CS4049
Bioinformatics

Spring 2025

Rushda Muneer

Types of databases

DNA & RNA
genes, genomes & variation

Gene expression
RNA, protein & metabolite expression

Proteins
sequences, families & motifs

Structures
Molecular & cellular structures

Systems
reactions, interactions & pathways

Chemical biology
chemogenomics & metabolomics

Ontologies
taxonomies & controlled vocabularies

Literature
Scientific publications & patents

Other software
cross-domain tools & resources

Biological Database

A database is an organized collection of related biological data, that can be easily stored, accessed and managed

1 **Bibliographic Database**

Contains article and research papers of different journals **Pubmed**

2 **Sequence Database**

Contains protein and nucleotide sequence **GenBank, DDBJ, PIR**

3 **Structure Database**

Contains 3D structure of proteins and nucleic acids **PDB**

4 **Metabolic Database**

Contains data about various biological pathways **KEGG MetaCyc**

5 **Model organism Database**

Contains indepth biological data of studied model organism. **Flybase, RGD**

6 **Enzyme Database**

Contains data about structure and function of various enzymes **BRENDA**

7 **Disease Database**

Disease related information

OMIM

8 **Chemical Database**

Data on several small organic molecules **PubChem**

9 **Microarray Database**

Gene expression data from microarray experiments **GEO**

10 **Taxonomic Database**

Database that provides information on earths species of animals, plants

Catalogue of life

www.biologyexams4u.com

Nucleotide Databases

EMBL

European Molecular Biology Laboratory

ndb NUCLEIC ACID DATABASE

NCBI

GenBank

GSA

Genome Sequence Archive

NCBI

dbSNP

DDBJ

DNA Data Bank of Japan

An official website of the United States government Here's how you know

National Library of Medicine
National Center for Biotechnology Information

Log in

All Databases | Search

- NCBI Home
- Resource List (A-Z)
- All Resources
- Chemicals & Bioassays
- Data & Software
- DNA & RNA
- Domains & Structures
- Genes & Expression
- Genetics & Medicine
- Genomes & Maps
- Homology
- Literature
- Proteins
- Sequence Analysis
- Taxonomy
- Training & Tutorials
- Variation

Welcome to NCBI

The National Center for Biotechnology Information advances science and health by providing access to biomedical and genomic information.

About the NCBI | Mission | Organization | NCBI News & Blog

Submit

Deposit data or manuscripts into NCBI databases

Download

Transfer NCBI data to your computer

Learn

Find help documents, attend a class or watch a tutorial

Develop

Use NCBI APIs and code libraries to build applications

Analyze

Identify an NCBI tool for your data analysis task

Research

Explore NCBI research and collaborative projects

Popular Resources

- PubMed
- Bookshelf
- PubMed Central
- BLAST
- Nucleotide
- Genome
- SNP
- Gene
- Protein
- PubChem

NCBI News & Blog

New! Introducing the Multiple Comparative Genome Viewer (MCGV) Beta Release

16 Jan 2025

NLM's NCBI is excited to introduce the

An updated bacterial and archaeal reference genome collection is available!

14 Jan 2025

EMBL-EBI home | Services | Research | Training | About us | EMBL-EBI

EMBL's European Bioinformatics Institute

EMBL-EBI

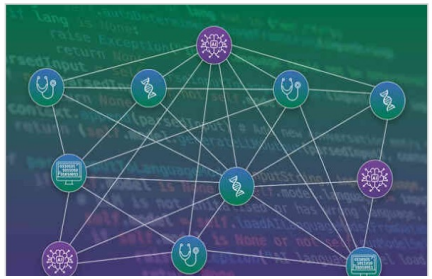
Unleashing the potential of big data in biology

