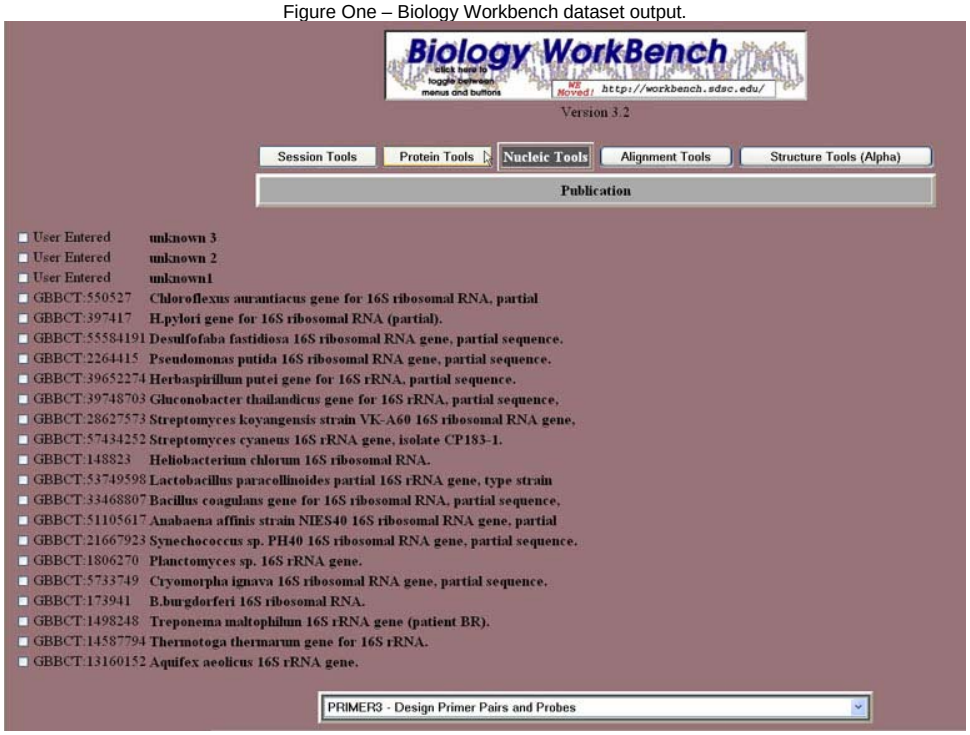
A Laboratory Class Exploring Microbial Diversity and Evolution Using On-Line Databases, the Biology Workbench, and Phylogenetic Software

Authors: S.M. Boomer, K.L Shipley, B.E. Dutton, and D.P. Lodge
Address: Western Oregon University, Department of Biology, Monmouth, OR 97361

Abstract
Students assemble and align bacterial datasets using DNA information downloaded from the National Center for Biotechnology Information (NCBI) website and Biology Workbench (BW). Specifically, they compare unknown original DNA sequences (from, in our case, hot spring communities) to a backbone of diverse bacterial control sequences representing 11 different phyla. Students use aligned datasets to obtain phylogenetic trees with Phylogenetic Analysis Using Parsimony (PAUP). In performing this exercise, students form predictions about bacterial phylogenetic relationships, enhance their understanding of bacterial diversity, and better appreciate the large body of research about bacterial diversity using DNA information.

Figure One – Biology Workbench dataset output.



Activity

INTRODUCTION

Learning Objectives.
Upon completion of this activity, students will be able to (1) learn how molecular data is stored, annotated, and accessed through NCBI/GenBank; (2) have a broader understanding of bacterial diversity, emphasizing current research using DNA approaches; (3) be able to gather, assemble, and align a 16S rRNA gene datasets using on-line software; (4) analyze phylogenetic trees using parsimony-based methods; and (5) understand and interpret phylogenetic trees to address microbiology classification issues.

Background.
This laboratory exercise has been carried out in a variety of formats at Western Oregon University (WOU) and was included as one component of our poster about computational biology curriculum, "A Bioinformatics Course Emphasizing Molecular Microbial Diversity," at the 2004 American Society for Microbiology General Meeting. Here, it received considerable attention from undergraduate microbiology instructors interested in developing comparable exercises and/or courses. Although an increasing body of literature exists about bioinformatics curricula for undergraduates, most is directed at genome or protein structure topics, typically emphasizing human or eukaryotic systems (2, 3, 5).

This curriculum was first developed in 1998 as a 2-week unit for an advanced elective course, Molecular Biology. In 2000, it was integrated into a new elective course, Computational Biology/Bioinformatics, that also featured microbiology-driven genome and protein explorations (all curricula available on-line at http://www.wou.edu/~boomers/Bi301/comp04cover.htm). Finally, in 2002, this exercise was adapted for the laboratory component of General Microbiology, a course that all Biology Majors at WOU are required to take. This report presents the most basic version of the exercise, currently used for General Microbiology. Students complete this laboratory exercise during the final 2 weeks of the 10-week course, following extensive lectures about microbial diversity, ecology, metabolism, evolution, and genetics. Prior lectures importantly include a summary of Woese's 16S work (10). The course uses/requires Brock Biology of Microorganisms (6), an invaluable resource for bacterial diversity - particularly during this exercise.

The same instructor (Boomer) who has developed lectures also runs the laboratory. Students receive all instructional materials at the beginning of the term, and these documents are provided electronically on in-lab computers. Pre-lab lecturing is limited to an overview of on-line tools and software function understanding, as summarized in lab handouts. This class requires two sessions, each 2-3 hours in duration. During the first laboratory session, students gather their control dataset using NCBI/GenBank (www.ncbi.nlm.nih.gov/) , and upload all control sequences and unknowns to accounts on BW (http://workbench.sdsc.edu/). During the second laboratory session, students align their datasets using BW-provided ClustalW (9), and then search and evaluate phylogenetic trees using Phylogenetic Analysis Using Parsimony (PAUP) (8).

PROCEDURE

Materials.
This exercise should be run with no more than 2 people working together. Each pair needs:

- One computer with internet access and disk space for file storage
- PAUP – available for purchase at http://paup.csit.fsu.edu/ ($500/10 users)
- TreeView – freeware (http://taxonomy.zoology.gla.ac.uk/rod/treeview.html)
- Text Program (e.g. Microsoft Word)
- Ruler

Before considering this exercise, instructors should understand several advanced phylogenetics concepts, two recommended resources for which are Molecular Evolution, A Phylogenetic Approach and Bioinformatics, A Practical Guide to the Analysis of Genes and Proteins (1, 7). Minimally, instructors should recognize how alignment tools work with an emphasis on ClustalW (9), distance vs. discrete phylogenetic approaches (e.g. understanding ClustalW invokes some simple distance methods, and the PAUP protocol invokes discrete parsimony methods), tree searching methods (e.g. this protocol employs a heuristic approach), and basic tree terminology (e.g. this protocol generates unrooted, ultrametric trees with legends). The extent to which instructors teach advanced concepts should be at their discretion. For this General Microbiology exercise, students are only expected to understand basic principles included in provided handouts. For Computational Biology/ Bioinformatics (in which Molecular Evolution, A Phylogenetic Approach provides the course text), students are expected to fully master all advanced concepts for the phylogenetics unit.

