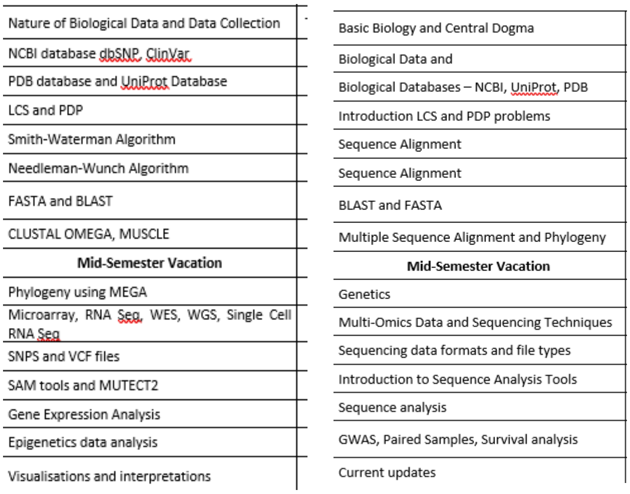
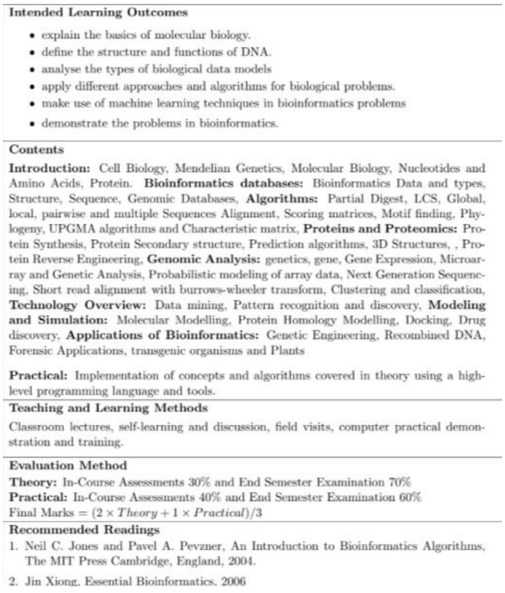
Curriculum



Nature of Biological Data and Data Collection	Basic Biology and Central Dogma
NCBI database dbSNP, ClinVar	Biological Data and
PDB database and UniProt Database	Biological Databases – NCBI, UniProt, PDB
LCS and PDP	Introduction LCS and PDP problems
Smith-Waterman Algorithm	Sequence Alignment
Needleman-Wunch Algorithm	Sequence Alignment
FASTA and BLAST	BLAST and FASTA
CLUSTAL OMEGA, MUSCLE	Multiple Sequence Alignment and Phylogeny
Mid-Semester Vacation	**Mid-Semester Vacation**
Phylogeny using MEGA	Genetics
Microarray, RNA Seq, WES, WGS, Single Cell RNA Seq	Multi-Omics Data and Sequencing Techniques
SNPS and VCF files	Sequencing data formats and file types
SAM tools and MUTECT2	Introduction to Sequence Analysis Tools
Gene Expression Analysis	Sequence analysis
Epigenetics data analysis	GWAS, Paired Samples, Survival analysis
Visualisations and interpretations	Current updates



Intended Learning Outcomes

- explain the basics of molecular biology.
- define the structure and functions of DNA.
- analyse the types of biological data models
- apply different approaches and algorithms for biological problems.
- make use of machine learning techniques in bioinformatics problems
- demonstrate the problems in bioinformatics.

Contents

Introduction: Cell Biology, Mendelian Genetics, Molecular Biology, Nucleotides and Amino Acids, Protein. **Bioinformatics databases:** Bioinformatics Data and types, Structure, Sequence, Genomic Databases, **Algorithms:** Partial Digest, LCS, Global, local, pairwise and multiple Sequences Alignment, Scoring matrices, Motif finding, Phylogeny, UPGMA algorithms and Characteristic matrix, **Proteins and Proteomics:** Protein Synthesis, Protein Secondary structure, Prediction algorithms, 3D Structures, , Protein Reverse Engineering, **Genomic Analysis:** genetics, gene, Gene Expression, Microarray and Genetic Analysis, Probabilistic modeling of array data, Next Generation Sequencing, Short read alignment with burrows-wheeler transform, Clustering and classification, **Technology Overview:** Data mining, Pattern recognition and discovery, **Modeling and Simulation:** Molecular Modelling, Protein Homology Modelling, Docking, Drug discovery, **Applications of Bioinformatics:** Genetic Engineering, Recombined DNA, Forensic Applications, transgenic organisms and Plants

Practical: Implementation of concepts and algorithms covered in theory using a high-level programming language and tools.

Teaching and Learning Methods

Classroom lectures, self-learning and discussion, field visits, computer practical demonstration and training.

Evaluation Method

Theory: In-Course Assessments 30% and End Semester Examination 70%
Practical: In-Course Assessments 40% and End Semester Examination 60%
Final Marks = $(2 \times Theory + 1 \times Practical)/3$

Recommended Readings

1. Neil C. Jones and Pavel A. Pevzner, An Introduction to Bioinformatics Algorithms, The MIT Press Cambridge, England, 2004.
2. Jin Xiong, Essential Bioinformatics, 2006

Objective

- What is the use of computing in biology
- What are Biological data
  - define DNA, RNA, Protein and Amino Acid
  - demonstrate sequence alignment
- How biological data generated
- What are Biological databases
- How to store and analyse biological data
- Sequence Alignment
- Assumptions on biological analysis
- Phylogeny Tree
- Targets and drug design concepts

Bioinformatics (Motivation)

Biology quickly has 500 years of exciting problems to work on

5

5

Keywords

Actin Filaments Animal cells Plant cells

Plasma Membrane

Extracellular Matrix

Golgi Apparatus

Lysosome mitosis DNA replication

Centrioles

Ribosomes Prokaryotes Eukaryotes Apoptosis

Centrosomes

Mitochondria

Rough Endoplasmic Reticulum Genoptype

Smooth Endoplasmic Reticulum

Nucleus

DNA methylation

Histones epigenetics

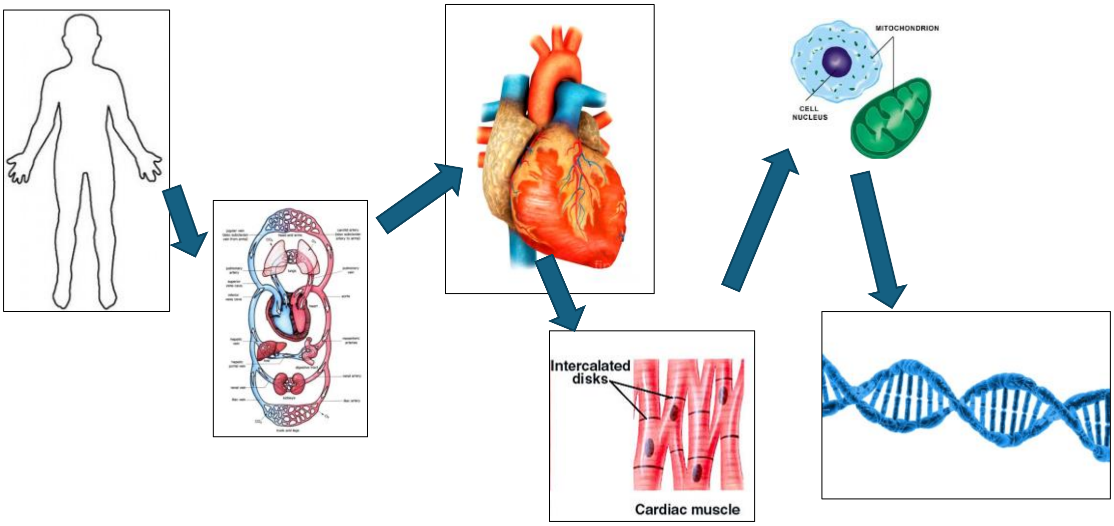
Cell Cycle (G0, G1, M and G2 phases) Genetic finger printing

