**Worksheet 1 - Molecular Tools for Yeti Hunting**

**Learning Objectives**

*At the end of this worksheet, students should be able to...*

* describe the necessary features of the 12S rRNA mitochondrial gene that make it a common target for sequence typing
* describe how a simplified overview of the nucleotide BLAST algorithm identifies sequence similarity
* explain what DNA barcoding is and how variation in a DNA sequence can be used to identify an organism
* evaluate sequence alignments and understand potential limitations

*Disclaimer - Historic names have been included where appropriate, but the more modern participants in this story are fictitious.*

After finding a Yeti footprint, Eric Shipton explored the Himalayas for almost 50 years but was never able to find convincing evidence of this bigfoot-like creature. His granddaughter, Dr. Margaret Shipton, now wishes to use modern genetic approaches to continue her grandfather’s search in hopes of providing more definitive proof for a creature unknown to science. Intrigued, you have recently volunteered to join her research team. In preparation for your research, you need to learn more about how to test purported Yeti artifacts using this modern approach.

**Evidence for Bigfoot**

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 (Figure 1) have been reported through a wide range of habitats and go by different names (Table 1). 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.



Figure 1. Bigfoot cartoon (courtesy of Tom Stiglich).

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, Yeti, or related samples, corroborating scientific evidence is absent. Biological artifacts have been reportedly collected - including hair, bone and tissue samples - from a variety of locations. Additional evidence includes eyewitness 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 skeletons have ever been collected. Recently a procedure known as **DNA barcoding** has been developed that might be useful in the molecular identification of artifacts.

**Modern genetic approaches to interrogate biological 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. Bioinformatics is the use of computer programs to analyze DNA (and protein) sequences.

Sequencing of DNA from 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.

**DNA barcoding**

DNA barcoding is designed to create a uniform molecular approach to the characterization and cataloging of biodiversity focused on the analysis of well-conserved regions of the genome. Environmental DNA collection and analysis has been used to help identify species that exist in a location that may be difficult to physically spot or collect (Figure 2).

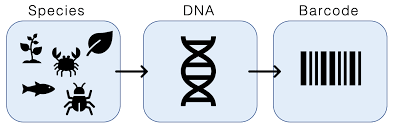


Figure 2. Schematic representation of DNA barcoding. Image Credit: Larissa Fruehe, CC BY-SA 4.0 (Wikipedia).

Scientists have identified a number of genes for DNA barcoding that have well-conserved regions as well as variable regions between species. The variable regions allow scientists to distinguish between genus and in some cases species. In animals, one target sequence that is commonly investigated is the 12S rDNA sequence from the mitochondrial genome (Figure 3). The 12S rRNA gene encodes the 12S rRNA, which is the small subunit rRNA used in mitochondrial polypeptide synthesis. This is a well-conserved gene, although since it encodes a non-coding RNA, variation is more common.

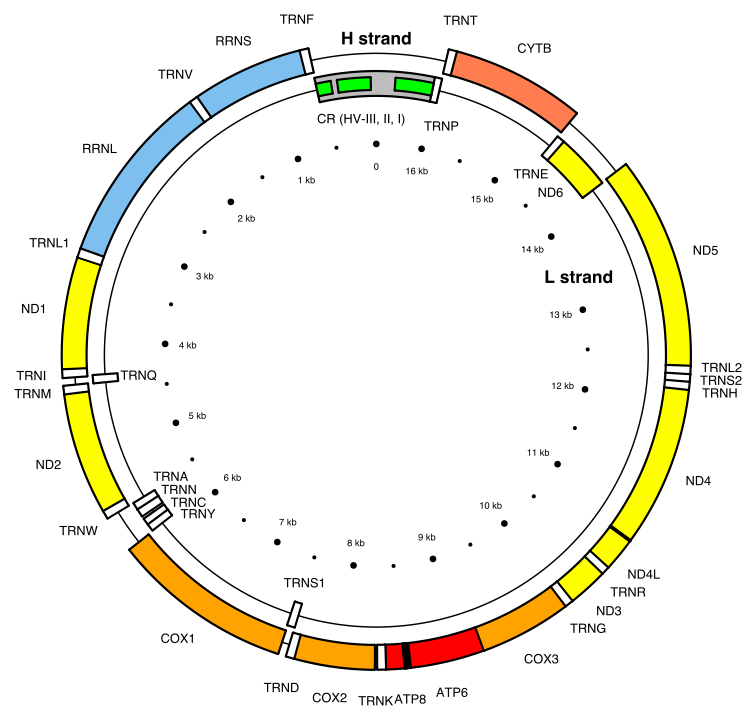


Figure 3. Map of the human mitochondrial genome. The mitochondrial genome contains some protein coding genes as well as the mitochondrial large (16S) rRNA (RRNL in Figure 3) and small (12S) rRNA (RRNS in Figure 3) genes (blue regions highlighted in figure). Human ancestry can be determined by sequencing a separate hypervariable region (HVR) (shown in green above). Image credit: Emmanuel Douzery, CC BY-SA 4.0, Wikipedia.

The well-conserved region of the 12S rRNA gene allows researchers to use the polymerase chain reaction to amplify these regions for DNA sequencing. Once the 12S rRNA gene sequence of the unknown sample has been determined, it is compared to sequences from known species to search for a match. There are currently thousands of 12S and 16S rRNA sequences within the Genbank database.

