Exercise 17
AUTOMATED NUCLEOTIDE SEQUENCING AND ELECTROPHEROGRAM EVALUATION

Introduction

DNA sequencing, the process of determining the sequence or arrangement of nucleotides (bases) in a sample of DNA molecules, can play a significant roll in the identification of bacterial isolates, and in most cases involves the PCR. Commercial sequencing facilities currently use highly sophisticated, automated systems that employ dideoxynucleotides and the chain termination method (Sanger sequencing) for nucleotide sequencing. A **dideoxynucleoside triphosphate** (**ddNTP**) is an analog of a dNTP that lacks a hydroxyl group on the third carbon of its sugar (the 3' end). When incorporated into DNA strands during replication, ddNTPs are unable to form phosphodiester bonds with incoming nucleotides, and thus efficiently terminate DNA synthesis. By adding small amounts of ddNTPs to a PCR reaction-mixture containing template DNA, dNTPs and a single primer, it is possible to generate populations of nucleotide strands with different ones terminating at every position in the template strand. In Big Dye sequencing systems, the four different types of ddNTPs each carry a different colored fluorescent label, and the nucleotide strands generated are subjected to electrophoresis within capillary tubes. Lasers are used to excite the fluorescent labels, and a camera captures the color patterns generated. Data collection software and computer analysis is then used to generate a visual record of the DNA sequence data called an **electropherogram**,

The electropherograms generated by automated sequencers are four-color chromatograms displaying sequencing results as a series of peaks. Each peak represents an accumulation of oligonucleotides ending with a specific base as designated by color, and peak heights (intensity of signal) indicate the relative number of oligonucleotides present in each size category. In addition to the data represented by colored peaks, sequencing machines generate text files showing their interpretation of this data (typically 500bp with 98% accuracy). The machines cannot verify the validity of the text files generated, so human interpretation and editing is necessary.

Nucleotide sequence data is initially edited by sequencing facility personnel, and later by persons intending to use the data. Automated sequencing methods typically generate considerably more short oligonucleotides than long, so the peak heights near the beginning of a sequence would be much greater than those toward the end, if not modified (the sequence would be top heavy). One job of sequencing facility personnel is to apply an analysis program that converts the raw data or "true profile" into a "flat profile" so the four-colored peaks appear to be more or less equal in height along the entire length of the sequence. They also edit the data for miscalled bases and other anomalies recognized as machine error. The sequence data sent to users typically meets specifications as indicated above e.g., 500bp of sequence with 98% accuracy, but this means that as much as 2% of the data may be wrong. It is the users responsibility to evaluate the data and make changes as necessary to improve accuracy.

For this exercise we will use computer technology to access, evaluate and edit nucleotide sequence data obtained from bacterial isolates. The bacteria represented are those assigned as Physiological Unknown #2 (PUNK2), and the DNA sequences are portions of 16S ribosomal DNA amplified with the PCR. Since the sequence data in each file is maximally accurate for only about 500 bases, and the 16S ribosomal-DNA gene is more than 1500 bases in length, three different primers were used to generate sequence data for each organism type. In most cases the primers used were Bacteria 8-Forward, 533-Forward and 1530-Reverse, but in some instances the reverse primer used was 149Power -Reverse. The data sets appear just as they were received from the **College of Biological Sciences sequencing facility at UC Davis**, and will require some editing.

Materials

Computer (Power Book or equivalent) equipped with Macintosh OSX and viewing software package 4Peaks and Electronic versions of nucleotide sequences.
Note – Our sequences were run on an ABI 3730 Capillary Electrophoresis Genetic Analyzer with ABI BigDye Terminator v3.1 sequencing chemistry.

Procedure

1. Obtain a lap top computer from the cart provided, access the airport and go to the Bio. Sci. 4 (Microbiology) web site as previously instructed.

2. Click on **Web-Based Laboratory Assignments**, and then **PUNK2 Electropherograms (EPGs)**. Verify that the semester and year are correct.

