**Yeti bioinformatics lesson introduction**

Bioinformatics is a powerful tool in modern biology. The ability to obtain DNA sequence through a variety of methods has led to an abundance of data that can be analyzed for numerous purposes. Introducing bioinformatics in a basic biology course (advanced high school or introductory undergraduate course) allows students to learn how to gather basic information regarding the power of sequence analysis.

Using DNA sequencing to identify species is a powerful research tool, whether it be in a microbiology laboratory to understand diversity in a sample (McCabe et al, 1999), in environmental DNA identification (Kelly *et al*. 2014), in crime scene investigation (Melton *et al.,* 2012) or in the identification of unknown species (Sykes *et al*., 2014). Isolation of DNA samples from the environment, known or reportedly known samples, or crime scenes has advanced to be able to use very small samples of material to isolate even degraded DNA (Rohland *et al*, 2018). One common method to characterize the DNA is to use PCR to amplify a small region of the mitochondria genome. Specifically, researchers are amplifying portions of the well-conserved 12S or 16S rRNA genes from the mitochondria (Melton and Holland, 2007; Melton *et al*., 2012; Yang *et al*., 2014). While these are powerful approaches to compare samples to previously characterized sources of DNA, caution needs to be taken when this approach is used for small fragments as there can be some overlap between species and degradation can lead to errors in the obtained sequence (Yang *et al.,* 2014).

This pair of assignments utilizes the mythical Yeti to illustrate DNA sequence analysis and then bioinformatics to characterize the samples. The assignments stem from the research published by Sykes *et al*. (2014) that investigated putative Yeti artifacts that have been collected and stored in museums, personal collections and other repositories. With the emergence of DNA sequencing technologies and the ability to sequence small amounts of material - along with the willingness of those who possess these rare artifacts to allow testing - scientists can assign identities of the specimens to either a known species or to the Yeti as a yet unknown DNA sequence (not matching any previously submitted database sequence).

I have given these out-of-class assignments in my introductory biology course after having introduced the basics of DNA replication, PCR and sequencing. I have used an online homework resource to post the questions from the assignment. Some of the questions are automatically graded while others require manual grading. Much of the work that the students do, especially in the second assignment, creates alignments that are used to answer the subsequent questions.

In the first assignment, students are asked basic questions regarding DNA inheritance, methods of basic sequencing technology as compared with prokaryotic DNA replication, as well as given the opportunity to read an electropherogram. This assignment is generally a review for the basic information, but allows students to read a relatively basic and clear electropherogram. It is important that the students be able to see this in color.

In the second assignment, students are introduced to BLAST searches (Altschul *et al*., 1990) using FASTA formatting as well as ClustalOmega multiple sequence alignment tool (Goujon *et al.*, 2010; Siever *et al.*, 2011). The six sequences, termed ‘Yeti’ in the FASTA file, are partial sequences from Sykes *et al*., 2014. The GenBank identities of these files are shown in Table 1. Sykes *et al*. (2014) generated sequence data from 30 different samples, however, analysis of 30 samples seems too great for this demonstration. Sykes *et al*. (2014) demonstrate that all of the artifacts that were sampled matched to known (and even presumably extinct) species. The connection of samples to extinct species has been contested by several authors (Edwards and Barnett, 2014; Gutiérrez and Pine, 2015; Lan *et al.*, 2017). A ‘breaking news’ addition to part 2 provides a new (currently fictional) sequence (called Alaska1). This created sequence is similar to the primate sequences in the database, but was selectively edited to not match anything in the database. The sequence currently does not match any database submission, although a synthetic sequence similar to the human sequence does appear in the database and does cause some confusion. This ‘breaking news’ should remind students that the absence of a match just indicates that we have not identified every organism and that new entries may suggest the discovery of a new species or support the presence of a mythical animal.

**Reference**

Altschul, SF, Gish, W, Miller, W, Myers, EW, and Lipman, DJ. 1990. Basic local alignment search tool. J. Mol. Biol. 215: 403-410.

Edwards, CJ, and Barnett, R. 2014. Himalayan ‘yeti’ DNA: polar bear or DNA degradation? A comment on ‘Genetic analysis of hair samples attributed to yeti’ by Sykes *et al.* (2014). Proc. R. Soc. B. 282: 20141712.

