Chapter 4

Bioinformatics, Virtual Labs, and the Human Genome Project

Anne Cordon and Donna Messersmith

Anne Cordon
Department of Botany
University of Toronto
Toronto, ON M5S 3B2
(416) 978-7431
cordon@botany.utoronto.ca

Donna Messersmith, PhD
Precollege Science Education Initiatives
Howard Hughes Medical Institute
301-215-8896
messersmithd@hhmi.org

Anne is a Senior Lecturer at the University of Toronto. She has been the lab and course coordinator for the 1100-student second year course BIO250Y, Introduction to Molecular Cell Biology and the tutorial and course coordinator of the 650-student third year course JLM349S, Eukaryotic Molecular Biology since their creation in 1991 and 1997 respectively. Anne received the Outstanding Teaching Award 2001 and has been the recipient of several Dean's Excellence awards over the years.

Donna received her B.S. in Zoology from the University of Maryland, College Park, and her Ph.D. in Anatomy and Cell Biology from Georgetown University. Donna designs multimedia scientific education resources, including animations and virtual labs, at the Howard Hughes Medical Institute. These resources are available at www.biointeractive.org and on free CD-ROMs. Donna has taught molecular biology and biochemistry at Georgetown University and conducted neuroscience research at the National Institutes of Health and Uniformed Services University of the Health Sciences.

Reprinted From: Cordon, A. and D. Messersmith. 2002. Bioinformatics, virtual labs, and the human genome project. Pages 43-67, 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.

Contents

Introduction

This lab is intended to provide students with an introduction into the important, relatively new area of bioinformatics used increasingly for not only basic research in molecular biology, taxonomy, conservation biology, forensics, and medicine, but also for commercial applications in the pharmaceutical industry and agricultural biotechnology. Bioinformatics is a discipline combining mathematics and biology and the technology supporting bioinformatics includes computational tools and databases.

The Howard Hughes Medical Institute's (HHMI) virtual lab on pathogenic bacterial identification provides an interesting and relevant case study to launch the introduction of some bioinformatics search and analysis tools (specifically Entrez, BLAST, and ClustalW).

Notes for the Instructor

Rationale:

A previous lab we designed for students involved identifying "unknown" sequences (which were taken from various genes in the trytophan operon in *E. coli* studied in lecture) using BLAST, but this exercise was far less effective motivating us to redesign the approach. Not only is the medical case study interesting to students, the reaction to the virtual lab was extremely positive. The lab demonstrators also confirm that the students understand the concepts and bioinformatic tools (BLAST and ClustalW) much better than they did with the previous exercise. This "dry" lab is combined with a pre-existing "wet" lab in which students isolate prokaryotic RNA and view the resulting 16S rRNA after electrophoresis on agarose gels.

Other ABLE workshops have approached bioinformatics differently, and have provided valuable resources for comparison, alternative strategies, or additional background. Of particular note are

1) "What I could teach Darwin using 'Darwin 2000', an interactive web site for student research into the evolution of genes and proteins" (Hershberger 2000). This is another web-based lab, with excellent explanations of the various tools and useful, clear examples building on the evolution of the hemoglobin molecule.
2) "DNA sequencing used to illustrate mutations and evolution" (Gurney *et al.* 2001). This lab is both a wet lab preparing the PCR products and analytical lab using the web databases and software.
3) "Exploring important biological concepts using *Biology Workbench"* (Ball *et al.* 2002). The Biology Workbench is a collection of online tools applicable for many types of studies. The web site *http://workbench.sdsc.edu/* brings together in one site many of the DNA and protein databases as well as software for searching the databases, aligning sequences, and creating inferred phylogenetic trees.

Level and background of students:

To get the most out of the bacterial identification virtual lab and the sequence analysis using BLAST and ClustalW, students should have some background in DNA structure and preferably some lab experience especially doing gel electrophoresis, the polymerase chain reaction, and DNA sequencing. This lab might be used in introductory courses in molecular biology, genetics, or even evolutionary biology. Although we devote lab time to the discussion of this material, it could also be used as a lecture or tutorial supplement because the students do the exercises independently. I feel (and this was substantiated by comments from workshop participants on the evaluation forms) that this lab could be adapted for use in all levels from introductory to advanced, depending on the context and expectations defined by the instructor for the students.

Alternative sites for multiple sequence alignments and phylogenetic tree generation:

During the workshop, Ted Gurney who is very familiar with this software, suggested *http://www.toulouse.inra.fr/multialin.html* as an alternative to ClustalW, because it has useful and different color highlights for similar and different regions. In addition, an alternative site for constructing phylogenetic trees is *http://taxonomy.zoology.gla.ac.uk/rod/treeview.html*. We do not develop the evolutionary relationship theme in this lab, but other instructors might want to expand this area.

The scientists at HHMI responsible for the online material have been exceptionally helpful, and encourage the [free!] incorporation of their material into teaching. Based on our work and suggestions of how to make the site even better, the bacterial ID vlab has been expanded to include additional samples among other proposed additions.

HHMI Virtual Labs and other Interactive Learning Tools

Dennis Liu and Satoshi Amagai, Howard Hughes Medical Institute, Chevy Chase, MD

BioInteractive site (www.BioInteractive.org)

The BioInteractive site has a variety of educational components, all of which are available on the web in some form. The quality and effectiveness of our materials depends on a multidisciplinary team of scientific and educational advisors, animators and software designers, content developers, and evaluation coordinators. Our materials have proven very popular with an international audience spanning the educational continuum from middle school through University. The web site includes teacher guides, streaming video, animations, interactive science demonstrations (Click and Learn), virtual exhibits, and virtual labs. Most of our materials are also available in print, video, and CD-ROM. For the purposes of this brief paper, we'll discuss Animations, Click and Learn "interactives," and Virtual labs. You can access these features through the BioInteractive site, or through the URLs listed below.

Animations (www.hhmi.org/grants/lectures/biointeractive/animations)

Animations are originally developed to accompany lectures and are not initially formatted for online or CD use. Although not intrinsically interactive, animations are useful for revealing hidden worlds, clarifying anatomical relationships, or showing complex processes such as molecular interactions. We have developed both 2D and 3D animations that can be found online and viewed using a tool that we call the Animation Console. The Animation Console is designed to provide a common format and set of tools for all our animations: a single place where they can be indexed and searched. We hope that this makes the animations more standard and easier for an instructor or individual to use. The Console can help to make animations more interactive by use of context specific tabs in the browser window for: teaching tips, background information, and references. We intend to expand our use of these tabs to provide context for using the animations for different educational levels and backgrounds, and also to add inquiry-oriented exercises and questions related to each animation.

Holiday Lecture program (www.holidaylectures.org)

Each December since 1993, the Howard Hughes Medical Institute (HHMI) presents the Holiday Lectures on Science. The program began as a lecture series, delivered by HHMI scientists for high schools students in the greater Washington metropolitan area. Over the years a variety of features and components have been added to reach a wider audience and provide durable year-round resources for science education.

