BIOINFORMATICS

For a better understanding of the experiment, perform the following bioinformatics exercises before you analyze your results.

I. Use BLAST to Find DNA Sequences in Databases (Electronic PCR)

1. Perform a BLAST search as follows:
   a. Do an Internet search for "ncbi blast."
   b. Click on the link for the result "BLAST: Basic Local Alignment Search Tool..." This will take you to the Internet site of the National Center for Biotechnology Information (NCBI).
   c. Click on the link "nucleotide blast" (blastn) under the heading "Basic BLAST."
   d. Enter both primer sequences into the search window:

The following primer set was used in the mtDNA experiment:	
Forward Primer	5′-TTAACTCCACCATTAGCACC-3′
Reverse Primer	5′-GAGGATGGTGGTCAAGGGAC-3′

   e. Omit any nonnucleotide characters from the window because they will not be recognized by the BLAST algorithm.
   f. Under "Choose Search Set," select the "Nucleotide collection (nr/nt)" database from the drop-down menu.
   g. In the text box beside Entrez Query, type "complete genome."
   h. Under "Program Selection," optimize for "Somewhat similar sequences (blastn)."
   i. Click on "BLAST." This sends your query sequences to a server at NCBI in Bethesda, Maryland. There, the BLAST algorithm will attempt to match the primer sequences to the millions of DNA sequences stored in its database. While searching, a page showing the status of your search will be displayed until your results are available. This may take only a few seconds or more than 1 minute if many other searches are queued at the server.
2. Analyze the results of the BLAST search, which are displayed in three ways as you scroll down the page:
   a. First, a graphical overview illustrates how significant matches (hits) align with the query sequence (shown in red). Matches are indicated by colored bars, with a color key at the top of the graphic indicating the range of scores assigned to each color. With primers, which are short, it is impossible to get high scores, as the score is dependent on the length of match between the query and

hit. **What do you notice about the lengths (and colors) of the matches (bars) as you look from the top to the bottom? What does this mean?**

b. This is followed by a list of significant alignments (hits) with links to the corresponding accession numbers. (An accession number is a unique identifier given to a sequence when it is submitted to a database such as GenBank.) Note the scores in the "E value" column on the right. The Expectation or E value is the number of alignments with the query sequence that would be expected to occur by chance in the database. So, an E value of 1 means that a search with your sequence would be expected to turn up one match by chance. The lower the E value, the higher the probability that the hit is related to the query. E values will decrease as the length of matched sequence increases or the percent identity increases. **What is the E value of the most significant hit(s) and what does this mean?**

c. Third is a detailed view of each primer (query) sequence aligned to the nucleotide sequence of the search hit (subject, abbreviated "Sbjct"). Note that the forward primer (nucleotides 1–20) and the reverse primer (nucleotides 21–40) often align within the sequence entry having the same accession number.

3. Click on the accession number link to open any hit that is labeled "complete genome."

   a. At the top of the report, note basic information about the sequence, including its length (in base pairs, or bp), database accession number, source, and references to papers in which the sequence is published. **What is the source and size of the sequence in which your BLAST hit is located?**

   b. In the middle section of the report, the sequence features are annotated, with their beginning and ending nucleotide positions ("xx..xx"). These features may include genes, coding sequences (CDS), ribosomal RNA (rRNA), and transfer RNA (tRNA). You will examine these features more closely in Part II.

   c. Scroll to the bottom of the data sheet. There you will find the nucleotide sequence to which the term "Sbjct" refers.

   d. Return to the BLAST results page.

4. Predict the length of the product that the primer set would amplify in a PCR (*in vitro*) as follows:

   a. Scroll down to the alignments section (third section) of the BLAST results page. Select a hit that is labeled "complete genome." Both primers should align to the sequence.

   b. **To which positions do the primers match in the subject sequence?**

   c. The lowest and highest nucleotide positions in the subject sequence indicate the borders of the amplified sequence. Subtract

the lowest nucleotide position in the subject sequence from the highest nucleotide position in the subject sequence. **What is the difference between the coordinates?**

d. **Note that the actual length of the amplified fragment includes both ends, so add 1 nucleotide to the result that you obtained in Step 4.c. to obtain the exact length of the PCR product amplified by the two primers.**