Prophase, Metaphase, Anaphase and Telophase

6

6

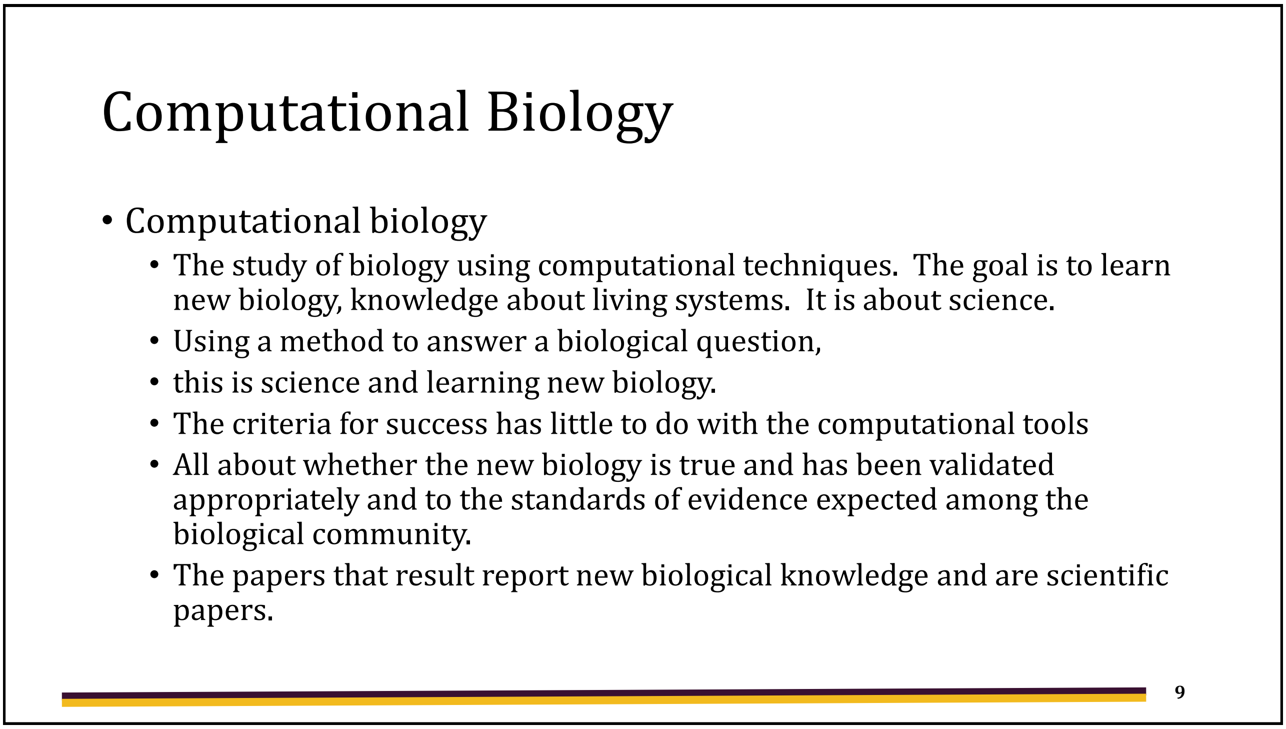
Human Body and System



BIOINFORMATICS

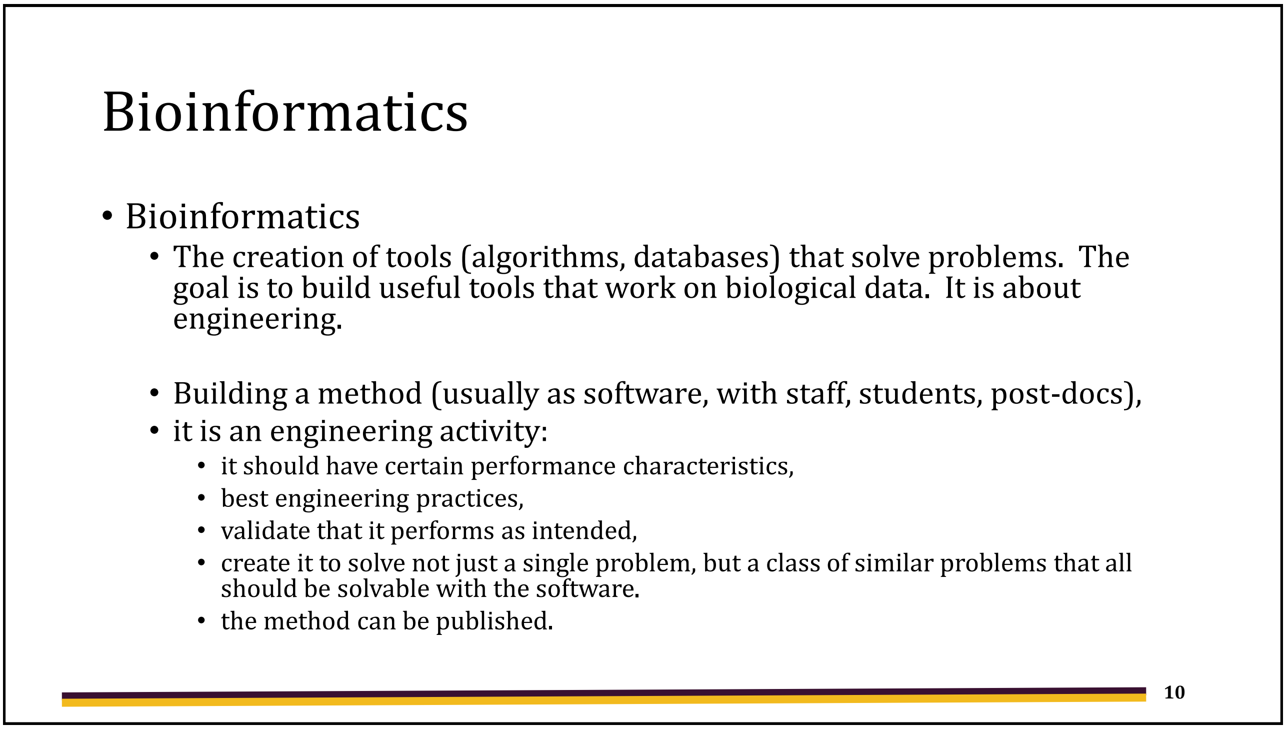
=

COMPUTATION
AL BIOLOGY?



Computational Biology

- Computational biology
  - The study of biology using computational techniques. The goal is to learn new biology, knowledge about living systems. It is about science.
  - Using a method to answer a biological question,
  - this is science and learning new biology.
  - The criteria for success has little to do with the computational tools
  - All about whether the new biology is true and has been validated appropriately and to the standards of evidence expected among the biological community.
  - The papers that result report new biological knowledge and are scientific papers.



Bioinformatics

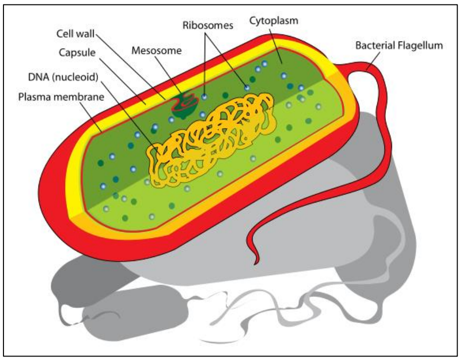
- Bioinformatics
  - The creation of tools (algorithms, databases) that solve problems. The goal is to build useful tools that work on biological data. It is about engineering.

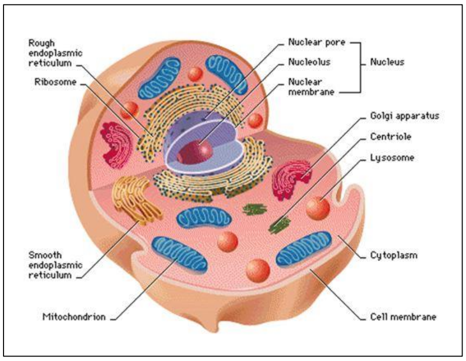
  - Building a method (usually as software, with staff, students, post-docs),
  - it is an engineering activity:
    - it should have certain performance characteristics,
    - best engineering practices,
    - validate that it performs as intended,
    - create it to solve not just a single problem, but a class of similar problems that all should be solvable with the software.
    - the method can be published.

DNA and Genes

- DNA is where the genetic information is stored
- Genes contain the information as a sequence of nucleotides
- Genes are abstract concepts – like longitude and latitudes in the sense that you cannot see them separately
- Genes are made up of nucleotides

Cell



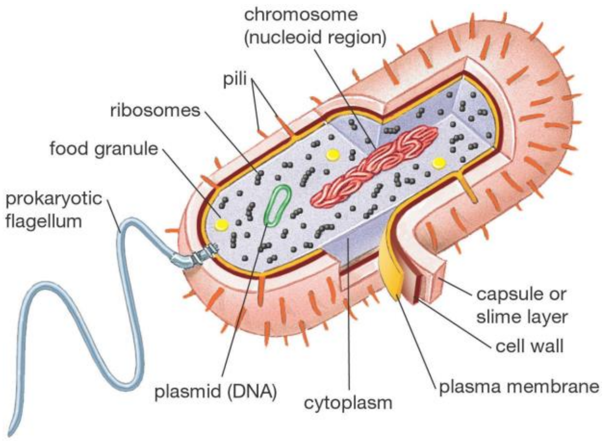


Prokaryotic Vs Eukaryotic

- **Bacteria**
- **Size:** 1-10mm
- Cell Wall (murein)
- **No distinct subcellular organelles**
- **Circular chromosome – nucleoid**
- Often plasmids, RNA and Ribosomes
- **Unicellular or multicellular**
- Escherichia Coli (*E. Coli*) is most studied bacterium

