Sequence Analysis of the Human Immunodeficiency Virus Type 1 Genome

Stephanie Lucas, University of San Francisco
Mentor: Takis Benos, University of Pittsburgh
June 16th, 2006

Introduction:

The field of genomics is built upon the idea of taking an organisms' genome and comparing that genome to other sequences of the same species or closely related species in order to draw conclusions about particular genes and understand its phylogenetic history (Mount 2001). These principles of sequence analysis will be used on the Human Immunodeficiency Virus Type 1 genome in order to (1) track if there are any differential selective pressures on parts of the genome, (2) identify areas of the genome that have higher or lower variability, and (3) predict and confirm transcription factors that might bind in the promoter sequence.

The HIV-1 genome contains nine genes and is approximately 9181 base pairs long (from the reference sequence from the RefSeq). The virus is circular, and (in the reference sequence entry) the U3/R/U5 control region connects the first gene (GAG-POL) and the last gene (NEF). The genes of the HIV-1 genome are presented below (van Opijnen et al. 2004):

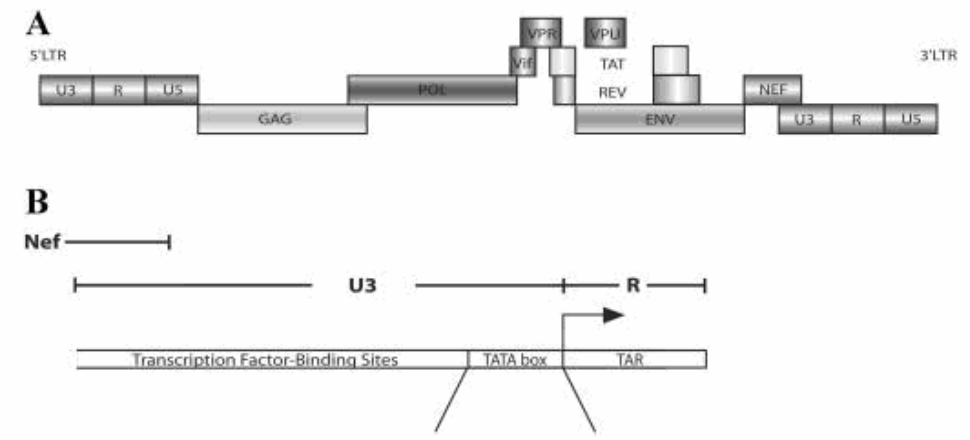

Figure 1: A) The complete genome of the HIV-1 virus. Note: The U3/R/U5 region is the same due to it being a circular molecule. GAG-POL in the sequence from the PDB is equivalent to the POL and GAG above. **B)** The U3/R region of the promoter sequence.

The U3/ R region includes many of the regulatory elements as shown above. A number of transcription factor binding sites have been predicted for nine different subtypes of HIV-1and are shown below (Jeeninga et al. 2000):

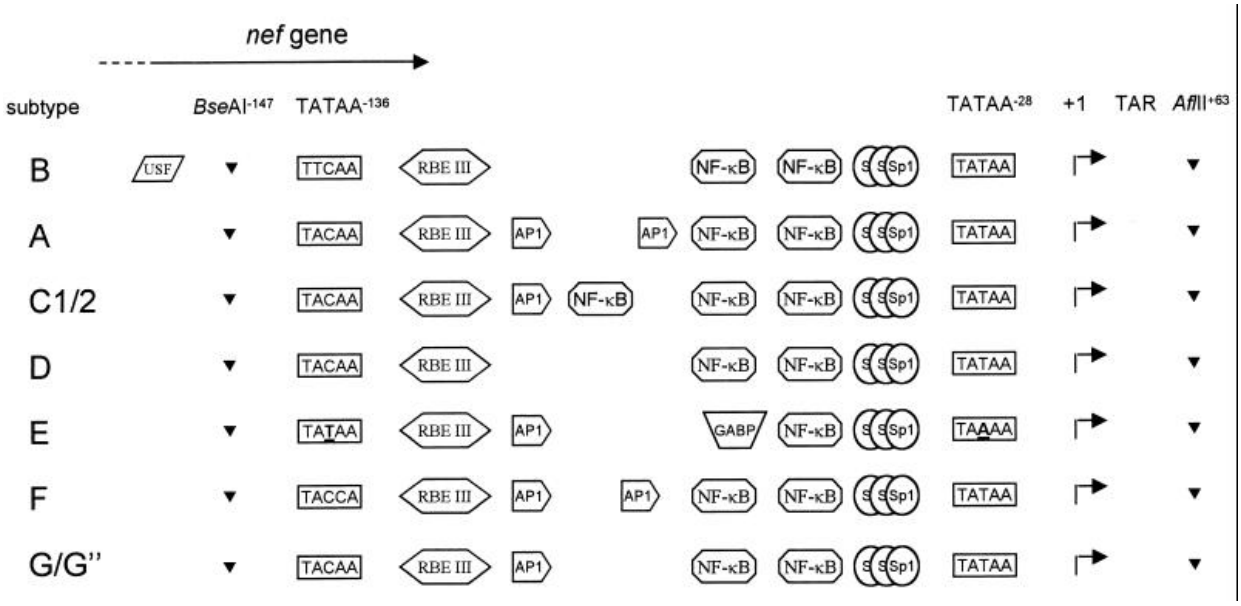

Figure 2: Expected transcription factor binding sites within 9 subtypes of HIV-1.

