Bioinformatics: An Introduction

Swakkhar Shatabda

Department of Computer Science and Engineering
BRAC University

Your Course Teacher

- Swakkhar Shatabda, Professor
  - BSc in CSE, BUET 2007
  - PhD, Griffith University, Australia, 2014
- Contact
  - Email: swakkhar.shatabda@bracu.ac.bd
  - Room: 4G#28
  - Counselling Time: Sun-Mon-Tue-Wed
- Research Interests:
  - Applied Machine Learning
  - Artificial Intelligence Search
  - Computational Biology

BRAC UNIVERSITY
Inspiring Excellence

Course Outline

1. Introduction to Bio-informatics
2. Sequence Alignment, Dynamic Programming
3. Hidden Markov Models, HMM Profile, Coding Regions
4. Genome Assembly: Read Mapping, Variant Calling, De-Bruijn Graphs, Suffix Trees
5. Gene Expression Analysis: Clustering, Visualization, Dimensionality Reduction
6. Single Cell RNA-Seq and Spatial Transcriptomics
7. Phylogenetic Trees
8. Biological Networks
9. Gene Ontology and Biological Pathways
10. Evolutionary Genomics, Comparative Genomics
11. Representation Learning, Contrastive Learning and Alignment in Bioinformatics
12. Drug and Protein Structures, Proteomics, Generative AI

BRAC UNIVERSITY
Inspiring Excellence

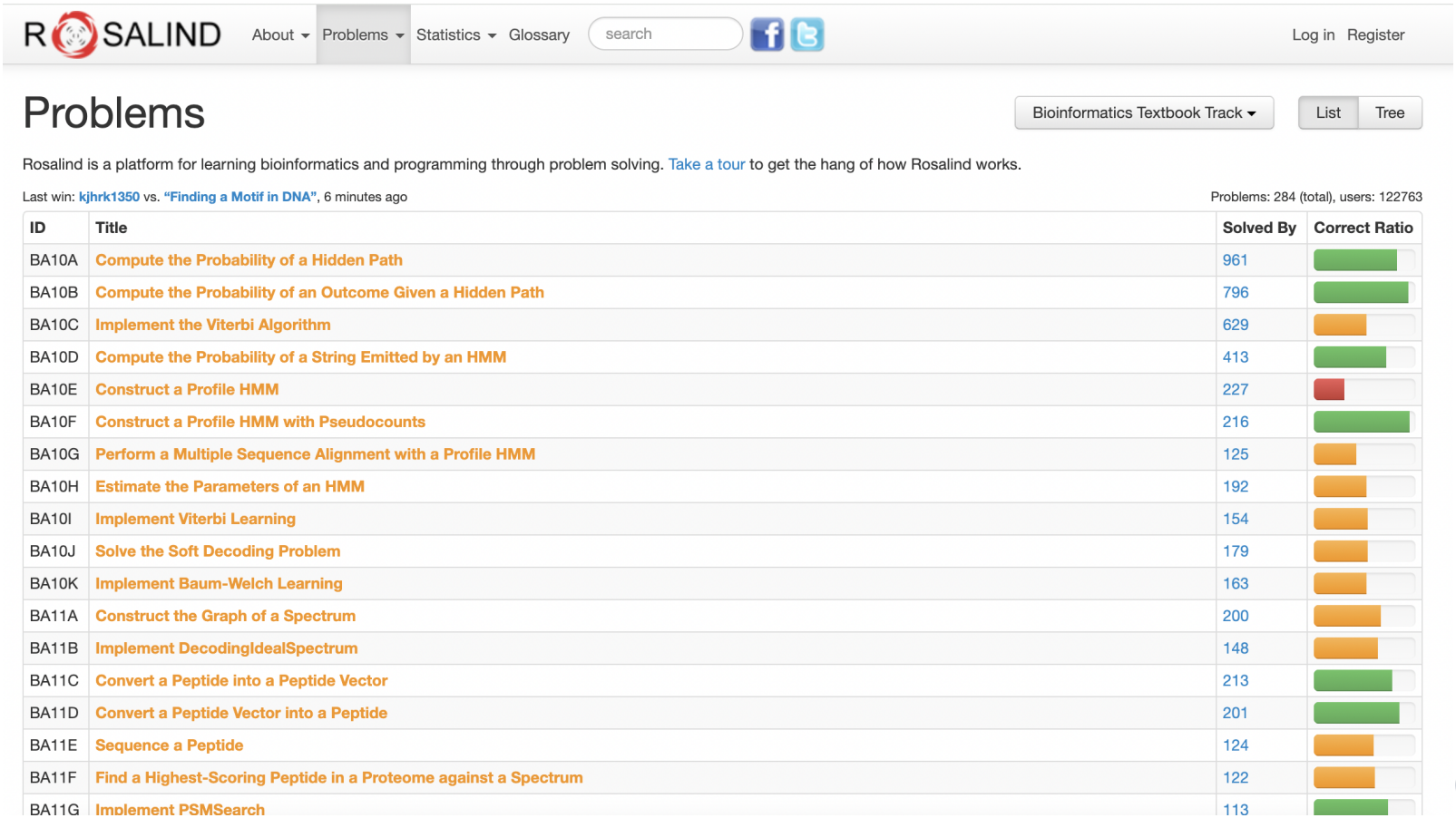
Assessments

- Assignments - 10%
  - 2 Assignments (Programming)
- Quiz - 15%
  - Best 3 out of 4
- Mid Exam - 25%
- Final Exam - 30%
- Term Project - 20%
- For any missed exams (mid and final), please follow the procedure from the department.
- No makeup exams will taken taken for quiz.

