DETAILS OF THE ACTIVITIES PERFORMED DURING THE 2025 INTERNSHIP AT THE NATIONAL INSTITUTE OF BIOTECHNOLOGY

INDEX

Sl. No.	Activity Title	Description / Sub-Points
1	**DNA Isolation**	Extraction of genomic DNA from biological samples for downstream analysis.
2	**Polymerase Chain Reaction (PCR)**	Amplification of specific DNA sequences using thermal cycling.
3	**Real-Time Quantitative PCR (Real-Time qPCR)**	Quantitative analysis of gene expression using fluorescent probes.
4	**Bacterial Colony Identification by the Streak Plate Method**	Isolation of pure bacterial colonies for morphological and biochemical analysis.
5	**DNA Sequencing**	Determination of nucleotide sequence using Sanger or Next-Generation Sequencing technologies.
6	**RNA Sequencing (RNA-Seq)**	High-throughput sequencing of transcriptomes to study gene expression.
7	**RNA Reverse Sequencing**	Conversion of RNA to cDNA followed by sequencing to understand expression profiles.
8	**Antibiotic Resistance and Sensitivity Testing of UTI Isolates**	Detection of antimicrobial resistance in 23 bacterial strains from urinary tract infections.
9	**Electrophoresis**	Separation of DNA, RNA, or proteins based on size using agarose or polyacrylamide gels.
10	**In Silico Molecular Docking for Drug Discovery**	Computational docking of ligands to target proteins to predict binding affinity and drug efficacy.
11	**Linux-based Genome Assembly**	Assembly of raw genome sequencing reads using Linux-based bioinformatics tools.
12	**Protein Sequencing Analysis**	Analysis of amino acid sequences, structural domains, and physicochemical properties of proteins.

1. DNA Isolation

The process of **DNA isolation** was conducted, starting from blood collection and proceeding through sample preparation, concentration, measurement, and quality assessment. DNA isolation involves the extraction of deoxyribonucleic acid (DNA) from cells or tissues by lysing cells, removing proteins and contaminants, and purifying the DNA.

This procedure is fundamental to several applications, including:

- Genetic analysis
- Molecular biology experiments
- Diagnostic and forensic investigations
- Biotechnology research

DNA Isolation Protocol Followed:

1. 200 µL of blood sample was mixed with 200 µL of binding buffer and 20 µL of Proteinase K.
2. The mixture was briefly vortexed and centrifuged to collect the contents.
3. The sample was incubated at 55°C for 10 minutes using a heat block.
4. Subsequently, 200 µL of 70% ethanol was added and vortexed thoroughly.
5. The entire mixture was transferred into a spin column placed in a collection tube.
6. Centrifugation was performed at 10,000 × g for 1 minute; the flow-through was discarded.
7. 500 µL of Wash Buffer I was added to the column, followed by centrifugation at 10,000 × g for 1 minute; the flow-through was discarded.
8. 500 µL of Wash Buffer II was then added, and centrifugation was performed at 13,000 × g for 3 minutes; the flow-through was discarded.
9. The spin column was transferred to a new 1.5 mL collection tube.
10. 200 µL of Elution Buffer was added directly to the column matrix.
11. Final centrifugation was conducted at 10,000 × g for 1 minute to collect purified DNA.
12. The isolated DNA was stored at -20°C for long-term preservation.

2. Polymerase Chain Reaction (PCR)

Polymerase Chain Reaction (PCR) is a molecular biology technique utilized to amplify specific DNA segments by cycling through distinct temperature phases to denature DNA, anneal primers, and extend new DNA strands with the help of the enzyme Taq polymerase.

During the training, two types of PCR techniques were employed: **Conventional PCR** and **Quantitative PCR (qPCR)**.

Conventional PCR Procedure:
Conventional PCR was performed to amplify target DNA sequences, followed by endpoint analysis through gel electrophoresis. This method allows detection of the presence or absence of the DNA target but does not quantify the initial DNA amount.

The protocol followed was:

1. All reagents including Taq polymerase, dNTPs, primers, DNA template, and buffer were thawed on ice.
2. The PCR reaction mix was prepared in sterile PCR tubes.
3. Tubes were gently vortexed and briefly spun down to collect the contents.
4. The PCR tubes were placed into a thermal cycler.
5. Thermal cycling conditions were set as follows:
   - Initial denaturation at 94–95°C for 2–5 minutes
   - 20–30 cycles of:
     - Denaturation at 94–95°C for 30 seconds
     - Annealing at 50–65°C (temperature depending on primer specifics)
     - Extension at 72°C for 30 seconds to 1 minute (dependent on amplicon size)
   - Final extension at 72°C for 5–10 minutes
   - Hold at 4°C indefinitely
6. Upon completion, PCR tubes were removed and briefly centrifuged.
7. A volume of 5–10 µL of the PCR product was loaded onto a 1–2% agarose gel containing an appropriate DNA ladder.
8. Electrophoresis was carried out at 80–120 V for 30–45 minutes.
9. DNA bands were visualized under UV light following staining with ethidium bromide or an alternative safe DNA .

2. Real-time PCR (qPCR)

Real-time PCR, also known as quantitative PCR (qPCR), is a molecular biology technique used not only to amplify DNA but also to quantify the amount of amplified product in real-time during each cycle of PCR. This method facilitates simultaneous detection and quantification of nucleic acids, including DNA or RNA (converted to cDNA).

Principle:

- PCR amplification proceeds as in conventional PCR, with the addition of a fluorescent reporter molecule.
- Fluorescence increases proportionally with the accumulation of PCR product in each cycle.
- Fluorescence is measured at the end of each cycle, enabling real-time quantification of DNA.
- Two main detection chemistries were utilized:
  - DNA-binding dyes (e.g., SYBR Green), which fluoresce upon binding double-stranded DNA.
  - Sequence-specific probes (e.g., TaqMan probes), which emit fluorescence when cleaved during amplification.

Applications of Real-time PCR:

- Gene expression analysis
- Pathogen detection and quantification
- Genotyping and mutation detection
- Copy number variation studies

Real-time PCR Protocol Followed:

Materials:

- DNA or cDNA template
- Forward and reverse primers
- Hot-start DNA polymerase
- Fluorescent dye or probe
- qPCR master mix (buffer, dNTPs, MgCl2)
- Nuclease-free water
- Real-time PCR instrument

Procedure:

1. A master mix was prepared containing:
    - qPCR buffer with DNA polymerase
    - Primers (200–500 nM each)
    - Fluorescent dye or probe
    - Template DNA or cDNA (1–100 ng depending on assay)
    - Nuclease-free water to a final volume of 20 µL
2. Samples and controls (no-template and positive controls) were pipetted into PCR plates or tubes, which were then sealed with optical caps or films.
3. Thermal cycling was conducted on the real-time PCR machine with the following program:
    - Initial denaturation at 95°C for 2–10 minutes to activate polymerase and denature DNA
    - 35–45 cycles consisting of:
        - Denaturation at 95°C for 10–15 seconds
        - Combined annealing and extension at 55–65°C for 30–60 seconds (depending on primer design)
4. Fluorescence measurements were taken at the end of each extension phase, and amplification curves were generated in real-time.
5. Data analysis was performed to determine cycle threshold (Ct) values—the cycle number at which fluorescence exceeds a preset threshold. DNA quantification was achieved relative to standards or controls using either the ΔΔCt comparative method or absolute quantification via standard curves.

3. Real-Time Quantitative PCR (Real-Time qPCR)

Real-Time quantitative PCR (qPCR) is a sensitive molecular biology technique used to amplify and simultaneously quantify a specific DNA sequence during the PCR cycles. Unlike conventional PCR, where detection is performed post-reaction, real-time qPCR enables real-time monitoring of DNA amplification through fluorescent dyes (such as SYBR Green) or sequence-specific probes (such as TaqMan probes).

