Dual Multiple Change Point Model Leads to More Accurate Recombination Detection

Vladimir N. Minin, [1] **Karin S. Dorman,** [2] **Marc A. Suchard** [3]

Keywords: Recombination, HIV evolution, MCMC, phylogenetics

Recombination is an important force in the evolution of HIV, promoting the developing of multiple drug resistant strains and deterring the construction of therapeutic vaccines. Consequently, research interests expand in recombination detection methods using multiple sequence alignments that accurately identify the recombination sites. We extend a computational method to detect homologous recombination [3] to improve its recombination detection resolution. In the original model an alignment is partitioned into K segments, where K is an unknown parameter. Each partition $k \in 1, \ldots, K$ has a vector of phylogenetic parameters (Θ_k, τ_k) associated with it, where Θ_k is a vector of parameters describing the nucleotide substitution process and τ_k is a bifurcating tree topology describing evolutionary relationships between sequences. End points of the partitions ξ_k are called change-points. Recombination is inferred if there is at least one change-point ξ_k such that $\tau_{k-1} \neq \tau_k$. This model was successfully applied to test recombination hypotheses in HIV strains by Suchard et al. [2]. Modeling spatial variation of all parameters with a single change-point process results in prior correlation between sites where substitution parameters vary and sites where topologies change. This can lead to loss of accuracy of recombination site identification, when recombination occurs near the boundary of regions with varying evolutionary pressures. Here, we develop a dual Multiple Change Point (MCP) model that decouples substitution parameters change-points from topology break-points by introducing two *a priori* independent change-point processes to describe the variation. We use reversible jump MCMC sampling to approximate the posterior distribution of model parameters [3].

To demonstrate improved accuracy, we start with a previously used test example involving mtDNA sequences from 4 primates and generate a series of datasets with simulated recombination events near a site where evolutionary pressures change greatly. In Figure 1 we plot inferred most probable recombination sites against simulated recombination sites for the single and dual MCP models. The single MCP model shows strong attraction between inferred recombination sites and the evolutionary pressure change. The dual MCP model yields more accurate inference with small variation about the diagonal caused by randomly distributed informative sites. For better interpretation of this variation, we divided all informative sites into two classes: supportive or contradictory of the recombinant structure, denoted in Figure 1 as light and dark gray circles respectively. As expected, the greatest inaccuracies in detection occur when the simulated recombination event is located in an uninformative region, especially those bordered by contradictory sites.

We also apply the dual MCP model to the *gag* gene sequenced from HIV-1 isolate VI557 and 9 different subtype consensus sequences. VI557 is a reported recombinant, with ambiguous support [1]. Figure 2 depicts the marginal posterior probabilities of the different possible topologies for each site in the alignment. We see one region near the 5' end in the alignment with large uncertainty in the topology, but no change in the most probable topology. This suggests little support for recombination in this alignment.

[1]Department of Biomathematics, UCLA, Los Angeles, CA. E-mail: vminin@ucla.edu
[2]Department of Statistics, Iowa State University, Ames, IA. E-mail: kdorman@iastate.edu
[3]Department of Biomathematics, UCLA, Los Angeles, CA. E-mail: msuchard@ucla.edu

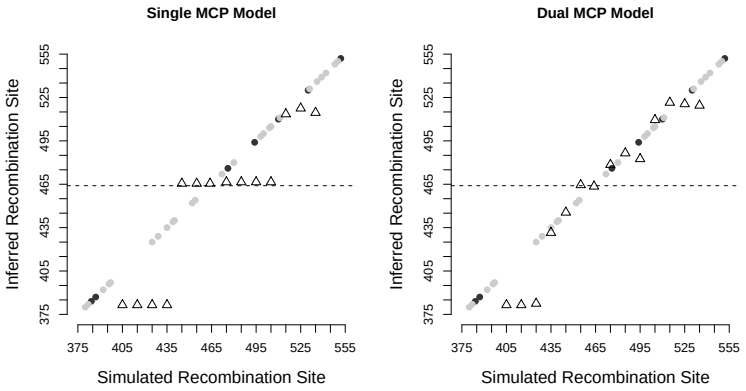

Figure 1: Simulation results. Inferred most probable recombination sites are plotted against simulated recombination sites near a substantial change in evolutionary pressures (fixed at site 464, dashed line). Circles on the diagonal denote informative sites that support (light gray) or contradict (dark gray) the recombinant structure.

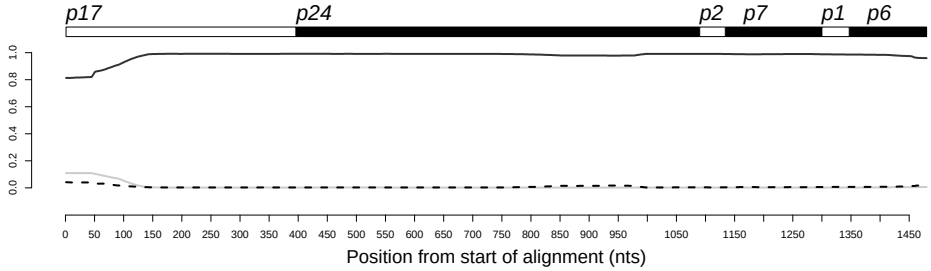

Figure 2: HIV results. Plot shows the locations of the gene products within the *gag* gene and the marginal posterior probabilities of the two most likely tree topologies (light gray, dark gray) and the sum of the marginal posterior probabilities of all other topologies for each site in the alignment(dashed line).

The dual MCP model inherits a major strength of the original MCP model in its realistic modeling of spatial phylogenetic variation using a parsimonious number of parameters. Using two change-point processes results in better sampling of topologies during MCMC simulations (not shown) and increases the accuracy of recombination site identification.

References

[1] Siepel, A. and Korber, B. 1995. Scanning the database for recombinant HIV-1 genomes. Pages III 35-60 in Human retroviruses and AIDS compendium.

[2] Suchard, M. A., Weiss, R. E., Dorman K. S. and Sinsheimer, J. S. 2002. Oh brother, where art thou? A Bayes factor test for recombination with uncertain heritage. *Systematic Biology* 51(5):715–728.

[3] Suchard, M. A., Weiss, R. E., Dorman K. S. and Sinsheimer, J. S. 2003. Inferring spatial phylogenetic variation along nucleotide sequences: A multiple change-point model. *Journal of the American Statistical Association* 98:427–437.