Proteins: Historians of Life on Earth

Garry A. Duncan, Eric Martz, and Sam Donovan

Video VI: Microbial Evolution

Introduction

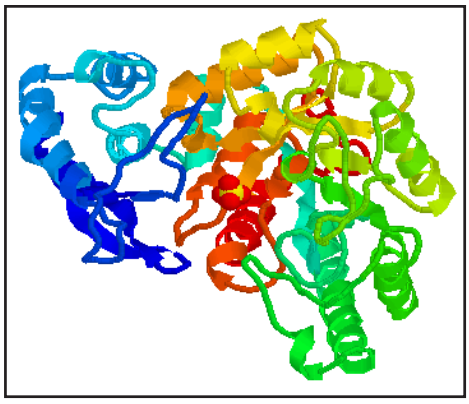

Figure 1. Yeast enolase (4ENL) structure. The helices and sheets are displayed as ribbons and the remainder of the molecule is displayed only as a backbone trace. The coloring shades from blue (amino terminus) to red (carboxyl terminus). There is a sulfate molecule, displayed as a CPK structure in the active pocket of the enzyme. (A color version is available on the *Microbes Count!* CD.)

Prior to the 1980's, one of the most commonly accepted taxonomic hypotheses in biology was that all organisms belonged to one of two domains: (1) the eukaryotes, which included organisms whose cells contain a well-formed nucleus; and, (2) the prokaryotes, which included unicellular organisms whose cells lacked a nucleus, such as the bacteria. In recent years there has been a fundamental rethinking of how to organize the diversity of life. Recent molecular evidence has led to a new hypothesis—that the prokaryote domain is actually comprised of two distinct domains. Some bacteria-like organisms look like normal bacteria but may have had a distinct phylogenetic history. Consequently, these bacteria-like organisms may comprise a distinct domain, given the name Archaebacteria, or more simply, Archaea. The name reflects an untested conjecture about their evolutionary status. Recent phylogenetic evidence suggests that the Archaebacteria may be at least as old as the other major domains; hence, it now seems possible that the most recently categorized group of organisms may actually be the oldest. It is important to note that not all scientists agree with the three domain hypothesis. The bibliography section contains some suggested reading on this debate.

Changes in the nucleotide sequence of DNA and amino acid sequence in proteins can be thought of as molecular fossils. That is, these changes act as historical records of evolutionary events and give us clues about the relatedness of different species in much the same way that changes in morphological characters, preserved in the form of fossils, give us clues about change over time. The extraordinary growth of sequence databases, along with the development of tools to explore and mine these databases, has radically enhanced the ability of biologists to uncover the patterns of organic evolution that have occurred throughout the history of life on Earth.

The following investigations can act as a springboard for you to pose evolutionary questions that might be answered by analyzing molecular data. To accomplish this end, it is important to have a user-friendly, web-based interface that enables you to access DNA and protein databases, perform alignments and construct phylogenetic trees. The *Biology Workbench* (hereafter known simply as the Workbench), developed at the National Center for Supercomputing Applications (NCSA), provides this user-friendly, web-based interface.

Investigation 1: Explorations in Evolution Through Protein Sequence Alignments and Phylogenetic Tree Construction

Objectives: (1) Gain experience using bioinformatics tools and databases, primarily through the *Biology Workbench*. (2) Use protein sequence data and

analysis tools to evaluate the two hypotheses described above regarding different ways to classify the domains of life. In order to accomplish objective 2, a ubiquitous protein must be selected. For this investigation you will examine and compare the protein sequences of enolase, an enzyme involved in the last stage of glycolysis during which 3-phosphoglycerate is converted into pyruvate and a second molecule of ATP is formed. Enolase is found in all organisms, because they all utilize glycolysis to produce ATP for metabolism. You will compare the amino acid sequences of enolase from the seven species in Table 1 along with several species of your own choosing.

Table 1. Species information and web sites.