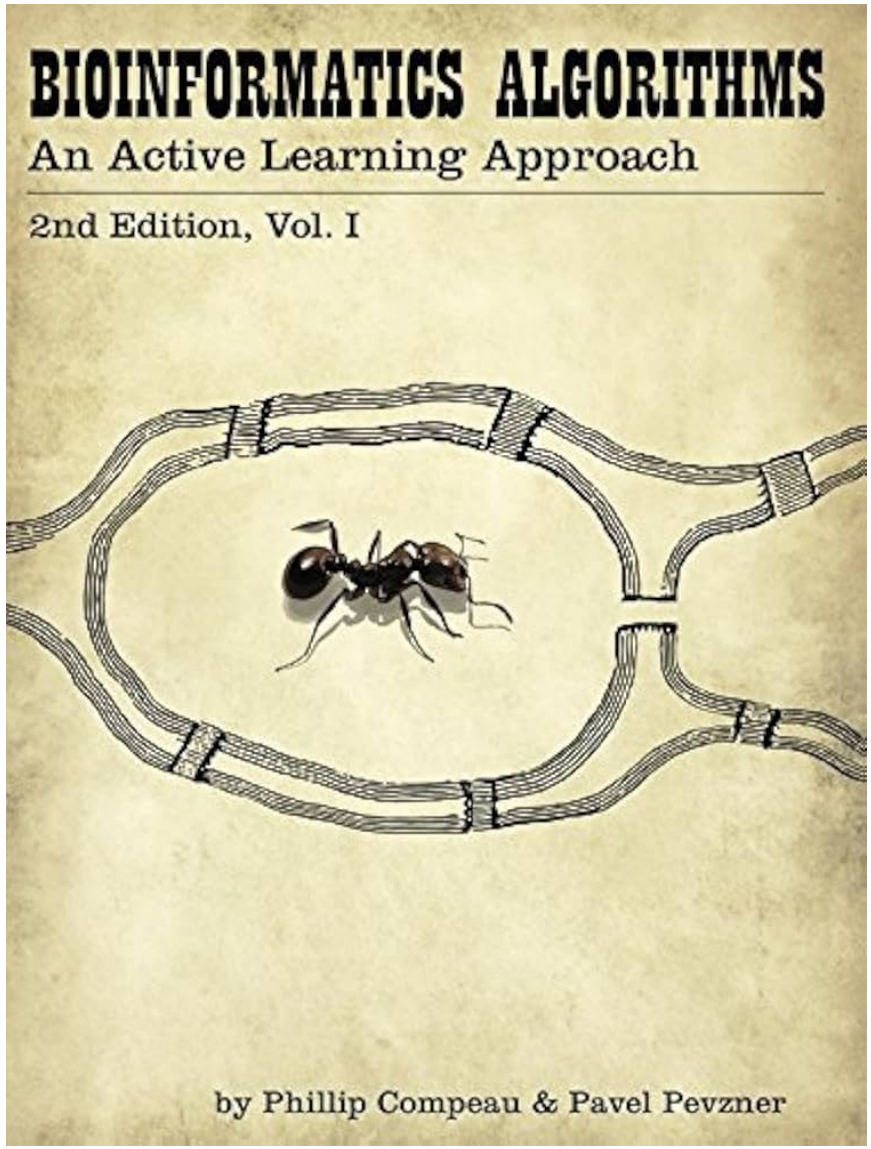
Find a gene, protein or chemical | All | Search

