

## Chapter 6

# Exploring Important Biological Concepts Using Biology Workbench

*Mary V. Ball<sup>1</sup>, Stacey Kiser<sup>2</sup>, Garry A. Duncan<sup>3</sup>*

<sup>1</sup>Department of Biology  
Carson-Newman College  
CNC Box 72044  
Jefferson City, TN 37760  
*mball@cn.edu*

<sup>2</sup>Biology Department  
Lane Community College  
4000 E. 30<sup>th</sup> Ave.  
Eugene, OR 97405  
*kisers@lanec.edu*

<sup>3</sup>Biology Department  
Nebraska Wesleyan University  
5000 St. Paul Ave.  
Lincoln, NE 68504-2794  
*gduncan@nebrwesleyan.edu*

Mary is a Professor of Biology at Carson-Newman College. She did her graduate work in population genetics. Her interests include teacher education, environmental education, and application of the Internet in teaching.

Stacey is a biology instructor starting her sixth full time year at Lane Community College. She received her BS from Oregon State University in 1991 and her MS from the University of Oregon in 1993. She worked as a research assistant for the Workshop Biology program at the University of Oregon for two years. She teaches the general survey for non-biology majors and the introductory zoology sequence for life science majors.

Garry received his BS and MS degrees in Zoology from Arizona State University and his PhD in Genetics from the University of Arizona. He has been on the faculty at Nebraska Wesleyan University for the past 22 years where he is currently a professor of biology and chair of the biology department. His primary teaching and research interests are in genetics, molecular biology and evolution. He has been the recipient of two major teaching awards. His research efforts have focused mostly on the population biology of the fruit fly *Drosophila* and on the black fly *Simulium*.

©2002 Mary V. Ball, Stacy Kiser, Garry A. Duncan

## Using Biology Workbench

**Reprinted From:** Ball, M. V., S. Kiser, and G. A. Duncan. 2002. Exploring important biological concepts using biology workbench. Pages 85-109, in *Tested studies for laboratory teaching*, Volume 23 (M. A. O'Donnell, Editor). Proceedings of the 23rd Workshop/Conference of the Association for Biology Laboratory Education (ABLE), 392 pages.

- Copyright policy: <http://www.zoo.utoronto.ca/able/volumes/copyright.htm>

Although the laboratory exercises in ABLE proceedings volumes have been tested and due consideration has been given to safety, individuals performing these exercises must assume all responsibility for risk. The Association for Biology Laboratory Education (ABLE) disclaims any liability with regards to safety in connection with the use of the exercises in its proceedings volumes.

### Contents

Introduction to Biology Workbench .....	86
Activity I: Comparison of Preproinsulin Sequences Across Species (Ball) .....	92
Introduction for Instructors .....	92
Materials .....	92
Notes for Instructors .....	92
Student Handout.....	92
Acknowledgements.....	96
Activity II: Comparison of Preproinsulin Sequences Across Species (Kiser)...	97
Introduction for Instructors .....	97
Materials .....	97
Notes for the Instructor .....	97
Student Handout.....	98
Acknowledgements.....	102
Activity III: Mining DNA and Bringing Molecules to Life .....	103
Introduction for Instructors .....	103
Materials .....	103
Notes for Instructors .....	103
Student Handout.....	104

### Introduction

There now exist numerous databases containing DNA and protein sequences. The Biology Workbench (<http://workbench.sdsc.edu/>) website offers “one-stop shopping” in the sense that it brings together in one site many of these databases, along with online software for searching the databases, aligning sequences, and creating inferred phylogenetic trees. The Workbench site can be used by students to simulate the types of ongoing research that use sequence databases. In addition, a number of important biological concepts can be developed through accessing and aligning sequence data.

#### Uses of Sequence Databases in Research

There are a number of ways that sequence databases are being used by researchers. A few examples are described below:

### *Constructing Phylogenies*

The Linnean system of classification places species into categories (genus, etc.) but does not attempt to designate the closeness of relationships within a taxon. Sequence alignments across species are being used to infer phylogenies. Some genes that have become widely used for this purpose are the 12S rRNA and the 16S rRNA genes, and the cytochrome c gene.

As an example, molecular data indicate that Asiatic black bears and American black bears are more closely related to each other than to the other species in the genus *Ursus*, brown bears (which includes the grizzly bear) and polar bears (see PubMed abstract number 8744769).

### *Revising Taxonomic Classifications*

The North American elk has been traditionally considered a subspecies of red deer. However, recent comparisons of mitochondrial DNA sequences suggest that the N. American elk is more similar to Sika deer than to red deer (see PubMed abstract 9878235).

### *Classifying Genes into Gene Families*

Gene duplication and divergence are expected over time to give rise to families of genes descended from a common ancestral gene. The comparison of gene organization, protein spatial structure, and amino acid/DNA sequences allows the recognition of potentially related genes. One interesting gene family is the lysozyme/alpha-lactalbumin family, which includes lysozyme c, a bacteriolytic agent in tears, saliva, and milk, stomach lysozyme, a variant in ruminants that breaks down fermenting bacteria (making the products of digestion available to the animal), and alpha-lactalbumin, a protein in milk required for the synthesis of milk sugar (see PubMed abstract 9307874).

### *Developing Plausible Evolutionary Scenarios*

Comparison of genes and proteins across species has been used to attempt to infer which living even toed ungulates (deer, antelopes, pigs, hippos, and giraffes) are most closely related to whales. Data for cytochrome b and milk casein proteins suggest that whales are more closely related to hippos than to other even-toed hoofed animals, but the data for 12S rRNA fail to support this hypothesis (see PubMed abstracts 8752004 and 9159933).

### *Understanding Genetic Diseases*

More than 100 different mutations have been found to cause G6PD-deficiency, an X-linked genetic disorder that is associated with hemolytic anemia. Correlating human mutant amino acid sequences with the amino acid sequences from 42 species revealed that relatively few were at fully conserved sites or a poorly conserved sites, consistent with the prediction that mutations at fully conserved sites would be lethal (for this particular gene) and mutations at poorly conserved sites would not cause disease (see PubMed abstract 1069863).

## **Biology Workbench Tools**

The Biology Workbench site offers a growing list of online software and it is beyond the scope of this introduction to describe the capabilities of every available program or to thoroughly describe the capabilities of any single program. Those that were used in the activities provided in our workshop are briefly described below.

## Using Biology Workbench

### *BLAST*

BLAST stands for “**B**asic **L**ocal **A**lignment **S**earch **T**ool”, but a mental image to use to remember what it does is that of a shotgun blast, in which there are many “hits”, or a dynamite blast, which produces numerous small particles. This program allows you to search selected databases for those sequences that to some degree match the sequence under consideration. Since you must have one sequence to begin with, one approach is to first use the following search program to locate the sequence you want to try to match, so you can copy it rather than having to type it in. There is a BLAST Tutorial available at:

<http://www.ncbi.nlm.nih.gov/Education/blasttutorial.html>

### *Ndjinn*

The name “Ndjinn” appears to be simply a funky spelling of “engine” (and is pronounced the same way). This basic search tool allows you to search selected databases by keyword, such as by enzyme name. Some of the “hits” may not be similar in sequence but only be ones in which the keyword appears somewhere in the record.

### *CLUSTALW*

This program takes 2 or more sequences and aligns them. Conceptually, it “slides” the sequences to best match each other, creating “gaps” if appropriate. There are, however, a variety of possible settings other than the “default settings”, so in practice, using the default settings may not necessarily produce the best match. (For example, tyrosinase from *Neurospora crassa* is supposed to be homologous to tyrosinase from vertebrates, but no homology is apparent from an alignment created using the default settings.)

The output for a CLUSTALW alignment shows the alignment, with some color-coding and symbols designating totally conserved sites, but the colors and symbols do not print out well, especially in black and white. Workbench will automatically produce an unrooted “guide tree” diagram for each alignment, unless you change that setting.