Species	General Information
Methanococcus jannaschii	Archaebacterium; thermophile (48-94 C); strict anaerobic that lives at pressures of over 200 atmospheres; autotroph that gets its energy from hydrogen and carbon dioxide producing methane. Source: http://jura.ebi.ac.uk:8765/ext-genequiz/genomes/mj/
Pyrococcus horikoshii	Archaebacterium; hyperthermophilic (optimal growth at 98 C, pH 7.0 and NaCl concentration 2.4%). Source: http://jura.ebi.ac.uk:8765/ext-genequiz/genomes/ph0004/
Escherichia coli	Bacterium; Gram negative; rod-shaped; facultative anaerobe; common inhabitant of the gut of warm blooded animals. Source: http://jura.ebi.ac.uk:8765/ext-genequiz//genomes/ec0005/index.html
Bacillus subtilis	Bacterium; Gram positive; rod-shaped; aerobe; nonpathogenic bacterium commonly found in the soil. Source: http://jura.ebi.ac.uk:8765/ext-genequiz//genomes/bs0005/index.html http://www.biojudiciary.org/glossary/index.asp?flt=b
Saccharomyces cerevisiae (yeast)	Unicellular eukaryote; fungus; economically important microbe because it has the ability to ferment glucose into ethanol and CO_2; its biochemistry and genetics are well known. Source: http://jura.ebi.ac.uk:8765/ext-genequiz//genomes/sc0006/index.html
Drosophila melanogaster (fruit fly)	Multicellular eukaryote; commonly known as the fruit fly; feeds on decaying plant matter. One of the best understood genetic organisms. Source: http://jura.ebi.ac.uk:8765/ext-genequiz//genomes/dm0006/index.html http://www.ceolas.org/fly/intro.html
Homo sapiens (human)	Multicellular eukaryote; mammalian primate; omnivore.

Overview of operations

Since you do not have an amino acid sequence of enolase for comparison, you must search for one. Once you have a sequence, you will do the following:

1. Generate a list of proteins with similar sequences by conducting a BLAST search for similar sequences;
2. Select a wide variety of species, representing all the major groups (Table 1, plus one or more selections of your own);
3. Create a multiple sequence alignment using your enolase sequences with ClustalW; and,

4. Construct a phylogenetic tree based on the sequence differences in the alignment. The *Biology Workbench* provides access to all of the databases and tools for these operations.

Using the *Biology Workbench*

The instructions below contain some of the information you will need to use the *Biology Workbench*. Please see the "Orientation to the *Biology Workbench*" document on the *Microbes Count!* CD for a broader overview of what the *Biology Workbench* is and how it is organized.

1. Entering the *Biology Workbench*:

   a. Launch your web browser and go to the following URL for the *Biology Workbench:* http://workbench.sdsc.edu/.

   b. If you have already set up an account on the Workbench, go to Step c now. If, however, this is your first time utilizing the Workbench, then click on the Click Here hyperlink to set up an account. Fill out the account information and click the Submit button and go to Step 2 below.

   c. Click on the hyperlink Enter the *Biology Workbench* 3.2

   d. Enter your user id and password and then click the Submit button.

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

2. Starting a new session or resuming an old session:

   Before you can utilize the Workbench, you need to begin a New session or Resume a previous session, just as you need to begin a new file for word processing or continue a previous file in work processing. In other words, you cannot use the Protein Tools, Nucleic Tools, or Alignment Tools until you have resumed an old session or started a new session. Scroll down the page and click on the Session Tools button.

   a. To start a new session, click (i.e., select) Start New Session in the scrollable window and then click the Run button. On the new web page that appears, you need to name the session (= file) you are about to begin. In this case, we are going to name the session ***Enolase*** since we are going to be conducting a protein search, amino acid sequence alignment and tree construction for enolase. Now click the Run button. The page that now comes up is the same as the one that you were on a moment ago, except that your new session (i.e., Enolase, which is now a file name on a remote server) is now listed with your previous sessions, if you have any. (You may have to scroll down the page to see it.) If the radio button for the Enolase session is not already selected, click it now.

   b. You are now ready to begin searching for amino acid sequences. So, click the Protein Tools button near the top of the page.

3. Selecting a sequence:

   a. Now you are in the Protein Tools window. You need access to protein databases in order to perform your search. To do so, select Ndjinn – Multiple

Database Search in the scrollable window and then click the Run button. When the new web page appears, you need to indicate what protein you are searching for and what database(s) you wish to search. For this investigation, type enolase into the blank field following the word Contains. Now select the PDBFINDER database in the scrollable window, and click the Search button at the top of the web page. (You are selecting this database because the 3D structures are known for all of the proteins in this database.)