Methodology:

- All reagents including the qPCR master mix, primers, probes, and cDNA/DNA template were thawed on ice prior to use.
- The reaction mixture was prepared on ice in sterile PCR tubes or plates containing:
  - X µL of 2× qPCR master mix (e.g., SYBR Green or TaqMan)
  - Y µL of forward primer (typically 0.2–0.5 µM final concentration)
  - Y µL of reverse primer (typically 0.2–0.5 µM final concentration)
  - Z µL of probe (when applicable, for TaqMan assays)
  - W µL of template DNA or cDNA (generally 1–100 ng per reaction)
  - Nuclease-free water to reach the final reaction volume (typically 20 µL)
- The reaction mixtures were gently vortexed and briefly centrifuged to collect the contents.
- The tubes or plates were sealed using optical caps or films.
- The samples were loaded into the real-time PCR instrument.
- The thermal cycler was programmed with the following conditions:
  - Initial denaturation at 95°C for 2–10 minutes to activate the polymerase and denature the DNA template.
  - 40–45 cycles of:
    - Denaturation at 95°C for 10–15 seconds
    - Combined annealing and extension at 60°C for 30–60 seconds, during which fluorescence data were collected.
- Fluorescence signals were monitored in real time during the extension phase of each cycle.
- After completion of the cycling program, Ct (cycle threshold) values were analyzed.
- Quantification of target DNA was performed using either standard curves for absolute quantification or the ΔCt method for relative gene expression analysis.

4. Bacterial Colony Identification by the Streak Plate Method

The streak plate method is a microbiological technique employed for isolating pure bacterial cultures and obtaining well-isolated colonies from a mixed population.

Objectives:

1. To obtain a pure culture of bacteria from a mixed culture.

2. To obtain well-isolated bacterial colonies.
3. To propagate bacteria for further study.

By streaking, a dilution gradient is established across the agar plate surface. This gradient results in confluent growth in areas where bacterial cells are not sufficiently separated, while well-isolated, discrete colonies develop in regions with fewer bacteria deposited. Each isolated colony is assumed to have originated from a single bacterium, thus representing a clonal pure culture.

Procedure:

1. The inoculating loop was sterilized by placing it in the flame of a Bunsen burner until it was red hot, or by using a micro-incinerator, followed by allowing it to cool.
2. An isolated colony was picked from the agar plate culture and spread over the first quadrant (approximately one-quarter of the plate) using close, parallel streaks.
3. Immediately afterward, the inoculating loop was streaked gently over the first quadrant using a back-and-forth motion.
4. The loop was flamed again and allowed to cool before streaking was extended from the edge of the first quadrant into the second quadrant of the plate.
5. The flaming and cooling process was repeated before streaking was further extended from the second quadrant into the third quadrant.
6. After flaming and cooling the loop once more, streaking was extended from the third quadrant into the central fourth quadrant of the plate.
7. Finally, the loop was flamed again to complete the procedure.

5. DNA Sequencing

DNA sequencing is the process by which the exact order of nucleotides (adenine, thymine, cytosine, guanine) in a DNA molecule is determined. This process is fundamental in genomics, molecular biology, diagnostics, and biotechnology.

Sanger Sequencing (Chain Termination Method):
Developed by Frederick Sanger in 1977, this first-generation sequencing technique, also known as the dideoxy method, was widely used in the Human Genome Project. The method relies on the selective incorporation of chain-terminating dideoxynucleotides (ddNTPs) by DNA polymerase during in vitro DNA replication.

- **Principle:**
  DNA polymerase incorporates normal nucleotides (dNTPs) that allow elongation due to the presence of a 3'-OH group, whereas the incorporation of modified dideoxynucleotides (ddNTPs), which lack the 3'-OH group, terminates chain elongation.
- **Protocol:**
  1. The template DNA is denatured.
  2. A reaction mixture is prepared containing template DNA, a single primer, DNA polymerase, normal dNTPs, and fluorescently labeled ddNTPs of different colors.
  3. DNA polymerase extends the primer and randomly incorporates ddNTPs, generating fragments of varying lengths.

4. Fragments are separated by size using capillary gel electrophoresis.
5. Fluorescent labels are detected via laser scanning, producing a chromatogram that indicates the sequence based on color peaks.

- **Applications:**
  Small DNA segments sequencing (up to ~1000 bp), mutation analysis, plasmid and gene verification.

Cycle Sequencing:
A variant of Sanger sequencing utilizing PCR-like thermal cycling to amplify the sequencing signal.

- **Principle:**
  Thermal cycling, similar to PCR, is performed with ddNTPs included, allowing simultaneous amplification and sequencing.
- **Protocol:**
  1. Reaction mix containing a single primer, DNA template, DNA polymerase, dNTPs, and fluorescent ddNTPs is prepared.
  2. Thermal cycling includes denaturation (95°C), annealing (50–60°C), and extension (72°C).
  3. Post-PCR cleanup removes excess dyes and unincorporated nucleotides.
  4. Capillary electrophoresis is performed to separate fragments.
  5. Chromatogram output is generated for sequence analysis.
- **Advantages:**
  High sensitivity, automation compatibility, and suitability for low template quantities.

Next-Generation Sequencing (NGS):
Second-generation sequencing technology enables massive parallel sequencing of millions of DNA fragments simultaneously, dramatically increasing throughput.

- **Working Principle:**
  DNA is fragmented and adapters are ligated to enable attachment to a flow cell. Clonal amplification occurs (e.g., bridge amplification), followed by sequencing by synthesis (SBS), where base incorporations emit fluorescent signals detected in real time.
- **General Workflow:**
  1. Library preparation with DNA fragmentation and adapter ligation.
  2. Clonal amplification on flow cells.
  3. Sequencing by synthesis.
  4. Bioinformatics analysis for read alignment and variant calling.

- **Common Platforms and Features:**

Platform	Method	Output
Illumina	SBS + bridge amplification	High accuracy, short reads
Ion Torrent	pH change detection	Faster, cost-effective
Roche 454	Pyrosequencing (discontinued)	Longer reads
SOLiD	Ligation-based	High accuracy
BGI/MGI	Combinatorial Probe-Anchor	Competitor to Illumina

- **Applications:**
  Whole-genome sequencing, transcriptomics (RNA-Seq), epigenetics, metagenomics.

Comparison of Sequencing Methods:

Feature	Sanger Sequencing	Cycle Sequencing	NGS (Illumina, Ion Torrent)	Other (PacBio, Nanopore)
Generation	First-gen	First-gen (variant)	Second-gen	Third-gen (mostly)
Template Type	Cloned/purified DNA	PCR product/plasmid	Fragmented genomic DNA	Genomic DNA (long fragments)
Amplification	None (single primer)	Thermocycling (single primer)	Bridge/emulsion PCR	Sometimes none (single molecule)
Read Length	500–1000 bp	600–800 bp	100–300 bp (paired-end)	PacBio: 10–50 kb, Nanopore: >1 Mb
Accuracy	>99.99%	>99.99%	~98–99.9%	PacBio: ~99%, Nanopore: 95–98%
Throughput	Low	Low to moderate	Very high (millions of reads)	High
Detection Method	Fluorescent ddNTPs + CE	Same as Sanger	Fluorescent imaging, pH-based	Light, electrical current
Time Required	~5–6 hours	~4–6 hours	1–2 days (run + analysis)	Minutes to hours
Cost per Sample	High	Moderate	Very low per base	Moderate to low per base
Hands-on Labor	High	Moderate	Low (automation)	Low to moderate
Sample Multiplexing	Limited	Limited	High (barcoding/indexing)	High
Applications	Small gene sequencing	Mutation detection	Whole-genome, transcriptome	Long-read WGS, structural variants
Data Output	Chromatograms	Chromatograms	Massive FASTQ files	Long-read FASTQ/BAM files

Primer

Definition:
A primer is a short nucleic acid strand, typically 18–25 bases in length, that serves as the initiation site for DNA synthesis. It provides a free 3'-OH group necessary for DNA polymerase to add nucleotides during replication or amplification.

Types of Primers:

- **Forward and Reverse primers:** Used in PCR to flank the target DNA region.
- **Random hexamers:** Used in cDNA synthesis to prime at random locations along RNA.
- **Oligo-dT primers:** Bind specifically to the poly-A tail of mRNA during reverse transcription.

Applications:

- PCR amplification
- DNA and RNA sequencing
- Reverse transcription
- Quantitative PCR (qPCR)

Ready Reaction Mix (RRM)

Definition:
Ready Reaction Mix (RRM) is a pre-formulated solution containing key components for nucleic acid amplification or sequencing reactions.