Interactive Demonstrations (www.hhmi.org/grants/lectures/biointeractive/demos)

"Click and Learn," our name for web-based interactive demonstrations, are "a notch up" on the inquiry-interactivity scale. They are designed to function online or on CD, appealing to individual users, but they also may be excellent lecture enhancements; we know of instructors who use them this way. "Click and Learn" interactives are useful for making graphs or illustrations interactive and for helping the learner make complex comparisons. They can include illustrations and animations.

For example, the vertebrate circulatorium allows a user to compare different aspects of the heart and circulatory system among various vertebrate groups. Hopefully a user can be encouraged to think in evolutionary terms and be led to ask good comparative questions.

Virtual Labs (www.hhmi.org/grants/lectures/biointeractive/vlabs)

Virtual Labs are our most interactive products and potentially high on the inquiry scale as well. We know that vlabs provide a very satisfying experience for an individual learner. The vlabs also can be used in a more formal educational setting as supplemental material to prepare or reinforce a wet lab, or even to provide an experience when a wet lab is not possible. Vlabs are useful for revealing science as a process and for carrying a learner through that process treating concepts and methods/technology hand in hand. Vlabs can be made complex like real experimentation, and of course, remediation can be programmed. To date we have labs relating to immunology, neurophysiology, cardiology, and bioinformatics, all found on CD or online at

Bacterial ID Vlab

Bioinformatics is covered in a vlab that we call the Bacterial ID Lab. The Bacterial ID Lab is based on the fact that PCR can be used to sequence the 16s ribosomal subunit from any bacterial species. A large database of sequences then allows that bacterium to be identified, or even previously unknown specimens to be placed in evolutionary relationship to other bacterial species. The lab is largely tutorial in nature currently, but students do get the results from their sequencing experiments and perform an actual BLAST search online with their results; then they must interpret their results (the CD version allows the option of going online or running a simulated BLAST search from the CD). Currently the lab features a single scenario yielding a single sequence from a single unknown. We plan to add other scenarios and unknown sequences. The design of the lab also makes it possible for instructors to furnish their own scenarios and sequences.

One of us (AC) has evaluated the lab with a large class of second-year life science majors at the University of Toronto.

Student Outline

Introduction

Bioinformatics is a discipline combining mathematics and biology. Bioinformatics technology includes the computational tools and databases that support genomic and related research, which encompasses the study of DNA structure and function, gene expression, and protein production. Bioinformatics technology enables the extraction of information that can be used in basic molecular research as well as commercial applications such as drug discovery, clinical diagnostics, and agricultural biotechnology. The feature article by Ken Howard (2000) in the July 2000 issue of Scientific American says it all -- The Bioinformatics Gold Rush.

The spectacular rise of the commercial genomics industry and the broadening application of molecular techniques in biology and medicine have created a commercial market for bioinformatics software, hardware, and services. Jason Reed of the investment banking firm Ascar Gruss and Son in New York City estimates that the total market for bioinformatics tools and services, including custom databases, could exceed $2.0 billion dollars within five years (Howard 2000; Reed 2000).

The U.S. Human Genome Project (HGP) began in 1990, coordinated by the U.S. Department of Energy (DOE) and the National Institutes of Heath (NIH). This mammoth undertaking sparked the initiation of many large-scale genomic research projects including the bacterium *Escherichia coli*

(E. coli), worm *Caenorhabditis elegans (C. elegans)*, fruit fly *Drosophila melanogaster*, and the laboratory mouse (*Mus musculus*). Model organisms offer a cost-effective way to follow the inheritance of genes through many generations in a relatively short time among other applications.

One of the first and most important problems encountered in these genome projects was how to acquire, store, and analyze massive amounts of DNA sequence information. *GenBank*, a major public repository of DNA sequence data, has grown to include roughly 4.86 million individual sequence records representing about 3.86 billion base pairs (as of early 2000) as compared to 0.56 million records in 1995! GenBank contains the full and partial genome sequences of over 670 different organisms, including 27 complete genomes (Reed 2000)

Where did all of these data come from? From local and international academic and government research groups as well as commercial, privately owned companies. Almost all private companies conducting genomic research, such as Celera Genomics, Incyte, Human Genome Sciences, Millennium Pharmaceuticals, have sequenced stretches of human and other organisms' DNA. Some of this privately-generated sequence data has been submitted to public databases like GenBank, while some data remain proprietary.

The Human Genome Sequence

A public consortium made up of research groups at Washington University (St. Louis), Baylor College of Medicine, the Whitehead Institute and the Sanger Center in England announced the completion of the 'rough draft' of the human genome in June 2000 -- approximately 5 years ahead of schedule. The private company Celera Genomics announced the completion of a rough draft in April 2000. The basic sequencing may be nearly over, but the generation of data is accelerating.

Where are we going with this new information? Reporting on presentations given at the eighth international conference on *Bioinformatics and Genome Research* held in June 1999, H. Lim (2000) predicts that as we move from the pre-genomic to the post-genomic era, research methodology has changed from, for example, analysis of a biological problem of a single gene, to that of analysis of biological problems in multiple dimensions. Presenters at the conference predict that the challenges in the post-genomic era will be pharmacogenomics (drugs tailored to an individual's specific genome thus individualized medicine), proteomics (derivations of quantitative information on proteins present in different tissues, cells and sub-cellular fractions) and other “omics” which involve the combination of different molecular complements in the cell.

For good references, see:

Howard, K. 2000 (July). The Bioinformatics Gold Rush. Scientific American. http://www.sciam.com/ (accessed July, 2000).

Lim, H. 2000 (April). Bioinformatics in the Pre- and Post-Genomic Eras. Trends in Biotechnology (TIBTECH) 18:133-135.

Reed, J. 2000 (March). Trends in Commercial Bioinformatics. http://www.oscargruss.com/reports.htm (accessed July, 2000).

General Background

Molecular sequencing

Through the 1980s and 1990s, one of the most important technique available to the molecular biologist was DNA sequencing, by which the precise order of nucleotides in a piece of DNA is determined. Prior to the mid 1970's, it was much easier to sequence proteins and RNA than DNA. For example, the complete amino acid sequence of insulin was deduced by Sanger's laboratory in 1954. This landmark achievement represented the first complete sequence of a protein. With the help of a variety of ribonucleases and chemical strategies, the complete sequences of 5S ribosomal RNAs were deduced for a variety of organisms in the early 1960s. However, only since the late 1970s has rapid and efficient DNA sequencing been possible.

