Bayesian phylogenetic mapping of recombination hot-spots

Vladimir N. Minin

Department of Statistics
University of Washingont, Seattle

Bayesian Phylogeny, June 2008

Motivation

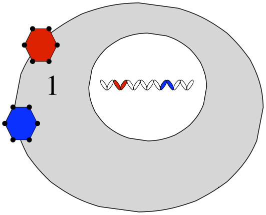
Reasons to Study HIV Recombination

- Allows for “discontinuous” jumps in evolution
- Has immediate medical applications (HIV drug resistance)
- Complicates phylogenetic reconstruction
- Not as rare as thought before (43 circulating recombinant forms (CRFs) in the Los Alamos HIV database)

What do we want to know about HIV recombination?

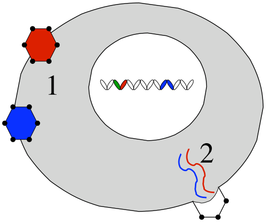
1. Spatial distribution of recombination break-points
2. Biochemical and selective forces that shape this distribution

Steps in HIV Recombination



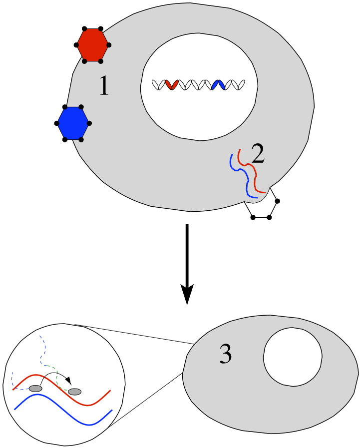
1. Co-infection of host cell by 2 distinct subtypes
2. Co-packaging of 2 distinct RNAs into a single virion
3. Strand jumping during reverse transcription
4. Release of recombinant virus

Steps in HIV Recombination



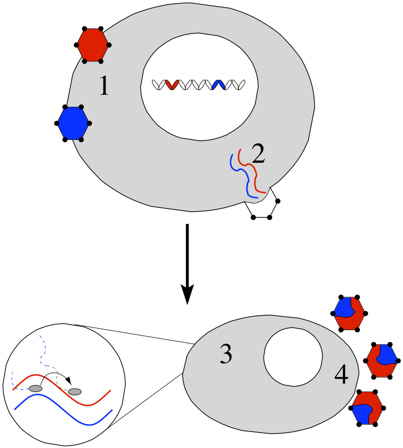
1. Co-infection of host cell by 2 distinct subtypes
2. Co-packaging of 2 distinct RNAs into a single virion
3. Strand jumping during reverse transcription
4. Release of recombinant virus

Steps in HIV Recombination



1. Co-infection of host cell by 2 distinct subtypes
2. Co-packaging of 2 distinct RNAs into a single virion
3. Strand jumping during reverse transcription
4. Release of recombinant virus

Steps in HIV Recombination



1. Co-infection of host cell by 2 distinct subtypes
2. Co-packaging of 2 distinct RNAs into a single virion
3. Strand jumping during reverse transcription
4. Release of recombinant virus

Evolutionary Histories with Recombination

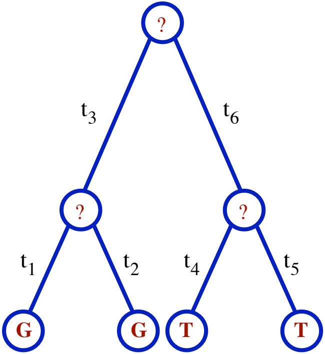
Example

- Hein's parsimony algorithm (Hein, 1990)
- Various sliding window approaches (Salminen et al., 1995, McGuire et al., 1997, Husmeier et al., 2001, ...)
- Hidden Markov models (Husmeier et al., 2003, 2005)
- Multiple change-point models (Suchard et al., 2002, 2003, Minin et al., 2005)

Dual Multiple Change-Point (DMCP) model, Part 1

organism 1	**G**	**C**	**T**	**A**	**A**	**...**
organism 2	**G**	**C**	**T**	**A**	**A**	**...**
organism 3	**T**	**G**	**T**	**T**	**A**	**...**
organism 4	**T**	**G**	**T**	**T**	**C**	**...**