BRAC UNIVERSITY
Inspiring Excellence

Problem Solvers!

https://rosalind.info/problems/list-view/?location=bioinformatics-textbook-track



ROSALIND About Problems Statistics Glossary search Log in Register

Problems

Bioinformatics Textbook Track | List | Tree

Rosalind is a platform for learning bioinformatics and programming through problem solving. Take a tour to get the hang of how Rosalind works.

Last win: kjhrk1350 vs. "Finding a Motif in DNA", 6 minutes ago

Problems: 284 (total), users: 122763

ID	Title	Solved By	Correct Ratio
BA10A	Compute the Probability of a Hidden Path	961	
BA10B	Compute the Probability of an Outcome Given a Hidden Path	796	
BA10C	Implement the Viterbi Algorithm	629	
BA10D	Compute the Probability of a String Emitted by an HMM	413	
BA10E	Construct a Profile HMM	227	
BA10F	Construct a Profile HMM with Pseudocounts	216	
BA10G	Perform a Multiple Sequence Alignment with a Profile HMM	125	
BA10H	Estimate the Parameters of an HMM	192	
BA10I	Implement Viterbi Learning	154	
BA10J	Solve the Soft Decoding Problem	179	
BA10K	Implement Baum-Welch Learning	163	
BA11A	Construct the Graph of a Spectrum	200	
BA11B	Implement DecodingIdealSpectrum	148	
BA11C	Convert a Peptide into a Peptide Vector	213	
BA11D	Convert a Peptide Vector into a Peptide	201	
BA11E	Sequence a Peptide	124	
BA11F	Find a Highest-Scoring Peptide in a Proteome against a Spectrum	122	
BA11G	Implement PSMSearch	113	

BRAC UNIVERSITY
Inspiring Excellence

Text Books and Materials

B1 Bioinformatics Algorithms: An Active Learning Approach Vol I&II Book by Pavel A. Pevzner and Phillip Compeau

B2 Genome-Scale Algorithm Design Bioinformatics in the Era of High-Throughput Sequencing, Second Edition, Makinen et al.

B3 Molecular Biology of the Cell, Bruce Alberts et al.

B4 A First Course in Systems Biology Book by Eberhard Voit

P1 Research Papers

H1 Technical Handouts

BRAC UNIVERSITY
Inspiring Excellence

Term Project

G Group Project (4 members maximum per group)
`https://forms.gle/YATyE5cC8y4teYwLA` (Group Confirmation)

T Select a contemporary topic to explore

P Understand the problem statement

L Read necessary literature

E Perform some experiments on data

R Report the findings.

Template and Topics

- Report Template:
  `https://www.overleaf.com/read/xmcdckpnpmms#437fed`
- Project Ideas: `https://docs.google.com/document/d/1dv0iuV0ekr7imHwgBRRm52nvLEgwtm5t/edit?usp=sharing&ouid=112757978227538035208&rtpof=true&sd=true`

Bioinformatics and Computational Biology

Merriam-Webster: Bioinformatics

Bioinformatics is the collection, classification, storage, and analysis of biochemical and biological information using computers especially as applied to molecular genetics and genomics.

ChatGPT: Bioinformatics

Bioinformatics is an interdisciplinary field that combines biology, computer science, mathematics, and statistics to analyze and interpret biological data, especially large-scale datasets like genomic sequences, protein structures, and gene expression profiles.

ChatGPT:Computational Biology

Computational Biology is a field that applies mathematical models, algorithms, and computational techniques to understand and simulate biological systems and processes. It focuses on biological problem-solving using computational methods, often involving modeling, simulations, and large-scale data analysis.

How Different Are We?

We are 99.9% Identical. What makes us diverse?

Genotype and Phenotype

- A gene is a segment of DNA that provides the cell with instructions for making a specific protein, which then carries out a particular function in your body.
- Nearly all humans have the same genes arranged in roughly the same order and more than 99.9% of your DNA sequence is identical to any other human.
- On average, a human gene will have 1-3 letters that differ from person to person.
- These differences are enough to change the shape and function of a protein, how much protein is made, when it's made, or where it's made.
- They affect the color of your eyes, hair, and skin. More importantly, variations in your genome also influence your risk of developing diseases and your responses to medications.