### *BOXSHADE*

This program creates a much more colorful color-coded alignment, starting with the CLUSTALW output. Try changing the font size to 14 and selecting the “ruler” option. To create a transparency, use the right mouse button to select “view image” so all you print is the alignment.

### *DRAWTREE and DRAWGRAM*

These programs produce either unrooted or rooted trees (respectively) from a CLUSTALW alignment. However, in some cases, the trees do not match those produced directly by CLUSTALW. Also, the rooted tree “puts the root somewhere in the middle,” which can result in trees that are drastically different from what you would expect in terms of taxonomy.

### **Using Workbench in Teaching:**

The uses of Workbench vary in terms of how much of the work is done by the instructor and how much by the students. At one extreme, the instructor could use Workbench to create printouts, transparencies, and/or PowerPoint presentations without any student involvement. At the other extreme, students could be responsible for the entire investigation. In between these two extremes, the instructor might provide the sequence data (in a computer file) and have the students “take it from there”.

### **Other Online Sources of Data and Useful Information**

*PubMed* (<http://www.ncbi.nlm.nih.gov/entrez/>)

PubMed allows online searching of the biomedical literature. (You might expect that searching for non-medical info would produce little valuable info, but this is not the case!) In most cases an abstract is available and in some cases there is a link to a free online version of the entire article.

To explore an example of the resources available through PubMed, go to the website and type in the number 2315305. Either click on Go, or press Enter. In the resulting list, click on the authors' names to access the abstract. Clicking on the button labeled "FREE full text article..." will allow you to download the actual article.

In the upper right hand corner of the abstract page, note the choices Related Articles, Books, Protein, Nucleotide, and Link Out. (Not all choices are available for every abstract.) Clicking on Related Articles produces a list of citations to articles that are somehow related to the one displayed (although not necessarily according to what you are interested in). Clicking on Books produces a copy of the abstract, with certain terms "hot-linked". Clicking on any of these takes you to a list of references to the term in the online version of the book *Molecular Biology of the Cell*, 3rd edition. Clicking on Protein or Nucleotide takes you to a list of one or more sequences that the authors have appended to the abstract. These may be entered directly into Workbench (see directions included).

*Online Mendelian Inheritance in Man* (<http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=OMIM>)

In some abstracts, another choice is shown in the upper right hand corner – OMIM. This links to an entry in the online version of Victor McKusick's *Mendelian Inheritance in Man*. The website can also be accessed and searched separate from a PubMed search.

*Human Gene Mutation Database* (<http://www.uwcm.ac.uk/uwcm/mg/hgmd0.html>)

There is no comprehensive list of which genes are included in this database. You just have to use the search feature for the genes you are interested in to determine if data for those genes have been incorporated. For those genes that are included in the database, you can get a detailed list of the known mutations, grouped by type. The output gives the codon number, the "original" *sense DNA* codon, and the mutant sequence, as well as specifying how the mutation alters the amino acid sequence. The three-letter abbreviations for amino acids are used.

*GeneCards* (<http://mach1.nci.nih.gov/cards/index.html>)

According to the site, "GeneCards is a database of human genes, their products and their involvement in diseases." It offers concise information about the functions of all human genes that have an approved symbol, as well as selected others [gene listing].

### **Important Biological Concepts**

In exploring online data and interpreting the results of aligning and comparing sequence data, students are likely to encounter a number of important biological concepts that tend to be mentioned briefly if at all in a general biology course. A few examples are given below:

## Using Biology Workbench

*Cell Structure: Mitochondria contain genes.*

Mitochondria contain both ribosomes and genes. Among the mitochondrial genes are genes coding for 12S rRNA and cytochrome b. Try searching PubMed for “12S rRNA phylogeny” or for “cytochrome b phylogeny” for abstracts that can be used as the basis for a Workbench activity.

*Protein Structure: Tertiary structure is frequently stabilized by disulfide bridges.*

The folding of an amino acid chain into a three-dimensional shape is often stabilized by *cross-linking* between *cysteines*. Have students routinely examine alignments for conserved cysteine residues.

*Genetic Code: Missense mutations may result in a chemically similar amino acid being inserted at that position in an amino acid chain.*

Mutations that result in the substitution of one amino acid for another are termed *missense mutations*. Due to the relationship among codons in the genetic code, many missense mutations result in the substituted amino acid having properties similar to the original one. A BOXSHADE plot color-codes the alignment by the degree of similarity.

*Protein Synthesis: Secreted proteins contain a sequence called the signal sequence.*

Proteins that are secreted are translated by ribosomes that become associated with the rough endoplasmic reticulum. This is accomplished by the presence of an initial sequence of nucleotides in the mRNA that codes for an amino acid sequence called the *signal sequence*. Many protein sequences retrievable with Workbench indicate the signal sequence.

*Protein Synthesis: The DNA sequence that is transcribed is called the antisense sequence.*

The DNA sequence that is transcribed is called the *antisense strand*, while its complementary sequence, not transcribed, is the *sense strand*. The sequence in the *sense strand* matches that of the RNA, except that T occurs in the DNA sequence wherever U occurs in the RNA sequence. In textbooks, it is common for the hypothetical *antisense* sequence to be given and for students to be asked to specify the complementary mRNA sequence, group the sequence into triplet codons, and then look each mRNA codon up in a Genetic Code chart based on mRNA codons. However, DNA sequences in sequence databases generally are the *sense* sequences. Most general biology books use the three-letter abbreviations for amino acids (*e.g.*, MET – methionine). However, protein sequence data uses a one-letter abbreviation system. A Genetic Code chart that is based on *sense DNA* codons and that gives both the three-letter and the one-letter abbreviations is available at <http://psyche.uthct.edu/shaun/Sblack/geneticd.html>.

*Evolution: Gene duplication and differentiation is expected over time to produce families of related genes.*

Unequal crossing over results in one homolog carrying two copies of some genes and the other lacking those genes. Among descendants, duplication homozygotes may occur. Although the duplicate genes are initially on the same chromosome, over time translocation may result in their becoming unlinked. Over time, duplicate genes are expected to diverge in sequence as a consequence of mutation. Genes descended from a common ancestor are said to belong to the same *gene family*. Database records can be used to identify which gene family a gene belongs to and then Workbench can be used to align sequences to observe conserved regions among related genes.

Some mutations may alter the gene to the point that it becomes a nonfunctional *pseudogene*. Others may alter protein function so that a protein with a new function may evolve. (There are other mechanisms that can result in gene duplication, such as the reverse transcription of a messenger RNA molecule, followed by reinsertion of the sequence. Lacking a promoter, this sequence is unlikely to be transcribed and is in this case called a *processed pseudogene*.) Some pseudogene sequences are available and can be compared to the sequence of the corresponding functional gene by using Workbench.

*Enzyme Action: Changes in the active site are likely to affect enzyme function.*

Enzymes function in a lock and key manner in that there is a complementary fit between an *active site* on the enzyme and the substrate. Missense mutations that affect the active site are more likely to alter function than mutations affecting other regions of the molecule, so residues in the active site are more likely to be highly conserved across species. This phenomenon can be verified by using Workbench to determine the positions of highly conserved residues, and then relating these data to the 3-D structure of an enzyme using Protein Explorer, a user-friendly protein-structure exploration program available at <http://www.umass.edu/microbio/chime/explorer/>.

*Genetic Diseases: For many genetic diseases, there are dozens of different alleles that may result in the disease.*

Although many introductory texts create the impression that all persons with albinism have the same genotype, in fact there are dozens of known mutant alleles at the tyrosinase locus that result in albinism in homozygotes. This is true for most genetic diseases, with sickle cell anemia being an exception. The data from the Human Gene Mutation Database can be compared to alignments created with Workbench to demonstrate that mutations that destroy enzyme function tend to occur at highly conserved sites.

