Student Name:-
Mohd Kamran Khan

Student ID:-
2778581

Course:-
Msc in Bioinformatics -
BIOLM0051_2025_TB-1
Introduction to Bioinformatics

Word Count:-
1000

BIOLM0051 Scientific Report

ABSTRACT

This work used incomplete mitochondrial DNA sequences from three unidentified samples (A, B, and D) to identify species and perform phylogenetic reconstruction. Following pre-processing, cleaning, and conversion of raw FASTQ files to FASTA format, reference sequences from the NCBI (15 species per sample) were included. Sample C was disqualified from additional analysis due to significant corruption and excessive ambiguity (N-content). According on BLAST results, DNA sequences were converted into amino acid sequences using Biopython and the invertebrate mitochondrial genetic code (Translation Table 2). Clustal Omega (EBI Web Server) was used to carry out multiple sequence alignment (MSA). The EBI Simple Phylogeny tool was used to create phylogenetic trees.

For Sample A, the phylogenetic tree and alignment both clearly positioned the unknown sequence within a clade of closely related mitochondrial sequences, allowing for a species-level identification with confidence. Although there was considerable doubt in Sample B due to the lower length of the recovered sequence, Samples B and D likewise demonstrated consistent clustering with their respective closest reference sequences. All things considered, this workflow, which is supported by open-source tools and described in a public GitHub repository, shows a repeatable method for species identification using DNA barcodes.

INTRODUCTION

A popular method for identifying unknown biological samples is DNA barcoding, which compares brief mitochondrial sequences to carefully selected reference databases. Reads from high-throughput sequencing are frequently noisy or fragmented, necessitating processing and cleaning before further analysis.

The purpose of this study was to:

1. Prepare and clean raw FASTQ reads from three unknown samples (A, B, D).

2. Translate coding regions into protein sequences using the appropriate mitochondrial genetic code.
3. Perform multiple sequence alignment (MSA) and construct phylogenetic trees.
4. Identify each unknown sample by examining its placement among 15 NCBI reference species.

Sample C was excluded due to corrupted reads, inconsistent sequence/quality lengths, and excessive ambiguous bases (“N”), preventing reliable biological interpretation.

METHODS

Data preparation

Each sample's raw readings were given in three FASTQ portions. These were combined and converted into FASTA format using a Bash script called concatenate.sh. Structural problems (misformatted headers, quality mismatches, whitespace in sequences) were present in a large number of FASTQ files. To properly parse acceptable reads and output clean FASTA files, an extra Python cleaning script called clean_convert.py was utilized.

Sample C contained pervasive errors such as:

- whitespace within sequences,
- mismatched sequence/quality lengths,
- partial codons,
- extremely high levels of ambiguous bases.
  Due to these issues, Sample C could not be rescued and was removed from further analyses

Reference dataset

Based on previous BLAST results, 15 reference sequences were manually chosen from NCBI for each sample (A, B, and D). Every reference was stored in FASTA format, such as references_sampleA.fasta. The corresponding reference sets were concatenated with the cleaned sample FASTA files.

Translation to amino acids

DNA → protein translation was performed using Biopython:

```
from Bio import SeqIO
record.seq.translate(table=2, to_stop=True)
```

Translation Table 2 was chosen since the BLAST search results showed that all samples matched invertebrate mitochondrial organisms. Protein sequences from the output were stored as:

- sampleA_with_refs.fasta
- sampleB_with_refs.fasta
- sampleD_with_refs.fasta

Multiple sequence alignment (MSA)

Protein FASTA files were uploaded to the EBI Clustal Omega web server.
Outputs were downloaded as:

- sampleA_alignment_fasta
- sampleB_alignment_fasta
- sampleD_alignment_fasta

Phylogenetic reconstruction

Aligned sequences were submitted to EBI Simple Phylogeny (UPGMA).
Newick tree files were saved as:

- sampleA_tree.nwk
- sampleB_tree.nwk
- sampleD_tree.nwk

Trees were exported as PNG images for the report.