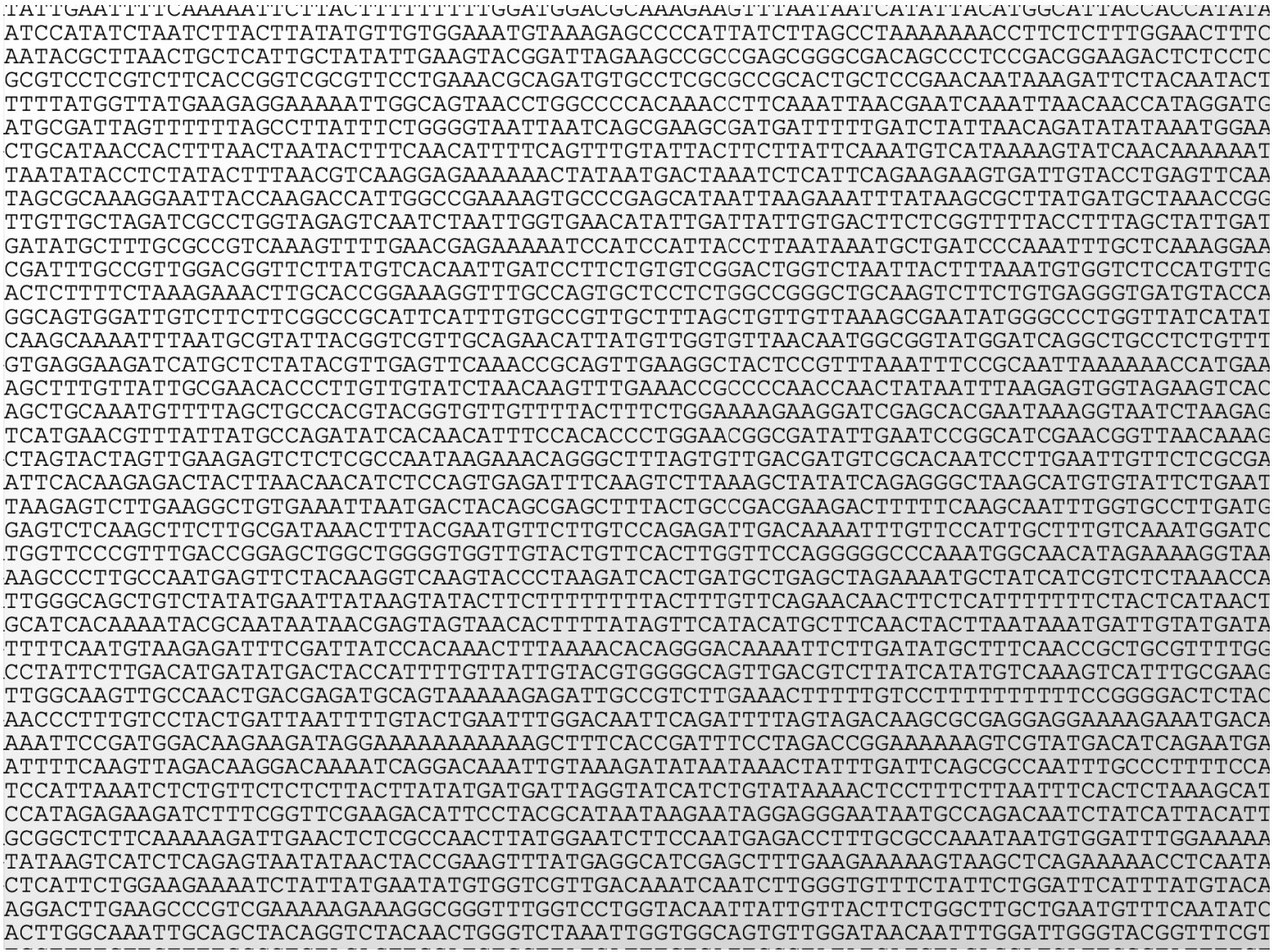
BRAC UNIVERSITY
Inspiring Excellence

Genome: Structure and Function

- Genome is a fancy word for all your DNA. From potatoes to puppies, all living organisms have their own genome.
- If all the DNA from a single human cell was stretched out end-to-end, it would make a six-foot-long strand comprised of a six billion letter code.
- The DNA in a cell is not a single long molecule. It is divided into a number of segments of uneven lengths. At certain points in the life cycle of a cell, those segments can be tightly packed bundles known as chromosomes.
- Half of your genome comes from your biological mother and half from your biological father, making you related to each, but identical to neither. Your biological parents' genes influence traits like height, eye color, and disease risk that make you a unique person.

