CS-4049 Bioinformatics

Spring 2025

Rushda Muneer

Contact Information

- Rushda Muneer
  - MS Computer Science, Hanyang University, South Korea
  - BSc Computer Engineering, UET, Lahore
- Research
  - Bioinformatics, Advanced Machine Learning, AI in Healthcare
- Office
  - F-024 (CS Department)
  - Tuesday & Thursday (2:00 -3:00 pm) (Send an Email before appointment)
- Phone
  - 042-111 128 128 Ext: 632
- Email
  - rushda.muneer@lhr.nu.edu.pk

Grading Policy

Quiz (Theoretical)	• 10
Assignments (Programming)	• 10
Sessional I	• 15
Sessional II	• 15
Project	• 10
Final	• 40
Total	• 100

Reference Books

Bioinformatics Algorithms: An Active Learning Approach
Phillip Compeau & Pavel Pevzner, 2nd Edition

Bioinformatics and Functional Genomics, Jonathan Pevsner, 2nd Edition

Bioinformatics Computing by Bryan Bergeron, Harvard Medical School and MIT Massachusetts, USA.

Term Project

Submission of 1-2 pages **Proposal** with project title, group info, contact details, abstract, tentative/algorithms or tools, outcome or deliverables, references

Due Date: February 7, 2025 (4:00pm) [Hardcopy & Soft copy]

Submission of **Mid Evaluation Report**

Due Date: March 21, 2025 (4:00 pm)

Final Report submission [Problem formulation, existing solutions, proposed solution/methodology/framework, implementation/experiments, results, conclusion, references] will be held in the last week of semester (**May 9,2025. 4:00pm**)

Presentation in second last week of semester.

2-3 students in a group (Role of each member must be reflected individually)

Tentative Course Outline

Week 1 Introduction to Bioinformatics , Origin of Replication
Week 2 Molecular Biology, Motif Search Algorithms for Molecular clocks
Week 3 Biological Databases, Gene Ontology
Week 4 Genomics, Genome Assembly
Week 5 Sequence Analysis (Antibiotics), Mass Spectrometry
Sessional 1
Week 6 Sequence Similarity and Homology
Week 7 Global and Local Alignment
Week 8 Dynamic Programming

Tentative Course Outline

Week 9 Human Genome, Sorting by Reversals, DotPlot Matrices
Week 10 Phylogenic Analysis
Sessional 2
Week 11 Transcriptomics, Clustering, Gene Expression Analysis
Week 12 Multiple Pattern Matching, BLAST
Week 13 Hidden Markov Models, HIV Phenotyping
Week 14 Proteomics, Peptide Sequencing
Week 15 Computational Annotation of Biomolecules (Deep Learning and Image-Based Methods for Understanding Biomolecules)

Introduction to Bioinformatics

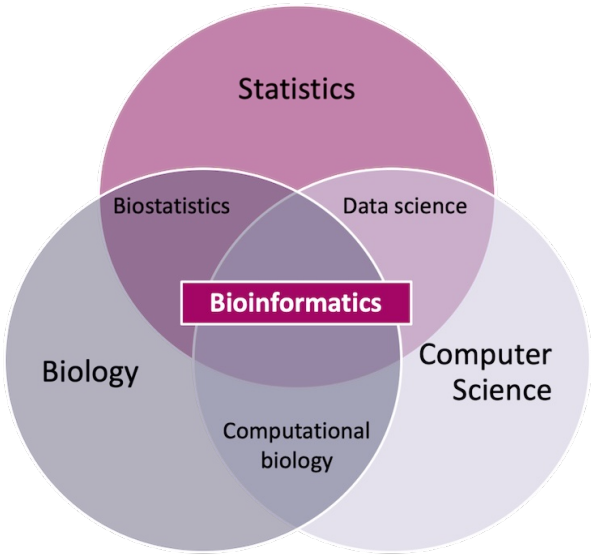
Introduction to Bioinformatics

What is Bioinformatics?

- Bioinformatics is the interdisciplinary field that combines **biology**, **computer science**, and **statistics** to analyze and interpret biological data.
- It involves the application of computational techniques to understand **genomics**, **proteomics**, **systems biology**, and more.

