

Bioinformatics I

Lecture 1: Introduction to Bioinformatics

Dr Manaf A Guma
University Of Anbar- college of Applied sciences-Hit
Department of applied chemistry

1

What does bioinformatic study?



Bioinformatics and computational biology are addressing biological problems with computational methods



It uses the tools and terminology of biology to describe the properties of living organisms and their genes



It answers some questions in genetics fields:



How are we different from others?

2

Bioinformatics:

Genetics:

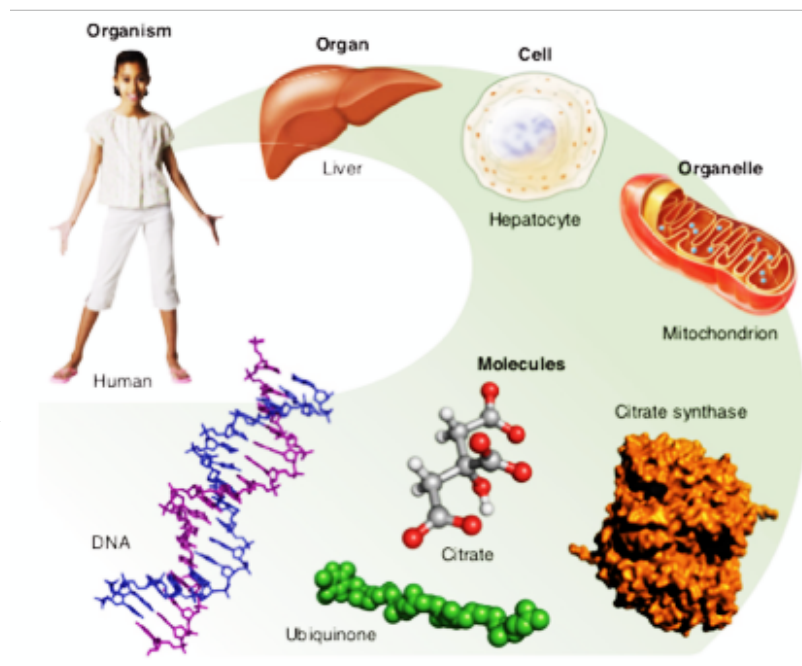
- DNA and RNA analysis and Structure.
- Includes techniques.
- PCR, Cloning and mutagenesis.
- Statistics of different variants of species

Proteins:

- Expression and purification of protein.
- Structural and functional of proteins.
- Chemistry & biology of the protein.
- Biophysics.
- NMR, X-ray Crystallography, Electron microscopy and enzymes and others' kinetics
- Interactions and affinities.
- Drug design and discovery

3

**What we do consist of ?
What are the small molecules that code us?**



4

The Storage of Genetic Information: What is the meaning of the terms?

- Bioinformatics are driven by simple information that are built up from **the sequences of DNA, RNA and proteins (which are the Storage of Genetic Information)**
- **DNA** consists of four nucleotides that store genetic information.
- The base sequence encodes the necessary information **to generate proteins.**
- The entirety of genomic DNA in any organism is known as a **genome.**
- The total pool of mRNA in any organism is referred to as a **transcriptome.**
- The entire pool of proteins in any organism is referred to as the **proteome.**
- A **genome** comprises genes that contain the information to build proteins.
- **Genome** is the entire DNA sequence of an organism.

5

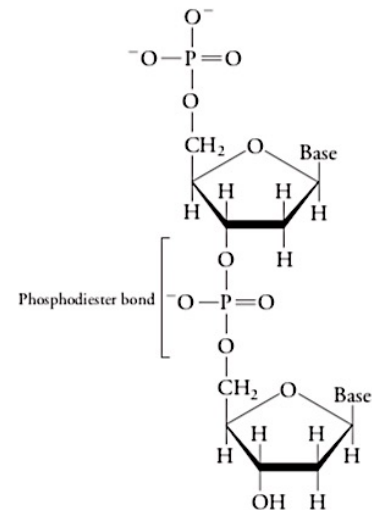
What are Macro (proteins) and micro (DNA) biomolecules ??