### **Some Quirks of Workbench:**

Although multiple users can log on simultaneously to one account, the Workbench developers discourage this, explaining that the site was not designed to work in this fashion. On the other hand, if each student or team of students has an account, there is currently no way to share data among accounts. Also, Workbench saves sequences and CLUSTALW alignments, but it does not save any BOXSHADE plots or trees that you create, so these need to be printed out, saved to a file, or just recreated as needed.

Another quirk relates to doing several alignments in a row by selecting different subsets from the same list of sequences. Although you can check and uncheck boxes in front of each sequence name with a click of the mouse, doing so may result in an output different from what you thought you selected! What you have to do is choose “Deselect All Sequences” from the menu and then start all over to choose another set of sequences.

Another quirk is that there are three alternative versions of the screen display! On the site page, you will note the comment in the banner that states, “click here to toggle between menus and buttons.” There does not appear to be a “default” view, so in a class different computers may be showing different views! Good advice is to decide which view you like, write any directions to fit that view, and make sure everyone using those directions has that view selected.

### **A Final Comment:**

## Using Biology Workbench

The “Full” Workbench site at the San Diego Supercomputer website is the site which offers the greatest number of tools, but is often slow to access. Also, first-timers tend to find it confusing that clicking on Run does not immediately give you the output – another page comes up first that requires you to click Submit. Explaining in advance that this is to be expected and that the justification is that this step allows you to alter the default settings as needed may help avoid confusion. [The Student version of Workbench (<http://inquiry.uiuc.edu/bioweb/perl/bioweb.cgi>) is on a faster site and does not offer the user the option of changing default settings, so overall the same result can be obtained much more quickly.]

### **Activity I: Comparison of Preproinsulin Sequences Across Species (Ball)**

#### **Introduction for Instructors**

In this activity, students use the Student Version of Biology Workbench, which is easier to use, but has limited features. Also, they search for sequences using a keyword instead of using a sequence. For an activity in which students search using a sequence and use the complete Biology Workbench, see Stacey Kiser’s activity, Activity II. (Stacey and Mary independently had the idea to use preproinsulin as the focus for a Workbench activity and independently developed their activities.)

#### **Materials**

This activity requires computers with Internet access.

#### **Notes for the Instructor**

Workbench is designed to run on Netscape and tends to crash with Internet Explorer. Since the Workbench site is subject to being changed, the instructor should work through the activity ahead of time and identify any modifications that need to be made in the instructions. Also, the instructions were written for PCs and may need to be modified slightly for use on Macintoshes. Since Internet access is unpredictable, the instructor may want to print out sample results on transparencies, for use in the event the Internet is not available at the time of the lab.

Even if there are enough computers available for students to work alone, there is a benefit from having them work in pairs. Ideally, being able to print out results in color facilitates completion of the datasheet. If color printers are not available, taping a blank sheet of transparency film over the computer monitor and using a transparency pen to mark the beginning and end of each sequence can simplify data collection.

The handout tells students to include the human sequence and to ignore any sequences that are not specifically for preproinsulin (or insulin precursor). Failure to follow these instructions will complicate interpretation of results.

Have student teams put their data on the board (species compared and calculated percentages) and lead the class to look for patterns in the results related to which species were compared.

### **Student Outline**

#### **Comparing Preproinsulin Sequences Across Species**

**Introduction: (Everything you wanted to know about insulin but were afraid to ask)**



### *The Function of Insulin:*

Insulin is a small protein (polypeptide) hormone consisting of 51 amino acids. It is secreted by groups of cells called beta cells of the Islet of Langerhans in the pancreas. Following ingestion of carbohydrates (or sugars), insulin is secreted by the pancreas in response to increase in blood sugar. Most cells of the body have insulin receptors that will bind the insulin. Once insulin is bound to the cell surface receptor, other receptors that absorb glucose (sugar) from the bloodstream into the cell (which lowers blood sugar levels) are activated. When a pancreas does not secrete insulin, a person becomes diabetic and is often required to take daily injections of insulin.

The importance of insulin to your body is one of energy use. The cells of your body get most of their fuel from glucose and the blood glucose levels are regulated primarily by insulin. The brain uses about 25% of the total body glucose but does not store it. Therefore there must be glucose available in the bloodstream to maintain adequate brain function (hence many of the problems associated with hypoglycemia and hyperglycemia). Insulin also controls the rate of glucose used in muscle, fat and liver cells. Muscle requires insulin to move glucose across its cell membrane. Insulin also promotes the uptake of amino acids and prevents the breakdown of protein. Insulin affects the liver by promoting the uptake of glucose by reducing the rate of glycogen breakdown.

Without insulin, you can eat lots of food and actually be in a state of starvation since many of our cells cannot access the calories contained in the glucose very well without the action of insulin. This is why Type 1 diabetics who do not make insulin can become very ill without insulin shots. Those who develop a deficiency of insulin must have it replaced via shots or pumps (Type 1 Diabetes). More commonly, people will develop insulin resistance (Type 2 Diabetes) rather than a true deficiency of insulin. In this case, the levels of insulin in the blood are similar or even a little higher than in normal, non-diabetic individuals. However, many cells of Type 2 diabetics respond sluggishly to the insulin and therefore their cells cannot absorb the sugar molecules as well. This leads to blood sugar levels that run higher than normal. Type 2 diabetics will occasionally need insulin shots, but most of the time other methods of treatment will work.

### *How Insulin is Produced in the Body:*

The biologically active insulin consists of 51 amino acids, three disulfide cross-links, and is comprised of two separate chains, termed A and B. Insulin is not synthesized in active form, but is first made as a single inactive peptide chain called preproinsulin. Preproinsulin has no cross-links, and in addition to the A and B chain, has two additional portions called the signal sequence and the connecting (C) peptide. The signal sequence informs the cell that the finished proinsulin should be deposited outside the cell. The C-peptide is necessary to allow proinsulin to fold in the correct conformation to ultimately produce active insulin. Preproinsulin is processed by a two step procedure; in the first step, the signal sequence is cleaved by a peptidase as it is transported into the rough endoplasmic reticulum, then two of the three cross-links are formed to give a new but still inactive peptide called proinsulin. A second peptidase then cleaves the C-peptide, and an internal disulfide forms to produce insulin. Endogenous (self-produced) insulin has a half-life of about four minutes in the bloodstream. C-peptide lasts about 30 minutes. By measuring the amount of C-peptide in the blood, scientists can determine the amount of insulin produced by the pancreas.

### *Insulin Firsts:*

- First Hormone Identified

## Using Biology Workbench

Insulin was the first hormone identified (late 1920's), which won the doctor and medical student who discovered it (Banting and Best) the Nobel Prize. They discovered insulin by tying a string around the pancreatic duct of several dogs. When they examined the pancreases of these dogs several weeks later, all of the pancreas digestive cells were gone (they died and were absorbed by the immune system) and the only thing left was thousands of pancreatic islets. They then isolated the protein from these islets and discovered insulin. Note that there are other hormones produced by different types of cells within pancreatic islets (glucagon, somatostatin, *etc.*) but insulin is produced in far greater amounts under normal conditions making the simple approach used by Banting and Best quite successful.

- First Protein Sequenced

Insulin was the first protein to have its complete amino acid sequence determined, reported by Frederick Sanger in 1956 (Sanger later won the Nobel Prize for this work). Sanger's work not only resulted in the sequence of insulin, but he also demonstrated several facets of protein structure. His experiments showed that proteins have a unique amino acid sequence, all molecules of a given protein are identical, and the sequence of each different protein is unique. He also showed for the first time that all amino acids in mammalian proteins are in the L (or S)-configuration, that the peptide bond is an amide bond, and that amino acids have alpha amino groups and alpha carboxyl groups. We now know that proteins are made when a section of DNA is read (a process called transcription) and a complimentary molecule of RNA is formed. This RNA is then used to specify the amino acid sequence of a given protein through a process called translation. Thus, the sequence of a protein is encoded in DNA.