The ability to sequence DNA was dependent on several technical innovations: the development of molecular cloning by Boyer, Cohen and Berg in the early 1970s; the development of rapid DNA sequencing technologies by Sanger and Coulson in the UK (the chain termination method); and the chemical degradation method by Maxam and Gilbert from the USA in the mid 1970s. It became much easier to sequence DNA in comparison to sequencing protein or RNA. Consequently, an exponential increase in the amount of DNA sequence information became available, from which RNA and protein sequences could be directly deduced. The DNA sequence is now the first and most basic type of information to be obtained from a cloned gene. Sequencing DNA 10-20 kb is routine, and most research laboratories have the necessary expertise to generate this amount of information.

One of the major challenges faced in the Human Genome project was obtaining the primary sequence data in a rapid and cost-effective way. New ways to automate the process were developed and the “shotgun” and “directed shotgun” methods of sequencing were perfected. Using these automated methods, prior mapping is not necessary. The DNA is broken into fragments that are then sequenced, followed by assembly done by matching chain termination sequences and known sequence tagged sites (STSs).

Nucleotide Sequence Databases

GenBank is the National Institute of Health's (NIH) genetic sequence database, an annotated collection of all publicly available nucleotide and protein sequences. The unit records represent single contiguous (gap free) stretches of DNA or RNA with additional comments. Presently, all records in GenBank are generated from direct submissions to the DNA sequence databases from the original scientists, who volunteer their records as part of the publication process. (See Appendix 2 to see a sample GenBank record, *E. coli* gene lacZ). Most of the input data are DNA sequences from which protein or RNA is inferred.

GenBank, built by the National Center for Biotechnology Information (**NCBI**) at NIH in Bethesda, Maryland, is part of the International Nucleotide Sequence Database Collaboration, along with its two partners, the DNA Database of Japan (**DDBJ**, Mishima, Japan) and the European Molecular Biology Laboratory (**EMBL**) nucleotide database from the European Bioinformatics Institute (EBI, Hinxton, England). All three centers are separate points of data submission, but all of them exchange updated information daily.

In the early 1980s, there was no common format or electronic submission of data, which slowed the collation of information tremendously. In 1988, the three groups (American, Japanese and European) met and agreed to use a common format for data elements whereby each database update only the records that were submitted to it. This means that each record is owned by the database that created it, that “update clashes” and overwriting records are prevented, and most

importantly, the information in all three databases is compatible as well as accessible to a global community. These database centers are also computational biology centers as it became clear that sequence data couldn't be generated simply by automated means, but need to be proofread by biologists. Additional tools were developed to analyze the information.

Database Searching tools: similarity searching

In this lab, you will use BLAST to search the sequence database. This program is an important research tool because it is a good combination of speed, sensitivity, flexibility, and statistical rigor. BLAST is an acronym for *Basic Local A1ignment Search Tool.* The BLAST search algorithm takes your input sequence and compares it to all known genetic sequences (DNA or protein), identifying the known molecules that have similar sequences.

BLAST is sometimes referred to as a "one-against-all" homology search algorithm, since the input is a single sequence which is compared against all other known sequences. This is in contrast to the *Multiple Sequence Alignment, or MSA* homology search, a "many-against-each-other" search in which a small, defined set of sequences are compared only against each other, not against the entire database.

There are various BLAST programs designed for either nucleotide sequence queries or protein sequence queries. You will be using BLASTN for this lab: *BLASTN* takes a nucleotide sequence (the query or unknown sequence) and its reverse compliment, and searches them against a nucleotide sequence database. (See Appendix 3 for a sample BLASTN search result.)

Another extremely powerful tool is The National Center for Biotechnology Information's (NCBI) Entrez search engine *(http://www.ncbi.nlm.nih.gov/Entrez).* Entrez not only lets you search for genetic sequence database records but interfaces with several databases. Entrez connects to databases of nucleotide or protein sequences, 3-dimensional structures of macromolecules, and even the MedLine bibliographic database via PubMed -- all from this single site.

Did you know?
That the Databases of Human Genes and Inherited Diseases set up at John Hopkins University in Maryland in 1989, moved to the Hospital for Sick Children in Toronto in 1999.

Dr. Lap-Chee Tsui, geneticist-in-chief at the Toronto Hospital for Sick Children is the president of HUGO (Human Genome Organization).

Outline and summary of this bioinformatics project:

Section 1: Bacterial ID Virtual Lab
You will perform this section before coming to lab by connecting to the HHMI web site and using their "virtual Bacterial Identification Lab." The vlab demonstrates how one would isolate and purify a specific bacterial DNA sequence using PCR. You will use the isolated DNA from this vlab in Section 2.

Section 2: Identify unknown bacterium from its DNA sequence of 16S RNA using BLAST
Using the search tool BLAST, you will sequence the DNA (from section 1) and identify the bacterium.

Section 3: Multiple sequence alignment using ClustalW
For this exercise, you will compare the DNA sequences of 16S RNA from five unknown bacteria using Multiple Sequence Alignment tool ClustalW. After conducting multiple sequence alignments, you will construct a phylogenetic "tree" diagram using JalView.

Section 4: Presentations on model organisms
For this section, you will research a model organism used in genomic research, and give a 3-minute presentation along with a one-page written summary. Alternatively, you will do your research (and resulting presentation) on some aspect of the impact of the human genome project.

Bioinformatics Project Evaluation (Total marks = 30)

20 Quiz covering:
- Virtual lab (Section 1): isolation, amplification (PCR), purification (PCR), sequencing, identification of unknown bacterium
- Blast Searches (Section 2) and reading GenBank record
- MSA Searches (Section 3) using ClustalW

[Sample questions are in the Appendix]

5 Oral presentation (Section 4):

5 Written summary (Section 4) of your presentation

Section 1: "Virtual Bacterial Identification Lab": Isolation and Purification of 16S rDNA

Background

For this part of the lab, which you will do on your own before you come to lab, you will connect to the Howard Hughes Medical Institute's (HHMI) web site and use their "Virtual Bacterial Identification Lab." Explanatory sections are summarized or reprinted here from their virtual lab (with permission). The purpose of the HHMI virtual lab is to familiarize you with the science and techniques used to identify different types of bacteria based on their DNA sequence. In the process, you will see how bacteria are grown and specific DNA is isolated and purified using the molecular techniques Polymerase Chain Reaction (PCR) and DNA sequencing. These techniques are fundamental tools and often used in molecular research as well as forensic analysis.

In this specific example, the sequence of DNA used for identifying the bacterium is the region that codes for the 16S subunit of the ribosomal RNA (16S rDNA). From genomic studies, it has been found that different bacterial species have unique 16S rDNA, thus a comparison of this region may be used as a diagnostic test. Imagine you are a pathologist or a pathology lab technician at a well-equipped research hospital. Your task is to identify a bacterial sample received from a clinician of a very sick patient who needs to be on the correct drug regime as quickly as possible.

Why use molecular techniques to identify the pathological (disease-causing) bacterium instead of traditional methods?