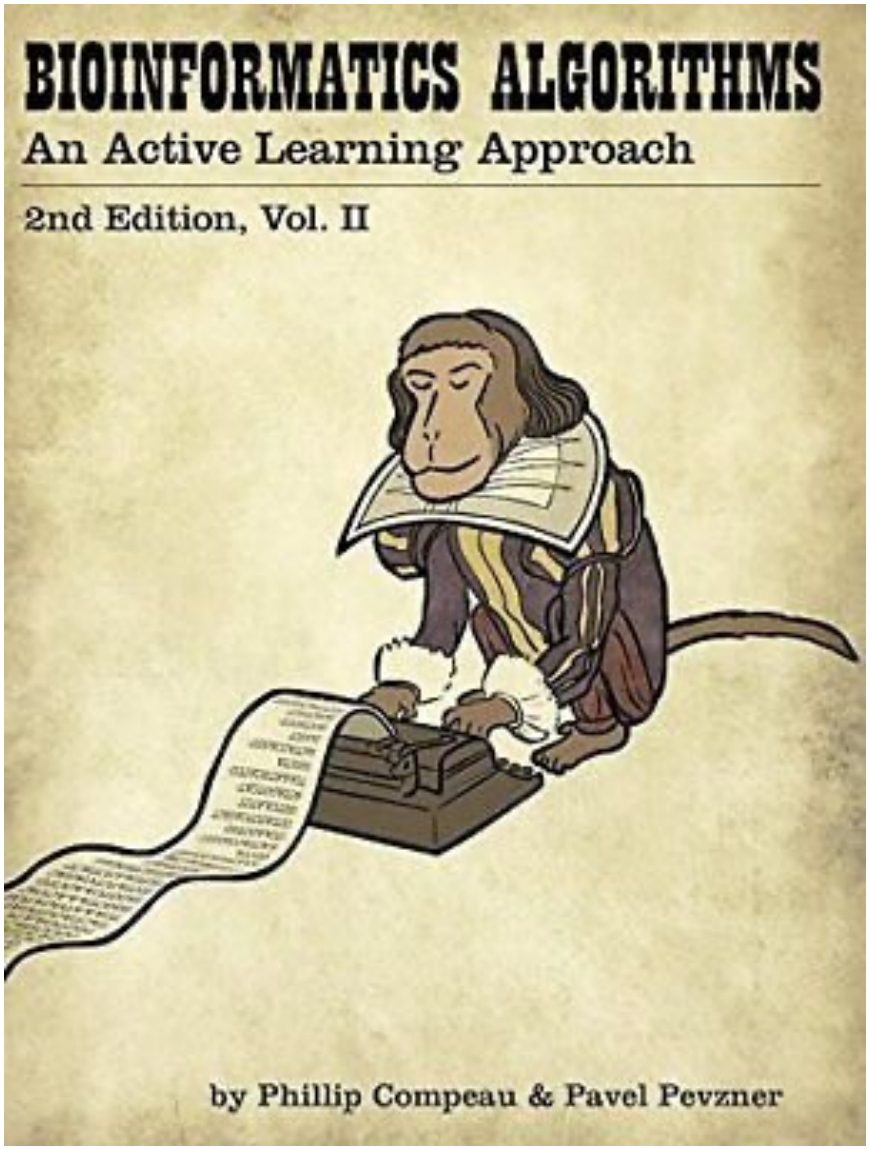
Example searches: blast keratin bfl1 | About EBI Search

Find data resources | Submit data | Explore our research | Train with us

Latest news




BIOINFORMATICS ALGORITHMS
An Active Learning Approach
2nd Edition, Vol. I
by Phillip Compeau & Pavel Pevzner


BIOINFORMATICS ALGORITHMS
An Active Learning Approach
2nd Edition, Vol. II
by Phillip Compeau & Pavel Pevzner

Circadian clock

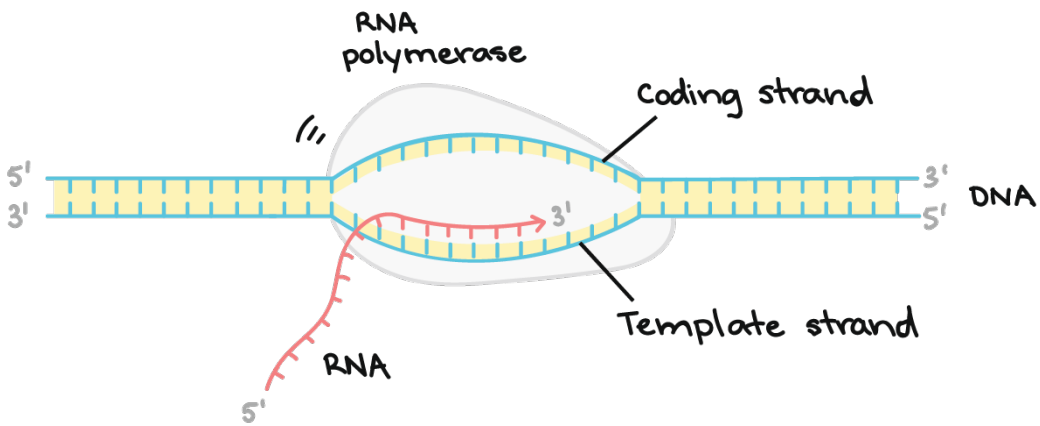
- The daily schedules of animals, plants, and bacteria are controlled by **an internal timekeeper** called the **circadian clock**.
- This clock regulates **activity** and **rest**.
- It functions at a **molecular level**, influencing **gene expression** and **protein production**.
- Mutations in circadian genes can lead to **disorders** like **Delayed Sleep-Phase Syndrome (DSPS)**.
- In plants, over 1,000 genes follow a circadian rhythm, regulating photosynthesis, light reception, and flowering.

