

# Exploring HIV Evolution: An Opportunity for Research

Sam Donovan and Anton E. Weisstein

Video IV: Reading the Code of Life

Human Immunodeficiency Virus (HIV), like other retroviruses, has a much higher mutation rate than is typically found in organisms that do not go through reverse transcription (the copying of RNA into DNA). Mansky and Temin (1995) estimated the rate of point mutations in HIV to be  $3 \times 10^{-5}$  errors per base per replication cycle. The impact of HIV as a pathogen is due in large part to this high mutation rate which, among other things, causes the surface proteins to change and avoid normal immune detection and suppression. A great deal has been learned about the evolution of HIV because it is relatively easy to sample the rapidly changing population of viruses within an infected individual and look at the patterns of molecular change over time.

In this activity you will study aspects of sequence evolution by working with a set of HIV sequence data from 15 different subjects (Markham, et al., 1998). You will first learn about the dataset, then study the possible sources of HIV for these subjects, and then design and pursue your own research project.

Basic HIV biology such as life history characteristics of the virus and its interactions with the human immune system are not discussed here but are important background for placing these exercises into a broader biological and clinical context. A few suggested resources for reviewing basic HIV biology are provided in the reference section.

## An orientation to the HIV sequence data

The HIV genome is very small and relatively simple. It is made up of nine genes and about 9,500 nucleotides. In this lab, you will use sequences from the envelope gene, *env* (see Figure 1). The envelope gene codes for two membrane proteins (gp120 and gp41) that extend from the cell membrane and are involved in identifying target cells for the HIV to infect (see Figures 2 and 3). The HIV surface proteins are also sites that the immune system can sometimes detect, making it possible to destroy that HIV virus particle (see Figure 4). For the dataset you will be working with, the researchers identified different forms, or clones, of HIV based on differences in the nucleotide sequence of a short, 285 base pair, region of gp120 called V3 (see Figure 1). The V3 region is known to be highly variable and involved with both host cell and antibody recognition. Characterizing the population of HIV in a subject based on the different versions of the V3 region sequence gave researchers a measure of HIV evolution that has potential clinical significance.

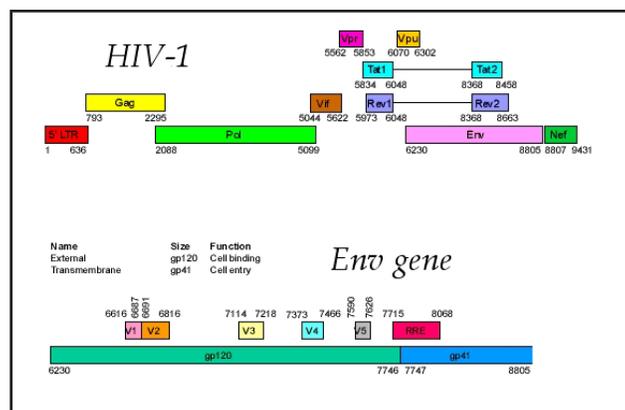


Figure 1: Schematics of the HIV-1 genome (top) and *env* gene (bottom). Note the overlap in the coding regions of genes. The variable regions (V1-V5) of *env* gene are loops that extend from the core of the protein.

Figure 2: A schematic of the HIV virus. Note the positioning of the *env* gene products gp41 and gp120 on the surface of the particle.