BRAC UNIVERSITY
Inspiring Excellence

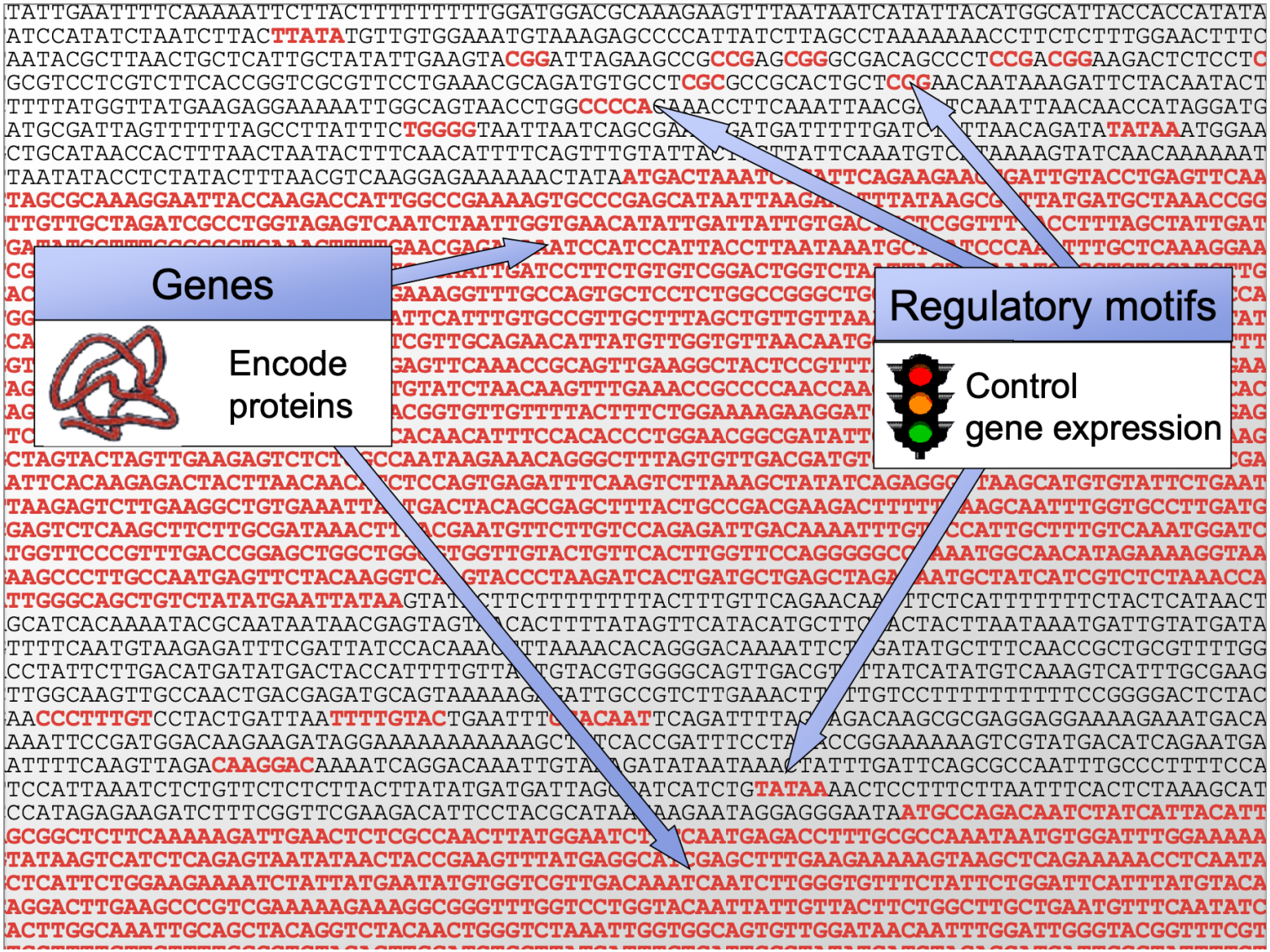
Yeast Genome



BRAC UNIVERSITY
Inspiring Excellence

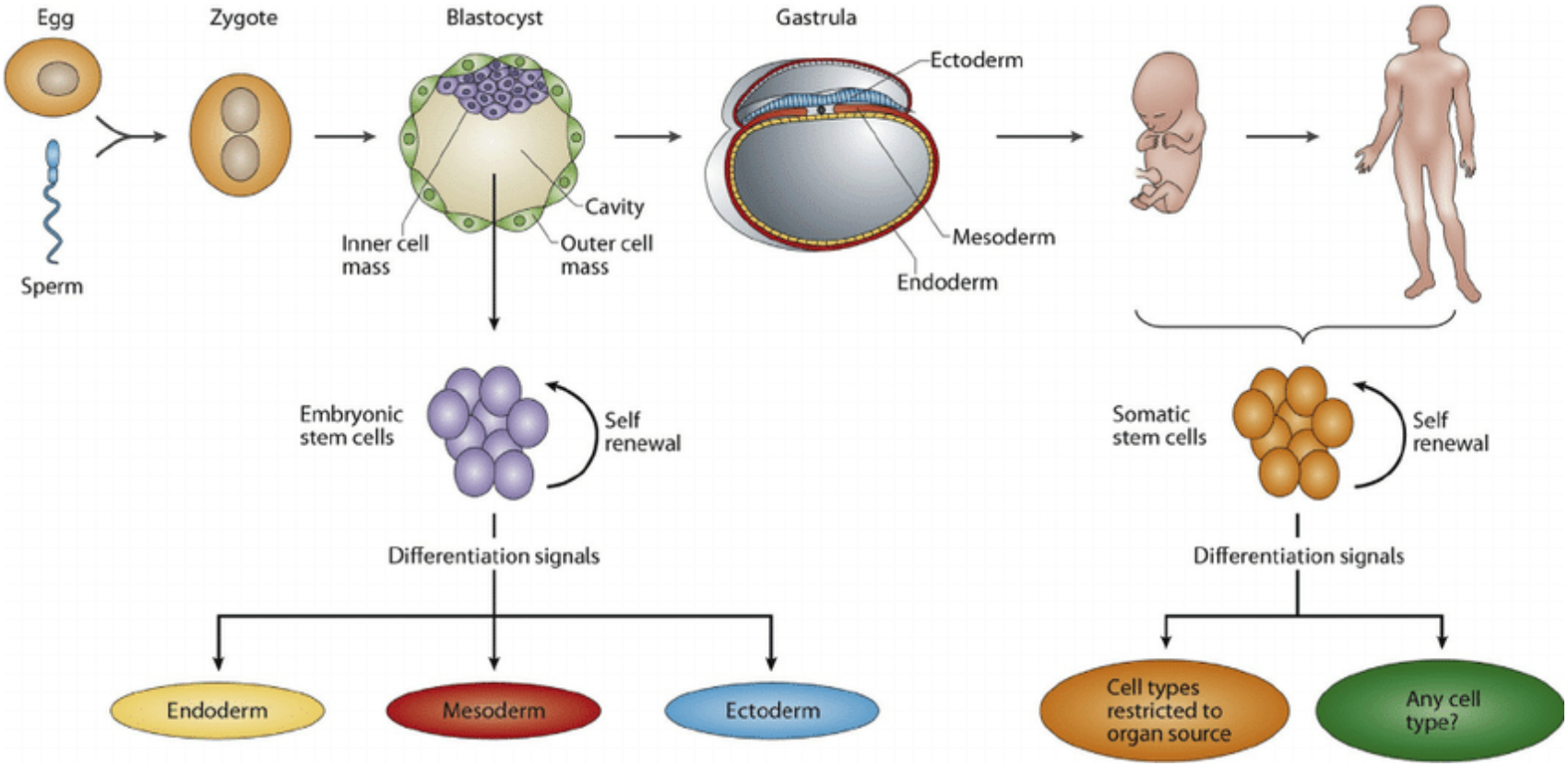
Yeast Genome

Yeast Genome



BRAC UNIVERSITY
Inspiring Excellence

How do our bodies develop?