Typical Components:

- DNA polymerase
- Reaction buffer
- Deoxynucleotide triphosphates (dNTPs)
- Magnesium chloride ($MgCl_2$)
- Sometimes primers and stabilizers
- In sequencing, may also include fluorescently labeled dideoxynucleotides (ddNTPs)

Purpose:

- Simplifies reaction setup by combining multiple reagents into one mix
- Ensures consistency and reproducibility across reactions
- Reduces pipetting errors and contamination risk

Common Uses:

- Sanger sequencing (with fluorescent ddNTPs)
- PCR kits for amplification

6. RNA Sequencing (RNA-Seq)

Definition:
RNA-Seq is a Next-Generation Sequencing (NGS) technique used to analyze the transcriptome—the complete set of RNA transcripts expressed in a cell or tissue at a given time.

Workflow:

1. RNA Isolation
2. mRNA Enrichment or rRNA Depletion
3. RNA Fragmentation
4. cDNA Synthesis via Reverse Transcription
5. Library Preparation (adapter ligation)
6. Sequencing (e.g., Illumina platform)
7. Data Analysis (alignment, expression quantification, differential gene expression)

Applications:

- Gene expression profiling
- Transcript discovery
- Alternative splicing analysis
- Non-coding RNA analysis

7. RNA Reverse Sequencing

Clarification:
The term "RNA Reverse Sequencing" is not standard in molecular biology. It may refer to one of the following concepts:

Possible Interpretations:

A. Reverse Transcription Followed by Sequencing (RT-Seq):

- RNA is reverse transcribed into complementary DNA (cDNA).
- The cDNA is then sequenced using NGS or Sanger methods.
- This process forms the basis of standard RNA-Seq.

B. Antisense RNA Sequencing:

- Sequencing of antisense RNAs (non-coding regulatory RNAs) after reverse transcription.

C. Strand-specific RNA Sequencing (Directional RNA-Seq):

- A technique that preserves the original RNA strand orientation during sequencing.
- Important for identifying antisense transcripts and overlapping genes.

8. Antibiotic Resistance and Sensitivity Testing of 23 Bacterial Isolates from Urinary Tract Infections (UTIs)

Background:
UTIs are commonly caused by bacterial pathogens such as *Escherichia coli*, *Klebsiella* spp., and *Proteus* spp. Effective treatment depends on the accurate determination of antibiotic susceptibility to guide clinical therapy and prevent resistance spread.

Objective:
To determine the antibiotic susceptibility profiles of 23 bacterial isolates from UTI patients using the Kirby-Bauer disk diffusion method.

Principle

The disk diffusion method evaluates bacterial sensitivity to antibiotics by measuring the inhibition of bacterial growth around antibiotic-impregnated disks placed on an inoculated agar surface. The diameter of the zone of inhibition is interpreted according to CLSI guidelines to classify isolates as Sensitive (S), Intermediate (I), or Resistant (R).

Materials and Methods

Sample Collection and Bacterial Isolation:

- 23 urine samples collected aseptically from patients suspected of UTI.
- Cultured on cystine lactose electrolyte-deficient (CLED) agar.
- Incubated at 37°C for 24 hours.
- Colonies identified via morphological, biochemical, and Gram staining methods.

Preparation of Bacterial Suspension:

- Pure colonies suspended in sterile normal saline.
- Turbidity adjusted to 0.5 McFarland standard (~1.5 × 10^8 CFU/mL).

Media and Antibiotic Disks:

- Mueller-Hinton agar plates used.
- Antibiotics tested included:
    - Ciprofloxacin (5 µg)
    - Nitrofurantoin (300 µg)
    - Ampicillin (10 µg)
    - Trimethoprim-sulfamethoxazole (1.25/23.75 µg)
    - Others as needed

Disk Diffusion Procedure

1. Using sterile swabs, bacterial suspension evenly spread on Mueller-Hinton agar plates to form a uniform lawn.
2. Antibiotic disks placed aseptically with sufficient spacing to avoid overlapping inhibition zones.
3. Plates incubated inverted at 35–37°C for 23 hours under aerobic conditions.

Measurement and Interpretation

- After incubation, zone diameters measured in millimeters using a ruler.
- Interpretation based on CLSI 2023 standards:

Antibiotic	Resistant (R) Zone Diameter (mm)	Intermediate (I) Zone Diameter (mm)	Sensitive (S) Zone Diameter (mm)
Ampicillin (10 µg)	≤ 13	14 – 16	≥ 17
Ciprofloxacin (5 µg)	≤ 15	16 – 20	≥ 21
Nitrofurantoin (300 µg)	≤ 7	8 – 10	≥ 11
Trimethoprim-Sulfamethoxazole (1.25/23.75 µg)	≤ 10	11 – 15	≥ 16
Amoxicillin-Clavulanate (20/10 µg)	≤ 13	14 – 17	≥ 18
Gentamicin (10 µg)	≤ 12	13 – 14	≥ 15
Cephalexin (30 µg)	≤ 14	15 – 17	≥ 18

Note: Zone diameter breakpoints may vary slightly based on bacterial species and updated CLSI guidelines.

9. Electrophoresis

Definition:
Electrophoresis is a laboratory technique that separates charged molecules such as nucleic acids, proteins, or small ions based on their size, shape, and charge by applying an electric field. The molecules migrate through a medium, typically a gel, allowing separation and analysis.

Types of Electrophoresis

1. Gel Electrophoresis

Principle:
Molecules migrate through a porous gel matrix under an electric field. Smaller molecules move faster through the gel pores than larger ones.

Common Gels:

- **Agarose Gel:** Mainly used for separating DNA and RNA fragments of various sizes.
- **Polyacrylamide Gel (PAGE):** Primarily used for protein separation and small nucleic acid fragments; can be run under denaturing (SDS-PAGE) or native conditions.

Applications:

- DNA fingerprinting and plasmid analysis
- RNA quality assessment
- Protein size determination and purity analysis

2. Chromatography Electrophoresis (Capillary Electrophoresis)

Principle:
Separation occurs inside a narrow capillary filled with electrolyte under high voltage. Molecules separate based on their charge-to-size ratio.

Advantages:

- High resolution and rapid separation
- Automated with minimal sample volume required

Applications:

- Analysis of amino acids, nucleotides, peptides, and small proteins
- Clinical diagnostics and pharmaceutical quality control

3. Other Types of Electrophoresis

- **Isoelectric Focusing (IEF):** Separates proteins by their isoelectric point (pI) using a pH gradient. Proteins focus where their net charge is zero. Useful for protein characterization.
- **Pulsed-Field Gel Electrophoresis (PFGE):** Used for very large DNA molecules; the electric field periodically changes direction to improve resolution. Commonly used in bacterial strain typing.
- **Two-Dimensional Electrophoresis (2-DE):** Combines IEF and SDS-PAGE to separate proteins by pI and molecular weight. Widely used in proteomics.
- **Free-Flow Electrophoresis (FFE):** Continuous separation in a flowing system with an applied electric field, used for preparative separations and cell sorting.

Agarose Gel Electrophoresis (for DNA/RNA)

Materials:

- Agarose powder
- TAE or TBE buffer
- DNA/RNA samples with loading dye
- DNA stain (e.g., ethidium bromide or GelRed)
- Gel casting tray and comb
- Power supply and electrophoresis chamber

Protocol:

1. **Prepare agarose gel:**
   - Weigh appropriate agarose (0.8–2% depending on fragment size).
   - Dissolve in TAE or TBE buffer by heating until fully melted.
   - Cool to ~50°C, add DNA stain if not staining post-run.
   - Pour into casting tray with comb; let solidify (~20–30 minutes).
2. **Load samples:**
   - Mix DNA/RNA with loading dye.
   - Carefully pipette into wells.
3. **Run electrophoresis:**
   - Place gel in chamber filled with running buffer.
   - Connect electrodes (DNA migrates toward positive/anode).
   - Run at 80–120 V for 30–90 minutes.
4. **Visualize bands:**
   - Use UV transilluminator or gel documentation system.