3. Select the grid number corresponding to the number of your physiological unknown to access the correct data file, and click the number only once. A folder will appear on the desktop “behind” the Web document, and will not be immediately visible.

4. Minimize the Web document (click on the yellow circle at the upper left corner). Then click on the file folder containing your electropherogram files. Each folder should contain three .ab1 files.

5. Click on each file (one at a time), and the 4Peaks program will automatically open displaying the electropherograms and associated text files. Position these files as instructed.

6. Observe the contents of your data files as instructed below, but do not edit (make changes to) the nucleotide sequences contained within them.

    a. Move the colored bubble about ¼ the distance along the scroll bar and observe the electropherogram in this region (210-260bp). You should see evenly spaced peaks, each with just one color. There will be some variation in peak height (a normal phenomenon), but sequence peaks can readily be distinguished from any background "noise" present (i.e., small peaks near the baseline).
    b. Notice that each electropherogram peak is accompanied by a colored letter (A, G, C or T), at the top of the viewing window, and that peak colors and letter colors match. Nucleotide strands terminating with adenine are represented by green peaks and letters; those terminating with guanine are represented by black; those terminating with cytosine by blue; and those terminating with thymine, by red.
    c. Move the colored bubble about ¾ the length of the scroll bar and observe the peaks in this region (600-700bp). You are likely to find broad, ragged-looking peaks separated by "valleys" that do not contact the baseline. The sequence data in this region is likely to contain errors, and as you continue moving to the right, they will increase in number.
    d. Return to the beginning of your sequence, and you will notice that although there is a nucleotide sequence indicated by colored letters, the peak data may be missing, very small, or overlapping. This is inaccurate sequence data and some of it will probably be edited out.
    e. In some cases, additional letters may be included in the sequence and are colored white against a red background. These are IUB-codes indicating more than one base can be found at the position indicated. These can be interpreted using the information below.

Oligonucleotide IUB-Codes (Codes of the International Union of Biochemistry)

Single nucleotide bases, IUB Coding for **A** = Adenine, **C** = Cytosine, **G** = Guanine, **T** = Thymine, **U** = Uridine, and **I** = Inosine.

For mixed (wobble) bases (either of two bases possible) IUB Coding for **M** = A and C, **R** = A and G, **W** = A and T, **S** = G and C, **Y** = C and T, and **K** = G and T.
For mixed (wobble) bases (any of three bases possible) IUB Coding for **V** = A, G and C, **H** = A, C and T, **D** = A, G and T, and **B** = G, T and C.
The letter N indicates that any one of four bases is possible.

7. Open the Microsoft Word program and select a new blank document. Label this document Nucleotide Sequence for PUNK2 # (enter your number as selected).

8. Position your Word document and the 4Peaks viewing window so that both are visible on your screen. Placing the 4Peaks window just above the word file will usually work.

9. Click on the Bacteria 8-Forward file, copy the nucleotide sequence and paste it into your Word file. To accomplish this, click on 4Peaks, move your cursor to "Edit" (upper right), click and hold while you scroll down to "Copy Sequence" and then release. Then click on your word file (top of document), Click and hold on "Edit" (upper right), scroll down to "Paste", and release. Your nucleotide sequence now appears as a text file in your Word document. Label this file.

10. Copy and paste the nucleotide sequence from the 533-Forward, 4Peaks file into your word file below the Bacteria 8-Forward sequence. Be sure to leave a space between the two and to label the 533-Forward sequence.

11. Click on the 1530-Reverse file (Note that in some folders this may be replaced with a 1492-Reverse file, but you will deal with it in the same manner.) You must "**flip**" or "**reverse**" this sequence before you copy and paste it into your Word file. The "flip" or "reverse" option will appear when you scroll downward under "Edit". The electropherogram will then be labeled as a reverse complement sequence.

12. Evaluate the data presented in your electropherograms (as indicated above), and then edit the beginning (left edge of sequence) of your Bacteria 8-Forward file and the end (right edge of sequence) of your 1530-Reverse (or 1492-Reverse) file to eliminate sequencer errors. Remember that you want to keep as much of the sequence as possible.