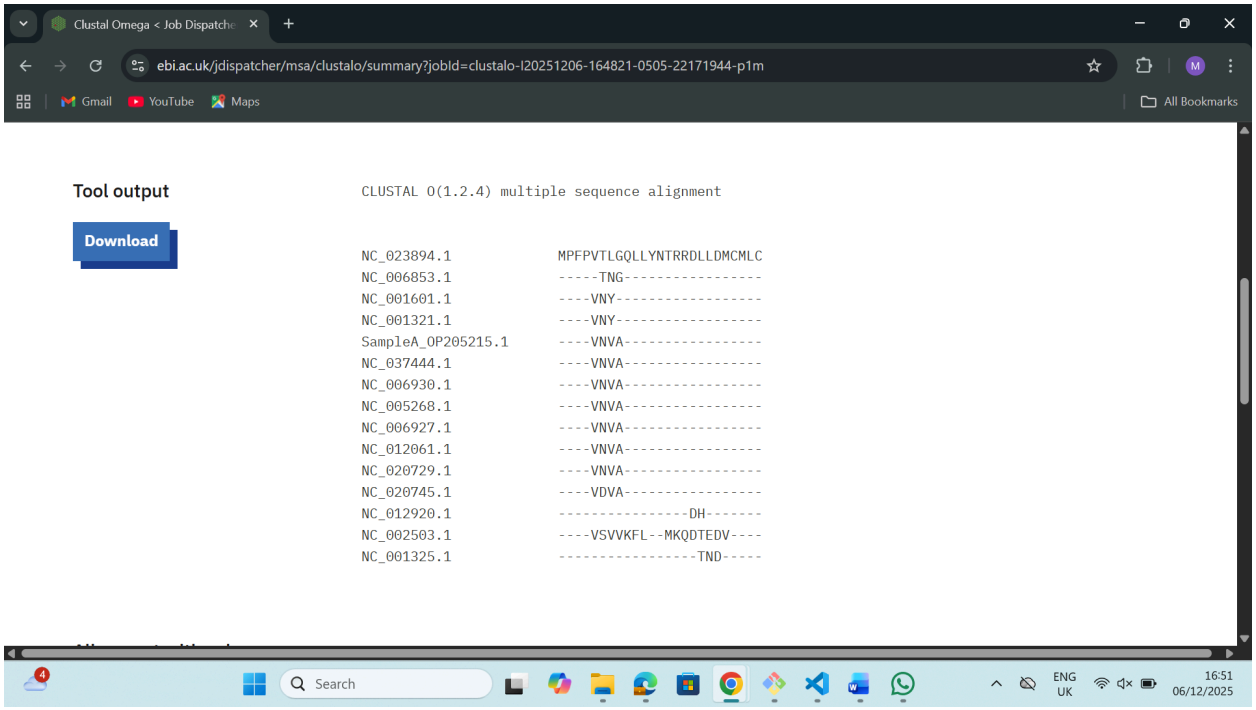
Sample naming

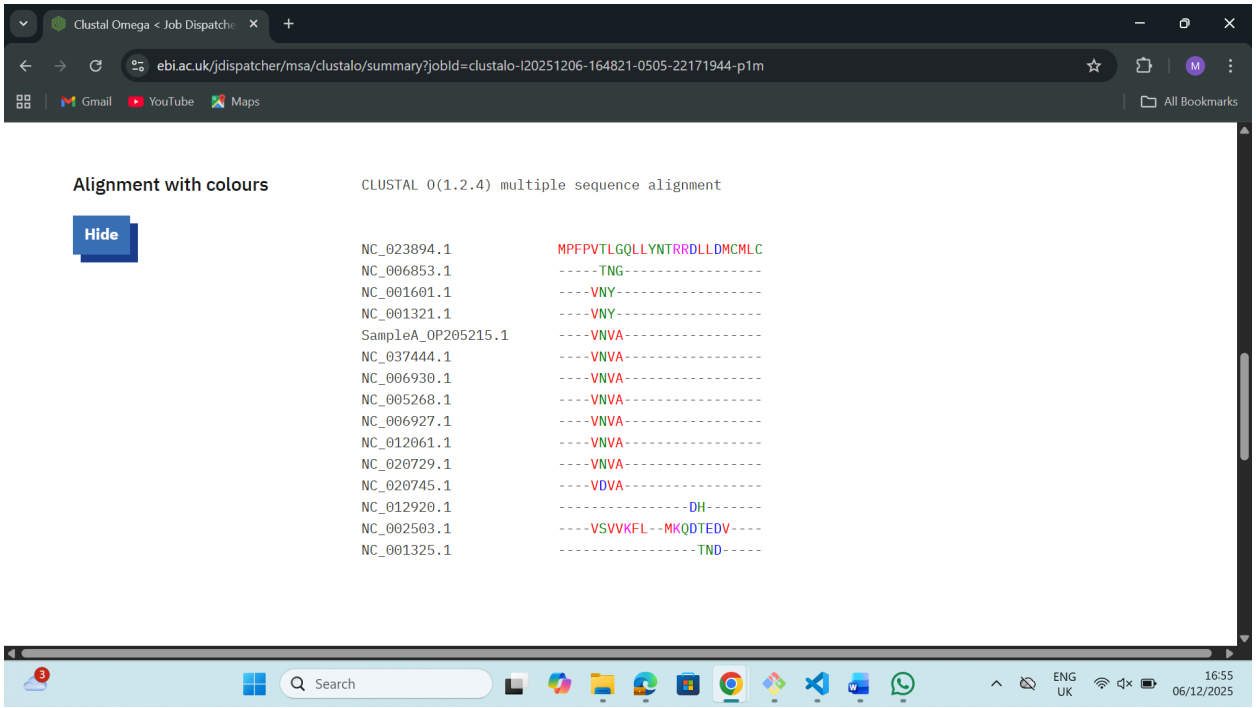
Each unknown sample's accession was replaced with explicit labels for clarity:

