

# Tree of Life: An Introduction to Microbial Phylogeny

Beverly Brown, Sam Fan, LeLeng To Isaacs, and Min-Ken Liao

Video VI: Microbial Evolution

## Introduction

Bioinformatics tools allow microbiologists and evolutionary biologists to take a closer look at the evolutionary relationships between species by quantifying the differences between genetic sequences from many organisms. New hypotheses about the history of life on earth, including the re-rooting of existing phylogenetic trees—that is, identifying the earliest forms of life—are based on new sequence data that further defines the molecular similarities and differences between species

This activity will help familiarize you with the use of internet-accessible bioinformatics tools, methods, and data. You will have the opportunity to extend your understanding of phylogenetic relationships, explore the advantages and disadvantages of different molecular markers, and probe the complexity of establishing relationships between organisms.

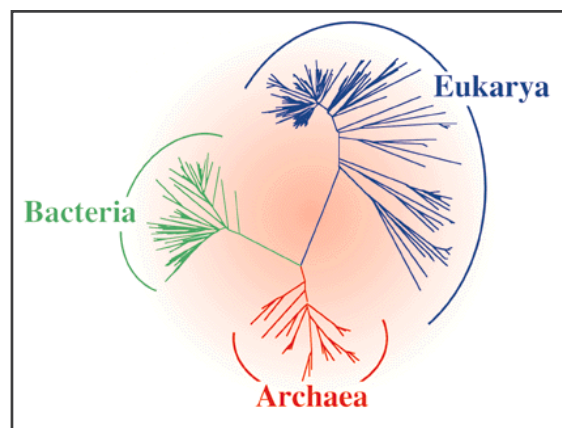


Figure 1. A phylogenetic tree showing the three domains, or superkingdoms, encompassing all life on earth.

## Concepts

- The basics of constructing phylogenetic trees.
- The use of 16S rRNA and other important molecular markers for classification.
- Advantages and disadvantages of traditional vs. molecular methods for classification

## Objectives

- Learn how to retrieve DNA sequence information using the internet.
- Understand how phylogenetic trees are constructed.
- Understand the kinds of information conveyed in a phylogenetic tree.

## Background

Bioinformatics involves the use of computers to mine the vast amount of molecular information available to scientists. This information includes DNA and protein sequences for a wide variety of organisms. Phylogenetics, the development of hypotheses about the evolutionary history of a group of species, is an important application of bioinformatics.

Before the use of sequence data to develop evolutionary trees, phylogenies were based on morphological and metabolic characteristics of organisms. A phylogeny is typically represented as resolved or partially resolved bifurcating tree, consisting of nodes and branches (Figure 2a, 2b). The relative lengths of different branches

A *resolved* tree is one where every taxon has one and only one sister group. Under these conditions each branch *bifurcates*, or splits into two. A *taxon* is a group of organisms that is treated as a unit of analysis in a phylogenetic tree.

represent the genetic similarity, producing a weighted tree (Figure 2a). However, if the investigator is interested only in the respective groupings of species, sometimes the branches are shown as being of (arbitrary) equal length, yielding an unweighted tree (Figure 2b). If we know the root of an unweighted tree, we can draw it as a *cladogram*; like an unweighted tree, a cladogram's branch lengths are arbitrary. The numbers represent species which are nodes. There are also nodes where the lines join. The lines are branches.

Figure 2a. An unrooted weighted tree.

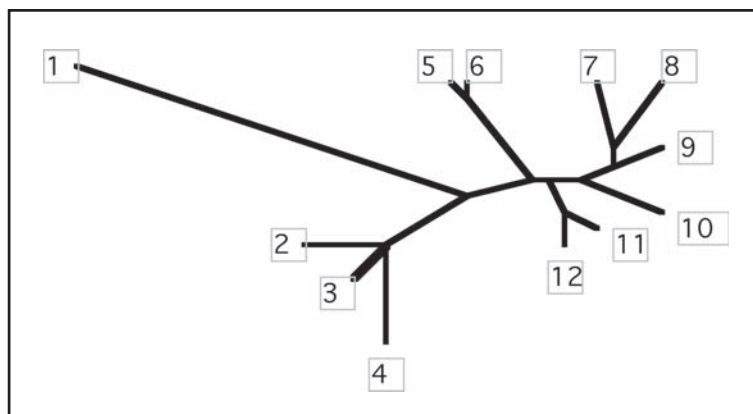
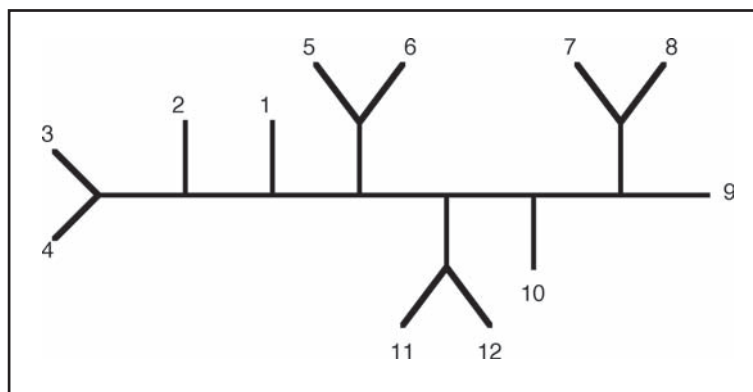