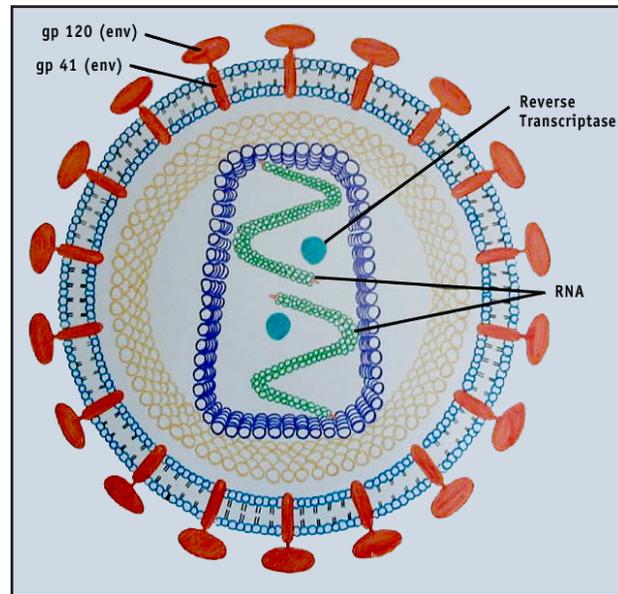
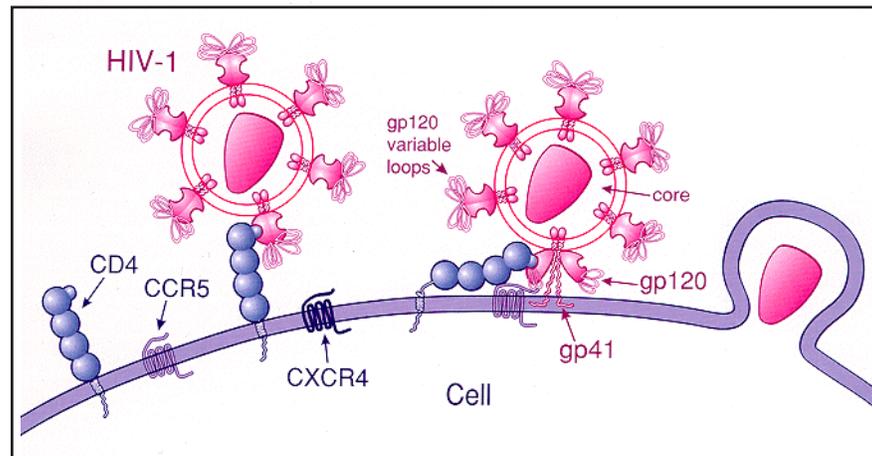


Figure 3: Schematic showing the role of gp120 in recognition of host cells.



### Activity 1: Looking at the NCBI Resources and HIV sequence data

As previously mentioned, these data were originally published as part of a research study looking at HIV evolution in different subjects. To learn a little more about the study and the data itself, this first activity involves searching the National Center for Biotechnology Information (NCBI) databases for additional information related to this research.

#### Part 1: PubMed

The NCBI provides access to a variety of databases including PubMed (a literature database) and GenBank (a nucleic acid sequence database). This federally funded research resource is very useful in part because the databases are linked together. This means that finding information in one area will allow you to look up related information in other areas. You should find the PubMed record for the Markham et al. article and then follow the links option for that record to find all the nucleotide sequences associated with that article.

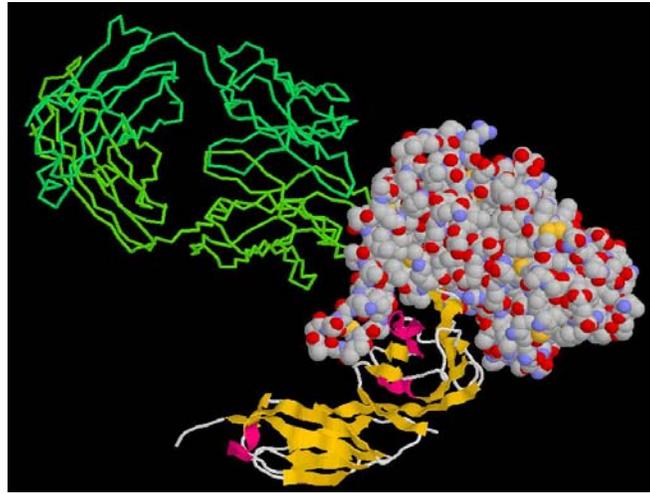


Figure 4: A representation of the gp120 protein structure (CPK space fill, right) and its interactions with the CD4 protein (ribbon, bottom) and an antibody (backbone trace, upper right). Structure published by Kwong et al., 1998.

