Molecular Forensics

Sam Donovan

Video IV: Reading the Code of Life

Introduction

Molecular data are routinely used to determine paternity and to link physical evidence at a crime scene to individuals. The specificity of DNA sequence data makes it possible to link tissue samples—even very small amounts of hair, blood, saliva or semen—to the individual who is the source of the material, with a high degree of confidence and accuracy. The idea of a DNA "fingerprint" involves comparing a set of "molecular markers" from a tissue sample to samples from known subjects. The identity, or lack of identity, between the known and unknown samples can then be used as evidence in a court of law.

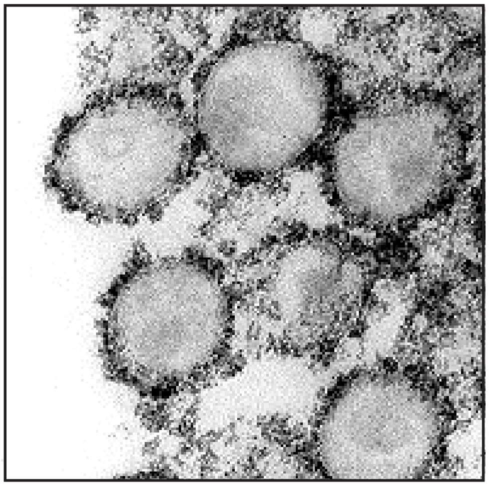


Figure 1. HIV particles

In addition to using molecular data to find identity matches, DNA evidence has been used in courtrooms in other ways as well. In this exercise we will look at how the analysis of viral evolution can be used to make biological and legal arguments about the transmission of the virus between individuals. Instead of looking for an identity match you will use the similarity between sequences to group strains of the virus and make inferences about their historical relationships. We know that a population of HIV viruses within a person evolves quickly due to its rapid generation time and frequent mutations. We can use the pattern of these changes to look for evidence that people share the same virus source.

This activity is based on an actual case which broke new ground with respect to the use of molecular evidence in a courtroom. While the exercises outlined here use some of the actual data collected and analyzed as part of the court proceedings, they do not address the complexity nor the depth of the evidence. They do, however, provide you with an opportunity to practice some of the reasoning used to go from sequence data to biological inference, and to translate that biological understanding into evidence that could be used in a court of law. For more information about the history of this case see the Additional Resources section at the end of this activity.

Background

In the spring of 1990, Kimberly, a 22 year-old living in Fort Pierce, Florida, tested positive for HIV. This was surprising because she had no identifiable risk factors for contracting the virus. Epidemiological research focused on an invasive dental procedure performed by an HIV positive dentist several years earlier. Searching the dentist's records revealed a number of other HIV positive patients, several of whom also had no known risk factors for contracting the virus. The

Centers for Disease Control and Prevention (CDC) became involved and the case received a great deal of media attention due to the public's concern that HIV+ health care workers might be a threat to their patients. Multiple lawsuits were filed by patients claiming that they were infected by the dentist and seeking damages.

The following exercises provide an introduction to working with sequence data and drawing inferences from sequence alignment and distance trees. By comparing the genetic sequences of HIV virus genes isolated from blood samples from the dentist, his patients, and other HIV+ individuals in the community who did not have contact with the dentist (local controls), scientists worked to determine if there was a relationship between the dentist's and patients' viruses. As you try to establish whether there is evidence that the dentist was responsible for the HIV infection in his patients you will also explore:

- characteristics of nucleic acid sequence data;
- how a multiple sequence alignment can be used to visualize similarities and differences between sequences; and,
- how a distance graph (unrooted tree) represents the differences between sequences and can be used to develop hypotheses about their evolutionary relationships.

Part 1 - Comparing Raw Sequence Data

The data in Table 1 below includes 3 sequences from different patients (E, F and G), a dentist sequence, and 2 sequences from HIV+ individuals who lived in the area but had not had contact with the dentist (local controls). Each sequence is from the same part of the HIV genome (the GP120 gene, V3 region) — a variable region of a gene that makes a viral coat protein.

Study the data in Table 1 and discuss the questions below with your group members. Be prepared to share your findings and any questions that arise.

- What sorts of patterns do you see within/between these sequences?
- How are these sequences similar (different)?
- Are they all similar/different in the same ways?
- How do you think this information could be used to determine if the dentist was the source of the HIV in the patients?