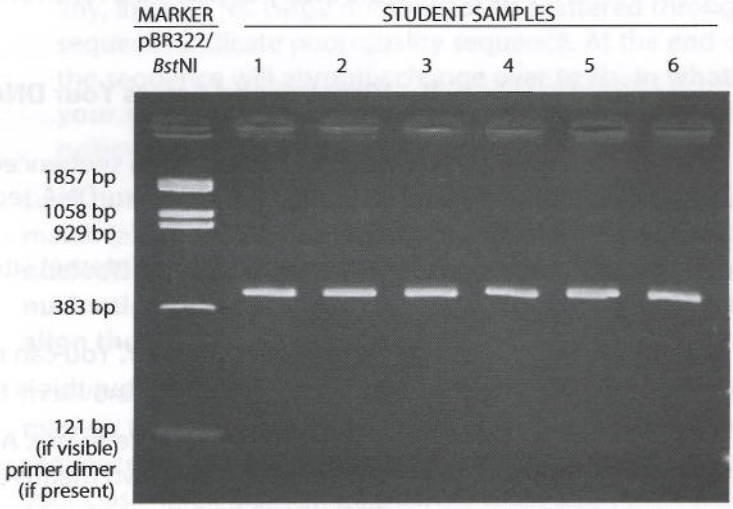
II. Examine the Structure of the Mitochondrial Chromosome and Its Genes

1. Continue working with the same hit that you used in Steps 4 and 5 of Part I. Open the data sheet for that hit. To help you answer the questions below, print the report or copy the report into a text document.
2. Look at the features section in the middle of the report: D-loop region, gene, coding sequence (CDS), ribosomal RNA (rRNA), and transfer RNA (tRNA) features may be annotated. Next to each feature are its nucleotide coordinates on the mitochondrial chromosome ("xx..xx").
3. Use the information in the features section and the nucleotide positions that match the primers (from Step 4.b. in Part I above) to identify the features that the amplicon spans.
   a. **How many features does the amplicon span? What are they called? What is the size of each feature? Be sure to include all of the sequences of each feature in your calculations.**
   b. **The largest feature that you identified in Step 3.a. has several names. Do some research to learn and understand these names.**
4. Study the general features of the mitochondrial genome.
   a. Look at the features list for your "complete genome" hit. Remember that genes can be located on either the forward or reverse strand; in the features section, the word "complement" indicates that a feature is transcribed from the reverse strand. **List the genes that are transcribed from the forward strand and those that are transcribed from the reverse (complement) strand. On which strand are most genes found?**
   b. **What kind of product do most of the genes produce? For what biological function are these products used?**
   c. **What is the spacing like between genes? Is there much intergenic DNA between two adjacent genes?**
   d. Focus on one of the protein-coding genes. Compare the nucleotide coordinates for the gene and the CDS (coding sequence). **What do you notice? What does this tell you about the structure and origin of mitochondrial genes?**

RESULTS AND DISCUSSION

I. Interpret Your Gel and Think About the Mitochondrial Genome

1. Observe the photograph of the stained gel containing your PCR samples and those from other students. Orient the photograph with the sample wells at the top. Use the sample gel shown below to help to interpret the band(s) in each lane of the gel.



2. Locate the lane containing the pBR322/*Bst*NI markers on the left side of the gel. Working from the well, locate the bands corresponding to each restriction fragment: 1857, 1058, 929, 383, and 121 bp. The 1058- and 929-bp fragments will be very close together or may appear as a single large band. The 121-bp band may be very faint or not visible.
3. Scan across the row of student results. You should notice that virtually all student lanes contain one prominent band. The amplification product of 440 bp should roughly align with the 383-bp marker.
4. It is common to see a diffuse (fuzzy) band that runs ahead of the 121-bp marker. This is "primer dimer," an artifact of the PCR that results from the primers overlapping one another and amplifying themselves. **How would you interpret a lane in which you observe primer dimer but no bands, as described in Step 3?**
5. The mitochondrial control region mutates at approximately 10 times the rate of nuclear DNA. **Propose a biological reason for the high mutation rate of mtDNA.**
6. **The high mutability of the mitochondrial genome means that it evolves more quickly than the nuclear genome. This makes the mitochondrial control region a laboratory for the study of DNA evolution. However, can you think of any drawbacks to this high mutation rate when studing evolution?**