For all of these exercises, we have made every effort to recommend stable resources given the ephemeral nature of many computer-based resources. These include NCBI, supported by the National Institutes of Health and the National Library of Medicine since 1988, and BW, supported by the San Diego Supercomputer Center and National Science Foundation since 1996. Where ClustalW can be used on-line at BW, the more computationally demanding PAUP software package must be purchased. Although PHYLIP (Phylogeny Inference Package) is available as a free package for on-line downloading (4), we have found it less user-friendly in terms of operation, instructional documentation, and maintenance because it not provided through a commercial vendor. A useful summary and comparison of these 2 packages can be found in reference (1). Instructors who would prefer to adapt these curricula for PHYLIP have enough information to make appropriate decisions about parameter settings for alternative protocols. Given that PHYLIP encompasses over 30 different programs with hundreds of pages of on-line technical documentation links, however, discussing its operation is beyond the scope of this paper. It is also worth mentioning that some PHYLIP software (e.g. TreeDraw) is included on BW. Unfortunately, TreeDraw is a simplistic program that imports ClustalW's distance-based alignment data to produce a tree (notably lacking a legend). Although distance-based methods generate trees, they cannot compare with sophisticated discrete methods (e.g. parsimony or maximum likelihood, both of which are options on PAUP and PHYLIP) that retain alignment information and, in so doing, provide legend information that relates to actual nucleotide changes. In general, instructors need to be wary of "phylogenetic" software that can be run on-line as most represent distance-based methods that will typically result in a different and less-informative tree.

Student Handouts.

Lab Session One
Appendix 1 – Introduction and Procedures (see end of this report)
Appendix 2 – NCBI Control Worksheet* (see end of this report
Appendix 3 – Unknown Sequences* (see end of this report)

*These need to be available in an electronic or on-line form because students need to copy/paste into or from these documents during the lab exercises.

Lab Session Two
Appendix 4 – Introduction and Procedures

Instructor Version.
This class extends over two sessions, each 2-3 hours in duration. In our course, students have been required to use computers throughout lab to generate reports, and during lecture to research genomes and epidemics using on-line tools and resources (including NCBI). Although student pairs typically take 2 hours to complete these exercises, individual students and students lacking comparable experience should be provided 3 hours. If students have not been exposed to information about microbial diversity and Woese's 16S-based approach, an additional hour of lecture time should be developed to address these topics.

Instructors with no experience using any of these resources or tools should spend 3-5 hours understanding NCBI data and this dataset, working with Biology Workbench, and mastering PAUP operation. Although instructors are encouraged to use this all-purpose 16S dataset and provided controls, they should explore creating their own datasets - if only because it provides good experience for mastering lab concepts and procedures. Minimally, instructors can vary provided controls by selecting alternative 16S bacterial sequences directly from GenBank, cutting/pasting raw sequences as Word files to students without names included.

Dataset Findings
Using NCBI/GenBank, students should be able to solve the GenInfo. Identifier sequence information number (GI#), Kingdom/Phylum, and Genus/Species for each accession number. Some students are confused by GenBank annotations for Kingdom/Phylum and will record different levels of classification. GenBank taxonomists also use some terminology that is not consistent with some undergraduate textbooks (e.g. Firmicutes instead of Gram Positive). Given that some sequences represent uncultured environmental DNA samples, this terminology often appears, generating additional questions about proper naming. Lastly, GenBank annotations do not always contain obvious references to the actual sample source, metabolism, or ecology of the isolate. Thus, completing the final column of requested data requires that students perform some detective work, searching the publication or project title, the abstract link (if available), or their textbook. The Instructor Dataset Key **(Appendix 5, see below)** indicates summary information for each accession number based on information available on NCBI: GI#, Classification, Genus/Species Name, Publication or Project Title, and Key Points from the Abstract (if available). In the last column of this key, comments are provided on whether students will have to consult their text. This key should be viewed as highly exhaustive, with typical students recording only a subset of the requested information. For example, students often recognize that Pseudomonas (accession = U71007) is a chemotroph, but do not specify that it degrades hydrocarbons, an important metabolic skill for bioremediation applications. In terms of providing predictions about the relationships of control sequences, students often erroneously predict that phylogeny follows metabolism (e.g. they may incorrectly predict that all phototrophs will cluster together). Students may also propose that the tree will follow cell structure or morphology which, in some cases, will be supported (e.g. Gram Positives, a coherent phylum, should form a common cluster).

Accession	**GI#**	**Kingdom/Domain Phylum, Class**	**Genus/Species**	**Title Information** (Cut/Pasted from NCBI)	**Abstract Utility**	**Text Utility**
AJ309733	13160152	Bacteria; Aquificae;	Aquifex aeolicus	The complete genome of the	Identifies organism as ancient,	Confirms NCBI. Will provide