- One of the First Signal Sequences Determined

Cells make many different proteins that must be correctly sorted to proper cellular locations. One of these sorting mechanisms involves signal sequences. Proteins destined for secretion from the cell typically possess these signal sequences (generally 15-30 amino acids located at the amino terminus of the protein). As the protein is made, it is transported into the rough endoplasmic reticulum with removal of the signal sequence. Although preproinsulin was not the first protein to have its signal sequence determined, it was one of the early ones. In 1999, in recognition of his work on protein-sorting mechanisms, Dr. Gunter Blobel received the Nobel Prize in Medicine.

- First Biotech Product to be Approved by the FDA

The first successful insulin preparations came from cows (and later pigs). The pancreatic islets and the insulin protein contained within them were isolated from animals slaughtered for food. Bovine (cow) and porcine (pig) insulin worked very well (and still do) for the vast majority of patients, but some could develop an allergy or other types of reactions to the foreign protein (since they differ slightly in amino acid sequence, cow and pig insulin are sometimes perceived as being foreign and attacked by the immune system). In the 1980's technology had advanced to the point where we could make human insulin. The advantage would be that human insulin would have a much lower chance of inducing a reaction because it is not a foreign protein. The technology that made this approach possible was the development of recombinant DNA techniques. In simple terms, the human gene that codes for the insulin protein was cloned (copied) and then put inside of bacteria. A number of tricks were performed on this gene to make the bacteria constantly produce insulin. This was one of the first human genes to be cloned into bacteria. Large vats of genetically engineered bacteria now make vast quantities of human insulin. From this, pharmaceutical companies can purify human insulin, sold as Humulin.

**Question for Exploration:**

Insulin is derived from a much longer chain called preproinsulin, which contains a "signal sequence" at the beginning (common in secreted proteins) and a "middle section" that is later cut out. The question to be explored is: Are the differences in preproinsulin sequences among different species equally distributed among the various regions?

## Using Biology Workbench

### **Rationale:**

*It seems logical that those positions in the protein sequence that are part of the final insulin molecule are more likely to be invariant among different species.*

### **Procedure:**

#### *Log In:*

- Log In to Biology Workbench at <http://inquiry.uiuc.edu/bioweb/perl/bioweb.cgi>, by entering your username and password and then clicking on *Submit*. (If you do not already have a username and password, register. Use your first initial, second initial, last name as your username, for example, *btsmith*, and enter the password *biology*, so the instructor can examine your account at a later date.)

#### *Start a New Session:*

- Click on *Session Tools*. (Wait for a response.)
- Click on *New*. (Wait for a response.)
- Type in *Insulin Study* in the Session Description box and click on *Start New Session*.

#### *Search the Protein Databases and Import Desired Sequences:*

- Click on *Protein Tools*. (Wait for a response.)
- Click on *Ndjinn*. (Wait for a response.)
- Unselect PBDfinder and Select SWISSPROT by clicking next to the names. Type in preproinsulin, and then click on Search. (Wait for a response.)
- Use the control key to highlight the entries for *insulin precursor from different species*, then click on *Show Records* (ignore other entries).
- Briefly examine some of the information given. Particularly, *note the common names of the species that correspond to the species codes used*. Use these data to complete the chart on the worksheet.
- Go back to the previous window and click on *Import Sequences*.

#### *Align the Imported Sequences:*

- *Scroll down and click on the box on the left of each of 7 entries that you choose to align, then scroll back up and click on CLUSTALW.* (Be prepared to justify your choices.) *Include the human sequence* so that you can make comparisons (see worksheet).
- Scroll down and click on *Import Alignments*.

#### *Create a Shaded Alignment Chart:*

- Scroll down and click in the box on the left of the list of proteins, then click on *BOXSHADE*. (Wait for a response.)

#### *Examine the Results:*

- The user can set the color-coding for the resulting alignment chart. In this case, those sites that are identical in all the species aligned are color-coded *green*. Where not all, but more than half of the species match, the predominant amino acid is color-coded *yellow*. *Blue* indicates a mismatch, but one in which the blue-coded amino acid is similar to the more common, yellow one. *No Color* indicates a mismatch in which the amino acids are not chemically similar or a position at which there is a gap. *Thus, a green column indicates absolute identity across the species being compared and a column with only yellow and blue indicates strong similarity across species, but one with any uncolored letters indicates greater variability at that position.*

*Create a Protein Similarity Diagram:*

- Go back to the *Alignment Tools* by clicking on the *Return* button at the bottom of the BOXSHADE window. Click in the box on the left of the list of proteins (if not already selected) and click on *DRAWTREE*. (Wait for a response.)
- *Print out the resulting Protein Similarity Diagram. Label each species by its common name. Answer question 4 on the worksheet.*

*Additional Explorations:*

- Click the *Return button* at the bottom of the *DRAWTREE window* with the diagram and click on the *Protein Tools button* at the top of the page to get back to the list of protein sequences. *Select and compare a different set of species. Are your conclusions the same in both cases? Explain.*

**Worksheet**

**Team Members:**

I. List the SWISSPROT code and common names of the species you chose to compare:

**Example:**                                 SWISSPROT Code                                 Common Name  
   CANFA   dog

<b>Species you compared:</b>	<u>SWISSPROT Code</u>	<u>Common Name</u>
	_____	_____
	_____	_____
	_____	_____
	_____	_____
	_____	_____
	_____	_____
	_____	_____

Briefly explain why you selected these particular species.

II. Complete the following table by counting the number of positions for which the same amino acid is present in ALL species and express the numbers as percentages:

<u>Part of the Sequence</u>	<u>Human Sequence</u>	<u># Identical</u>	<u>% Identical</u>
signal sequence	<u>M - AAA</u>	_____	_____
insulin B chain	<u>FVN - PKT</u>	_____	_____
C peptide	<u>RRE - OKR</u>	_____	_____
insulin A chain	<u>GIVE - CN</u>	_____	_____

III. Drawing Conclusions:

1. Was the percentage of positions with variability across the species you compared highest for the A chain, the C peptide, or the B chain? Does this make sense biologically? Explain your reasoning.
2. How similar was the signal sequence across the species you compared? Do these observations make sense biologically? Explain your reasoning.
3. What other patterns do you observe? For example, are columns with at least one blue equally common in all four parts of the sequence? If not, does the observed pattern make sense?
4. How do the groupings (clustering of species) in the Protein Similarity Diagram compare to the way these species are grouped taxonomically? (Any surprises?)

**Acknowledgments**

This exercise was co-authored by Mary Ball and Steve Karr. Steve is the chairman of the Biology Department at Carson-Newman College and teaches cell biology, immunology, and microbiology.

## **Activity II: Comparison of Preproinsulin Sequences Across Species (Kiser)**

### **Introduction**

Stacey has her students work directly on the Biology WorkBench web site, rather than using the Student Version. She explains that this web site is used by researchers in laboratories, and hopefully gives students more confidence that they too can “do” science. Stacey runs this exercise with both the life-science majors and the non life-science majors. She does it with the non life-science majors right after a lab in which they model the process of protein synthesis. The objectives are to reinforce the structure of proteins and introduce the idea of molecular evolution.

This is a high-level lab that takes a lot of support from the instructor at certain points. Stacey goes from group to group (there are eight computers in the lab room) and asks students to summarize what they have done to that point, or to interpret for her the information on the screen. It usually takes three hours for students to get the printout of a tree. In the next class meeting Stacey asks groups to write out a list of the steps that lead to the tree. This forces students to reflect and summarize in their own words. It takes away (somewhat) from the “black box” syndrome of working with complicated data on the computers.