7. There are numerous insertions of mtDNA in nuclear chromosomes. Most of these insertions are very ancient and are the result of a mitochondrial genome reduction process that transferred functional mitochondrial genes to the nuclear genome. A 540-bp insertion of the mitochondrial control region on chromosome 11 is an example of a relatively recent event that occurred about 350,000 years ago. **Would you expect any difference in the mutation rates of the control region sequence in the mitochondrial genome versus the chromosome 11 insertion? What implication does this have for the study of human evolution?**

II. Visualize and Assess Your DNA Sequence

If your DNA has not been sequenced, you can follow these directions to assess another student's mtDNA sequence.

1. Open the *BioServers* Internet site at the DNA Learning Center (www.bioservers.org).
2. Enter *Sequence Server*. You can register if you want to save your work for future reference, but this is not required.
3. The interface is simple to use: Add or obtain data using the top buttons and pull-down menus and then work with the data in the work space below.
4. Click on "MANAGE GROUPS" at the top of the page.
5. Scroll down the default "Classes" list to find your class and click on the check box to select it.
6. Click on "OK" to move your class data into the work space.
7. Select your sample number from the pull-down menu. Click "OPEN" to view your sequence.
8. To view a chromatogram of your sequence, you will need to download the sequence data (trace file) as well as viewer software.
   a. Follow the "Trace File" link and the associated directions and save the .abi file to your desktop.
   b. Follow the link to "DNA Sequencing Core (University of Michigan)" for a list of viewers to download. Alternatively, search the Internet for Geospiza's FinchTV, another viewer.
9. Double-click on the trace file to launch your sequence in the chromatogram viewer. (If this does not work, first open the chromatogram viewer, click on "File" and then on "Open," and navigate to select your .abi trace file.)
10. Inspect your chromatogram. Remember that the primers amplify a 440-nucleotide sequence, so it is physically impossible to generate a sequence (a read) longer than this.

a. Each peak represents the fluorescence measured at that nucleotide position. Whenever possible, the software "calls" each peak as an A, T, C, or G. **What does the amplitude (height) of each peak represent? What do you notice about the amplitude of the peaks at the beginning of the sequence?**

b. Every sequence will begin with nucleotides (A, T, C, G) interspersed with Ns. In "clean" sequences, where experimental conditions were near optimal, the initial Ns will end within the first 25 nucleotides. The remaining sequence will have very few, if any, internal Ns. Large numbers of Ns scattered throughout the sequence indicate poor-quality sequence. At the end of the read, the sequence will abruptly change over to Ns. **In what part(s) of your sequence read are the most Ns found? What do you notice about the trace at N positions?**

11. Load two different trace files into different windows and attempt to match (align) the DNA sequences. Compare the two reads. **Do the nucleotide numbers (positions) correspond to the same nucleotide sequence in both reads? What trick did you use to align them?**

12. In several percent of cases, a clean sequence ends abruptly about midway through the read and then gives way to a poor-quality sequence with many Ns. Check to see if this situation describes any of your classmates' sequences; if so, manually align their sequence with a clean sequence. **What sequence feature coincides with the abrupt transition from a clean to a poor-quality sequence?**

III. Assess the Extent of Mitochondrial Variation in Modern Humans

1. If you are not already in the *Sequence Server* of the *BioServers* Internet site, open it as described in Steps 1 and 2 of Part II. Then click on "MANAGE GROUPS" at the top of the page.

2. Select "Modern Human mtDNA" from the "Sequence sources" pull-down menu in the upper right-hand corner. Click in the check boxes to select the African, Asian, and European mitochondrial DNA samples and click "OK" to move these groups into the work space.

3. If your class data are not already in the work space, follow Steps 4–6 from Part II to add them. (If your DNA was not sequenced, you can follow these instructions to add data from another class.)

4. Use the pull-down menus and check boxes to select a student sequence to compare. If you determined in Step 10 of Part II that your sequence read was clean, make sure that the check box next to your sequence is marked. (If your sequence was not clean, choose another student's sequence that was clean.) Then choose one African, Asian, or European reference sequence and click on the check box to select it. Uncheck all other boxes; make sure that you end up with only two sequences checked.

