Analysis of the Molecular Clock Hypothesis for the Brazilian HIV-I Subtype

M. Carneiro, M. Pinto and C. J. Struchiner

Programa de Computação Científica (PROCC) – FIOCRUZ – Rio de Janeiro – Brazil

INTRODUCTION

Detailed knowledge about the rate of the genetic variation is vital for understanding how HIV induces disease and develops resistance, as well as for studies on the molecular epidemiology and origin of the virus. The molecular clock hypothesis in molecular evolutionary studies, has been controversial since it was first proposed (Zuckerland and Pauling 1965). A variety of methods have been developed to test the molecular clock and these have been applied to many studies, including the molecular evolution of viruses. The many discussions about the existence of a molecular clock suitable to the different genes of the HIV-1 have been based on very small data sets, what could likely create biased results. In this study, we tested 179 HIV-1 sequences carefully extracted from the GenBank, with the motifs most expressive in Brazil (Monica, 2003). The resulting Likelihood Ratio Test (Felsenstein, 1981) rejected the molecular clock hypothesis for the env V3 region. This casts doubt on the validity of recent attempts to date the origin of the epidemic.

TEST CONDITIONS

The Sequences were extracted from the GenBank using a simple Blast query that matched every HIV-1 sequence with more than 300 nucleotides. Then several filters were applied in order to select the sequences with the complete envelope region of the HIV genome, then selecting the sequences with the motifs most expressive in Brazil (GPGR and GWGR), and trying to avoid recombinants. The 179 best candidates were submitted to multiple sequence alignment through Clustalx and the phylogeny was estimated through Maximum Likelihood method, using the best fitted model of evolution (HKY), with the tree-puzzle software, it also compared then, the likelihood of the clocklike model as the null hypothesis against the non-clocklike model (more general).

FIRST RESULTS

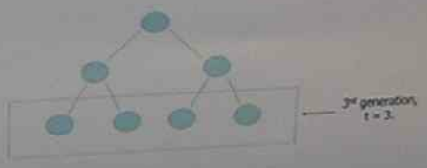
The simpler (clocklike) model was rejected on a significance level of 5%, for the log-likelihood with the more complex model (non-clocklike) has significantly increased. The fact that the sequences were randomly taken from the GenBank, without any information about its dates, should be taken into into consideration before interpreting this result as a non-clocklike evolutionary model.
But this consideration also brings the question whether this is a good test to infer the clocklike behaviour, and if it is, what's the best way we can provide data to it to expect the best results?

ANALYSIS OF THE LIKELIHOOD RATIO TEST

In order to analyse the Likelihood Ratio Test as a good test for the Molecular Clock Hypothesis, we first had to create clocklike evolved strains. For that we used the clockgen software which simulates the perfect clocklike evolution from a randomized strain. With the whole clocklike phylogeny, we could sample entire groups from different generations and use the Likelihood Ratio Test to check wether or not it would favor the simpler (clocklike) model within the critical significance level.

SAMPLING BY GENERATION

The generation sampling is the most perfect data that could be expected by a Phylogenetic Estimation program. It would contain exactly all strains that coexist in a given time t. Here we call this time, the Generation t.



SAMPLING BY GENERATION AND N-STRAINS

Although the results of the test of sampling by generation shows a fine exponential behaviour, we should also test the generations using the same degrees of freedom. While packing all the strains of one generation, we are comparing different n's, what is statistically incorrect. For that, we sampled exactly the same number of strains from each generation, and plotted another graph.

RESULTS AND FUTURE WORK

For the generation sampling we got an expected curvature showing that the farther the generation is, the harder it is for the phylogeny reconstructed in a clocklike fashion to get higher likelihood ratios, being almost always overwhelmed by the more complex model (graph 1). The "same degrees of freedom test" showed a less behavioured curvature, but still well formed, which could imply that the result of the first test was not (at least not badly) mistaken by the comparison of different degrees of freedom (graph 2). More tests and simulations can me made out of this framework in order to test population bottlenecks, random generation sampling, and other intra-host / inter-host phenomena.