The Results page indicates that you have matched 15 unique records. (The number of unique records may be larger than 15 records since new records are being added on a continuous basis.) Click the box in front of the one that says: PDBFINDER:4ENL (from yeast). This will be the enolase sequence in which we will anchor the rest of our searches. (For later use in *Protein Explorer*, you will need to write down the PDB id number, which is the four digit alpha-numeric code—in this case, 4ENL.) Now click the Import Sequence(s) button, and continue to Step b.

b. Before going further, you should find out more about this enolase molecule. Click the box in front of the protein, select View Database Records of Imported Sequences from the scrollable window and then click the Run button. In the web page that now appears, select the Formatted radio button and then click the Show Record(s) button. You will find a wealth of information about this protein, including its amino acid sequence, its enzyme code number, citations, etc. You can even view the molecule in 3D (upper right of page). After viewing, click the Return button at the bottom of the web page.

4. Searching for records with similar sequences using *BLASTP* (Basic Local Alignment Search Tool for Proteins):

   a. If it isn't already selected, click the box in front of pdbfinder:4enl_carbon-oxygen lyase.

   b. In the scrollable window, select BLASTP – Compare a PS to a PS DB, and then click the Run button. Select the database SwissProt in the scrollable window. As you scroll to the bottom of the web page, you will note that you can specify a number of search criteria. For our purposes, we can use all of the default selections. At the bottom of the web page, click the Submit button. The BLASTP tool will find other similar protein sequences in the SwissProt protein database

5. Selecting records for alignment

   a. Scroll down the BLASTP results page. For this activity, select the following six records (the yeast record, which you have already selected, acts as the seventh record and does not need to be selected again). In addition, select at least one or more species of your choice, and make a mental prediction (hypothesis) about where you think your species will fit on the phylogenetic tree.

Locus Name	Organism
ENO_DROME	*Drosophila melanogaster*
ENOB_HUMAN	Human
ENO_METJA	*Methanococcus jannaschii*
ENO_PYRHO	*Pyrococcus horikoshi*
ENO_BACSU	*Bacillus subtilis*
ENO_ECOLI	*E. coli*
Your sequences:	

b. Scroll back up to near the top of the web page and click the Import Sequences button. This action will import the amino acid sequences of all of the records (= sequences) you have selected. (The yeast record with which you started was already imported.) In the next step of the investigation, you will align the sequences.

6. Conducting an Alignment Using CLUSTALW – Multiple Sequence Alignment tool

a. Click the boxes of all of the records you wish to align, including the yeast record.

b. Select the CLUSTALW tool in the scrollable window. Now click the Run button. The CLUSTALW page appears, which contains all of the different settings you can alter when running this analysis. For this investigation, you will use all of the default settings. So, just click the Submit button. It will take the computer a few moments to develop the alignments.

c. Now scroll down the CLUSTALW Multiple Sequence Alignment page and see the alignments. At the top of the alignment, note the color coding key. The first alignment group contains the alignment for amino acids 1-60, while the second alignment group contains the alignment for amino acids 61-120, etc. You will also note some dashes within the alignment, which indicate missing amino acids. Scroll down below the alignment and you will note an unrooted tree. (We will reconstruct this tree in Step 7 below.) Continue scrolling and you will find additional information, including the number of amino acids in the enolases for each species. To save this alignment as a file for future viewing and for further analysis, scroll to near the bottom of the page, click the Import Alignments button.

7. Constructing a tree

a. In the Alignment Tools page, click the box in front of the CLUSTALW-Protein file of the aligned sequences. (You may have to scroll down the web page in order to see this box.) Selecting this box acts to select the entire list of records that have been aligned.

b. Select the *DRAWTREE* application tool in the scrollable window, and then click the Run button. (This tool draws an unrooted phylogenetic tree.) The *DRAWTREE* page appears, which contains all of the different settings you can alter. Again, we will use all of the default settings. Click the Submit button. Note that you can print out a copy of the tree, which should look similar to the one in Figure 2 below.