- **Plants, animals, fungi and protists (algae and protozoa)**
- **Size:** 10-100mm
- Cell Wall – only plants, fungi and protists (cellulose)
- **Well defined subcellular compartments bounded by lipid membranes**
- **Cytoplasm** consists of organelles, ribosomes, cytoskeleton (shape, movement and organises many metabolic functions)
- **Cytoskeleton:** microtubules made of tublin & microfilaments made of actin.
- **Most are multicellular**
- **Differentiate to specialized tissue/cells**

Prokaryotic Cell

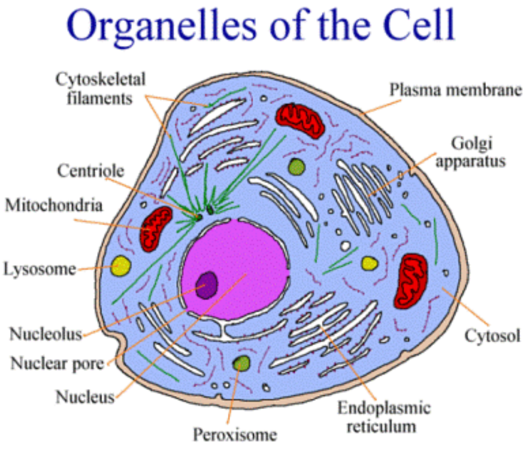


Copyright © 2005 Pearson Prentice Hall, Inc.

Prokaryotic Cell

- **Nucleoid** – composed of circular double-stranded DNA.
- **Plasmid DNA** – Short circular DNA and replicates independently of the cell genome.
- **Mesosome** – Folds of the plasma membrane with associated respiration enzymes. Instead of mitochondria.
- **Ribosomes** – Smaller, scattered throughout the cytoplasm
- **Pilli** – protein rods for cell-cell attachment and DNA transfer.
- **Flagellum** – Motility of many bacteria
- **Cell Wall** – Rigid and made up of murein (polysaccharide cross-linked by peptide chains). Gram-positive thicker walls compared to Gram-negative. Protection from lysozymes and penicillin.
- **Capsule** – slime layer of mucilage and helps bacteria form colonies.

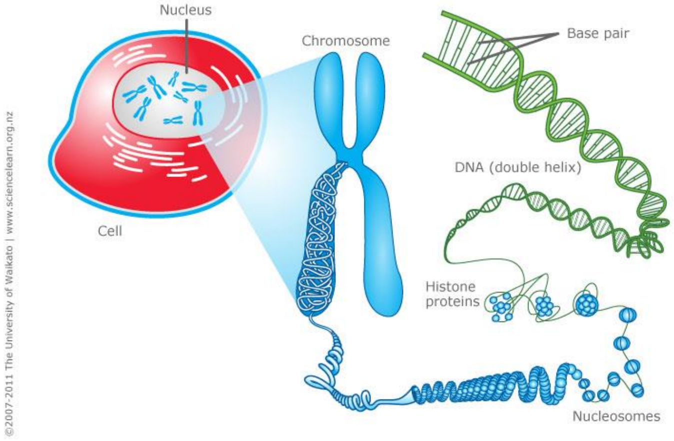
Eukaryotic Cell



Eukaryotic Cell

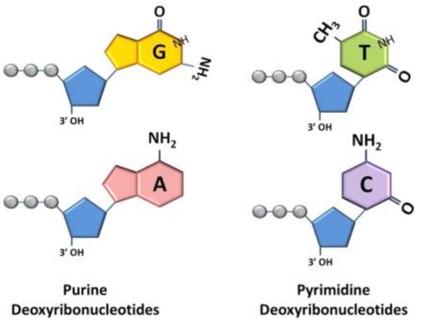
- **Nucleus** – Cellular DNA. Transcription & processing of RNA. Nuclear pores within the nuclear membrane.
- **Mitochondria** – Cellular respiration, the oxidation of nutrients to generate energy in the form of adenosine 5'-triphosphate (ATP). 1-2mm in diameter. 1000-2000 per cell. Smooth outer membrane & Inner folded membrane (cristae). Derived from prokaryotes and retain DNA (circular), RNA and protein machinery.
- **Endoplasmic Reticulum (ER)** – Cytoplasmic membrane system for lipid biosynthesis and xenobiotic metabolism. Smooth and Rough ER. Rough ER has ribosome attached for protein synthesis.
- **Golgi Apparatus** – Protein and lipids produced are packaged in the Golgi for final destination.
- **Lysosomes** – Small membrane-bound organelles & bud off from the Golgi. Consist of degradative enzymes for proteins, nucleic acid, lipids and carbohydrates (macromolecules).
- **Centrioles** – Regulator of the cell cycle and cytoskeletal organisation.

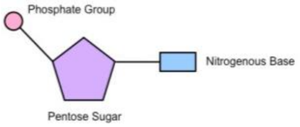
DNA

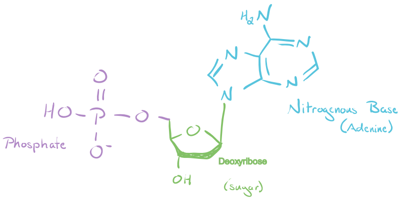


©2007-2011 The University of Waikato | www.sciencelearn.org.nz

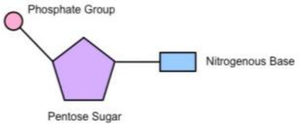
Nucleotide (nt)





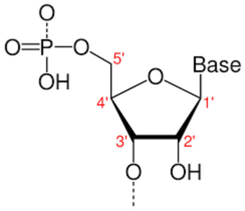


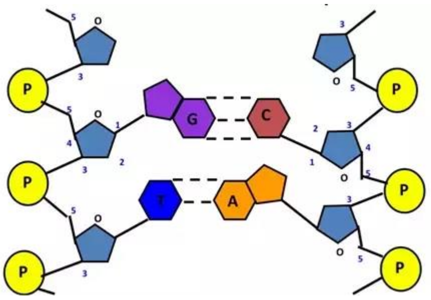
Nucleotide (nt)

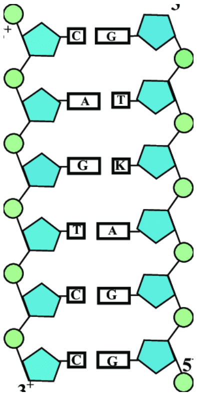


- Each nt is made up of
  - Sugar
  - Phosphate group
  - Base
- The base it (nt) contains makes the only difference between one nt and the other
- There are 4 different bases
  - G(uanine),A(denine),T(hymine),C(ytosine)
- The information is in the order of nucleotide and the order is the info
- Genes can be many thousands of nt long
- The complete set of genetic instructions is called genomes

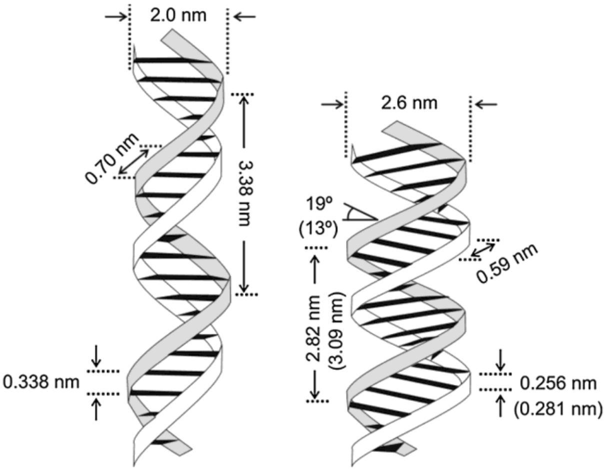
DNA


O
O=P–O
OH
5'
Base
O
4'
1'
3'
2'
O
OH

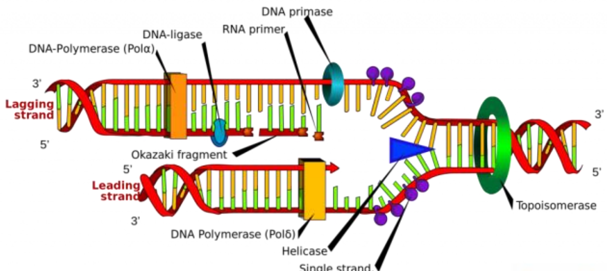

P
P
P
G
C
T
A
P
P
P

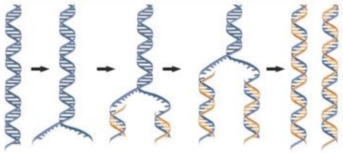

C
G
A
T
G
K
T
A
C
G
C
G
5'
3'