Over the years, a battery of tests has been developed to categorize and identify bacteria. Tests include staining and growing bacteria under a variety of conditions. Such procedures typically require vigorously and reliably growing bacterial cultures. Many pathogens grow poorly on solid medium while others grow only in liquid culture, making identification through traditional techniques difficult or impossible. With the aid of molecular methods, however, these limitations can be overcome. In addition, some species of bacteria cannot be differentiated from closely related species through traditional methods. For these species, molecular methods offer the only reliable and convenient means of identification.

Procedure

START NOW, by connecting to

http://www.biointeractive.org/grants/lectures/biointeractive/vlabs.html

As you go through this exercise, you will "perform" these basic steps (all steps are described in the virtual lab you will do.) *[For the quiz, you should be able to answer the questions in italics.]:*

1. Prepare a sample from the patient and isolate whole bacterial DNA. *[How was this done in the vlab?]*
2. Make many copies of the desired piece of DNA. [*What technique do you use? How do you separate the desired DNA sequence from the rest of the DNA? How do you purify your sample?]*
3. Sequence the DNA. *[Briefly describe how this is done.]*
4. In Section 2 below, we will discuss how you will use the DNA sequence information to identify the bacterium isolated from the patient.

Section 2: "Virtual Bacterial Identification Lab": BLAST Search to Identify Bacterium

Background

As stated earlier, it has been found (from genomic studies) that different bacterial species have unique 16S rDNA, thus a comparison of this region may be used as a diagnostic test. Bacterium identification relies on matching the unknown sequence from a particular sample against a database of all known 16S rDNA sequences using a program called **BLAST**.

After identifying the bacterium described in the online virtual lab (SAMPLE A), identify other 16S rDNA sequences (choose from samples B, C, D, F, G) using BLAST.

Why would you use BLAST? (Adapted from HHMI "Virtual Bacterial Identification Lab")

Under what situations would a scientist search sequence databases? As an example, sequence matching can be used to determine whether a newly identified DNA sequence is part of a known gene. In the simplest scenario, if a new sequence is identical or almost identical (except for a few nucleotide changes) to that of a gene in the sequence database, it is reasonable to predict that the new sequence is either part of the same gene or of a closely related gene. But what if two sequences that appear to be different share sections that are identical? How do you know whether the identical sections are due to chance or indicate some meaningful relationship between the two sequences? Sequence analysis using BLAST or another program provides a "similarity score" to help answer this question. (The "similarity score" is discussed further below.)

If the function of a particular DNA sequence is already known (*e.g.*, the 16S rRNA gene we will be working with in this lab), comparing its sequence with that of the same gene from another species of bacterium provides information about the evolutionary relationship between the two bacterial species. The assumption here is that the number of positions that differ in the nucleotide sequence is proportional to the time elapsed, since the two species formed their own lines of descent from a common predecessor. However, not all DNA sequences change at a constant rate over time. For example, it is not at all clear whether all organisms experience similar mutation rates from purely environmental factors (from increased UV exposure, for example). If the DNA sequence has or has had at some point in evolution a functional role, the rate of evolution and selection — which may be related to population size among other things — can affect its rate of change. Moreover, in some cases, mutations are caused by deletions, insertions, and substitutions of long sequences of DNA rather than by single nucleotide changes. Finally, some sequences of DNA encode proteins with very specific structural requirements, and any change may prove unfavorable to the organism. Such sequences therefore do not tolerate change well and tend to remain the same for long periods of time. These are referred to as *"conserved" regions.* In contrast, sequences that can accommodate change more easily are referred to as *"variable" regions.*

Similarity score: Interpreting BLAST search results (Reprinted from HHMI "Virtual Bacterial ID lab.") [You should understand the basic distinction between the BLAST Score and the E value]

Let's consider how one might go about assigning a numerical value to the degree of similarity between two DNA sequences. Suppose we have two sequences as follows:

CGGCAT
CGCGAT

Let's assign one point for each base pair that matches exactly and 0 point for each base pair that does not. We have C-C (match), G-G (match), G-C (no match), C-G (no match), A-A (match), and T-T (match) for a total of 4 points. Under this hypothetical system, the more nucleotides that match up, the higher the score.

When comparing two DNA sequences, it's important to remember that because of evolutionary history, the sequences may have diverged not only by substitution of bases but also possibly by deletions or insertions of bases. This means that the sequences that are being matched may not be exactly the same length but might have gaps. In practical terms, for these two sequences, the best match (for a total of 5 points) is:

CGGC-AT
CG-CGAT

Another possible alignment (for a total of 5 points) is:

CG-GCAT
CGCG-AT

From the simple example above, you can imagine how rapidly sequence comparisons can become complicated as DNA length increases. The statistics for comparing two sequences of DNA are thus highly complicated. Here we cover just the bare essence of the topic so that you can interpret the response from your sequence query.

Let's suppose you do a BLAST search of the following sequence:

TATCGCGTATTGCC

BLAST will come back with a result, starting with the reference of the search program, the number of letters in your sequence, the number of letters in the database, a graphic representation of the sequence matches, and a list of matches. The list of matches is sorted with the best matching sequences shown first. For the sequence we used, the list starts with the following:

Sequences producing significant alignments:	Score (bits)	E Value
gb|AC012156.14|AC012 Homo sapiens chr 12..	28	5.8
ref|NC_001142.1 Saccharomyces cerevisiae...	28	5.8

What does this mean? "Score" is a numerical score assigned by BLAST. In the simple example, we used earlier, we simply assigned 1 point for matches, and 0 point for non-matches. In BLAST, the scoring system uses "bits" as the measure of information. For DNA, each position can be occupied by either T, A, C, or G. Each match therefore contains 2 bits of information (only 1 is correct out of

4 possible). For a 14-nucleotide-long sequence like ours, the maximum match score then is 28 bits. The higher the score, the better the match.

"E-value" is the number of hits one can expect to see just by chance when searching a database of a particular size. The value is defined as

$$E = N/n * m * n * 2^{-S}$$

where m and n are the length of the two nucleotide sequences (measured in base pairs), S is the bit score, and N refers to the total length of all sequences in the database. The formula should make intuitive sense. For example, if S is higher (*i.e.*, better matches), you would expect to see fewer "hits." On the other hand, if m or n are larger (*i.e.*, one or the other sequence is longer), then you would expect to see more hits purely by chance. Finally, if the database contains more sequences (*i.e.*, N is larger), then you would expect to see more hits. In any case, if BLAST returns an E-value that is very small or close to zero, then you probably have a meaningful match that is not due to random chance. To interpret the matches, you therefore need to pay attention to whether the E-value is reasonably small. E-value is related to the P-value by the following formula:

$$P = 1 - e^{-E}$$

So for a P-value of 0.95 (the statistically significant level), the E-value is around 3. Thus, in your search, an E-value of 3 or less would be an acceptable match.

You should also keep in mind that there are a lot of sequences in the database and that some of them are from the same species and therefore might be very similar. In some cases, the name of the organism may have changed after it was originally reported; accordingly, two or more sequences may match extremely well but appear to belong to completely different species.