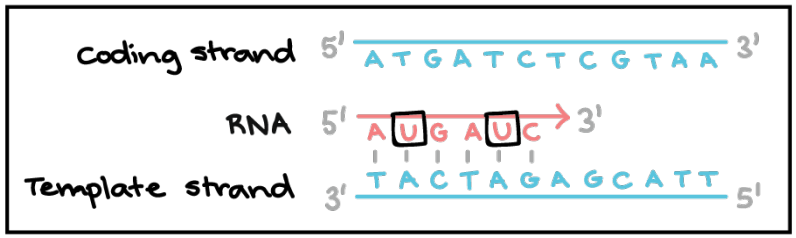
Gene Expression and Protein Regulation

- Genes encode proteins, which dictate cell function.
- **"Genes encode proteins"** means that a gene, which is a specific **sequence of DNA**, contains the **instructions** necessary to **build** a particular **protein**.
- Essentially acting as a blueprint for the protein's amino acid sequence, which determines its structure and function within a cell
- Cells control protein levels to respond to environmental changes.
- The flow of genetic information: DNA → RNA → Protein.
- Regulation occurs at two main stages:
  - **Transcription** (DNA to RNA)
  - **Translation** (RNA to Protein)

Transcription (DNA → RNA)

- The process of copying genetic information from DNA into a complementary RNA molecule.
- **Enzyme involved: RNA polymerase**
- **Process:**
  - **Initiation:** RNA polymerase binds to the **promoter** region of DNA.
  - **Elongation:** RNA polymerase moves along the DNA, synthesizing a **messenger RNA (mRNA)** strand.
  - **Termination:** RNA polymerase reaches a **terminator sequence**, releasing the mRNA.