### **Materials**

- Internet accessible computers

### **Notes for the Instructor**

Stacey also includes screen captures at various points in the student handout. Using a drawing program she adds arrows to indicate important features students need to find, buttons to click on, or places to place text for a search. She did not include them in this manuscript, but urges instructors to use them to assist the students during the lab.

The Biology WorkBench web site is very dynamic. Be prepared to walk through your handout a week before doing it in class to look for changes in the website. Adjust your instructions and figures accordingly.

For assessment and tracking purposes, Stacey chooses to set up an account and have all the students log in under that account instead of setting up their own accounts (see “Starting a Session” section). They still have total internet access, but it gives the instructor the ability to review their work by selecting their files.

The WorkBench web sites can also be modified to present the choices as buttons, menus within scrolling screens, or pulldown menus. You can toggle between these choices by clicking on the title bar at the top of any menu screen. This version is written as if students are working with the scrolling window menus, but it should be pointed out to students that they can switch if they want (and may find themselves in an altered visual state of reality if they accidentally click on the title bar).

## Student Outline

### Comparison of Preproinsulin Sequences Across Species Tracking Differences In Protein Sequences

#### Outcomes:

At the end of this lab, you will be able to:

1. Name the subunits of a protein.
2. Compare similarities and differences among proteins.
3. Describe how the comparison can lead to the creation of a phylogenetic tree.
4. Describe what differences among proteins mean in terms of microevolution.

#### Biology WorkBench address: <http://workbench.sdsc.edu>

We will be working with the molecule *insulin* today. Insulin is produced by cells in the pancreas. It is a protein that acts as a hormone to signal other cells in the body to import glucose from the blood stream and store it as glycogen for later use.

**Question:** What disease(s) do you know of that involve a problem with insulin?

When the protein is first synthesized near the cell's rough endoplasmic reticulum, it is called *preproinsulin* and consists of three distinct sections. The A-chain and B-chain will be modified to construct the active insulin molecule. The C-chain, which is situated between the A-chain and the B-chain in the preproinsulin molecule, is cut out in the processing of the protein. On one end is a sequence (the "pre" section) that signals the endoplasmic reticulum to allow the molecule to enter. It is quickly cut off and the molecule consists of just the A, B, and C chains. It is called proinsulin when it is in this state.

**Question:** Name specific area(s) in the cell that might do this modification to the protein.

#### A special note about this activity:

You will be using a powerful supercomputer site that is used by research scientists around the world to answer research questions. It is designed to be as flexible for research, but we will not be using many of the more advanced features. Be prepared to:

- A. Be patient.
- B. Scroll up and down to find the buttons within certain windows and to check on all the settings.
- C. Use the help buttons and other web sites to help you make decisions and evaluate your results.

#### Step 1: Starting a Session

- Go to the Biology WorkBench website address listed at the beginning of this lab. Click on the "Enter the Biology Workbench" button to begin a session. If the instructor has provided your class with an account, log in using the name and password they give you in class.
- If your instructor does not give you an account name, start your own session by clicking on the "set up a free account" option. The next screen asks you for information, including a password. Make this password something easy to remember the next time you log in. Enter the Biology WorkBench by clicking on the highlighted link.
- Scroll down through the first screen in order to find the tool bar. The most basic structure of the site asks you to choose a set of "Tools." When you select one of the specific tools, you will be

## Using Biology Workbench

provided with a narrow set of options. Each time you log on, go to the Session Tools first to either begin a new session or log on to a previous session. Choose “*Session Tools*” to begin a new session for your lab group.

- To start a new session, select the “*Start New Session*” option in the choice window and click the “*Run*” button. It will ask you to type in a name for your session. Type in a name that clearly identifies your group, and then hit the “*Start New Session*” button. You will then return to the Session Tools web page. Your new session name should be highlighted. If it is not highlighted, scroll down through the list of session names and click on the gray button next to your session name to highlight it.
- You all can access your group’s session from any internet connection again (at home?) by simply logging in to my account and selecting your group’s session name at this web page. Select the “*Resume Session*” choice then “*Run*” to get back into your previous session.

### Step 2: Selecting a protein

- We first need to find a protein sequence for human preproinsulin, and then search the multiple databases for preproinsulin sequences from other organisms. Make sure that the button next to your new session name is selected, scroll to the Tool bar, then click on the “*Protein Tools*” button to start searching for a protein.
- The “Protein Tools” web page gives you a variety of tools to search the global protein sequence databases. Notice you do not have any protein sequences so far, so the window shows you “*Empty*.” You will need to search for a preproinsulin protein sequence from one of the databases.
- Select “*Ndjinn – Multiple Database Search*” from the window. Click on “*Run*” to get to the protein database’s search engine web site.
- In the new search window, enter “*preproinsulin*” into the search window. (Note: preproinsulin has a short sequence that acts as a signal to get the protein into the lumen of the rough endoplasmic reticulum. The “pre” sequence gets cut off immediately). Next, scroll down through the page and look at the choices of protein databases. The notes next to each database give a brief description of the database. One database that will find human preproinsulin is “*GBPRI*” GenBank Primate SequencesSequences. It is about one third of the way down the list of databases. Select this database by clicking on the gray square next to its name.
- Submit the search string to search by clicking on the “*Search*” button back at the top of the page.
- The search engine returns a number of sequences. All of the sequences will be selected (checks in the boxes in front of each name) and you will have to select just one to import into your WorkBench session. Usually the data is arranged in a ranked order, with the best matches listed first. Select the first sequence on the list that describes a *human* preproinsulin molecule by *clicking off* the other protein sequences until only the human preproinsulin sequence has a check in the box next to it.
- Many of the databases contain extensive records about the sequences. Use the buttons near the bottom of the screen to “*Show Record(s)*”.
- *Note:* This brings up a new window with much more data about the sequences, including the original scientists who found the sequence, when and where it was published, and classification information for the organism. Read through this, paying close attention to the “Definitions” section. Make sure they describe this sequence as a human sequence.



- Once you have checked the sequence information, return to the WorkBench page by closing the window. This returns you to the WorkBench Protein Tools search results window.
- The final step is to import the human preproinsulin sequence into the WorkBench by selecting the “*Import Sequence(s)*” button. This moves the sequence from the Ndjinn database to your session account on the WorkBench. You will now return to the Protein Tools page and be able to search for proteins that are similar to your imported human preproinsulin sequence.

### Step 3: Searching for similar sequences

- Once you have a sequence imported to the WorkBench, you can search multiple databases for similar sequences. You will stay in the Protein Tools section of Workbench and use “*BLASTP*” to search for other protein sequences. This search engine compares our sequence to other selected protein databases and returns any sequences that are wholly or partially similar to our search sequence.
- Click the box in front of your imported human preproinsulin sequence. Scroll down the window and select the “*BLASTP*” computer program. This compares a protein sequence (PS) to a protein sequence database (PS DB). Click the “*Run*” button.
- You will get a new BLASTP window that allows you to make some choices and set some parameters for the search. Make the following changes:
  1. Select “*SWISSPROT*” in the database selection window.
  2. Click the “*Disable Gapped Alignment*” button.
  3. Scroll down and change the “*Alignments*” button to 250 (from 50).
- Now click the “*Submit*” button at the bottom of the page.
- You now have a whole selection of similar sequences of which you should pick five or six to align. It is better to pick ones with a higher “*Score*” for alignment. You can select more than one sequence on Macintosh computers by holding down the Apple key while choosing each new sequence to add to your set. You can select more than one sequence on a WINDOWS machine by holding down the control key and clicking on each sequence.
- Once you have five or six selected, click the “*Show Record(s)*” button to find out the names of the organisms that are the sources of the insulin proteins you selected to compare.