NCBI website URL: <http://www.ncbi.nlm.nih.gov>

Original paper: Markham, R. B., W. C. Wang, A. E. Weisstein, Z. Wang, A. Munoz, A. Templeton, J. Margolick, D. Vlahov, T. Quinn, H. Farzadegan, X. F. Yu (1998). Patterns of HIV-1 evolution in individuals with differing rates of CD4 T cell decline. *Proceedings of the National Academy of Sciences* 95(21):12568-73. Pub Med ID: 98445411

- How did you search for the PubMed entry?
- What other ways might you have searched?
- What other types of related information are available?

## Part 2: GenBank

In this section you will take a closer look at a GenBank record and the type of data that is stored there. Once you reach the nucleic acid data associated with the Markham et al. paper you will see that there are a variety of different ways to view the data.

The data you will be working with is coded to help you recognize its source. While all of the data are HIV sequences, each sequence is identified based on the subject it was taken from, the visit during which it was collected, and its clone number. Thus, each sequence has a code like S4V2-4 that can be read as subject 4, visit 2, 4th clone. Each clone is a unique sequence collected during a particular visit. Over 600 different HIV sequences were identified in these 15 subjects and published electronically in GenBank. Choose one of the GenBank records and view both the full record and the FASTA formatted sequence.

- What was the accession number of the sequence you chose?
- Which subject of the study was that HIV sequence from? Which section of the record contains information about who the HIV was collected from?

- Download several (4 to 6) sequences in FASTA format to your local hard drive by selecting several at the same time in the summary view so they are saved into a single text file. Be careful to remember where you put the file and what you name it so that you can find it later.
- Open the file that you saved with a word processor to confirm that you have the sequences and that they are in the FASTA format. In the FASTA format each sequence is preceded by a label which begins with the greater than sign (>).

### Part 3: Introduction to the *Biology Workbench*

In order to analyze sequence data we will use the *Biology Workbench*, an on-line suite of bioinformatics tools. This section contains a brief introduction to the *Biology Workbench* that is intended to get you up and running quickly. There is more information about how the *Biology Workbench* is organized and how to use various tools in the “Orientation to the *Biology Workbench*” supplement on the *Microbes Count!* CD.

- Log in to the *Biology Workbench*: <<http://workbench.sdsc.edu>>.
- If you do not already have an account, you will need to set one up by following the Set up an account link.
- Once you have logged in, scroll down until you see the 5 buttons that take you to the different tool sets. This exercise uses nucleic sequence data, so follow the appropriate link.
- You should now see a scrolling list of tools for working with nucleic sequence data. Select Add new sequences and press the Run button. The next window allows you to enter sequences in a variety of ways. You can type them in directly, paste them in from another file, or upload a text file containing sequence data.
- Choose the Browse button to select the file you saved earlier from NCBI. Once you have selected the file use the Upload button to open it and read the data into this page.
- Once the labels and sequences appear in the data fields, choose Save to import that data into your *Biology Workbench* session.
- Each sequence should now appear as a data line below the list of analysis tools. Select one sequence and use the command View the Sequence to confirm that the sequence was successfully imported.

Now that you have been introduced to the *Biology Workbench* interface and procedures for uploading data files, you are ready to use the same procedures to upload a real dataset for analysis in the next activity.

- Look in the list of nucleic acid tools and find the ClustalW tool. Highlight the tool and then select Help to see more information about what this tool does.

- Select all of your sequences using the appropriate command and run a multiple sequence alignment using ClustalW.
- Look over the output and see if you can relate the differences in the sequences to the topology of the unrooted tree diagram and the pairwise similarity scores.

Note that the sequence labels on your tree will be the accession numbers for those sequence records. With just the accession numbers to describe the sequences, it may be difficult to think about the biological basis for the comparisons you just performed. Given what you have learned in Activity 1 you should be able to go back to GenBank and find additional information about the sequences you selected. We will use the ClustalW tool extensively in Activity 2. Look over this practice tree and think about how you will interpret your experimental trees to draw biological inferences from them.