    Note – It is not necessary to edit the ends of the 533-Forward sequence, as this will overlap both of the other two sequences.

13. Remove overlap from Bacteria 8-Forward and 533-Forward sequences as follows:
    a. Count down about 500 bases in your Bacteria 8-Forward sequence (about 10 lines of text) and look for the sequence "GGGCGTAAA". Insert two or more "returns" to break the sequence just beyond the last "A".
    b. Look for this same sequence near the beginning of the 533-Forward sequence (it will often appear as "GGGCGTAA", i.e., one "A" is missing). Insert two or more "returns" to break the sequence just beyond the last "A".
    c. Check to see if the sequences beyond your "cut" sites are the same for both the Bacteria 8-Forward and 533-Forward sequences. If they are, delete the overlapping sequence.

14. Remove overlap from 1530-Reverse and 533-Forward sequences as follows:
    a. Count up about 500 bases in your 1530-Reverse sequence (about 10 lines of text) and look for the sequence "GGGGGCCC" (sometimes GGGGACCC). Insert two or more "returns" to break the sequence just after the last "C".
    b. Look for this same sequence near the middle of the 533-Forward sequence. Insert two or more "returns" to break the sequence just beyond the last "C".
    c. Check to see if the sequences beyond your "cut" sites are the same for both the 1530-Reverse and 533-Forward sequences. If they are, delete the overlapping sequence.

15. Join your three sequences into one **contiguous sequence** by eliminating any spaces existing between them. Your goal is to obtain an error-free nucleotide sequence of maximum length.

16. Use the "character count" option under "Tools" in the menu bar to determine the length of your sequence. In most cases it will be around 1500 bases in length. You will use this file in the next exercise.

Questions

1. What is automated nucleotide sequencing?

2. What are dideoxynucleoside triphosphates (ddNTPs) and how do they affect DNA replication if they are incorporated into growing nucleotide strands?

3. What is an electropherogram?

4. Why is it necessary to edit nucleotide sequence data that has been obtained from an automated sequencer?

5. What are IUB Codes (other than A,G,T and C), and what do they represent when present within a nucleotide sequence?

6. What portions of the data presented in an electropherogram are most likely to contain errors and therefore to require careful evaluation?

7. Why was it necessary to "flip" or "reverse" the 1530-Reverse (or 1492-Reverse) sequence before copying it to the word file? What would occur if you failed to do this?

Exercise 18
GENOMICS, PROTEOMICS & BIOINFORMATICS
Web-Based Activities & Assignments

Introduction

The term **genome** may be defined as the complete gene compliment of an organism that is contained within multiple linear chromosomes (eukaryotic cells), one or more circular chromosomes (prokaryotic cells) or one or more DNA or RNA molecules (viruses or viroids). **Genomics** is the study of multiple genomes as a functional unit. In practice, genomics is the study of very large numbers of genes (genomes from multiple organisms) undertaken simultaneously. Unlike genetics, which deals primarily with individual genes or gene clusters within a single population of organisms, genomics attempts to include all genes as a dynamic system and looks at changes in these genes over time. Genomics also attempts to determine how genes interact and how they influence biological processes e.g., metabolism, physiology, behavior, etc. in all living organisms.

The term **proteome** (coined in 1995) refers to the entire protein complement of an organism type, a tissue or a cell type. The proteome of an organism is dependent upon the genome, but unlike the genome, can change considerably over time. Although the genome of an organism can be changed by mutation or genetic exchange, it is generally fairly stable over time. The proteome is not. The type and number of proteins expressed by an organism can and typically does change dramatically in response to external and internal environmental stimuli. **Proteomics** is the study of all the proteins encoded by genomes, and deals with gene expression at the protein level, or the link between genomes and proteins, i.e., factors that influence gene expression and protein formation. It also involves the analysis of proteins and protein complexes, their structural and functional properties and their interactions and modifications within living organisms over time.