5. Click on "COMPARE" in the gray bar. (The default operation is multiple sequence alignment using the CLUSTAL W algorithm.) The checked sequences are sent to a server at Cold Spring Harbor Laboratory, where the CLUSTAL W algorithm will attempt to align each nucleotide position. This may take only a few seconds or more than a minute if a lot of other searches are queued at the server.
6. The results will appear in a new window. Examine the sequences, which are displayed in rows of 25 nucleotides. The alignment typically begins with yellow-shaded dashes (–), indicating an initial stretch of nucleotides that is present in one of the samples but not in the other. This occurs because the sequence read usually begins at different points in different samples. Yellow highlighting also denotes mismatches between the sequences. A gray-shaded "N" indicates a sequence error, a position in one or both sequences where a nucleotide could not be determined.
7. Scan into the sequence beyond any initial sequence errors or patches of mismatch. These mismatches are due to poor sequence near the sequencing primer and should not be counted as polymorphisms. This typically occurs at position 25–40. Once you are into clean sequence, count the number of polymorphisms (differences) between the two individuals as follows:
   a. Count all yellow-shaded nucleotide differences. These are single-nucleotide polymorphisms (SNPs).
   b. Count all deletions (–). If you see a string of internal dashes (– – – –), this likely arose from a single mutation event and thus should be scored as a single polymorphism.
   c. Do not count any Ns. Sequences with more than one or two internal Ns or with multiple nucleotide differences on every line are not reliable. If you detect this, select other sequences with which to work.
   d. Stop counting when you again encounter frequent Ns and/or sequence mismatches at the end of the alignment. This should occur at about position 375–400.
8. Record the names of your samples and the total number of polymorphisms that you counted in Step 7. In addition, record the number of nucleotides of clean sequence you counted (typically 325–400).
9. Use the pull-down menus and check boxes in your work space to select a different pair of individuals and repeat Steps 5–8. Continue until you have made at least five comparisons that sample a variety of populations and recorded your results. (You do not need to use the sequences from your class for all comparisions; the idea is to compare many pairs of modern human mtDNA sequences.)

10. For each pair of sequences, calculate the percent difference:

$$\text{difference (\%)} = \frac{\text{number of nucleotide differences}}{\text{total number of nucleotides counted}} \times 100.$$

11. Share class data by comparing many different sequence pairs. Exclude any "outliers" with suspiciously large numbers of SNPs; these comparisons likely include low-quality sequence for one or both sequences. **Use these data to determine the average number of mutations, as well as the average percent difference, between pairs. What do these numbers tell us about human variation?**

12. Mutations accumulate one at a time. If one assumes a constant mutation rate, then we can estimate how often a new mutation arises. However, this "mutational clock" needs to be set by some external event. In this case, we can use anthropological data that suggest that modern humans arose about 200,000 years ago. **If this is so, what is the approximate mutation rate for the mitochondrial control region? Why is your calculation only approximate?**

IV. Assess Mitochondrial Variation in Ancient Samples

A. Otzi the Iceman

1. If you are not already in the *Sequence Server* of the *BioServers* Internet site, open it as described in Steps 1 and 2 of Part II. Then click on "MANAGE GROUPS" at the top of the page.
2. Select "Ancient Human mtDNA" from the "Sequence sources" pull-down menu in the upper right-hand corner.
3. Click in the check box to select "Otzi the Iceman mtDNA," and click "OK" to move this sequence onto the work space.
4. Use the pull-down menus and check boxes to select a modern human from your class, Africa, Asia, or Europe to compare with Otzi. (If these sequences are not already in your work space, follow Steps 4–6 from Part II and Step 2 of Part III to add them.)
5. Perform Steps 5–8 and Step 10 of Part III to compare the number of differences and percent difference between Otzi and the chosen sequence.
6. Use the pull-down menus and check boxes in your work space to select a different individual to compare to Otzi and repeat Steps 5–8 and 10 of Part III. Continue until you have compared Otzi to each of at least five modern humans and recorded your results.
7. **Pool the class data to determine the average number of differences and average percent difference between Otzi and modern humans. What does this tell you about Otzi and the DNA mutational clock?**

