Molecular Biology Basics

Bioinformatics

- Analysis, prediction, and modeling of biological data with the help of computers.

- Focuses more on the engineering side o and the creation of tools that work with biological data to solve problems.

Why Bioinformatics?

- DNA sequencing technologies have created massive amounts of information that can only be efficiently analyzed with computers.
- Protein structure prediction
- Model (that represent biological information).
- As information becomes ever so larger and more complex, more computational tools are needed to sort through the data.

Bioinformatics vs. computational Biology

☒ Bioinformatics is generally defined as the analysis, prediction and modeling of biological data with the help of computers.

☒ computational biology is about studying biology using computational techniques, which further the understanding of the science.

Bioinformatics	Computational Biology
1. Definition	1. Definition
2. Focuses on the development of tools, methods and software for managing, analyzing and interpreting biological data.	2. Involves the use of computational methods and mathematical models to understand biology system and process.
3. It deals with organizing and making sense of large datasets.	3. It emphasizes answering biological questions through simulations, modeling and analysis.
4. Develop efficient tools and pipelines to process and organize biological data.	4. Use computational techniques to understand and predict biological behaviours.
5. Relies on data science, database design, programming and statistical analysis.	5. Involves mathematical modeling, simulations and hypothesis testing. ~~co~~
6. Application: Genome assembly and annotation, sequence alignment and analysis, functional genomics etc	6. Application: Modeling cellular and molecular systems, drug design and protein-ligand interaction simulations, population genetics.

(bacteria) or other materials.	cellulose or chitin, not peptidoglycon.
(VIII) Examples: Bacteria and archaea.	(VIII) Example: Animals, plants, fungi and protists!

**

Some terminology: Page-18

Genome: ~~an~~

The genome is an organism's complete set of DNA.

→ A bacteria contains about 600,000 DNA base pairs

→ Human and mouse genomes have some 3 billion.

Gene: A discrete units of hereditary information located ~~m~~ on the chromosomes and consisting of DNA.

Chromosome:

A long DNA molecule with part of the genetic material of an organism.

→ Human genome has 46 (23 pairs) distinct chromosomes.

→ Each chromosome contains many genes.

Genotype:

The genetic makeup of an organism.

Phenotype:

The physical expressed traits of an organism.

Why we can not always construct a multiple alignment from pairwise alignments?

Ans:

A multiple sequence alignment is an alignment of more than two sequence, whereas a pairwise alignment involves only two sequences. We can derive pairwise alignments from a multiple alignment. For example, given the multiple alignment:

X: A C – G C G G – C
Y: A C – G C – G A G
Z: G C C G C – G A G

We can extract the following pairwise alignments:

X: A C G C G G – C Y: A C G C – G A G	Y: A C – G C G A G Z: G C C G C G A G

X: A C – G C G G – C
Z: G C C G C – G A G

However, we cannot always construct a multiple alignment because pairwise alignments may be ~~incosi~~ Incosistent with each other. From an optimal multiple alignment, we can infer pairwise alignments ~~are not~~ between all pairs of sequences. But these pairwise alignments are not

~~nees~~ necessarily optimal. When aligning two sequences at a time, the algorithm optimizes their alignment independently. However, when aligning all ~~the~~ three sequences together, gaps placed optimally for one pair may not work for another pair. Additionally, even if sequence1 aligns ~~wi~~ well with sequence2 and sequence2 aligns well with sequences3, it does not always mean sequence1 align well with sequence3 in the same way. This inconsistency makes it impossible to always reconstruct a valid multiple alignment from pairwise alignments.

Profile: ~~Pr~~

Profile is usually a probability for each letter to occure in each colum~~mn~~n.

Multiple Alignment (Greedy Approach):

Steps:

1. At first calculate all possible pairwise alignment of the given multiple sequence.
2. Then we will find two closest sequence among all of them (the pairwise alignment which score is the greatest among all of them.
3. Then we will join these two sequence into one profile.
4. Them will add the new sequence with other sequences.

Manhattan Tourist Problem (Pseudocode):

```
• MT (n, m)
if n=0 or m=0
        return MT (n,m)
X ← MT (n-1, m) ~~to~~ + length of the edge from (n-1,m) to
        (n,m)
Y ← MT (n, m-1) + length of the edge from (n,m-1) to
        (n,m)
return max ~~(n,m)~~ (X,Y
```

Greedy Sorting ~~Psu~~(Pseudocode):

```
GreedySorting(P)
  approxReverersalDistance ← 0
  for k ← 1 to |P|
    if element k is not sorted
      apply the k-sorting reverersal to P
      approxReverersalDistance ← approxReverersalDistance+1
      if the k-th element of P is -k
        apply the k-sorting reversal to P
        approxReverersalDistance ← approxReverersalDistance+1
  return approxReverersalDistance
```

~~Adja~~

Adjacencies and Breakpoints:

