29. In silico markovian prediction of variable and constant regions of lentivirus genomes

Aurélia Boissin-Quillon *, Didier Piau**, Caroline Leroux*

** UMR754 INRA-ENVL-UCBL-IFR 128 "Rétrovirus et Pathologie Comparée", Université Claude Bernard Lyon 1, 69007 Lyon*
*** UMR5208 "Institut Camille Jordan", Université Claude Bernard Lyon 1, 69007 Lyon*

Variation accumulates in the env gene coding the envelope proteins of lentiviruses, such as HIV (Human Immunodeficiency Virus), EIAV (Equine Infectious Anemia Virus), SRLV (Small Ruminant Lentiviruses) and SIV (Simian Immunodeficiency Virus), particularly in the surface SU glycoprotein part. This is due to the high level of viral replication, the low fidelity of the RT (reverse transcriptase) during the retrotranscription of the RNA viral genome into DNA, the lack of proofreading activity of the RT, and recombination events in coinfected cells. Interestingly the SU mutations do not appear uniformly along the env gene, but are clustered in restricted areas. Based on multiple sequences alignment, studies of the variation of the env gene led to the definition of variable (V) regions, framed by more stable constant (C) regions. To determine if the constant and variable regions were characterized by specific signatures, we developed a simple and robust mathematical method using the Hidden Markov Models (HMMs) to differentiate them.

HMMs are efficient statistical tools to analyze the heterogeneity along genomes, such as a succession of variable and constant regions. They are able to break down a heterogeneous sequence in a succession of different locally homogeneous regions, called states. Each state (or region) is described by a specific Markov chain. In a Markov chain, the value at position x depends on the values at positions x-1 to x-m, where m is the order of the model. The succession of states is itself ruled by a master first order Markov chain, called the hidden chain. Once the order of the model and the number of states have been fixed, each state is described by the specific transition probabilities between nucleotides, and the succession of states is described by the transition probabilities between states. These intra- and inter-state probabilities are usually estimated by an expectation-maximization (EM) algorithm that provides a maximum likelihood estimate of the parameters with no foreknowledge of either the observations or the states. In our case, this algorithm failed to predict the C and V regions of the lentiviruses. Thus, we skipped the usual estimation of the transition probabilities between nucleotides by the EM algorithm. We used as nucleotide transition probabilities, the ones estimated by counting the nucleotide transitions for the different regions from the alignment of the training set of sequences. We estimated iteratively the state transition probabilities with the EM algorithm, keeping every nucleotides transition probabilities at their calculated value. We named this new algorithm the fixed EM algorithm.

Based on sequences of the SU coding region of env, we defined HMMs using 187 EIAV sequences, 155 HIV sequences, 68 SRLV sequences, 61 SIV sequences. The HMMs have been developed on a training set composed of the half of the sequences, the other part of the sequences used to test the models. For each lentivirus, we developed HMMs with 2 hidden states, corresponding to a possible distinction between C and V regions. Using a first order (m=1) HMM on deduced amino acid sequences or a HMM of order 5 on DNA sequences, we obtained a clear and accurate delimitation of the known C and V regions on all the test sequences of the corresponding virus. However, the models trained on one lentivirus failed to identify the C and V regions of the others virus. Based on a training set composed of EIAV, HIV, SIV, and SRLV sequences, combined HMMs with 2 hidden states were developed. A HMM of order 1 on deduced amino acid sequences and a HMM of order 5 on DNA sequences allow to predict the C and V regions of all studied virus. Moreover, the combined models were able to correctly identify some constant and variable regions of the env gene of BIV (Bovine Immunodeficiency Virus) and FIV (Feline Immunodeficiency Virus), which were not used to train the models. This results show that the constant and variable regions of the lentiviruses can be identified with mathematical models, suggesting a different composition in words of nucleotides of these two types of region. We are at present attempting to extract the nature of the statistical signals delimiting these regions.