		Aquificales		hyperthermophilic bacterium Aquifex aeolicus	thermophilic, and performing chemolithotrophic autotrophy.	students with ideas about ecology.
AB039769	14587794	Bacteria; Thermotogae; Thermotogales	Thermotoga thermarum	…Hyperthermophilic bacteria from the Kubiki oil reservoir Niigata, Japan	Identifies organism as from a Japanese oil reservoir, and performing anaerobic fermentative heterotrophy.	Confirms NCBI. Will provide students with more ideas about ecology and metabolism.
X87140	1498248	Bacteria; Spirochaetes; Spirochaetales	Treponema maltophilum	Treponema maltophilum sp. nov., a small oral spirochete isolated from human periodontal lesions	Only describes media development issues for dental isolates.	Text necessary for ecology and metabolism.
M64310	173941	Bacteria; Spirochaetes; Spirochaetales	Borrelia burgdorferi	Phylogenetic analysis of the genus Borrelia: a comparison of North American and European isolates of Borrelia burgdorferi	Only describes DNA detection and comparison methods for examining geographically isolated isolates.	Text necessary for ecology and metabolism.
AF170738	5733749	Bacteria; Bacteroidetes; Flavobacteria	Cryomorpha ignava	…Cryomorpha ignava….novel flavobacteria isolated from various polar habitats	Identifies organism as from Antarctic/Southern Ocean particulates and quartz stone subliths, and as strict aerobes.	Confirms NCBI. Will provide students with more ideas about metabolism.
X85249	1806270	Bacteria; Planctomycetes; Planctomycetacia	Planctomyces sp.	…planctomycete bacteria from postlarvae of the giant tiger prawn…	Focuses on media development issues for Planctomyces.	Confirms NCBI. Will provide students with more ideas about metabolism.
AF513475	21667923	Bacteria; Cyanobacteria; Chroococcales.	Synechococcus sp. PH40	The Hawaiian Archipelago: A Microbial Diversity Hotspot	UNPUBLISHED - none	Text necessary for ecology and metabolism.
AY701541	51105617	Bacteria; Cyanobacteria; Nostocales	Anabaena affinis	Phylogenetic analysis of planktonic species of Anabaena	UNPUBLISHED - none	Text necessary for ecology and metabolism.
AB116143	33468807	Bacteria; Firmicutes; Bacillales	Bacillus coagulans	Low GC Gram Positive bacterium isolate from compost	UNPUBLISHED - none	Text necessary for ecology and metabolism.
AJ786665	53749598	Bacteria; Firmicutes; Lactobacillales	Lactobacillus paracollinoides	Taxonomic note 'L. pastorianus' …a former synonym for L. paracollinoides	Only resolves classification issue.	Text necessary for ecology and metabolism.
M11212	148823	Bacteria; Firmicutes; Clostridia	Heliobacterium chlorum	Gram-positive bacteria: possible photosynthetic ancestry	Identifies this unique anoxygenic Gram Positive phototroph, mentioning unique pigments.	Confirms NCBI. Will provide students with ideas about ecology.
AY079156	28627573	Bacteria; Actinobacteria; Actinobacteridae	Streptomyces koyangensis	Streptomyces koyangensis sp. nov., a novel actinomycete that produces 4-phenyl-3-butenoic acid	Identifies organism as from Korean soil, emphasizing Streptomyces-like traits.	Text useful for more information about ecology and metabolism.
AJ871293	57434252	Bacteria; Actinobacteria; Actinobacteridae	Streptomyces cyaneus	…Diversity of biological soil crusts in the Colorado Plateau studied by molecular fingerprinting and intensive cultivation	UNPUBLISHED - none	Text useful for more information about ecology and metabolism.
AB128051	39748703	Bacteria; Proteobacteria; Alphaproteobacteria	Gluconobacter thailandicus	…An acetic acid bacterium in the alpha-Proteobacteria	Identifies organism as from Thai flowers, emphasizing carbon utilization as compared with close relatives.	Text useful for more information about ecology and metabolism.
AB109890	39652274	Bacteria; Proteobacteria; Betaproteobacteria	Herbaspirillum putei	Proposals of Herbaspirillum putei sp. nov. for bacterial strains isolated from well water.	Only resolves classification issue.	Text necessary for more information about ecology and metabolism.
U71007	2264415	Bacteria; Proteobacteria; Gammaproteobacteria	Pseudomonas putida	Molecular characterization of an n-alkane-degrading bacterial Community…	Identifies organism as from contaminated Venice lagoon, emphasizing carbon utilization as compared with close relatives.	Text useful for more information about ecology and metabolism.
AY268891	55584191	Bacteria; Proteobacteria; Deltaproteobacteria	Desulfofaba fastidiosa	Characterization of the marine propionate-degrading, sulfate-reducing bacterium Desulfofaba fastidiosa sp. nov.	Identifies organism as from sulfate-methane transition zone of marine sediment, emphasizing anaerobic reduction of sulfur to support chemoheterotrophy.	Text useful for more information about ecology and metabolism.
Z25741	397417	Bacteria; Proteobacteria; Epsilonproteobacteria	Helicobacter pylori	Analysis of ribosomal RNA genes of Helicobacter pylori	UNPUBLISHED - none	Text necessary for more information about ecology and metabolism.
D38365	550527	Bacteria; Chloroflexi; Chloroflexales	Chloroflexus aurantiacus J-10-fl	…A filamentous phototrophic bacterium which forms dense cell aggregates by active gliding movement	Identifies organism as from Japanese hot spring mats, emphasizing ability to grow as anaerobic photoheterotrophs or aerobic chemoheterotrophs.	Text useful for more information about ecology and metabolism.

Uploaded Dataset Check

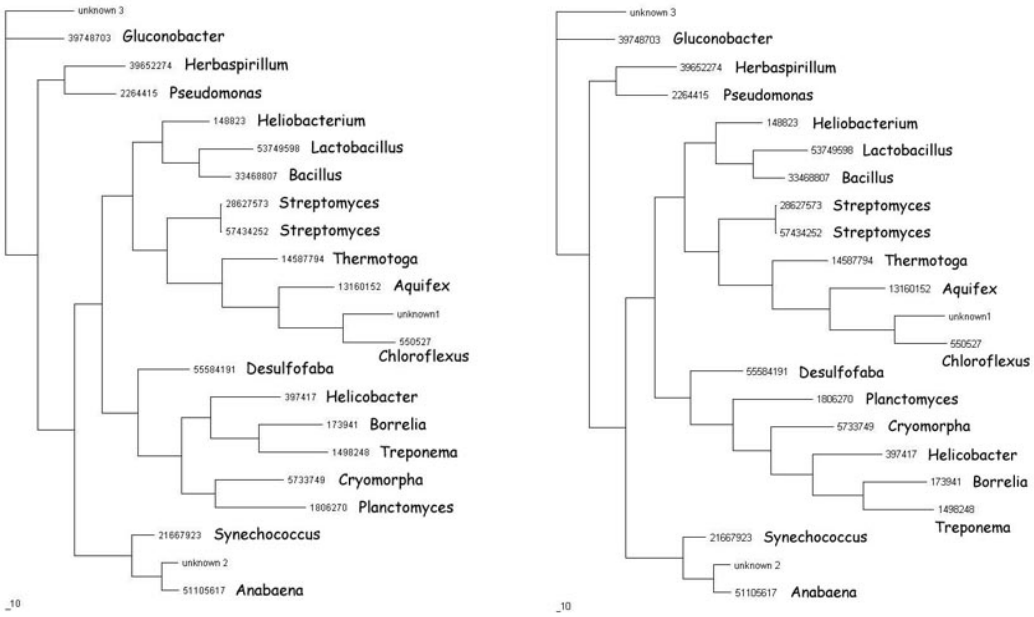
Students typically encounter initial difficulties working with Biology Workbench, forgetting to select correct sessions, projects, or programs, and - during Ndjinn database searching - failing to select the GBBCT (GenBank Bacterial Sequences) database. Instructors should, minimally, confirm that students have correctly loaded all dataset members before the next lab session. A sample of the expected dataset printout is shown in **Figure 1.**

Alignment Data

In addition to needing some reminders about session, project, and program selection, students may have questions about ClustalW settings options, none of which are changed from default. Minimally, instructors should confirm that students have correctly aligned the dataset. A sample of the expected alignment printout (page 1 only, as requested from students) is shown in **Figure 2.** Instructors should confirm that the complete dataset

appears on the top of the page. They should also note that the order of the dataset has changed, a result of ClustalW alignment algorithms ranking each member by similarity. Students may note this and ask why.

