

Experiment 1: Exploring DNA Sequence Databases, followed by Retrieving and Analyzing DNA Sequences

AIM

To explore major DNA sequence databases (GenBank, EMBL, NCBI), retrieve DNA sequences in FASTA format, and perform basic sequence analysis, including determination of sequence length, GC content, nucleotide composition, and reverse complementation.

Theory

The foundation of modern bioinformatics rests on publicly accessible repositories of nucleotide sequences. In 1982, the **National Center for Biotechnology Information (NCBI)** introduced **GenBank** as the first large-scale open archive of annotated DNA sequences. Almost at the same time, the **European Molecular Biology Laboratory (EMBL)** launched its sequence archive, and Japan's National Institute of Genetics established the **DNA Data Bank of Japan (DDBJ)**. These three resources joined forces to form the **International Nucleotide Sequence Database Collaboration (INSDC)**, ensuring that every sequence submitted to any one of them is shared daily among all three. This collaboration guarantees that researchers worldwide have access to the same up-to-date data, formatted according to a common standard that includes unique accession numbers and detailed feature annotations.

Although **GenBank** is commonly mentioned alongside “**NCBI**,” it is not synonymous with the institute itself. NCBI is a division of the U.S. National Library of Medicine charged with developing and maintaining a suite of bioinformatics tools—such as PubMed, BLAST, and Taxonomy—from which **GenBank, the nucleotide archive**, is one flagship resource. Every entry in GenBank begins with a header line that specifies the accession number, organism name, and a brief description. Subsequent lines contain the nucleotide sequence. This format, known as **FASTA**, was developed in the mid-1980s to facilitate human readability and computational parsing. Its simplicity and universal acceptance have made FASTA the de facto standard for exchanging sequence data between databases, analysis software, and laboratories.

Each FASTA record/files consists of:

Header Components:

- Accession number (unique identifier)
- Organism name (species designation)
- Gene/sequence description
- Additional metadata (length, type, etc.)

Sequence Lines: Raw nucleotide/amino acid sequence

FASTA File extensions: file.fa, file.fasta, file.fsa

Example of FASTA format:

```
>NC_000006.12:151654148-152129619 Homo sapiens chromosome 6, GRCh38.p13 Primary Assembly
TATTGATTTTGTGTAACATGTGTTGTATATATCTATAACGAGAACTCAAGTCATACTGTAATCCTAT
TTTGAAACTGACTTTTTTCTTTATCAGTATATCAAGATTATTTCCACATCATTTGACATTTTTTCT
ACAGTGTAATTTAATGGCTACATTGTTTCTATCTATGAATATATCAAACCTATTTCTAAAAACCTA
CTCAGGGATTTTAAAAAATAAAACGATGTTTTAATATTATAAAGATTCAAGTATATTTCTTATACG
TACACATTTCTAAGGTTTGAGTCTTACAAGATGCTGAAGTACTGAGTCTGCTCATCTGTCAC
ATAGGGAAAAATTATAGAAGGAAACATCAAGATTGGAAAAATCTGTGAGAATTGTTTGCATTAGTGT
GTAGGTGTGTGTTGGGGTGGTGCAGCTTGGGGCAGAGGCCTCAGGTGTGGCTGTGGAGTGATCA
GATAGAGTTTTTGGAGTTTGGCTTTTGGCCAGGACACTTGGTGCCTGCCCCAGAGCTGCAGCCAGAA
GGCGTTCTCAGAGGTGAAGTCCAGGCAGTGAGGAGCTGTCTGCCAGTAGGCAGTTGAAGAAAAAATG
AGCTAGAGGAAAAAACAACAAAGAACTCTCTTCTAATGTGCCAGGCTGCCGGGAGCTGGAATGA
AGCACTGACAGGAGTGGGTATTTTATGTTGAAGGGAATAATCAACTGGTTTTTTTGGTACCCAAGACTTT
CCACCTTCACACACACATGAGATGCTTTGAAATAAAGATAGTCACTTGACTTAGTAAAGTTTGTGAC
ATAAAAAATAGAGAAATACCAAAGAATACAAAAAGGAAACCTTCGTTAATATTATTCAGACTTAAATTC
CAGATTGTATCAACATTAAGGGGTTGATGAAAACATGGGAGAAAGCAAGGGACGTGAGATCGGGCTCA
ATTCTTGACTTGCTGGGGGAAGGTATCAACACAGAACTTTAAGAATTAGAAGGCATTAAAAAGAAATAG
AAATCCTGAATCAAATTGAAACAGTAAAAATAAATAGTCCAAAGATGTGTAATATATCACTATCACAAT
```