- The same program, the same code, somehow it's executing a different program to make it into a neuron versus a skin cell.

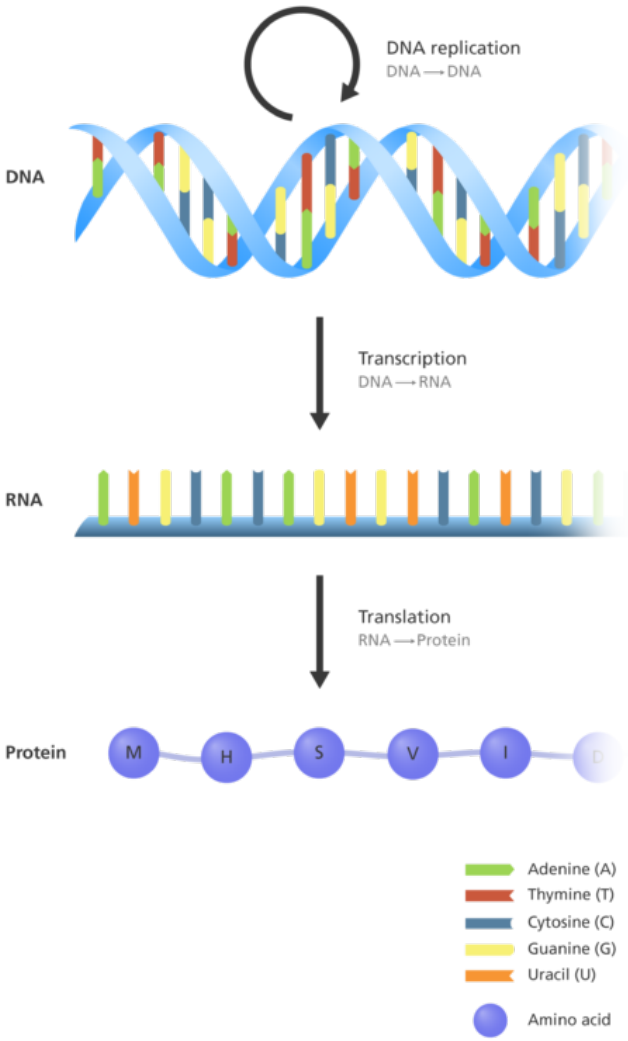
BRAC UNIVERSITY
Inspiring Excellence

Genetic Diseases

- Cancer, a common phenotype, that is, they all have a common feature that they're dividing without control.
- A genetic disease is caused by a change in the DNA sequence.
- Some diseases are caused by mutations that are inherited from the parents and are present in an individual at birth.
- Other diseases are caused by acquired mutations in a gene or group of genes that occur during a person's life.
- Changes in the DNA sequence are called genetic variants.
- The majority of the time genetic variants have no effect at all. But, sometimes, the effect is harmful: just one letter missing or changed may result in a damaged protein, extra protein, or no protein at all, with serious consequences for our health.
- Additionally, the passing of genetic variants from one generation to the next helps to explain why many diseases run in families

BRAC UNIVERSITY
Inspiring Excellence

Does Genome Determine Everything?



- The genome does not determine everything.
- While genes provide the blueprint for life, various epigenetic, environmental, and stochastic (random) factors influence how traits develop and function.

BRAC UNIVERSITY
Inspiring Excellence

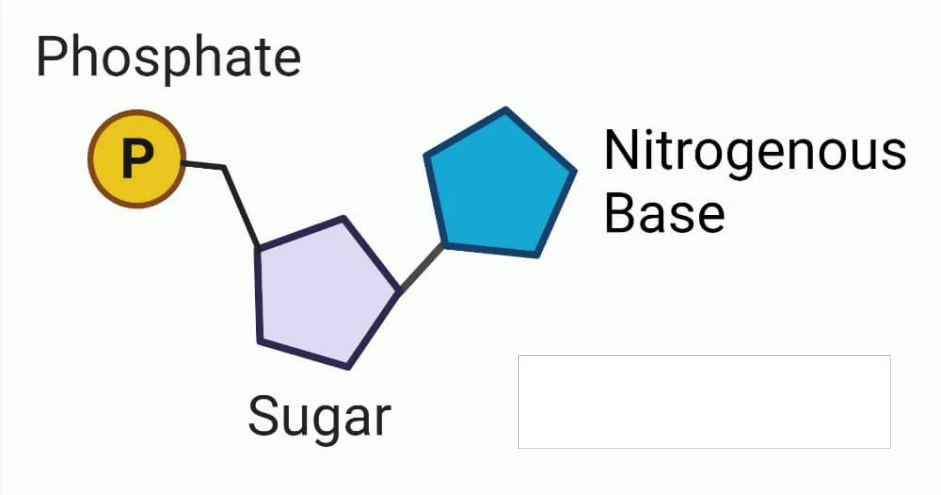
DNA

- The DNA molecule stores the genetic information of an organism.
- DNA contains regions called genes, which encode for proteins to be produced.
- Other regions of the DNA contain regulatory elements, which partially influence the level of expression of each gene.
- Within the genetic code of DNA lies both the data about the proteins that need to be encoded, and the control circuitry, in the form of regulatory motifs.