Figure 2b. An unrooted unweighted tree.



Evolutionary tree building seeks to find the tree that provides the most biologically sensible explanation of how diverse groups of species evolved. There can be many possible bifurcating tree shapes to consider for even a relatively small numbers of species. For example, the 8 taxa represented in Figure 3 below could be organized in 10,395 different unrooted trees.

Some methods of tree building use some secondary criteria to evaluate how well each of the possible tree shapes fit the data. However, an alternative approach that works better for larger numbers of species is to first construct a table containing measures of difference between pairwise comparisons of species (e.g. perhaps count the number of mutations that differ between the sequences of different species). The next step then uses the values in the table to determine the order in which species are clustered into a tree (Figure 3).

This procedure produces just one tree, and does not evaluate this tree against other possible trees, which may be almost as well supported by the data.

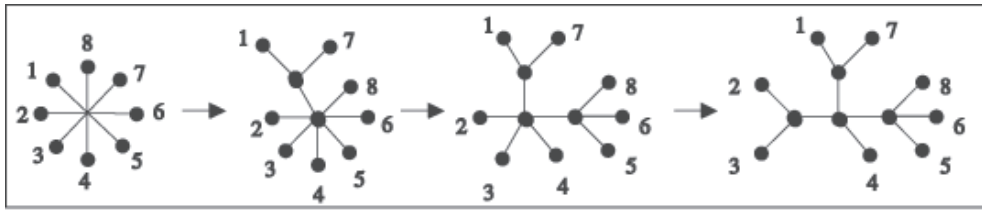


Figure 3. Phylogenetic tree building by pairwise comparisons. The starting assumption represented in the left-most tree is that all of the species are equally similar to one another. As different characters are considered, small groups of species are grouped, resulting in a tree that displays a hypothesis about how the species may be differentially related to one another (i.e., species 2 and 3 may be more closely related to each other than either of them is to species 1 or 7.)

There are several factors to consider when constructing a tree:

- Does this tree represent real relationships? Does it make sense?
- Can we expect the patterns in the data to produce a single unique tree?
- Can we infer a species phylogeny from the comparative analysis of only small sections of the genomes?
- Will protein or DNA or rRNA sequences be better for inferring phylogeny?
- Is it valid to compare homologous sequences if they are from different genome compartments (e.g. from chloroplasts, mitochondria, nuclei, nucleomorphs)?
- How far back in history can we reliably infer phylogeny from sequence data?
- What are the assumptions of the different tree-building methods? Which ones would be best? How would I decide?

## Bioinformatics Tools

The basic resources consist of a bank of data, in our case RNA sequences, a program to retrieve sequences that we wish to use, and a program to compare the similarities and differences between sequences. These programs also display the comparisons of sequences in the form of a tree diagram. There are several online resources that provide access to these tools. For this activity we will use the *Biology Workbench*, developed at the National Center for Supercomputing Applications (NCSA). Note that simply generating the tree is not the end of the story. Interpretation is a time-consuming and extensive process, so you should expect to spend most of your time interpreting your trees.

The RNA sequences for this activity are stored in the GenBank database, an annotated collection of all publicly available DNA sequences. The GenBank records are organized into a variety of smaller databases, such as the GBBCT database that you will be searching. We will use the *Biology Workbench* to access the “Ndjinn” procedure. Using Ndjinn, we will retrieve sequences from GenBank databases by performing a text search for exact matches anywhere in the data record. We can search either by the sequence’s GenBank accession number or by other characteristics such as species and gene names. In either case, searching with Ndjinn will return all records that contain the search text. Each record label in your search results will contain that record’s accession number, so knowing the accession number of the particular sequence you’re interested in can make it much easier to identify that sequence.