Goujon, M, McWilliams, H, Li, W, Valentin, F, Squizzato, S, Paern, J, Lopez, R. 2010. A new bioinformatics analysis tools framework at EMBL-EBI. Nucleic Acids Research 38: Suppl W695-9.

Gutiérrez, EE, and Pine, RH. 2015. No need to replace an “anomalous” primate (Primates) with an “anomalous” bear (Carnivora, Ursidae). ZooKeys 487: 141-154.

Kelly, RP, Port, JA, Yamahara, KM, Martone, RG, Lowell, N, Thomsen, PF, Mach, ME, Bennett, M, Prahler, E, Caldwell, MR, Crowder, LB. 2014. Harnessing DNA to improve environmental management. Science 344: 1455-1456.

Lan, T, Gill, S, Bellemain, E, Bischof, R, Nawaz, MA, Lindqvist, C. 2017. Evolutionary history of enigmatic bears in the Tibetan Plateua – Himalaya region and the identity of the yeti. Proc. R. Soc. B 284: 20171804.

McCabe, KM, Zhang, YH, Huang, BL, Wagar, EA, McCabe, ER. 1999. Bacterial species identification after DNA amplification with a universal primer pair. Mol. Genet. Metab. 66: 205-211.

Melton, T, and Holland, C. 2007. Routine forensic use of the mitochondrial 12S ribosomal RNA gene for species identification. J. Forensic Sci. 52: 1305-1307.

Melton, T, Holland, C, and Holland, M. 2012. Forensic mitochondrial DNA: current practice and future potential. Forensic Science Review 24: 110.

Rohland, N, Glocke, I, Aximu-Petri, A, Meyer, M. 2018. Extraction of highly degraded DNA from ancient bones, teeth and sediments for high-throughput sequencing. Nature Protocols 13: 2447-2461.

Sievers, F, Wilm, A, Dinee, DG, Gibson, TJ, Karplus, K, Li, W, Lopez, R, McWilliams, H, Remmert, M, Söding, J, Thompson, JD, Higgins, D. 2011. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Molecular Systems Biology 7: 539.

Sykes, BC, Mullis, RA, Hagenmuller, C, Melton, TW, and Sartori, M. 2014. Genetic analysis of hair samples attributed to yeti, bigfoot and other anomalous primates. Proc. R. Soc. B. 281: 20140161.

Yang, L, Tan, Z, Wang, D, Xue, L, Guan, M, Huang, T, Li, R. 2014. Species identification through mitochondrial rRNA genetic analysis. Scientific Reports 4: 4089.

**Yeti or not: Part 1**

**Introduction**

Cryptozoology is the study of creatures that live in folklore (and maybe your backyard). These creatures include the Loch Ness monster (Scotland), the chupacabra (Puerto Rico, Central and South America), and bigfoot (United States and Canada). There are many other cryptid creatures found throughout folklore. While some of these creatures have a history in a specific region of the world, bigfoot-like creatures have been reported through a wide range of habitats and go by different names (see table below). Cryptozoologists have reported physical and behavioral differences in these bigfoot-like creatures, suggesting that they may not all be the same type of animal. While the reports are not uniform, they often describe a large, upright ape-like creature.

Table 1. Local names of bigfoot-like cryptid animals and locations. (https://exemplore.com/cryptids/Names-for-Bigfoot-Around-the-World)

|  |  |
| --- | --- |
| Local name | Location of ‘sightings’ |
| Bigfoot/sasquatch | Northern United States and Canada |
| Yeti | Himalayas |
| Almasy | Mongolia |
| Yowie | Australian outback |
| Skunk ape | Florida and southern United States |
| Grassman | Ohio |
| Wendigo | Canada |
| Orange pendek | Sumatra |
| Mapingauri | South America |
| Yeren | China |

While artifacts have been collected from a variety of sites and are reportedly bigfoot or yeti (or related) samples, supporting scientific evidence has been lacking. Biological artifacts that have been reportedly collected - including hair, bone and tissue samples - from a variety of locations. Additional artifacts include eye witness accounts as well as photographic evidence and castings of footprints. Many cryptozoologists have been in search of more conclusive evidence of these legendary creatures, but no entire individuals, bodies or specimens have ever been collected, only reported artifacts.