Once retrieved, a sequence's basic properties—length, base composition, and GC content—provide immediate biological insights. The **total number of nucleotides** influences experimental designs such as primer selection for PCR and considerations for cloning strategies. **Base composition**, defined by the proportion of adenine, thymine, guanine, and cytosine, can reveal organism- or region-specific biases that affect codon usage and genomic stability. **GC content**, in particular, reflects the fraction of guanine plus cytosine bases; because G–C pairs form three hydrogen bonds versus two for A–T pairs, DNA regions with high GC content exhibit greater thermal stability. In eukaryotic genomes, GC-rich stretches often coincide with gene-dense areas and regulatory elements.

Additionally, reverse complementation is a critical preprocessing step for many molecular biology applications. It involves generating the reverse complement of a DNA strand by first replacing each base with its complement ($A \leftrightarrow T$, $G \leftrightarrow C$) and then reversing the resulting sequence. This operation yields the sequence of the opposite strand in the 5'→3' orientation. Reverse complements are essential for designing primers, probes, and for *in silico* analyses that require examination of both strands of a genomic region.

Taken together, the DNA sequence databases and the FASTA standard file format provide a reliable, globally synchronized framework for storing and sharing raw sequence data. By learning the retrieval and basic analysis of DNA sequences, from understanding how databases operate, how to retrieve sequences in desirable formats, to interpreting length, composition, and reverse complementation, one acquires essential skills that underpin advanced bioinformatics applications.

DATABASE LINKS & RESOURCES

Primary Databases

Database	URL	Description
NCBI GenBank	https://www.ncbi.nlm.nih.gov/genbank/	Genetic sequence database
NCBI Nucleotide	https://www.ncbi.nlm.nih.gov/nucleotide/	Collection of sequences
ENA (EMBL)	https://www.ebi.ac.uk/ena/browser/home	European nucleotide database
DDBJ	https://www.ddbj.nig.ac.jp/index-e.html	Japanese nucleotide database