**Question:** List the organisms’ common names for the selected insulin sequences.

	SWISSPROT ID Name	Common Name
1.		
2.		
3.		
4.		
5.		
6.		

**Question:** Before you do any further comparisons, *predict* which organisms are most closely related based on what you know about them right now:

**Question:** Which will be most distantly related? (farther apart on an evolutionary tree)

**Question:** Briefly justify your predictions above by writing down what characteristics you used to make your predictions.

## Using Biology Workbench

- Close the BLASTP help window and return to the BLASTP results window on the WorkBench. Import the sequences into the WorkBench for comparison by scrolling down to the bottom of the page and clicking the “*Import Sequence(s)*” button. This will automatically import all the sequences in the highlighted records you are viewing at the time.

### Step 4: Aligning sequences

- The next step is to use a different series of web pages to align the imported preproinsulin sequences. Alignment lines up the selected protein sequences in a way that maximizes the similarities and reduces the amount of differences among all the sequences. It highlights areas of similarity so you can quickly visualize overall patterns among the sequences.
- Make sure to select all of the records that you want to align from the list that you just imported by clicking on the boxes in front of the appropriate sequences until there is a check mark in front of all the sequences you want to align. Scroll down through the window and select “*CLUSTALW - multiple sequence alignment.*” Click the “*Run*” button.
- The CLUSTALW web page appears with the list of your sequences at the top. Included are another series of settings for alignment. We will run CLUSTALW using the default settings.

Before you look at the aligned sequences, make the following predictions:

**Question:** Recall that the preproinsulin protein is made up of three chains: two chains (A and B) will covalently bond to create the hormone insulin, and one chain (C) will be cut out completely (see diagram). Which of the chain(s) do you predict will be more similar among your different preproinsulin sequences?

**Question:** Which chain(s) will vary the most?

**Question:** Briefly justify your predictions.

- Click on the “*Submit*” button to align the selected preproinsulin sequences. The output will show you the areas of the protein that are the same in all of the sequences that you submitted to the alignment computer. These will show up as colored letters.

**Question:** The letters in the output represent individual monomers of the polypeptide sequences. What do individual letters represent? (Hint: think about the structure of proteins.)

**Question:** How closely matched are your sequences? Can you find ones that are more similarly matched with others, or some that don't seem to match as well?

**Question:** Does alignment output have different areas where the proteins are similar to each other, and areas where the proteins differ from each other?

**Question:** Can you find the three different chains of the protein from the diagram on the back page in the alignment on the computer screen? (Hints: They may not be in order, or arranged the same way front to back. You may have to decode a short piece of the output. Or you could think carefully about areas of similarity and differences in your sequences, and how those might relate to the A, B, and C chains.)

**Question:** Did your output match your predictions for the chain(s) that will be more similar and the chain(s)?

### Step 5: Building a phylogenetic tree with your aligned sequences

- Next, you can build a tree that indicates the degree of relatedness based on the number of differences among all of the sequences that you aligned. Select “*Import Alignment(s)*” to import your aligned sequences. When you import the aligned preproinsulin sequences, you will automatically switch to the “*Alignment Tools*” web page. This part of the web site gives you a series of tools to analyze your aligned sequences. You should see a series of the five or six sequences you just aligned with a single check box in front of them.
- Check the box, then select the “*DRAWTREE – Draw Unrooted Phylogenetic Tree from Alignment*” choice in the window. Click the “*Run*” button. Change the “*Send Results To*” button on the toolbar to “*New Window*” and click the “*Submit*” button.
- Scroll down the window and look at the unrooted tree generated from the aligned sequence data. Print a copy of the tree to examine, and answer the final questions.

**Question:** Which protein sequences diverged most recently from each other? How can you tell this from your tree?

**Question:** How did the computer use the alignment data to decide how to create a tree? (Hint: What was it comparing from the aligned sequences?)

**Question:** Does your tree support your predictions in question 4? Were you correct in predicting the most closely related organisms? The least closely related organisms?

**Question:** What would you do as a scientist if you constructed a tree like this and it did not support your predictions?

**Question:** Which set of data (current relationships versus new protein sequences) would you most likely believe? Why?

### Step 6: Extra Credit Projects

A. There is another alignment tool that allows users to visualize not only amino acids that are identical to each other, but also analyzes amino acids that are chemically similar to each other. It uses chemical rules much like on page 69 of your textbook.

*Return through the web pages to the Alignment Tools page and try using “BOXSHADE” tool to align your sequences. Once you have created an alignment, write up a paragraph about how this type of alignment gives us different (more useful?) information. Include a description of why certain amino acid substitutions are “similar” to each other.*

B. You can now navigate the web site on your own. Choose another protein from the readings (papers or the textbook) and conduct a similar experiment. Build phylogenetic trees and compare them to predicted trees. If possible, compare the same organisms that you did for preproinsulin. Do the trees come out the same? How does *your* protein differ from the preproinsulin protein? What kinds of differences might be important when comparing for evolutionary hypotheses?

### Acknowledgments

Stacey would like to thank BioQUEST for introducing her to the Biology Workbench. Joyce Cadwallader and Peter Woodruff helped her refine her initial lab at a BioQUEST summer workshop in 2000. Mary Ball and Garry Duncan are always pushing the envelope and causing the lab to further evolve.

### **Activity III: Mining DNA and Bringing Molecules to Life**

#### **Introduction for Instructors**

The following exercises under this activity listing will act as a springboard to empower students who wish to pose evolutionary questions that can be solved by analyzing molecular data. To accomplish this end, students must have a user-friendly and web-based interface that enables them to access DNA and protein databases, perform alignments and phylogenetic trees in order to answer their questions. The Biology Student Workbench (hereafter known simply as the Workbench), developed by NCSA, provides this user-friendly, web-based interface.

The predominant taxonomic hypothesis in biology has been that all organisms belong 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 this view. Some recent evidence has led to the hypothesis that the prokaryote domain is actually comprised of two distinct domains. Some bacteria-like organisms look like normal bacteria but they seem to have a distinct phylogenetic history. Consequently, these bacteria-like organisms comprise a distinct domain, called the Archaeobacteria, or more simply, Archaea. The name reflects an untested conjecture about their evolutionary status. Recent phylogenetic evidence suggests that the Archaeobacteria are at least as old as the other major domains; hence, it now seems possible that the newest group of organisms is actually the oldest. It is important to note that not all scientists agree with the three-domain hypothesis.

DNA and protein sequencing have dramatically improved the ability of biologists to uncover the pattern of organic evolution through geologic time. Sequence substitutions in DNA and proteins can be considered molecular fossils; that is, these changes form a fossil record much like those changes in physical characters that have been observed in the traditional fossil record. Looking for differences in the sequences and discovering patterns in the degree of difference is an indicator of how long ago in the past any two organisms may have shared a common ancestor.

#### **Materials**

Like Activities I and II, this set of investigations requires Internet access.

#### **Notes for Instructors**

See the comments for the other activities in this chapter. The instructor should work through the entire exercise in advance. Make sure to visit the Protein Explorer website in advance and to verify that your computers are set up to be able to use this software.

## Student Outline

### Mining DNA and Bringing Molecules to Life

#### *Investigation 1: Explorations in Evolution Through DNA Sequence Alignments and Phylogenetic Tree Construction*

**Purpose:** Investigate whether the phylogenetic trees developed from DNA sequence alignments support the two-domain hypothesis or three-domain hypothesis.

**Procedure:** In this activity, students will examine and compare the amino acid sequence of an enzyme called *enolase*. We find *enolase* 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 a ubiquitous enzyme since all organisms utilize glycolysis. The students will compare the amino acid sequences of *enolase* from six species (Table 1). The complete genomic sequences have been determined for all six of these species.