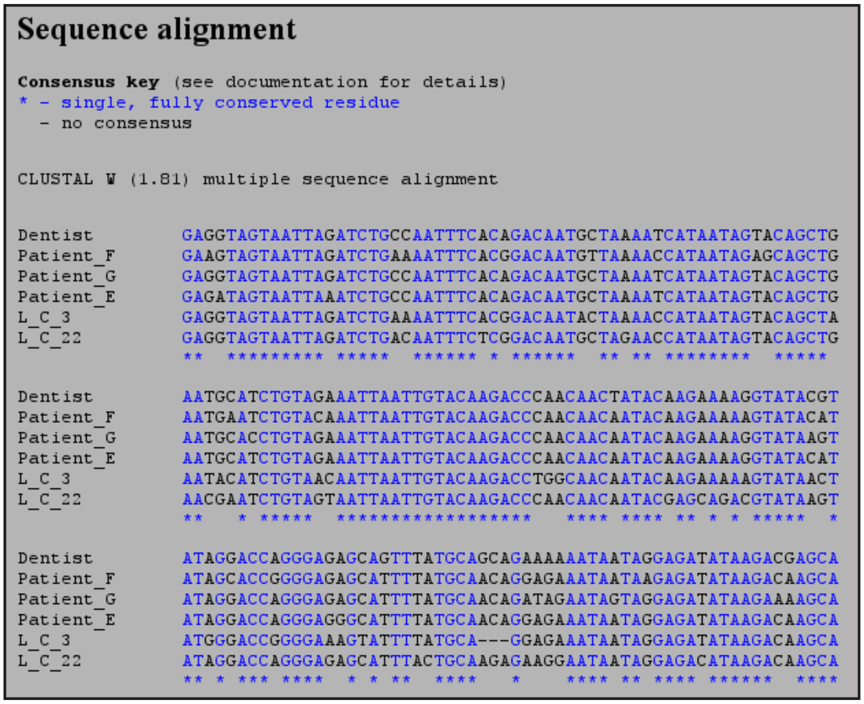
Table 1: HIV nucleic acid sequences from 6 subjects.

Dentist	GAGGTAGTAATTAGATCTGCCAATTTCACAGACAATGCTAAAATCATAATAGTACAGCTGAATGCATCTGTAGAAATTAATTGTACAAGACCCAACAACTATACAAGAAAAGGTATACGTATAGGACCAGGGAGAGCAGTTTATGCAGCAGAAAAAATAATAGGAGATATAAGACGAGCACATTGTAACATTAGTAGAGAAAAATGGAATAATACTTTAAAACAGGTAGTTACAAAATTAAGAGAACAATTTGTGAATAAAACAATAATCTTTACTCACCCCTCAGGAGGGGACCCAGAAAT
Patient E	GAGATAGTAATTAAATCTGCCAATTTCACAGACAATGCTAAAATCATAATAGTACAGCTGAATGCATCTGTAGAAATTAATTGTACAAGACCCAACAACAATACAAGAAAAGGTATACATATAGGACCAGGGAGGGCATTTTATGCAACAGGAGAAATAATAGGAGATATAAGACAAGCACATTGTAACATTAGTGGAGAAAAATGGAATAATACTTTAAAACAGGTAGTTACAAAATTAAGAGAACAATTTGGGAATAAAACAATAATCTTTAATCACTCCTCAGGAGGGGACCCAGAAAT
Patient F	GAAGTAGTAATTAGATCTGAAAATTTCACGGACAATGTTAAAACCATAATAGAGCAGCTGAATGAATCTGTACAAATTAATTGTACAAGACCCAACAACAATACAAGAAAAAGTATACATATAGCACCGGGGAGAGCATTTTATGCAACAGGAGAAATAATAAGAGATATAAGACAAGCACATCGTAACCTTAGTAGCATAAAATGGAATAACACTTTAAGACAGATAGCTAAAAAATTAAAAGAACAATTTGGAAATAAAACAATAATCTTTAATCAATCCTCAGGAGGGGACCCAGAAAT
Patient G	GAGGTAGTAATTAGATCTGCCAATTTCACAGACAATGCTAAAATCATAATAGTACAGCTGAATGCACCTGTAGAAATTAATTGTACAAGACCCAACAACAATACAAGAAAAGGTATAAGTATAGGACCAGGGAGAGCATTTTATGCAACAGATAGAATAGTAGGAGATATAAGAAAAGCATATTGTAACATTAGTAGAGAAAAATGGAATAATACTTTAAAACTGGTAGTTACAAAATTAAGAGAACAATTTGTGAATAAAACAATAATCTTTAATCACTCCTCAGGAGGGGACCCAGAAAT
Local Control 3	GAGGTAGTAATTAGATCTGAAAATTTCACGGACAATACTAAAACCATAATAGTACAGCTAAATACATCTGTAACAATTAATTGTACAAGACCTGGCAACAATACAAGAAAAAGTATAACTATGGGACCGGGGAAAGTATTTTATGCAGGAGAAATAATAGGAGATATAAGACAAGCACATTGTAACCTTAGTAGAACAGCATGGAATGACACTTTAGAACAGATAGTTGGAAAATTACAAGAACAATTTGGGAATAAAACAATAGTCTTTAATCACTCCTCAGGAGGGGACCCAGAAAT
Local Control 22	GAGGTAGTAATTAGATCTGACAATTTCTCGGACAATGCTAGAACCATAATAGTACAGCTGAACGAATCTGTAGTAATTAATTGTACAAGACCCAACAACAATACGAGCAGACGTATAAGTATAGGACCAGGGAGAGCATTTACTGCAAGAGAAGGAATAATAGGAGACATAAGACAAGCACATTGTAACATTAGTGGAGCAGAATGGGAAAGCACTTTAAAACGGATAGTTGAAAAATTAGGAGAACAATTTAAGAATAAAACAATAGTCTTTAATCACTCCTCAGGAGGGGACCCAGAAAT