Procedure (BLAST Search to identify your bacterium)

1. At this point, you need to select and copy the DNA sequence of your unknown bacterium.
2. Link to NCBI's BLAST Sequence Homology Search server (*http://www.ncbi.nlm.nih.gov/BLAST*) and select

   Nucleotide BLAST ?

   - Standard nucleotide-nucleotide BLAST [blastn]

Because we wish to look for nucleotide sequences homologous to the 16S rDNA, you will use the **blastn** program and the **nr** (non-redundant) database. You may leave all of the other default settings. Paste your DNA sequence into the box provided, then click on **BLAST** to submit your query.

3. Interpret the BLAST results and select the most likely identity of your unknown. *After the server computer conducts your analysis, the results are presented three ways: as a graphic, a table of "hits" (identified similarities), and a series of sequence alignments.*
   A. The graphic has lines showing the positions and ranges of identity/similarity between your sequence (the query) and other possible sequences in the database. The location and length of each line indicates the extent of similarity (how close the match is, also shown as the line's color as well as length).
   B. Under the text "Sequences producing significant alignments" is the table of "hits." Each database entry similar to the query sequence is presented, beginning at the top with the closest match and ending at the bottom with the weakest. Clicking on the code on the left

of each line (*e.g.*, emb|V00296|ECLACZ) links you to the GenBank entry for the sequence. Clicking on the number (the score) at the right end of the line will jump you downward within the file to the sequence alignment.

C. Each matched sequence is presented as a separate alignment with the query sequence. Only the similar/identical regions of each molecule's sequence are presented here. The numbers after the words "Query and Sbjct" indicate the position within each database entry to which the nucleotides on that line correspond. This display is where one can analyze in detail the nucleotide differences between the query and its homologue. *You do NOT need to print out your BLAST search.* You need to understand what the results mean and what is the most likely match and why. Usually the best match is the result at the top of the list with the highest score and lowest E value.

4. Find the GenBank file record for your bacterium and learn how to read a GenBank file. On the left column of your BLAST search, click on the identification number of the bacterium you think best matches your 16S rDNA sample. You should now be linked to the GenBank record of your chosen bacterium. Look at the GenBank record of your chosen bacterium. Acquaint yourself with the parts of the GenBank database record for your nucleotide sequence and be able to identify the information in your record that is bolded below. (See Appendix 1 for definitions of these terms; see Appendix 2 for a sample GenBank record.)

   What types of information are contained in the following parts of the record?

   - Locus
   - **Definition**
   - Keywords
   - **Accession and NID**
   - **Source**
   - **Organism**
   - Reference(s) **--earliest record**
   - Medline
   - Comments
   - Features
   - **CDS *Why is there no coding sequence for the 16S RNA??***
   - /translation **Why is your sequence NOT translated?**
   - /db_xref
   - mutation
   - variation
   - exon **Why do you NOT expect to find either exons or introns in your sequence?**
   - intron
   - precursor_RNA
   - mRNA
   - **Base Count**
   - Sequence

Section 3: Identifying Conserved Sequences using Multiple Sequence Alignment (MSA): ClustalW

Background

You have seen that the 16S rDNA region is sufficiently different from one bacterium to another to use the differences as a means of identification. However, how different/similar are the regions, and are some bacterial 16S rDNA sequences more similar than others? You might have predicted that genes coding for molecules with such a vital function as being part of the ribosome would not have such variability between species. Start to generate questions about what parts of the 16S rDNA have variable and conserved regions and why this might be so; look in your textbook for the structure of 16S rRNA and see if you can predict variable and conserved regions on the DNA. (H*int: look for regions that are double-stranded and single-stranded; where are the active sites, etc*.)

The Multiple Sequence Alignment (MSA) algorithm ClustalW takes a set of input sequences and aligns them so that homologous regions (the features that are common to the entire set of sequences) are highlighted. This serves to identify the nucleotides or amino acids within the sequences that have been conserved during their evolutionary divergence. Natural selection tends to select against changes that result in loss of molecular function, thus conserved residues identified in an MSA are presumed to be important for the structure and function of the molecule.

The Multiple Sequence Alignment, or MSA, homology search algorithm is sometimes called a "many-against-each-other" search because the input is a small, defined set of sequences that are compared only against each other, not against an entire database. This is in contrast to the BLAST similarity search algorithm, a "one-against-all" similarity search, in which the input is a single sequence that is compared against all other known sequences listed in the database. Thus, the starting point for an MSA is a set of sequences that are already presumed to be homologous.

Procedure

1. For this comparison you will compare the DNA from all 5 unknown bacteria. Select and copy the sequences from *"Other SAMPLE Unknown Sequences"* from the BIO250Y course web page (notice there is one option for "all unknown sequences"), or make ONE text file of the five unknown sequences in Vlab.
2. To use the CLUSTALW software for multiple sequence alignment, the DNA sequences must be in a specific format:
   >bacterium1 [Each new sequence must start with a ">" and have no spaces in the title]
   ATGCTTAAA….. [DNA sequence starts on a new line]
   >bacterium2
   CGGTAAACT
3. Access the European Bioinformatics Institute's ClustalW MSA server *(http://www2.ebi.ac.uk/clustalw/)* to conduct multiple sequence alignments.

You do NOT need to print out your alignment results. You need to know how to read the results, e.g., what portions of the sequences are identical; where are there differences?

4. Interpret the results.
   (The results of the MSA are a series of stacked lines, each line representing one of the sequences in the query set. Gaps (dashes) are introduced as necessary to maximize the alignment of

identical or similar residues among the set of sequences. Insertions and/or deletions reflect important evolutionary events. At the bottom of each stack of aligned sequences are symbols that summarize the alignment at that position in the sequence. An asterisk denotes a position at which all query sequences have the exact same amino acid. Dots indicate the degree of homology when there is not complete sequence conservation.)

5. Create a phylogenetic tree.
   - For a graphical view of the alignment, click on the gray button labeled "**JalView**." (For instruction on all of JalView's features, click on the text link "Use JalView.") A new browser window will open (don't close the old one!).
   - Colored boxes group homologous residues in the JalView window. The darker the color, the greater percentage of sequences within the set that have the same residue at that position. Notice that this view lets you quickly spot broad regions of high homology, and note individual sequences that are non-homologous at a given position.
   - Click on "Tools" from the JalView menu and click on "tree diagrams."
   - Draw or print your tree. Which bacteria are most similar/different?

Section 4: Individual presentations based on model organisms

The evolution of present-day organisms (and thus their component cells) from common ancestors has important implications for cell and molecular biology as an experimental science. Because the fundamental properties of all cells have been conserved during evolution, the basic principles learned from experiments performed with one type of cell are generally applicable to other cells. Another reason for the use of model systems is that many kinds of experiments can be more readily undertaken with one type of cell than another.