Before the advent of molecular phylogenetics, relationships among organisms (and trees that summarize those relationships) were inferred from the organisms' morphological and metabolic characteristics. Now that the tools and data for analyzing DNA and protein sequences are available, it's also possible to examine the genetic relationships between microbes. In cases where molecular analysis yields trees similar to those obtained by non-molecular methods, we can gain added confidence that both data sets correctly describe the species' evolutionary history. In other cases, molecular and non-molecular trees may differ significantly, suggesting that processes such as convergent evolution may be occurring.

### Overview of Operations

1. Choose the organisms that you will use to construct your phylogenetic tree.
2. Retrieve the sequences for your organisms.
3. Generate a phylogenetic tree from your database of sequences.

### Selecting your organisms

The first step is choosing a diverse set of bacteria for your phylogenetic analysis. The 1984 edition of *Bergey's Manual of Systematic Bacteriology* (familarly, *Bergey's*) is a four-volume reference that groups bacteria using primarily morphological and metabolic criteria (the 1984 edition is the last edition that does not extensively use molecular criteria). You can reasonably use the categories of species in each volume as the basis for building a morphological/metabolic tree.

Table 1 lists representative species from each major group in *Bergey's*, plus a eukaryotic example, *Saccharomyces cerevisiae*, for a total of eight groups. Choose at least two examples from each bacterial group, for a total of 14 species. You will also need to include *Saccharomyces cerevisiae*, the eukaryotic example, so you will end up with a total of 15 organisms for analysis. Enter the information for your organisms into the Species Data Table (Table 2). (The Species Data Table is also available on the *Microbes Count!* CD.) The GenBank numbers for the additional species in each group are not listed in Table 1 but you will learn how to find them in a later step.

Table 1. Representative species from each major group in *Bergey's Manual of Systematic Bacteriology*.

Volume 1A (Gram-negative bacteria)	
<i>Escherchia coli</i>	GBBCT #174375
<i>Helicobacter pylori</i>	GBBCT #402670
<i>Salmonella typhi</i>	GBBCT #2826789
<i>Serratia marcescens</i>	GBBCT #4582213
<i>Treponema pallidum</i>	GBBCT #176249
Additional species: <i>Agrobacterium tumefaciens</i> , <i>Bordetella pertussis</i> , <i>Thermus aquaticus</i> , <i>Yersinia pestis</i>	

Table 1. continued

<b>Volume 1B (Rickettsias and endosymbionts)</b>	
<i>Bartonella bacilliformis</i>	GBBCT #173825
<i>Chlamydia trachomatis</i>	GBBCT #2576240
<i>Rickettsia rickettsii</i>	GBBCT #538436
Additional species: <i>Coxiella burnetii</i> , <i>Thermoplasma acidophilum</i>	
<b>Volume 2A (Gram-positive bacteria)</b>	
<i>Bacillus subtilis</i>	GBBCT #8980302
<i>Deinococcus radiodurans</i>	GBBCT #145033
<i>Staphylococcus aureus</i>	GBBCT #576603
Additional species: <i>Bacillus anthracis</i> , <i>Clostridium botulinum</i> , <i>Lactobacillus acidophilus</i> , <i>Streptococcus pyogenes</i>	
<b>Volume 2B (Mycobacteria and nocardia)</b>	
<i>Mycobacterium haemophilum</i>	GBBCT #406086
<i>Mycobacterium tuberculosis</i>	GBBCT #3929878
Additional species: <i>Mycobacterium bovis</i> , <i>Nocardia orientalis</i>	
<b>Volume 3A (Phototrophs, chemolithotrophs, sheathed bacteria, gliding bacteria)</b>	
<i>Anabaena sp.</i>	GBBCT #39010
<i>Cytophaga latercula</i>	GBBCT #174236
<i>Nitrobacter winogradskyi</i>	GBBCT #402722
Additional species: <i>Heliothrix oregonensis</i> , <i>Myxococcus fulvus</i> , <i>Thiobacillus ferrooxidans</i>	
<b>Volume 3B (Archeobacteria)</b>	
<i>Methanococcus jannaschii</i>	GBBCT #175446
<i>Thermotoga subterranea</i>	GBBCT #915213
Additional species: <i>Desulfurococcus mucosus</i> , <i>Halobacterium salinarium</i> , <i>Pyrococcus woesei</i>	
<b>Volume 4 (Actinomycetes)</b>	
<i>Actinomyces bowdenii</i>	GBBCT #6456800
<i>Actinomyces neuii</i>	GBBCT #433527
<i>Actinomyces turicensis</i>	GBBCT #642970
<b>Eukaryotic representative</b>	
<i>Saccharomyces cerevisiae</i>	GBPLN #172403