Polyacrylamide Gel Electrophoresis (SDS-PAGE for Proteins)

Materials:

- Acrylamide/bis-acrylamide solution
- SDS-PAGE running buffer
- Sample buffer with SDS and reducing agent
- Protein samples
- Gel casting apparatus
- Electrophoresis chamber and power supply
- Protein stain (Coomassie blue or silver stain)

Protocol:

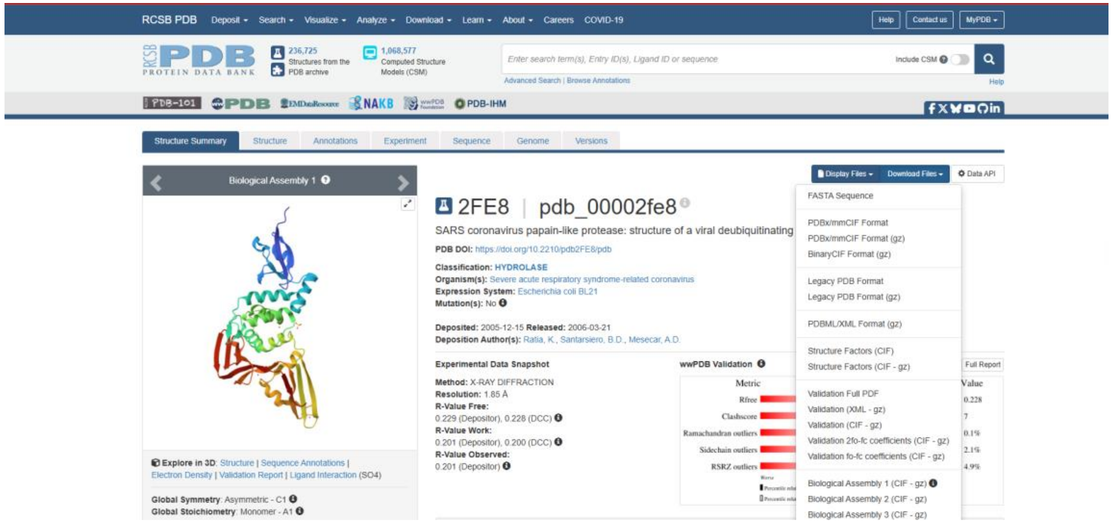
1. **Prepare gels:**
   - Prepare separating gel (8–15% acrylamide) and pour into plates.
   - Overlay with isopropanol to smooth surface; polymerize.
   - Pour stacking gel (4–5% acrylamide) on top; insert comb.
2. **Prepare samples:**
   - Mix protein with SDS sample buffer.
   - Heat at 95°C for 5 minutes to denature proteins.
3. **Load samples and run:**
   - Load samples and molecular weight ladder.
   - Run electrophoresis at constant voltage (100–150 V) until dye front reaches bottom.
4. **Stain and visualize:**
   - Stain gel with Coomassie blue or silver stain to detect protein bands.

10. In Silico Molecular Docking for Drug Discovery

Evaluation of PLPro of SARS-CoV-2 NIB-1 and GRL0617 docking:

1. Protein Retrieval:

The target protein was retrieved from the RCSB Protein Data Bank (rcsb.org). The protein name or PDB ID was searched, and the correct entry was selected. The structure file was downloaded in PDB format, and the file was saved locally for further preparation.



2FE8 | pdb_00002fe8

SARS coronavirus papain-like protease: structure of a viral deubiquitinating

PDB DOI: https://doi.org/10.2210/pdb2FE8/pdb

Classification: HYDROLASE
Organism(s): Severe acute respiratory syndrome-related coronavirus
Expression System: Escherichia coli BL21
Mutation(s): No

Deposited: 2005-12-15 **Released:** 2006-03-21
Deposition Author(s): Ratia, K., Santarsiero, B.D., Mesecar, A.D.

Experimental Data Snapshot

Method: X-RAY DIFFRACTION
Resolution: 1.85 Å
R-Value Free:
0.229 (Depositor), 0.228 (DCC)
R-Value Work:
0.201 (Depositor), 0.200 (DCC)
R-Value Observed:
0.201 (Depositor)

wwPDB Validation

Metric	Value
Rfree	0.228
Clashscore	7
Ramachandran outliers	0.1%
Sidechain outliers	2.1%
RSRZ outliers	4.9%

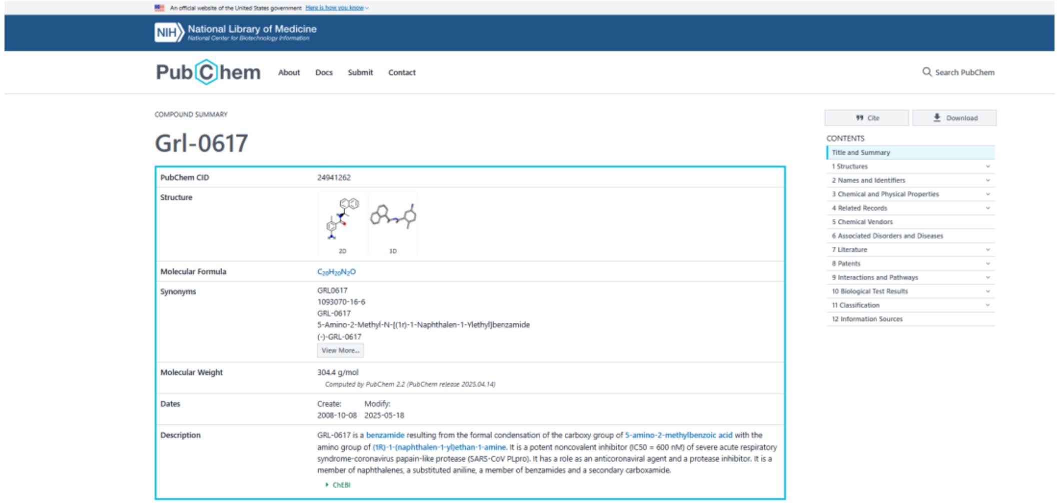
Explore in 3D: Structure | Sequence Annotations | Electron Density | Validation Report | Ligand Interaction (SO4)

Global Symmetry: Asymmetric - C1
Global Stoichiometry: Monomer - A1

FIgure: Protein form RCSB PDB

2. Ligand Selection and SMILES Extraction:

The desired ligand was located using the PubChem database (pubchem.ncbi.nlm.nih.gov). After the compound was selected, its Canonical SMILES string was obtained from the "Chemical Structure" section. This string was copied for use in the next step.



An official website of the United States government Here is how you know
NIH National Library of Medicine
National Center for Biotechnology Information
PubChem
About
Docs
Submit
Contact
Search PubChem
Cite
Download
CONTENTS
Title and Summary
1 Structures
2 Names and Identifiers
3 Chemical and Physical Properties
4 Related Records
5 Chemical Vendors
6 Associated Disorders and Diseases
7 Literature
8 Patents
9 Interactions and Pathways
10 Biological Test Results
11 Classification
12 Information Sources
2D
3D
View More...
ChEBI

COMPOUND SUMMARY

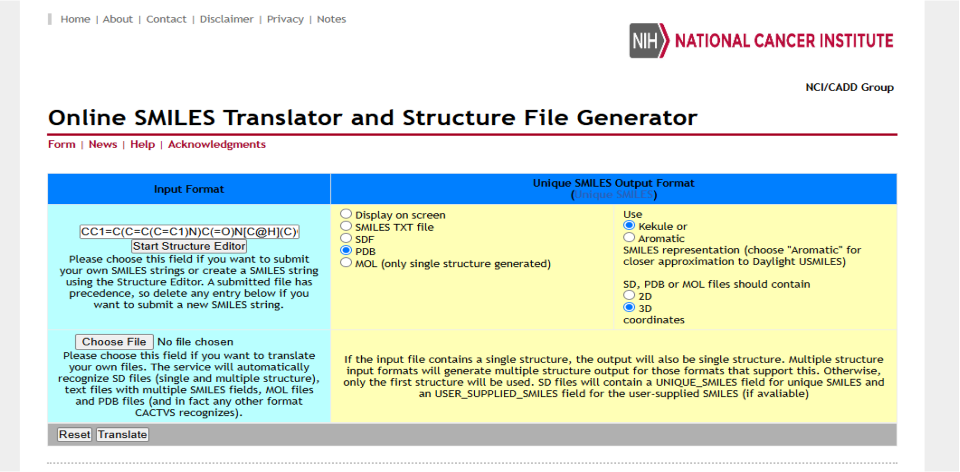
Grl-0617

PubChem CID	24941262
Structure	
Molecular Formula	$C_{20}H_{20}N_2O$
Synonyms	GRL0617 1093070-16-6 GRL-0617 5-Amino-2-Methyl-N-[(1r)-1-Naphthalen-1-Ylethyl]benzamide (-)-GRL-0617
Molecular Weight	304.4 g/mol Computed by PubChem 2.2 (PubChem release 2025.04.14)
Dates	Create: 2008-10-08 Modify: 2025-05-18
Description	GRL-0617 is a benzamide resulting from the formal condensation of the carboxy group of 5-amino-2-methylbenzoic acid with the amino group of (1R)-1-(naphthalen-1-yl)ethan-1-amine. It is a potent noncovalent inhibitor (IC50 = 600 nM) of severe acute respiratory syndrome-coronavirus papain-like protease (SARS-CoV PLpro). It has a role as an anticoronaviral agent and a protease inhibitor. It is a member of naphthalenes, a substituted aniline, a member of benzamides and a secondary carboxamide.