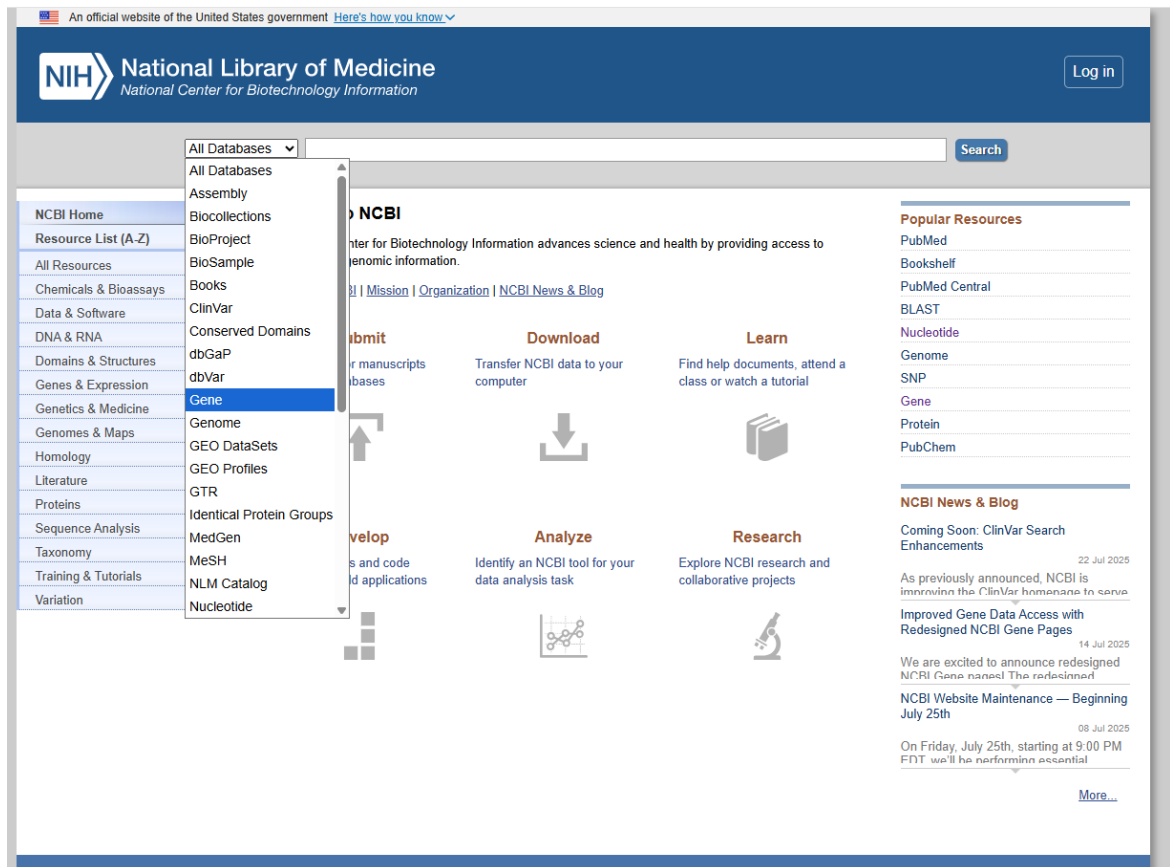
Search and Analysis Tools

Tool	URL
NCBI Global Search	https://www.ncbi.nlm.nih.gov/search/
GC Content Calculator	https://jamiemcgowan.ie/bioinf/gc_content.html
GC Content Plot	https://jamiemcgowan.ie/bioinf/gc_content_plot.html
Nucleotide Composition	https://jamiemcgowan.ie/bioinf/gc_content.html
Reverse Complement	https://jamiemcgowan.ie/bioinf/complement.html

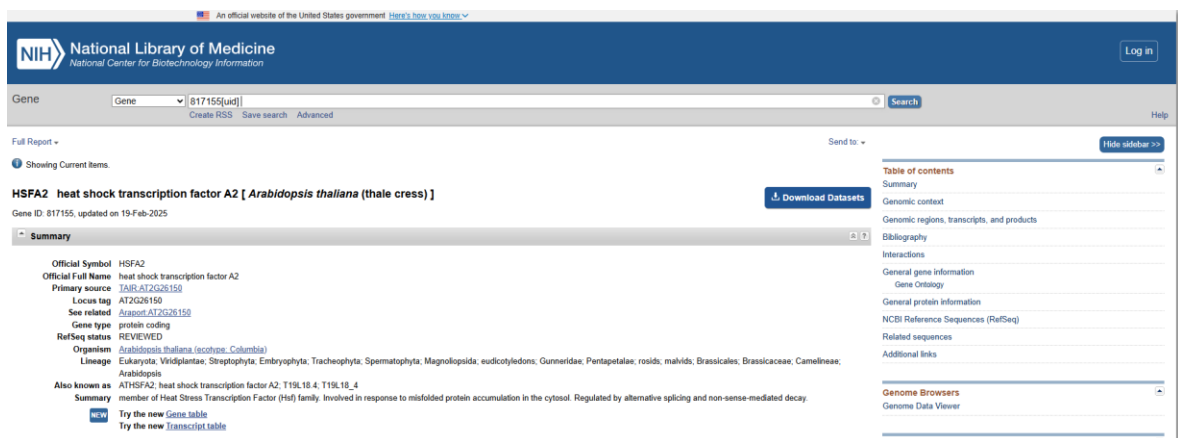
TASKS AND STEPS

1. Database Navigation

1. Open NCBI home: <https://www.ncbi.nlm.nih.gov/>
2. In the “Database” drop-down (upper left), select **Gene**.



3. In the search box, enter **817155** and click **Search**.



B. mRNA Sequence

1. Scroll to the NCBI Reference Sequences (RefSeq) section; find **NM_001124916.2** (Arabidopsis thaliana heat shock transcription factor A2 (HSFA2), mRNA).

