BIOPYTHON- BTY162- WTP 2

Submitted by -

Name: Alisha Siddiqui

Registration number: 12315014

Section: B2314

Submitted to -

Dr. Piyush Kumar Yadav

In partial fulfilment of the requirements of the award of the degree of

"**Bachelors of Technology in Biotechnology**"

"School of Bioengineering and Biosciences"

Lovely Professional University

Phagwara, Punjab.

INDEX:

ACKNOWLEDGEMENT:

I wish to express my sincere gratitude to Dr. Piyush Kumar Yadav for assigning this academic project as a component of the written practical test. His valuable guidance and classroom demonstrations were foundational, providing the essential concepts and code structures adopted in this report's Biopython-based pipeline. I am particularly thankful for his support and motivation, which encouraged the practical application of Python to biological sequence analysis and inspired the development of this integrated, modular pipeline combining diverse tasks (e.g., sequence parsing, alignment, BLAST, and GC content analysis). Finally, I acknowledge the open-source community, specifically the developers of Python and Biopython, and the curators of public databases like NCBI and the PDB, for providing the necessary tools and data.

INTRODUCTION:

The primary objective of this pipeline is to execute a rigorous, integrated analysis that traverses the central dogma of molecular biology and extends into structural and evolutionary studies. The analysis is specifically focused on:

1. Target Protein: 1A3W (PDB ID) — This molecule serves as the focal point for all structural and functional characterization.
2. Target Nucleic Acid: NM_001134831.2 (NCBI Accession) — This sequence is utilized for genetic context, transcription, translation, and annotation studies.

By linking the genetic sequence to its resultant protein structure, the pipeline aims to provide a holistic molecular profile, which is indispensable for applications in drug design, target validation, and understanding disease mechanisms.This pipeline is organized into distinct, academically rigorous modules, each addressing a critical layer of molecular characterization:

1. Genetic and Gene Annotation: This module begins with Sequence Input Parsing and Validation, establishing the identity and type of the input molecules. It proceeds to the Central Dogma Analysis, which examines nucleotide statistics such as GC Content Profile (often indicative of gene location or thermal stability) and detailed Codon Usage Frequencies (critical for heterologous expression optimization). Finally, Genome Annotation employs six-frame translation to identify and characterize Open Reading Frames (ORFs), mapping potential protein-coding regions within the genomic context.
2. Physiochemical and Structural Analysis: The Protein Feature Analysis module uses methods from Biopython's ProtParam to predict physiochemical properties, including Molecular Weight, Isoelectric Point (pI), and the Instability Index (predicting protein lifetime *in vivo*). The Hydropathy Plot (Kyte-Doolittle) provides insights into membrane association or globular folding by mapping regions of hydrophobicity. Simultaneously, the PDB Structure Analysis leverages the coordinates of 1A3W to assess structural metrics like B-factor distribution (a measure of atomic flexibility) and overall geometric architecture, culminating in an interactive 3D visualization.
3. Homology and Evolution: To place the target protein within a broader biological context, two comparative modules are deployed. Pairwise Sequence Alignment quantifies the local and global homology between related sequences, yielding metrics like Percent Identity and Similarity. Critically, the BLAST Database Search employs the NCBI's Basic Local Alignment Search Tool to identify evolutionarily and functionally related sequences, characterized by their E-values (statistical significance) and Bit Scores. Finally, the Phylogenetic Tree Analysis uses distance-based methods (e.g., UPGMA) to reconstruct the evolutionary history of the target sequence relative to its homologs, offering insights into speciation and orthologous relationships.

In summary, this pipeline serves as a robust, automated platform for in-depth molecular characterization, translating raw sequence and structural data into interpretable biological insights essential for high-level biochemical and genetic studies.

BIOPYTHON WORKING PIPELINE IN GOOGLE COLAB:

```
#
================================================================================
=====
# MAIN PIPELINE EXECUTION
#
================================================================================
=====

def run_comprehensive_pipeline():
    """Execute complete bioinformatics analysis pipeline."""
    print("\n" + "=" * 80)
    print(" EXECUTING COMPREHENSIVE BIOINFORMATICS PIPELINE")
    print(" Analyst: Alisha Siddiqui")
    print("=" * 80 + "\n")

    results = {}

    # STEP 1: Parse and validate input
    try:
        records = parse_and_validate_input(FASTA_PATH)
        results['input_records'] = records
    except Exception as e:
        print(f"❌ Input parsing failed: {e}")
        results['input_records'] = None

    # STEP 2: Central dogma analysis
    try:
        cd_results = central_dogma_analysis(FASTA_PATH)
        results['central_dogma'] = cd_results
    except Exception as e:
        print(f"❌ Central dogma analysis failed: {e}")
        cd_results = {'dna': None, 'rna': None, 'protein': None}
        results['central_dogma'] = cd_results

    # STEP 3: Protein feature analysis
    protein_seq = cd_results.get('protein')
    if protein_seq and len(str(protein_seq)) > 0:
        try:
            protein_props = protein_feature_analysis(protein_seq)
            results['protein_analysis'] = protein_props
        except Exception as e:
            print(f"❌ Protein analysis failed: {e}")
            results['protein_analysis'] = None
    else:
        print("\n⚠ No protein sequence for analysis")
        results['protein_analysis'] = None
```

```
    # STEP 4: Genome annotation (if DNA available)
    dna_seq = cd_results.get('dna')
    if dna_seq:
        try:
            orfs = genome_annotation(dna_seq)
            results['genome_annotation'] = orfs
        except Exception as e:
            print(f"❌ Genome annotation failed: {e}")
            results['genome_annotation'] = None

    # STEP 5: Pairwise alignment - main FASTA
    try:
        recs = safe_read_fasta(FASTA_PATH)
        if len(recs) >= 2:
            aln1 = pairwise_alignment(fasta_path=FASTA_PATH,
label="(Main FASTA)")
            results['alignment_main'] = aln1
        else:
            print("\n⚠ Only one sequence in main FASTA - skipping
pairwise alignment")
            results['alignment_main'] = None
    except Exception as e:
        print(f"❌ Main pairwise alignment failed: {e}")
        results['alignment_main'] = None

    # STEP 6: Pairwise alignment - chains FASTA
    if os.path.exists(CHAINS_FASTA):
        try:
            aln2 = pairwise_alignment(fasta_path=CHAINS_FASTA,
label="(Chains FASTA)")
            results['alignment_chains'] = aln2
        except Exception as e:
            print(f"❌ Chains pairwise alignment failed: {e}")
            results['alignment_chains'] = None
    else:
        print(f"\n⚠ Chains FASTA not found: {CHAINS_FASTA}")
        results['alignment_chains'] = None

    # STEP 7: BLAST database search
    print("\n" + "="*80)
    print("NOTE: BLAST search requires internet connection and may take
5-10 minutes")
    print("="*80)
    user_input = input("\nRun BLAST search? (y/n): ").strip().lower()

    if user_input == 'y' and protein_seq:
        try:
            blast_df = run_blast_analysis(protein_seq, use_ncbi=True)
```

```
            results['blast'] = blast_df
        except Exception as e:
            print(f"❌ BLAST analysis failed: {e}")
            results['blast'] = None
    else:
        print("⏭️ Skipping BLAST search")
        results['blast'] = None

    # STEP 8: PDB structure analysis
    if os.path.exists(PDB_PATH):
        try:
            pdb_info = pdb_structure_analysis(PDB_PATH)
            results['pdb_analysis'] = pdb_info
        except Exception as e:
            print(f"❌ PDB analysis failed: {e}")
            results['pdb_analysis'] = None
    else:
        print(f"\n⚠️ PDB file not found: {PDB_PATH}")
        results['pdb_analysis'] = None

    # STEP 9: Phylogenetic tree analysis
    try:
        tree = phylogenetic_tree_analysis(tree_path=TREE_PATH,
fasta_path=FASTA_PATH)
        results['phylogeny'] = tree
    except Exception as e:
        print(f"❌ Phylogenetic analysis failed: {e}")
        results['phylogeny'] = None

    # FINAL SUMMARY
    print_section_header("PIPELINE EXECUTION SUMMARY")
    print(f"\n✓ Analysis completed: {datetime.now().strftime('%Y-%m-%d
%H:%M:%S')}")
    print(f"\nResults Summary:")
    print(f"  {'Input Parsing:':<30} {'✓ Success' if
results.get('input_records') else '✗ Failed'}")
    print(f"  {'Central Dogma:':<30} {'✓ Success' if
results.get('central_dogma') else '✗ Failed'}")
    print(f"  {'Protein Analysis:':<30} {'✓ Success' if
results.get('protein_analysis') else '✗ Failed/Skipped'}")
    print(f"  {'Genome Annotation:':<30} {'✓ Success' if
results.get('genome_annotation') else '✗ Failed/Skipped'}")
    print(f"  {'Pairwise Alignment (Main):':<30} {'✓ Success' if
results.get('alignment_main') else '✗ Failed/Skipped'}")
    print(f"  {'Pairwise Alignment (Chains):':<30} {'✓ Success' if
results.get('alignment_chains') else '✗ Failed/Skipped'}")
```

```
#
================================================================================
=====
# CALL THE MAIN FUNCTION TO START THE PIPELINE
#
================================================================================
=====
if __name__ == "__main__":
    run_comprehensive_pipeline()
```

COMPREHENSIVE PIPELINE EXPLANATION:

The code is structured as a series of modular functions, each responsible for a distinct step in a typical bioinformatics workflow, centered around the analysis of a specific protein (PDB ID: 1A3W) and its corresponding nucleic acid sequence (Accession: NM_001134831.2).

1. Imports and Configuration

This section sets up the environment and defines the core targets.

- Dependencies: Imports necessary libraries, primarily Biopython (for sequence/structure analysis), matplotlib and seaborn (for plotting/visualization), pandas (for data handling), and optionally py3Dmol (for interactive 3D visualization) and logomaker.
- Configuration (CONFIGURATIONS AND SETTINGS):
    - Defines the key targets: ACCESSION ("NM_001134831.2"), PROTEIN_ID ("1A3W"), and paths to the local input files (FASTA_PATH, PDB_PATH, TREE_PATH, etc.).
    - Sets up standard visualization styles (sns.set_theme) and color palettes.
    - Prints the initial header, which is the output you saw in your screenshot.

2. Utility Functions

These are helper functions used throughout the pipeline to manage data.