$P_{i+1} - P_i = 1$ → Adjacency

Central Dogma

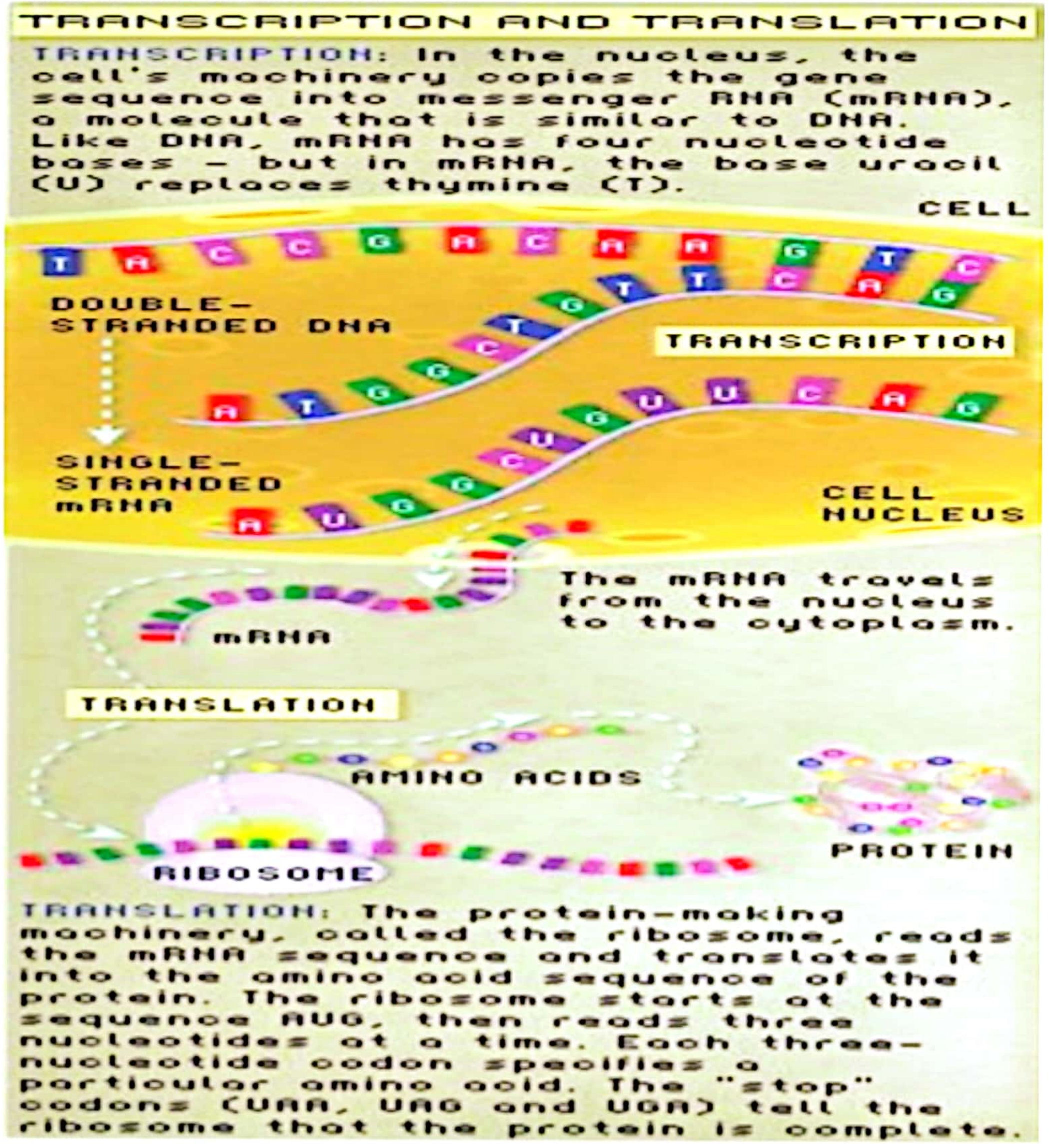
(DNA□RNA□protein) The paradigm that DNA directs its transcription to RNA, which is then translated into a protein.

- **Transcription**

  (DNA□RNA) The process which transfers genetic information from the DNA to the RNA.

- **Translation**

  (RNA□protein) The process of transforming RNA to protein as specified by the genetic code.



TRANSCRIPTION AND TRANSLATION

TRANSCRIPTION: In the nucleus, the cell's machinery copies the gene sequence into messenger RNA (mRNA), a molecule that is similar to DNA. Like DNA, mRNA has four nucleotide bases – but in mRNA, the base uracil (U) replaces thymine (T).

The mRNA travels from the nucleus to the cytoplasm.

TRANSLATION: The protein-making machinery, called the ribosome, reads the mRNA sequence and translates it into the amino acid sequence of the protein. The ribosome starts at the sequence AUG, then reads three nucleotides at a time. Each three-nucleotide codon specifies a particular amino acid. The "stop" codons (UAA, UAG and UGA) tell the ribosome that the protein is complete.

Nucleic Acid:

Biological molecules (RNA & DNA) that allow organisms to reproduce.

Proteins:

→ Make up the cellular structure.

→ Large, complex molecules made up of smaller subunits called amino acids.

The code of life:

- The structure and the four genomic letter codes for all living organisms.

- Adenine, Guanine, Thymine and Cytosine which pair A-T and C-G on complimentary strands.

Cell Information:

Definition:

The **two-break distance** between two genomes is the smallest number of **two-break operations** required to convert one genome into another. A **two-break operation** is a type of genomic mutation where two cuts are made in a genome, followed by rearrangement and rejoining of the resulting segments.