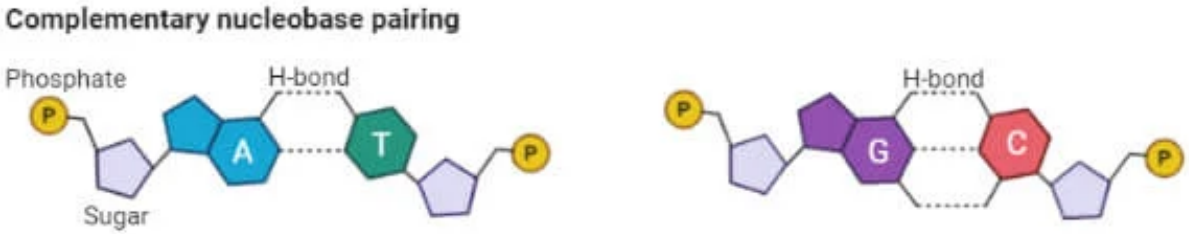
BRAC UNIVERSITY
Inspiring Excellence

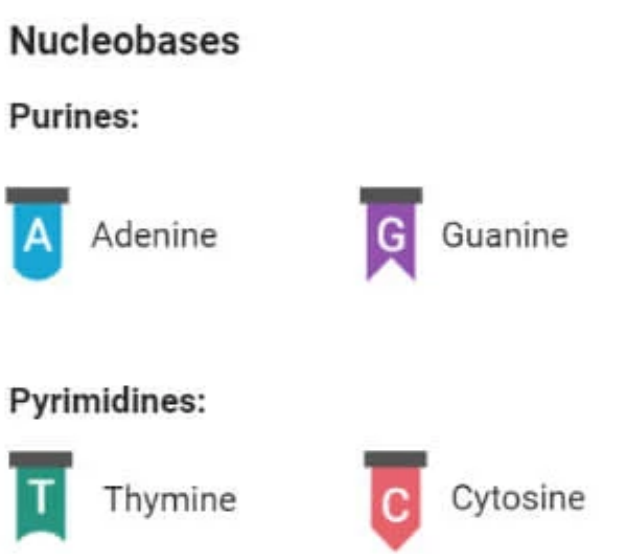
Nucleotides

- DNA is composed of four nucleotides: A(adenine), C(cytosine),T (thymine), and G (guanine).
- A and G are purines, which have two rings, while C and T are pyrimidines, with one ring.
- A and T are connected by two hydrogen bonds, while C and G are connected by three bonds. Therefore, the A-T pairing is weaker than the C-G pairing.



DNA - Structure



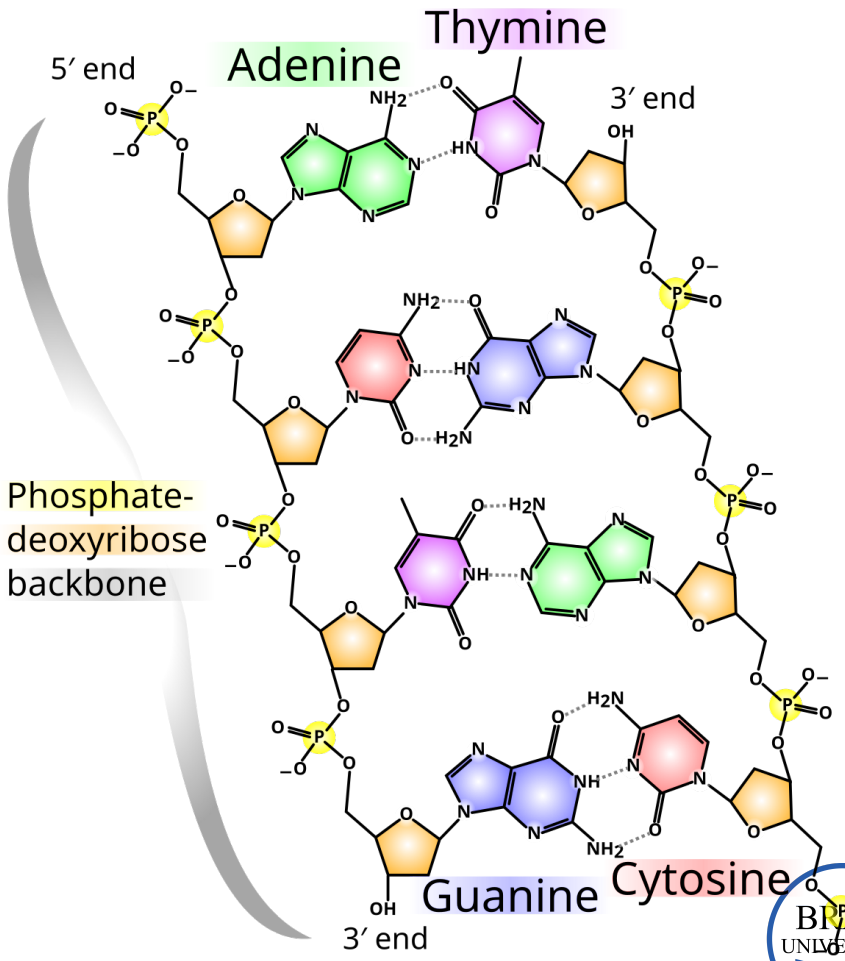


- The two DNA strands in the double helix are complementary, meaning that if there is an A on one strand, it will be bonded to a T on the other, and if there is a C on one strand, it will be bonded to a G on the other.

BRAC UNIVERSITY
Inspiring Excellence

DNA - Directionality

- The DNA strands also have directionality, which refers to the positions of the pentose ring where the phosphate backbone connects.
- This directionality convention comes from the fact that DNA and RNA polymerase synthesize in the 5' to 3' direction.



