論文の内容の要旨

生産環境生物学専攻

平成２０年度博士課程入学

氏名：Leonardo de Oliveira Martins

指導教員名：岸野洋久

Bayesian Inference of Viral Recombination: Topology distance between DNA segments and its distribution

ウイルスゲノム組換えのベイズ推定：DNA部分配列の間のトポロジー距離とその分布

The phylogenetic inference is the problem of reconstructing the ancestrality betweeen a group of DNA or protein sequences, and is classically represented by a phylogenetic tree. These sequences may represent different species, or different genes from a same species (or both), and the underlying assumption is they share a common ancestral. To achieve consistency - the certainty that we approach the true phylogeny as more data becomes available - we would like to collect and analyze large genomic sequences. The complication is that besides the natural limitation of the genome sizes, organisms can exchange material between themselves, rendering the topological interpretation innacurate. One example of such an exchange is recombination.

In HIV-1, the reverse transcriptase switches RNA templates on average 3 times per replication cycle, yielding an average of about one recombinational strand transfer event per 3000 base pairs. A similar rate is also found in HIV-2 and murine leukemia viruses. Recombination also have been found to play a role in severe acute respiratory syndrome coronaviruses, hepatitis, enteroviruses and other primate lentiviruses. Recombinations lead to emergence of the resistant mutants to multiple drugs and may increase the chance that mutant-free individuals arise among the population of individuals with deleterious mutant genes. Reassortment is a similar type of genetic exchange in RNA viruses, where whole RNA molecules constituents of the segmented viral genome are swapped between individuals, and are responsible for antigenic shift in influenza A viruses.

In the case of HIV-1, it was observed that some sequences always clustered together, and this was used to classify HIV-1 in subtypes. As more data were collected, it became evident that disagreements from this classification appeared depending on the gene used to do the subtyping (inference of the subtype). This discordance was then attributed to recombination, and sequences with similar mosaic structure (region-dependent clustering) present in unrelated patients started being classified as Circulating Recombinant Forms (CRF). These recombinants are nowadays routinely detected by phylogenetic methods based on a local sequence similarity between the putative recombinant and all possible parentals. These so-called parentals are reference sequences from the original subtype classification.

Genomic regions involved in recombination may support distinct topologies, and phylogenetic analyses should incorporate this heterogeneity. If we have such a scenario of sporadic recombination, then phylogenetic methods to detect recombination can be employed. Recombination can therefore be detected by comparing inconsistency in topologies between adjacent segments, taking account of uncertainty in the phylogenetic inference. On the other hand, when recombination is more common than substitutions, this phylogenetic signal may be completely lost - thus every site would follow a distinct phylogenetic tree. In this cases we should give up the topological description and focus on populational parameters (like the recombination rate, population expansion, or divergence times).

So far inference of recombination under the phylogenetic approach has been restricted to the presence or absence of recombination break-points between sites, and detection of recombination hot-spots relied on unusual clustering patterns of these break-points along the genome. Many techniques of recombination detection are based on sliding window procedures that compare the topology of one segment against neighbouring segments or the whole alignment. These methods are sensitive to ancestral recombination events and moderate contribution of recombination. Variation in the selective pressure should be considered when estimating recombination events, since it may also lead to conflicting spatial phylogenetic signal. Bayesian change point models identify recombination breakpoints and differentiated substitution rates as change points of topologies and evolutionary rate parameters. Short segments may not have enough phylogenetic signal to discriminate between competing topologies, and large segments may miss the recombination breakpoints.

We developed a distance measure between unrooted topologies that closely resembles the number of recombinations. Despite the relation between a distance metric between topologies (called the Subtree Prune-and-Regraft distance, or SPR distance) and the amount of recombination is well known, there is still no definitive way of calculating it. To achieve that we needed to devise an approximation to this distance, which is a conservative estimate of the number of recombinations between two segments based on the distance between their inferred topologies. By introducing a prior distribution on these recombination distances, a Bayesian hierarchical model was devised to detect phylogenetic inconsistencies occurring due to recombinations. Our procedure assumes that recombination is moderate, and we focus on detectable changes in the phylogeny. An attractive argument in favor of Bayesian procedures is that instead of having a single point estimate of the parameter of interest, we have its distribution, posterior to observing the data. Other advantages include the possibility of exploiting

arbitrarily complex models and choosing the prior distributions to achieve a manageable level of abstraction. The disadvantage is the complexity of implementing the algortithm to draw samples from this posterior distribution. Since these samples should not be correlated, our algorithm creates the posterior samples by running heated chains serially and in parallel.

In our model the topological distance between segments (where one segment may one or a few sites) is modelled according to a modified Poisson distribution. By modelling the recombination distance between segments we penalize recombination scenarios where neighboring regions can only be explained by an excessive number of recombinations. This model relaxes the assumption of known parental sequences, still common in HIV analysis, allowing the entire dataset to be analyzed at once. We furthermore remove one possible source of noise from the phylogenetic inference which are the individual branch lengths (amount of evolution along the tree). This removal is achieved by averaging the topology over all possible branch lengths assuming they are independent realizations of an exponential distribution. This marginalization over individual branches and the assumption of indepence among segments should accomodate for rate heterogeneity among lineages and sites.

