probability parameters:
  \( \nu_S, \nu_R \)
Sampling from the posterior distribution

- **Sampling** from

\[ P(S, R, r, w, \nu_S, \nu_R | D) \]
Sampling from the posterior distribution

- **Sampling** from
  \[ P(S, R, r, w, \nu_S, \nu_R | D) \]
- **Gibbs sampling**
Sampling from the posterior distribution

- **Sampling from**
  \[ P(S, R, r, w, \nu_S, \nu_R | \mathcal{D}) \]

- **Gibbs sampling**
  \[- \quad S \sim P(S | R, r, w, \nu_S, \nu_R, \mathcal{D}) \]
Sampling from the posterior distribution

- **Sampling** from
  \[ P(S, R, r, w, \nu_S, \nu_R | D) \]
- **Gibbs sampling**
  - \( S \sim P(S | R, r, w, \nu_S, \nu_R, D) \)
  - \( R \sim P(R | S, r, w, \nu_S, \nu_R, D) \)
Sampling from the posterior distribution

- **Sampling** from
  \[ P(S, R, r, w, \nu_S, \nu_R | \mathcal{D}) \]

- **Gibbs sampling**
  - \( S \sim P(S | R, r, w, \nu_S, \nu_R, \mathcal{D}) \)
  - \( R \sim P(R | S, r, w, \nu_S, \nu_R, \mathcal{D}) \)
  - \( \ldots \)
Sampling from the posterior distribution

- **Sampling from**
  \[ P(S, R, r, w, \nu_S, \nu_R | \mathcal{D}) \]

- **Gibbs sampling**
  - \( S \sim P(S | R, r, w, \nu_S, \nu_R, \mathcal{D}) \)
  - \( R \sim P(R | S, r, w, \nu_S, \nu_R, \mathcal{D}) \)
  - \( \ldots \)

- **\( \nu_S, \nu_R \):** Sample from Beta distribution
Sampling from the posterior distribution

- **Sampling** from
  \[ P(S, R, r, w, \nu_S, \nu_R | \mathcal{D}) \]
- **Gibbs sampling**
  - \( S \sim P(S | R, r, w, \nu_S, \nu_R, \mathcal{D}) \)
  - \( R \sim P(R | S, r, w, \nu_S, \nu_R, \mathcal{D}) \)
  - \( \ldots \)
- \( \nu_S, \nu_R \): Sample from Beta distribution
- \( S, R \): Stochastic forward–backward algorithm
Sampling from the posterior distribution

- **Sampling from**
  \[ P(S, R, r, w, \nu_S, \nu_R | D) \]

- **Gibbs sampling**
  - \( S \sim P(S| R, r, w, \nu_S, \nu_R, D) \)
  - \( R \sim P(R| S, r, w, \nu_S, \nu_R, D) \)
  - \( \ldots \)

- \( \nu_S, \nu_R \): Sample from Beta distribution

- \( S, R \): Stochastic forward–backward algorithm

- \( w, r \): Metropolis-Hastings
Synthetic simulation study
Phylo-HMM: $P(S_t|\mathcal{D})$
FHMM

- 10 rate states
- Fixed rate factors
- Between $r = 0.001$ and $r = 100$ approximately uniform on a log scale
Phylo-FHMM. Left: $P(S_t|\mathcal{D})$. Right: $P(R_t|\mathcal{D})$
Neisseria (Zhou & Spratt, 1992)

DNA alignment, 787 nucleotides (argF gene)

1) Neisseria gonorrhoeae  3) Neisseria cinerea
2) Neisseria meningitidis  4) Neisseria mucosa
FHMM, Neisseria