**Table 1.** List of species used in the investigation.

Species name	Group
<i>Methanococcus jannaschii</i>	Archaea
<i>Eschericia coli</i>	Bacteria (Gram negative)
<i>Bacillus subtilis</i>	Bacteria (Gram positive)
<i>Drosophila melanogaster</i>	Eukarya
<i>Homo sapiens</i>	Eukarya
<i>Saccharomyces cerevisiae</i> (yeast)	Eukarya

*Procedures for Using the Biology Student Workbench to search DNA databases, construct DNA sequence alignments, and construct phylogenetic trees:*

#### 1. Entering the Biology Student Workbench:

- a. Launch your graphics browser and go to the following URL for the Biology Student Workbench: <http://peptide.ncsa.uiuc.edu/>. (Or, if you are an experienced user, you can enter the Biology Workbench, which is for advanced users and in which you can manipulate many different searching/weighting parameters: <http://workbench.sdsc.edu/>).
- b. Click on the button at the top of the window titled *Curriculum Materials*.
- c. Click on the hyperlink *Biology Workbench for Students* (second bullet, with green background).
- d. Click on the next hyperlink, which is also titled *Biology Workbench for Students*. (If you would like, you could examine the tutorials before going on.)
- e. If you have already set up an account on the Workbench, go to step f. 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 *Register* button.
- f. Enter your user id and password and then click the *Register* button.

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

- a. 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 to continue a previous file in work processing. In other words, you cannot use the *Protein Tools*, *Nucleic Tools*, or *Alignment Tools* until either you have resumed an old session or started a new session. So, click the *Session Tools* button at the top of the left frame.

## Using Biology Workbench

- b. You can tell that you are in the Sessions window (even though the window does not have this title) since all of the buttons refer to the various things you can do with a session. You are now ready to start a new session or resume a previous session.

- 1) To start a new session, click the *New* button near the top of the page. 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 *Start New Session* 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) 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. You are now ready to begin searching for amino acid sequences. So, click the Protein Tools button near the top left frame of the page. Now go to step 3.

OR

- 2) If, however, you wish to *resume a session*, click the radio button before the session you wish to resume, then click the *Resume* button. Now click the *Protein Tools* button. (If this is your first session, you won't have any sessions in which you can resume.)

*Overview of remaining operations:* Since we do not have an amino acid sequence of enolase for comparison, we must search for one. Once we have done that, we will do the following: (1) generate a list of proteins with similar sequences by conducting a search for similar sequences, (2) select sequences we wish to align and then align them, and (3) finally, construct a phylogenetic tree and distance matrix based on the sequence alignments.

### 3. Selecting a sequence:

- a. Now that you are in the 'Protein Tools' window, click the *Ndjinn* button, which will enable you to search multiple databases for a protein sequence of your choosing — in this case, enolase — since we don't have our own sequence to start with.
- b. You now have a decision to make. If you don't care whether there is a 3D structure of the protein you are interested in, then you should proceed to *option 1* below since the protein database is very large. If, however, you wish to examine the 3D structure of a protein and/or import alignments into the software *Protein Explorer* (discussed in the next section of the paper) to visualize where mutations or conservation of amino acids occur in 3D, then you should select *option 2* below.

*Option 1:* In the window of the *Ndjinn Multiple Database Search* query form, scroll down and select *SWISSPROT*, which, as the last part of its name suggests, is a protein database. If your search is unsuccessful, then you should search the *PDFFINDER* and *PIR* databases. Before clicking the *Search* button, you must enter your search term in the query form above the search button. Type in *enolase*. Leave the pop-up menu on the default selection (i.e., Contains). The search will then look for any database submission that contains the search term in any part of the submission record. Now click the *Search* button.

The *RESULTS* page indicates that we have matched 120+ unique records. (The number of unique records may be larger for your search since records are being added on a continuing basis.) In the window containing the list of the 120+ records, select *swissprot*:

*eno1\_yeast* – *enolase 1 (ec 4.2.1.11)*. This entry record contains the amino acid sequence for which we will base the rest of our search. Now go to step c.

*Option 2:* If you wish to import your alignments into the software Protein Explorer to visualize where mutations or conservation of amino acids occur in 3D, then select the *PDBFINDER* database, type in the name of the protein, and click the *Search* button. (In order for Protein Explorer to function in this type of visualization capacity, there must be a known 3D structure.)

Click on one of the results (there are 13 for enolase). For example, click on the one that says: *[1] PDBFINDER:4ENL – CARBON-OXYGEN LYASE*. For later use in *Protein Explorer*, you will need to write down the PDB id number, which is the middle four letters of the record—in this case, 4enl. Now go to step c.

- c. Before you import this sequence into the Biology Workbench, click the *Show Record(s)* button, which will give you a wealth of information about the protein, including its amino acid sequence, its enzyme code number, citations, etc. You can even view the molecule in 3D. (Be sure to write down the molecule's id number, in this case it is 4enl. You need this protein data bank id number for the software Protein Explorer to visualize where mutations or conservation of amino acids occur in 3D.) Now click the *Back* button on your browser.

Before we can search for records containing similar sequences, we must import to the workbench the entry you selected above. To do this, click the *Import Sequence(s)* button. You are now ready to begin a search for proteins with similar amino acid sequences.

#### 4. Searching for records with similar sequences:

- a. Click the box in front of the record you wish to anchor your search. In this case, click the box in front of *SWISSPROT.ENO1\_YEAST* if you did option 1 above, or *PDBFINDER:4ENL \_\_ CARBON-OXYGEN LYASE* if you did option 2 above. (You may have to use the scroll bar in the window until you see this entry, which will be below all of the buttons.)
- b. Click the *BLASTP* button, which enables us to compare our protein sequence (ps) to a protein sequence database (ps db).
- c. You should now be in the BLASTP interface page. Note near the top of the page that our *Selected Sequence(s)* is *ENOLASE 1 (EC 4.2.1.11) (2-PHOSPHOGLYCERATE DEHYDRATASE) (2-PHOSPHO-D- GLYCERATE HYDRO-LYASE) Saccharomyces cerevisiae (Baker's yeast)* if you selected option 1 above or *CARBON-OXYGEN LYASE*, if you selected option 2 above. Now perform the following steps in order to complete the BLASTP search:
  - Click the *Choose One or More Databases* radio button and then select *SWISSPROT*.
  - Now click the *Submit* button, which will now execute your BLASTP search. (BLASTP finds other proteins in the SWISSPROT data bank that have similar amino acid sequences.)

#### 5. Selecting records for alignment:

- a. Find the small window with a scroll bar. Just above the window, the small font should say *Sequences producing significant alignments*. In this window, we want to select the sequences that we wish to align to the yeast sequence. After clicking the first selection, you will need to hold down on the command key (Apple key on the Mac) while making additional selections. For the *enolase* activity, select the following five records (the yeast record, which you have already selected, acts as the sixth record and does not need to be selected again):

## Using Biology Workbench

ENO_DROME	(Drosophila melanogaster)
ENOB_HUMAN	(human)
ENO_METJA	(Methanococcus jannaschii)
ENO_BACSU	(Bacillus subtilis)
ENO_ECOLI	(E. coli)

- b. Now click the *Import Sequences* button. This action will import the amino acid sequences of the five records (= sequences) you have selected, as well as the yeast record (sequence) with which you started, so that you can do alignments.
6. Conducting a sequence alignment:
  - a. Click the boxes of all of the records you wish to align—in the case of our enolase alignment, click the boxes of all six records.
  - b. Select the *CLUSTALW – multiple sequence alignment* option. It will take the computer a few moments to develop the alignments.
  - c. Now scroll down the page and see the alignments. (You will also see that an unrooted tree has already been constructed.) Continue to scroll to the bottom of the page and click the *Import Alignments* button so that you will then be able to launch applications to construct rooted and unrooted phylogenetic trees, and/or change sequence formats (e.g., if you wish to change from an MSF format to Fasta format).
7. Tree Construction:
  - a. Under *Sequence Alignments*, click the box in front of the *CLUSTALW-Protein* list near the bottom of the page. (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 (i.e., all six records, in the case of our enolase exercise).
  - b. Click the *DRAWTREE* button, which draws an unrooted phylogenetic tree. Note that you can print out a copy of the tree, which should look similar to the one in Figure 1.