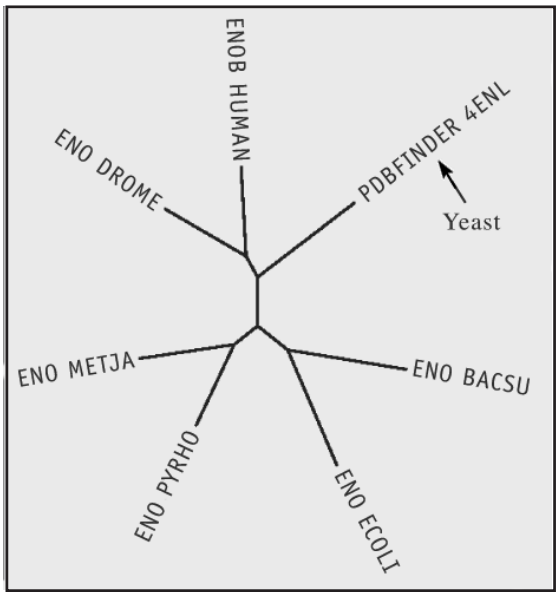

Figure 2. Unrooted tree. The branch points in the tree are called nodes, while the lines are called branches. The length of the branch is a direct measure of the amount of change that has occurred in this protein. Note that the two species of bacteria, *Bacillus subtilis* and *E. coli*, share a common branch of the tree; the two archaea, *Methanococcus jannaschii* and *Pyrococcus horikoshii*, share a common branch; and, the three eukarya—*Saccharomyces cerevisiae* (yeast), *Drosophila melanogaster*, and human—share a common branch.

Questions for discussion:

1. Where would you expect *Methanococcus* and *Pyrococcus* to split off of the unrooted tree if the two domain (i.e., Bacteria and Eukarya) hypothesis is correct?
2. Where would you expect *Methanococcus* and *Pyrococcus* to split off of the unrooted tree if the three domain (i.e., Bacteria, Archaea and Eukarya) hypothesis is correct?
3. Did the species you added to the investigation appear on the tree where you predicted?
4. Which hypothesis does the tree in Figure 1 support?

8. Exporting sequence alignments in Fasta format

Click the Return button at the bottom of the web page in order to return to the Alignment Tools web page.

a. Click the box in front of the CLUSTALW-Protein file of the aligned sequences. In the scrollable window, select View Aligned Sequence(s) and then click the Run button. This will take you to the View window.

b. Scroll down the web page to view the sequences.

c. If you wish to import the sequence into *Protein Explorer*, you will need to change the format to Fasta. The Format window will probably say MSF; click the arrow for the dropdown window to open; select FASTA. This selection will automatically change the format to FASTA. Once the format has been changed, you can either Save the sequences to a file or Copy/Paste them into *Protein Explorer*. If you are going to continue on to Investigation 2 right now, then the Copy/Paste method is the easiest. Simply highlight and copy the entire group of sequences, including the > sign and the enzyme/species names, so that they can eventually be pasted into *Protein Explorer* (Step 3.c in Investigation 2 below).

Investigation 2: Visualizing the Evolution of Protein Structure in 3D

Objectives: (1) Become familiar with the many capabilities of *Protein Explore*r (PE); (2) be able to load a 3D structure of a protein into PE for viewing; (3) be able to place aligned sequences (Fasta format) into PE; and, (4) be able to visualize in 3D the evolutionary changes within the protein structure.

1. Opening *Protein Explorer (PE)*

Launch *Protein Explorer* in a new browser window at:
http://proteinexplorer.org.

On this web page, note all the different ways in which you can learn about PE, including a 1-2 hour tour that will give you a better idea of all of the capabilities of PE.

2. Loading the protein's 3D structure

a. Now go back to the PE home page (http://proteinexplorer.org). Locate the field where you enter the PDB Identification code. Type in the PDB ID for the enolase from yeast (i.e., 4ENL). (This is the code you were instructed to write down in step 3.a. in Investigation 1.) Now click the Go button. Be very patient. This takes awhile, particularly if this is the first time your computer has used PE. Click OK on any windows that come up. The 3D structure of the protein is now shown, but it does NOT indicate where there are any amino acid changes. This won't happen until the amino acid alignments you copied while working in the *Biology Workbench* (Step 8 in Investigation 1) are pasted into PE in step 3.c. below.

b. You should now be in the FirstView frame. Click the item Explore More with QuickView.

c. You should now be in the QuickViews frame. Near the bottom of this frame, click the hypertext where it says: Go to Advanced Explorer.