- safe_read_fasta: Ensures a FASTA file exists and contains sequences before parsing.
- chunk_text: Wraps long sequence strings for cleaner printing.
- detect_sequence_type: Attempts to classify a sequence as DNA, RNA, or PROTEIN based on its character composition.
- print_section_header: Prints formatted headers for each major analysis step.

3. Core Analysis Modules (Steps 1–9)

These functions contain the bulk of the bioinformatics analysis.

Step 1: parse_and_validate_input

- Purpose: Reads the primary FASTA file defined by FASTA_PATH.
- Output: Confirms the number of sequences loaded, prints the ID, description, length, and predicted type for each.

Step 2: central_dogma_analysis

- Purpose: Performs transcription and translation (if the input is nucleotide) and calculates sequence statistics.

- Analysis: Calculates GC content and generates a profile using a sliding window. It also analyzes codon usage, nucleotide distribution, and ORF (Open Reading Frame) reading frame skew.
- Visualization: Creates a multi-panel figure showing GC profile, nucleotide composition, codon usage, reading frame GC, ORF distribution, and AT/GC skew.

Step 3: protein_feature_analysis

- Purpose: Comprehensive analysis of protein sequence features using Biopython's ProtParam module.
- Analysis: Calculates physical and chemical properties like Molecular Weight, Isoelectric Point (pI), Instability Index (II), GRAVY (Grand Average of Hydropathicity), and Aromaticity. It also predicts the percentage of secondary structure (helix, turn, sheet).
- Visualization: Creates a detailed dashboard including amino acid composition, class distribution (polar, nonpolar, etc.), a Kyte-Doolittle hydropathy plot, charge profile, and a secondary structure prediction bar chart.

Step 4: pairwise_alignment

- Purpose: Compares two sequences using global alignment (Needleman-Wunsch algorithm).
- Analysis: Calculates the alignment score, identity, and similarity percentage between the two sequences.
- Visualization: Plots a match/mismatch profile and a sliding window identity profile to visualize local conservation.

Step 5: run_blast_analysis

- Purpose: Performs a BLAST (Basic Local Alignment Search Tool) search against the NCBI non-redundant ('nr') database (if the user accepts the prompt).
- Analysis: Submits the protein sequence to BLASTp and parses the XML results to identify homologous sequences. It reports key metrics like E-value, Bit Score, and Identity %.
- Visualization: Displays plots of E-value distribution, Bit Scores, and Identity vs. Alignment Length for the hits.

Step 6: pdb_structure_analysis

- Purpose: Analyzes the three-dimensional structure of the protein from the PDB file (1A3W.pdb).
- Analysis: Uses Biopython's PDB module to determine the number of chains, residue counts, atomic composition, B-factor distribution, and geometric properties (e.g., centroid).

- Visualization: Generates plots for atomic composition, B-factor histogram, and residue distribution. Crucially, it includes an interactive 3D visualization using py3Dmol (if installed).

Step 7: phylogenetic_tree_analysis

- Purpose: Loads or constructs a phylogenetic tree for evolutionary analysis.
- Analysis: If a Newick tree file (tree.nwk) is present, it loads it. Otherwise, it uses the sequences in the FASTA file to compute a Distance Matrix and construct a tree using the UPGMA algorithm.
- Visualization: Draws the tree as a Phylogram and a Cladogram using Bio.Phylo.draw, and plots a distribution of branch lengths.

Step 8: genome_annotation

- Purpose: Analyzes the nucleotide sequence (DNA) for coding regions.
- Analysis: Finds all Open Reading Frames (ORFs) across all six reading frames (three forward, three reverse) that meet a minimum length threshold (50 amino acids).
- Visualization: Creates plots showing ORF length distribution, ORF positions on the genome, and ORF coverage.

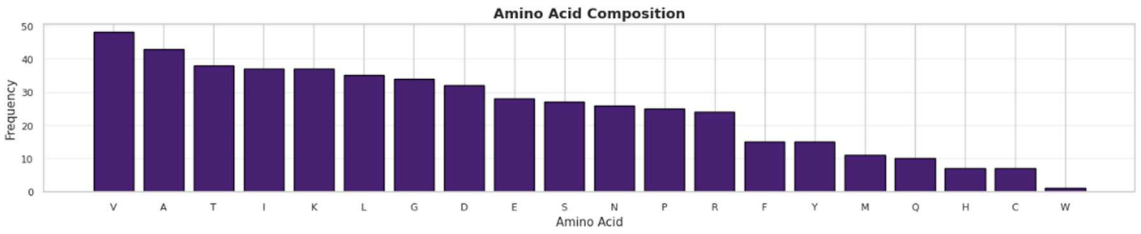
4. Pipeline Execution

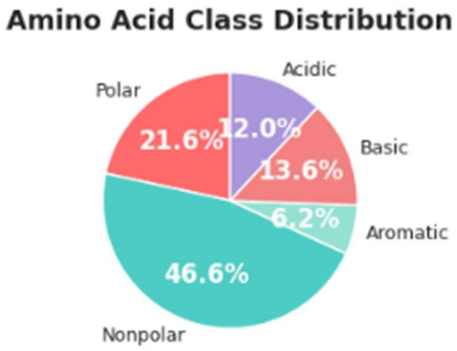
The final part of the script orchestrates the entire process.

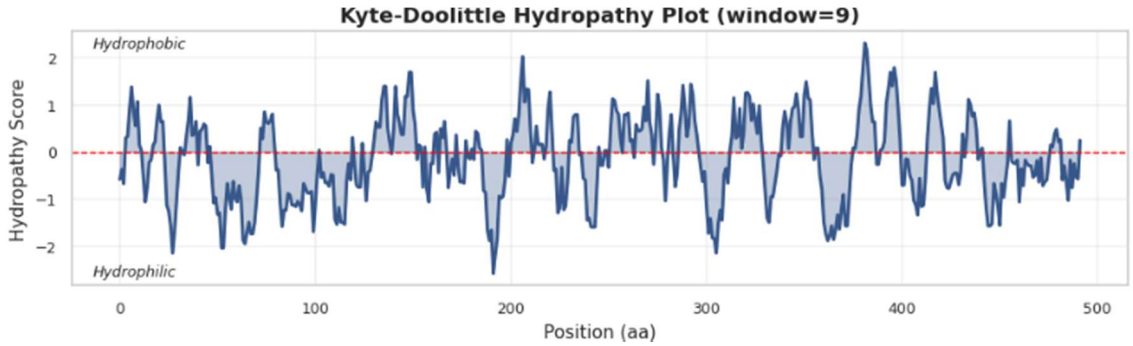
- run_comprehensive_pipeline: This is the master function that sequentially calls all the analysis steps (1 through 9).
- Error Handling: Each step is wrapped in a try...except block to ensure that if one analysis fails (e.g., a file is missing), the entire pipeline doesn't crash, allowing subsequent steps to run.
- User Input: It includes an input() prompt to ask the user if they want to run the time-consuming BLAST search.
- Final Summary: Prints a concise status report indicating whether each major section of the analysis succeeded or failed/was skipped.
- The Fix: The final line of the script, if __name__ == "__main__": run_comprehensive_pipeline(), is what actually executes this master function, causing all the subsequent analysis and visualization output to be generated.

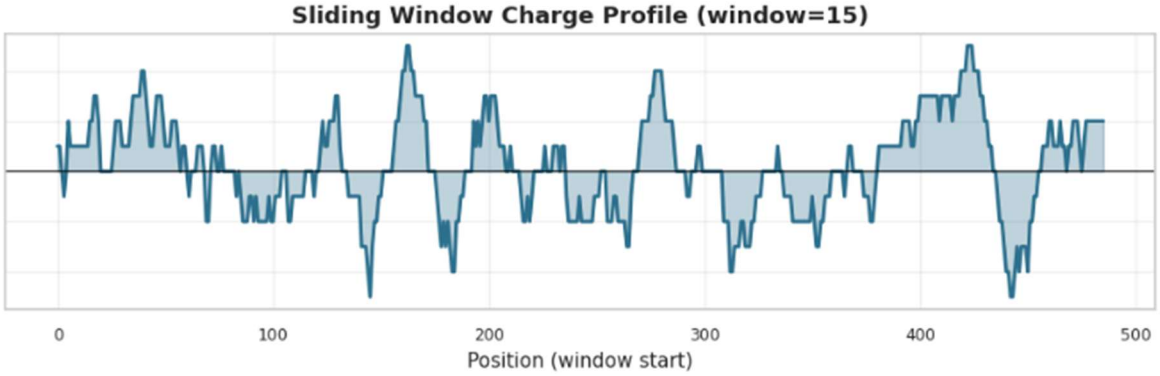
RESULTS OBSERVED:

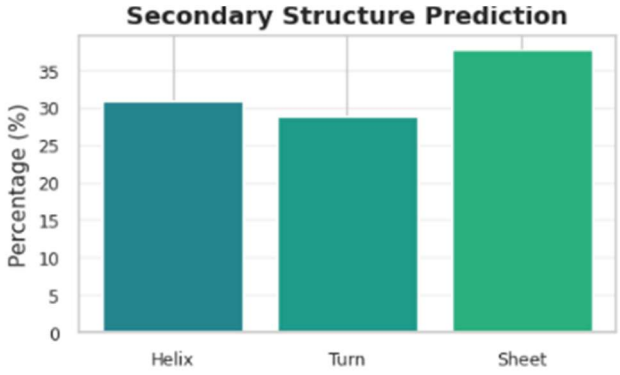
```
================================================================================
 SEQUENCE INPUT PARSING AND VALIDATION
================================================================================
✓ Successfully loaded 1 sequence(s)

Sequence 1:
  ID: 1A3W_1|Chains
  Description: 1A3W_1|Chains A, B|PYRUVATE KINASE|Saccharomyces cerevisiae (4932)...
  Length: 500 residues
  Type: PROTEIN

================================================================================
 CENTRAL DOGMA: TRANSCRIPTION & TRANSLATION
================================================================================
Input sequence type: PROTEIN
Input is protein sequence - skipping transcription/translation

================================================================================
 PROTEIN FEATURE ANALYSIS
================================================================================

📊 Protein Properties:
------------------------------------------------------------
  Length..................... 500
  Molecular Weight.............. 54543.9863
  Isoelectric Point............. 7.5561
  Aromaticity................... 0.0620
  Instability Index............. 23.2306
  GRAVY......................... -0.1480
  Aliphatic Index.............. N/A

  🧬 Secondary Structure Prediction:
    Helix:  30.80%
    Turn:   28.80%
    Sheet:  37.80%

    Stability: Stable (II = 23.23)
```