- Go to the Session Tools and create a new session to store the sequences you will be working with in Activity 2.

## Activity 2: Looking at the sources of HIV across subjects

For this activity you will work with HIV sequence data collected from 15 individuals from an injection drug using population in Baltimore. The goal of the study will be to determine if the HIV isolated from particular subgroups of subjects derives from a common source. In order to approach this question you will need to characterize the populations of HIV within an individual and quantify the differences between individuals. The work will be distributed among groups to make the research more manageable. You can then combine your own results with those from other groups to try to draw some conclusions about the entire population of subjects.

Most of your analyses will involve building multiple sequence alignments and distance-based unrooted trees. A multiple sequence alignment involves lining up several sequences so that comparisons can be made across sequences for each nucleotide position. Differences between related sequences can then be interpreted as nucleotide substitutions. Building alignments sometimes involves inserting gaps that represent areas where an insertion or deletion has taken place in one or more of the sequences. You will be using the ClustalW tool to build your alignments and distance-based unrooted trees.

### Part 1: Looking at clustering across subjects

Upload the “visit\_1\_S1\_S9.txt” and “visit\_1\_S10\_S15.txt” files into the nucleic acid tool set of the *Biology Workbench* session you created in the last exercise. There are 2 data files because the *Biology Workbench* only allows up to 64 sequences to be uploaded at a time. Together these 2 files contain 97 sequences from the 15 subjects’ first visits (see Table 1). The sequence labels for these data include information about the subject, visit and clone number for each sequence.

This research was a prospective study, meaning that the subjects were receiving regular HIV screening and the first visit for this dataset indicates the first time



In tracing the branches between any two terminal nodes, or sequences, the total length traced reflects the genetic distance between those sequences. We can use the genetic distances as a first-order approximation of the evolutionary relationships between the sequences. The greater the genetic distance, the longer the time since they shared a common ancestor. However, because it is an unrooted tree we can't make any inferences about the direction of the evolutionary change. An important feature of unrooted trees are long internal branches. Long internal branches (i.e., branches that don't connect to terminal nodes) separate clusters of sequences. For example, Figure 5a suggests that sequences 1 and 2 are much more similar to each other than either of them is to sequences 3 or 4. On the other hand, Figure 5b does not provide strong evidence for grouping sequences.

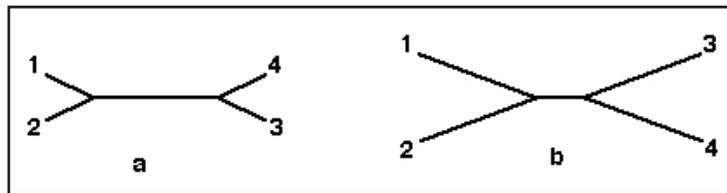


Figure 5: Hypothetical unrooted distance trees showing the difference between long (a) and short (b) internal branch lengths.

- Sketch your distance tree on paper so that you can make notes on it.
- Do the clones from each subject cluster together?
- Do some subjects' clones show more diversity than others?
- Do some of the subjects cluster together?
- Write a brief description of your tree and how you interpret the clustering pattern with respect to the similarities and potential evolutionary relationships between subjects' HIV sequences.
- Copy and paste your tree from the *Biology Workbench* into a word processor and print it out to share with the class.

## Part 2: Quantifying diversity within and between subjects

This section will introduce several additional ways to quantify sequence similarity and difference. These measures will help you design and interpret your research in the next section.

Table 1 illustrates that different subjects have different numbers of clones in their initial samples. This is one crude way to quantify the diversity of the HIV within that individual at that time, assuming that each subject's sample contains approximately the same total number of viruses. The  $S$  statistic can also be used to quantify the diversity of sequence in a population. The value of  $S$  is simply the number of positions that vary, or are not identical, across all the sequences in an alignment.

- Select all the clones from one subject and align them. From the alignment calculate  $S$  by counting the number of positions where there is at least one nucleotide difference across the collection of clones. Enter your data into Table 3 below.
- Import your alignment so that you can use it in a later section.
- Run the same analysis for a second and third subject and record your results in Table 3.