DNA

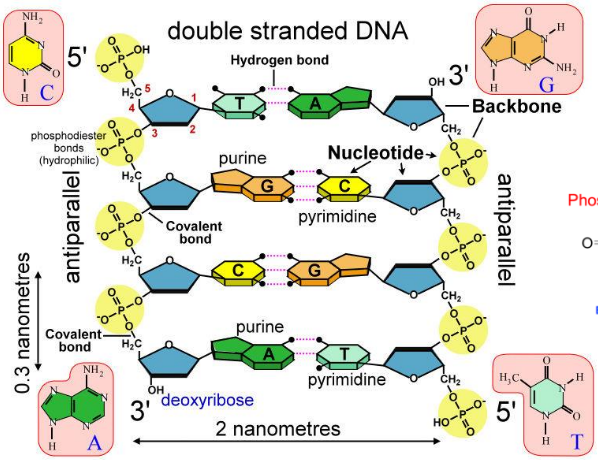

2.0 nm
0.70 nm
3.38 nm
0.338 nm
2.6 nm
19º
(13º)
0.59 nm
2.82 nm
(3.09 nm)
0.256 nm
(0.281 nm)

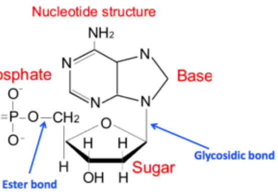
DNA Replication



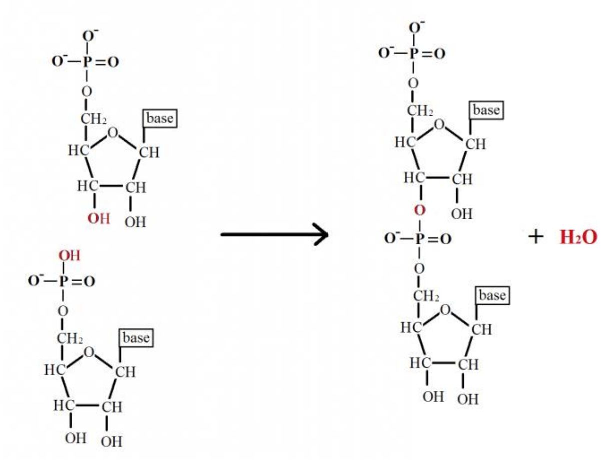


DNA Bonds





DNA Bonds



DNA Packaging

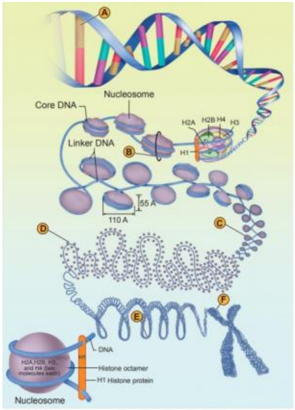
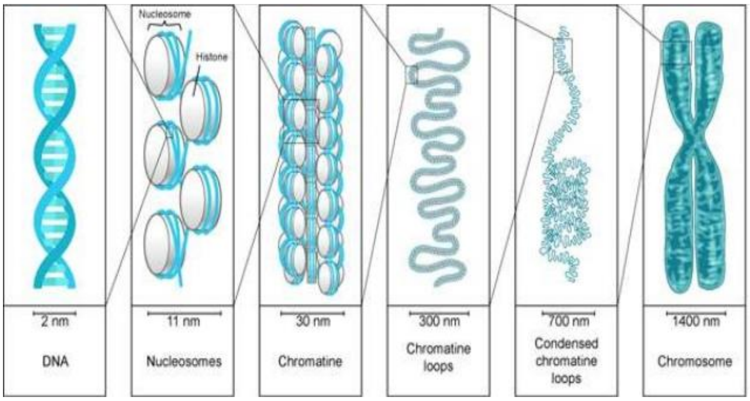
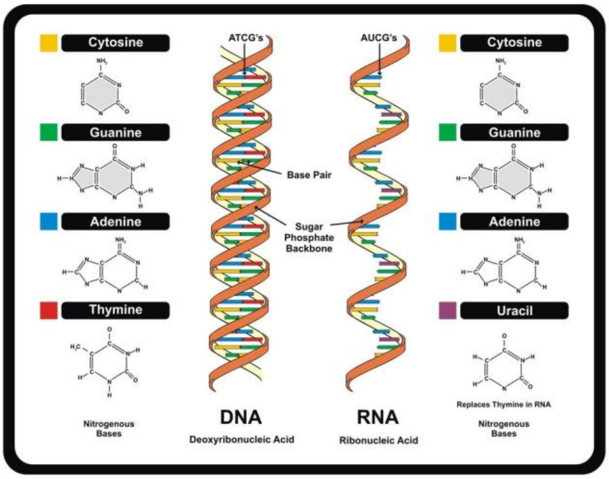


Fig. 5.3 Condensation of DNA - A - DNA, B-Nucleosomes and Histones, C- Chromatin fiber, D- Coiled chromatin fiber, E- Coiled coil, F- metaphase chromatid

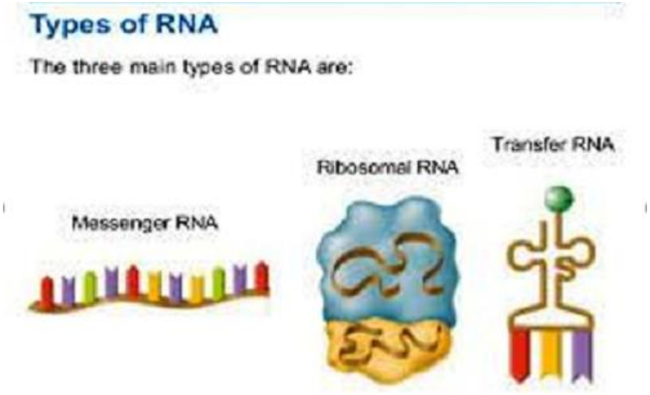
DNA Packaging



DNA Vs RNA



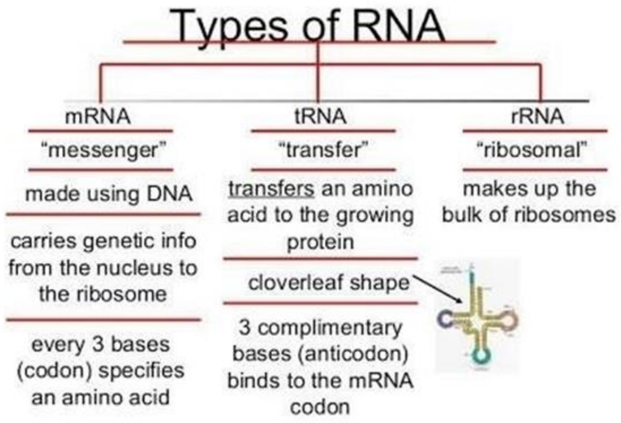
Types of RNA



Types of RNA

The three main types of RNA are:

Types of RNA



Types of RNA

mRNA

"messenger"

made using DNA

carries genetic info from the nucleus to the ribosome

every 3 bases (codon) specifies an amino acid

tRNA

"transfer"

transfers an amino acid to the growing protein

cloverleaf shape

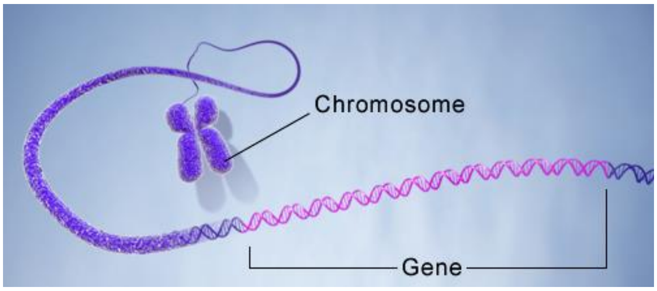
3 complimentary bases (anticodon) binds to the mRNA codon

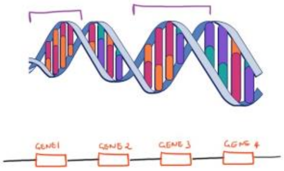
rRNA

"ribosomal"

makes up the bulk of ribosomes

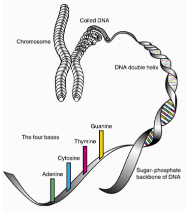
Gene



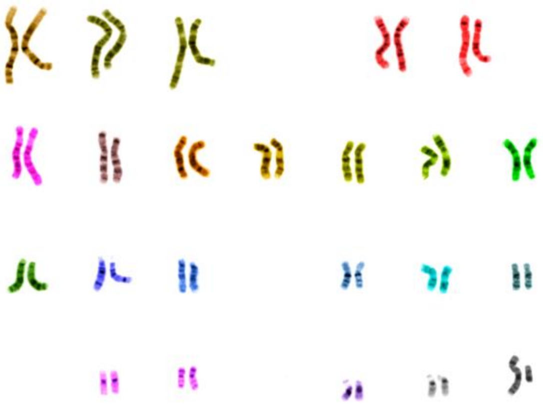


Chromosomes

- DNA strings make chromosomes
- Analogy
  - Letters - nt
  - Sentences – genes
  - Individual volumes of Britannica encyclopedia – chromosomes
  - All volumes together - Genome



Genome

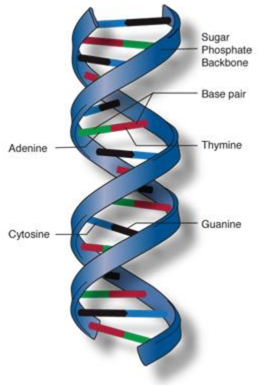


Genetic and Evolution

- Mutation
  - The changing of the structure of a gene, resulting in a variant form that may be transmitted to subsequent generations, caused by the alteration of single base units in DNA.
- Natural selection
  - The process whereby organisms better adapted to their environment tend to survive and produce more offspring.
- Genetic Drift
  - Variation in the relative frequency of different genotypes in a small population.