DNA - Replication

- The structure of DNA, with its weak hydrogen bonds between the bases in the center, allows the strands to easily be separated for the purpose of DNA replication (the capacity for DNA strands to be separated also allows for transcription, translation, recombination, and DNA repair, among others).
- This was noted by Watson and Crick as It has not escaped our notice that the specific pairing that we have postulated immediately suggests a possible copying mechanism for the genetic material.
- In the replication of DNA, the two complementary strands are separated, and each of the strands are used as templates for the construction of a new strand.
- DNA polymerases attach to each of the strands at the origin of replication, reading each existing strand from the 3' to 5' direction and placing down complementary bases such that the new strand grows in the 5' to 3' direction.

BRAC UNIVERSITY
Inspiring Excellence

DNA - Replicaiton

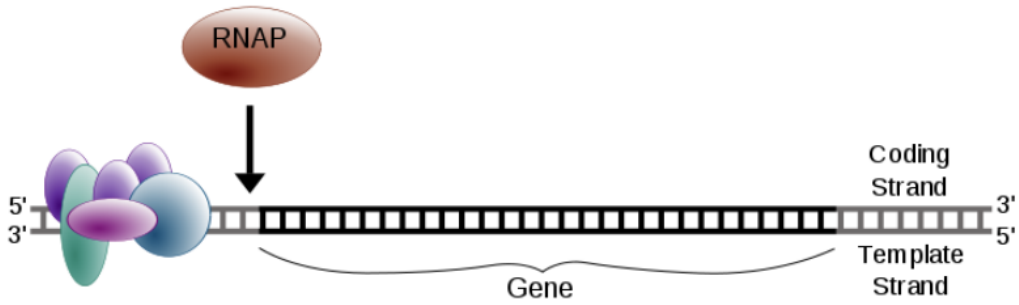
- Because the new strand must grow from 5' to 3', one strand (the leading strand) can be copied continuously, while the other (the lagging strand) grows in pieces which are later glued together by DNA ligase. The end result is 2 double-stranded pieces of DNA, where each is composed of 1 old strand, and 1 new strand.
- Genes can occur on either strand of DNA. The DNA before a gene (in the 5' region) is considered upstream whereas the DNA after a gene (in the 3' region) is considered downstream.

BRAC UNIVERSITY
Inspiring Excellence

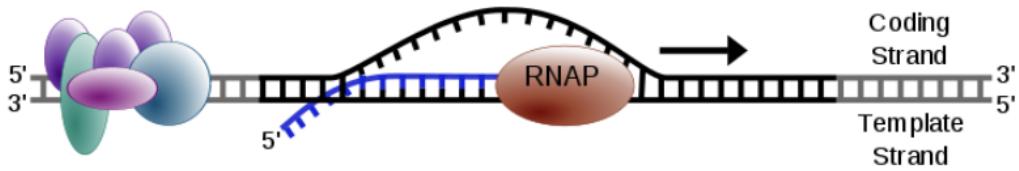
Transcription

- Transcription is the process by which RNA is produced using a DNA template.
- The DNA is partially unwound to form a bubble, and RNA polymerase is recruited to the transcription start site (TSS) by regulatory protein complexes.
- RNA polymerase reads the DNA from the 3' to 5' direction and placing down complementary bases to form messenger RNA (mRNA).
- RNA uses the same nucleotides as DNA, except Uracil is used instead of Thymine.
- mRNA in eukaryotes experience post-translational modifications, or processes that edit the mRNA strand further.
  - Splicing removes introns, intervening regions which don't code for protein, so that only the coding regions, the exons, remain.
  - The 5' end is capped with a modified guanine nucleotide. At the 3' end, adenine residues are added to form a poly(A) tail.

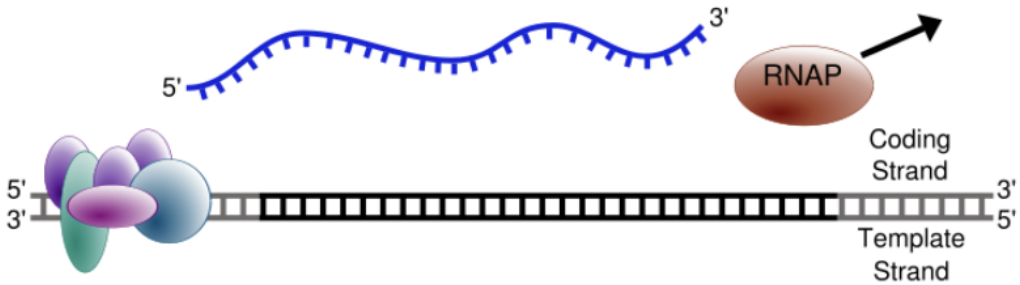
Transcription



(a) Transcription initiation



(b) Transcription elongation



(c) Transcription termination

BRAC UNIVERSITY
Inspiring Excellence

RNA world!

- RNA is produced when DNA is transcribed. It is structurally similar to DNA, with the following major differences:
  1. The nucleotide uracil (U) is used instead of DNA's thymine (T).
  2. RNA contains ribose instead of deoxyribose (deoxyribose lacks the oxygen molecule on the 2' position found in ribose).
  3. RNA is single-stranded, whereas DNA is double-stranded.
- There are many different types of RNA, including:
  - **mRNA (messenger RNA)** contains the information to make a protein and is translated into protein sequence.
  - **tRNA (transfer RNA)** specifies codon-to-amino-acid translation. It contains a 3 base pair anti-codon complementary to a codon on the mRNA, and carries the amino acid corresponding to its anticodon attached to its 3' end.
  - **rRNA (ribosomal RBA)** forms the core of the ribosome, the organelle responsible for the translation of mRNA to protein.
  - **snRNA (small nuclear RNA)** is involved in splicing (removing introns from) pre-mRNA, as well as other functions.