Figures Three and Four – Two most probable phylogenetic trees based on these parameters, this dataset.



PAUP Analysis and TreeView
Students typically encounter initial difficulties working with PAUP, most often because they fail to precisely follow directions. To ameliorate frequent problems that arise during the targeted saving of the .tre file, it is strongly recommended that instructors carefully plan and assign the final destination address for these files – preferably a non-networked location. Instructors who use a shared network location need to be aware of the fact that different student teams could over-write one another's files if they save using common names. An annotated sample of the two expected phylogenetic trees is shown in **Figures 3 and 4**; Comic Sans text annotations represent genus names that students should have hand-written adjacent to respective accession numbers. Instructors may ask students to print out and both trees and compare them for differences as part of the final discussion question. If this is done, students should note some discrepancies in the cluster that includes the Spirochetes.

In terms of analyzing the final tree, discussion questions effectively lead students toward understanding basic phylogenetic concepts while reinforcing microbial diversity. After dealing with all the precise and complex information and procedures in this exercise, students are often surprised at how simple tree-based identification of each unknown is (question 2). The predicted identity of each unknown can be deduced by examining the closest neighbor. For example, unknown 1's nearest neighbor is Chloroflexi (phylum)/Chloroflexus (genus), an anoxygenic phototroph. Unknown 2's nearest neighbor is Cyanobacteria (phylum)/Anabaena (genus), an oxygenic phototroph. Unknown 3's nearest neighbor is Proteobacteria (phylum)/Gluconobacter, a chemoheterotrophic acetic acid bacterium.

In terms of question 3 (how genetically similar is each unknown to each of its nearest neighbors?), students will have to use rulers to measure the total horizontal line distance between each unknown and its nearest neighbor. They will then have to compare this value to the legend to determine how many nucleotide differences are represented by this line length. Typical mistakes encountered include: measuring both horizontal and vertical distances, and failing to convert line length to nucleotide distance using the legend length. The legend line corresponding to 10 nucleotides is approximately 1.7-8 mm. Given this, the approximate answers for each unknown are as follows: Unknown 1 is about 260 nucleotides different from Gluconobacter; Unknown 2 is about 150 nucleotides different from Chloroflexus, and Unknown 3 is about 70 nucleotides different from Anabaena. Using more sophisticated applications within PAUP and TreeView, it is possible to determine and print exact nucleotide differences between dataset members but a discussion of these methods is beyond the scope of this report.

Student responses to question 4 (Did your tree support all your predictions about the relationships between control bacteria? Explain. What does the tree suggest about how useful metabolism and ecology are in terms of phylogeny? What does the tree suggest about the evolution of photosynthesis, in particular?) will vary widely. If students have proposed metabolism-driven hypotheses (e.g. phototrophs will all cluster together), they should come to recognize that metabolic types are, in fact, widespread. For example, phototrophic microbes include distantly related Gram Positives (Heliobacterium), chemotrophic thermophiles (Chloroflexus), and Cyanobacteria (Synechococcus and Anabaena). If students have hypothesized certain cell structures or shapes will form distinct groupings, they will observe varying results. For example, although bacteria with a Gram Positive wall will cluster together, bacteria with a Gram Negative wall are widely dispersed. Although Spirochete bacteria form a distinct cluster, other shapes (e.g. filaments and rods) are widely dispersed. One accurate prediction students may make is that all Proteobacteria will cluster together. Unfortunately, this selected dataset splits Proteobacteria into three regions (Alpha, Beta/Gamma, and Delta/Epsilon), a result of using a small and extremely diverse dataset.

Safety Issues.
None.

ASSESSMENT and OUTCOMES

Suggestions for Assessment.
Student pairs turn in the following lab-generated assignments for grading: (1) Control Dataset Worksheet (10 points); (2) Biology Workbench Upload Summary Print-out (2 points); (3) Aligned Data, page 1 only (2 points); (4) Final Tree, Including Hand-Labeling (6 points); and (5) Discussion (10 points). The combined value of this assignment is 10% of the lab assignment grade (30/300 total points).

Of these assignments, the Control Dataset defines the stage for all further data interpretation and should be graded the most carefully. Students should be rewarded for acute detective work in researching less obvious features of the sequences that regard source, ecology, and metabolism. For example, students who simply put should phrases like "environmental sample" instead of more advanced information (e.g. from Japanese oil reserve) should not receive full credit because a key element of this exercise is understanding the breadth of DNA-based research that is occurring in many unusual places around the world.

The Final Tree and Discussion should also be checked carefully in terms of comments made in the previous section (PAUP Analysis/TreeView). The following point distribution is used: 3 pts. for question 2, 2 pts. for question 3, and 5 pts. for discussing their hypotheses in light of actual tree data.

At the end of the term, students are assessed via a written lab exam, with 10% (15/150 total points) covering this lab exercise. Students are asked a combination of multiple choice and short answer questions dealing with the function and purpose of key computational resources used in this exercise: NCBI, GenBank, BW, ClustalW, and PAUP. Although students are not asked to physically use computers during this exam, GenBank data record print-outs are provided and students are asked to analyze them for specific information (e.g. Phylum, Genus, source).

Field Testing.
Since being developed in 1998, approximately 100 junior- or senior-level undergraduate Biology Majors have completed this curriculum. Of these, about 20 represent Molecular Biology students, 20 Computational Students, and 60 General Microbiology students. Most (50-60%) students were pursuing careers in the health sciences. The remaining students sought careers in secondary education and research (academic, government, or biotechnology).