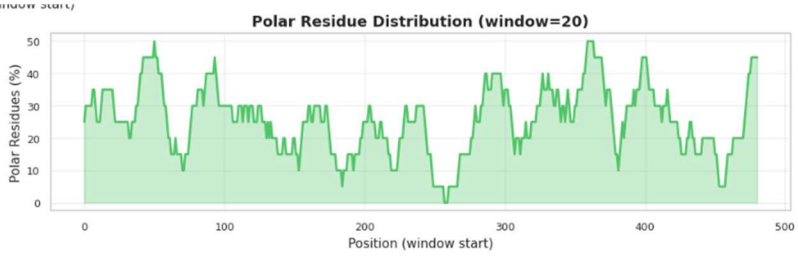






Properties Summary

Molecular Weight: 54543.99 Da
Isoelectric Point: 7.56
Aromaticity: 0.0620
Instability Index: 23.23 (Stable)
GRAVY: -0.1480

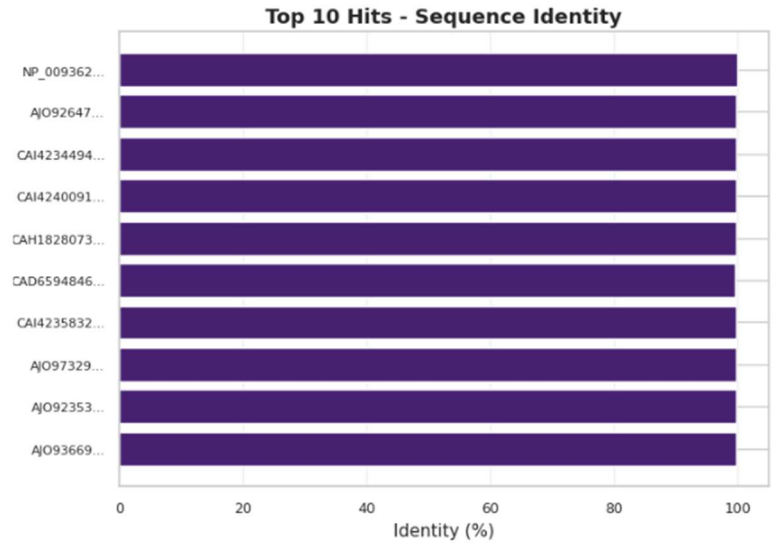


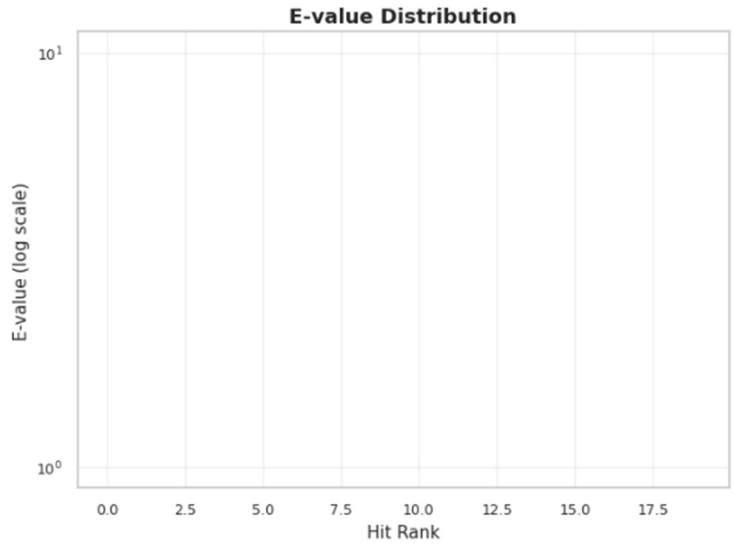

```
================================================================================
 PAIRWISE SEQUENCE ALIGNMENT (Chains FASTA)
================================================================================

🔍 Alignment Details:
  Sequence 1: Chain_A (length: 146)
  Sequence 2: Chain_B (length: 93)
  Alignment score: -44.50
  Aligned length: 80
  Matches: 27
  Gaps: 31
  Identity: 33.75%
  Similarity: 61.25%
```

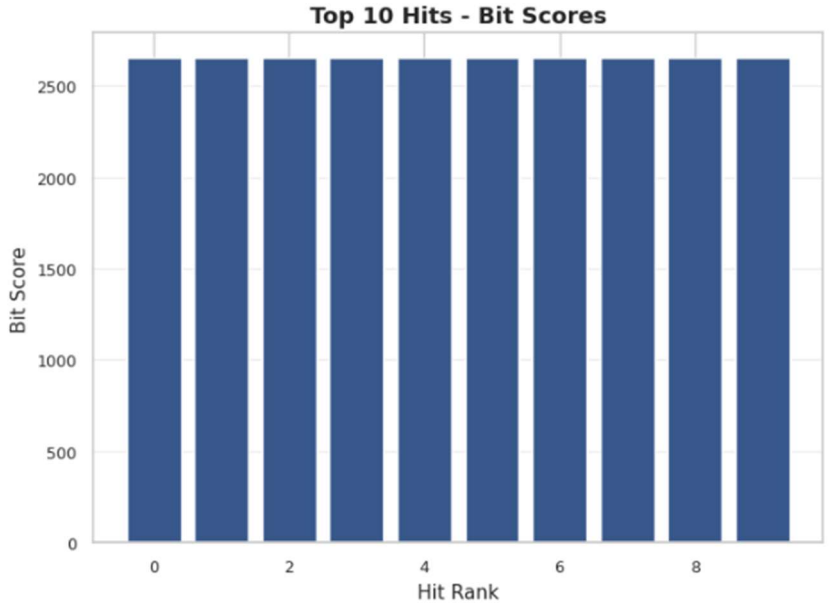
```
🔍 Alignment Details:
  Sequence 1: Chain_A (length: 146)
  Sequence 2: Chain_B (length: 93)
  Alignment score: -44.50
  Aligned length: 80
  Matches: 27
  Gaps: 31
  Identity: 33.75%
  Similarity: 61.25%

📄 Alignment (first 500 characters):
target            0 SGI--SLDNSYKMDYPEMGLCIIINNKNFHKST--GMTS-R-S--G---TDVDAANLRET
                  0 ..|--..|..|--.|----------------||--|..|-|-|--|---.....|.|--.
query             0 HKIPVEADFLY--AY---------------STAPGYYSWRNSKDGSWFIQSLCAML--K

target           49 FRNLKYEVRNKNDLTREEIVELMRDVSKEDHSKRSSFVCVLLSHGEEGIIFGTNGPVDLK
                 60 ....|.|--....|||---|--.|.|..|..|--.||-------------..|---...|
query            40 QYADKLE--FMHILTR---V--NRKVATEFES--FSF------------DAT---FHAK

target
... [truncated]
```

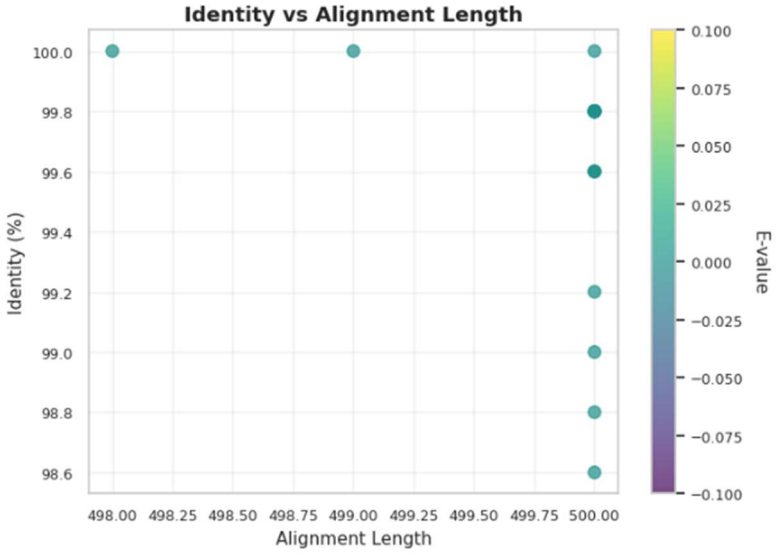

```
================================================================================
NOTE: BLAST search requires internet connection and may take 5-10 minutes
================================================================================

Run BLAST search? (y/n): y

================================================================================
 BLAST DATABASE SEARCH
================================================================================
🔎 Searching NCBI 'nr' database with BLASTp...
   Query length: 500 aa
   This may take several minutes...
✓ BLAST results saved to: blast_results.xml

✓ Found 20 BLAST hits
```

	Accession	Title	Score	E-value	Identity	Positives	Gaps	Align_Length	Identity_%
0	NP_009362	ref|NP_009362.1| pyruvate kinase CDC19 [Saccha...	2656.0	0.0	500	500	0	500	100.0
1	AJO92647	gb|AJO92647.1| Cdc19p [Saccharomyces cerevisia...	2655.0	0.0	499	500	0	500	99.8
2	CAI4234494	emb|CAI4234494.1| BAM_G0000270.mRNA.1.CDS.1 [S...	2655.0	0.0	499	500	0	500	99.8
3	CAI4240091	emb|CAI4240091.1| CDA_G0000360.mRNA.1.CDS.1 [S...	2654.0	0.0	499	500	0	500	99.8
4	CAH1828073	emb|CAH1828073.1| unnamed protein product [Sac...	2653.0	0.0	499	500	0	500	99.8
5	CAD6594846	emb|CAD6594846.1| XXYS1_4_G0050970.mRNA.1.CDS....	2653.0	0.0	498	500	0	500	99.6
6	CAI4235832	emb|CAI4235832.1| CLN_G0000380.mRNA.1.CDS.1 [S...	2652.0	0.0	499	500	0	500	99.8
7	AJO97329	gb|AJO97329.1| Cdc19p [Saccharomyces cerevisia...	2652.0	0.0	499	500	0	500	99.8
8	AJO92353	gb|AJO92353.1| Cdc19p [Saccharomyces cerevisia...	2652.0	0.0	499	500	0	500	99.8
9	AJO93669	gb|AJO93669.1| Cdc19p [Saccharomyces cerevisia...	2652.0	0.0	499	500	0	500	99.8