The screenshot shows the NCBI Reference Sequences (RefSeq) page. At the top, there's a header "NCBI Reference Sequences (RefSeq)" with a "NEW Try the new Transcript table" link. Below this is a "Genome Annotation" section. The main content area is titled "Reference assembly" and contains two sections: "Genomic" and "mRNA and Protein(s)". The "mRNA and Protein(s)" section lists the entry "1. NM_001124916.2 → NP_001118388.1 heat shock transcription factor A2 [Arabidopsis thaliana]". It includes links for "See identical proteins and their annotated locations for NP_001118388.1", "Status: REVIEWED", "UniProtKB/TrEMBL: A0A178YQK8, B3H5P6", and "Conserved Domains (1) summary". The conserved domain is "cd12113 HSF_DNA-bind; HSF-type DNA-binding" with a location of "49 → 79".

2. Click FASTA next to **NM_001124916.2**.

The screenshot shows the NCBI GenBank page for NM_001124916.2. The header includes the NIH logo and "National Library of Medicine National Center for Biotechnology Information". There's a "Log in" button. The main content area is titled "Arabidopsis thaliana heat shock transcription factor A2 (HSFA2), mRNA". It includes links for "FASTA" and "Graphics". The "FASTA" link is highlighted. Below this is a "Go to" link. The "FASTA" section shows the sequence: "NM_001124916 1876 bp mRNA linear PLN 20-OCT-2022". The "DEFINITION" is "Arabidopsis thaliana heat shock transcription factor A2 (HSFA2), mRNA". The "ACCESSION" is "NM_001124916". The "VERSION" is "NM_001124916.2". The "DBLINK" is "BioProject: PRJNA116" and "BioSample: SAMN03081427". The "KEYWORDS" is "RefSeq". The "SOURCE" is "Arabidopsis thaliana (thale cress)". The "ORGANISM" is "Arabidopsis thaliana". The "GenBank" section shows the sequence: "NM_001124916.2 1876 bp mRNA linear PLN 20-OCT-2022". The "DEFINITION" is "Arabidopsis thaliana heat shock transcription factor A2 (HSFA2), mRNA". The "ACCESSION" is "NM_001124916". The "VERSION" is "NM_001124916.2". The "DBLINK" is "BioProject: PRJNA116" and "BioSample: SAMN03081427". The "KEYWORDS" is "RefSeq". The "SOURCE" is "Arabidopsis thaliana (thale cress)". The "ORGANISM" is "Arabidopsis thaliana". The "GenBank" section shows the sequence: "NM_001124916.2 1876 bp mRNA linear PLN 20-OCT-2022". The "DEFINITION" is "Arabidopsis thaliana heat shock transcription factor A2 (HSFA2), mRNA". The "ACCESSION" is "NM_001124916". The "VERSION" is "NM_001124916.2". The "DBLINK" is "BioProject: PRJNA116" and "BioSample: SAMN03081427". The "KEYWORDS" is "RefSeq". The "SOURCE" is "Arabidopsis thaliana (thale cress)". The "ORGANISM" is "Arabidopsis thaliana".

3. Click on "Send to" and then select "File". *Make sure that you select the correct file format (FASTA)
4. Save the sequence as At_HSFA2_mRNA.fasta.

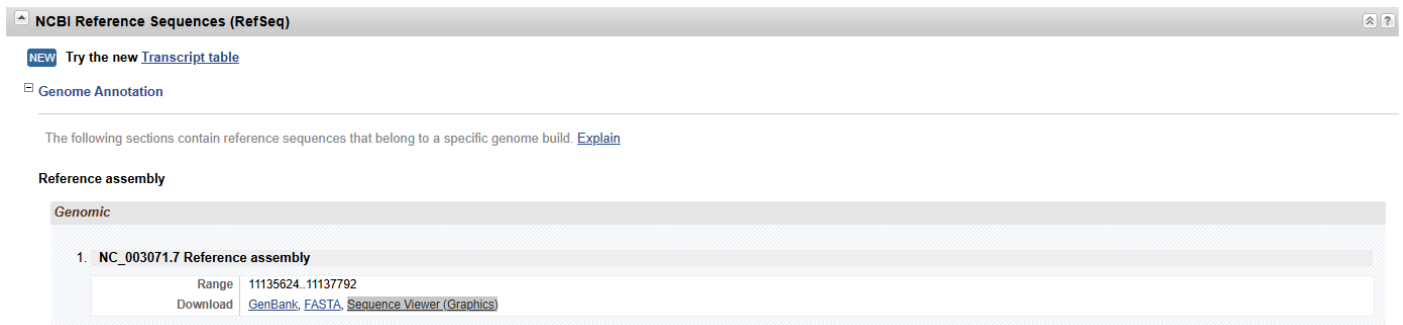
3. Sequence Record Analysis & File Management

1. Examine the GenBank record (for both Gene Sequence & Mrna Sequence):
 - Accession number
 - Organism name
 - Sequence length
 - Definition line
2. Record observations
3. Create directory UBT518_Omics_Lab_Expt01/.
4. Place At_HSFA2_gene.fasta and At_HSFA2_Mrna.fasta inside.