B. Neandertals

1. If you are not already in the *Sequence Server* of the *BioServers* Internet site, open it as described in Steps 1 and 2 of Part II. Then click on "MANAGE GROUPS" at the top of the page.
2. Select "Neanderthal mtDNA" from the "Sequence sources" pull-down menu in the upper right-hand corner.
3. Click in the check boxes to select several samples of Neandertal mtDNA and click "OK" to move these sequences onto the work space.
4. Using the pull-down menus and check boxes, select a modern human from your class, Africa, Asia, or Europe to compare with one of the Neandertal samples. (If these sequences are not already in your work space, follow Steps 4–6 from Part II and Step 2 of Part III to add them.)
5. Perform Steps 5–8 and 10 of Part III to compare the number of differences and percent difference between the Neandertal sequence and the modern human sequence.
6. Use the pull-down menus and check boxes in your work space to select a different Neandertal–modern human pair and repeat Steps 5–8 and 10 of Part III. Continue until you have compared at least five Neandertal–modern human pairs and recorded your results.
7. Compare each possible pair of Neandertals to assess the extent of variation in Neandertals across their habitat range.
8. **Pool the class data to determine the average number of differences and average percent difference between Neandertals and modern humans—and between Neandertals. What does this tell you about the relationship between Neandertals and modern humans?**
9. **If modern humans arose 200,000 years ago, approximately how long ago did humans and Neandertals share a common mitochondrial ancestor?** (Assume that the mutational clock is running at the same rate as in Step 12 of Part III.)
10. Scientists estimate that Neandertals lived on Earth for approximately 300,000 years. **What does the level of genetic variation in Neandertals suggest about Neandertal population changes during the course of their existence on Earth?**

C. Denisovans

1. Recently, the mtDNA sequence was determined for an ancient fragment of finger bone and tooth found in Denisova Cave, a remote location in the Altai mountains of Siberia. This cave is known to have been occupied by Neandertals and modern humans. Using sequence analysis, determine whether these fragments are Neandertal or modern human.
2. Select "Unknown Hominid Ancestor" from the "Sequence sources" pull-down menu in the upper right-hand corner.

3. Click in the check box to select the sequences and click "OK" to move them onto the work space.
4. Perform Steps 5–8 and 10 of Part III to compare the number of differences and percent difference between each of these sequences and a modern human sequence. **Pool the class data to determine the average number of differences and average percent differences. Does your result suggest that the samples are modern human?**
5. **If modern humans arose 200,000 years ago, approximately how long ago did modern humans and the people from Denisova Cave share a common ancestor?**
6. Perform Steps 5–8 and 10 of Part III to compare the number of differences and percent difference between each of these sequences and a Neandertal sequence. **Pool the class data to determine the average number of differences and average percent differences. What does this suggest? Discuss.**
7. **Does the mitochondrial DNA tell the whole story? Explain.**
8. Next-generation sequencing has now made it possible to analyze the nuclear DNA from many different modern humans, Neandertals, and Denisovans. **Do a literature search and summarize what we now know about the relationship of these two extinct hominids and modern human populations.**

V. Discover What DNA Says About Human Evolution

Use the chart below to record your answers to the questions that follow.

African 1	African 2	European or Asian	African different	European or Asian different

1. If you are not already in the *Sequence Server* of the *BioServers* Internet site, open it as described in Steps 1 and 2 of Part II. Then follow Step 2 of Part III to retrieve the African, Asian, and European mtDNA samples if they are not already in your work space.
2. Use the pull-down menus and check boxes to select two African samples and one European or Asian sample. Record the names of the sample populations in the chart.
3. Click on "COMPARE" (Align: CLUSTAL W) to perform a multiple sequence alignment.

4. Evaluate each yellow-shaded mismatch, where one sample is different from the other two (see Steps 6 and 7 of Part III). For each position, determine whether one of the Africans is different or the European/Asian is different. Record tick marks in the chart for each mismatch.
5. Use the pull-down menus and check boxes in your work space to select a different set of individuals and repeat Steps 2–4. Continue until you have made a total of five comparisons and recorded your results in the chart.
6. **What do your results show? Is this what you expected?**
7. **Pool the class results to confirm or refute your own results.**
8. **What does your analysis say about African mtDNA diversity? What population factors could account for this?**
9. **What does your analysis say about European and Asian mtDNA diversity? What population factors could account for this?**