Genomics and proteomics are relatively new areas of study within the biological sciences. Although awareness of inheritance dates from ancient times, extensive investigations into factors influencing heredity were not reported until the mid 1800s (1856-1863 by **Gregor Mendle**). The significance of nucleic acids as the hereditary material of living organisms was not recognized until 1944. Prior to that, proteins were considered the most likely candidates. The structure of the DNA double helix was identified in 1953, and the **genetic code** was deciphered during the early 1960s. These findings provided the foundation for genomics and proteomics, as the basic relationships between DNA, RNA and protein became clear. Methods and materials for the manipulation of DNA in vitro (**recombinant DNA techniques**) became available during the 1970s, and greatly increased the pace of discovery. Techniques for determining the nucleotide sequences of nucleic acids (**nucleotide sequencing**) were developed in 1977, and the ability to reproduce these sequences in vitro using **polymerase chain reaction** technology (PCR) was developed in 1985-86. The complete nucleotide sequence for an entire genome (human mitochondria) was published in 1981, making structural genomics a reality. The **Human Genome Project**, proposed in 1986, was begun in 1990, along with studies of other organisms. The first complete genome for an entire organism, *Haemophilus influenzae*, was elucidated in 1995. In the early part of 2003, as biologists celebrated the 50th anniversary of the discovery of DNA, complete genomes for more than 1000 viruses and over 100 microbes had been recorded along with those from numerous multicellular organisms. Since then, the number of whole genomes recorded for prokaryotic organisms has increased dramatically.

The next logical step for biologists was functional genomics, i.e., determining the functions of all genes identified, but this is no easy task. Variations in genetic code preference, and the occurrence of post-transcriptional and post-translational modification events complicate the process. Functional genomics deals with **phenotype**, the observed characteristics of organisms, as influenced by the expression of genes, but the links between genes and proteins with similar function are often difficult to identify. The study of gene expression in time and space, and the structure, function and interactions of all proteins (proteomics) is an enormous undertaking. Not surprisingly, the quantity of data generated by genomics and proteomics is immense, and not easily interpreted. Storage of this data required the establishment of numerous interactive databases, and the development of efficient methods for viewing and analyzing the data accumulated. As systems for data acquisition, management, search and interpretation have developed; a new discipline called **bioinformatics** has emerged. Bioinformatics requires the use of computers and computational skills with an emphasis on statistical and machine learning techniques. Researchers develop and implement algorithms to facilitate the understanding of biological processes. Through the use of bioinformatics, researchers can compare genes and proteins to determine the evolutionary relationships of organisms, they can predict the structure and functions of proteins and create models for dynamic physiological systems. Bioinformatics can be applied to the development of new food crops, new vaccines, new drug treatments and new ways to diagnose disease. It will undoubtedly impact the progress and direction of biological investigations for many years to come.

This multi-part exercise is designed to provide students with an introduction to genomics, proteomics and bioinformatics by applying these methods to the identification and classification of unknown bacteria. Exercises, “A” and “B”, will be completed in association with different laboratory activities assigned at different times during the semester.

Materials/Information Sources

1. Computer with processing speed of 700MHz
2. Interactive databases including but not limited to:
    a. **NCBI** = National Center for Biotechnology Information (USA)
    b. **KEGG** = Kyoto Encyclopedia of Genes and Genomes (Japan)
    c. **TIGR** = The Institute for Genome Research (USA)
    d. **EMBL** = European Molecular Biology Laboratory (Europe)
    e. ExPASy = Expert Protein Analysis System (Switzerland)
3. Demonsrtation/Instructional Materials
    a. **4Peaks** = Software from Netherlands Cancer Institute – student project

Section A – Application of PCR & BLAST in the Identification of Physiological Unknown #2.