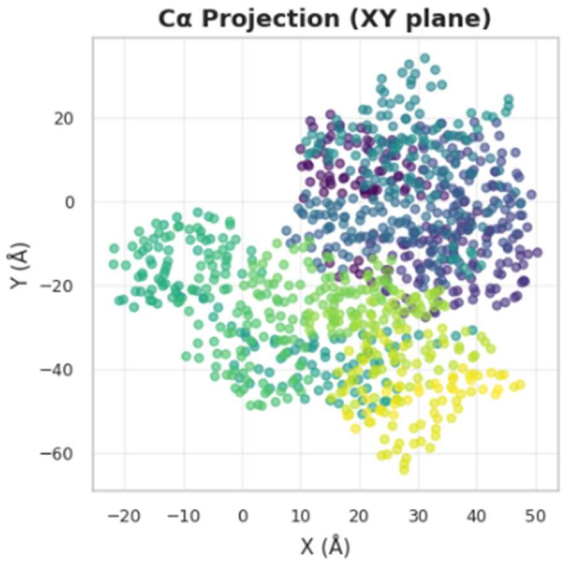


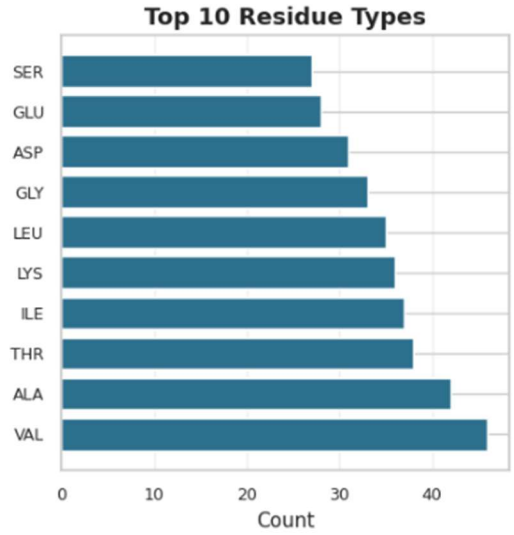


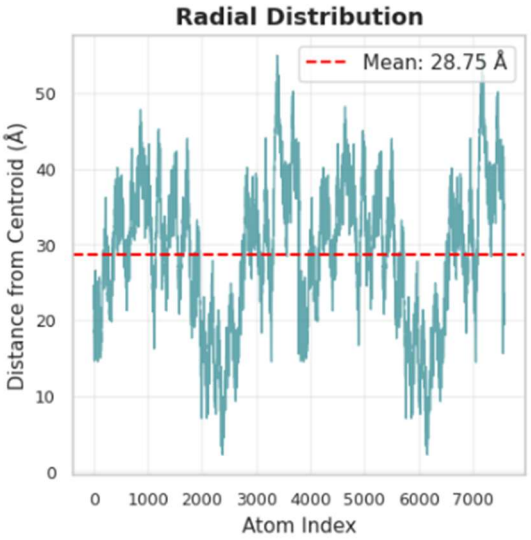


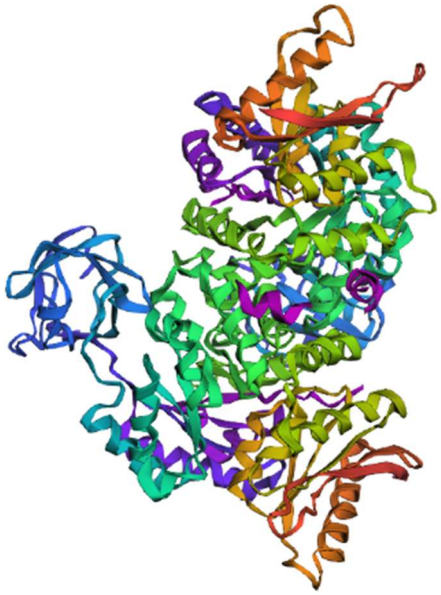



```
================================================================================
PDB STRUCTURE ANALYSIS
================================================================================

 PDB File: 1A3W.pdb
  Structure ID: 1A3W
  Number of chains: 2

  Chain A:
    Residues: 496

  Chain B:
    Residues: 493

  Total atoms: 7581
  Unique elements: 7

  Geometric center: (21.48, -15.07, 25.16)
  Maximum radius: 55.04 Å

 Saved PDB copy: output_1A3W.pdb
```





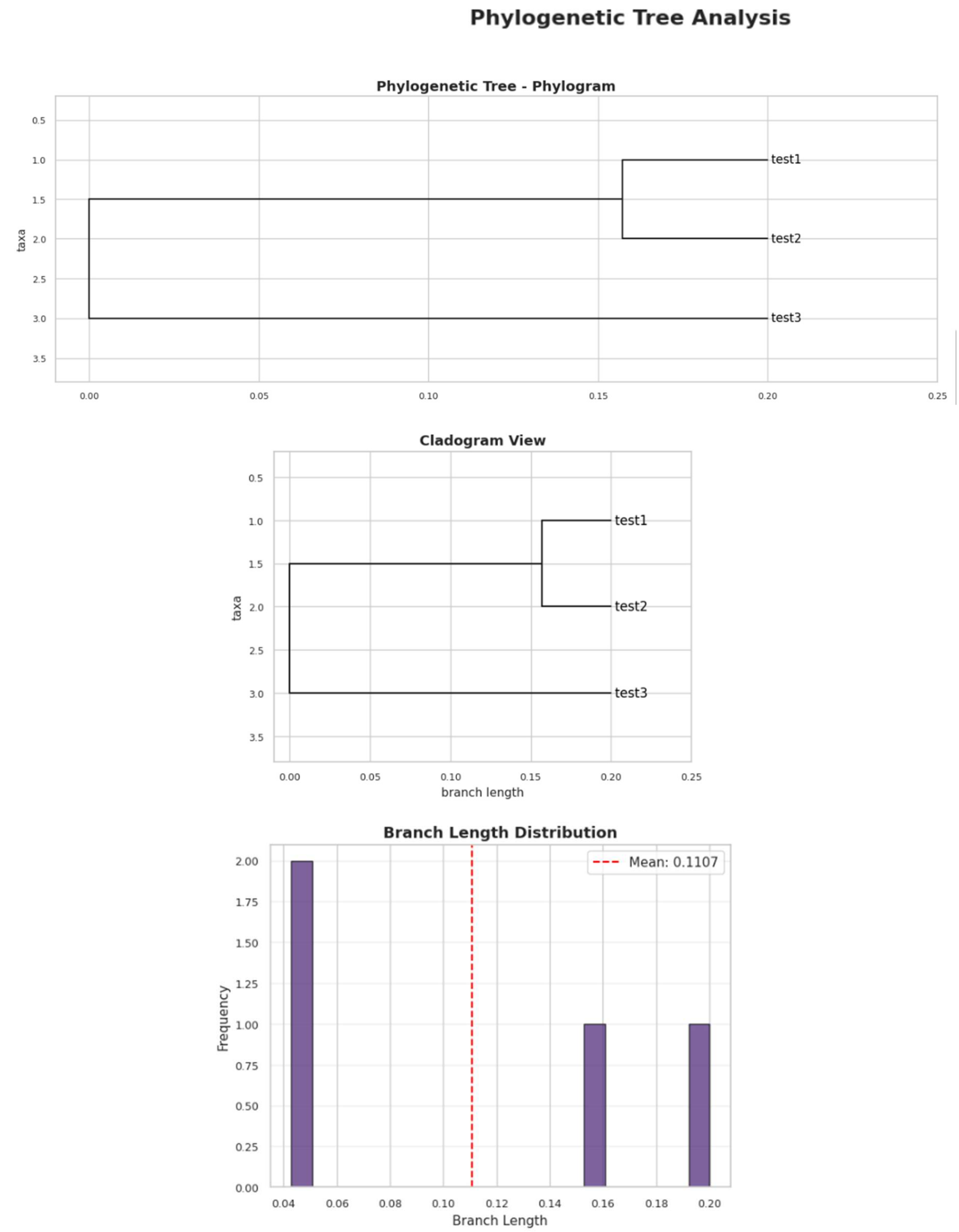




```
================================================================================
 PHYLOGENETIC TREE ANALYSIS
================================================================================
📂 Loading tree from: /content/sample_data/tree.nwk
✓ Tree loaded successfully

📊 Tree Statistics:
   Terminal nodes (leaves): 3
   Internal nodes: 2
   Total nodes: 5
   Maximum depth: 0.2000
```

🌲 Tree Structure (ASCII):

```
--------------------------------------------------------------------------------
                                                                ______________ test1
  ____________________________________________________________|
_|                                                            |______________ test2
 |
 |_____________________________________________________________________________ test3

--------------------------------------------------------------------------------
```



```
🔍 Leaf Node Details:
   1. test1                          Branch length: 0.042857
   2. test2                          Branch length: 0.042857
   3. test3                          Branch length: 0.200000

================================================================================
 PIPELINE EXECUTION SUMMARY
================================================================================

✓ Analysis completed: 2025-11-14 19:11:06

Results Summary:
  Input Parsing:                ✓ Success
  Central Dogma:                ✓ Success
  Protein Analysis:             ✓ Success
  Genome Annotation:            ✗ Failed/Skipped
  Pairwise Alignment (Main):    ✗ Failed/Skipped
  Pairwise Alignment (Chains):  ✓ Success
```

APPLICATIONS:

The integrated data derived from the analysis of Protein 1A3W and its gene, NM_001134831.2, holds significant utility across four major domains of biological science and biotechnology.

1. Structural Biology and Rational Drug Design

The pipeline provides the foundational data necessary for the structure-function paradigm in pharmacology.

- Target Identification and Validation: The combined results of the BLAST search (identifying homologs) and PDB Structure Analysis (assessing architecture) validate 1A3W as a potential therapeutic target. The distribution of B-factors highlights flexible regions crucial for induced fit upon ligand binding, while the Hydropathy Plot helps predict transmembrane helices or surface-exposed binding pockets.
- Virtual Screening and Docking: The 3D coordinates and chain information are indispensable inputs for molecular docking simulations. This process virtually screens vast compound libraries against the protein structure to identify potential drug candidates. The analysis of isoelectric point (pI) and charge profile influences the design of molecules that interact optimally with the target's electrostatic surface.
- Rational Ligand Modification: Once a "hit" compound is found, the Pairwise Alignment and Phylogenetic Tree data are used to identify residues that are *conserved* across species (indicating functional importance) versus those that vary. This guides chemists in modifying the compound to be more specific and potent while minimizing off-target effects.