4. Visualizing the Sequence

A. Gene Sequence:

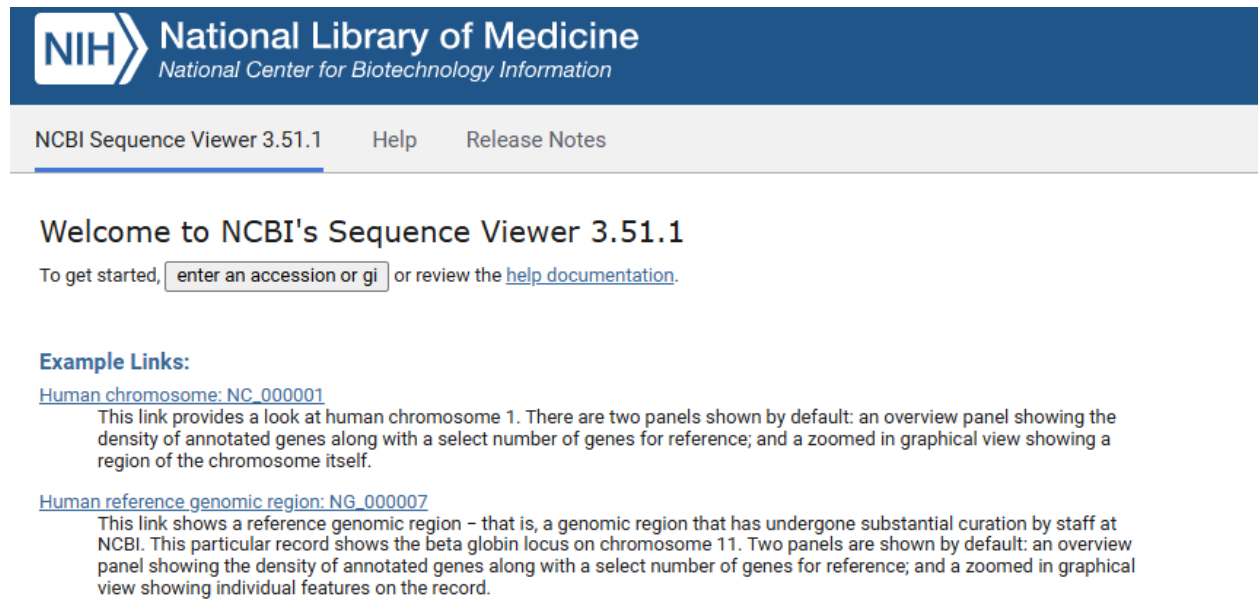
1. In the same NCBI Reference Sequences (RefSeq) section; there is a section Genomic.
2. Within that section click on “Sequence Viewer (Graphics)”



3. Screenshot the view and save as “HSFA2_viewer_gene.png” in the folder “UBT518_Omics_Lab_Expt01”.
4. Explore the sequence view to identify different features.

B. mRNA Sequence:

1. Open NCBI Sequence Viewer <https://www.ncbi.nlm.nih.gov/projects/sviewer/>



NCBI Sequence Viewer 3.51.1 Help Release Notes

Welcome to NCBI's Sequence Viewer 3.51.1

To get started, or review the [help documentation](#).