For this section, you are to research one of the organisms listed below that you signed up for in a previous lab. If you do not have a topic, contact your TA or the office. Your report should include:

- a brief description of the organism,
- why the genome of this organism would be a good one to sequence,
- the current status of sequencing its genome (complete or partially sequenced),
- who is working on the sequencing, and
- what they have found so far.

[Hint: One way to find information quickly is to use the search engine "Google" found at www.google.com and type [name_of_organism] AND genome. Be aware that google is a general search engine accessing both credible and non-credible sources on the web, so take care in selecting your information].

I actually list about 25 organisms, including representatives from all kingdoms; here is a partial list:

Escherichi coli (bacterium)
Saccharomyces cerevisiae (yeast)
Caenorhabditis elegans (nematode worm)
Drosophila melanogaster (fruit fly)
Arabidopsis thaliana (mustard plant)
Xenopus laevis (frog)
Mus musculus (mouse)

You will be expected to give a *three- minute (maximum) presentation* to the class and give out a *one-page (maximum) typed summary of your topic* to classmates and your TA. Your mark for this section will be based on both your presentation (5 marks) and your printed summary (5 marks). Include your references on your printed summary. You will be assessed on how effective your presentation is by the logical, convincing, and accurate development of evidence; appropriate use of the chalk board or overheads, and the appropriate use of resources (both what you use and how you use references).

General Resources:

The following are some excellent general sites for information about model organisms and genomics.

1. Information on both terms and links to useful sites: *http://www.genomicglossaries.com/default.asp*
2. The Institute for Genomic Research at *http://www.tigr.org*
3. See the "Genome Gateway" from the home pages of the scientific journals Nature or Science.
4. The NCBI have information on model genomes at *http://www.ncbi.nlm.nih.gov/Genomes/*
5. Cold Spring Harbor Laboratories at *http://www.cshl.org/*
6. See the various pages from the Weizman Institute, especially Searching Molecular Biology Databases, Frequently Asked Questions at *http://bioinformatics.weizmann.ac.il/mb/faq/*
7. The Munich Information Center for Protein Sequences (MIPS) at *http://mips.gsf.de*

Alternative Section 4: Bioinformatics and The Human Genome Project Presentations

For this section, you are to research one of the topics listed below related to bioinformatics, genomics, or the human genome project and prepare a short presentation. You were asked to sign up for a topic in a previous lab to ensure that a variety of topics will be covered. If you do not have a topic, contact your TA or the office. URLs are given with each topic. You may use others, but be sure of the credibility and reliability of your source.

You will be expected to give a three-minute (maximum) presentation to the class and give out a one-page (maximum) typed summary of your topic to classmates and your TA. Your mark for this section will be based on both your presentation (5 marks) and your printed summary (5 marks). Include your references on your printed summary.

Evaluation: you will be assessed on how effective your presentation is by the logical, convincing, and accurate development of evidence; appropriate use of the chalk board or overheads, and the appropriate use of resources (both what you use and how you use references).

For the presentations: Use may use the chalkboard or acetate overhead sheets to summarize your findings. You should prepare these summaries and any graphic material before the lab.

For general references, use these special volumes:
The Human Genome. (15 February) 2001 Nature 409:745-964 [Celera]

The Human Genome (16 February) 2001 Science 291 [public consortium]

You will select a topic within one of these five areas

1. Bioinformatics -

a) Current and future prospects in the public (academic & government) sector versus private (commercial) sector:
See US government website address *http://www.ornl.gov/hgmis*
See Celera website : *http://www.celera.com/*
See Trends in Commercial Bioinformatics. A report issued March 13, 2000 by Jason Reed of Oscar Gruss & Son. Available at *http://www.oscargruss.com/reports.htm*

b) Bioinformatic tools: Explain and demonstrate the application of some of the other tools, *e.g.*, one of the molecular graphic tools below. Download atomic coordinate files from online structure databases, and study 3D representations of macromolecules with molecular graphics software to see the relationships between the structure and function of a protein, *i.e.*, how the sequence of subunits and their folding in three-dimensional space determine its proper function.

- Molecular Simulations Inc.'s WebLab ViewerLite molecular graphics software *(www.msi.com/solutions/products/weblab/viewer/register/lite/download_lite.htm*l)
- MDL Information Systems' Chemscape Chime molecular graphics browser plug-in *(www.mdli.com/support/chime/chimefree.ht*m)
- Research Collaboratory for Structural Bioinformatics' (RCSB) Protein Data Bank *(www.rcsb.org/pdb/*)
- National Center for Biotechnology Information's (NCBI) Molecular Modeling Database *(www.ncbi.nlm.nih.gov/Structure/*)
- Tools listed in Reed, J. 2000 (March). Tends in Commercial Bioinformatics *(http://www.oscargruss.com/reports.htm)*

c) Additional ideas: See the various pages from the Weizman Institute, especially Searching Molecular Biology Databases, Frequently Asked Questions. *(http://bioinformatics.weizmann.ac.il/mb/faq/)* and look at GeneCards (human genes, maps, proteins, and diseases (*http://bioinformatics.weizmann.ac.il/cards/)*

d) Interview an expert, *e.g.*, summarize the interview by Scientific American with Stuart Kauffman (of the Bios Group and Cistem Molecular in Santa Fe), considered to be among the pioneering pre-eminent scientists in bioinformatics. The interview is available online at the Scientific American web site *(http://www.sciam.com/)*. Go to Interviews and then look for Kauffman.

2. History, time-course, and significance of the Human Genome Project

a) Initial source: *http://www.ornl.gov/hgmis/faq/faqs1.html*

- Lots of other information on the site at the basic address *http://www.ornl.gov/hgmis*

(The Human Genome Program of the U.S. Department of Energy (DOE) funds this web site.)

- You might want to read "Evolution of a Vision: Genome Project Origins, Present and Future Challenges, and Far-Reaching Benefits" at *http://www.ornl.gov/hgmis/publicat/hgn/v7n3/02smithr.html*

b) Also see the web page for the **Human Genome Organization (HUGO)**. Dr. Tsui, geneticist-in-chief at the Toronto Hospital for Sick Children is the current president. *(http://www.gene.ucl.ac.uk/hugo/)*

c) Also, look at the Celera Genomics page to see the private company's description of the project: *(http://www.celera.com/)*

3. Model organisms and the relevance of their genomics

The NCBI have information on model genomes at *http://www.ncbi.nlm.nih.gov/Genomes/*
Some model organisms with resources are listed below:

a) Arapidopsis (mustard plant): *http://www.arabidopsis.org/*

b) *Drosphila melanogaster* (fruit fly)
FlyBase-the Drosophila sequence database at *http://flybase.bio.indiana.edu/*

c) *Mus musculus* (mouse)

- Start with "Leaping Across Genomes Comparing Mouse and Human DNA" *(http://www.ornl.gov/hgmis/publicat/hgn/v7n6/08mice.html)*

- Then look at the mouse genome page: *http://www.informatics.jax.org/*

d) Pufferfish in addition to the NCBI page try http://www.celera.com/
Go to Celerascience and news and scroll down to June of 2000.