Key Areas of Bioinformatics

- **Genomics**
  - Study of genomes (complete sets of DNA)
  - Sequence alignment, genome assembly, variant analysis
- **Proteomics**
  - Study of proteins and their functions
  - Protein structure prediction, protein-protein interaction analysis
- **Systems Biology**
  - Analysis of biological systems as a whole
  - Modeling and simulating biological processes

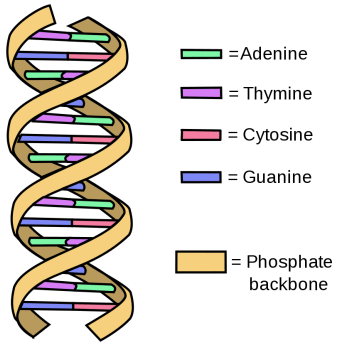


Introduction to Bioinformatics

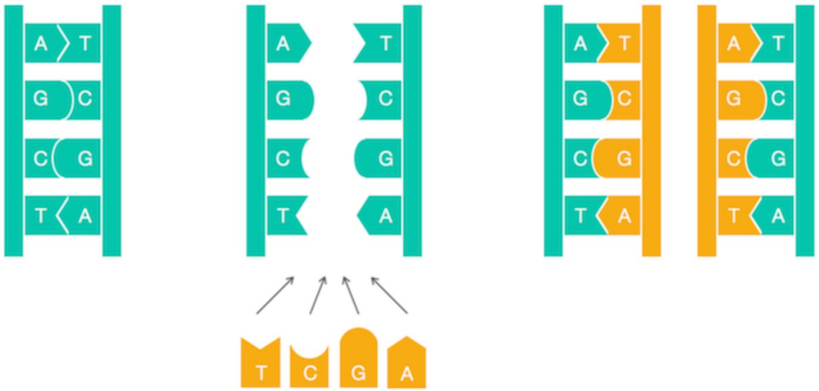
- **Importance of Bioinformatics**
  - Facilitates the understanding of **complex biological systems**.
  - Helps in **drug discovery**, **personalized medicine**, and **disease diagnosis**.
  - Enables the storage, retrieval, and analysis of vast amounts of biological data.
- **Databases** such as GenBank, Ensembl, and Protein Data Bank (PDB).
- **Programming languages** like Python, R, and Perl.

Genome Replication: A Vital Cellular Process

- Genome replication is essential for cell division.
  - Ensures each daughter cell inherits a complete genome.
- Landmark discovery by James Watson and Francis Crick (1953):
  - DNA double helix structure.
  - *“The specific pairing we have postulated immediately suggests a possible copying mechanism for the genetic material.”*
- **Mechanism:**
  - Parent DNA strands unwind.
  - Each strand acts as a template for synthesizing a complementary strand.
- **Result:**
  - Starts with one pair of complementary strands.
  - Ends with two pairs of complementary strands.



DNA



Computational Relevance of Genome Replication

- **Initial Perception:**
  - At first glance, replication might seem irrelevant to computational problems.
  - Simple mimicry involves copying a genome string algorithmically.
- **Biological Insights Provide Computational Challenges:**
  - Replication starts at a specific genomic region called the **replication origin (ori)**.
  - DNA polymerases, molecular copy machines, execute replication.
- **Biomedical Applications of Locating ori:**
  - Essential for understanding replication mechanisms and gene therapy.
  - **Gene Therapy Example:**
    - Viral vectors with artificial genes treat genetic disorders.
    - 1990: First successful gene therapy saved a child with Severe Combined Immunodeficiency Disorder.
    - Applications in agriculture (e.g., frost-resistant tomatoes, pesticide-resistant corn).
- **Significance of ori in Gene Therapy:**
  - Ensures replication of the vector genome inside cells.
  - Biologists must locate ori to maintain functional replication after genetic manipulation.

Computational Problem

- **Input:** A DNA string representing the genome.
- **Output:** The location of the **ori** in the genome.
- **STOP and Think:**
  Does the **Finding Origin of Replication Problem** represent a well-defined computational problem?
- Let's discuss and analyze.