Modern molecular biology has advanced to a point where analysis of biological artifacts can be used to look for biological macromolecules, including DNA. DNA that is of sufficient quality for Next Generation sequencing can be isolated from artifacts and ancient DNA samples. Sequencing of ancient hominids (e.g., Neanderthal and Denisovan- Noonan *et al*., 2006; Meyer et al., 2012) and mammoths (Miller *et al*., 2008) have been reported. Initially scientists focused on the sequence of the mitochondrial genome and more recently on nearly complete nuclear genomes. The analysis of these genomes is ongoing, but is providing insight into evolution of modern, related species.

Several groups have taken the molecular approach to identify purported samples of bigfoot/yeti origin to confirm the identity. In this assignment, you will be asked basic questions about DNA inheritance and molecular methods. In the following assignment, you will be using sequencing data to identify the samples as well as compare samples.

Answer the questions below based on your knowledge of the topics. You might have to do some research to answer the questions, but should be able to reference a basic biology textbook or genetics book to obtain the answers.

**Questions**

1. Inheritance of DNA in mammals
	1. How is nuclear DNA inherited in mammals?
	2. How is mitochondrial DNA inherited in mammals?
2. Compare the molecular method of DNA sequencing with prokaryotic DNA replication.
	1. How are the DNA polymerases used in nature (DNA replication) and sequencing different?
	2. What other steps in DNA sequencing are different than prokaryotic DNA replication?
	3. What are the special nucleotides used in sequencing, how are they different than the nucleotides used in prokaryotic DNA replication and why are they useful?
	4. How is prokaryotic DNA replication similar to PCR? How is it different?
3. Using the diagram, determine the sequence that is illustrated by this electropherogram (*this is not a sequence related to the next part of the case study)*.



* G = black, C = blue
* A = green, T = red

(read peaks left to right – 5’ to 3’)

|  |  |  |  |
| --- | --- | --- | --- |
| Enter sequence  | 5’ |  | 3’ |

**References**

Meyer, M, Kircher, M, Gansauge, M-T, Li, H, Racimo, F, Mallick, S, Schraiber, JG, Jay, F, Prüfer, K, de Filippo, C, Sudmant, PH, Alkan, C, Fu, Q, Do, R, Rohland, N, Tandon, A, Siebauer, M, Green, RE, Bryc, K, Briggs, AW, Stenzel, U, Dabney, J, Shendure, J, Kitzman, J, Hammer, MF, Shunkow, MV, Derevianko, AP, Patterson, N, Andrés, AM, Eichler, EE, Slatkin, M, Reich, D, Kelso, J, Pääbo, S. 2012. A high-coverage genome sequence from an archaic Denisovan individual. Science 338: 222-226.

Miller, W, Drautz, DI, Rata, A, Pusey, B, Qi, J, Lesk, AM, Tomsho, LP, Packard, MD, Zhao, F, Sher, A, Tikhonov, A, Raney, B, Patterson, N, Lindblad-Toh, K, Lander, ES, Knight, JR, Irzyk, GP, Fredrikson, KM, Harkins, TT, Sheridan, S, Pringle, T, Schuster, SC. 2008. Sequencing the nuclear genome of the extinct woolly mammoth. Nature 456: 387-390.

Noonan, JP, Coop, G, Kudaravalli, S, Smith, D, Krause, J, Alessi, J, Chen, F, Platt, D, Pääbo, S, Pritchard, JK, Rubin, EM. 2006. Sequencing and analysis of Neanderthal genomic DNA. Science 314: 1113-1118.

**Yeti bioinformatics case study**

**Part 2**

In order to validate some biological Yeti artifacts, a group of researchers was able to isolate DNA from the artifacts and determine the sequence of a gene within the mitochondrial genome. This gene is the 12S rRNA gene involved in ribosome function within the mitochondria. (*The mitochondrial genome encodes rRNA and some ribosomal proteins. There is translation that occurs within the mitochondria that uses these different rRNA and ribosomal proteins. The 12S rRNA is the ‘small’ rRNA in the mitochondria of eukaryotic cells.)*

Bioinformatics is the use of computer programs to analyze DNA (and protein) sequences. Bryan Sykes and his research group collected samples, isolated DNA and obtained DNA sequences from 30 potential Yeti artifacts (Sykes *et al*., 2014). These sequences were compared to each other and to databases to determine if they were a match to a known sample (public sequence databases contain millions of contributed sequences and the 12S rRNA gene is a common sequence for comparison of organisms. Using a partial sequence (small region) of the 12S rRNA gene, called a hypervariable region, a researcher can determine which organism provided the DNA. Using this approach, Sykes group was able to compare the samples with known samples in the database.