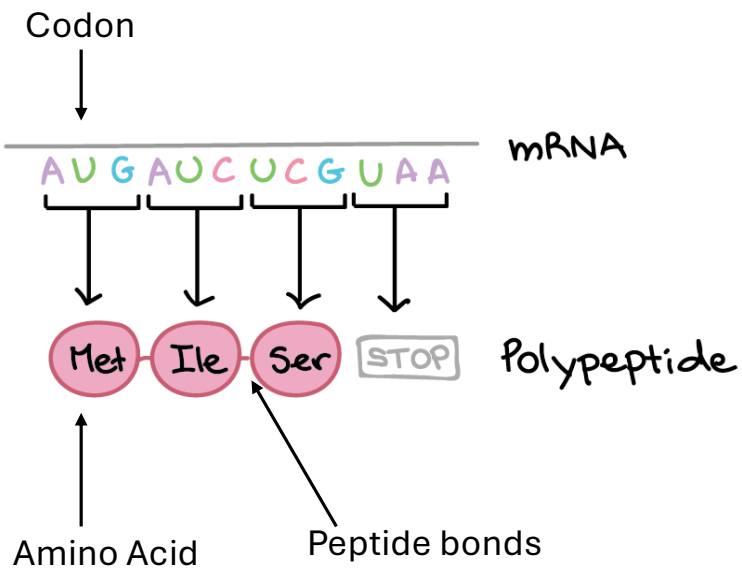


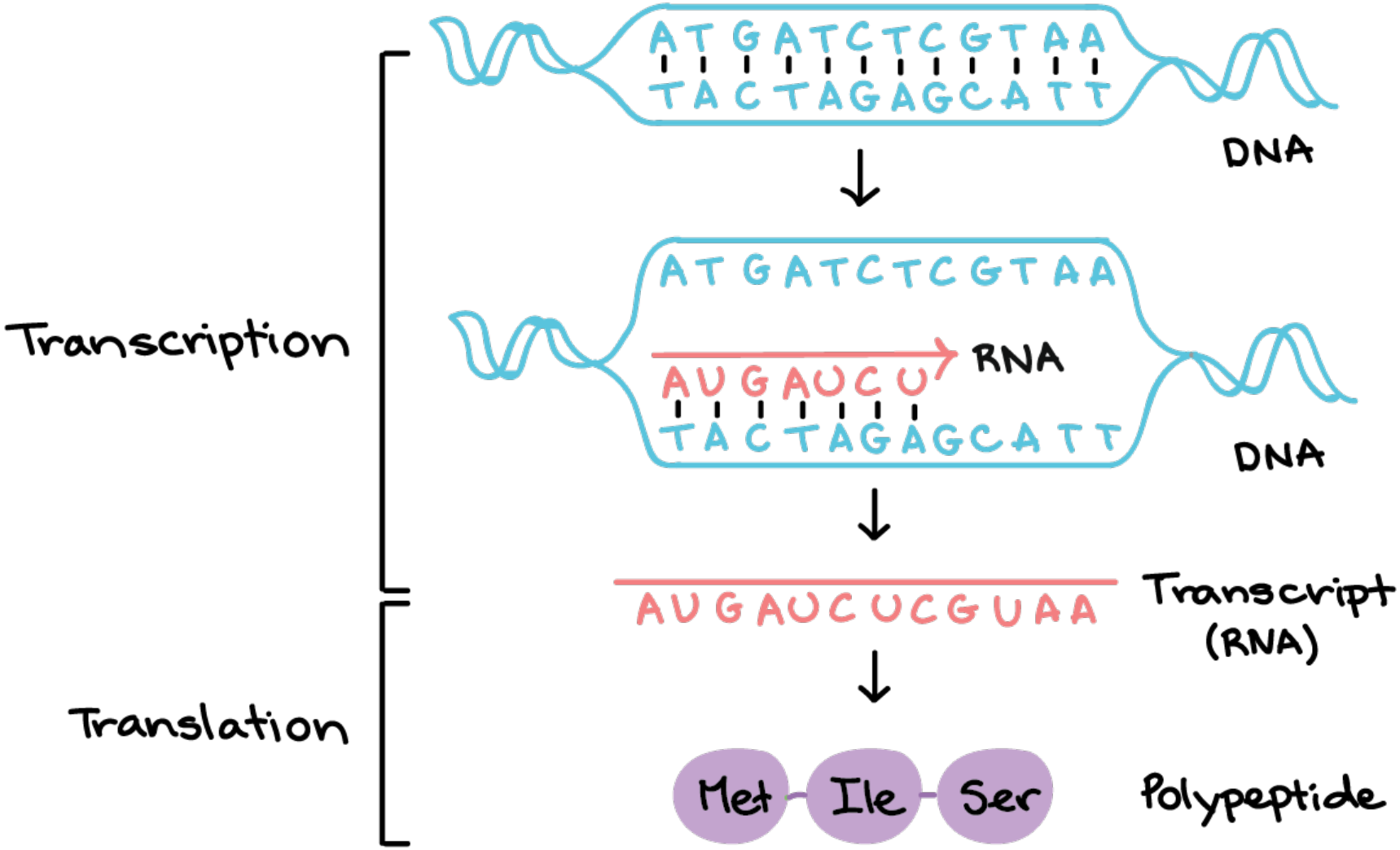


In RNA, Uracil (U) is used instead of Thymine (T) as a pair with Adenine (A)

Translation (RNA → Protein)