FIgure:Grl-0617 Ligand

3. SMILES Conversion to 3D Structure Using CACTUS:

The SMILES string was submitted to the CACTUS online translator (cactus.nci.nih.gov/translate/). The format was specified as either SDF or PDB, and the conversion was performed by the server. The resulting 3D structure was downloaded and saved.



Home | About | Contact | Disclaimer | Privacy | Notes
NIH NATIONAL CANCER INSTITUTE
NCI/CADD Group
Form | News | Help | Acknowledgments
Start Structure Editor
Choose File
Reset
Translate

Online SMILES Translator and Structure File Generator

Input Format	Unique SMILES Output Format (Unique SMILES)	
CC1=C(C=C(C=C1)N)C(=O)N[C@H](C) Please choose this field if you want to submit your own SMILES strings or create a SMILES string using the Structure Editor. A submitted file has precedence, so delete any entry below if you want to submit a new SMILES string.	○ Display on screen ○ SMILES TXT file ○ SDF ● PDB ○ MOL (only single structure generated)	Use ● Kekule or ○ Aromatic SMILES representation (choose "Aromatic" for closer approximation to Daylight USMILES) SD, PDB or MOL files should contain ○ 2D ● 3D coordinates
No file chosen Please choose this field if you want to translate your own files. The service will automatically recognize SD files (single and multiple structure), text files with multiple SMILES fields, MOL files and PDB files (and in fact any other format CACTVS recognizes).	If the input file contains a single structure, the output will also be single structure. Multiple structure input formats will generate multiple structure output for those formats that support this. Otherwise, only the first structure will be used. SD files will contain a UNIQUE_SMILES field for unique SMILES and an USER_SUPPLIED_SMILES field for the user-supplied SMILES (if avaliable)	

Figure: Conversion into 3D format

4. Protein modification:

To prepare the protein for docking, PyMOL was used to visually inspect and modify the structure, specifically to remove water molecules that were not essential for ligand binding.

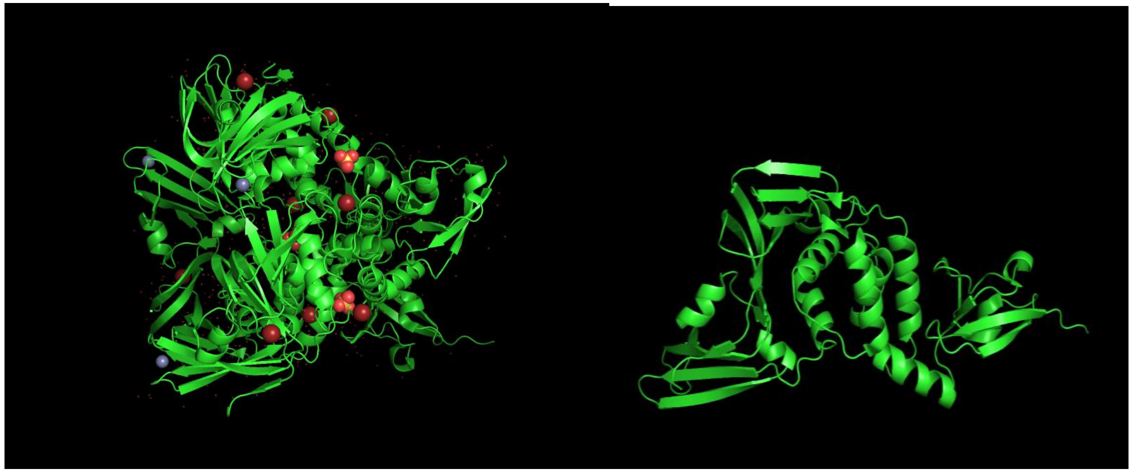


Figure: Cutting water and isolating the 2nd strand of protein

1. Autodocking:

Docking was carried out using AutoDock Vina, following a structured protocol demonstrated during the webinar session. A suitable grid box was defined around the active or binding site of the protein, ensuring that the ligand had enough space to explore various conformations during docking.

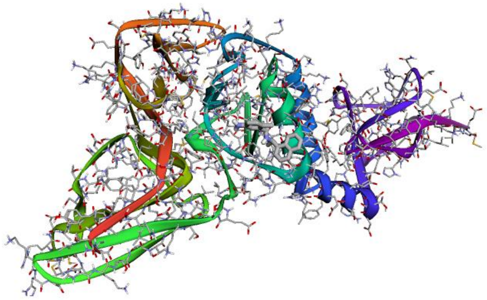


Figure: Mode 1 alignment after docking

Mode	Binding Affinity (kcal/mol)	RMSD Lower Bound (Å)	RMSD Upper Bound (Å)
1	-7.524	0.000	0.000
2	-7.195	28.230	29.844
3	-7.122	31.440	34.618
4	-6.786	27.884	30.111
5	-6.747	11.264	12.559
6	-6.674	2.021	2.280
7	-6.484	28.207	32.662
8	-6.478	36.165	38.246
9	-6.368	2.316	3.900

Table: Mode 1 has the lowest (best) binding energy of -7.524 kcal/mol, indicating the most favorable binding pose.

Proteomics:

Proteomics is the characterization/study of proteome, including expression, structure, functions, interactions and modifications of proteins at any stage. Proteome is the collection of all translated proteins in a cell: secretory proteins, target proteins or extracellular proteins.

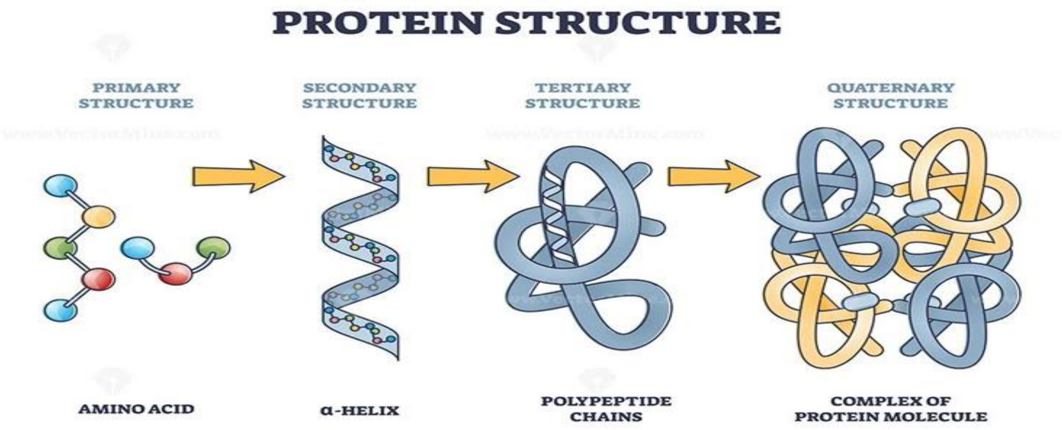
Protein products are the translated polypeptides from genes, which may undergo changes

- Post-translational modifications (PTMs) (e.g., phosphorylation, glycosylation)
- Folding into active 3D structures
- Complex formation with other biomolecules

Proteins have four levels of structure:

- **Primary** – amino acid sequence
- **Secondary** – α-helices, β-sheets (local folding)
- **Tertiary** – 3D structure of a single polypeptide
- **Quaternary** – structure formed by multiple subunits (if applicable)

Proteins have four levels of structure that determine their shape and function. The primary structure is the linear sequence of amino acids in a polypeptide chain, held together by peptide bonds. This sequence dictates how the protein will fold. The secondary structure refers to localized folding into patterns like alpha-helices and beta-sheets, stabilized by hydrogen bonds between backbone atoms. Building on this, the tertiary structure is the overall 3D shape of a single polypeptide, formed by interactions among side chains, including hydrogen bonds, ionic bonds, hydrophobic interactions, and disulfide bridges. Some proteins also have a quaternary structure, which involves the assembly of multiple polypeptide subunits into a larger functional complex, as seen in proteins like haemoglobin. Each level of structure is essential for the protein's stability and biological activity.



