**Yeti or not: Do they exist?**

**Teaching notes:**

‘Yeti or not’ introduces students to DNA barcoding, bioinformatics (BLAST, FASTA, Multiple Sequence Alignment) and interpretation of the data collected. The case uses data from existing Yeti artifacts as well as supplemental information to direct students through DNA sequence analysis and interpretation of the resulting data. Students are directed to collect information from the National Center for Biotechnology Information (http://[www.ncbi.nlm.nih.gov](http://www.ncbi.nlm.nih.gov)) regarding specific DNA sequences deposited from the research. Students will compare some of the sequences (perform BLAST alignments) and gather information regarding the similarity and differences between the sequences. In the last (assessment) worksheet, three sequences are provided - including one that is not a match to sequences in the database - for students to identify using the bioinformatics approaches in the earlier worksheets.

**NIBLSE Core Competencies Addressed**

* C1 - Explain the role of computation and data mining in addressing hypothesis-driven and hypothesis-generating questions within the life sciences.
* C2 - Summarize key computational concepts, such as algorithms and relational databases, and their applications in the life sciences.
* C9 - Interpret the ethical, legal, medical and social implications of biological data.

**CourseSource Bioinformatics Learning Framework**

**Computation in the life sciences**

* + What is the role of computation in hypothesis-driven discovery processes within the life sciences?
	+ What computational concepts are important in bioinformatics?

**DNA - information storage (genomics)**

* + Where are data about the genome found (e.g., nucleotide sequence, epigenomics) and how are they stored and accessed?
	+ How can bioinformatics tools be employed to analyze genetic information?

**Ecology and evolution (metagenomics)**

* + How can bioinformatics tools be employed to examine ecological niches?

**Computational skill**s

* + How do biologists employ software development as part of the scientific discovery process?
	+ What higher-level computational skills can be used in bioinformatics research?

**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). DNA barcoding has been used to catalog diverse species from local ecosystems to biodiversity. 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 rRNA from the mitochondrial genome or 16S rRNA from the nuclear genome (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). Other genes that have been used in DNA barcoding are shown in Table 1.

Table 1. Some genes targeted for DNA barcoding.

|  |  |
| --- | --- |
| **Organism group** | **Marker gene/locus** |
| Animals | nuclear - 16S rRNAmitochondrial - cytochrome oxidase subunit I (COI), 12S rRNA |
| Plants | chloroplast - ribulose-1,5-bisphosphate carboxylase/oxygenase large subunit (rbcL) |
| Bacteria | 16S rRNA, RNA polymerase ß subunit (rpoB) |

These worksheets utilize the cryptid 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 a yet unknown DNA sequence (not matching any previously submitted database sequence).

These assignments have been used in an introductory biology course after having introduced the basics of DNA replication, PCR and sequencing. An online homework resource was used 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 third worksheet, creates alignments that are used to answer the subsequent questions. A fourth worksheet is included as a summative assessment of students abilities to perform bioinformatics.

In an optional review assignment (background supplement), 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 (unrelated to any sequences in the second assignment). It is important that the students be able to see this in color.

Learning objectives

Students will be able to

* Describe how the authenticity of cryptozoology specimens can be probed using modern molecular biology techniques
* Relate the molecular genetic techniques of DNA sequencing and polymerase chain reaction to prokaryotic DNA replication (review worksheet)
* Compare and contrast the molecular mechanisms of prokaryotic DNA replication, PCR and sequencing
* Determine how DNA sequence information is determined from electropherograms (review worksheet)
* Explain DNA barcoding and the requirements for conserved and variable sequences (worksheet 1)
* Explain why mitochondrial DNA loci are frequently used to type organisms
* Interpret sequence alignment results BLAST and multiple sequence alignments (MSA) (worksheet 2)
* Explain the results of a sequence alignment and validate the mismatches within an alignment (worksheets 2 and 3)
* Summative assessment of bioinformatics skills presented (worksheet 4)

Prerequisite knowledge

Students should be taught the basics of DNA replication, sequencing and PCR prior to being introduced to this case study. Similarities and differences between DNA polymerases used in prokaryotic DNA replication, (Sanger) DNA sequencing and PCR are made. A review assignment has been provided as a supplement to cover this material.

Worksheet 1 - students are introduced to BLAST searches (Altschul *et al*., 1990) and led through interpretation of results.

Worksheet 2 - students are introduced to FASTA format, perform a BLAST search and collect data from GenBank files identified in the BLAST search. The search involves six sequences from Sykes *et al*., (2014). Additional sequences are available in the supplemental files. Sykes *et al*., (2014) reported 30 artifact DNA sequences.

Worksheet 3 - students are introduced to multiple sequence alignment and perform an MSA on the sequences analyzed in worksheet 2. Students are introduced to using ClustalOmega (Goujon *et al*., 2010; Siever *et al.,* 2011).

Worksheet 4 - summative assessment of the bioinformatics skills presented in the worksheets. 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.

**Recommended delivery**

It is recommended that there is a review of basic sequencing technologies prior to beginning this assignment. Supplemental documents are provided to address related background topics. Additional background information is provided in the next section.

The first worksheet can be done in class with a group of students with a short period for group discussion followed by a classroom discussion.

The second and third worksheets should be provided as take-home materials as this will require the use of the internet (freely available programs). Since some of the questions ask for alignments, providing the document as a Word (or other text) document facilitates the students copying and pasting of results from the browser into the document for recording and analysis. Additional FASTA sequences may be found in the Yeti supplemental analysis document. These sequences represent all of the sequences from Sykes *et al.,* 2014. The specific sequences used in Worksheet 2 and 3 are indicated and can be replaced with additional sequences from this document.

The fourth worksheet is designed as a summative assessment and can be used to determine if the students have learned how to perform basic bioinformatics analysis and interpret the results. One of the sequences has been manipulated to not match the current database sequences, which should lead students to suggest that this might represent a new sequence and be molecular evidence for a new species.

**Additional background information:**

A comparison between DNA replication and sequencing DNA polymerases would include:

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.

Differences between DNA replication and sequencing might include:

|  |  |  |
| --- | --- | --- |
|  | **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 |

Students are encouraged to review DNA replication, sequencing and PCR in their assigned textbook or using [OpenStax Biology 2e](https://openstax.org/books/biology-2e/pages/14-2-dna-structure-and-sequencing.). A movie overview of DNA sequencing can be found [here](https://www.youtube.com/watch?v=wdS3j0TgbjM). *Both links are provided in the first assignment.*

**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.