Student Data.
Since 2003, we have completed assessment of lab curricula in General Microbiology, which serves a maximum of 16 students per term. Twenty-two students rated this curriculum on a 10-point scale (10 = Excellent; 1 = Poor) in Fall 2003 and Spring 2004, as summarized below:

Please Rate The Statement: This Lab…	Average Rating
Made Connections Beyond Microbiology	7.75
Improved My Awareness of Microbial Diversity	9
Improved My Interest in Microbial Diversity	8.4
Enhanced My Interest In Scientific Research	9.2
Enhanced My Ability To Use Computers	7.6
Exposed Me To New Technology	9.3
Enhanced My Organizational Skills	7.7
Enhanced My Writing Skills	6.4
My Overall Rating Of This Lab Is	9

SUPPLEMENTARY MATERIALS

Possible Modification.
This exercise can be modified in many ways, some based on direct experience running more advanced versions of this exercise in other courses. As previously suggested, instructors can use these procedures to develop alternative datasets and unknowns to address different questions in microbiology. For example, we have developed specific new datasets and trees that examine the phylogenetic relationships between different metabolic groups (e.g. only photosynthetic bacteria, nitrogen cycling bacteria, sulfur cycling bacteria, etc.). Similar datasets could be developed to address epidemiological issues using viruses or pathogenic bacterial datasets. In addition to dataset modifications, we have provided two additional modifications for instructors who either have either less or more time to devote to this topic.

Shorter Modification
For instructors who do not have adequate time or PAUP software licenses, it is possible to provide students with just a final tree (showing accession or GI numbers and a legend) and have them use NCBI/GenBank to solve the dataset and interpret the unknowns (i.e. completing the Control Dataset Worksheet and the Discussion as written). This exercise has been effectively run with both undergraduate biology majors, as well as adapted for on-line distance-education curriculum for secondary science teachers and pre-college students.

Year	Student	Research Project Topic	Career Interests
2001	Junior	Bacterial flagellin – horizontal gene transfer	Government/Research
2001	Senior	Herpes protease – acyclovir resistance over space	Pre-Veterinary
2001	Junior	Polio protease – analysis of Haiti epidemic	Industry/Biotechnology
2001	Senior	Influenza HA – vaccine prediction	Industry/Biotechnology
2002	Senior	Photosystem I P70– evolution of photosynthesis	Academic/Research
2002	Senior	Luciferase – prokaryotic vs. eukaryotic homologs	Education
2004	Sophomore	Multi-drug efflux pumps – horizontal gene transfer	Pre-Medicine
2004	Post-Baccalaureate	HIV RT – geographical relationships	Pre-Medicine
2004	Junior	West Nile – evolution over geographical distance	Undecided
2004	Senior	SARS protease – origin/structure of proteases	Pre-Medicine
2004	Junior	Hydrogenase - prokaryotic vs. organellar homologs	Government/Research
2004	Senior	Electron Transport Chain PDX – evolution of ETC	Pre-Dentistry

Longer Modification
As has been done in our Computational Biology/Bioinformatics Course (Biology 301), students can fully develop and analyze their own datasets to address a microbiological question that interests them. A summary of microbiology-driven projects students have undertaken in this course is included in **Appendix 6 (shown above)**. Although meaningful, this exercise is extremely challenging and should only be attempted with senior-level biology majors and by instructors with advanced training in phylogenetic analysis. Common student mistakes include choosing non-homologous sequences (e.g. mixing hemagglutinin and neuraminidase sequences during an influenza project), and including variable-length sequences (e.g. mixing full-length 1500 base pair sequences, 200-300 base pair PCR products, and whole genome entries). Including and emphasizing these parameters on dataset-building worksheets will help. Instructors who embark on this project should give students 4-6 additional lab sessions to research and select chosen projects and datasets. In our full course, we spend 4 weeks (8 full labs) on this project, requiring that students include and analyze relevant protein structure data alongside their alignments and final trees.

Acknowledgements.
This work was supported, in part, by an NSF Microbial Observatories/Research at Undergraduate Institute grant (NSF-MO/RUI 0237167).

References.

1. **Baxevanis, A.D., Ouellette, B.F.F.** 1998. Bioinformatics, A Practical Guide to the Analysis of Genes and Proteins, First ed., John Wiley & Sons, Inc. New York, NY.

2. **Campbell, A. M.** 2003. Public access for teaching genomics, proteomics, and bioinformatics. Cell Biol Educ. **2:**98-111.

3. **Cohen, J.** 2003. Guidelines for Establishing Undergraduate Bioinformatics Courses. Journal of Science Education and Technology **12:**449-456.

4. **Felsenstein, J.** 2005. PHYLIP (Phylogeny Inference Package) version 3.6. Distributed by the author at Department of genome Sciences, University of Washington, Seattle (http://evolution.genetics.Washington.edu/phylip.html).

5. **Honts, J. E.** 2003. Evolving strategies for the incorporation of bioinformatics within the undergraduate cell biology curriculum. Cell Biol Educ. **2:**233-237.

6. **Madigan, M. T., J. M. Martinko, and J. Parker.** 2003. Brock Biology of Microorganisms, Tenth ed. Prentice Hall, Upper Saddle River, NJ.

7. **Page, R.D.M., Homes, E.C.** 1998. Molecular Evolution, A Phylogenetic Approach, Third ed., Blackwell Science, University Press, Cambridge.

8. **Swofford, D. L.** 2000. PAUP 4.0b10. Sinauer Associates Inc., Sunderland, MA.

9. **Thompson JD, H. D., Gibson TJ.** 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res. **22:**4673-80.

10. **Woese, C. R.** 1987. Bacterial evolution. Microbiol Rev **51:**221-71.

Additional Appendices.

APPENDIX ONE - Lab Session One Procedures

1. Solving the Dataset Using NCBI

About the National Center for Biotechnology Information (NCBI)

NCBI, supported by federal funding (NIH/NLM) since 1988, archives molecular information via several databases, including Nucleotide (GenBank), Protein, Genome, and Structure. Each data entry is minimally annotated with the following information - all of which is searchable. It is especially useful to click available publication links for published sequences because abstracts have additional information.

- LOCUS = name, length, molecule, date
- ACCESSION = unique assigned identifier (assigned by NCBI at the time of acceptance)
- VERSION = accession version plus GI:number (submission number; e.g. 1 = first submitted)
- SOURCE = organism of origin, common name
- ORGANISM = whole organism, detailed classification and taxonomy
- AUTHORS = submitting scientists
- TITLE = publication, manuscript, and/or project title, links to abstract (minimally), possibly article
- FEATURES = biological markers including, in some cases, physical source and structural motifs

Task – Identifying and Understanding Your Controls Using NCBI

In this lab, your overall goal will be to use known bacterial sequences to identify new/unknown bacterial sequences from Yellowstone hot springs. The first step – understanding your controls.

Procedures

Go to NCBI and note the "menu bar" items at the top; set the search bar to nucleotide (i.e. GenBank).
Enter each accession number listed below – one at a time - and hit "go."
One "hit" will come up; click on the accession number - this will take you to the annotated flatfile.
Read the annotations and your text to determine requested information to complete the table
Repeat for each accession number, recording all pertinent information as you go.
You may want the NCBI Control Worksheet open so you can cut/paste from NCBI as you work.