Domains and Motifs:

A **domain** is a distinct, independently folding part of a protein that often has a specific function, such as binding to DNA, other proteins, or small molecules. Proteins can have one or multiple domains, and each domain can contribute to the overall function of the protein. For example, an enzyme might have a catalytic domain and a regulatory domain.

A motif is a short, conserved sequence or structural pattern that appears in different proteins and is usually associated with a specific function, such as a zinc finger motif involved in DNA binding or a helix-turn-helix motif in transcription factors. Motifs are not independently stable, while domains are independently stable.

To find the domains of a protein, interpro scan is used

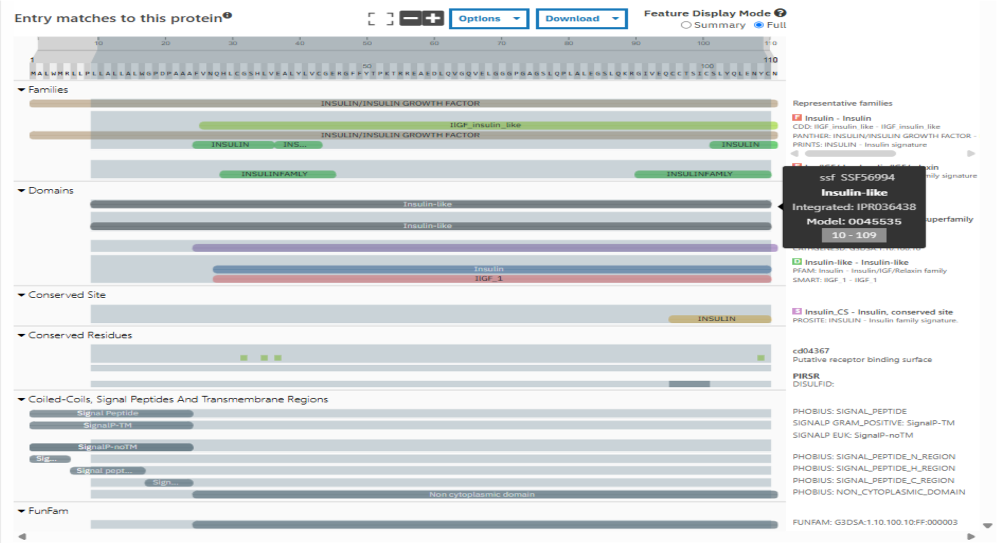

Entry matches to this protein
Options
Download
Feature Display Mode
○ Summary ● Full
MALWMRLLPLLALLALWGPDPAAAFVNQHLCGSHLVEALYLVCGERGFFYTPKTRREAEDLQVGQVELGGGPGAGSLQPLALEGSLQKRGIVEQCCTSICSLYQLENYCN
Families
INSULIN/INSULIN GROWTH FACTOR
IIGF_insulin_like
INSULIN/INSULIN GROWTH FACTOR
INSULIN
INS...
INSULIN
INSULINFAMLY
INSULINFAMLY
Domains
Insulin-like
Insulin-like
Insulin
IIGF_1
Conserved Site
INSULIN
Conserved Residues
Coiled-Coils, Signal Peptides And Transmembrane Regions
Signal Peptide
SignalP-TM
SignalP-noTM
Sig...
Signal pept...
Sign...
Non cytoplasmic domain
FunFam
Representative families
F Insulin - Insulin
CDD: IIGF_insulin_like - IIGF_insulin_like
PANTHER: INSULIN/INSULIN GROWTH FACTOR -
PRINTS: INSULIN - Insulin signature
ssf SSF56994
Insulin-like
Integrated: IPR036438
Model: 0045535
10 - 109
uperfamily
D Insulin-like - Insulin-like
PFAM: Insulin - Insulin/IGF/Relaxin family
SMART: IIGF_1 - IIGF_1
S Insulin_CS - Insulin, conserved site
PROSITE: INSULIN - Insulin family signature.
cd04367
Putative receptor binding surface
PIRSR
DISULFID:
PHOBIUS: SIGNAL_PEPTIDE
SIGNALP GRAM_POSITIVE: SignalP-TM
SIGNALP EUK: SignalP-noTM
PHOBIUS: SIGNAL_PEPTIDE_N_REGION
PHOBIUS: SIGNAL_PEPTIDE_H_REGION
PHOBIUS: SIGNAL_PEPTIDE_C_REGION
PHOBIUS: NON_CYTOPLASMIC_DOMAIN
FUNFAM: G3DSA:1.10.100.10:FF:000003

Figure: the domain and motif analysis

Insulin is a well-characterized peptide hormone with strong evidence supporting its role as an insulin-like signaling molecule. It has a conserved structure closely resembling that of human insulin and insulin-like growth factors (IGFs), reflecting its essential biological function. The presence of a signal peptide confirms that insulin is secreted into the extracellular environment, consistent with its role in endocrine signaling. It lacks transmembrane domains, indicating that it is not membrane-bound. Conserved cysteine residues and a distinct insulin signature motif are critical for forming disulfide bonds that stabilize its functional three-dimensional structure. A predicted receptor-binding site further supports its ability to interact with specific insulin receptors, triggering downstream signaling pathways. Its classification in multiple structural and functional family databases confirms insulin’s evolutionary conservation and well-established role in metabolic regulation.

Protein-protein docking:

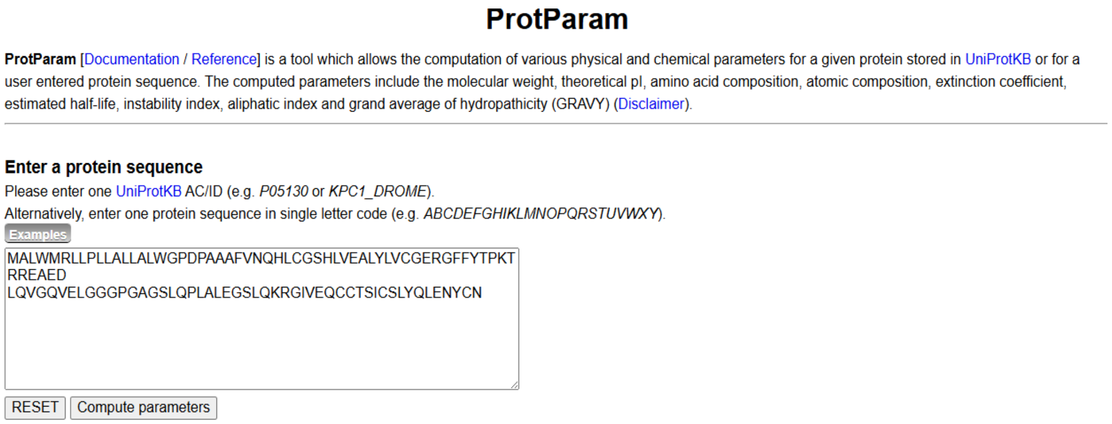
Protein-protein docking is a computational method used to predict the structure of a complex formed by two or more interacting proteins. This technique helps in understanding protein interactions, which are crucial for many biological processes such as signal transduction, enzyme activity regulation, and immune responses. The docking process involves predicting the binding mode between proteins based on their three-dimensional structures, typically obtained from experimental data like X-ray crystallography or NMR spectroscopy.