Subject	Number of Clones	$S$	Theta	Min difference	Max difference

Table 3

Next you will calculate  $\theta$  (theta), which is an estimate of the average pairwise genetic distance. It is an estimate because it does not take into account the frequency of the different clones in the sample.  $\theta$  is given by the formula

$$\theta = S / \left( \sum_{i=1}^{n-1} \frac{1}{i} \right),$$

where  $S$  is the statistic you just calculated above and the denominator of  $\theta$  is a partial harmonic sum. For a sample of  $N = 5$  the denominator would be calculated by  $1 + 1/2 + 1/3 + 1/4 = 25/12$ . The harmonic sum is used because we do not expect the number of differences to increase linearly with the number of sequences analyzed.

- Calculate  $\theta$  for the three subjects you chose to work with and enter the results in the datatable.

The last measures we will investigate are the minimum and maximum differences between any pair of sequences in an alignment. While  $S$  and  $\theta$  provide some measure of difference across a set of sequences, the minimum and maximum pairwise distances look only at the extremes in terms of similarity and difference. You could obtain these values by looking directly at parts of larger alignments but it is easier to let the computer provide the pairwise comparisons for you.

- Use the Clustdist tool in the alignment tool set to generate a distance matrix for an alignment you saved.
- Select the highest and lowest pairwise scores and convert that percentage difference score into the raw number of differences by multiplying by the length of the sequence (285 in most cases).

- Round to the nearest integer, record your results in Table 3 and repeat the analysis for the other 2 subjects.

This measure can also be used to look at maximum and minimum differences across subjects.

- First, create a new alignment with all of the sequences from 2 subjects.
- Next, use Clustdist to generate a pairwise distance matrix for the alignment across subjects and find the minimum and maximum differences between the subjects sequences. Remember, you want to look only at pairwise distances across subjects, not within one subject.
- Record your data in Table 4 and repeat the analysis for the other pairs of subjects.

Pair of subjects being compared	Min difference	Max difference

Table 4.

- Post your results from this preliminary analysis so that other research groups can review your work as they design the research project in Activity 3.

### Part 3: Design an analysis for looking at the sources of HIV

You now have the basic tools to compare and quantify groups of sequences. Look over the class results from the different groups' preliminary analyses and generate a list of possible specific questions to pursue. Design a small research project that will address one of your specific questions and contribute to the class-wide knowledge about the possible sources of HIV in this population.

- Prepare a report that communicates the question or idea you were testing, how you approached it with data collection and analysis, a summary of your findings, and an interpretation of your results.

### Activity 3: HIV evolution research project

The data you used in Activity 2 was just a fraction of the data available from this study. The Markham et al. research followed this populations of 15 subjects over time and continued to sample and characterize their HIV populations at 6-month

intervals for up to four years. In addition to studying the HIV population, the researchers estimated the immunological function of the subjects at each visit by measuring the number of CD4 T-cells in their blood samples. The CD4 T-cell count is an indicator of the health of the immune system because it represents a class of immune cells that are attacked by the HIV virus. A CD4 count below 200 cells/microliter is one of the diagnostic indicators used for defining Acquired Immune Deficiency Syndrome (AIDS).

There is a table summarizing the entire dataset on the CD. There are also files containing all the data from all the visits made by each subject. These data files are saved in FASTA format with labels like those used in the dataset you used in Activity 2.

- Print and review the data summary table for the complete Markham et al. data set.
- Discuss with your group any patterns you see in the summary data table and possible research questions related to those patterns.
- Design a small research study and prepare a scientific poster to share your project with the class.

### Web Resources Used in this Activity

*Biology Workbench* (<http://workbench.sdsc.edu>)

The *Biology Workbench* was originally developed by the Computational Biology Group at the National Center for Supercomputing Applications at the University of Illinois at Urbana-Champaign. Ongoing development of version 3.2 is occurring at the San Diego Supercomputer Center, at the University of California, San Diego. The development was and is directed by Professor Shankar Subramaniam.