4. Ethical, Social and legal

Template for Ethical Decision Making

- What is the ethical question?
- Who are the stakeholders? What are the relevant principles or values?
- What are the possible solutions? How are stakeholders affected? Are the ethical principles supported?
- What is the best solution and how is it justified?

http://www.ornl.gov/hgmis/resource/elsi.html

- Also see the Human Genome Organization : *http://www.gene.ucl.ac.uk/hugo/*
- Also the legal perspective covered in "Introducing the Human Genome Project: Its Relevance, Triumphs , and Challenges" *(http://www.ornl.gov/TechResources/Human_Genome/publicat/judges/drell.html)*

5. Potential Application of genomics & bioinformatics

Many possible sources. You might want to start with:

http://www.ornl.gov/hgmis/project/benefits.html

a. Molecular medicine. *Pick one of the following applications* and find and explain examples of how genomic/bioinformatics has been used.
   - medical diagnostics
   - commercial drug discovery
   - pharmacogenomics
   - gene therapy [See 2001 Student web pages from JLM349S 2000 at *http://dragon.zoo.utoronto.ca/jlm349/*]

b. Industrial biotechnology, for example
   - biofuels
   - environmental monitoring of pollutants

c. Agbiotech, for example
   - Genetically modified food resistant to disease, pests or environmental factors [See 2000 Student web pages from JLM349S 2000 at *http://dragon.zoo.utoronto.ca/~jlm349/*]
   - Environmental cleanup

d. DNA forensics: Forensic DNA analysis: technology and application
   (http://www.parl.gc.ca/information/library/PRBpubs/bp443-e.htm)
   (http://www.ornl.gov/hgmis/eductation/student.html)

Table 1. Biocomputing tools (Hershberger 2000)

Search tool	internet location	function
[word or topic search] **Entrez search** Finding Sequences by name in Genbank e.g., ask for 16S rDNA	National Center for Biotechnology Information's (NCBI) GenBank database www.ncbi.nlm.nih.gov/Entrez	Access DNA and protein sequence data and read a sequence database record to obtain specific information on mutations, gene structure, functional sites within proteins, etc.
[sequence search, start with nucleotide or amino acid sequence**]** **BLAST Homology Searches**	National Center for Biotechnology Information's (NCBI) BLAST server www.ncbi.nlm.nih.gov/BLAST	Identify DNA and protein sequences within the GenBank sequence database that share regions of similarity with your sequence of interest. Allows you to investigate the concept of **sequence homology**, or **similarity** within the sequences of different genes or proteins. "one-against-all"
[sequence comparison] **Multiple Sequence Alignment**	European Bioinformatics Institute's (EBI) ClustalW MSA server www2.ebi.ac.uk/clustalw/	Identify conserved sequences and consensus sequences among genes and proteins. Allows you to investigate the concepts of **sequence conservation** and **consensus sequences**, or the patterns of similarity among sequences of genes or proteins that share common three-dimensional structures or functions. "many-against-eachother"
Molecular Graphics **(not required in this laboratory)**	Research Collaboratory for Structural Bioinformatics' (RCSB) Protein Data Bank www.rcsb.org/pdb/ National Center for Biotechnology Information's (NCBI) Molecular Modeling Database www.ncbi.nlm.nih.gov/Structure/	Download and study 3D representations of macromolecules with molecular graphics software to see the relationships between the **structure and function** of a protein i.e., how the sequence of subunits and their folding in three-dimensional space determine function.

Bioinformatics Lab Evaluation

Please complete the top part of this form and give to your TA before your presentation.

Student #:_____ Last name:_____

First name: _____ Topic:_____

Presentation: ____/5 (average of 4 components below)

Criteria	**0-2 = below expected** **3-4 = expected** **5 = exceptional**	**Comments** *(include what is most notable and ideas for improvement)*
Logical and convincing development of evidence; good sources used effectively		
Coherent, integrated, insightful synthesis		
Appropriate level of detail; selective detail (listens to other talks & avoids repeating material)		
Effective presentation including use of overheads/chalkboard		

Printed Summary: ____ / 5 (average of 5 components below)
Appropriate (credible, reliable) source material **___**
Logical and convincing development of evidence___
Coherent, integrated, insightful synthesis ___
Highlighted key points of presentation with the appropriate level of detail____
References sited correctly

Comments:

(Student Outline) Appendix A: GenBank record file-- definitions

http://ncbi.nlm.nih.gov/genbank/gbrel.txt

The following is a brief description of each entry field. Detailed information about each field may be found in Sections 3.4.4 to 3.4.14**.**

LOCUS - A short mnemonic name for the entry, chosen to suggest the sequence's definition.
DEFINITION - A concise description of the sequence.
ACCESSION - The primary accession number is a unique, unchanging code assigned to each entry. (Please use this code when citing information from GenBank.)
NID - The unique nucleic acid identifier that has been assigned to the current version of the sequence data that are associated with the GenBank entry identified by a given primary accession number.
KEYWORDS - Short phrases describing gene products and other information about an entry.
SEGMENT - Information on the order in which this entry appears in a series of discontinuous sequences from the same molecule.
SOURCE - Common name of the organism or the name most frequently used in the literature.
ORGANISM - Formal scientific name of the organism (first line) and taxonomic classification levels (second and subsequent lines).
REFERENCE - Citations for all articles containing data reported in this entry. Includes four subkeywords and may repeat.

AUTHORS - Lists the authors of the citation.
TITLE - Full title of citation. Optional sub keyword (present)
in all but unpublished citations)/one or more records.
JOURNAL - Lists the journal name, volume, year, and page numbers of the citation.

MEDLINE - Provides the Medline unique identifier for a citation.
REMARK - Specifies the relevance of a citation to an entry.
COMMENT - Cross-references to other sequence entries, comparisons to other collections, notes of changes in LOCUS names, and other remarks.
FEATURES - Table containing information on portions of the sequence that code for proteins and RNA molecules and information on experimentally determined sites of biological significance.
BASE COUNT - Summary of the number of occurrences of each base code in the sequence.
ORIGIN - Specification of how the first base of the reported sequence
is operationally located within the genome. Where possible, this includes its location within a larger genetic map
- *The ORIGIN line is followed by sequence data (multiple records).*

Feature Key Names

The first column of the feature descriptor line contains the feature key. It starts at column 6 and can continue to column 20. The list of valid feature keys is shown below.