Several online tools and software packages are available to perform protein-protein docking, including HADDOCK, ClusPro, and ZDOCK. These platforms use algorithms that consider shape complementarity, electrostatics, and energetic scoring to generate and rank possible interaction models. The output provides insights into the binding interface, affinity, and stability of the protein complex, aiding in drug design and functional annotation of proteins.

Physicochemical properties of protein:

Analysing Physicochemical Properties (Using ProtParam)

Go to ProtParam: https://web.expasy.org/protparam/ Paste the FASTA sequence into the box. Click “Compute Parameters”.



ProtParam

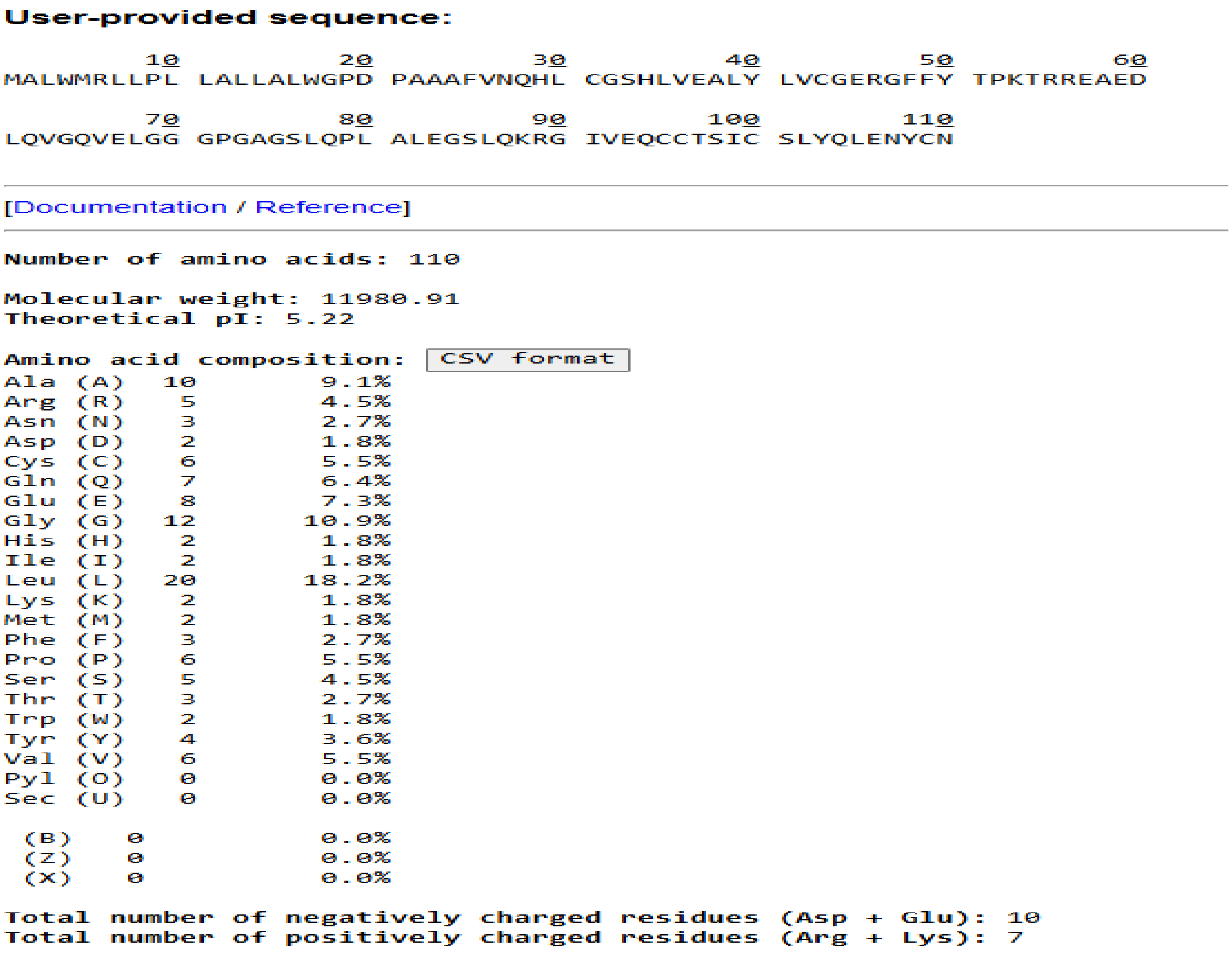
ProtParam [Documentation / Reference] is a tool which allows the computation of various physical and chemical parameters for a given protein stored in UniProtKB or for a user entered protein sequence. The computed parameters include the molecular weight, theoretical pI, amino acid composition, atomic composition, extinction coefficient, estimated half-life, instability index, aliphatic index and grand average of hydropathicity (GRAVY) (Disclaimer).

Enter a protein sequence

Please enter one UniProtKB AC/ID (e.g. *P05130* or *KPC1_DROME*).

Alternatively, enter one protein sequence in single letter code (e.g. *ABCDEFGHIKLMNOPQRSTUVWXY*).

```
MALWMRLLPLLALLALWGPDPAAAFVNQHLCGSHLVEALYLVCGERGFFYTPKT
RREAED
LQVGQVELGGGPGAGSLQPLALEGSLQKRGIVEQCCTSICSLYQLENYCN
```



User-provided sequence:

```
         10         20         30         40         50         60
MALWMRLLPL LALLALWGPD PAAAFVNQHL CGSHLVEALY LVCGERGFFY TPKTRREAED

         70         80         90        100        110
LQVGQVELGG GPGAGSLQPL ALEGSLQKRG IVEQCCTSIC SLYQLENYCN
```

[Documentation / Reference]

```
Number of amino acids: 110

Molecular weight: 11980.91
Theoretical pI: 5.22

Amino acid composition: [CSV format]
Ala (A)  10         9.1%
Arg (R)   5         4.5%
Asn (N)   3         2.7%
Asp (D)   2         1.8%
Cys (C)   6         5.5%
Gln (Q)   7         6.4%
Glu (E)   8         7.3%
Gly (G)  12        10.9%
His (H)   2         1.8%
Ile (I)   2         1.8%
Leu (L)  20        18.2%
Lys (K)   2         1.8%
Met (M)   2         1.8%
Phe (F)   3         2.7%
Pro (P)   6         5.5%
Ser (S)   5         4.5%
Thr (T)   3         2.7%
Trp (W)   2         1.8%
Tyr (Y)   4         3.6%
Val (V)   6         5.5%
Pyl (O)   0         0.0%
Sec (U)   0         0.0%

 (B)   0            0.0%
 (Z)   0            0.0%
 (X)   0            0.0%

Total number of negatively charged residues (Asp + Glu): 10
Total number of positively charged residues (Arg + Lys): 7
```

Properties table:

Property	**Value**	**Description**
Amino Acids	110	Total number of residues
Molecular Weight	11,980.91 Da	Total protein mass
Theoretical pI	5.22	pH where net charge is zero
Net Charge at pH 7	–3	Slightly acidic
Instability Index	40.33	>40 = unstable
Aliphatic Index	102.91	High = thermostable
GRAVY	0.193	Slightly hydrophobic
Half-life (mammal)	30 hours	In vitro
Half-life (yeast)	>20 hours	In vivo
Half-life (*E. coli*)	>10 hours	In vivo
Extinction Coeff. (oxidized)	17,335 $M^{-1}\cdot cm^{-1}$	UV absorbance at 280 nm
Formula	$C_{535}H_{841}N_{143}O_{153}S_{8}$	Molecular formula
Total Atoms	1,680	Atom count in molecule
Acidic Residues (D+E)	10	Negative charge contributors
Basic Residues (R+K)	7	Positive charge contributors

Methodological Workflow for Identifying Druggable Targets in S. pneumoniae (or Similar)