Example Links:

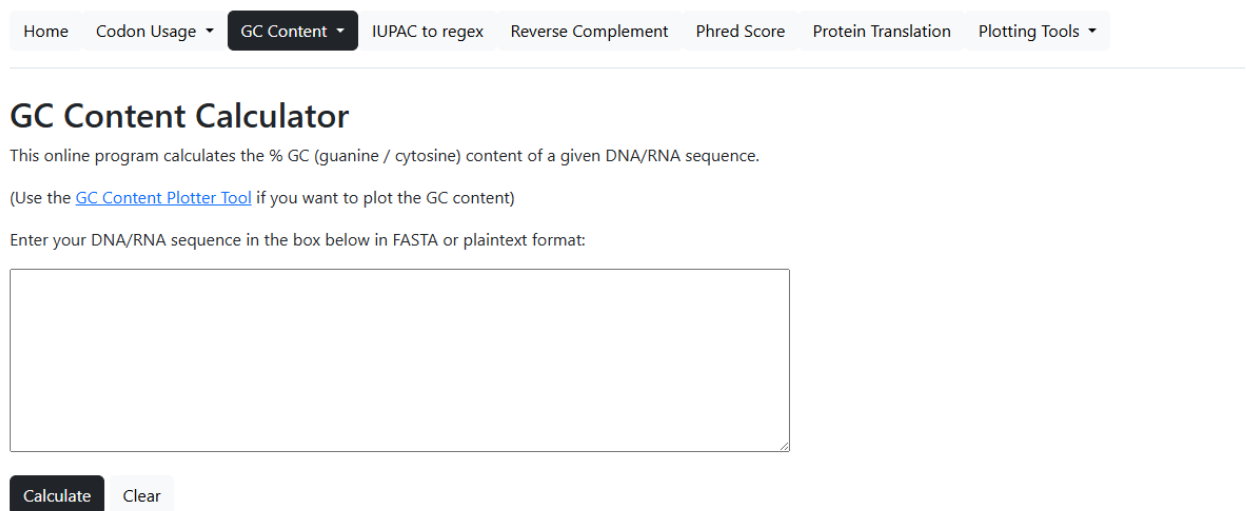
[Human chromosome: NC_000001](#)
This link provides a look at human chromosome 1. There are two panels shown by default: an overview panel showing the density of annotated genes along with a select number of genes for reference; and a zoomed in graphical view showing a region of the chromosome itself.

[Human reference genomic region: NG_000007](#)
This link shows a reference genomic region – that is, a genomic region that has undergone substantial curation by staff at NCBI. This particular record shows the beta globin locus on chromosome 11. Two panels are shown by default: an overview panel showing the density of annotated genes along with a select number of genes for reference; and a zoomed in graphical view showing individual features on the record.

2. Click on “enter and accession or gi”
3. For the Mrna sequence, enter **NM_001124916.2** and click on **OK**
3. Screenshot the view and save as “HSFA2_viewer_mRNA.png” in the folder “UBT518_Omics_Lab_Expt01”.
4. Explore the sequence view to identify different features.

5. Sequence Analysis — Length, Base Composition and GC Content %

1. Open https://jamiemcgowan.ie/bioinf/gc_content.html



Home Codon Usage **GC Content** IUPAC to regex Reverse Complement Phred Score Protein Translation Plotting Tools

GC Content Calculator

This online program calculates the % GC (guanine / cytosine) content of a given DNA/RNA sequence.

(Use the [GC Content Plotter Tool](#) if you want to plot the GC content)

Enter your DNA/RNA sequence in the box below in FASTA or plaintext format:

Calculate Clear

2. Open the mRNA fasta file, copy and paste the mRNA sequence in the box, and click “**Calculate**”.
3. Note length, GC content%, counts, and percentages of A, U (T), G, C.
4. Repeat this for the gene sequence as well.