Turn In/Assignment

Complete the attached NCBI Control Worksheet (**APPENDIX 2**).

2. Uploading the Dataset Using BW

About the Biology Workbench (BW)

BW, free since 1996, is supported by the San Diego Supercomputer Center (SDSC), with additional federal funding (e.g. NSF). It provides 150,000 sessions per month, with each person creating/using individual accounts. BW provides programs and interfaces with other databases (like NCBI). You will use TWO tool bars today: SESSION (start/choose projects) and NUCLEIC (uploading data). Within each, there are MANY pieces of software for different functions. We will only use a few of these options. If you want to learn more, download and read the available on-line BW tutorial (http://peptide.ncsa.uiuc.edu/tutorials_current/How3.2/).

Task – Uploading Your Controls to Biology Workbench (BW)

NCBI archives information but provides limited capacity for storing and analyzing complex datasets. For this analysis, scientists turn to an array of different software, ranging from extremely expensive to on-line/free. We will use BW to build our dataset today. Next time, we will use different BW tools to analyze our dataset.

Procedures – Session Tools
Open BW, http://workbench.sdsc.edu/, and set up an account. Record your user name/password.
Using your account, select and open SESSION TOOLS
From the SESSION TOOLS box, click on "Start New Session" and hit the RUN button.
A new screen should open, asking for a description. Type "16S Tree" and hit START NEW SESSION.
A new screen should open showing 16S Tree has been created and is selected

Procedures – Nucleic Tools/Ndjnn
Now select and open NUCLEIC TOOLS. A new window with a new box of different tools will appear.
Select Ndjinn - Multiple Database Search and hit RUN button (Ndjinn = Engine, as in search engine).
Enter the accession number in the text box and select/check GBBCT (GenBank Bacterial Sequences).
GBBCT is one of many databases that BW can link to; you will have to scroll to find it on the screen.
This tells BW to connect to NCBI/GenBank. **Click the SEARCH button at the top of page.**
A new window with one hit should come up. Confirm it is correct using your worksheet records.
Click the checkbox next to the sequence hit and select IMPORT SEQUENCE(S).
You will be taken back to the NUCLEIC TOOLS page, now showing the newly imported sequence.
Repeat these steps for all accession numbers within your dataset.
When finished, do not shut BW down - proceed to part three, next…

3. Uploading the Unknowns

Task – Uploading Your Unknowns to Biology Workbench (BW)
In addition to the controls, you will now cut/paste three unknown DNA sequences to your 16S Tree Session. In our case, unknowns represent new isolates from Yellowstone hot spring communities.

Procedures
From within 16S Tree/NUCLEIC TOOLS, select "Add New Nucleic Acid Sequence" and hit RUN.
After the new screen comes up, type in the appropriate label (e.g. Unknown 1).
Now, cut/paste the each sequence from the Unknown Sequences document (**APPENDIX 3**)
Make sure that you cut/paste all information from each sequence entry, starting with >name
Each Unknown should look like this in BW. After confirming, click SAVE.

ADDED SEQUENCE	**Label:** Unknown 1 **Sequence:** >Unknown 1 TAGAGGTTGATCCTGGCTCAGAACGAACGCTGGCGGCAGGCCTAACACATGCAAGTCCGAGCGCACCCTT CGGGGTGAGCGGCGGACGGGTGAGTAACGCGTGGGAATATACCCTTCTCTACGGAATAGTCTCGGGAAAC…

Note About Your Controls
Control Labels appears as: Aquifex pyrophilus 16S ribosomal RNA gene, partial sequence. Control Sequences appear as: >37222674 (return) TTCCCT… In this case, >37222674 is the GI number. In other cases, it may be the Accession Number. Regardless, it is this value (what appears after the ">") that will carry through on your final tree. Thus, it is important that you keep track of all levels of data (GI, accession, and name). If you lose information, you can re-search NCBI/GenBank.

Turn In/Assignment
Print out your final uploaded 16S Tree dataset - controls and unknowns – by going to File and selecting Print. Although you have no basis for predicting what your unknowns are at this time, you should be able to speculate on some controls that will cluster together and why. Formulate predictions about THREE groups that should form distinct clusters and explain your reasoning, using evidence from your NCBI Control Worksheet. At least one of your predictions should focus on photosynthetic microbes.

APPENDIX TWO - NCBI Control Worksheet

Name(s):
Date:

Accession	GI#	Genus	Phylum	Where Did Isolate Come From?	Metabolism/Ecology *May need to consult your text*
AJ309733					
AB039769					
X87140					
M64310					
AF170738					
X85249					
AF513475					
AY701541					
AB116143					
AJ786665					
M11212					
AJ871293					
AY079156					

AB109890					
U71007					
AY268891					
Z25741					
D38365					

Did you remember to attach the final print-out of your Biology Workbench loading exercise?
Did you summarize hypotheses about control clustering and why?

APPENDIX THREE - Unknown Sequences

>Unknown 1
TAGAGGTTGATCCTGGCTCAGAACGAACGCTGGCGGCAGGCCTAACACATGCAAGTCCGAGCGCACCCTT
CGGGGTGAGCGGCGGACGGGTGAGTAACGCGTGGGAATATACCCTTCTCTACGGAATAGTCTCGGGAAAC
TGGGGGTAATACCGTATACGCCCTTCGGGGGAAAGATTTATCGGAGAAGGATTAGCCCGCGTTGGATTAG
GTAGTTGGTGGGGTAATGGCCTACCAAGCCGACGATCCATAGCTGGTTTGAGAGGATGATCAGCCACACT
GGGACTGAGACACGGCCCAGACTCCTACGGGAGGCAGCAGTGGGGAATCTTAGACAATGGGCGCAAGCCT
GATCTAGCCATGCCGCGTGAGTGACGAAGGTCTTAGGATCGTAAAGCTCTTTCGCTGGGGAAGATAATGA
CTGTACCCAGTAAAGAAGTCCCGGCTAACTCCGTGCCAGCAGCCGCGGTAATACGGAGGGGACTAGCGTT
GTTCGGAATTACTGGGCGTAAAGCGCACGTAGGCGGATTAGCAAGTTAGGGGTGAAATCCCGGGGCTCAA
CCCCGGAACGGCCCTTAAAACTGCTAGTCTAGAGTTCGAGAGAGGTGAGTGGAATTCCGAGTGTAGAGGT
GAAATTCGTAGATATTCGGAGGAACACCAGTGGCGAAGGCGGCTCACTGGCTCGATACTGACGCTGAGGT
GCGAAAGCGTGGGGAGCAAACAGGATTAGATACCCTGGTAGTCCACGCCGTAAACGATGAATGCCAGACG
TCGGCAAGCATGCTTGTCGGTGTCACACCTAACGGATTAAGCATTCCGCCTGGGGAGTACGGTCGCAAGA
TTAAAACTCAAAGGAATTGACGGGGGCCCGCACAAGCGGTGGAGCATGTGGTTTAATTCGAAGCAACGCG
CAGAACCTTACCAACCCTTGACATGGTTATCGTAGTTACCAGAGATGGTTTCGTCAGTTCGGCTGGATAA
CACACAGGTGCTGCATGGCTGTCGTCAGCTCGTGTCGTGAGATGTTCGGTTAAGTCCGGCAACGAGCGCA
ACCCACACCCTTAGTTGCCAGCATTCARTTGGGCACTCTAGGGGAACTGCCCGTGATAAGCGGGAGGAAG
GTGTGGATGACGTCAAGTCCTCATGGCCCTTACGGGTTGGGCTACACACGTGCTACAATGGTAGTGACAA
TGGGTTAATCCCAAAAAGCTATCTCAGTTCGGATTGGGGTCTTGCAACTCGACCCCATGAAGTCGGAATC
GCTAGTAATCGCGTAACAGCATGACGCGGTGAATACGTTCCCGGGCCTTGTACACACCGCCCGTCACACC
ATGGGAGTTGGGTCCACCCGAAGGCCGTGCGCCAACCAGCAATGGGGGCAGCGGACCACGGTGAGCTTAG
CGACTGGGGTGAAGTCGTAACAAGGTAACCAGGTGG