A common program that can be used to compare a sample DNA sequence to a database of sequences is known as BLAST (basic local alignment sequence tool) (Altschul *et al*., 1990). The algorithm that compares the query sequence (your sample) with the database looks at short ‘words’ or a series of bases and compares that to everything in the database (millions of sequences). It continues to look at consecutive ‘words’ (Figure 1), building matches and scoring the quality of the match. In the end, the program can find the best matches within the database. In 1982, Genbank (one public database) was formed with 606 sequences in the database. In April 2019, the database has over 212 million sequences (over 320 trillion bases) (https://www.ncbi.nlm.nih.gov/genbank/statistics/). While the database is large, it is not untypical for a basic BLAST search to be completed within 10 seconds. While this is a powerful research tool, we do not need (at this time) to understand the algorithm.



Figure 1. Schematic of ‘word’ searching for the BLAST algorithm.

Sequences can be individually entered into the query box (more below) or as a group of sequences. One common means of presenting multiple sequences is in a file format called FASTA. In this format, the title of the sequence file is entered with a > before the name (see Figure 2). On a return line, the sequence is entered. Multiple sequences can be entered on their own lines in this format. Many bioinformatics programs (in addition to BLAST) can utilize the FASTA format. The sequence format provided in Figure 2 is in the FASTA format.

Using one of the collections of samples, you will be performing bioinformatic analysis on the results to determine if they come from a known or unknown organism.

Question 1. What might an ‘unknown’ sequence suggest?

Figure 2. ‘Yeti’ sequences from Sykes *et al*., 2014. This is a set of six sequences withdifferent identifiers than are provided in the manuscript. The FASTA file format uses a > preceeding a sequence identifier (Yeti in this example). On the subsequent line, the sequence is entered. This format allows for multiple sequences to be entered at the same time. Individual sequences can be posted using this same format.

>Yeti1

CTTAGCCTTAAACATAAATAATTTATTAAACAAAATTATTCGCCAGAGAACTACTAGCAACAGCTTAAAACTCAAAGGACTTGGCGGTGCTTTAAACCCTCCTA

>Yeti2

CTTAGCCTTAAACATAAATAATTTATTAAACAAAATTATTCGCCAGAGAACTACTAGCAACAGCTTAAAACTCAAAGGACTTGGCGGTGCTTTAAACCCTCCTA

>Yeti3

CTTAGCCCTAAACATAGATAATTTTACAACAAAATAATTCGCCAGAGGACTACTAGCAATAGCTTAAAACTCAAAGGACTTGGCGGTGCTTTATATCCCTCTA

>Yeti4

CTTAGCCCTAAACATAAATAATTGTAAAAACAAAATTATTCGCCAGAGTACTACCGGCAACAGCCCAAAACTCAAAGGACTTGGCGGTGCTTTATATCCATCTA

>Yeti5

CTTAGCCCTAAACACAGATAATTACATAAACAAAATTATTCGCCAGAGTACTACTAGCAACAGCTTAAAACTCAAAGGACTTGGCGGTGCTTTATATCCTTCTA

>Yeti6

CTTAGCCCTAAACTCTAATAGTTACATTAACAAAACCATTCGCCAGAGTACTACAAGCAACAGCTTAAAACTCAAAGGACTTGGCAGTGCTTTATATCCCTCTA

Figure 2. FASTA sequence file format of six Yeti artifacts.