6. Sequence Analysis — GC Content Plot

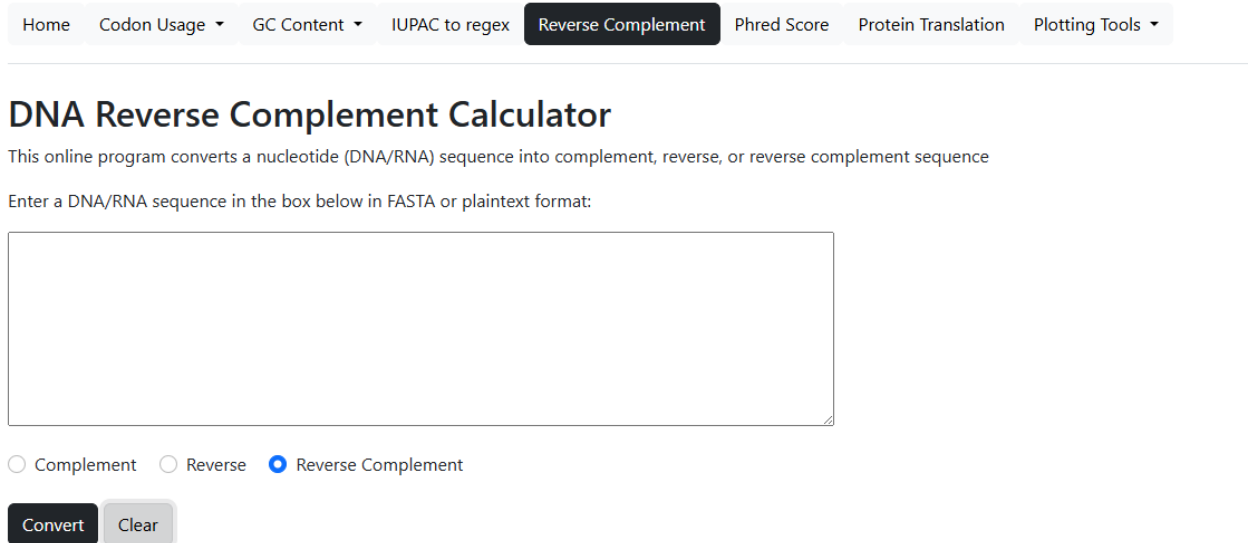
1. Open https://jamiemcgowan.ie/bioinf/gc_content.html and from the GC Content drop down menu, select “GC Content Plot”.

The screenshot shows the 'GC Content' dropdown menu open, with 'GC Content Plot' selected. The page title is 'GC Content Plot'. Below the title, it says 'This online program plots the GC content of a DNA/RNA sequence.' and '(Use the [GC Content Calculator](#) if you just want to calculate the overall % GC content)'. There is a text input field for the sequence, currently containing 'ATGC...'. Below the input field are two more input fields: 'Window size (bp):' with a value of '1000' and 'Step size (bp):' with a value of '500'. A note below these fields says '(Increase the step size to improve performance. Decrease the step size to show more data points)'. There is a checkbox labeled 'Show mean GC content (horizontal black line)' which is checked. At the bottom, there are two buttons: 'Plot' and 'Clear'.

2. Open the mRNA fasta file, copy and paste the mRNA sequence in the box.
3. Set the “Window Size (bp)” to 100 and “Step Size (bp)” to 50.
4. Click on “**Plot**”.
5. Hover your cursor on the plot, some icons will appear on your right. Click on “Download plot as a png”.
6. Save them with proper file naming in the folder “UBT518_Omics_Lab_Expt01”
7. Repeat this for the gene sequence as well.

7. Sequence Analysis —Reverse Complement

1. Open <https://jamiemcgowan.ie/bioinf/complement.html> “DNA Reverse Complement Calculator”



2. Open the mRNA fasta file, copy and paste the mRNA sequence in the box.
3. Select “Reverse Complement” and then click “**Convert**”
4. Save the Reverse Complement sequence as a FASTA file (Take a moment to think!)
5. Save the fasta file as “At_HSFA2_mRNA_revcomp.fasta” in the folder “UBT518_Omics_Lab_Expt01”.
6. Repeat this for the gene sequence as well.

DELIVERABLES

Files, tables, and figures

1. Content of FASTA Files to be copied in the reports:

- At_HSFA2_gene.fasta
- At_HSFA2_mRNA.fasta
- At_HSFA2_mRNA_revcomp.fasta
- At_HSFA2_gene_revcomp.fasta

2. Sequence Record Summary Table (including for both gene and mRNA):

Sequence	Accession	Length (bp)	A (%)	T/U (%)	G (%)	C (%)	GC (%)	Definition Line
----------	-----------	-------------	-------	---------	-------	-------	--------	-----------------

3. GC Content Plots (PNG images):

- HSFA2_mRNA_GC_plot.png
- HSFA2_gene_GC_plot.png

4. Sequence Viewer Screenshots (PNG images):

- HSFA2_viewer_gene.png
- HSFA2_viewer_mRNA.png
-

Report Document (PDF or DOCX) containing:

- Aim and theory concise summary
- Methodology steps (Database Navigation through Sequence Analysis)
- Sequence Record Summary Table with filled values
- Screenshots of FASTA retrieval, Sequence Viewer, GC content plots
- Discussion addressing differences between gene vs. mRNA sequences, GC content implications, and utility of reverse complements