BRAC UNIVERSITY
Inspiring Excellence

Translation

- The primary structure of the protein is determined by the sequence of amino acids of which it is composed.
- Since there are 20 amino acids and only 4 nucleotides, 3-nucleotides sequences in mRNA, known as codons, encode for each of the 20 amino acids.
- Each of the 64 possible 3-sequences of nucleotides (codon) uniquely specifies either a particular amino acid, or is a stop codon that terminates protein translation (the start codon also encodes methionine).

First letter	Second letter: U	Second letter: C	Second letter: A	Second letter: G	Third letter
U	UUU, UUC: Phe; UUA, UUG: Leu	UCU, UCC, UCA, UCG: Ser	UAU, UAC: Tyr; **UAA: Stop**; **UAG: Stop**	UGU, UGC: Cys; **UGA: Stop**; UGG: Trp	U C A G
C	CUU, CUC, CUA, CUG: Leu	CCU, CCC, CCA, CCG: Pro	CAU, CAC: His; CAA, CAG: Gln	CGU, CGC, CGA, CGG: Arg	U C A G
A	AUU, AUC, AUA: Ile; **AUG: Met**	ACU, ACC, ACA, ACG: Thr	AAU, AAC: Asn; AAA, AAG: Lys	AGU, AGC: Ser; AGA, AGG: Arg	U C A G
G	GUU, GUC, GUA, GUG: Val	GCU, GCC, GCA, GCG: Ala	GAU, GAC: Asp; GAA, GAG: Glu	GGU, GGC, GGA, GGG: Gly	U C A G

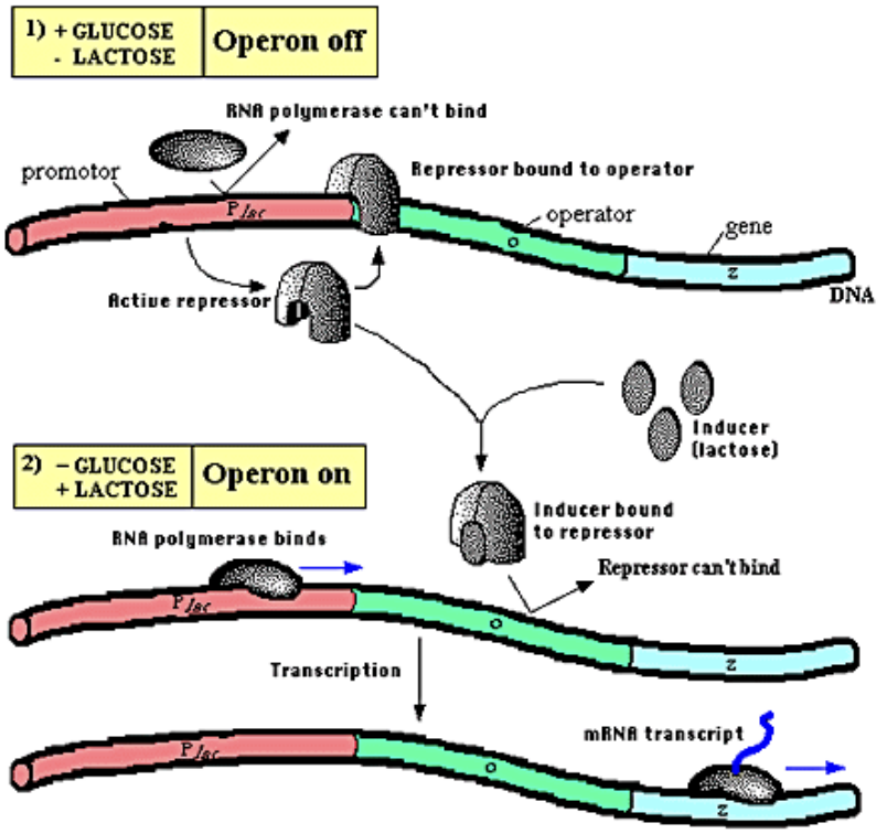
Proteins

- Protein is the molecule responsible for carrying out most of the tasks of the cell, and can have many functions, such as enzymatic, contractile, transport, immune system, signal and receptor to name a few. Like RNA and DNA, proteins are polymers made from repetitive subunits.
- Instead of nucleotides, however, proteins are composed of amino acids.
- Each amino acid has special properties of size, charge, shape, and acidity. As such, additional structure emerges beyond simply the sequence of amino acids (the primary structure), as a result of interactions between the amino acids.
- As such, the three-dimensional shape, and thus the function, of a protein is determined by its sequence.

BRAC UNIVERSITY
Inspiring Excellence

Regulation

- Not all genes are expressed at the same time in a cell.
- A regulatory network is involved to control expression level of genes in a specific circumstance.
- Transcription is one of the steps at which protein levels can be regulated. The promoter region, a segment of DNA found upstream (past the 5' end) of genes, functions in transcriptional regulation.
- The promoter region contains motifs that are recognized by proteins called transcription factors.



Induction of the *lac* Operon

Inspiring Excellence

Where to start?

Where does DNA starts its replication? (next class)
`https://forms.gle/24kDBLEJqrXT8YrT6`

BRAC UNIVERSITY
Inspiring Excellence