Why Computational Approaches Matter in Locating ori

- **Limitations of Experimental Methods:**
  - Biologists often plan experiments, such as segment deletions, to locate ori.
  - Experimental approaches are slow and costly.
  - ori has only been experimentally identified in a few species.
- **Computer Science Perspective:**
  - The **Finding ori Problem** is not well-defined computationally.
  - Computer scientists require clear inputs, outputs, and constraints to proceed.
- **Importance of Computational Methods in Biology:**
  - **Speed:** Faster than traditional experimental techniques.
  - **Interpretation:** Many experimental results require computational analysis for meaningful insights.
- **Goal:**
  - Develop computational solutions to predict ori efficiently.
  - Enable biologists to focus resources on other critical research tasks.

Finding ori in Bacterial Genomes: The Case of DnaA Boxes

- **Focus on Bacterial Genomes:**
  - Most bacterial genomes consist of a single circular chromosome.
- **Region encoding ori:**
  - Typically a few hundred nucleotides long.
  - **DnaA boxes** are specific DNA sequences that bind **DnaA protein** to initiate replication.
- **Research Approach:**
- Start with a known ori.
- Analyze this region to identify unique features for computational prediction in other bacteria.

DnaA Boxes & Origin of Replication (ori) in Bacteria

- **Example: Vibrio cholerae (Pathogen causing cholera)**
  - The **ori** region in *Vibrio cholerae* contains a known **DnaA box** sequence
  - Patterns in this sequence (e.g., repetitive motifs like "DnaA boxes") guide computational analysis.

```
atcaatgatcaacgtaagcttctaagcatgatcaaggtgctcacacagtttatccacaac
ctgagtggatgacatcaagataggtcgttgtatctccttcctctcgtactctcatgacca
cggaaagatgatcaagagaggatgatttcttggccatatcgcaatgaatacttgtgactt
gtgcttccaattgacatcttcagcgccatattgcgctggccaaggtgacggagcgggatt
acgaaagcatgatcatggctgttgttctgtttatcttgttttgactgagacttgttagga
tagacggtttttcatcactgactagccaaagccttactctgcctgacatcgaccgtaaat
tgataatgaatttacatgcttccgcgacgatttacctcttgatcatcgatccgattgaag
atcttcaattgttaattctcttgcctcgactcatagccatgatgagctcttgatcatgtt
tccttaaccctctatttttacggaagaatgatcaagctgctgctcttgatcatcgtttc
```

Finding the Hidden Message in ori

- **Replication and the Role of DnaA**
  - The *Vibrio cholerae* chromosome is **1,108,250 nucleotides** long.
  - Replication begins at a **short region (ori)**.
  - The **DnaA protein** binds to a specific sequence called the **DnaA box**.
  - The **DnaA box** acts as a **"bind here!" signal** for replication to start.
- **The Challenge**
  - How can we **find the hidden message** in **ori** without knowing what it looks like?
  - What **patterns or features** make **ori** stand out?

Hidden Message Problem

- **Goal:** Find a “hidden message” in the replication origin.
- **Input:** A string **Text** (representing the replication origin).
- **Output:** A **hidden message** in **Text**.
- **STOP & THINK**
- Does this represent a **clearly stated computational problem**?
- What defines a **hidden message**?
- How do we **systematically find it**?

Hidden Messages in "The Gold-Bug"

- **The Challenge of Hidden Messages**
  - The **Hidden Message Problem** lacks a clear computational definition.
  - The **ori** region of *Vibrio cholerae* is just as mysterious as the **parchment** in Edgar Allan Poe's *The Gold-Bug*.
- 📜 **Ciphered text from "The Gold-Bug":**

```
53‡‡†305))6·;48826)4‡.)4‡);806·;48†8^60))85;161;:‡·8
†83(88)5·†;46(;88·96·?;8)·‡(;485);5·†2:·‡(;4956·2(5
·—4)8^8·;4069285);)6†8)4‡‡;1(‡9;48081;8:8‡1;48†85;4
)485†528806·81(‡9;48;(88;4(‡?34;48)4‡;1‡(;:188;‡?;
```

- **Legrand's Approach to Decoding**
  1. **Identified recurring symbols**: ";48" appeared frequently.
  2. **Assumed pirates spoke English** and matched ";48" to "THE".
  3. **Substituted characters** (; → T, 4 → H, 8 → E), simplifying the text:
- 📜 **Partially Deciphered Text:**

```
53‡‡†305))6·THE26)H‡.)H‡)TE06·THE†E^60))E5T161T:‡·E
†E3(EE)5·†TH6(TEE·96·?TE)·‡(THE5)T5·†2:·‡(TH956·2(5
·—H)E^E·TH0692E5)T)6†E)H‡‡T1(‡9THE0E1TE:E‡1THE†E5TH
)HE5†52EE06·E1(‡9THET(EETH(‡?3HTHE)H‡T1‡(T:1EET‡?T
```