Table 2. The Species Data Table

<b>Volume 1A (Gram-negative bacteria)</b>	
	GBBCT #
	GBBCT #
<b>Volume 1B (Rickettsias and endosymbionts)</b>	
	GBBCT #
	GBBCT #
<b>Volume 2A (Gram-positive bacteria)</b>	
	GBBCT #
	GBBCT #
<b>Volume 2B (Mycobacteria and nocardia)</b>	
	GBBCT #
	GBBCT #
<b>Volume 3A (Phototrophs, chemolithotrophs, sheathed bacteria, gliding bacteria)</b>	
	GBBCT #
	GBBCT #
<b>Volume 3B (Archeobacteria)</b>	
	GBBCT #
	GBBCT #
<b>Volume 4 (Actinomycetes)</b>	
	GBBCT #
	GBBCT #
<b>Eukaryotic representative</b>	
<i>Saccharomyces cerevisiae</i>	GBPLN #172403

- Assuming that Bergey's classification accurately reflects microbial evolutionary relationships, sketch an unrooted tree for the species in your Species Data Table. You will need to make and justify judgments about which major groups are most closely related and which are more distant.

### Online access to sequence databases and alignment and tree-building software.

The instructions below contain the basic information that you will need to use the *Biology Workbench*. Please see the "Orientation to the *Biology Workbench*" on the *Microbes Count!* CD for a broader overview of what the *Biology Workbench*

is and how it is organized. You may also want to look at the “Proteins: Historians of Life on Earth” activity in Chapter 6 of this book.

### Entering the *Biology Workbench*

Go to <<http://workbench.sdsc.edu>>. If you do not already have a *Biology Workbench* account, click on the Click Here hyperlink to set up an account. Fill out the account information and click the Submit button. Once you have established an account, click on the link Enter the *Biology Workbench*, enter your user ID and password, and begin a new session.

*When working in the Biology Workbench, avoid using the browser's Back button; instead, use only the navigational buttons within the Workbench.*

If you leave the *Biology Workbench*, use the following directions to continue your unfinished work: Log into the *Workbench*. Click on “Session Tools.” Then highlight “Resume Session” in the scroll box. Click “Run.” The name of your folder should appear near the top of your screen.

### Building a sequence database

1. You are now ready to begin retrieving the sequences for your organisms. Click on the Nucleic Tools button near the top of the page.
2. Highlight Ndjinn – Multiple Database Search and click Run. This will bring up a page that gives you access to the many databases available through *Biology Workbench*. To do a search you will need to indicate what you are searching for and which database you wish to search.

In this project, we will use sequence data for the small protein subunit of ribosomal RNA (rRNA) from the species listed in Table 1. Ribosomal RNAs are labeled based on their sedimentation rates (S), which relates to their size. The small subunit in prokaryotes involves a 16S rRNA while eukaryotes have a slightly larger 18S rRNA. In retrieving sequence data for yeast, therefore, we will search for the 18S gene within the GBPLN database (GenBank Plant Sequences, which includes the fungi and algae). Sequence data for the 14 bacterial species, by contrast, will involve searching for the 16S gene within the GBBCT database (GenBank Bacterial Sequences).

3. Start by retrieving the gene sequence for the yeast.
  - Type in “*Saccharomyces cerevisiae* AND 18S”. This is the sole representative of Eukaryotes.
  - Select the GBPLN database. (You have to scroll down for this option.)
  - Click on Show 10 hits. Drag down and select Show 50 hits.
  - Click on Search. You will see a list of choices. Scroll down until you see gbpln:172403. Highlight this line and click on the Import Sequence(s) button located at the end of the first line of the interactive box. This will import the gene sequence for this microorganism.