3. Pasting aligned sequence from *Biology Workbench* into *Protein Explorer*

a. You should now be in the Advanced Explorer frame. Click the item that says MSA3D: Multiple Sequence Alignment Coloring.

b. In the MSA3D Procedure frame, click item 5, which says Paste the alignment into the MSA3D ALIGNMENT FORM.

c. You are now in the MSA3D Alignment Form window. Place the cursor in the Alignment Box and Paste the alignment sequences (*Fasta* format) that you copied from the *Biology Workbench* (i.e., step 8 in Investigation 1 above).

d. Now copy/paste the yeast sequence (i.e., 4ENL), which is the sequence for which the 3D structure is known, from the Alignment Box into the 3D Sequence Box. The yeast sequence (i.e., 4ENL) will be in both the Alignment Box and the 3D Sequence Box.

e. Click the Color Alignment & Molecule button just below the 3D Sequence Box. In a moment, a new browser page will open, showing the color-coded alignments for all of the species. (This process may take several moments, so be patient.) The legend for the color codings is indicated at the top of the page. For example, green indicates that an amino acid at a specific position is identical for all species.

4. View the evolutionary changes of the protein in 3D:

a. If your screen is large enough, you will see the 3D structure rotating on another web page. Click on that web page to bring it to the front.

b. The backbone trace of enolase has been colored as indicated. The results are more easily appreciated when the full structure, including side chains, is shown with all atoms "spacefilled" (to van der Waals radii). In the MSA3D Result frame, click on each of the three links—Identical, Similar, and Different (i.e., the first three bullets)—to spacefill all categories. The red balls are water molecules. Click the Water button so that the red balls (water oxygens) are hidden, enabling you to clearly see the protein itself. The 3D model is showing the several billion year evolutionary history of enolase.

c. Point to the 3D model, click and hold down on the mouse button and move the mouse. This action allows you to rotate the molecule. The catalytic site is marked by a brown zinc (Zn) ion (nearly buried) and an easily spotted red-and-yellow sulfate ion that happens to be bound there. Note that the active site is entirely green (complete identity), showing billions of years of evolutionary conservation, while the peripheral region of the molecule is almost entirely yellow because of amino acid substitutions.

Questions for discussion:

1. What does the complete conservation of the amino acids in the active site suggest to you?
2. Why do you think the peripheral region of the enolase molecule has varied so much over time in contrast to the stability of the active site?
3. Are there other regions on the enolase molecule highly conserved, besides the active site? (Hint: are there conserved regions on the peripheral part of the molecule?) What might be the role of those regions?

4. Do other enzymes in glycolysis show similar results? (To answer this, you would have to repeat the investigation, substituting other glycolytic enzymes for enolase.
5. What other proteins might also be shared by the taxa used in the above investigations? You could investigate these proteins to determine whether or not they show the same pattern of evolution as enolase.

Protein Explorer's MSA3D is used in this activity because it helps you to understand the steps involved in coloring a protein molecule to show the rate at which each amino acid evolves. A much more automated, and therefore easier, method of identifying conserved and rapidly mutating residues for any 3D protein structure is the *ConSurf Server* (http://consurf.tau.ac.il), which also uses a more sophisticated and robust method for calculating conservation scores for each residue. It is less sensitive to the choice of sequences in the alignment than is MSA3D. Take a look at the ConSurf Gallery at its website if you are interested.

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 occuring 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

"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: 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 VI: Microbial Evolution

Relevant Textbook Keywords

Active site, Enolase, Glycolysis, Gram negative bacteria, Gram positive bacteria, Molecular evolution, Mutations, Phylogeny

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

Bibliography

Margulis, L. and K. Schwartz (1998). *Five Kingdoms: An Illustrated Guide to the Phyla of Life on Earth*. W. H. Freeman and Company, New York.

Hagen, J. (1996). Robert Whittaker and the Classification of Kingdoms. In *Doing Biology*. Harper Collins College Publishers. New York.

Figure and Table References

Figure 1. Courtesy Sam Donovan

Figure 2. Modified from *Biology WorkBench* (http://workbench.sdsc.edu)