Counting Words in DNA Sequences

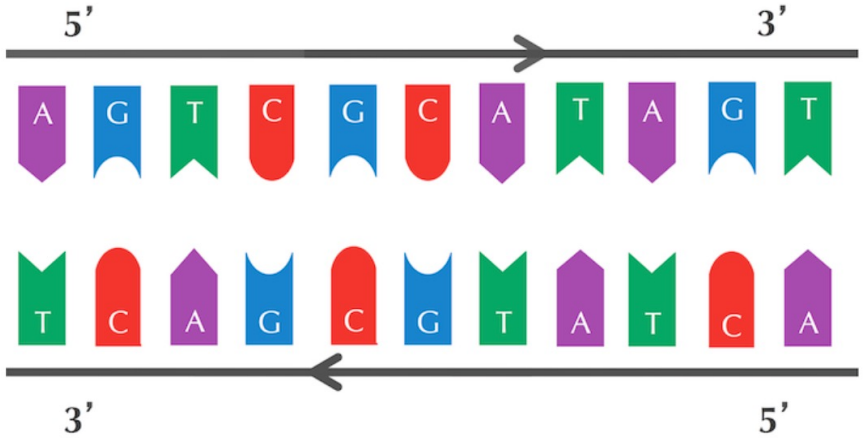
- **DNA as a Language**
  - DNA sequences can be analyzed **like text** to find **frequent patterns**.
  - Frequent **nucleotide strings** may indicate **protein binding sites**.
  - More occurrences **increase binding chances** and reduce mutation effects.
- **Finding Frequent k-mers**
  - A **k-mer** is a string of length **k**.
- **Count(Text, Pattern)** = number of times **Pattern** appears in **Text** (including overlaps).
- **Example**
- **Text:**
  ACAACTATGCATACTATCGGGAACTATCCT
- **Pattern:**
  ACTAT
- **Count(**ACAACTATGCATACTATCGGGAACTATCCT**,** ACTAT**) = ?**

The Frequent Words Problem

- **Definition**
  - A **Pattern** is a **most frequent k-mer** in **Text** if it has the **highest Count(Text, Pattern)** among all k-mers.
- **Examples**
- **Text:** ACAACTATGCATACTATCGGGAACTATCCT
  Most frequent 5-mer: ACTAT
- **Text:** CGATATATCCATAG
  Most frequent 3-mer: ATA
- **STOP & THINK**
  - Can a string have multiple most frequent k-mers?
  - Do we have a well define computational problem?
- **Frequent Words Problem**
  - **Input:** A **string (Text) ,** An **integer (k)**
  - **Output:** All most frequent k-mers in Text

DNA Complementary Strands & Directionality

- **Base Pairing Rules**
  - **A ↔ T** (Adenine pairs with Thymine)
  - **G ↔ C** (Guanine pairs with Cytosine)
- **DNA Replication**
  - A **template strand** guides the synthesis of its **complementary strand** using free nucleotides.
  - **Meselson & Stahl (1958)** confirmed this **semi-conservative replication** model.
- **Example**
  - **Template Strand:** AGTCGCATAGT
    Complementary Strand: ACTATGCGACT



Directionality in DNA

•DNA strands have **directionality** labeled **5' → 3'** (Five Prime to Three Prime).

•**Complementary strands run in opposite directions** (antiparallel).

•A strand **must be read in the 5' → 3' direction**.