1. Using the FASTA sequence provided (Figure 2), you will perform a BLAST search to determine the genus (and possibly species) of the origin of the sample.
	1. Open browser window to https://blast.ncbi.nlm.nih.gov.
	2. In the window, click on Nucleotide BLAST on the left hand side of the window. *The other ‘buttons’ are alternative versions of the BLAST program that either use amino acid sequences for the search or search the database in different manners. We just need the basic BLAST program that compares nucleotides to the nucleotide database. There are additional parameters that can be used for the search, but for this assignment all you need to do is use the default options.*
	3. Copy and paste your FASTA file (from Figure 2) into the white box.
	4. Scroll down and click on the BLAST button (blue).
	5. The computer algorithm will run for a short period of time and provide a new screen with the information needed. Near the top of the screen, there is a ‘Results for’ window with the identity of each query sequence entered. You will need to select different query sequences to complete the assignment. Each time you change the query in the search, it may take a short time to upload the results of the selected sample.
		1. You should see a Graphic summary on the screen and the information that you are looking for is below that graphic (scroll down).
		2. The next section is titled “sequences producing significant alignments”. This has some of the information. The computer program will (by default) provide up to 100 matches (best matches and newest matches appear at top of list).
		3. Using the results of the BLAST search, fill out the table below. To get the Identities information, you can either click on the Description (link) or scroll down to the first sample. (After you get the information for Yeti1, go to the top of the page, to the “Results for” and select the next query sequence. Repeat to complete the table.). Use the first identified (not synthetic or unidentified) entree

|  |  |  |  |
| --- | --- | --- | --- |
| Sample | Accession | Description (record just genus and species here) | Identities (record as #/#) |
| Yeti1 |  |  |  |
| Yeti2 |  |  |  |
| Yeti3 |  |  |  |
| Yeti4 |  |  |  |
| Yeti5 |  |  |  |
| Yeti6 |  |  |  |

Question 2 will come from the table above.

* + 1. Are any of the samples of unknown origin?
		2. Do any of the samples match the same genus/species description?
		3. What are the common names of the origin of the samples? *Hint: Use Google.com if you don’t know the common name.*

|  |  |  |
| --- | --- | --- |
| **Sample** | **Genus/species (from table above)** | **Common name** |
| Yeti1 |  |  |
| Yeti2 |  |  |
| Yeti3 |  |  |
| Yeti4 |  |  |
| Yeti5 |  |  |
| Yeti6 |  |  |

Question 3 will come from the table above

1. Now we can see how different these sequences are from each other. There are two different manners in which this can be done. One way is to compare each sequence to the others. We can do this using the BLAST program.
	1. At the top of the BLAST page that you are on, select ‘edit and resubmit’ near the top left side. This will bring you back to the start window.
	2. Below the window where you entered your FASTA file is a small box next to “Align two or more sequences”. Click on that button (a checkmark should appear). Paste your FASTA file into the second window (make sure that it is still in the first window).
	3. Run BLAST again. This time, you are comparing your query files against themselves.
	4. When the program has run, make note of which sequence is in the “Results for” box at the top of the page.
		1. Scroll down to the alignments and find the other sequences.
		2. Record your ‘identities’ for each OTHER sample in the table below.
		3. The table below has been partially filled out since Yeti1 will be a 100% match to Yeti1 (and so on).
		4. Some of the matches may not cover the entire sequence entered. At this time, we will not be dealing with that.
		5. In some cases, there may not be a match provided. Why do you think this might be?

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Yeti1 | Yeti2 | Yeti3 | Yeti4 | Yeti5 | Yeti6 |
| Yeti1 | 100% |  |  |  |  |  |
| Yeti2 |  | 100% |  |  |  |  |
| Yeti3 |  |  | 100% |  |  |  |
| Yeti4 |  |  |  | 100% |  |  |
| Yeti5 |  |  |  |  | 100% |  |
| Yeti6 |  |  |  |  |  | 100% |

1. A second approach is to perform a multiple sequence alignment (MSA) of the sequences. This approach aligns all of the sequences that you input into the program and creates a stacked alignment similar to the pairwise alignments seen in the BLAST searches.
	1. There are different programs available to do MSA. One common one is the CLUSTAL package. The algorithms that are employed with CLUSTAL have changed over time (and are not pertinent to this assignment). You will be using a newer version of CLUSTAL called CLUSTALOmega.
	2. Go to the website <https://www.ebi.ac.uk/Tools/msa/clustalo/>.
	3. Scroll down to Step 1 and in the box below that (which should say PROTEIN), select DNA. This changes the input format from amino acid sequences to nucleic acids.
	4. Paste your FASTA file into the window titled “sequences supported in any format”.
	5. Scroll down and click on the ‘submit’ button. You do not need to change any of the defaults.
	6. When the alignments are completed, copy and paste the MSA into the box below. If the formatting does not look aligned, change the text to Courier font (a uniform font width) and set the size of the font to 9 or 10.