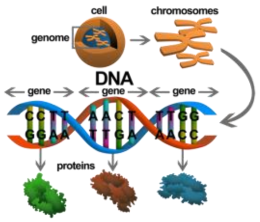
Double Helix

- The DNA is a double helix
- Each strand has complementary information
- Each particular base in one strand is bonded with another particular base in the next strand
  - G - C
  - A - T
  - For example -
  - AATGC one strand
  - TTACG other strand

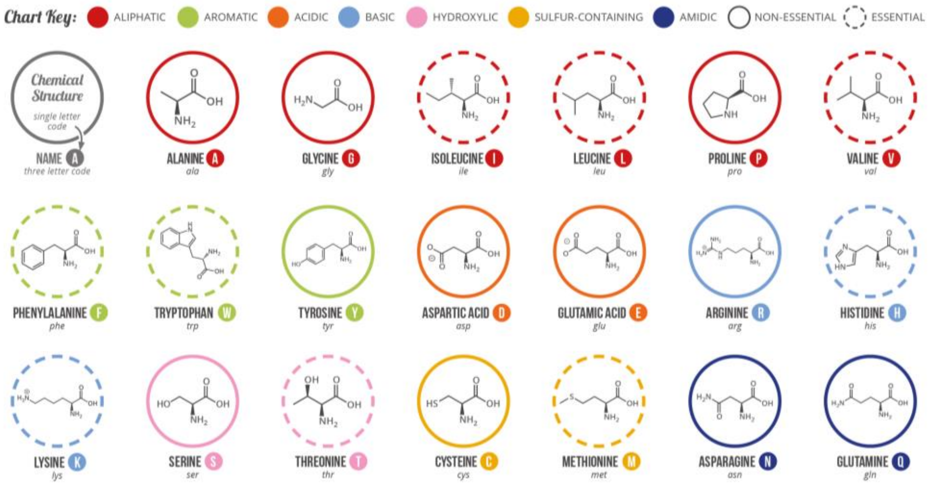


Protein

- Proteins are very important biological feature
- Amino Acids make up the proteins
- 20 different amino acids are there
- The function of a protein is dependant on the order of the amino acids



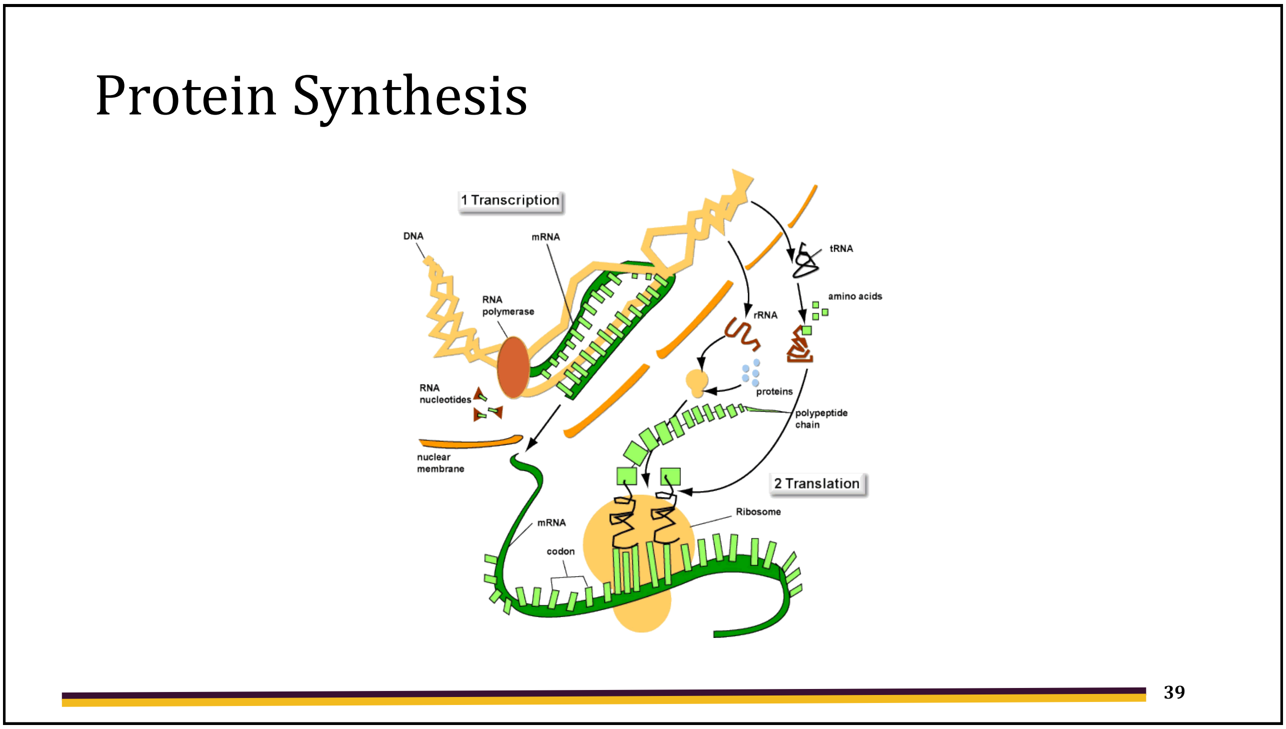
Amino Acids



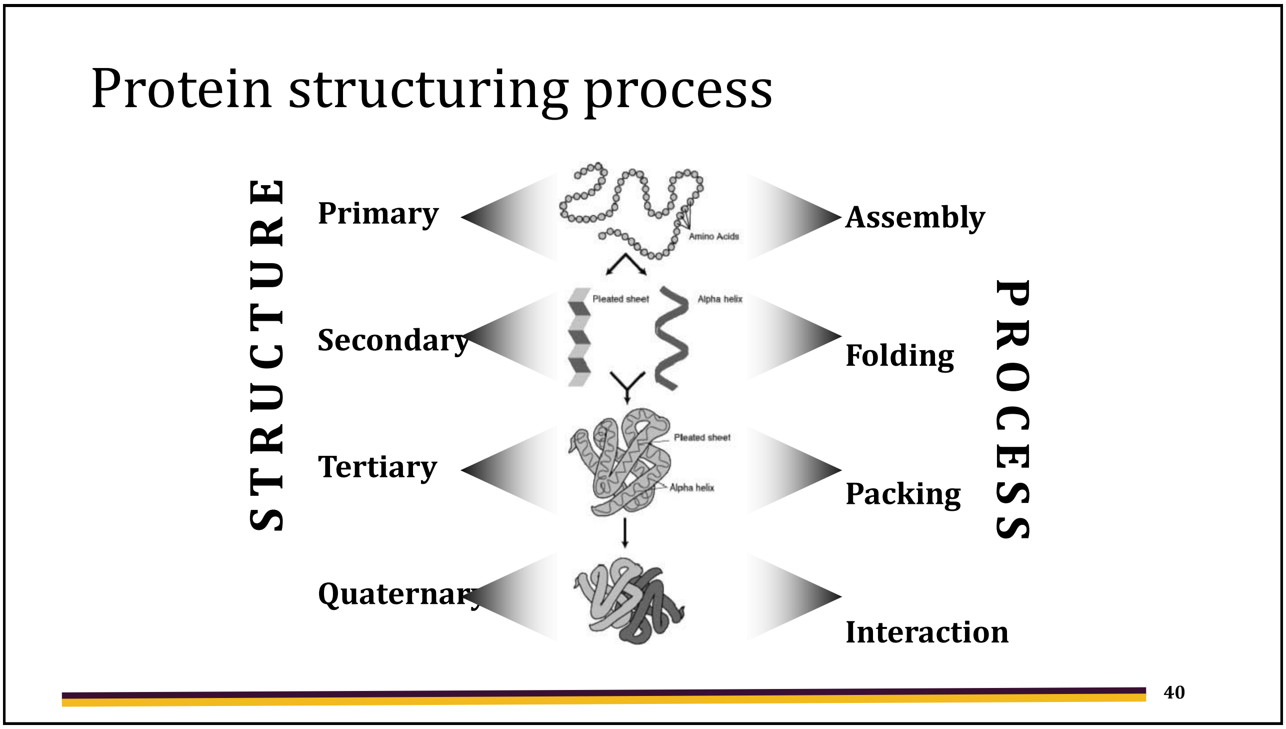
Note: This chart only shows those amino acids for which the human genetic code directly codes for. Selenocysteine is often referred to as the 21st amino acid, but is encoded in a special manner. In some cases, distinguishing between asparagine/aspartic acid and glutamine/glutamic acid is difficult. In these cases, the codes asx (B) and glx (Z) are respectively used.