Part 2 - Interpreting a Multiple Sequence Alignment

While it is possible to manually compare raw sequence data, it quickly gets unwieldy when you are working with long sequences or many different sequences. Luckily, computers are very efficient at following instructions and performing mathematical operations. In this section you will interpret the output from a program that has performed a multiple sequence alignment on the six sequences you looked at in Part 1. The ClustalW program is used to “align” sequences by finding the best ways to make their different nucleotide positions line up with one another and then color coding the positions (columns) to characterize the types of differences between sequences. Figure 2 shows part of a ClustalW multiple sequence alignment of the raw sequence data from Table 1.

Figure 2. Multiple sequence alignment.



Sequence alignment

Consensus key (see documentation for details)
* - single, fully conserved residue
  - no consensus

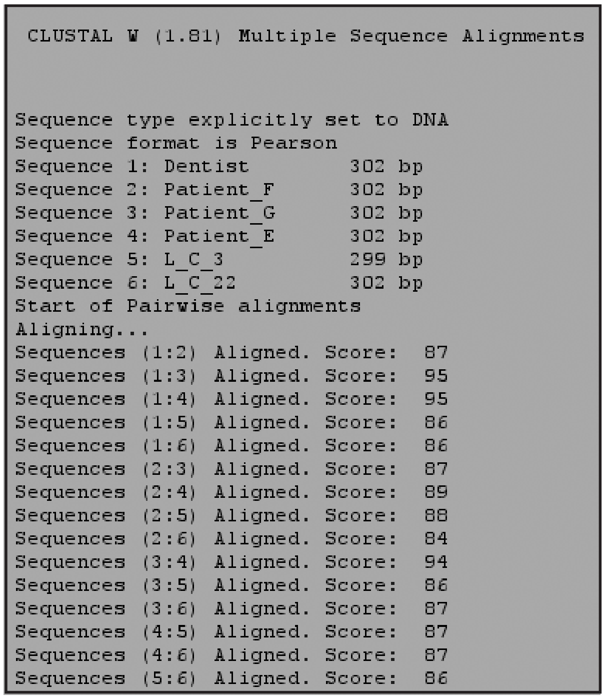
```
CLUSTAL W (1.81) multiple sequence alignment


Dentist        GAGGTAGTAATTAGATCTGCCAATTTCACAGACAATGCTAAAATCATAATAGTACAGCTG
Patient_F      GAAGTAGTAATTAGATCTGAAAATTTCACGGACAATGTTAAAACCATAATAGAGCAGCTG
Patient_G      GAGGTAGTAATTAGATCTGCCAATTTCACAGACAATGCTAAAATCATAATAGTACAGCTG
Patient_E      GAGATAGTAATTAAATCTGCCAATTTCACAGACAATGCTAAAATCATAATAGTACAGCTG
L_C_3          GAGGTAGTAATTAGATCTGAAAATTTCACGGACAATACTAAAACCATAATAGTACAGCTA
L_C_22         GAGGTAGTAATTAGATCTGACAATTTCTCGGACAATGCTAGAACCATAATAGTACAGCTG
               **  ********* *****  ****** * ******  ** ** ********  *****

Dentist        AATGCATCTGTAGAAATTAATTGTACAAGACCCAACAACTATACAAGAAAAGGTATACGT
Patient_F      AATGAATCTGTACAAATTAATTGTACAAGACCCAACAACAATACAAGAAAAAGTATACAT
Patient_G      AATGCACCTGTAGAAATTAATTGTACAAGACCCAACAACAATACAAGAAAAGGTATAAGT
Patient_E      AATGCATCTGTAGAAATTAATTGTACAAGACCCAACAACAATACAAGAAAAGGTATACAT
L_C_3          AATACATCTGTAACAATTAATTGTACAAGACCTGGCAACAATACAAGAAAAAGTATAACT
L_C_22         AACGAATCTGTAGTAATTAATTGTACAAGACCCAACAACAATACGAGCAGACGTATAAGT
               **   * *****  ******************   **** **** ** * * *****  *

Dentist        ATAGGACCAGGGAGAGCAGTTTATGCAGCAGAAAAAATAATAGGAGATATAAGACGAGCA
Patient_F      ATAGCACCGGGGAGAGCATTTTATGCAACAGGAGAAATAATAAGAGATATAAGACAAGCA
Patient_G      ATAGGACCAGGGAGAGCATTTTATGCAACAGATAGAATAGTAGGAGATATAAGAAAAGCA
Patient_E      ATAGGACCAGGGAGGGCATTTTATGCAACAGGAGAAATAATAGGAGATATAAGACAAGCA
L_C_3          ATGGGACCGGGGAAAGTATTTTATGCA---GGAGAAATAATAGGAGATATAAGACAAGCA
L_C_22         ATAGGACCAGGGAGAGCATTTACTGCAAGAGAAGGAATAATAGGAGACATAAGACAAGCA
               ** * *** ****  * * **  ****   *    **** ** **** ******  ****
```