Yeti6 CTTAGCCCTAAACTCTAATAGTTACATTAACAAAACCATTCGCCAGAGTACTACAAGCAA 60

Yeti4 CTTAGCCCTAAACATAAATAATTGTAAAAACAAAATTATTCGCCAGAGTACTACCGGCAA 60

Yeti1 CTTAGCCTTAAACATAAATAATTTATTAAACAAAATTATTCGCCAGAGAACTACTAGCAA 60

Yeti2 CTTAGCCTTAAACATAAATAATTTATTAAACAAAATTATTCGCCAGAGAACTACTAGCAA 60

Yeti3 CTTAGCCCTAAACATAGATAATTTT-ACAACAAAATAATTCGCCAGAGGACTACTAGCAA 59

Yeti5 CTTAGCCCTAAACACAGATAATTACATAAACAAAATTATTCGCCAGAGTACTACTAGCAA 60

 \*\*\*\*\*\*\* \*\*\*\*\*: :.\*\*\*.\*\* : \*\*\*\*\*\*\* \*\*\*\*\*\*\*\*\*\*\* \*\*\*\*\* .\*\*\*\*

Yeti6 CAGCTTAAAACTCAAAGGACTTGGCAGTGCTTTATATCCCTCTA 104

Yeti4 CAGCCCAAAACTCAAAGGACTTGGCGGTGCTTTATATCCATCTA 104

Yeti1 CAGCTTAAAACTCAAAGGACTTGGCGGTGCTTTAAACCCTCCTA 104

Yeti2 CAGCTTAAAACTCAAAGGACTTGGCGGTGCTTTAAACCCTCCTA 104

Yeti3 TAGCTTAAAACTCAAAGGACTTGGCGGTGCTTTATATCCCTCTA 103

Yeti5 CAGCTTAAAACTCAAAGGACTTGGCGGTGCTTTATATCCTTCTA 104

 \*\*\* \*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*.\*\*\*\*\*\*\*\*:\* \*\* \*\*\*

* 1. The asterisks below the aligned sequences indicate 100% consensus (all of the sequences have that base found in that position). *The : and . in the alignment indicate ‘similarity’, which in the nucleic acid alignment doesn’t mean too much for us.*
	2. At what position is the Yeti3 sequence shorter than the others (where is the missing base)? Question 4
	3. At what position (starting from the beginning) is the first mismatch? Question 5. Which sequence(s) is/are different than the others?
		1. It is harder to get a % match in this manner compared with the BLAST approach since you would have to look at all of the sequences rather than have the computer do so.
		2. But you do get a sense of where there are differences between all of the sequences?
	4. Looking at the MSA (you pasted in above), what bases (look at the mismatches) might be an indicator of the Yeti1 species compared with Yeti6? Determine the first two difference for Question 6.
1. You can also use the MSA program to generate a relationship tree of the samples.
	1. The phylogenic tree shows relationships between the samples. The more similar the samples, e.g., if they are identical – the closer they will be on the tree (shorter branches).
	2. Click on the Phylogenic Tree button above the alignment. Do not click on Send to Simple Phylogeny as this sends it to another program. The CLUSTALO version is sufficient.
	3. Looking at the tree and specifically the branches, which samples are closer matches? Does this correspond with the identities percent that you recorded from the BLAST search above? Discrepancies may exist since the BLAST program has the ability to not align the entire sequence (mismatches near the ends are left off and the full sequences are not aligned), while CLUSTALO aligns all of the sequences from the FASTA file.

**Breaking news:**

In a new discovery in the wilderness of Alaska, the skeletal remains of a large animal. The remains were largely dismantled by predators, and could not be identified in the field. Tissues samples were still present and were sent to a research laboratory for identification. Fragment analysis of the 12S rRNA was performed and the following sequence was obtained:

>Alaska1

CTTAGCCCTAAACATTAATAGTTACATTAACAAAATTATTCGCCAGAGTACTACAAGCAACAGCTTAAAACTCAAAGGACTTGGCAGTGCTTTATATCCCTCTA

1. Compare this sequence (using BLAST) to the database and identify the remains.
2. What information can you conclude from this sequence? (Short answer Question 7.)

**References**

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. 1990. Basic local alignment search tool. J. Mol. Biol. 215: 403-410.