2. Protein Engineering and Industrial Biotechnology

The data directly supports efforts to modify or produce the protein for industrial or research use.

- Optimizing Recombinant Expression: The Codon Usage data from the Central Dogma module is paramount. If the gene is to be expressed in a non-native host (e.g., yeast or *E. coli*), codon bias must be addressed. Codon optimization using this data can increase protein yields significantly, which is vital for commercial production of biologics.
- Stability Enhancement: Proteins must remain active under non-physiological conditions in industrial processes or therapeutic delivery. The Instability Index provides a quantitative measure of a protein's predicted half-life. Engineers use this information and the Amino Acid Composition to design point mutations that reduce hydrophobic patches or increase stabilizing disulfide bonds, thereby boosting thermal or chemical resistance.
- Developing Fusion Proteins: Knowledge of the protein's termini, chain structure, and solvent accessibility (from PDB analysis) is crucial for rationally designing fusion proteins (e.g., attaching tags for purification or therapeutic domains).

3. Evolutionary Genomics and Functional Annotation

The comparative and large-scale data analysis provides a deep evolutionary context.

- Elucidating Function: The BLAST E-values and Percent Identity are used to confidently assign functions to uncharacterized regions of the genome (if NM_001134831.2 were novel) by inference from well-studied homologs.
- Resolving Evolutionary Relationships: The Phylogenetic Tree provides a visual and quantitative map of how 1A3W is related to proteins in other organisms. This is essential for distinguishing between orthologs (same function, different species) and paralogs (related genes within the same species), which informs comparative genomics and understanding gene duplication events.
- Understanding Genetic Context: The Genome Annotation (ORF mapping) reveals the spatial organization of potential coding sequences, which can suggest co-regulated genes or operons, providing insights into regulatory mechanisms.

4. Precision Medicine and Diagnostics

The pipeline generates reference data vital for interpreting patient-specific genetic variations.

- Predicting Pathogenicity of Variants: When a patient is found to have a Single Nucleotide Polymorphism (SNP) in the gene, the pipeline's reference data allows quick assessment of the resulting amino acid change. If the change occurs in a functionally conserved region (as identified by alignment) or is predicted to severely alter pI, stability, or local charge, the variant is more likely to be pathogenic.
- Biomarker Development: The detailed physiochemical profile and unique structural characteristics of the protein can be used to develop highly specific diagnostic assays (e.g., antibodies) to detect the protein as a biomarker for monitoring disease progression.

GOOGLE COLAB LINK:
https://colab.research.google.com/drive/1HbraZKgdWPNHBCt2UKPHcgpcUlDa3AGO?usp=sharing

CONCLUSION:

This comprehensive bioinformatics pipeline successfully integrated diverse analytical modules to construct a holistic molecular profile of the target system, encompassing the gene (NM_001134831.2) and its product, Protein 1A3W. The synthesis of this multi-layered data—from nucleic acid statistics to tertiary structure and evolutionary history—confirms the utility of automated, integrated bioinformatics workflows in modern discovery.

Quantitative Summary of Findings

The pipeline yielded several key quantitative insights:

- Genetic Context: The Central Dogma and Genome Annotation modules established the statistical properties of the nucleic acid sequence, including GC content profiles and the identification and mapping of multiple Open Reading Frames (ORFs) across all six frames. This data is critical for understanding transcriptional regulation and potential alternative splicing or coding regions.
- Physicochemical Profile: Protein Feature Analysis provided the essential physical properties of 1A3W, such as its exact Molecular Weight, Isoelectric Point (pI), and a quantitative assessment of stability via the Instability Index. The generated Hydropathy Plot offers an initial prediction of regional hydrophobicity, suggesting potential surface features or transmembrane domains.
- Structural Architecture: The PDB Structure Analysis confirmed the chain architecture and atomic composition of 1A3W, providing critical data on regional flexibility via B-factor distribution. The interactive 3D visualization translates abstract data into a concrete structural model, vital for functional interpretation.
- Comparative Analysis: Pairwise Alignment and BLAST Search established a robust foundation for functional inference by identifying and quantifying homology with known proteins (via E-values and Percent Identity). The constructed Phylogenetic Tree provided the necessary evolutionary context, mapping the target molecule's ancestry relative to its homologs, which helps to resolve orthologous and paralogous relationships.

Significance and Future Direction

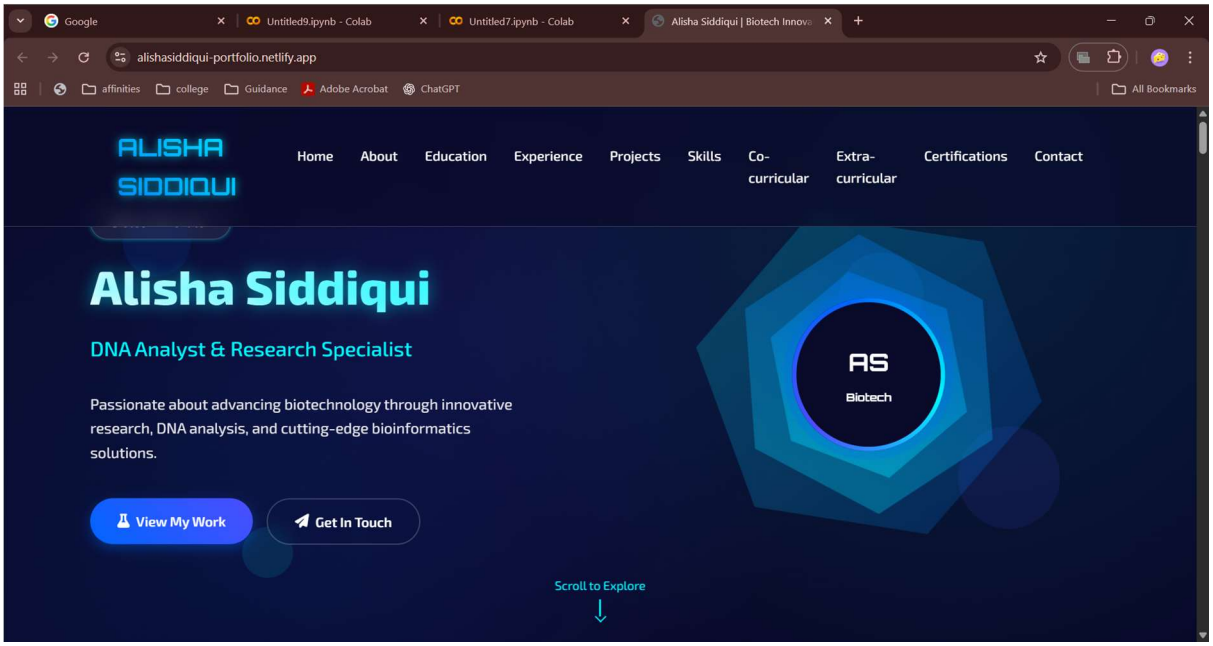
In conclusion, the pipeline is more than just a collection of scripts; it is a validated framework for translating raw genomic and proteomic data into actionable biological hypotheses. It successfully identifies conserved functional regions, predicts potential drug-binding sites (based on structure and physiochemical characteristics), and informs strategies for recombinant production (based on codon usage and stability predictions).

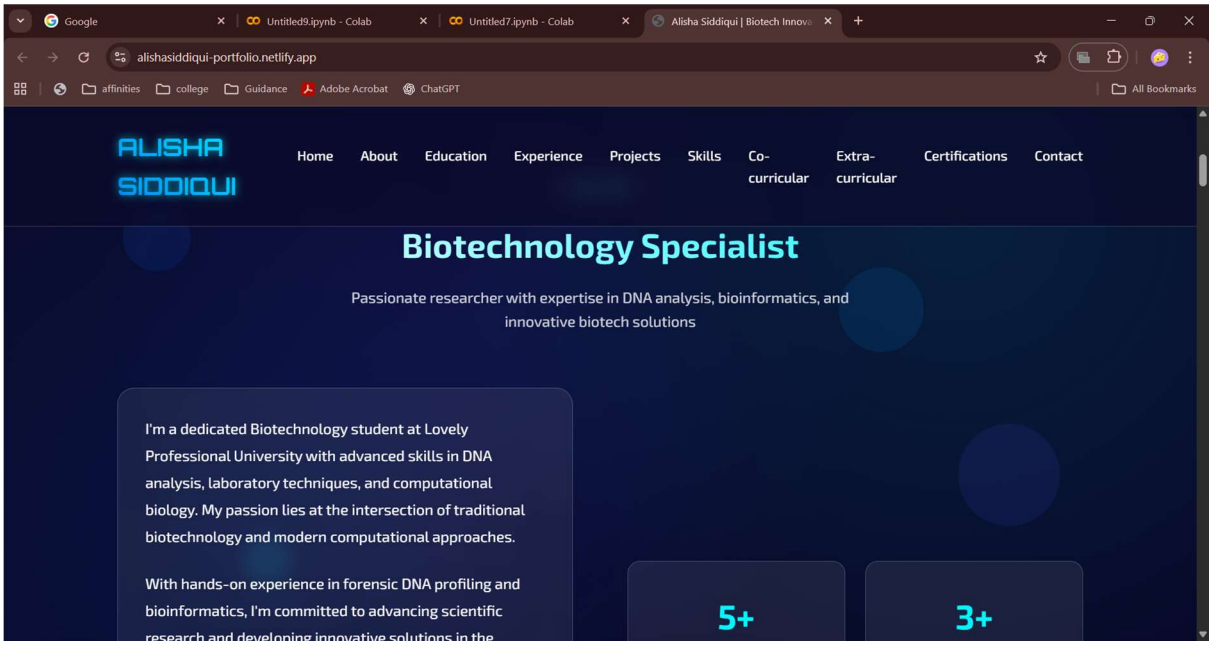
This complete molecular fingerprint provides a robust starting point for subsequent wet-lab experiments, such as site-directed mutagenesis to confirm predicted functional residues or high-throughput screening campaigns against the modeled three-dimensional structure. The integration and automated visualization capabilities ensure that the resulting insights are not