Step No.	Step Name	Web Tool / Resource	Why This Step? (Purpose)	How To Do It (Step-by-Step)	Requirements	Final Goal of the Step
1	Core Genome Retrieval	1. EDGAR 3.0 2. PATRIC 3. OrthoFinder 4. PanOCT 5. Roary	To find conserved genes common across many *Streptococcus* strains – ideal for broad-spectrum targets.	1. Go to EDGAR 3.0 2. Select at least 20 Streptococcus genomes 3. Set S. pneumoniae as reference 4. Run pan/core genome analysis 5. Download list of core genes	Access to sequenced genomes (RefSeq IDs)	Create a reliable dataset of conserved genes
2	Identify Hypothetical Proteins	1. NCBI Genome 2. UniProt 3. IMG 4. MicroScope 5. Ensembl Bacteria	To focus only on uncharacterized proteins (HPs) which may be unexplored drug targets	1. Open proteome file of core genes 2. Search for proteins annotated as "hypothetical protein" 3. Extract their locus IDs and sequences	Genome annotation in GenBank/FASTA	Isolate potential novel targets for further analysis
3	Functional Annotation	1. GO FEAT 2. InterPro 3. Pfam 4. SMART 5. eggNOG	To predict possible functions of HPs via domain/motif homology	1. Paste HP sequence into GO FEAT or InterPro 2. Run analysis 3. Record biological process, molecular function, cellular localization	HP sequences in FASTA format	Understand role of each HP biologically
4	Physicochemical Analysis	1. ProtParam 2. PepCalc 3. Protein Calculator v3.4 4. Compute pI/MW Tool (ExPASy) 5. ProPAS	To determine protein properties: size, charge, stability, hydrophobicity, etc.	1. Go to ProtParam 2. Paste protein sequence 3. Record MW, pI, instability index, aliphatic index, GRAVY	FASTA sequence	Filter stable, soluble proteins suitable for in vivo validation
5	Subcellular Localization	1. CELLO v2.5 2. PSORTb 3. SOSUI 4. SignalP 5. DeepLoc	To know where the protein works – cytoplasmic proteins are good drug targets, extracellular for vaccines	1. Submit sequence to CELLO & PSORTb 2. Compare predictions 3. Record predicted location	FASTA sequence	Classify HPs for drug or vaccine targeting
6	Unconventional Protein	1. SecretomeP 2. OutCyte	To check if protein is secreted without signal peptides –	1. Submit FASTA to OutCyte & SecretomeP	FASTA, internet access	Detect novel non-

	Secretion (UPS)	3. LipoP 4. PrediSi 5. Phobius	common in pathogenic virulence factors	2. Note secretion scores and presence/absence of signal peptide		classical secretory proteins
7	Essential Gene Prediction	1. DEG 2. OGEE 3. Geptop 4. EssenGene 5. BLASTP (against essential proteins)	Essential genes are those necessary for survival – top priority drug targets	1. Run BLAST of HPs against DEG 2. Set E-value < 0.0001, Bit score > 100 3. Record hits	BLAST interface; protein sequences	Prioritize proteins that bacteria cannot live without
8	Virulence/Pathogenicity Check	1. VirulentPred 2. MP3 3. VFanalyzer (VFDB) 4. PBIT 5. PHI-base	To identify proteins linked to infection and immune evasion	1. Paste each HP into VirulentPred 2. Repeat for MP3 3. Compare output: virulent, pathogenic or both	FASTA, internet access	Highlight HPs involved in disease mechanisms
9	Protein-Protein Interaction (PPI) Analysis	1. STRING 2. BioGRID 3. IntAct 4. Mentha 5. PrePPI	To find out how HPs interact with other proteins – hubs may control important processes	1. Use gene name or Uniprot ID 2. Select *S. pneumoniae* strain 3. Build and export interaction network	Annotated protein list	Discover central HPs that regulate cell functions
10	Druggability Prediction	DrugBank BLAST 1. DrugBank 2. ChEMBL 3. BindingDB 4. SwissTargetPrediction 5. TTD - Therapeutic Target Database	Check if a protein is similar to known drug targets – fast track to repurposing	1. BLAST HP sequence against DrugBank 2. Filter: E-value < 0.0001 and Bit score > 100 3. Record matched drugs	BLAST or local NCBI tools	Shortlist HPs with therapeutic potential
11	3D Structure Prediction	1. Robetta 2. SWISS-MODEL 3. AlphaFold DB 4. I-TASSER 5. Phyre2	To create models for docking if crystal structure is unavailable	1. Paste FASTA into Robetta 2. Wait for model 3. Validate using Ramachandran plot, MolProbity score	FASTA, email access	Get protein models for docking analysis
12	Molecular Docking	1. PyRx 2. AutoDock Vina 3. SwissDock 4. CB-Dock2 5. DockThor	Simulate drug binding to protein active sites	1. Convert SMILES to PDB (via CACTUS) 2. Load protein & ligand into PyRx 3. Run docking 4. Analyze binding energies	Ligands from DrugBank; software installed	Evaluate which drugs best bind to HP targets

13	Molecular Dynamics (MD)	1. GROMACS 2. AMBER 3. NAMD 4. CHARMM-GUI 5. Desmond (Schrödinger)	To test the stability of protein-drug complex under simulated physiological conditions	1. Prepare topology files 2. Neutralize, solvate, minimize 3. Run NVT, NPT, production simulation 4. Analyze RMSD, RMSF, SASA, Rg	GROMACS on Linux; terminal knowledge	Ensure stable and realistic binding behavior
14	Data Analysis & Visualization	1. R (ggplot2) 2. Python (matplotlib/seaborn) 3. Excel 4. GraphPad Prism 5. Tableau Public	Present your findings visually and statistically	1. Export docking & MD data 2. Plot graphs (RMSD, SASA, etc.) 3. Prepare tables and figures	CSV data files; coding or spreadsheet skills	Prepare figures for publication or thesis
15	Manuscript Writing	1. Overleaf (LaTeX) 2. Grammarly 3. Zotero (reference) 4. EndNote 5. Typeset.io	Compile research into a structured, readable scientific document	1. Write Abstract, Intro, Methodology, Results, Discussion 2. Cite using Zotero/Mendeley 3. Include tables, figures, supplements	Writing tools; time & patience	Produce a professional, publishable paper

11.Linux-based Genome Assembly

Genome assembly is the computational process of reconstructing a complete genome sequence from short DNA fragments (reads) generated by sequencing technologies. Due to the vast amount of data and complexity involved, genome assembly requires efficient and powerful computational tools. Linux-based systems are widely used for genome assembly because of their stability, flexibility, and support for bioinformatics software.

Genome assembly is the process of reconstructing the original genome sequence of an organism from smaller DNA fragments generated through sequencing technologies. Since most modern sequencing platforms produce short DNA reads, genome assembly involves computationally piecing these overlapping reads together to form longer contiguous sequences called contigs, which are further ordered and oriented into scaffolds representing chromosomes.

Types of Genome Assembly:

- **De novo Assembly:** Constructing a genome sequence without a reference, relying solely on overlaps between reads.
- **Reference-guided Assembly:** Using a known reference genome to align reads and build the genome sequence.

Steps in Genome Assembly:

1. **Quality Control:** Filtering and trimming raw sequencing data to remove low-quality reads and artifacts.

2. **Read Overlapping:** Identifying overlaps between reads using algorithms such as De Bruijn graphs or Overlap-Layout-Consensus (OLC).
3. **Contig Construction:** Merging overlapping reads into contiguous sequences.
4. **Scaffolding:** Ordering and orienting contigs using paired-end or mate-pair read information.
5. **Gap Filling and Polishing:** Refining the assembly to close gaps and correct errors.

Applications:

- Understanding genome structure and organization.
- Identifying genes, regulatory elements, and variants.
- Comparative genomics and evolutionary studies.
- Supporting functional genomics and molecular biology research.

12. Protein Sequencing Analysis

Protein sequencing analysis is a crucial technique in molecular biology and proteomics aimed at determining the precise amino acid sequence of a protein. This information is vital for understanding protein structure, function, post-translational modifications, and interactions, which are essential for elucidating biological processes and identifying therapeutic targets.

In my project, protein sequencing analysis is employed to characterize key proteins involved in [your specific organism, disease, or system — e.g., the Human T-cell Leukemia Virus type 1 (HTLV-1) proteome]. By determining the amino acid sequences, we aim to identify conserved domains, functional motifs, and potential drug-binding sites that can inform drug design and vaccine development.

The workflow typically involves:

- Protein extraction and purification from biological samples.
- Proteolytic digestion (e.g., trypsinization) to generate peptides.
- Mass spectrometry analysis for peptide identification.
- Computational reconstruction of the full protein sequence.
- Functional annotation using bioinformatics tools such as motif/domain identification and protein structure prediction.