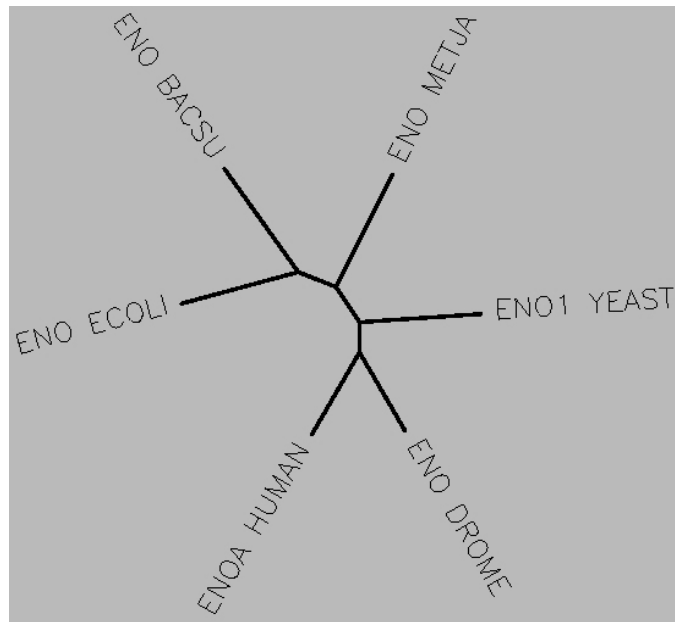


Figure 1. Unrooted tree. Note that the two species of Bacteria, *Bacillus subtilis* and *E. coli*, are on one side of the tree, while the Eukarya—yeast, *Drosophila* and human—are found on the other side of the tree, while *Methanococcus jannaschii* is found splitting off of the middle



## Using Biology Workbench

of the tree. The length of the line between the branches is a direct measure of the amount of change that has occurred. So, *M. jannaschii* is about half way between the Bacteria and the Eukarya.

## Using Biology Workbench

### Questions for discussion:

- (1) Where would you expect *Methanococcus* 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* to split off of the unrooted tree if the three domain (i.e., Bacteria, Archaea and Eukarya) hypothesis is correct?
- (3) Which hypothesis does the tree in Figure 1 support?

### 8. Exporting Sequence Alignments:

Click the *Return* button at the bottom of the web page in order to return to the previous web page. Click the *View* selection to view the sequences and change the format of how the sequence is viewed. If you wish to import the sequence into *Protein Explorer*, you will need to change the format to *Fasta*. To do so, click and hold on the *Format* popdown window; scroll down and release on *Fasta*. The format should automatically change in a few seconds. If not, click on the *Change Format* button. Once the format has been changed, you can either save the sequences to a file, or copy and paste them into *Protein Explorer*. If you are going to continue on to Investigation 2, then the copy/paste method is the easiest. Simply highlight and copy all of the sequences, including the > sign and the enzyme/species names, so that they can eventually be pasted into *Protein Explorer* (step 9 in Investigation 2 below).

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

*Background:* Amino acid sequences retrieved from databases can be aligned using software found in the Biology Workbench (NCSA), and the alignment of amino acid sequences can then be imported into Protein Explorer (PE) where mutations or conservation (= identity) may be visualized in 3D. The following procedures allow will allow us to view in 3D the aligned protein sequences of enolase (developed in Investigation 1 above) from *Homo sapiens*, *Drosophila melanogaster*, *Saccharomyces cerevisiae*, *Methanococcus jannaschii*, *Eschericia coli* and *Bacillus subtilis*. In essence, you will be visualizing in 3D the evolution of enolase.

1. Open Netscape and launch *Protein Explorer* (<http://www.umass.edu/microbio/chime/explorer/>). (You must use Netscape 4.04; sorry, Internet Explorer doesn't work and this is beyond the control of the person who developed Protein Explorer.)  
(If you do not have the 2.0 or greater version of Chime, you will be given a link to download it and add it as a plug-in to Netscape. *After you have downloaded it and installed it as a plug-in, you will need to exit from Netscape, and repeat step 1.*)
2. On this web page, note all the different ways in which you can learn about Protein Explorer, including a 1-2 hour tour that will give you a better idea of all of the capabilities of Protein Explorer. In the middle of the three gray columns, the one with the heading *Startup Options for Your Molecules*, click the fifth bullet, "*Bare*" *Protein Explorer*. When the new page comes up, click *Protein Explorer*.
3. Resize the window, as instructed; type in a *Session name* much like you would name a new word processing file (e.g., you could name the session Enolase Alignment); and, then click *Start Explorer session*.
4. You should now be in the *Load Molecule* window. You need to load a 3D protein molecule into Protein Explorer. Under *1. Load a Molecule*, enter the four-digit PDB ID for the enolase from yeast (i.e., 4enl). (This is the code you were instructed to write down in step 3.c. in Investigation 1.) Now click the *Load* button that follows your entry. 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 *Protein Explorer* in step 10 below.)

5. You should now be in the *FirstView* frame. Click the sixth bulleted item *Explore More!*
6. You should now be in the *QuickViews* frame. Near the bottom of this frame, click the hypertext where it says: *Go to Advanced Explorer.*
7. You should now be in the *Advanced Explorer* frame. Click item that says *MSA3D: Multiple Sequence Alignment Coloring.*
8. In the *MSA3D Procedure* frame, click item that says *MSA3D ALIGNMENT FORM.*
9. You are now in the *MSA3D Alignment Form* window where you will paste the amino acid sequences in FASTA format. To do so, place the cursor in the *Alignment Box* and paste the *Fasta* sequence you had copied into the clipboard from the *Biology Workbench* (i.e., step 8 in Investigation 1 above).
10. We need to copy/paste the yeast sequence, which is the sequence for which the 3D structure is known, from the *Alignment Box* into the *3D Sequence Box*. Since there is now a scroll bar, you will have to place the cursor somewhere in the *Alignment Box*, and then use the up or down arrows until the yeast sequence comes into view. Copy/paste this sequence into the *3D Sequence Box*.
11. Click the *Color Alignment & Molecule* button just below the *3D Sequence Box*. In a moment, a new Netscape page will open, showing the alignments for the six species. (This process may take several moments, so be patient.) The color codings are indicated at the top of the page. For example, medium green indicates that an amino acid at a specific position is identical for all 6 species. If your screen is large enough, you will see the 3D structure rotating on another Netscape web page. Click on that web page to bring it to the front.
12. 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, Different* (i.e., the first three bullets) — to spacefill all categories. The red balls are water molecules. Click the button *Water* button so that the red balls (water oxygens) are hidden, so you can see the backbone trace of the protein clearly.
13. The 3D model is showing the several billion year evolutionary history of enolase. 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 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 dark green (complete identity), showing billions of years of evolutionary conservation, while the peripheral region of the molecule is yellow.

*Possible questions for discussion:*

1. What would the peripheral region of the enolase molecule not need to be highly conserved, like the active site?
2. Are there other regions on the enolase molecule high conserved, besides the active site? What might be the role of those regions?
3. Do other enzymes in glycolysis show similar results? (To answer this, you would have to repeat this exercise, substituting other glycolytic enzymes for enolase.)
4. What other proteins might the six taxa used in the above investigations also share?