~iid



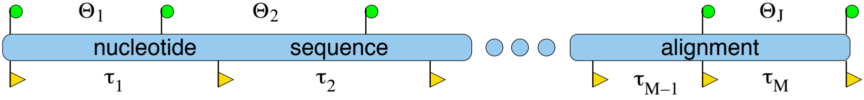
- $\mathbf{\Lambda} = \{\lambda_{ij}\}$ - substitution matrix
- $\mathbf{P}(t) = e^{t\mathbf{\Lambda}}$

HKY model:

$$\begin{pmatrix} - & \alpha\pi_G & \beta\pi_C & \beta\pi_T \\ \alpha\pi_A & - & \beta\pi_C & \beta\pi_T \\ \beta\pi_A & \beta\pi_G & - & \alpha\pi_T \\ \beta\pi_A & \beta\pi_G & \alpha\pi_C & - \end{pmatrix}$$

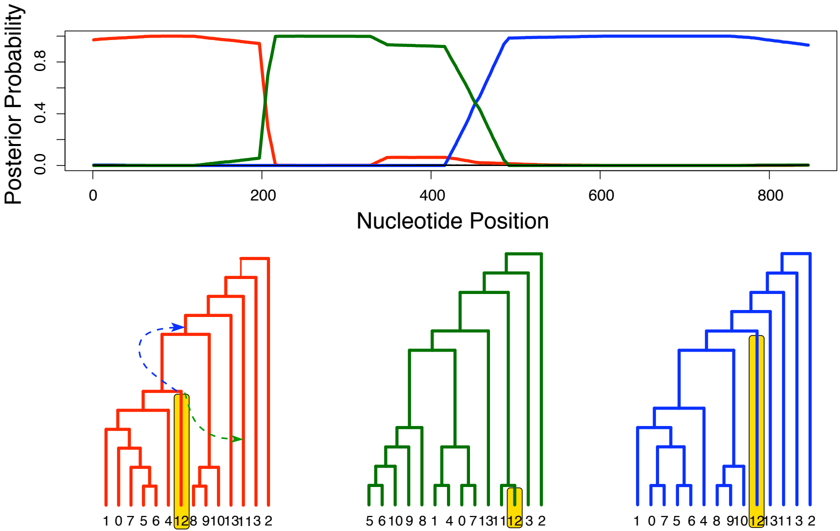
$$\pi = (\pi_A, \pi_G, \pi_C, \pi_T)$$

free parameter: $\kappa = \frac{\alpha}{\beta}$



- $\prod_s f(\mathbf{Y}_s | \tau(s), \Theta(s))$ - phylogenetic likelihood
- $1 = \xi_0 < \xi_1 < \cdots < \xi_M < \xi_{M+1} = S + 1, \forall\, s \in [\xi_{m-1}, \xi_m)$, $\tau(s) = \tau_m$, with $\tau_m \neq \tau_{m+1}$ - recombination break-points
- $1 = \rho_0 < \rho_1 < \cdots < \rho_J < \rho_{J+1} = S + 1, \forall\, s \in [\rho_{j-1}, \rho_j)$, $\Theta(s) \ = \Theta_j$ - substitution change-points
- Two change-point processes are independent

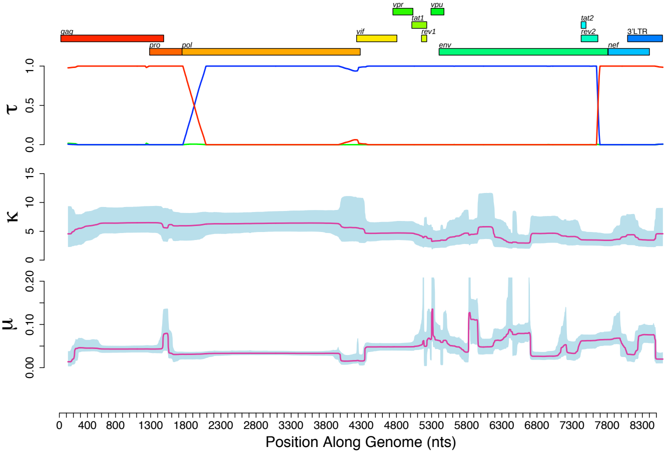
DMCP Analysis, Example 1



DMCP Analysis, Example 2 - HIV CRF

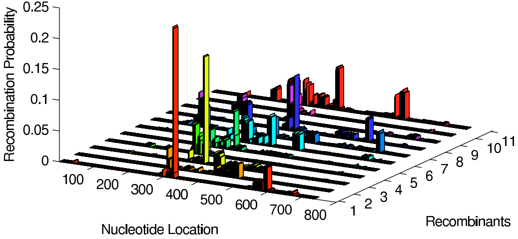
KAL153

(AB recombinant)