Protein

- The information required to make aa is stored in DNA
- DNA sequence determines amino acid sequence
- Amino Acid sequence determines protein structure
- Protein structure determines protein function
- A Substance called RNA is used to carry the Info stored in the DNA that in turn is used to make proteins
- Storage - DNA
- Information Transfer – RNA
- RNA is the message boy!



Protein Synthesis



Protein structuring process

STRUCTURE

Primary

Secondary

Tertiary

Quaternary

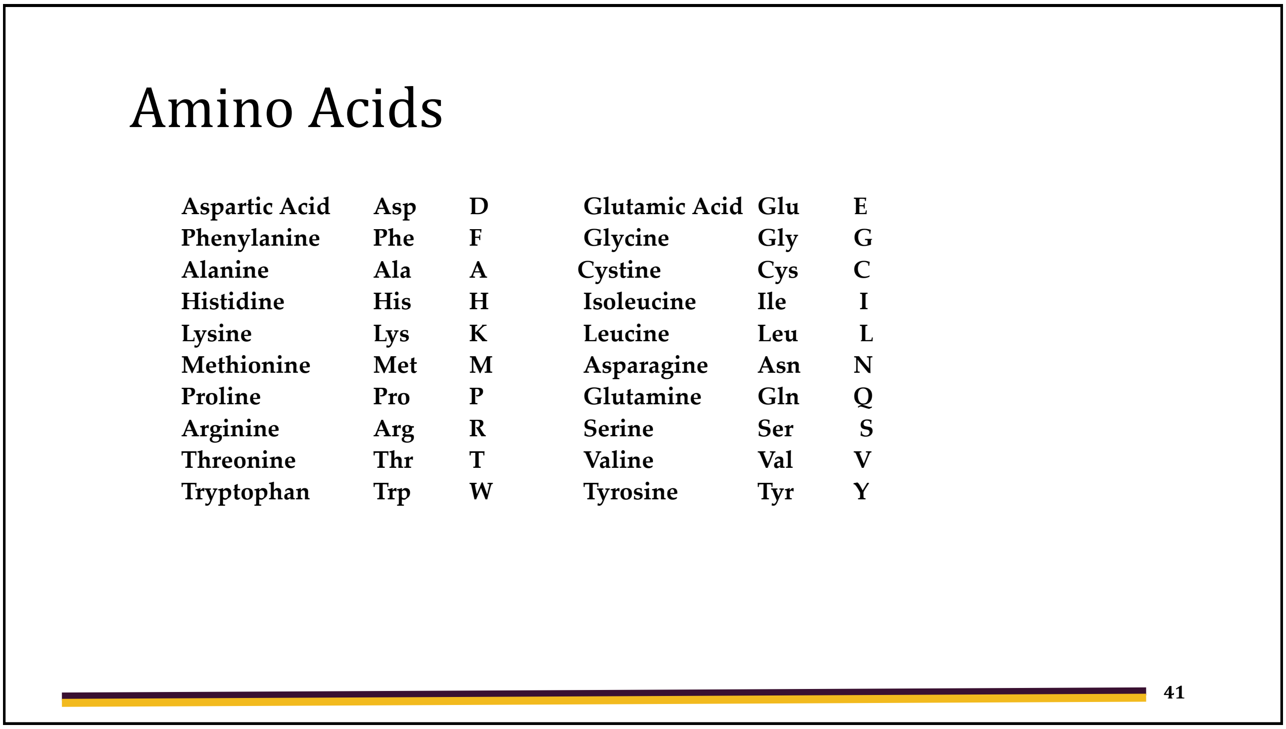
Assembly

Folding

Packing

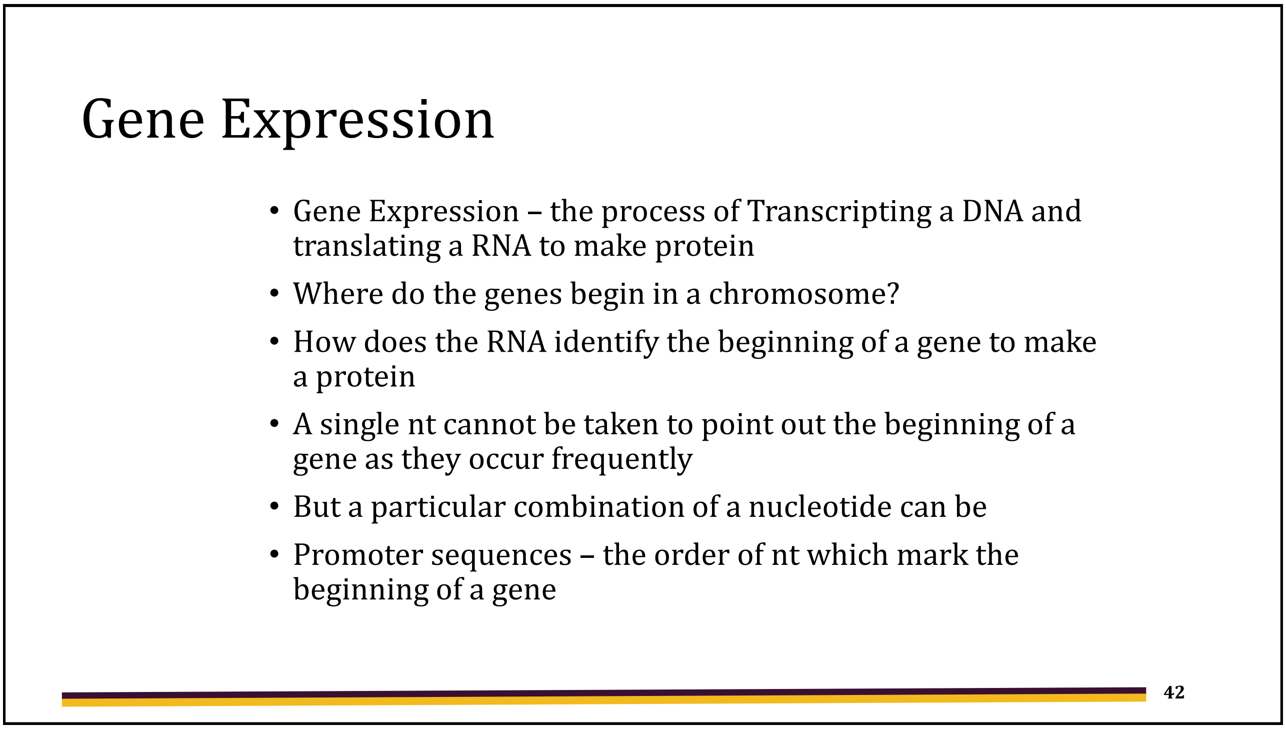
Interaction

PROCESS



Amino Acids

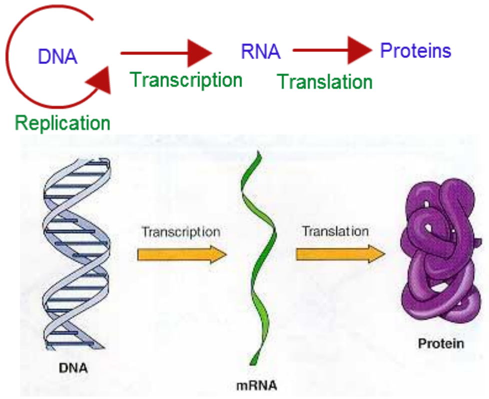
Aspartic Acid	Asp	D	Glutamic Acid	Glu	E
Phenylanine	Phe	F	Glycine	Gly	G
Alanine	Ala	A	Cystine	Cys	C
Histidine	His	H	Isoleucine	Ile	I
Lysine	Lys	K	Leucine	Leu	L
Methionine	Met	M	Asparagine	Asn	N
Proline	Pro	P	Glutamine	Gln	Q
Arginine	Arg	R	Serine	Ser	S
Threonine	Thr	T	Valine	Val	V
Tryptophan	Trp	W	Tyrosine	Tyr	Y