Sykes, B.C., Mullis, R.A., Hagenmuller, C., Melton, T.W., and Sartori, M. 2014. Genetic analysis of hair samples attributed to yeti, bigfoot and other anomalous primates. Proc. R. Soc. B. 281: 20140161.

**Answer Key**

**Yeti or not Part 1**

1. Inheritance of DNA in mammals
	1. How is nuclear DNA inherited in mammals?

Mammalian nuclear DNA is inherited as a haploid set of chromosomes from each parent.

* 1. How is mitochondrial DNA inherited in mammals?

Mammalian mitochondria (and mitochondrial DNA) is inherited from the material donor (through the egg cell). The sperm mitochondria is usually destroyed if it makes it to fertilization.

1. Compare the molecular method of DNA sequencing with prokaryotic DNA replication.
	1. How are the DNA polymerases used in nature (DNA replication) and sequencing different?

In prokaryotic DNA replication, the cell uses two different DNA polymerases (I and III) that are involved in the polymerizing of the nucleotides of the leading (DNA polymerase III) and lagging (DNA polymerase III and I) strands. Additionally, in replication, DNA polymerase III serves a proofreading function, having exonuclease activity.

In DNA sequencing, the DNA polymerase used can be DNA polymerase III or a similar DNA polymerase (the thermostable Taq DNA polymerase is commonly used now) and often there is no exonuclease activity. DNA polymerase I is not used in sequencing reactions.

* 1. What other steps in DNA sequencing are different than prokaryotic DNA replication?

|  |  |  |
| --- | --- | --- |
|  | Prok. DNA replication | DNA sequencing |
| Denaturation | Uses helicase | Uses heat |
| Primers | Multiple RNA synthesized by primase (one leading strand & one for each Okazaki fragment) | Usually uses only 1 (DNA) |
| Nucleotides | Deoxyribonucleotides (dNTPs) | Deoxyribonucleotides (dNTPs) and dideoxyribonucleotides (ddNTPs) |
| Polymerase | Discussed in previous question |

* 1. What are the special nucleotides used in sequencing, how are they different than the nucleotides used in prokaryotic DNA replication and why are they useful?

Sequencing uses dideoxyribonucleotides. Like deoxyribonucleotides, there is no 2’ OH (it is a 2’ H). The 3’ OH group on deoxyribonucleotides is involved in the formation of the phosphodiester bond by the DNA polymerase. In the dideoxyribonucleotides, the 3’ OH is replaced with a 3’ H and this terminates the polymerase reaction.

* 1. How is prokaryotic DNA replication similar to PCR? How is it different?

PCR is the amplification of a specific region of DNA from a template. Both replication and PCR use primers (replication multiple RNA primers – PCR usually 2 DNA primers). The DNA polymerase is often Taq polymerase for PCR, which is a thermostable version of DNA polymerase III, which is common in many prokaryotes. The denaturation of the template DNA in replication involves the activity of helicase, while heat is used to denature the template in PCR.

1. Using the diagram, determine the sequence that is illustrated by this electropherogram (*this is not a sequence related to the next part of the case study)*.



* G = black, C = blue
* A = green, T = red

(read peaks left to right – 5’ to 3’)

|  |  |  |  |
| --- | --- | --- | --- |
| Enter sequence  | 5’ | ccgaagatggccgtctga (18 mer) | 3’ |

**Yeti or not Part 2**

1.e.iii.

|  |  |  |  |
| --- | --- | --- | --- |
| Sample | Accession | Description (record just genus and species here) | Identities (record as #/#) |
| Yeti1 | KJ155697 | *Ursus maritimus* | 104/104 |
| Yeti2 | KJ155697 | *Ursus maritimus* | 104/104 |
| Yeti3 | MK937053 | *Canis lupus* | 103/103 |
| Yeti4 | NC\_045205 | *Capricornis rubidus* | 104/104 |
| Yeti5 | MN510465 | *Bos taurus* | 104/104 |
| Yeti6 | AC025627 | *Homo sapiens* | 104/104 |

Note that accession numbers may be different each year as the search presents the newest matching sequence on the top. The Yeti1 and Yeti2 samples are identical. The Yeti6 sample currently (January 2020) has the first match being a synthetic DNA construct (CP034495) that is also a perfect match to the human sequence.