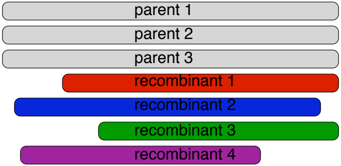
- Uncertainty in number and locations of break-points
- Variable dimensions ⇒ reversible jump MCMC

Multiple Recombinants



- Sparse data (# break-points ≪ sequence length)

- Unbalanced data (long indels)



Bayesian Hierarchical Model

Common Recombination Prior

- R_s - indicator of a recombination at site s, $p_s = \Pr(R_s = 1)$
- $\Pr(R_1 = r_1, \ldots, R_S = r_S) = \prod_{s=1}^{S} p_s^{r_s}(1 - p_s)^{1-r_s}$
- $M = \sum_{s=1}^{S} R_s \sim$ Poisson $\left(\sum_{s=1}^{S} p_s\right)$ (approximately)

Smoothing GMRF Hyper-Prior

- $\gamma_s = \ln\left(\frac{p_s}{1-p_s}\right)$ - recombination log-odds
- $\Pr(\gamma \mid \omega) \propto \omega^{(S-1)/2} \exp\left\{-\frac{\omega}{2}\sum_{s=1}^{S-1}(\gamma_s - \gamma_{s+1})^2\right\}$

Bayesian Hierarchical Model

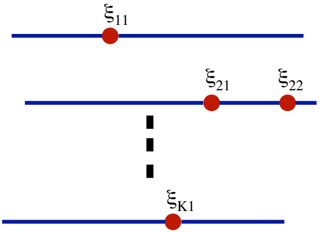
Common Recombination Prior

- R_s - indicator of a recombination at site s, $p_s = \Pr(R_s = 1)$
- $\Pr(R_1 = r_1, \ldots, R_S = r_S) = \prod_{s=1}^{S} p_s^{r_s}(1 - p_s)^{1-r_s}$
- $M = \sum_{s=1}^{S} R_s \sim$ Poisson $\left(\sum_{s=1}^{S} p_s\right)$ (approximately)

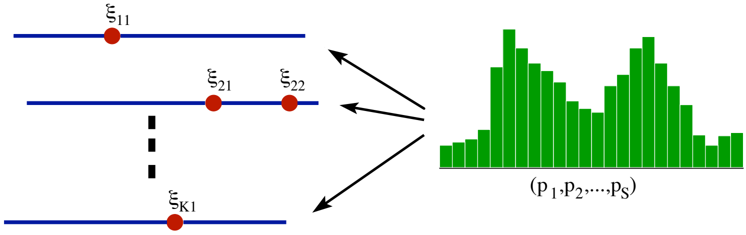
Smoothing GMRF Hyper-Prior

- $\gamma_s = \ln\left(\frac{p_s}{1-p_s}\right)$ - recombination log-odds
- $\Pr(\gamma \,|\, \omega) \propto \omega^{(S-1)/2} \exp\left\{-\frac{\omega}{2}\sum_{s=1}^{S-1}(\gamma_s - \gamma_{s+1})^2\right\}$