🧬 **Correct reading direction:**

•**Template Strand (5' → 3')** → **AGTCGCATAGT**

•**Complementary Strand (5' → 3')** → **TCAGCGTATCA**

DnaA Boxes in Vibrio cholerae: Finding Patterns

- **Observation**
- Among the four most frequent **9-mers** in the **ori region** of *Vibrio cholerae*, **ATGATCAAG** and **CTTGATCAT** are reverse complements of each other.
- Both occur **six times** in the **ori region**.
- **Importance of 9-mers**
- Finding a **9-mer** that appears **six times** in a DNA sequence of length 500 is unexpected and notable.
- This suggests that **ATGATCAAG** and **CTTGATCAT** are likely **DnaA boxes**.

atcaatgatcaacgtaagcttctaagcATGATCAAGgtgctcacacagtttatccacaac
ctgagtggatgacatcaagataggtcgttgtatctccttcctctcgtactctcatgacca
cggaaagATGATCAAGagaggatgatttcttggccatatcgcaatgaatacttgtgactt
gtgcttccaattgacatcttcagcgccatattgcgctggccaaggtgacggagcgggatt
acgaaagcatgatcatggctgttgttctgtttatcttgttttgactgagacttgttagga
tagacggtttttcatcactgactagccaaagccttactctgcctgacatcgaccgtaaat
tgataatgaatttacatgcttccgcgacgatttacctCTTGATCATcgatccgattgaag
atcttcaattgttaattctcttgcctcgactcatagccatgatgagctCTTGATCATgtt
tccttaaccctctatttttacggaagaATGATCAAGctgctgctCTTGATCATcgtttc

Biological Relevance

•The **DnaA protein** binds to **DnaA boxes** to initiate replication, regardless of the strand's orientation.

•Therefore, **both ATGATCAAG and CTTGATCAT** could represent **DnaA boxes**.

Next Step

•**Check if these sequences** appear in other regions of the Vibrio cholerae genome to confirm if they are unique to the ori region.

Looking for Hidden Messages in Multiple Genomes

- **Key Question**
- Can **ATGATCAAG** and **CTTGATCAT** be universal DnaA boxes across bacterial genomes?
- **Thermotoga petrophila Genome**
- A bacterium that thrives in **extremely hot environments** (temperature > 80°C).
- **Proposed ori region:**

```
aactctatacctccttttttgtcgaatttgtgtgatttatagagaaaatcttattaactga
aactaaaatggtaggtttggtggtaggttttgtgtacattttgtagtatctgatttttaa
ttacataccgtatattgtattaaattgacgaacaattgcatggaattgaatatatgcaaa
acaaacctaccaccaaactctgtattgaccattttaggacaacttcagggtggtaggttt
ctgaagctctcatcaatagactattttagtctttacaaacaatattaccgttcagattca
agattctacaacgctgttttaatgggcgttgcagaaaacttaccacctaaaatccagtat
ccaagccgatttcagagaaacctaccacttacctaccacttacctaccacccgggtggta
agttgcagacattattaaaaacctcatcagaagcttgttcaaaaatttcaatactcgaaa
cctaccacctgcgtcccctattatttactactactaataatagcagtataattgatctga
```

Finding Results

•**No occurrences** of **ATGATCAAG** or **CTTGATCAT** in the proposed ori region.

•This suggests that different bacteria may have **different DnaA boxes**.

Frequent 9-mers in Thermotoga petrophila ori

•**Six 9-mers** appear 3 or more times:

AACCTACCA	AAACCTACC	ACCTACCAC
CCTACCACC	GGTAGGTTT	TGGTAGGTT

Conclusion

•**DnaA boxes** are **bacteria-specific** and may vary across different species.

The Clump Finding Problem

- **Goal**
  - Instead of focusing on finding **specific hidden messages** (like ATGATCAAG or CTTGATCAT), let's look for **any k-mer that forms a clump** in the genome.
  - This approach can help find **ori regions** that may use different hidden messages in different bacterial genomes.
- **Clump Definition**
  - A **k-mer Pattern** forms an **(L, t)-clump** inside a string **Genome** if there is an interval of length **L** in **Genome** where this k-mer appears **at least t times**.
- **Example**
  - **Pattern**: TGCA
  - **Genome**:
    gatcagcataagggtccC**TGCA**A**TGCA**TGACAAGCC**TGCA**GTtgttttac
  - **Clump**: TGCA forms a **(25, 3)-clump** in the genome because it appears **3 times** within a short interval of length 25.
- **Known Clumps**
  - **Vibrio cholerae genome**: ATGATCAAG forms a **(500, 3)-clump**.
  - **Thermotoga petrophila genome**: CCTACCACC forms a **(500, 3)-clump**.