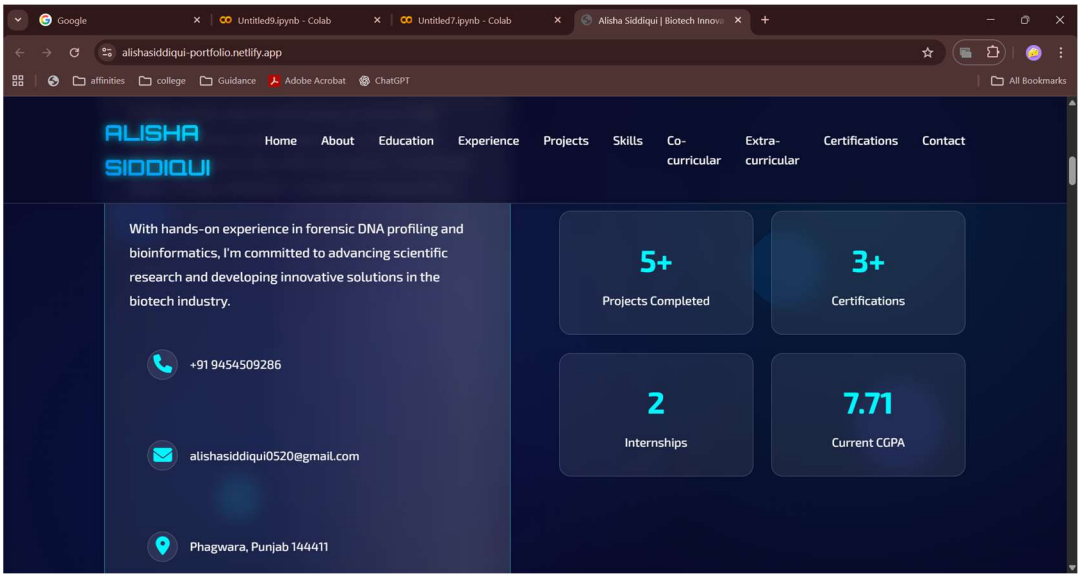
only precise but also immediately accessible and interpretable by researchers across disciplines.

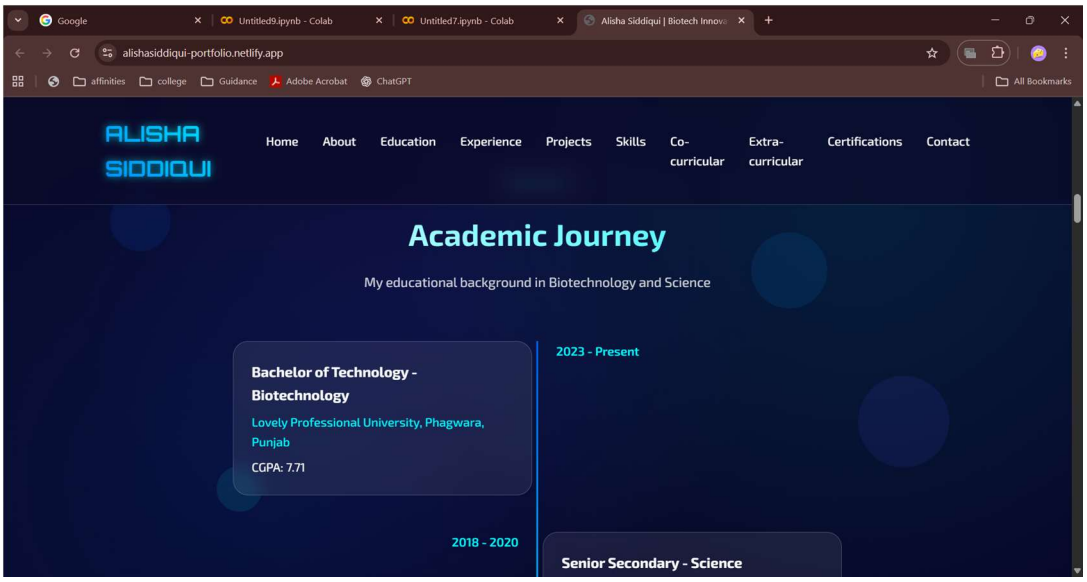
PORTFOLIO:

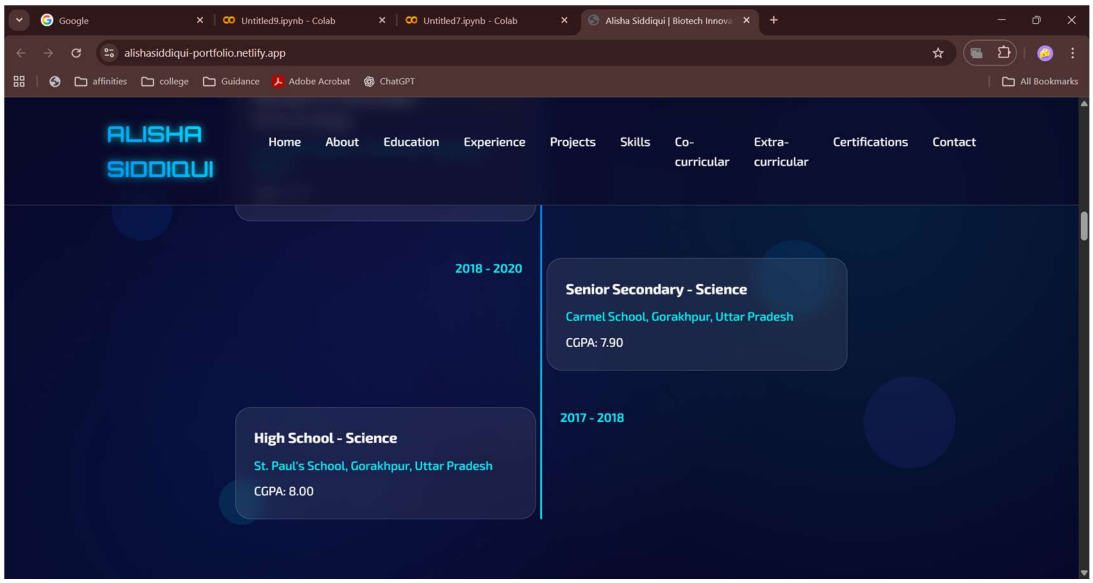
LINK: https://alishasiddiqui-portfolio.netlify.app/