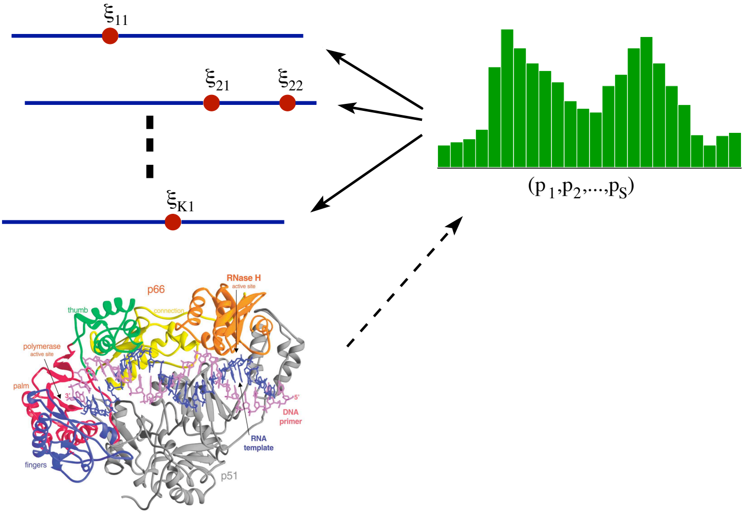
Graphical Model Representation



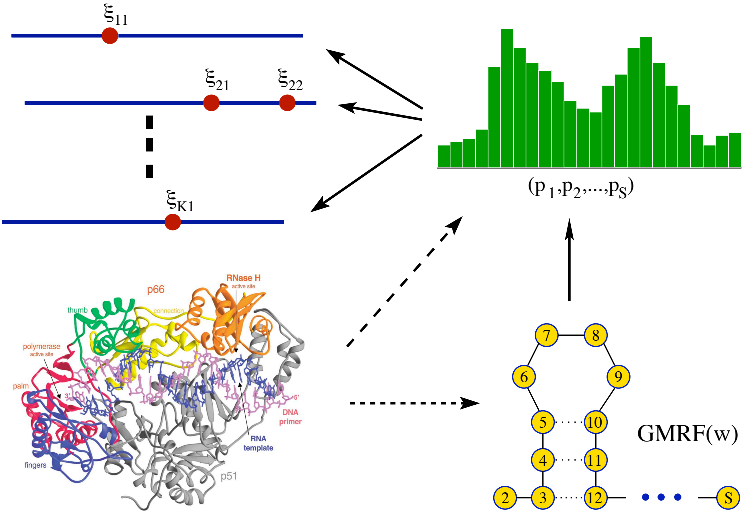
Graphical Model Representation



Graphical Model Representation



Graphical Model Representation



MCMC Algorithm

Metropolis-within-Gibbs with two major blocks

Updating individual-level parameters

- Condition on population-level recombination probabilities
- Use DMCP kernels with informative prior on recombination break-point locations

Updating population-level parameters

- Use fast GMRF sampling (Rue et al., 2001, 2004)
- Draw ω^* from an arbitrary univariate proposal distribution
- Use Gaussian approximation of $\Pr(\gamma \mid \omega^*, \mathbf{R})$ to propose γ^*
- Jointly accept/reject (ω^*, γ^*) in MH step

MCMC Algorithm

Metropolis-within-Gibbs with two major blocks

Updating individual-level parameters

- Condition on population-level recombination probabilities
- Use DMCP kernels with informative prior on recombination break-point locations

Updating population-level parameters

- Use fast GMRF sampling (Rue et al., 2001, 2004)
- Draw ω^* from an arbitrary univariate proposal distribution
- Use Gaussian approximation of $\Pr(\gamma \mid \omega^*, \mathbf{R})$ to propose γ^*
- Jointly accept/reject (ω^*, γ^*) in MH step

MCMC Algorithm

Metropolis-within-Gibbs with two major blocks

Updating individual-level parameters

- Condition on population-level recombination probabilities
- Use DMCP kernels with informative prior on recombination break-point locations

Updating population-level parameters

- Use fast GMRF sampling (Rue et al., 2001, 2004)
- Draw ω^* from an arbitrary univariate proposal distribution
- Use Gaussian approximation of $\Pr(\gamma \mid \omega^*, \mathbf{R})$ to propose γ^*
- Jointly accept/reject (ω^*, γ^*) in MH step

Implementing Constraints

Objective

$$\mathbf{Pr}(M > 0) = c \;\Rightarrow\; \sum_{s=1}^{S} p_s = -\ln(1-c).$$

Problem

Sum-of-probabilities constraint is non-linear in γ:

$$\sum_{s=1}^{S} e^{\gamma_s}/(1+e^{\gamma_s}) = -\ln(1-c)$$

Solution 1

Linearize constraint via Taylor expansion about arbitrary point $\mathbf{v}$. Sampling from GMRFs with linear constraints is easy (just re-centering). Choosing $\mathbf{v}$ is tricky, but feasible.

Solution 2

Renormalize the prior - not implemented yet!

$$p_s^* = \Pr(R_s = 1) = -\ln(1-c)p_s / \sum_{s=1}^{S} p_s$$

Implementing Constraints

Objective

$$\mathbf{Pr}(M > 0) = c \;\Rightarrow\; \sum_{s=1}^{S} p_s = -\ln(1-c).$$

Problem

Sum-of-probabilities constraint is non-linear in γ:

$$\sum_{s=1}^{S} e^{\gamma_s}/(1+e^{\gamma_s}) = -\ln(1-c)$$

Solution 1

Linearize constraint via Taylor expansion about arbitrary point $\mathbf{v}$. Sampling from GMRFs with linear constraints is easy (just re-centering). Choosing $\mathbf{v}$ is tricky, but feasible.