Alternatively, since you already know the GenBank number for the sequence you want, you can instead base your search on this number rather than on the species and gene names. Both this approach and the one described above will turn up multiple records containing the search text; changing the search option to Begins with rather than Contains can help reduce the number of spurious hits.

4. Now you will follow the same procedure to retrieve the sequences for the rest of your microorganisms from GenBank. Repeat the following steps for each of the bacterial species that you entered into your Species Data Table:
  - Highlight Ndjinn and click Run.
  - Since these are bacterial species, select the GBBCT (GenBank Bacterial Sequences) database.
  - Type in the name of your organism but this time use the 16S rather than the 18S gene.
  - Click on Search. Scroll down until you see the entry for your bacteria. (If the organisms you chose from Table 1 did not have GenBank numbers, this is where you will find their GenBank numbers. Enter the number into your Species Data Table now.) Highlight the entry and click on the Import Sequence(s) button.

At the end of your searches, you should have 14 bacterial sequences and one yeast sequence.

### Constructing a tree

Now you are ready to generate an unrooted tree.

5. Conducting an alignment using ClustalW – Multiple Sequence Alignment tool
  - Highlight Select All Sequences in the scroll box. Click Run. All the boxes in front of the organism names should be checked.
  - Scroll down and highlight CLUSTALW – Multiple Sequence Alignment. Click Run. The ClustalW page will appear. There are a number of different settings on this page; for this analysis you can simply use all of the default settings.
  - Click Submit. The screen will go blank and you may have to wait several minutes. Wait until a screen titled “CLUSTALW” with “Sequence alignment” appears.

Scroll down to examine the DNA sequences and how they align with each other. Make sure that the ends of the bacterial sequences do not have gaping holes (more than 20 or 30 bases). If a few do not align well, representing incomplete sequences or sequences beyond the rRNA coding region, you should delete them from the Ndjinn field and go back and import other sequences until you find some that align better. Otherwise your comparison will be invalid.

6. Scroll further down and you will see your tree! It should be similar to the tree in Figure 3.

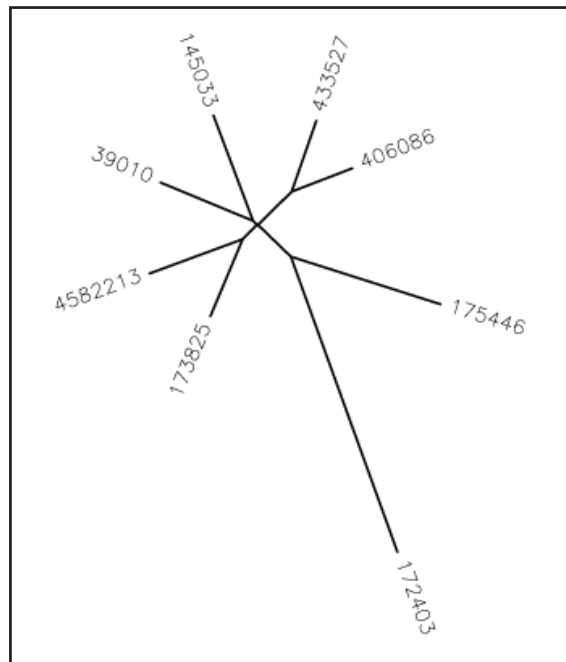


Figure 3. A weighted unrooted tree including a single species from each group in Table 1. Note that the longest branch is associated with the eukaryotic outgroup (*S. cerevisiae*, GB#172403), implying that of all sequences present, this sequence has the lowest genetic similarity to the others. Species labels correspond to GenBank accession numbers; in analyzing your own tree, you may wish to add species names (see text).

### Saving your tree for further analysis

7. You will need to copy your tree into a word processing application so that you can study it. The exact procedure for doing this will vary depending on the type of computer you are using and on your browser. In general, click on the tree image and hold until a menu appears. Highlight Copy this Image and release.
8. Now open Microsoft *Word* and open a document. You may want to choose the document that already has your Species Data Table in it. Choose Paste from the Edit menu and your tree will appear. You may want to adjust the size of your tree by selecting the tree image on one of the lower corners and dragging. Be sure to give your tree a label (such as Figure 1) and include a short description.
9. Your tree is labeled using the GenBank numbers for each species. It will be easier for you to interpret your tree if you add the species name. Look up each GenBank number in your Species Data Table and write the corresponding species name beside the number. You can do so by choosing TextBox under the Insert menu. Click where you want the textbox to be located and type the name of the species. Drag the textbox next to the corresponding number in your tree.