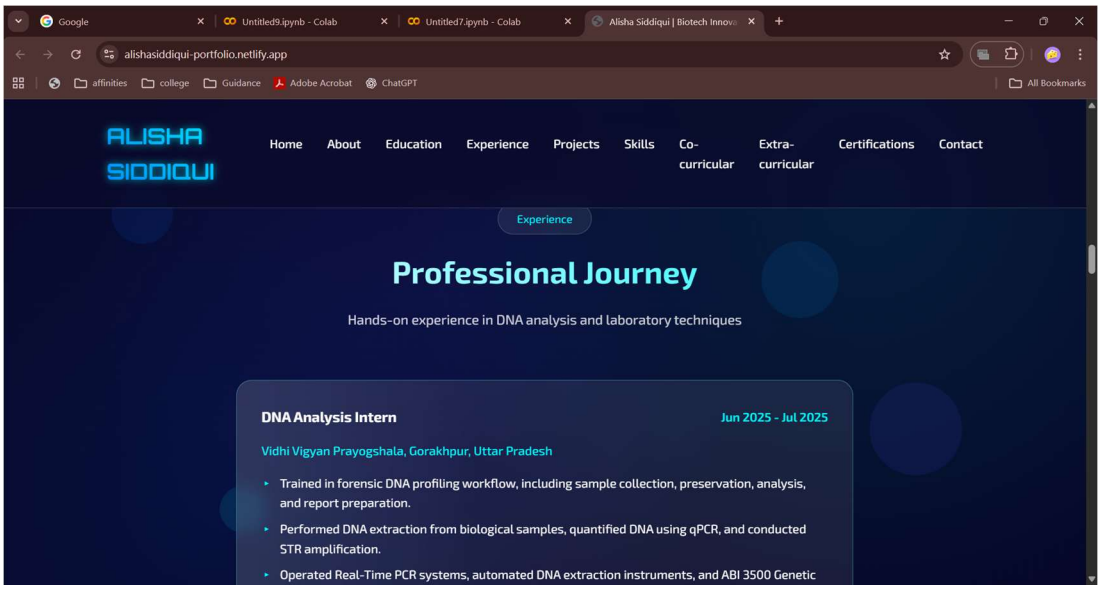


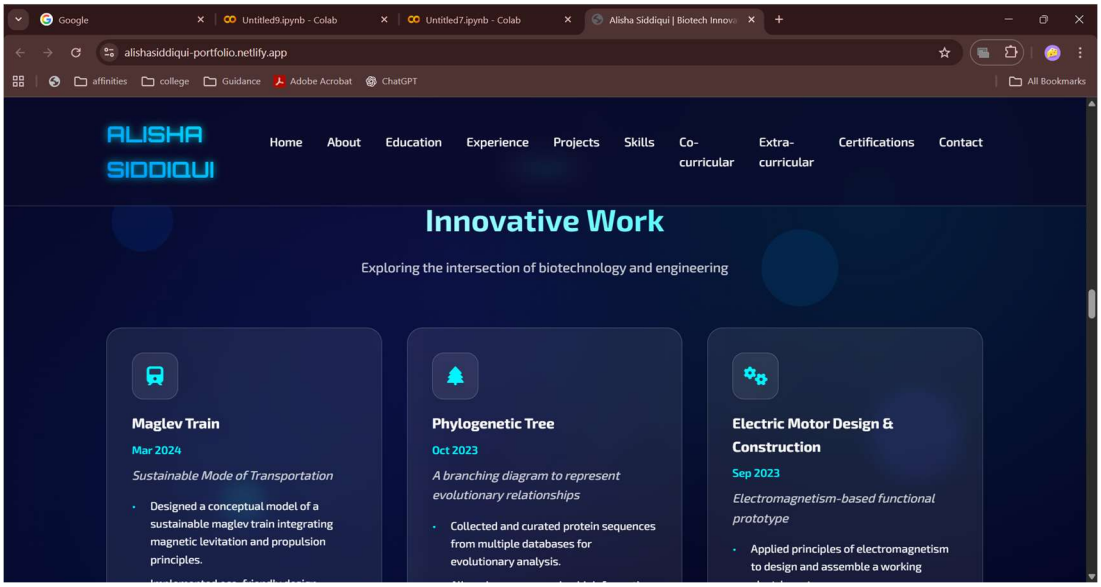


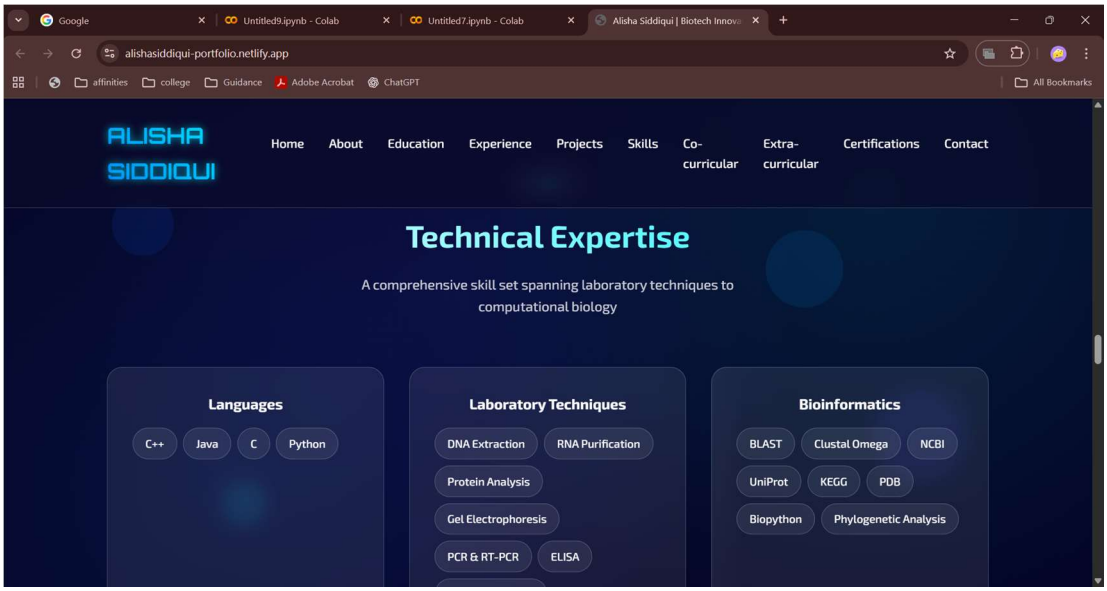

alishasiddiqui-portfolio.netlify.app
ALISHA SIDDIQUI
Home
About
Education
Experience
Projects
Skills
Co-curricular
Extra-curricular
Certifications
Contact
With hands-on experience in forensic DNA profiling and bioinformatics, I'm committed to advancing scientific research and developing innovative solutions in the biotech industry.
+91 9454509286
alishasiddiqui0520@gmail.com
Phagwara, Punjab 144411
5+
Projects Completed
3+
Certifications
2
Internships
7.71
Current CGPA

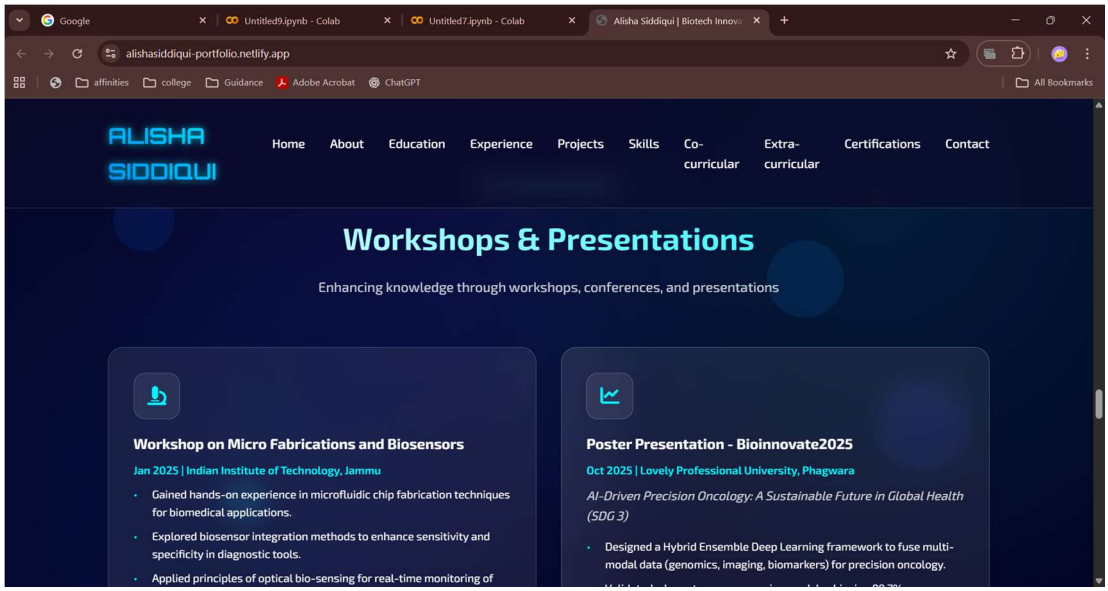

alishasiddiqui-portfolio.netlify.app
ALISHA SIDDIQUI
Home
About
Education
Experience
Projects
Skills
Co-curricular
Extra-curricular
Certifications
Contact
Academic Journey
My educational background in Biotechnology and Science
Bachelor of Technology - Biotechnology
Lovely Professional University, Phagwara, Punjab
CGPA: 7.71
2023 - Present
2018 - 2020
Senior Secondary - Science

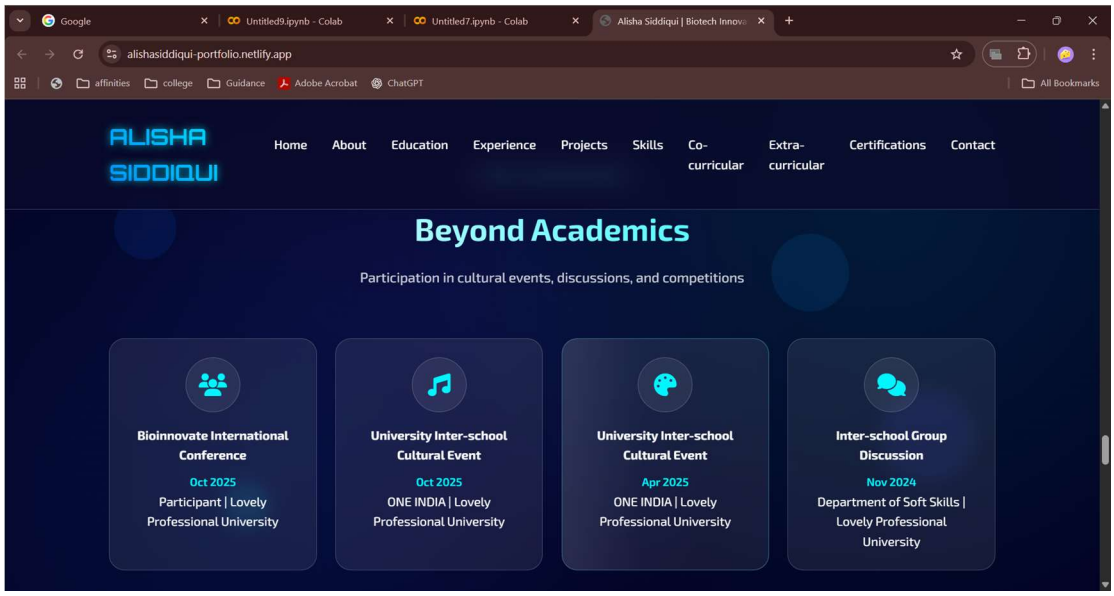

alishasiddiqui-portfolio.netlify.app
ALISHA SIDDIQUI
Home
About
Education
Experience
Projects
Skills
Co-curricular
Extra-curricular
Certifications
Contact
2018 - 2020
Senior Secondary - Science
Carmel School, Gorakhpur, Uttar Pradesh
CGPA: 7.90
2017 - 2018
High School - Science
St. Paul's School, Gorakhpur, Uttar Pradesh
CGPA: 8.00

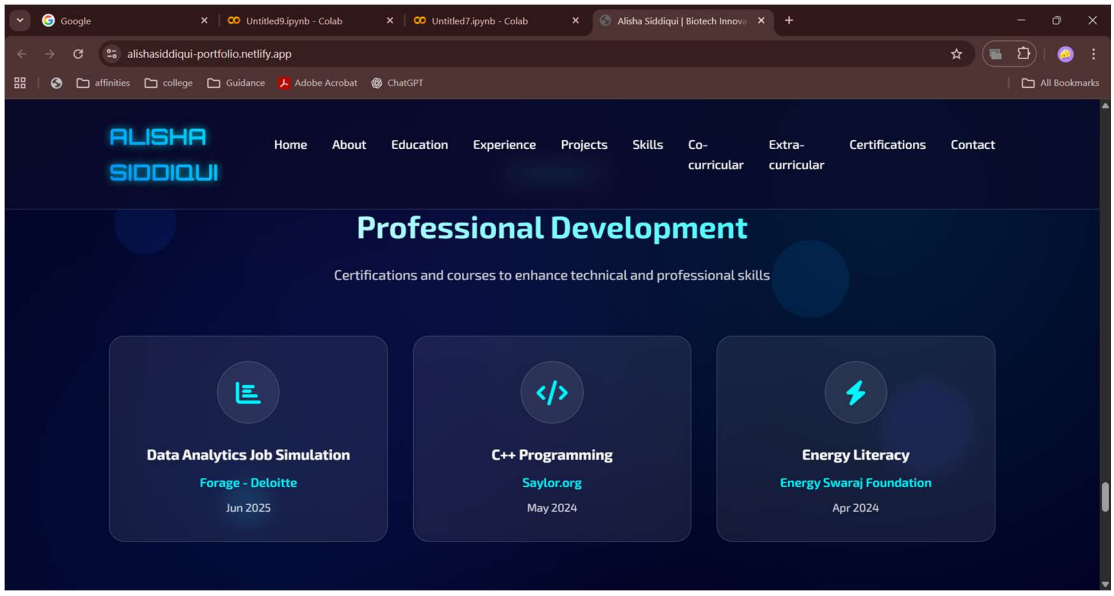

alishasiddiqui-portfolio.netlify.app
ALISHA SIDDIQUI
Home
About
Education
Experience
Projects
Skills
Co-curricular
Extra-curricular
Certifications
Contact
Experience
Professional Journey
Hands-on experience in DNA analysis and laboratory techniques
DNA Analysis Intern
Jun 2025 - Jul 2025
Vidhi Vigyan Prayogshala, Gorakhpur, Uttar Pradesh
Trained in forensic DNA profiling workflow, including sample collection, preservation, analysis, and report preparation.
Performed DNA extraction from biological samples, quantified DNA using qPCR, and conducted STR amplification.
Operated Real-Time PCR systems, automated DNA extraction instruments, and ABI 3500 Genetic