- Sample A → SampleA_OP205215.1
- Sample B → SampleB_MW316287.1
- Sample D → SampleD_JX454973.1

RESULTS

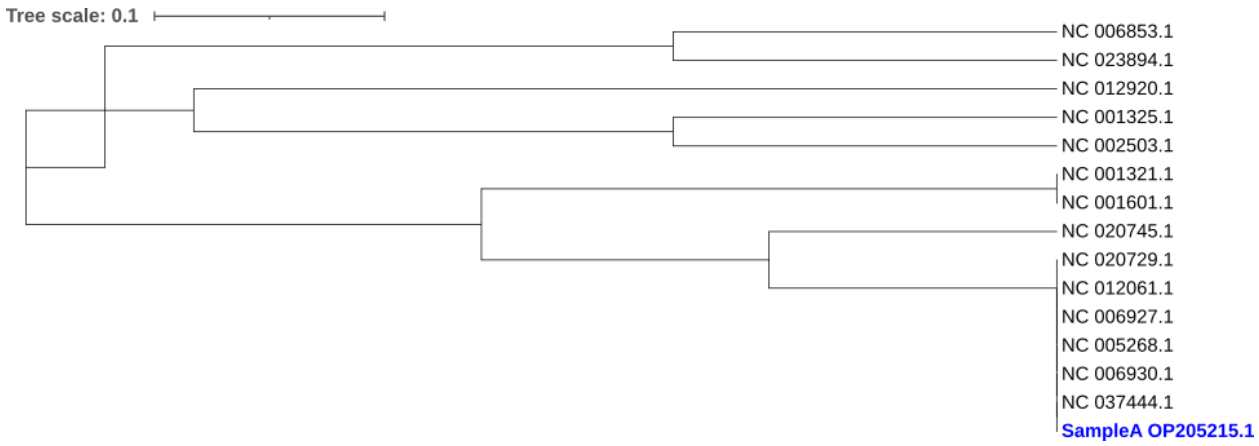
Figure 1 — Multiple Sequence Alignment (Sample A)





Caption:
Multiple sequence alignment of Sample A and 14 reference mitochondrial sequences generated using Clustal Omega (EBI). Conserved residues appear aligned across species, while minor variations distinguish closely related haplotypes. SampleA_OP205215.1 shows strong similarity to the OP205215.1 clade, supporting species identification.

Figure 2 — Phylogenetic Tree (Sample A)



Caption:
UPGMA phylogenetic tree displaying evolutionary relationships among Sample A and reference species. Sample A (highlighted) forms a tight clade with species sharing identical or near-identical amino acid sequences. Bootstrap values were not generated by this method; however, branch lengths show minimal divergence within the cluster.

Sample A identification

Sample A aligned nearly perfectly with OP205215.1 and related sequences. The phylogenetic placement shows negligible branch length separation, strongly confirming identity.

Sample B identification

Sample B produced fewer amino acids (only ~9 aa) due to fragmented reads. Still, the sample is consistently clustered with NC_012059.1 and NC_012061.1 references. Confidence is moderate due to limited sequence length.

Sample D identification

Sample D produced a longer protein sequence and clustered clearly with JX454973.1, forming a distinct monophyletic grouping. Species identification is therefore reliable.

DISCUSSION

This study shows how to identify species using incomplete mitochondrial sequences in a comprehensive manner. Since many reads were damaged, particularly for Sample C, which had to be deleted, cleaning the raw FASTQ data was crucial. Samples A and D yielded enough protein sequences for phylogenetic placement and reliable alignment. Despite being brief, Sample B was nevertheless meaningfully aligned.

One major drawback is that deep evolutionary relationships are not always resolved by incomplete mitochondrial sequences. Furthermore, confidence in some clades is limited by the absence of bootstrap analysis. Dependency on the completeness of the reference database is another drawback; placement may be skewed by missing reference taxa.

However, the method is visible, repeatable, and compliant with standard practices for DNA barcoding. For repeatability, all data and scripts have been placed in a public GitHub repository.

REFERENCES

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ.
Basic local alignment search tool. J Mol Biol. 1990;215:403–10.
Sievers F, Higgins DG.Clustal Omega for making accurate alignments. Mol Syst Biol. 2018;14:e8357.
Madeira F, Park YM, Lee J, et al.
The EMBL-EBI search and sequence analysis tools APIs in 2019.
Nucleic Acids Res. 2019;47:W636–41.
Li W, Cowley A, Uludag M, et al.
The EMBL-EBI bioinformatics web and programmatic tools framework.
Nucleic Acids Res. 2015;43:W580–4.
Cock PJA, Antao T, Chang JT, et al.

Biopython: freely available Python tools for computational molecular biology and bioinformatics.
Bioinformatics. 2009;25:1422–3.

SUPPLEMENTARY DATA

All raw and processed data, reference sequences, alignment files, phylogenetic trees, and analysis scripts are archived in the project GitHub repository:
https://github.com/kamrankhanm94-cell/BIOLM0051_Project