Each student was assigned a pure bacterial culture for Physiological Unknown #2 (PUNK2), was given a set of primers that could be used to amplify 16S ribosomal DNA from these bacteria, and was instructed in the application of the polymerase chain reaction (PCR). Following amplification, 16S ribosomal DNA from each isolate was concentrated, purified and taken to the College of Biological Sciences DNA Sequencing Facility at UC Davis. Sequence data from each unknown was returned electronically, and was posted on the Microbiology web site. Students were then assigned the task of evaluating the electronic data (electropherograms and text files) and editing as necessary to obtain a complete and accurate nucleotide sequence for each unknown. During this exercise, students will access stored copies of their nucleotide sequence text files and will use these to identify their unknown cultures using the BLAST algorithm available through NCBI.

NCBI = The National Center for Biotechnology Information. The National Center for Biotechnology Information is a massive database accessible to researchers throughout the world. NCBI advances science and health by providing access to biomedical and genomic information.

BLAST = **Basic Local Allignment Search Tool**. This is a tool used to compare nucleotide or amino acid sequences against others in public databases. For additional information about BLAST visit the NCBI Home Page and click on Tools, Data Mining.

PubMed Central = **An archive of life science journals**. Many of the entries are presented as free full-text articles (over 80 thousand in 100 journals). This is an outstanding source for up-to-date scientific information.

Procedure

1. Obtain a Power Book and connect to the **Biology Airport**. Airport symbol must be black.

2. Click on **Safari** (compass icon at bottom of screen) or Firefox (fox icon at the bottom of the screen) to access the browser. The browser window should open on the Sierra College Biological Sciences Webpage (http://biosci.sierracollege.edu).

3. Access your nucleotide sequence for Physiological Unknown #2 (as edited and stored during an earlier exercise) and copy it to the clipboard. To accomplish this, highlight your sequence, click and hold on "Edit", scroll down to "Copy" and then release.

4. Access the NCBI link (visible on the PUNK2 EPGS page under References on the lower left). On the NCBI homepage click on BLAST (under popular resources at the upper right). Under Basic BLAST, choose **nucleotide blast** (left side of page). Paste your nucleotide sequence into the BLAST search window. This can be accomplished as follows. Put your cursor at the upper right corner of the search window and click there, then click and hold on Edit (top right of screen), scroll down and release on **paste**. This will paste your nucleotide sequence into the search window (horizontal box near top of page).

5. In the **"Choose Search Set"** section, click on "**Others (nr-etc.)**" and then use the pull-down menu to select **"16S ribosomal RNA sequences (Bacteria and Archaea)"**. In the **"Program Selection"** section, click "**Highly similar sequences (Megablast)**" and then click on the **BLAST** button at the bottom of the window. Wait for your search results.

6. Scroll down the BLAST results page to see the results of your search. Observe the **Color Key for Alignment Scores** and notice that as you move your mouse over the colored bars, information about the alignment hits appears in the window just above the color key. Red bars indicate high degree of alignment, and identification information appears in the window. Scroll down and look under "**sequences producing significant alignments"** to see additional information about the identity of the unknown culture.

7. Record the name of the organism type showing the highest Score (usually the first one listed) and record the name (both genus name and specific epithet).

8. Click on the name or scroll down to the "**Alignments**" section to observe the nucleotide sequence for this entry (rows of bases lined up side-by-side), and then record the information indicated below.

    a. The length of the gene bank sequence your query sequence is aligned with (recorder just to the right of the accession number).
    b. The number of bases in your sequence that are identical to the gene bank sequence. This is listed under "**Identities**", and is a ratio (record both numbers).

c. The percent similarity between your sequence and the gene bank sequence.
d. The location and identity of bases that do not match (if there are any). Unmatched bases are indicated by white spaces (rather than lines) between bases. Their location can be determined by counting from the base number recorded either to the right or left of each alignment row (your sequence is the **Query**).

9. Scroll to the top of the sequence selected and then click on the accession number (within the blue-colored area under the technical name). This will open a new window containing specific information about the entry.

10. Record the information presented by answering the following questions:
    a. What is the **accession number** for this page of information? The accession number is the NCBI Reference Sequence (e.g., NR_028714.1).
    b. What is the **lineage** for the organisms listed? The taxonomic lineage is listed just below the genus and species names, which appear as colored print.
    c. What is the name of the first Author listed in association with this sequence, and what is the title of the reference article?