>Unknown 2
GATGAACGCTGGCGGCGTGCTTAACACATGCAAGTCGAACGAAGTCTTCGGACTTAGTGGCGGACGGGTG
AGTAACGCGTGAGGACCTACCCTAAGGACGGGGACAACAGTTGGAAACGACTGCTAATACCCGATGTGCC
GAGAGGTGAAAGATTTATCGCCTAAGGATGGACTCGCGTCAGATTAGCTAGTTGGTGTGGTAACGGCATA
CCAAGGCAACGATCTGTAGCTGGTCTGAGAGGATGATCAGCCACACTGGGACTGAGACACGGCCCAGACT
CCTACGGGAGGCAGCAGTGGGGAATTTTCCGCAATGGGCGAAAGCNTGACGGAGCAACGCCGCGTGTGGG
AGGAAGGTTTTAGGACTGTAAACCACTTTTCTCAGGGAAGAAGATCTGACGGTACCTGAGGAATCAGCAT
CGGCTAATTCCGTGCCAGCAGCCGCGGTAATACGGGAGATGCAAGCGTTATCCGGAATTATTGGGCGTAA
AGCGTCCGCAGGCGGTCTTATAAGTCTGTCGTTAAAGCACACAGCTTAACTGTGGGAGAGCGATGGAAAC
TGTGAGACTAGAGTGCGGTAGGGGTAGGGGGAATTCCCGGTGTAGCGGTGAAATGCGTAGATATCGGGAA
GAACACCAGCAGCGAAAGCGCCCTACTGGGCCGCAACTGACGCTCATGGACGAAAGCTAGGGGAGCGAAA
GGGATTAGATACCCCTGTAGTCCTAGCCGTAAACGATGGACACTAGGTGTTGTCTGTATCGACCCGGACA
GTGCCGTAGCTAACGCGTTAAGTGTCCCGCCTGGGGAGTACGCACGCAAGTGTGAAACTCAAAGGAATTG
ACGGGGGCCGCACAAGCGGTGGAGTATGTGGTTTAATTCGATGCAACGCGAAGAACCTTACCAGGGCTTG
ACATCCCGCGAATCTCTGTGAAAGTGGAGAGTGCTTCGGGAGCGCGGAGACAGGTGGTGCATGGCTGTCG
TCAGCTCGTGTCGTGAGATGTTGGGTTAAGTCCCGCAACGAGCGCAACCCTCGTTCTTAGTTGCCATCAT
TAAGTTGGGCACTCTAGGGAGACTGCCGGTGACAAACCGGAGGAAGGTGGGGACGACGTCAAGTCATCAT
GCCCCTTACGTCCTGGGCTACACACGTACTACAATGCTAGGGACAAAGAGCAGCCAACTCGCGAGAGTGA
GCTAATCTCATAAACCCTGGCTCAGTTCGGATTGCAGGCTGCAACTCGCCTGCATGAAGTAGGAATCGCT
AGTAATCGCAGGTCAGAATACTGCGGTGAATACGTTCCCGGGCCTTGTACACACCGCCCGTCACACCATG
GAAGTTGGCCACGCCCGAAGTCGCTACCCTAACCGTTCGCGGAGGGGGGCGCCGAAGGCAGGGCTGATGA
CTGGGGTG

>Unknown 3
GCTTGGTACCGAGCTCGGATCCACTAGTAACGGCCGCCAGTGTGCTGGAGTTCGCCCTTGCGGATCCGCG
GCCGCTGCAGAGTTTGATCCTGGCTCAGGACGAACGCTGGCGGCGTGCCTAATGCATGCAAGTAGCACGC
ACCCGCTTCGGCGGGTGAGTGGCGCACGGCTGAGTAACACGTGGGAACCCGCCCCCCGGTGGGGGATAAC
GCGACGAAAGTAGCGCTAATCCCGCATACGTCCCCAGGGAGAAAGCGCAGTGCGCGCCAGAGGGAGGAGC
CTGCGAGCCCATCAGGTCGTTGGTGGGGTAAGGGCCTACCAAGCCGATGACGGGTAGCTGGTCTGGGAGG
ATGACCAGCCAGACTGGGACTGAGACACGGCCCAGACTCCTACGGGAGGCAGCAGCAAGGAATTTTCGGC
AATGGGCGCAAGCCTGACCGAGCAACGCCGCGTGCAGGATGACGGCCTTCAGGGTTGTAAACCGCTTTTC
GGGGGGACGATGATGACGGTACCCTCGGAACAAGCCCCGGCTAACTCTGGGCCAGCAGCCGCGGTAAGAC
AGAGGGGGCGAGCGTTGTCCGGAGTCACTGGGCGTAAAGCGCGCGCAGGCGGCCAACTCAAGTGTTGTGT
GAAAGCCCCCGGCTCAACCGGGGGAGGTCATGGCAAACTGGGTCGACTCGAGCGTCGGAGAGGCCCCTCG
AATTGCCGGTGTAGCGGTAAAATGCGTAGAGATCGGCAGGAAGACCAAGGGGGAAGCCAGGGGGCTGGCC
GCTAGCTGACGCTGAGGCGCGACAGCGTGGGGAGCAAACCGGATTAGATACCCGGGTAGTCCACGCCGTA
AACGATGACCACTCGGCGTGTGGCGACTATTGACGTCGCGGCGCGCCTAGCTCACGCGATAAGTGGTCCG
CCTGGGAACTACGAGCGCAAGCTTAAAACTCAAAGGAATTGACGGGGGCCCGCACAAGCAGCGGAGCGTG
TGGTTTAATTCGACGCAACCCGAAGAACCTTACCCAGACTGGACATGACGGTGAAAACGGCGGAAACGTC
GTCGGCCTGCGAGGGTCCGTCACAGGTGCTGCATGGCTGTCGTCAGCTCGTGTCGTGAGATGTTGGGTTA
AGTCCCGCAACGAGCGCAACCCCTGCGGTTAGTTACTGCGTGTCTAACCGGACTGCCCTTTGGGGAGGAA
GGCGGGGATGACGTCAAGTCCGCATGGCCCTGACGCCTGGGGCGACACACACGCTACAATGGCGCCGACA
ATGCGTTGCCACCGCGTAAGCGGAGCGCGAATCGCCAAACGGCGCCGCAGTGCAGATCGGGGGCTGCAAC
TCGCCCCCGTGAAGGCGGAGTTGCTAGTAACCGCGTATCAGCCATGGCGCGGTGAATCCGTACCCGGGCC
TTGTACACACCGCCCGTCACGTCATGGGAGTTGTCAATGCCTGAAGTCCGTGGGCTAACCGCGTGAGCGG
AGGCAGCGGCCGAGGGCAGGGGCAGCGACTGGGACGAAGTCGTAAAAACCCCGGCGGCCCGCTCGAGCAA
GGGCGAATTCTGCAGATATCCATCACACTGGCGGCCGCTCGAGCATGCATCTAGAGGGCCCAA