Solution 2

Renormalize the prior - not implemented yet!

$$p_s^* = \Pr(R_s = 1) = -\ln(1-c)p_s / \sum_{s=1}^{S} p_s$$

Implementing Constraints

Objective

$$\mathbf{Pr}(M > 0) = c \;\Rightarrow\; \sum_{s=1}^{S} p_s = -\ln(1-c).$$

Problem

Sum-of-probabilities constraint is non-linear in γ:

$$\sum_{s=1}^{S} e^{\gamma_s}/(1+e^{\gamma_s}) = -\ln(1-c)$$

Solution 1

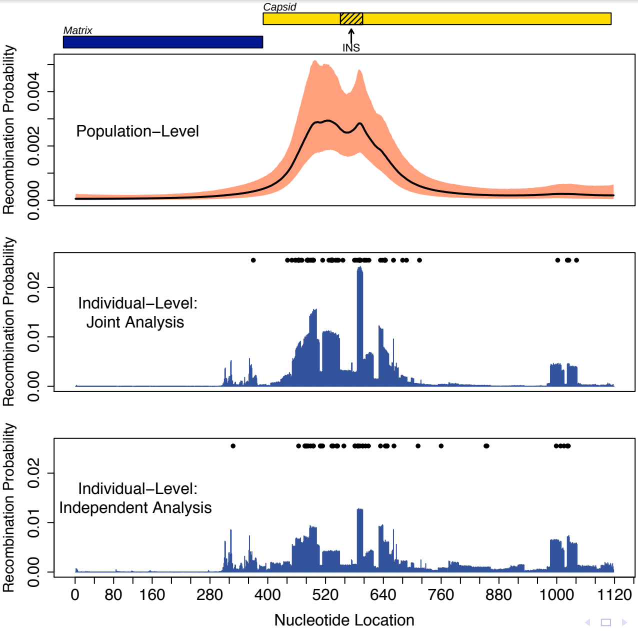
Linearize constraint via Taylor expansion about arbitrary point $\mathbf{v}$. Sampling from GMRFs with linear constraints is easy (just re-centering). Choosing $\mathbf{v}$ is tricky, but feasible.

Solution 2

Renormalize the prior - not implemented yet!

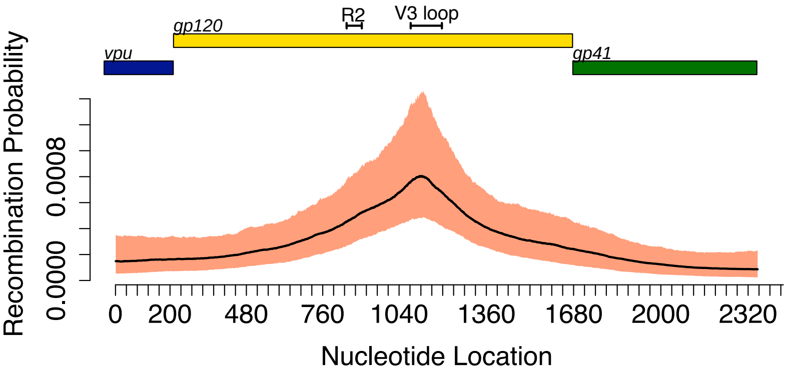
$$p_s^* = \Pr(R_s = 1) = -\ln(1-c)p_s / \sum_{s=1}^{S} p_s$$

Gag Recombinants



- 42 *gag* recombinants
- From 6 epidemiological studies
- All originated from A and G subtypes
- Length ranges from 562 to 820 nts (very unbalanced)
- INS-instability element, regulates expression

Env Recombinants



- 53 *env* conservatively selected recombinants
- Not controlled for subtype composition
- R2 is experimentally determined hot-spot (Galetto et al., 2004, 2006)

Acknowledgments

Collaborators

- Marc Suchard, UCLA
- Karin Dorman, Iowa State
- Fang Fang, Iowa State

Financial Support

- NIH grant R01 GM068955
- James B. Pendleton Charitable Trust
- UCLA AIDS Institute and Graduate Division