11. If there is a PubMed access, click on the colored number to the right of PubMed to access the abstract for this article. Explain briefly what the article is about. If there is no PubMed publication listed for the first entry selected, record the name and location of the facility responsible for the entry.

Section B. – The Application of PCR and BLAST in Association with Semester Projects.

Students wishing to apply nucleic acid analysis to the identification of unknown microorganisms cultured in this laboratory may use the PCR, nucleotide sequencing, electropherogram evaluation and the procedures outlined in exercise A above in conjunction with their semester projects. Once a pure culture of the organisms being investigated has been obtained, the Polymerase Chain Reaction can be used to amplify 16S Ribosomal DNA from the culture. Amplified DNA samples that have been concentrated and purified can be taken to the College of Biological Sciences DNA Sequencing Facility at UC Davis, and sequence data obtained may be compared to known sequences using the NCBI BLAST. Although DNA sequences from the organisms cultured may not match exactly with any sequence thus far recorded, students should be able to determine the identity of closely related species. **Note** – There is a cost for sequencing DNA samples so individual student will generally not be able to identify more than one culture.

Procedure:

1. Obtain a pure culture of the microorganisms to be identified by streaking a small sample taken from a single, well-isolated colony onto a new agar plate. Carefully observe the resulting culture to check for any variation in morphology. If variation is evident, repeat the process until a pure culture is obtained. **Note** – Identification will be compromised if the culture is not pure.

2. Determine the cell wall characteristic of your culture (Gram stain and KOH test).

3. Obtain a sample of template DNA and amplify the 16S ribosomal DNA by following the procedures provided in “Application of the PCR in Bacterial Identification” for Gram-negative or Gram-positive organisms.

4. Subject your sample of amplified DNA (PCR product) to gel electrophoresis using the procedure outlined in “Gel Electrophoresis of DNA Samples”. Arrange for an instructor or instructional assistant to obtain a gel and set up a chamber. You will load the entire volume of your PCR product into the gel.

5. Quality PCR product DNA will be cut from gels, weighed, carefully labeled and then purified prior to sequencing. Purification requires the use of QIAquick gel purification kits (Qiagen) and is a relatively expensive endeavor. Purification will be completed by an instructor or instructional assistant unless otherwise specified.

6. Clean DNA samples will be taken to the College of Biological Sciences DNA Sequencing Facility at UC Davis, and the sequence data obtained will be returned via email. This will be made available to students on the microbiology Webpage.

7. Use the information provided in “Automated Nucleotide Sequencing and Electropherogram Evaluation” to edit and combine the nucleotide sequences obtained from your unknown culture. Use the BLAST associated with the NCBI to determine which organism type(s) your sequence aligns with.

8. Use the Internet, the Bergey’s manual, periodicals and other resources to acquire information about your culture. It may be beneficial to contact other researchers to gain information.

9. Enzymatic testing will be required to verify the identity of your unknown. Check with an instructor to determine the availability of media and then prepare or obtain those materials required to complete the identification.

When you have completed your identification you will be expected to turn in a viable, pure culture of the organisms being investigated and a **written description** of their cell and colony morphology, as obtained from stained smears and the colonies growing on solid media. Be sure to specify which medium or media promoted optimal growth of the culture and describe variations obtained due to temperature, light or other factors. In order to receive full credit for your project, you must also turn in a **written report** as indicated in "Semester Project Write-up Guidelines". Partial credit may be given for submission of the "Semester Project Data Record" alone, but this will be determined at the discretion of individual instructors.

Questions

1. What is a genome and how is it related to genomics?

2. What is a proteome and how is it related to proteomics?

3. What is the relationship between computer technology and a new science known as bioinformatics?

4. What do the letters NCBI stand for in association with this exercise, and what kinds of information are available at the NCBI website?

5. What do the letters BLAST stand for and how was the BLAST algorithm used in association with this exercise?