On simulated datasets with up to 16 taxa, our method correctly detected recombination breakpoints and the number of recombination events for each breakpoint. With this correlation between sites even a single break-point has information about the minimum number of recombinations between the segments it comprises. This not only has a biological support but also makes the topology sampling problem computationally tractable, since sampling from the topological space is not trivial for more than a few taxa.

Our Bayesian hierarchical procedure not only detects the recombination breakpoints but also quantifies the disagreement between the trees. It therefore provides information regarding regions where recombinations occur frequently. We also compared the results of our procedure with other Bayesian methods, providing them with the real recombination breakpoints. The chance of correctly inferring the true tree is also higher than using other Bayesian procedures that neglect the similarity between trees on neighboring regions. Our simulated datasets contained variability of substitution rates along the trees for each site and across sites, and assuming a model of independent rates for each site and averaging over individual branch lengths proved to be useful in distinguishing recombination from non-random rate heterogeneity.

Distinguishing one ancestral recombination (shared among many sequences) from a recombination hotspot (many recombinations rising independently) can be difficult. The robustness of our procedure comes from the fact that a breakpoint cannot be pinpointed with arbitrary precision, and the prior on the SPR distance accommodates this compromise. The amount of recombination over a region can, therefore, be quantified regardless of the number of breakpoints just by looking at the sum of over this region. In the Bayesian framework, once we obtain the posterior distribution of the variables of interest it is straightforward to have point estimates ("best" configuration), credibility intervals ("best" ensemble of configurations) and to test hypothesis (likeliness of a given configuration).

Applying our method to the HIV-1 dataset we detected a higher number of recombination break-

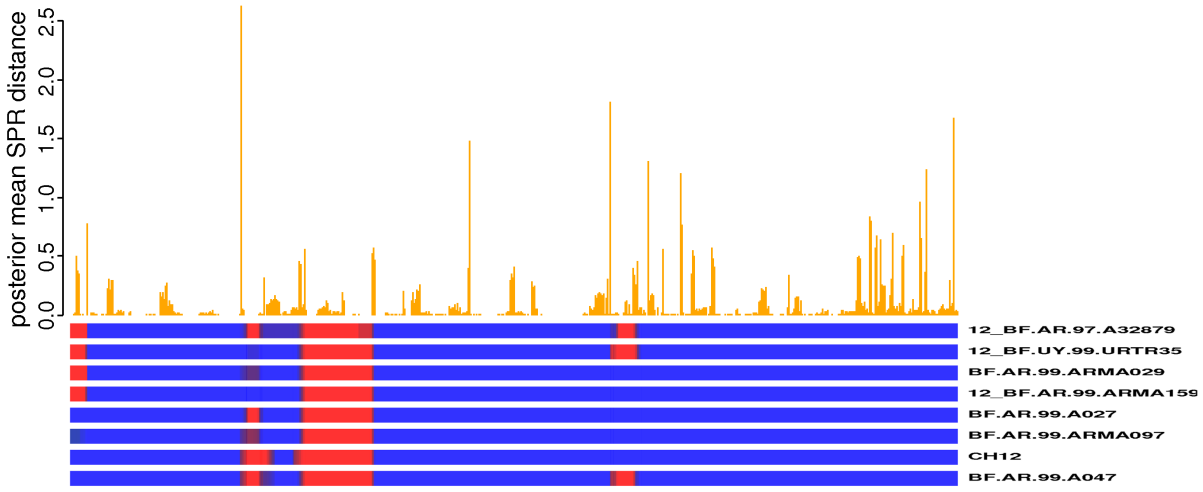

Figure 1: Posterior distribution of SPR distances among HIV-1 sequences. Below we have the genomic mosaic structure of each putative recombinant, where red means clustering with B subtype and blue indicates F subtype ancestrality.

points than that detected when parental sequences are assumed. This dataset was constructed by a systematic analysis of near full genome sequences from putative recombinant sequences from Brasil, Argentina and other South American countries. All of them were pre-analysed by bootscanning and determined to be variants of the subtype CRF12_BF. The procedure for choosing the recombinant sequences to be included in our analysis was thus by selecting sequences with the same recombination mosaic pattern, since in this case we can directly infer the monophyly of the recombinant sequences. We compared each putative recombinant sequence independently against reference subtypes F, B and C using the software DualBrothers. We utilized one reference parental sequence from each subtype to increase the detection power, avoiding contradicting signals. The sequences with the most similar mosaic structures as inferred by a hierarchical cluster analysis were then analyzed by our software, and the results are shown in Figure 1. In such a scenario we could confirm that all recombinations represented by the mosaic were reconstructed by our procedure, and the differences between the procedures reflected *de novo* recombination, that did not involve the reference parental subtypes.

The average of two recombinations per breakpoint, detected by noticing that the number of SPR moves was twice the number of breakpoints, is indeed an indication of existence of hot-spots. A scenario of one ancestral recombination giving rise to the diversity of a new recombinant subtype assumes that irrespective of intra-subtype recombination these recombinants should share a most recent common ancestor along all non-recombinant regions. Our results do not support a common ancestral origin for these recombinant sequences, at least for the chosen reference parental sequences, since the putative recombinants do not form a monophyletic group among segments.

We conclude that even for datasets displaying an identical recombination mosaic pattern, it is imperative to check for phylogenetic incongruences within the dataset. We must not rely on the breakpoints only as defined by the mosaic, since they are based on an arbitrary definition of sequences free from recombination.