- Does the information presented in the multiple sequence alignment support the patterns you saw when you looked at the raw sequence data?
- Why do you think that the Local Control 3 (L_C_3) sequence had a "gap" (-) inserted in its sequence?
- How do you think this information could be used to determine if the dentist was the source of the HIV in the patients?

Part 3 - Reading a Distance Tree

Another way to compare these sequences is to look at the genetic similarity between each pair of sequences. In this section this pairwise similarity is reported first as a table of pairwise % identities and then as a distance tree. The % identity table was generated by comparing each sequence with every other sequence, calculating the number of positions that are identical and then dividing that number by the total number of positions. The distance tree represents genetic distances between sequences as lengths between the tips of the branches. Thus, tracing the lines from one sequence, or branch tip, to another correlates with the pairwise distance reported in Figure 3.

Part of determining if the dentist is the source of the patients' HIV is seeing how the sequences group together based on their similarity. Because HIV evolves rapidly, it is very unlikely that the viral sequences in two subjects would be identical even if one person had transmitted the virus to the other. However, using the assumption that genetically similar sequences are more closely related to one another and therefore share a more recent common ancestor, it is possible to infer past identity from similarity.



```
 CLUSTAL W (1.81) Multiple Sequence Alignments


Sequence type explicitly set to DNA
Sequence format is Pearson
Sequence 1: Dentist        302 bp
Sequence 2: Patient_F      302 bp
Sequence 3: Patient_G      302 bp
Sequence 4: Patient_E      302 bp
Sequence 5: L_C_3          299 bp
Sequence 6: L_C_22         302 bp
Start of Pairwise alignments
Aligning...
Sequences (1:2) Aligned. Score:  87
Sequences (1:3) Aligned. Score:  95
Sequences (1:4) Aligned. Score:  95
Sequences (1:5) Aligned. Score:  86
Sequences (1:6) Aligned. Score:  86
Sequences (2:3) Aligned. Score:  87
Sequences (2:4) Aligned. Score:  89
Sequences (2:5) Aligned. Score:  88
Sequences (2:6) Aligned. Score:  84
Sequences (3:4) Aligned. Score:  94
Sequences (3:5) Aligned. Score:  86
Sequences (3:6) Aligned. Score:  87
Sequences (4:5) Aligned. Score:  87
Sequences (4:6) Aligned. Score:  87
Sequences (5:6) Aligned. Score:  86
```

Figure 3: Pairwise sequence similarities

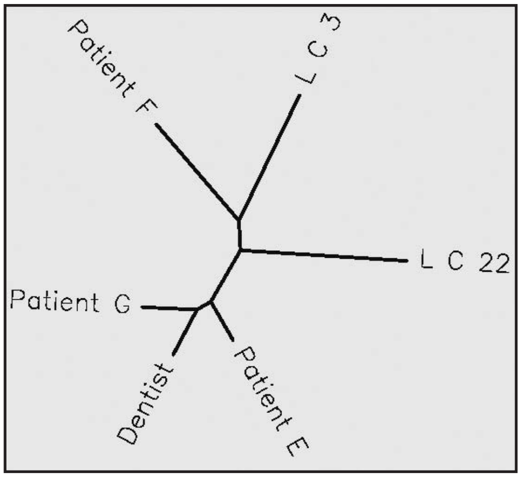


Figure 4: Distance tree.