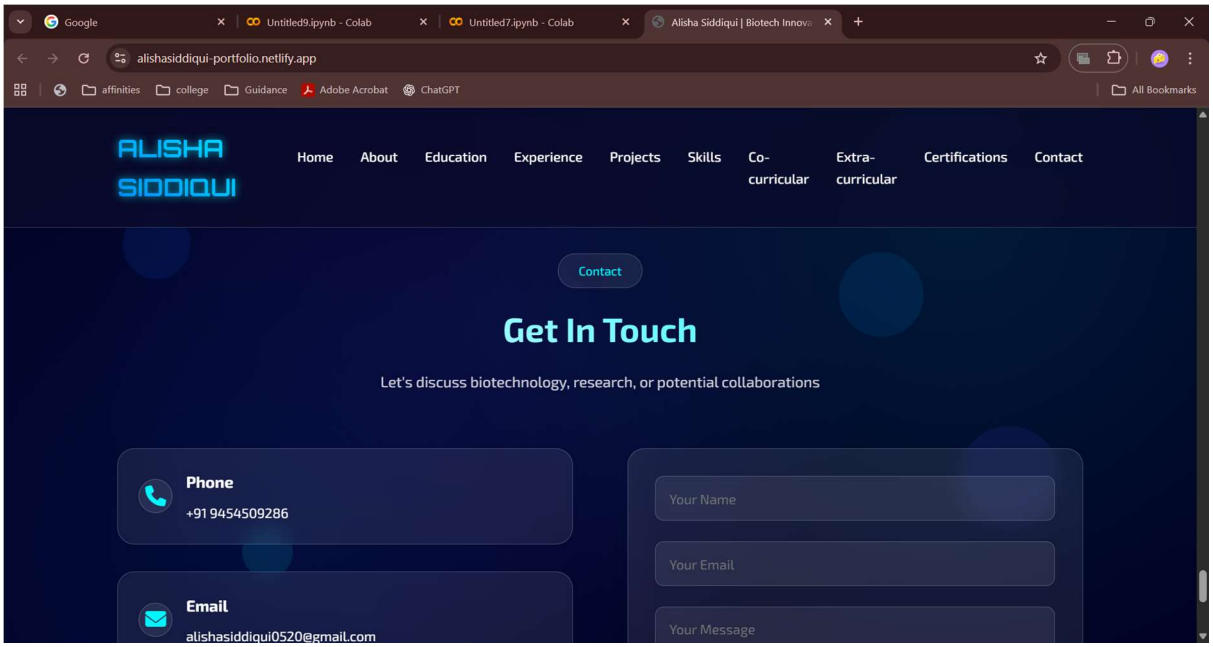

alishasiddiqui-portfolio.netlify.app
ALISHA SIDDIQUI
Home
About
Education
Experience
Projects
Skills
Co-curricular
Extra-curricular
Certifications
Contact
Innovative Work
Exploring the intersection of biotechnology and engineering
Maglev Train
Mar 2024
Sustainable Mode of Transportation
Designed a conceptual model of a sustainable maglev train integrating magnetic levitation and propulsion principles.
Phylogenetic Tree
Oct 2023
A branching diagram to represent evolutionary relationships
Collected and curated protein sequences from multiple databases for evolutionary analysis.
Electric Motor Design & Construction
Sep 2023
Electromagnetism-based functional prototype
Applied principles of electromagnetism to design and assemble a working

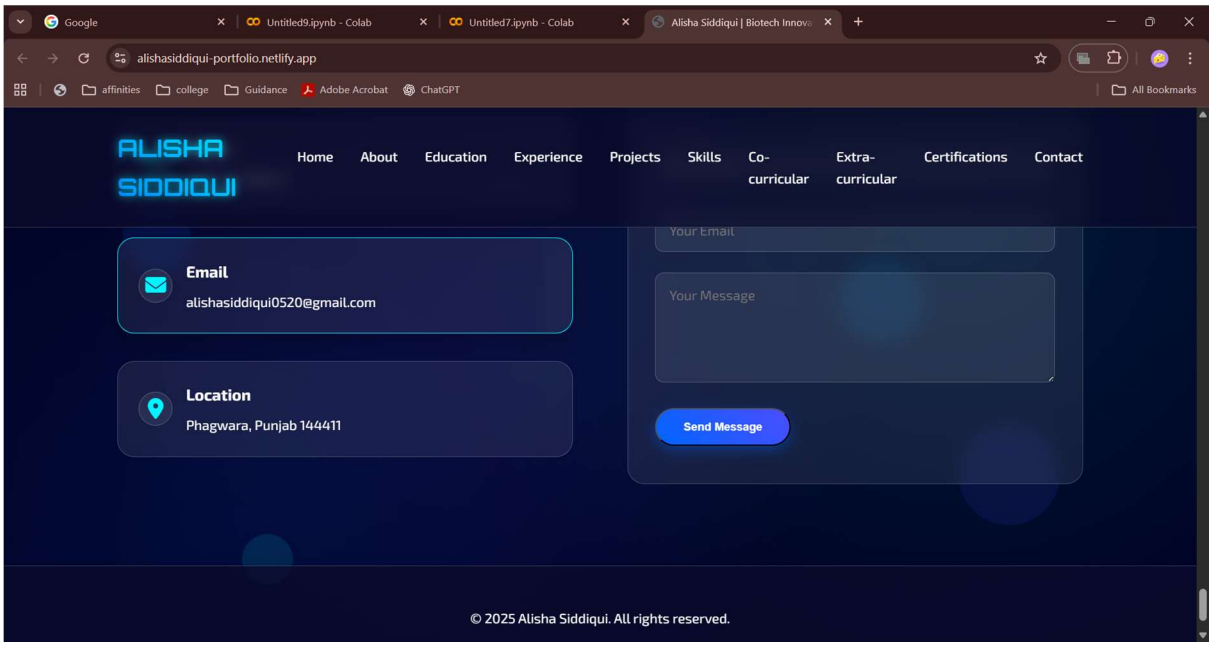

alishasiddiqui-portfolio.netlify.app
ALISHA SIDDIQUI
Home
About
Education
Experience
Projects
Skills
Co-curricular
Extra-curricular
Certifications
Contact
Technical Expertise
A comprehensive skill set spanning laboratory techniques to computational biology
Languages
C++
Java
C
Python
Laboratory Techniques
DNA Extraction
RNA Purification
Protein Analysis
Gel Electrophoresis
PCR & RT-PCR
ELISA
Bioinformatics
BLAST
Clustal Omega
NCBI
UniProt
KEGG
PDB
Biopython
Phylogenetic Analysis


alishasiddiqui-portfolio.netlify.app
ALISHA SIDDIQUI
Home About Education Experience Projects Skills Co-curricular Extra-curricular Certifications Contact
Workshops & Presentations
Enhancing knowledge through workshops, conferences, and presentations
Workshop on Micro Fabrications and Biosensors
Jan 2025 | Indian Institute of Technology, Jammu
Gained hands-on experience in microfluidic chip fabrication techniques for biomedical applications.
Explored biosensor integration methods to enhance sensitivity and specificity in diagnostic tools.
Applied principles of optical bio-sensing for real-time monitoring of
Poster Presentation - Bioinnovate2025
Oct 2025 | Lovely Professional University, Phagwara
AI-Driven Precision Oncology: A Sustainable Future in Global Health (SDG 3)
Designed a Hybrid Ensemble Deep Learning framework to fuse multi-modal data (genomics, imaging, biomarkers) for precision oncology.


alishasiddiqui-portfolio.netlify.app
ALISHA SIDDIQUI
Home About Education Experience Projects Skills Co-curricular Extra-curricular Certifications Contact
Beyond Academics
Participation in cultural events, discussions, and competitions
Bioinnovate International Conference
Oct 2025
Participant | Lovely Professional University
University Inter-school Cultural Event
Oct 2025
ONE INDIA | Lovely Professional University
University Inter-school Cultural Event
Apr 2025
ONE INDIA | Lovely Professional University
Inter-school Group Discussion
Nov 2024
Department of Soft Skills | Lovely Professional University


alishasiddiqui-portfolio.netlify.app
ALISHA SIDDIQUI
Home About Education Experience Projects Skills Co-curricular Extra-curricular Certifications Contact
Professional Development
Certifications and courses to enhance technical and professional skills
Data Analytics Job Simulation
Forage - Deloitte
Jun 2025
C++ Programming
Saylor.org
May 2024
Energy Literacy
Energy Swaraj Foundation
Apr 2024


Google
Untitled9.ipynb - Colab
Untitled7.ipynb - Colab
Alisha Siddiqui | Biotech Innova
alishasiddiqui-portfolio.netlify.app
affinities
college
Guidance
Adobe Acrobat
ChatGPT
All Bookmarks
ALISHA SIDDIQUI
Home
About
Education
Experience
Projects
Skills
Co-curricular
Extra-curricular
Certifications
Contact
Contact
Get In Touch
Let's discuss biotechnology, research, or potential collaborations
Phone
+91 9454509286
Email
alishasiddiqui0520@gmail.com
Your Name
Your Email
Your Message


Google
Untitled9.ipynb - Colab
Untitled7.ipynb - Colab
Alisha Siddiqui | Biotech Innova
alishasiddiqui-portfolio.netlify.app
affinities
college
Guidance
Adobe Acrobat
ChatGPT
All Bookmarks
ALISHA SIDDIQUI
Home
About
Education
Experience
Projects
Skills
Co-curricular
Extra-curricular
Certifications
Contact
Email
alishasiddiqui0520@gmail.com
Location
Phagwara, Punjab 144411
Your Email
Your Message
Send Message
© 2025 Alisha Siddiqui. All rights reserved.