#### Exercise 1: Traditional classification with the *Bergey's* manual

- Discuss the similarities and differences between the tree you sketched earlier in the activity and the computer generated trees. How do the morphological/metabolic relationships in the former relate to the molecular relationships identified in the latter?

**Exercise 2: Molecular classification**

One way to test the robustness of a phylogenetic tree is to draw a new tree after adding or subtracting one or more species from the analysis. This procedure may drastically change trees that aren't well-supported by the data, while leaving better-supported trees relatively unaltered. In the next two exercises, you will employ this technique to evaluate your tree.

- A. Based upon the molecular tree you've generated previously, predict what the tree might look like if you added two more bacterial species of your choice from Table 1. Where do you expect these new species to appear on the tree?
- B. Add the two bacterial species you chose in part A. Use the *Biology Workbench* to generate a tree. (Save a copy of the tree to your Word document as Exercise 2.) Does this agree with what you expected? Does it make sense? Why or why not?
- C. Which of the morphological and metabolic groups from *Bergey's* seem to be consistent with the molecular data? Which aren't? Use colors to indicate the organisms within each major group.

**Exercise 3:**

- A. Based upon your tree from Exercise 2, predict what the tree might look like if you removed one of the microorganisms listed in your Species Data Table.
- B. Remove the microorganism you chose in part A. Use the *Biology Workbench* to generate a tree. (Save a copy of the tree to your Word document as Exercise 3.) Does this agree with what you expected? Does it make sense? Why or why not?
- C. Did any microorganisms change their positions? Do any of them still stay together? Explain.

**Critical thinking questions**

On Taxonomy

- *E. coli* and *S. typhi* are closely related based on their positions in the tree, which indicates that they have similar genomic makeup. Does this mean they also have similar metabolic characteristics?
- *B. subtilis* is a spore former. How does spore formation benefit a microbe? How can you classify spore formers morphologically? What if the spore is an endospore?
- *S. marcescens* produces red pigments which are not expressed at all temperatures. Is this a reliable characteristic for identification under all conditions? How could pigment production be used as a reliable characteristic?

- Both *Mycoplasma* and *Haemophilus* can cause respiratory diseases. Does this piece of information help scientists determine their phylogenetic relationship?
- *E. coli* and *E. coli* O157:H7 are the same species. Why do they behave so differently? What kinds of genetic differences might be responsible? How could you test these hypotheses?

#### On Phylogenetics

- Why do we use 18S rRNA information for yeast and 16S for all the other prokaryotes?
- Why do we need to classify organisms? Why are phylogenetic trees important?
- What are important characteristics for classification? How would you choose? How are you going to get the information on the characteristics?
- Can you predict the metabolic characteristics of your unknown based on its position on the tree?
- You are the scientist on the expedition to the recently discovered planet of Lebesamin. Considering that Lebesamin is over 6 parsecs from Terra and given the atmospheric differences between the two planets, there is remarkable similarity in the genetic makeup of the life forms on the two planets. You are responsible for collecting and identifying all microorganisms on the planet. Using the skills you have developed through this exercise, write up a proposal for classifying the organisms.

### Web Resource Used in this Activity

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

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.

### Additional Resources

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

##### Text

A copy of this activity, formatted for printing

“Species Data Table” in MSWord format

“Orientation to the *Biology Workbench*”

**Related *Microbes Count!* Activities**

Chapter 2: Searching for Amylase

Chapter 4: Molecular Forensics

Chapter 4: Exploring HIV Evolution: An Opportunity for Research

Chapter 6: Proteins: Historians of Life on Earth

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 VI: Microbial Evolution

**Relevant Textbook Keywords**

Archaea, Bacteria, Eukarya, Evolution, Nucleic acids, Phylogenetic relationships, rRNA

**Related Web Sites** (accessed on 2/20/03)

*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**

Holt, J. G., Editor-in-Chief (1984) *Bergey's Manual of Systematic Bacteriology, Volume 1-4*. 1st Edition. Williams & Wilkins: Baltimore.  
<http://www.cme.msu.edu/bergeys/pubinfo.html>

**Figure and Table References**

Figure 1. Courtesy Mitchell Sogin, Josephine Bay Paul Center, Marine Biological Laboratory

Figure 3. Modified from *Biology WorkBench* (<http://workbench.sdsc.edu>)