Platform Compatibility: Requires an internet connection and a current browser.

### Additional Resources

#### Available on the *Microbes Count!* CD

##### Text

A PDF copy of this activity, formatted for printing

Data files from the Markham et al. (1998) research project

“Orientation to the *Biology Workbench*”

#### Related *Microbes Count!* Activities

Chapter 2: Searching for Amylase

Chapter 4: Molecular Forensics

Chapter 6: Proteins: Historians of Life on Earth

Chapter 6: Tree of Life: Introduction to Microbial Phylogeny

Chapter 6: Tracking the West Nile Virus

Chapter 6: One Cell, Three Genomes

Chapter 7: Visualizing Microbial Proteins

### **Unseen Life on Earth Telecourse**

Coordinates with Video IV: Reading the Code of Life

### **Relevant Textbook Keywords**

Antibody, CD4 cell, Evolution, Gene, HIV, Mutation, Phylogenetic relationships, Retrovirus

### **Related Web Sites** (accessed on 4/25/03)

Cells Alive HIV Tutorial

<http://www.cellsalive.com/hiv0.htm>

Centers for Disease Control and Prevention

<http://www.cdc.gov/hiv/dhap.htm>

Good info on HIV lifecycle

[http://www.thebody.com/niaid/hiv\\_lifecycle/virpage.html](http://www.thebody.com/niaid/hiv_lifecycle/virpage.html)

HIV Biology Background

<http://medlib.med.utah.edu/WebPath/TUTORIAL/AIDS/HIV.html>

Los Alamos HIV Lab

<http://hiv-web.lanl.gov/>

*Microbes Count!* Website

<http://bioquest.org/microbescount>

Unseen Life on Earth: A Telecourse

[http://www.microbeworld.org/htm/mam/is\\_telecourse.htm](http://www.microbeworld.org/htm/mam/is_telecourse.htm)

### **References** (accessed on 4/25/03)

Mansky, L. M. and H. M. Temin (1995). Lower in vivo mutation rate of human immunodeficiency virus type 1 than predicted by the fidelity of purified reverse transcriptase. *Journal of Virology*. 69:5087-5094. (Pub Med ID: 7541846)

Markham, R. B., W. C. Wang, A. E. Weisstein, Z. Wang, A. Munoz, A. Templeton, J. Margolick, D. Vlahov, T. Quinn, A. Farzadegan, X. F. Yu (1998). Patterns of HIV-1 evolution in individuals with differing rates of CD4 T cell decline. *Proc. Natl. Acad. Sci.* 95(21):12568-73. (Pub Med ID: 98445411)

Kwong, P. D., R. Wyatt, J. Robinson, R. W. Sweet, J. Sodroski, W. A. Hendrickson (1998). Structure of an HIV gp120 envelope glycoprotein in complex with the CD4 receptor and a neutralizing human antibody. *Nature* 393(6686):648-59. (Pub Med ID: 9641677)

### Figure and Table References

Figure 1. <http://hiv-web.lanl.gov/content/hiv-db/MAP/Env.html>  
<http://hiv-web.lanl.gov/content/hiv-db/MAP/landmark.html>

Figure 2. Courtesy of Sam Donovan

Figure 3. <http://www.niaid.nih.gov/daids/dtpdb/graphics/cellbin.gif>

Figure 4. Courtesy of Sam Donovan

Figure 5. Courtesy of Sam Donovan

For Figure 1.

Copyright Notice

For Scientific and Technical Information Only

Copyright © 1998-2003 The Regents of the University of California.

For All Information

Unless otherwise indicated, this information has been authored by an employee or employees of the University of California, operator of the Los Alamos National Laboratory under Contract No. W-7405-ENG-36 with the U.S. Department of Energy. The U.S. Government has rights to use, reproduce, and distribute this information. The public may copy and use this information without charge, provided that this Notice and any statement of authorship are reproduced on all copies. Neither the Government nor the University makes any warranty, express or implied, or assumes any liability or responsibility for the use of this information.