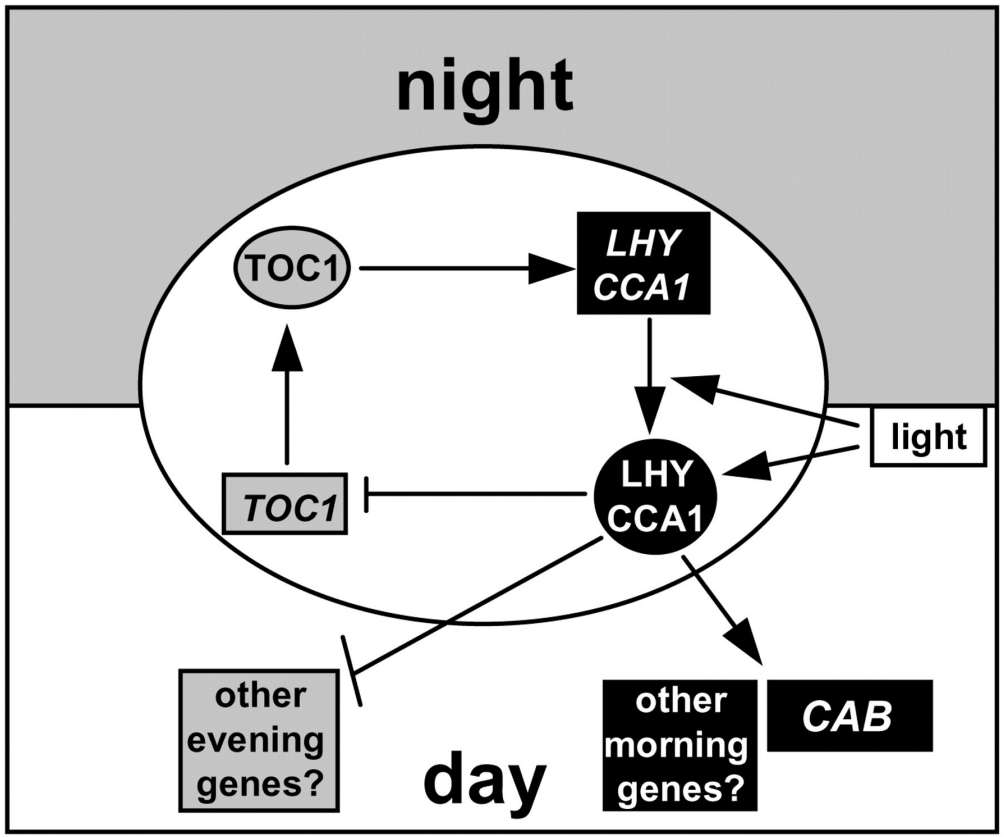
- The process of converting the mRNA sequence into a **polypeptide (protein)**.
- **Process:**
  - **Initiation:** The ribosome binds to the **start codon (AUG)** on the mRNA. (non-AUG codons can also occur based on the organism)
  - The ribosome reads **codons** (three-letter sequences).
  - Each codon matches a tRNA carrying a specific **amino acid**.
  - Amino acids are linked together by **peptide bonds**, forming a **polypeptide chain**.
  - **Termination:** When the ribosome reaches a **stop codon (UAA, UAG, UGA)**, translation ends, and the protein is released.




Transcription
ATGATCTCGTAA
TACTAGAGCATT
DNA
ATGATCTCGTAA
RNA
AUGAUCU
TACTAGAGCATT
DNA
AUGAUCUCGUAA
Transcript (RNA)
Translation
Met
Ile
Ser
Polypeptide

Plant Circadian Clock Regulation

- Each plant cell tracks day and night independently.
- Three key genes—LHY, CCA1, and TOC1—act as master timekeepers.
- **Regulation Mechanism:**
  - TOC1 promotes LHY & CCA1 expression at night.
  - LHY & CCA1 repress TOC1, forming a negative feedback loop during day.
  - Sunlight activates LHY & CCA1 in the morning, repressing TOC1.
  - As day continues LHY & CCA1 reduce repressing TOC1, eventually stopping.
  - At night, TOC1 activates increasing transcription, promoting LHY & CCA1 expression again in the morning.
  - The cycle continues.



Role of Transcription Factors:

- LHY, CCA1, and TOC1 **encode transcription factors (proteins)**.
- These transcription factors bind to specific **regulatory motifs** (short DNA sequences) in the **upstream region** of target genes (600–1000 nucleotides before the gene starts).
- **Example**: CCA1 transcripted proteins binds **AAAAAATCT** but can also bind to variant sequences (e.g. **AAGAACTCT**).
- **Motif Discovery in Bioinformatics:**
  - Regulatory motifs are **not** always **perfectly conserved**.
  - **Algorithms** are required to find these motifs **without prior knowledge**.
  - Motif finding helps identify **hidden regulatory sequences** shared across genes.

Identifying the Evening Element

- **Steve Kay's Research (2000):**
- Used DNA arrays to analyze gene activation in *Arabidopsis thaliana (a small plant from mustard family)*.
- Identified nearly 500 genes with circadian behavior.
- Extracted **upstream regions** to search for **common patterns**.
- The sequence "**AAAATATCT**" appeared **46 times** in the upstream regions.
- Suggests a potential regulatory motif for circadian control.
- Kay named "**AAAATATCT**" the **evening element**
- Performed a simple experiment to prove that regulatory motif is responsible for the circadian gene expression.
- He mutated the evening element in the upstream region of one gene
- As a result, the gene lost its circadian behavior (Motif Conservation).

NF-κB Binding Sites in Drosophila (Fruit Fly)

- If you infect a fly with a bacterium, the fly will switch on its immunity genes to fight the infection.
- **NF-κB activates immunity genes** in response to infection.
- Some immunity genes share a 12-mer sequence similar to "TCGGGGATTTCC"
- NF-κB binding sites show **more variability** compared to the **evening element**.
- Identifying such variable motifs requires **advanced computational methods**.