Question 2: What is the description (genus only) of Yeti#?

 Answers will vary based on the #. A random selection of 1-6 has been used in the past to minimize students sharing.

* + 1. What are the common names of the origin of the samples?

|  |  |  |
| --- | --- | --- |
| **Sample** | **Genus/species (from table above)** | **Common name** |
| Yeti1 | *Ursus maritimus* | Polar bear |
| Yeti2 | *Ursus maritimus* | Polar bear |
| Yeti3 | *Canis lupus* | Wolf/dog |
| Yeti4 | *Capricornis rubidus* | Red serow (goat-antelope) |
| Yeti5 | *Bos taurus* | Cow |
| Yeti6 | *Homo sapiens* | Human  |

Question 3: What is the common name of the organism for Yeti#?

Answers will vary based on the #. A random selection of 1-6 has been used in the past to minimize students sharing.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Yeti1 | Yeti2 | Yeti3 | Yeti4 | Yeti5 | Yeti6 |
| Yeti1 | 100% | 100% | 90% | 91% | 90% | 85% |
| Yeti2 |  | 100% | 90% | 91% | 90% | 85% |
| Yeti3 |  |  | 100% | 88% | 91% | 85% |
| Yeti4 |  |  |  | 100% | 90% | 85% |
| Yeti5 |  |  |  |  | 100% | 90% |
| Yeti6 |  |  |  |  |  | 100% |

Sometimes the Yeti6 doesn’t appear to line up with the other sequences. Repeating the BLAST run and selecting More Dissimilar Sequences solves this issue.

MSA

Yeti6 CTTAGCCCTAAACTCTAATAGTTACATTAACAAAACCATTCGCCAGAGTACTACAAGCAA 60

Yeti4 CTTAGCCCTAAACATAAATAATTGTAAAAACAAAATTATTCGCCAGAGTACTACCGGCAA 60

Yeti1 CTTAGCCTTAAACATAAATAATTTATTAAACAAAATTATTCGCCAGAGAACTACTAGCAA 60

Yeti2 CTTAGCCTTAAACATAAATAATTTATTAAACAAAATTATTCGCCAGAGAACTACTAGCAA 60

Yeti3 CTTAGCCCTAAACATAGATAATTTT-ACAACAAAATAATTCGCCAGAGGACTACTAGCAA 59

Yeti5 CTTAGCCCTAAACACAGATAATTACATAAACAAAATTATTCGCCAGAGTACTACTAGCAA 60

 \*\*\*\*\*\*\* \*\*\*\*\*: :.\*\*\*.\*\* : \*\*\*\*\*\*\* \*\*\*\*\*\*\*\*\*\*\* \*\*\*\*\* .\*\*\*\*

Yeti6 CAGCTTAAAACTCAAAGGACTTGGCAGTGCTTTATATCCCTCTA 104

Yeti4 CAGCCCAAAACTCAAAGGACTTGGCGGTGCTTTATATCCATCTA 104

Yeti1 CAGCTTAAAACTCAAAGGACTTGGCGGTGCTTTAAACCCTCCTA 104

Yeti2 CAGCTTAAAACTCAAAGGACTTGGCGGTGCTTTAAACCCTCCTA 104

Yeti3 TAGCTTAAAACTCAAAGGACTTGGCGGTGCTTTATATCCCTCTA 103

Yeti5 CAGCTTAAAACTCAAAGGACTTGGCGGTGCTTTATATCCTTCTA 104

 \*\*\* \*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*.\*\*\*\*\*\*\*\*:\* \*\* \*\*\*

Question 4. At what position is the Yeti3 sequence shorter than the others (where is the missing base)?

 Alignment base 26 is missing in Yeti3

Question 5. At what position (starting from the beginning) is the first mismatch?

 The first mismatch occurs at position 8 (C or T)

Question 6. Looking at the MSA (you pasted in above), what bases (look at the mismatches) might be an indicator of the Yeti1 species compared with the Yeti6? Determine the first two differences.

 Position 8 – Yeti1 is T and Yeti6 is C

 Position 14 – Yeti1 is A and Yeti6 is T

Question 7. What information can you conclude from this sequence?

The sequence is not a perfect match to any sequence in the GenBank database. This could suggest 1) errors in the sequencing or 2) this sequence belongs to a species that has not yet been introduced into the database.