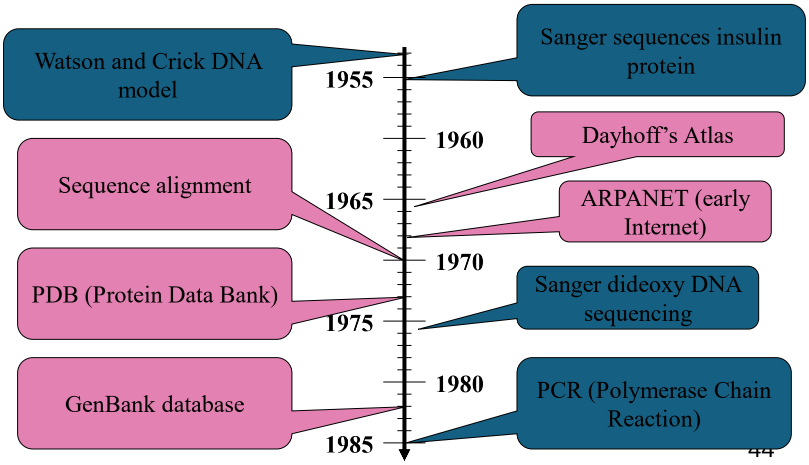
Gene Expression

- Gene Expression – the process of Transcripting a DNA and translating a RNA to make protein
- Where do the genes begin in a chromosome?
- How does the RNA identify the beginning of a gene to make a protein
- A single nt cannot be taken to point out the beginning of a gene as they occur frequently
- But a particular combination of a nucleotide can be
- Promoter sequences – the order of nt which mark the beginning of a gene

Central Dogma



From DNA to Genome


Watson and Crick DNA model

Sanger sequences insulin protein

Dayhoff's Atlas

Sequence alignment

ARPANET (early Internet)

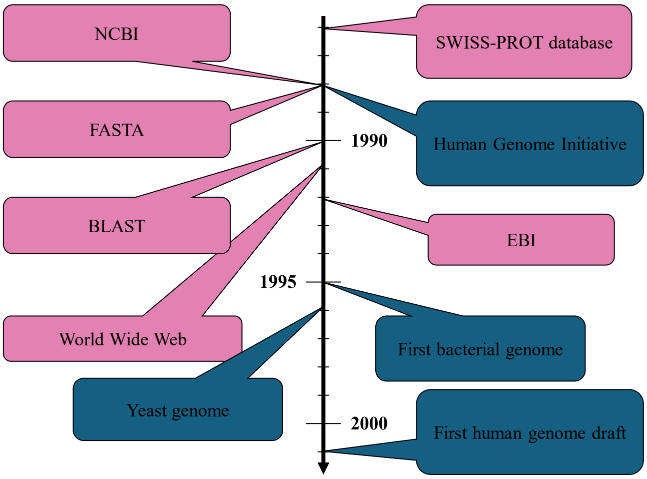
PDB (Protein Data Bank)

Sanger dideoxy DNA sequencing

GenBank database

PCR (Polymerase Chain Reaction)

From DNA to Genome



Data

No meaning

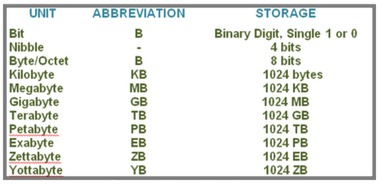
Raw Fact

Can't Understand

Unorganized

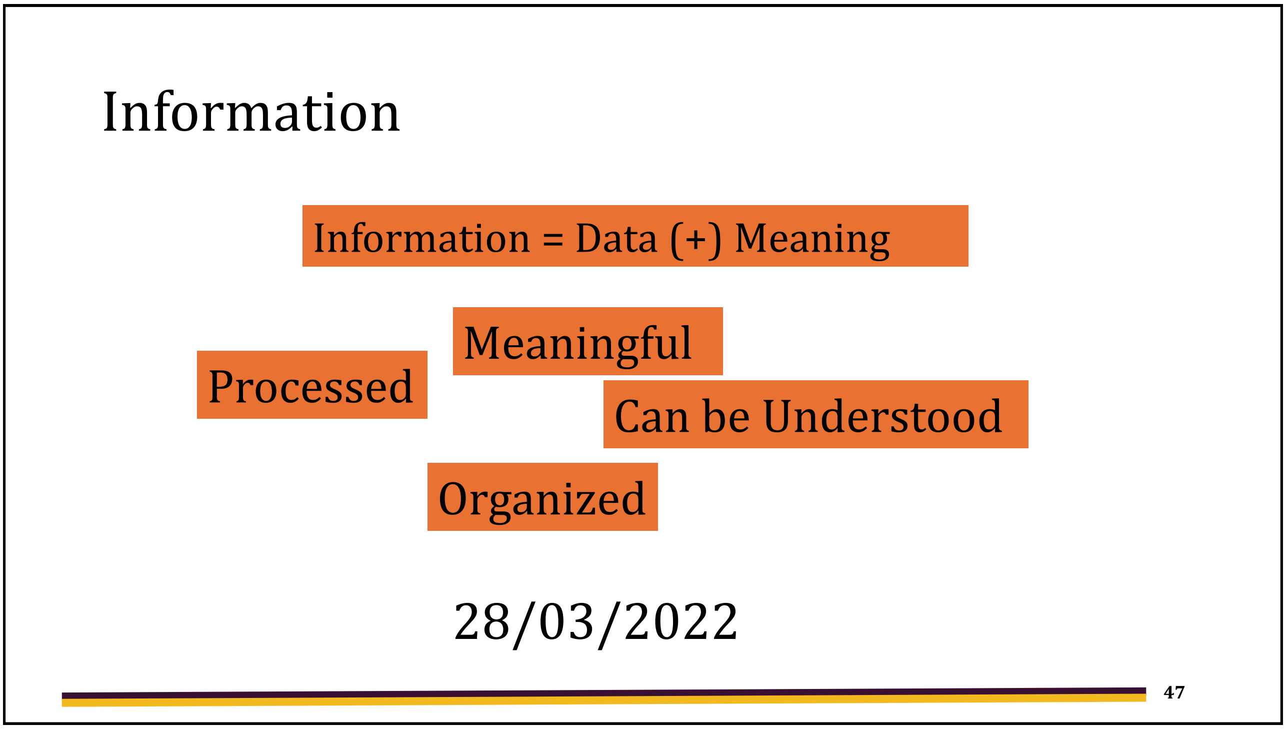
Need to be processed

Unfiltered



UNIT	ABBREVIATION	STORAGE
Bit	B	Binary Digit, Single 1 or 0
Nibble	-	4 bits
Byte/Octet	B	8 bits
Kilobyte	KB	1024 bytes
Megabyte	MB	1024 KB
Gigabyte	GB	1024 MB
Terabyte	TB	1024 GB
Petabyte	PB	1024 TB
Exabyte	EB	1024 PB
Zettabyte	ZB	1024 EB
Yottabyte	YB	1024 ZB