This project aims in the evolutionary study of the HIV-1 genome, in both the non-overlapping coding parts and the U3/R/U5 regulatory region. The parts of the U3/R/U5 region that are found to evolve slower than expected under Kimura's neutral model will be prime candidates for containing clusters of regulatory sites.

Methods

The first part of this project primarily consists of retrieving the reference sequence from the RefSeq. This sequence will be separated out into each of the nine individual genes and the U3/R/U5 region. The start and stop codons will be removed in order not to stray the conserved sequence count, since the start and stop codons are relatively invariable. Similarly, overlapping sequences must also be removed due to differential constraints applied to the individual bases. These parts may also skew the results. For example a shared nucleotide can be the second nucleotide position of one gene and the third for the other. In order to eliminate this complexity, only the non-overlapping ORF

sequences will be analyzed (nucleotides in multiples of three), leaving out 2 genes- GAG and TAT.

BLAST searches will be used on a database of over 1,100 sequenced HIV-1 genomes using the non-overlapping DNA sequences of the remaining seven genes and the U3/R/U5 region. In this way, a set of more than 1,110 independently evolving sequences will be obtained for each "part" of the HIV-1 genome.

These sequences will be imported and analyzed in the MEGA software package. MEGA allows us to detect instances of positive selection. Also, individual (gene-specific) sequence alignments and phylogenetic trees will be compared with the overall phylogenetic information. By doing so, any possible discrepancies- which will indicate differential evolutionary constraints in certain genes- will be detected. Another way to detect differential selective pressure in the various genes is by looking at the differences in their synonymous *vs.* non-synonymous changes as well as differences in the rate of transitions *vs.* transversions. We would expect that the amount of transitions will be twice as frequent as transversions.

Kimura's neutrality under the infinite sites model can be tested in *k* regions using a simple chi-square test (Kreitman & Hudson 1991). If neutrality fails for all the regions as a whole, we will test which region or combination of regions is responsible for the deviation from neutrality. This can be done by applying this same test in different sets of regions/genes.

It will also be interesting to see how the levels of polymorphisms in the U3/R/U5 region compare with the polymorphisms in each of the seven genes individually (at the synonymous positions) and all the genes together (average). The U3/R/U5 region can be divided in smaller regions for more accurate results. Although the ideal division length cannot be estimated, two strategies can be applied: (1) use the physical limits of the U3, R and U5 regions; (2) use a sliding window of a relatively arbitrary length (e.g., 100 bp)

The U3/R/U5 region will be further analyzed for potential transcription factor binding sites. A simple program like MATCH (comes the TRANSFAC database) can be used to predict sites of all the known mammalian transcription factors. We expect that many of these sites will be false positives. We plan to reduce the number of false positive

predictions by utilizing existing information about protein-protein interactions and by focusing on mutations that are characteristic of particular viral subtypes.

Expected Results:

This project is expected to yield information about the evolutionary constraints applied to the different parts of the HIV-1 genome. In addition, we will predict- with relative confidence- possible mammalian transcription factor binding sites that are used to regulate the expression of these viral genes.

Citations:

Kreitman, Martin, Hudson, Richard R. "Inferring the Evolutionary Histories of the *Adh* and *Adh-dup* Loci in *Drosophila melanogaster* From Patterns of Polymorphism and Divergence". Genetics 127 (1991): 565-582.

Jeeninga, Rienk E., Hoogenkamp, Maarten, Armand-Ugon, Mercedes, de Baar, Michel, Verhoef, Koen, Berkhout, Ben. "Functional Differences between the Long Terminal Repeat Transcriptional Promoters of Human Immunodeficiency Virus Type 1 Subtypes A through G". Journal of Virology 74:8 (2000): 3740-3751.

Mount, David W. 2001 Bioinformatics: Sequence and Genome Analysis. New York: Cold Spring Harbor Laboratory Press. 564p.

Van Opijnen, Tim, Kamoschinski, Joost, Jeeninga, Rienk E., Berkhout, Ben. "The Human Immunodeficiency Virus Type 1 Promoter Contains a CATA Box instead of a TATA Box for Optimal Transcription and Replication." Journal of Virology 78:13 (2004): 6883-6890.