allele	Related strain contains alternative gene form
attenuator	Sequence related to transcription termination
CDS	Sequence coding for amino acids in protein (includes stop codon)
enhancer	Cis-acting enhancer of promoter function
exon	Region that codes for part of spliced mRNA (in eukaryotes)
iDNA	Intervening DNA eliminated by recombination
intron	Transcribed region excised by mRNA splicing (in eukaryotes)
LTR	Long terminal repeat
mat_peptide	Mature peptide coding region (does not include stop codon)
misc_binding	Miscellaneous binding site
misc_difference	Miscellaneous difference feature
misc_recomb	Miscellaneous recombination feature
misc_RNA	Miscellaneous transcript feature not defined by other RNA keys
misc_signal	Miscellaneous signal
misc_structure	Miscellaneous DNA or RNA structure
modified_base	The indicated base is a modified nucleotide
mRNA	Messenger RNA

mutation	A mutation alters the sequence here
old_sequence	Presented sequence revises a previous version
precursor_RNA	Any RNA species that is not yet the mature RNA product
primer	Primer binding region used with PCR
promoter	A region involved in transcription initiation
protein_bind	Non-covalent protein binding site on DNA or RNA
RBS	Ribosome binding site
rep_origin	Replication origin for duplex DNA
repeat_region	Sequence containing repeated subsequences
repeat_unit	One repeated unit of a repeat_region
rRNA	Ribosomal RNA
STS	Sequence Tagged Site; operationally unique sequence that identifies the combination of primer spans used in a PCR assay
tRNA	Transfer RNA
unsure	Authors are unsure about the sequence in this region
variation	A related population contains stable mutation
-10_signal	`Pribnow box' in prokaryotic promoters
-35_signal	`-35 box' in prokaryotic promoters
3'UTR	3' untranslated region (trailer)
5'UTR	5' untranslated region (leader)

(Student Outline) Appendix B: Study Guide

This outline of questions and key concepts is provided as a study aid for your preparation for this lab on Bioinformatics .

1. Know basic terminology about bioinformatics as described in lab 6

- What is Bioinformatics?
- Recognize and be able to distinguish the names of places (and their location), repositories of data, and search and analysis tools
  e.g., National Centre for Biotechnology Information (NCBI) in Washington D.C;
  databanks include GENBANK (at NCBI), EMBL (Europe) DDBJ (Japan); search tools include Entrez, Blast, MSA

2. What basic information about the application of search and analysis tools used in this lab:

- **ENTREZ**
  1. What is Entrez and for what type of information would you use ENTREZ?
  2. Where is it and how do you access Entrez? Although Entrez is not directly used for this lab, it is a very useful tool to know about.
- **BLAST** (what does this stand for?)
  What is BLAST, what type of data do you input into this tool and what kind of information does the tool provide?
  1. When would you use BLAST (i.e., for what kind of scientific questions would BLAST be useful)
  2. Why is BLAST referred to as a "one against all similarity search"
  3. What do Score and E-value represent?
  4. Given a sample BLAST search result be able to locate basic types of information such as the Accession Number, and type of gene, sequence alignment as well as identify and explain your reasoning for the best match..
  5. Further thought question: What features of a sequence influence the minimum size (i.e., number of bases) to give a clear and unambiguous result in a BLAST search?
  6. Given some information about unknown/unidentified sequenced material, and BLAST search results identify the most likely match and explain why

- **Multiple Sequence Alignment (MSA) using ClustalW**
  1. What is MSA, what type of data do you input into the ClustalW tool and what kind of information does the tool provide?
  2. In what form must your data be to input into MSA?

3. When would you use MSA (i.e., for what kind of scientific questions would BLAST be useful)
4. Why is MSA referred to as "many against each other similarity search"?
5. Given a sample output identify homologous regions and variable regions. What are conservative substitutions and identify some on the output. What is the difference between conservative and non-conservative substitutions?
6. What general comments can you make about the degree of similarity between your two (or more) bacteria samples?
7. Suggest reasons why some regions of the 16S rDNA are highly conserved while others are not.

- Given a sample **Genbank record** be able to find the following information (see sample in the Appendix):
  1. What is the specific sequence? ("Definnition")
  2. Know what the permanent identifying code is for this sequence (accession number) and distinguish this from the unique nucleic acid identifier used for the current version of the sequence data. (NID).
  3. From what organism was this sequence isolated? (Source)
  4. What was the first reference describing this sequence ?
  5. For this first reference locate when, by whom, and if published .
  6. What are the total number of A,C,G,T nucleotides (Base Count) for this sequence and look for anomalies–e.g., are the number of each base similar or is there a predominance of G-C or A-T?
  7. What information can you get out of the Features section- e.g., CDS (coding sequence) indicates transcribed into mRNA for translation into amino acids. Why is there no CDS indicated for 16S rDNA?

3. What you need to look for in the Virtual Bactieria ID Lab

Know some basic information about the isolation, purification , and identification of bacterial DNA described in this exercise; the answers to all of these questions are given in the background text information in the vlab:

1. Why use molecular techniques to identify the unknown bacterium rather than standard microbiological tests?
2. Why is 16S rDNA used in this exercise to identify the unknown bacterium? Realizing that 16S rRNA has loops and folds gives you a clue to why such an important molecule has both conserved and variable regions. Look up the structure of 16S rRNA in your textbook
3. Isolate a bacterial sample from the patient [only in the simplest terms].
4. Isolate whole bacterial DNA.*[How was this done in the vlab?]*
5. Make many copies of the desired piece of DNA.
   - *What technique do you use?*
   - *How do you separate the desired DNA sequence from the rest of the DNA?*
   - *How do you purify your sample?*
6. Sequence the DNA. *Briefly describe how this is done*

Literature Cited

Ball, M., G. Duncan, D. Ranieri, and S. Kiser. 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.

Gurney, T., R. Ethel, D. Ratnapradipa, and R. Bossard. 2000. Introduction to the molecular phylogeny of insects. Pages 63-77, *in* Tested Studies for Laboratory Teaching, Volume 21 (S.J. Karcher, Editor). Proceedings of the 21st Workshop/Conference of the Association for Biology Laboratory Education (ABLE), 509 pages.

Gurney, T., R. LeMon, and K. Nolan. 2001. DNA sequencing to illustrate mutation and evolution. Pages 100-119, *in* Tested studies for Laboratory Teaching, Volume 22 (S.J. Karcher, editor). Proceedings of the 22nd Workshop/Conference of the Association for Biology Laboratory Education (ABLE), 489 pages.

Hershberger, R.P. 2000. What I could teach Darwin using "Darwin 2000," an interactive web site for student research into the evolution of genes and proteins. Pages 1-32 *in* Tested Studies for Laboratory Teaching, Volume 21 (S.J. Karcher, Editor). Proceedings of the 21st Workshop/Conference of the Association for Biology Laboratory Education (ABLE), 509 pages.

Howard, K. 2000. The bioinformatics gold rush. Scientific American July, 2000. Accessed online *(http://www.sciam.com)* July, 2000.

Lim, H. 2000. Bioinformatics in the pre- and post-genomic eras. Trends in Biotechnology (April) 18: 133-135.

Reed, J. 2000. Trends in commercial bioinformatics (March). Accessed online *(http://www.oscargruss.com/reports.htm)* July, 2000.