- Does the information presented in the table of pairwise similarity scores (Figure 3) and in the distance tree (Figure 4) support the patterns you saw when you looked at the raw sequence data and the multiple sequence alignment?

- Why do you think some of the lines are longer than others?

- Do you think the places where the lines connect with one another is important? How can the internal branches be interpreted?

- How do you think this information could be used to determine if the dentist was the source of the HIV in the patients?

Assignment

- Write a brief summary of the argument you would make to a judge or jury regarding the claims from these three patients that they received their HIV infections from the dentist. Refer directly to the data available in this exercise and be explicit about how you are interpreting it. Be careful not to overstate your conclusions.

Critical thinking questions

- What does the distance tree diagram tell you about the directionality of HIV transmission? What does it tell you about the likelihood of direct transmission between dentist and patient(s)? Does your language in the argument above reflect this understanding?
- This exercise has been simplified dramatically by including only one sequence from each subject in the analysis. In fact, each of these individuals had a variable population of HIV in their bodies. How might information about the similarities and differences between HIV populations for each individual shape your argument?
- Explain the importance of having local controls in your analysis. How are these sequences used to inform the case you make in the courtroom?

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

Teaching Notes

Dataset from lab, for use with the Teaching Notes

Related *Microbes Count!* Activities

Chapter 2: Searching for Amylase

Chapter 4: Exploring HIV Evolution: An Opportunity for Research

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

Amino acids, Bioinformatics, Disease, Epidemiology, Evolution, HIV, Nucleic acids, Phylogenetic relationships,Virus

Related Web Sites

Biology Workbench
http://workbench.sdsc.edu

Forensics and Genetics
http://ornl.gov/hgmis/elsi/forensics.html

Microbes Count! Website
http://bioquest.org/microbescount

Unseen Life on Earth: A Telecourse
http://www.microbeworld.org/htm/mam/is_telecourse.htm

References

Popular literature

Gentile, B. (July 1, 1991). Doctors with AIDS. *Newsweek* 48-56.

Scientific Literature

Ou, C. Y., C. A. Ciesislski, G. Myers, et al. (1992). Molecular epidemiology of HIV transmission in a dental practice. *Science* 256:1165–1171. (PMID: 1589796; UI: 92271245)

Bibliography

Barr, S. (1996). The 1990 Florida dental investigation: Is the case really closed? Annals of Internal Medicine 124:250-254. http://www.acponline.org/journals/annals/15jan96/flordent.htm

The Centers for Disease Control and Prevention (CDC) publish a weekly newsletter called the Morbidity and Mortality Weekly Report (MMWR). The following articles contain nice concise descriptions of the work involved in understanding the transmission of HIV in this case.

Possible Transmission of Human Immunodeficiency Virus to a Patient during an Invasive Dental Procedure. July 27, 1990, 39(29):489-493
http://www.cdc.gov/mmwr/preview/mmwrhtml/00001679.htm

Epidemiologic Notes and Reports Update: Transmission of HIV Infection during an Invasive Dental Procedure–Florida. January 18, 1991, 40(2):21-27, 33
http://www.cdc.gov/mmwr/preview/mmwrhtml/00001877.htm

Epidemiologic Notes and Reports Update: Transmission of HIV Infection During Invasive Dental Procedures–Florida. June 14, 1991, 40(23):377-381.
http://www.cdc.gov/mmwr/preview/mmwrhtml/00014428.htm

Investigations of Persons Treated by HIV-Infected Health-Care Workers–United States. May 07, 1993, 42(17):329-331,337
http://www.cdc.gov/mmwr/preview/mmwrhtml/00020479.htm

Figure and Table References

Figure 1.	www.ncbi.nlm.nih.gov
Figure 2.	Modified from Biology Workbench (http://workbench.sdsc.edu)
Figure 3.	Modified from Biology Workbench (http://workbench.sdsc.edu)
Figure 4.	Modified from Biology Workbench (http://workbench.sdsc.edu)
Table 1.	Sequences obtained from GenBank (http://www.ncbi.nlm.nih.gov/)