**Advantages of mtDNA for barcoding**

Mammalian mitochondrial DNA is maternally inherited and the human mitochondrial genome is circular and only 16,569 base pairs in size. The mitochondrial genome encodes genes for polypeptides that function within the mitochondria, as well as several rRNA genes involved in translation of the mitochondrially-encoded polypeptides. Mitochondria will contain one or more copies of the genome and cells may contain multiple mitochondria, making this one of the more abundant DNA molecules that can be found in mammalian samples.

The mitochondrial genome is relatively well conserved, allowing comparison between species and in some regions between individuals. Well-conserved regions allow scientists to amplify parts of the genome (using PCR) for sequence comparisons across more variable regions of the genome. The 12S rDNA is well-conserved within a species, but has demonstrated differences between species, allowing this region of the mitochondrial genome to facilitate the identification of the species from which the DNA was isolated. Using DNA sequences to identify species is called DNA barcoding.

Mitochondrial DNA is abundant in cells (multiple mitochondria in each cell with multiple copies of the genome in each organelle). This provides a more abundant source of genetic material than the comparable (although larger) nuclear genome. With historic (less than 100 years old) and ancient (more than 100 years old) DNA samples and the new technologies in DNA isolation and amplification using small amounts of DNA, mitochondrial sequences are being used to identify a variety of samples.

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). While this is a powerful research tool, we do not need (at this time) to understand the algorithm in depth. Briefly, the BLAST algorithm compares the query sequence (your sample) with the database by looking at short ‘words’ or a series of bases and comparing that to everything in the database (millions of sequences). It continues to look at consecutive ‘words’ (Figure 4), 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 2020, the database has over 216 million sequences (over 415 trillion bases). While the database is large, it is not untypical for a basic BLAST search to be completed within 30 seconds.

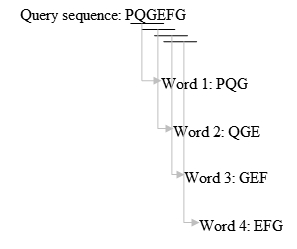


Figure 4. Schematic of ‘word’ searching for the BLAST algorithm. (Image from Wikipedia.)

The output of a BLAST search is an alignment between your query sequence and database subject (sbjct) sequence. While the computer algorithm can identify the quality of the match, you still need to understand what you are looking for with the results. Database (sbjct) sequences that are identical may be the same species or may be closely related species. Ideally, if you are looking for diversity, you are hoping to find matches that are not perfect. One consideration to make with a BLAST search is the location and length of the alignment. Short sequences may occur more frequently and be more conserved than longer sequences. The length of the sequence match does factor into the algorithm results (e-score).

Sequence alignment A

Query 1 CTTAGCCCTAAACCTCAACAGTTAAATCAACAAAACTGCTCGCCAGAACACTACGAGCCA 60

||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||

Sbjct 1 CTTAGCCCTAAACCTCAACAGTTAAATCAACAAAACTGCTCGCCAGAACACTACGAGCCA 60

Query 61 CAGCTTAAAACTCAAAGGACCTGGCGGTGCTTCATATCCCTCTA 104

||||||||||||||||||||||||||||||||||||||||||||

Sbjct 61 CAGCTTAAAACTCAAAGGACCTGGCGGTGCTTCATATCCCTCTA 104

**Question 1.** In Sequence alignment A (above), part of the human 12S rDNA sequence is used as the Query and has been used to search Genbank - one alignment result is shown. What can you infer from these sequences and your knowledge of the 12S rDNA locus?

When sequences are identical (as they are in Sequence alignment A for Question 1), this format (vertical lines showing identity) is a good format. However, when there are differences (which would be missing vertical lines), this gets more difficult to visualize, especially with longer sequences. In the figure below, the alignment data is presented in a way that indicates sequence identity (base is not shown) and only the altered bases (which would be mismatches or absent vertical lines) are shown here (below). Sequence alignment B (below) is a different alignment than Sequence alignment A.

Sequence alignment B

Query 708 CGTTCCAGTGAGTTCACCCTCTAAATCACCACGATCAAAAGGAACAAGCATCAAGCACGC 767

**Sbjct** 707 .**A**........................................**G**................. 766

Query 768 AGCAATGCAGCTCAAAACGCTTAGCCTAGCCACACCCCCACGGGAAACAGCAGTGATTAA 827

**Sbjct** 767 .**A**.......................................................**A**.**G** 826

**Question 2.** In Sequence alignment B, the same two 12S rDNA sequences were aligned, but a different region of the same gene is shown. The differences are highlighted in the second (sbjct) sequence. What might you infer from this alignment?

The sequence alignments generated by BLAST provide the best alignment of the query sequence with sequences within the database. Scientists can use this large database to help to identify biological samples. The program DNA Barcoding for Life is focused on identifying biological diversity within ecosystems. Environmental DNA collection and analysis has been used to help identify species that exist in a location that may be difficult to physically spot or collect.

**Question 3.** While DNA isolation and a BLAST search is a powerful manner in which to identify a specimen, there can be accurate sequences that do not match anything in the database.

What are two possible explanations that you could use to explain the lack of a perfect match between your query sequence and the database?

**Critical thinking question.** An alignment between the human 12S rDNA gene sequence (used in questions 1 and 2) with another Genbank file (EU626452) (alignment not shown) also produces a 100% match in the region shown in alignment A. However, EU626452 is identified in the database as an uncultured bacterial clone from a wetlands ecosystem sequencing project. Why do you think that this Genbank sequence identification may be incorrect?

**In the next worksheet, you will perform a BLAST search to try and identify the authenticity of biological artifacts.**