APPENDIX FOUR - Lab Session Two Procedures

1. Aligning Your Dataset

About ClustalW and Multiple Sequence Alignment

ClustalW, intended to be permanently free, was designed by Higgens *et al.* at the European Molecular Biology Laboratory (EMBL). ClustalW compares data in pairwise manner to establish percent similarity, and then ranks the sequences. The rankings are used to establish the order that the actual alignment will take place. The original, technical paper describing ClustalW is available on-line at: www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=pubmed&dopt=Abstract&list_uids=7984417. Although we will use the default ClustalW settings, it is possible to change some alignment parameters, including the penalty for opening and extending a gap (i.e. creating an insertion/deletion), the cost for aligning different amino acid residues (using various protein substitution matrices), and the cost for aligning any different nucleic acid residues .

Task – Preparing Your Alignments

Now that you understand and have uploaded your dataset onto BW, your next task will be to use ClustalW to produce an alignment. After obtaining the alignment data, you will convert the data to a format that can be read by the phylogenetic tree software (PAUP).

Procedures
Open your previously-created account on the BW site, http://workbench.sdsc.edu/
In SESSION TOOLS, select your 16S Tree dataset checkbox, "Resume Session," and hit Run.
A new window should open (still in SESSION), displaying all your entered data from the last lab.
Hit NUCLEIC TOOLS, "Select All Sequences," and hit Run; all sequences should be checked.
With all sequences checked, select "CLUSTALW – Multiple Sequence Alignment" and hit Run.
A new screen will come up with your dataset and several selection parameters; do not change.
Hit Submit – wait (2-5 minutes); once finished, click on Import Alignment(s).
A new window in ALIGNMENT TOOLS will come up.
After selecting your alignment dataset checkbox, select "View Aligned Sequence(s)" and hit Run.
In the Format pull-down box, select Paup/Nexus format; the data below will change accordingly.
Hit "Download/view all sequences in text format" and a new window will appear.
Perform a File/Save As (keeping .txt) and record the file name and where this has been saved.

Turn In/Assignment
Open your Paup/Nexus file in a text program and print ONLY the first page. Beware - the whole document is likely 20-30 pages long.

2. Searching and Analyzing Phylogenetic Trees

About PAUP
PAUP, Phylogenetic Analysis Using Parsimony, is a software package that uses alignment data to search and display phylogenetic trees. Many possible trees are initially constructed and searched using, in this exercise, heuristic (rapid, less computationally demanding) methods. PAUP then reports the most parsimonious trees – those with the fewest assumptions (e.g. the fewest number of assumed genetic changes). Resulting trees are viewed in TreeView, which includes a legend that defines the unit length corresponding to 10 nucleotide differences. On the kind of tree in this exercise, only the horizontal branch lengths are used to determine the nucleotide differences. You will need to use rulers analyze the relationships between the unknowns and their nearest neighbors.

Task – Opening Your Alignment in PAUP and Logging Output
Open PAUP – 3 windows will appear with an Open window in the foreground
Using the "Files of Type" pull-down box, change settings to "All files"
Browse to your .txt file (previously saved as text format for Paup/Nexus in BW)
Hit Open/Execute – and a new Display window with your previous alignment will appear.
Note command box at bottom where you can type things; this is where you will perform all future tasks.

Task – Generating Trees Using Parsimony Methods
In the command box type: **set criterion=parsimony** (and hit Execute)
To conduct a heuristic search, type in the command line: **hsearch** (and hit Execute)
A smaller search box will come up; after the search is complete, click on close when it is done
In the command line, type: **savetrees file=***.tre brlens=yes** (and hit Execute)
The display will tell you that the file has been saved
To view a low-resolution tree output: type **describetrees** (and hit Execute)
This will show up in the display buffer and also in the log file

** is the location; ** is the file name, which has to have the tree extension (.tre). Make sure you record this name and the location where it has been saved. This command line is VERY particular. It is useful to first locate your ultimate file destination and then copy the address line for this location – pasting it directly to the command line above.*

Task - Viewing and Printing Results
To print trees, open the TreeView Program, and then browse/open your tree file
From the Tree Menu, select Phylogram and use print commands to print

When opening your most parsimonious trees, you will have at least 2 trees. Scroll through them using the arrow keys in the upper left corner of the screen. Print at least 2 for turn-in and analysis..

Turn In Assignment/Discussion Questions
(1) Handwrite the GENUS for each of the tree branches.

(2) Based on simple branch location, what is the predicted identity of each of your unknowns. Speculate to the phylum and genus level, and comment on likely metabolism.

(3) Given your previous answer, how genetically similar is each unknown to each of its nearest neighbors. Explain.

(5) Did your tree support your predictions about the relationships between control bacteria? Explain. What does the tree suggest about how useful metabolism and ecology are in terms of phylogeny? What does the tree suggest about the evolution of photosynthesis?