1	T	C	G	G	G	G	g	T	T	T	t	t
2	c	C	G	G	t	G	A	c	T	T	a	C
3	a	C	G	G	G	G	A	T	T	T	t	C
4	T	t	G	G	G	G	A	c	T	T	t	t
5	a	a	G	G	G	G	A	c	T	T	C	C
6	T	t	G	G	G	G	A	c	T	T	C	C
7	T	C	G	G	G	G	A	T	T	c	a	t
8	T	C	G	G	G	G	A	T	T	c	C	t
9	T	a	G	G	G	G	A	a	c	T	a	C
10	T	C	G	G	G	t	A	T	a	a	C	C

- The ten candidate NF-κB binding sites appearing in the *Drosophila melanogaster* genome.
- The **colored upper-case letters indicate the most frequent nucleotide** in each column.

Hide and Seek with Motifs

- **Turning a Biological Challenge into a Computational Problem**
  - **Goal:** Identify regulatory motifs in DNA sequences.
  - Example: A **15-mer sequence** is hidden in **10 randomly generated DNA strings.**
  - Mimics a transcription factor binding site in gene upstream regions.

```
 1 "atgaccgggatactgataaaaaaaaagggggggggcgtacacattagataaacgtatgaagtacgttagactcggcgccgccg"
 2 "acccctatttttgagcagatttagtgacctggaaaaaaaatttgagtacaaaacttttccgaataaaaaaaaagggggggga"
 3 "tgagtatccctgggatgacttaaaaaaaagggggggtgctctcccgatttttgaatatgtaggatcattcgccagggtccga"
 4 "gctgagaattggatgaaaaaaaagggggggtccacgcaatcgcgaaccaacgcggacccaaaggcaagaccgataaaggaga"
 5 "tccctttgcggtaatgtgccgggaggctggttacgtagggaagccctaacggacttaataaaaaaaagggggggcttatag"
 6 "gtcaatcatgttcttgtgaatggatttaaaaaaaagggggggaccgcttggcgcacccaaattcagtgtgggcgagcgcaa"
 7 "cggttttggcccttgttagaggcccccgtaaaaaaaagggggggcaattatgagagagctaatctatcgcgtgcgtgttcat"
 8 "aacttgagttaaaaaaaagggggggctggggcacatacaagaggagtcttccttatcagttaatgctgtatgacactatgta"
 9 "ttggcccattggctaaaagcccaacttgacaaatggaagatagaatccttgcataaaaaaaagggggggaccgaaagggaag"
10 "ctggtgagcaacgacagattcttacgtgcattagctcgcttccggggatctaatagcacgaagcttaaaaaaaagggggggga"
```

Frequent Words Problem Algorithm

- Applying an algorithm for the Frequent Words Problem reveals the most frequent 15-mer.
- The implanted pattern is identified as the most frequent 15-mer.
- The short strings are randomly generated, so other frequent 15-mers are unlikely to exist.

```
1 "atgaccgggatactgatAAAAAAAAGGGGGGGggcgtacacattagataaacgtatgaagtacgttagactcggcgccgccg"
2 "acccctatttttgagcagatttagtgacctggaaaaaaaatttgagtacaaaacttttccgaataAAAAAAAAGGGGGGGa"
3 "tgagtatccctgggatgacttAAAAAAAAGGGGGGGtgctctcccgatttttgaatatgtaggatcattcgccagggtccga"
4 "gctgagaattggatgAAAAAAAAGGGGGGGtccacgcaatcgcgaaccaacgcggacccaaaggcaagaccgataaaggaga"
5 "tccctttgcggtaatgtgccgggaggctggttacgtagggaagccctaacggacttaatAAAAAAAAGGGGGGGcttatag"
6 "gtcaatcatgttcttgtgaatggatttAAAAAAAAGGGGGGGgaccgcttggcgcacccaaattcagtgtgggcgagcgcaa"
7 "cggttttggcccttgttagaggcccccgtAAAAAAAAGGGGGGGcaattatgagagagctaatctatcgcgtgcgtgttcat"
8 "aacttgagttAAAAAAAAGGGGGGGctggggcacatacaagaggagtcttccttatcagttaatgctgtatgacactatgta"
9 "ttggcccattggctaaaagcccaacttgacaaatggaagatagaatccttgcatAAAAAAAAGGGGGGGaccgaaagggaag"
10 "ctggtgagcaacgacagattcttacgtgcattagctcgcttccggggatctaatagcacgaagcttAAAAAAAAGGGGGGGa"
```