28032022



Information

Information = Data (+) Meaning

Meaningful

Processed

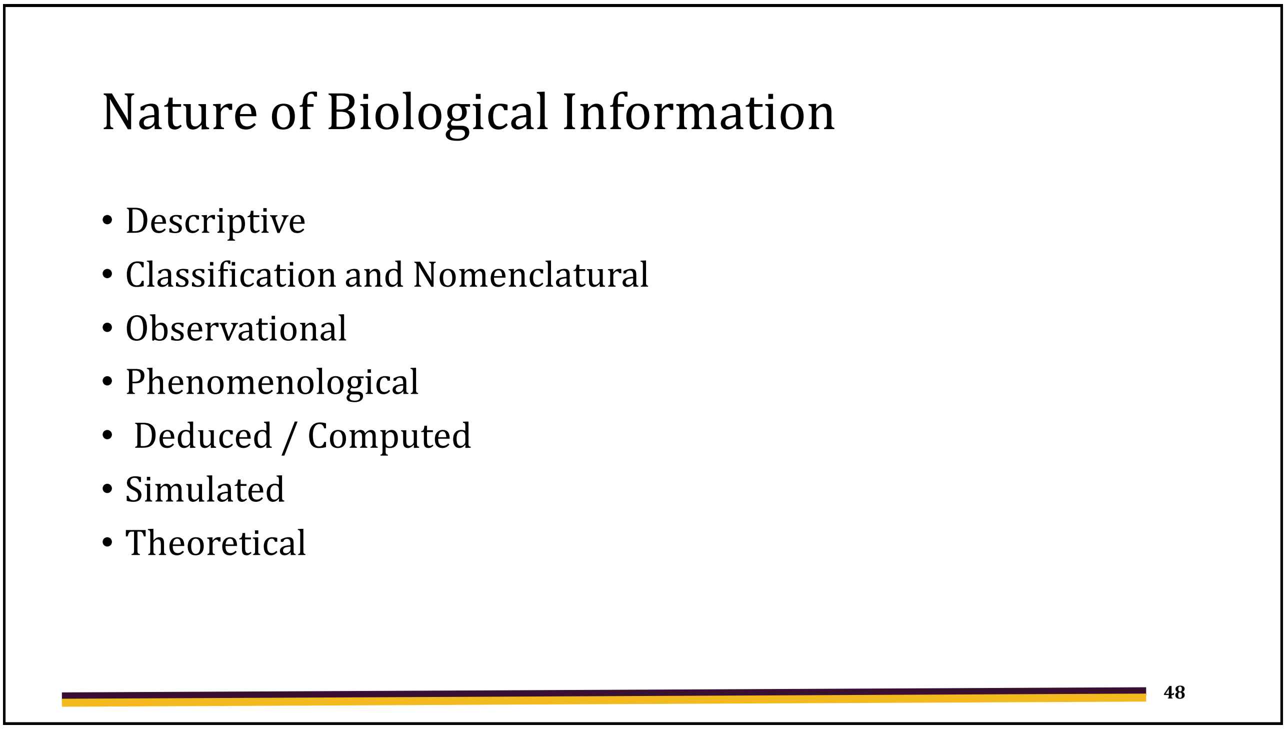
Can be Understood

Organized

28/03/2022

47

47



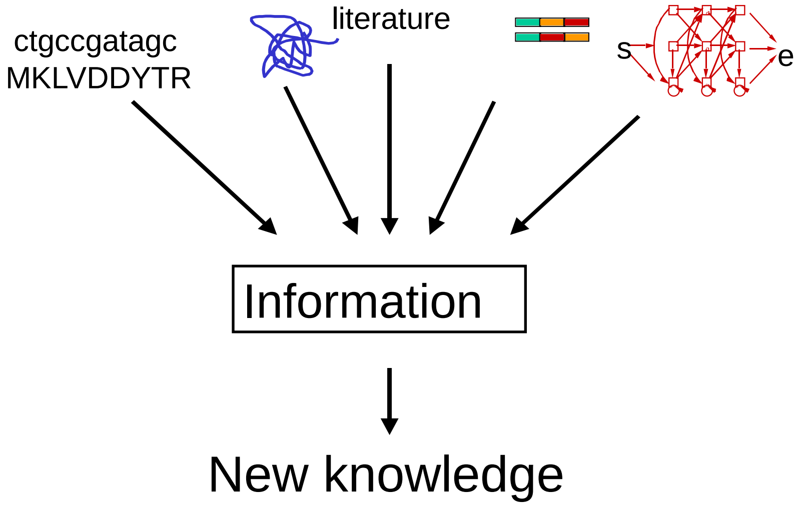
Nature of Biological Information

- Descriptive
- Classification and Nomenclatural
- Observational
- Phenomenological
- Deduced / Computed
- Simulated
- Theoretical

48

48

Biological Information



ctgccgatagc
MKLVDDYTR

literature

Information

New knowledge

Biological Problem

- **Patient**: has a unique MSP number, a Patient name, a Date of Birth, a Tissue Type and an indicator denoting whether the tissue is cancerous or normal.
- **A** patient library associates a patient with multiple tags
- **Each tag** has a unique tag number and a unique nucleotide sequence.

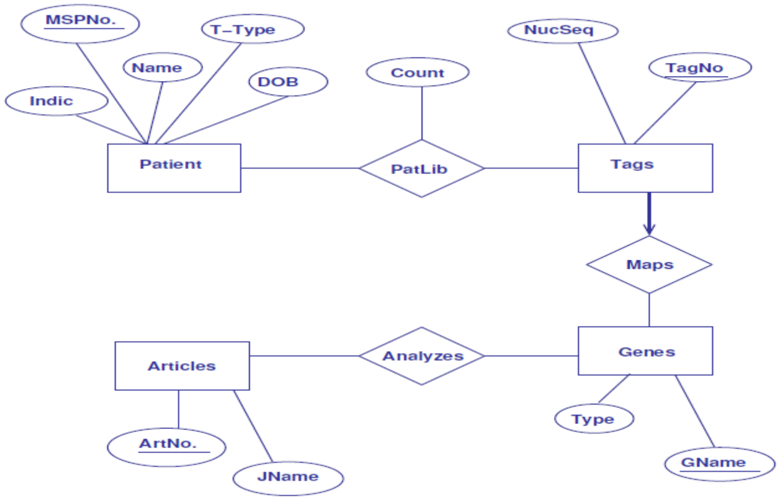
Cont..

- **For each tag in the patient library**, a count is given to record the number of times the tag occurs in the library. In general, the same tag can be associated with any number of patients.
- **A tag may be mapped** to a gene. Each gene has a unique gene name and a type.
- **In general, multiple tags may be mapped to the same gene**. However, two different genes cannot be mapped to the same tag.
- **Finally**, an article is identified by a unique article number and a journal name. An article may analyze multiple genes and a gene may be analyzed by multiple articles.

Diagram

- Entities
  - Patients –
    - MSP Number, Name, DOB, Tissue Type and Indicator
  - Tags –
    - Tag Number and Nucleotide Sequence
  - Genes –
    - Gene Name and Type
  - Articles –
    - Article Number and Journal Name
- Relationships:
  - Patient Library - Many to Many, Has an attribute Count
  - Map - Many to 1 from Tags to Genes
  - Analyses - Many to Many

Diagram



Bioinformatics topics

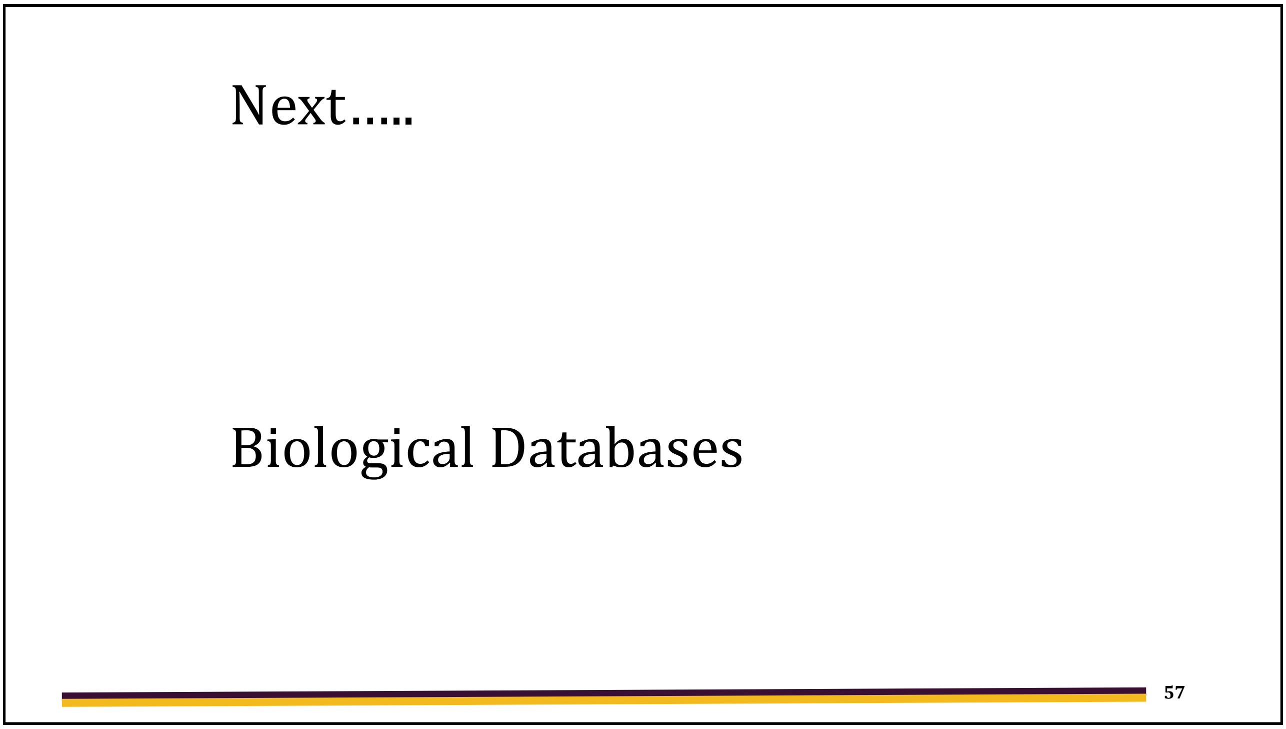
- Genomics
- Proteomics
- Metabolomics
- Phylogenetic
- Oncogenomics
- String matching
- Medical Record Analysis
- Survival Analysis

Techniques

- Sequence Analysis
- Prediction
- Patter Recognition
- Classification
- Clustering
- Gene Expression Analysis

Databases in Biology

- Sequence databases
- Sequence analysis
- Functional genomics
- Literature databases
- Structural databases
- Metabolic pathway databases
- Specialized databases



Next…..

Biological Databases