Handling Mutated Patterns

- **Scenario**:
- Instead of implanting the exact same pattern, we **mutate the implanted pattern** by randomly changing nucleotides at **four selected positions** in each sequence.
- **Example:** "**AAAAAAAAGGGGGGG**" no longer appears in the sequences due to mutations.

```
 1 "atgaccgggatactgatAgAAgAAAGGttGGGggcgtacacattagataaacgtatgaagtacgttagactcggcgccgccg"
 2 "acccctattttttgagcagatttagtgacctggaaaaaaaatttgagtacaaaacttttccgaatacAAtAAAAcGGcGGGa"
 3 "tgagtatccctgggatgacttAAAAtAAtGGaGtGGtgctctcccgatttttgaatatgtaggatcattcgccagggtccga"
 4 "gctgagaattggatgcAAAAAAAGGGattGtccacgcaatcgcgaaccaacgcggacccaaaggcaagaccgataaaggaga"
 5 "tccctttgcggtaatgtgccgggaggctggttacgtagggaagccctaacggacttaatAtAAtAAAGGaaGGGcttatag"
 6 "gtcaatcatgttcttgtgaatggatttAAcAAtAAGGGctGGgaccgcttggcgcacccaaattcagtgtgggcgagcgcaa"
 7 "cggttttggcccttgttagaggcccccgtAtAAAcAAGGaGGGccaattatgagagagctaatctatcgcgtgcgtgttcat"
 8 "aacttgagttAAAAAAtAGGGaGccctggggcacatacaagaggagtcttccttatcagttaatgctgtatgacactatgta"
 9 "ttggcccattggctaaaagcccaacttgacaaatggaagatagaatccttgcatActAAAAAGGaGcGGaccgaaagggaag"
10 "ctggtgagcaacgacagattcttacgtgcattagctcgcttccggggatctaatagcacgaagcttActAAAAAGGaGcGGa"
```

Limitations of the Frequent Words Problem

- Applying the **Frequent Words Problem algorithm** is not helpful, as the mutated pattern doesn't appear.
- The **Frequent Words with Mismatches algorithm** may be an option but becomes **slow with longer motifs** and **more mutations**.
- **Biological Insight**:
- The Frequent Words Problem doesn't model the biological motif-finding process accurately.
  - **DnaA boxes**: Patterns that frequently appear within a short genome interval.
  - **Regulatory motifs**: Can appear scattered throughout the genome, with variation.

A brute force algorithm for Motif Finding

- **What is Brute Force?**
- A general problem-solving technique that explores all possible solution candidates.
- While **easy to design** and **guaranteed to find the correct solution**, it may take an **immense amount of time** due to the **large number of candidates** to check.
- **Brute Force for the Implanted Motif Problem**:
- **Observation**: Any (k, d)-motif is **at most d mismatches** away from some **k-mer** in the **first string** in the dataset.
- **Approach**:
  - Generate **all possible k-mers** from the **first** string.
  - Check which of these **k-mers** are (k, d)-motifs by comparing with the **rest of the strings.**

Problems with Brute Force Motif Finding

- **Slow** for large k and d.
- **Pairwise Similarity Issue:** Finding similar k-mers in DNA may not reveal **implanted pattern**.
- **Example:**
- Implanted 15-mers "**AgAAgAAAGGttGGG**" and "**cAAtAAAAcGGGGcG**", each of which differs from correct motif "**AAAAAAAAGGGGGGG**" by **4 mismatches**.
- These implanted motifs may have **8 mismatches** between them, making detection difficult.

```
AgAAgAAAGGttGGG
||  ||    |  ||  |
cAAtAAAAcGGGGcG
```

Problems with Brute Force Motif Finding

- **Subtle Motif Problem:**
  - A 15-mer implanted with **4 mutations** in 10 sequences (600 nucleotides each).
  - **Pairwise comparison fails** as thousands of random 15-mers are < 8 mismatches apart.
  - Prevents us from identifying the true implanted motifs by pairwise comparisons
- **Conclusion:**
  Pairwise similarity is unreliable for detecting subtle motifs, requiring more advanced algorithms.

Your Task

- Think about how this problem can be resolved?
- Discuss in the next class.