- DNA and RNA ?? Which are polymers of nucleotides, each of which consists of a purine or pyrimidine base, deoxyribose or ribose, and phosphate.
- Amino acids ??? Are coded by nucleotides>>>>
- Proteins ?? Poly amino acids
- **Note: Micro-** is a prefix which may be applied to word when describing something that is small scale. **Macro-** is a prefix which **means** large.

6

Nucleic Acids:

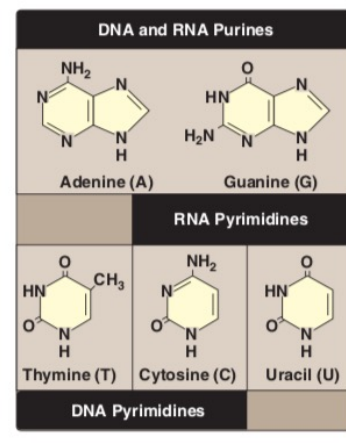
- Polymers of nucleotides are termed **polynucleotides** or **nucleic acids**, better known as DNA and RNA.
- RNA contain the bases adenine, cytosine, guanine, and uracil, whereas the residues in DNA contain adenine, cytosine, guanine, and thymine.
- Polymerization involves the phosphate and sugar groups of the nucleotides, which become linked by **phosphodiester bonds**.



7

Purine or pyrimidine base

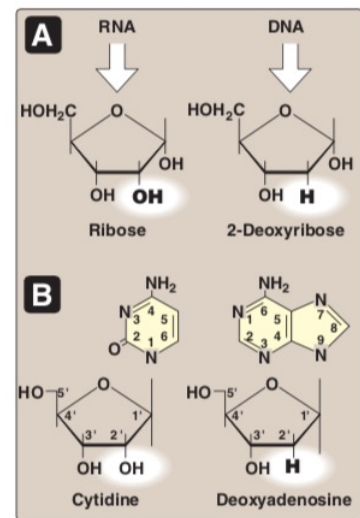
- Nucleotides are composed of a nitrogenous base, a pentose monosaccharide, and one, two, or three phosphate groups.
- The nitrogen-containing bases belong to two families of compounds: the purines and the pyrimidines.
- Both DNA and RNA contain the same purine bases: adenine (A) and guanine (G). Both DNA and RNA contain the pyrimidine cytosine (C), but they differ in their second pyrimidine base: DNA contains thymine (T), whereas RNA contains uracil (U). T and U differ in that only T has a methyl group



Purines and pyrimidines commonly found in DNA and RNA.

8

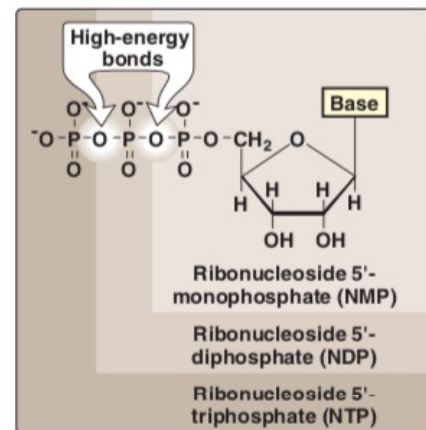
- **Nucleosides**
- The addition of a pentose sugar to a base produces a nucleoside.
- If the sugar is ribose, a ribonucleotide is produced; if the sugar is 2-deoxyribose, a deoxyribonucleoside is produced
- **Nucleotides**
- The addition of one or more phosphate groups to a nucleoside produces a nucleotide.



A. Pentoses found in nucleic acids.
B. Examples of the numbering systems for purine- and pyrimidine-containing nucleosides.

9

- The first phosphate group is attached by an ester linkage to the 5'-OH of the pentose. Such a compound is called a nucleoside 5'-phosphate or a 5'-nucleotide.
- The second and third phosphates are each connected to the nucleotide by a “high-energy” bond.

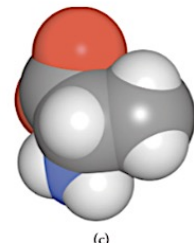
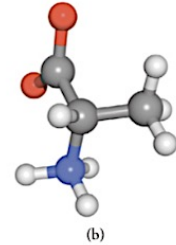
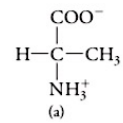


Ribonucleoside monophosphate, diphosphate, and triphosphate.

10

Amino Acids:

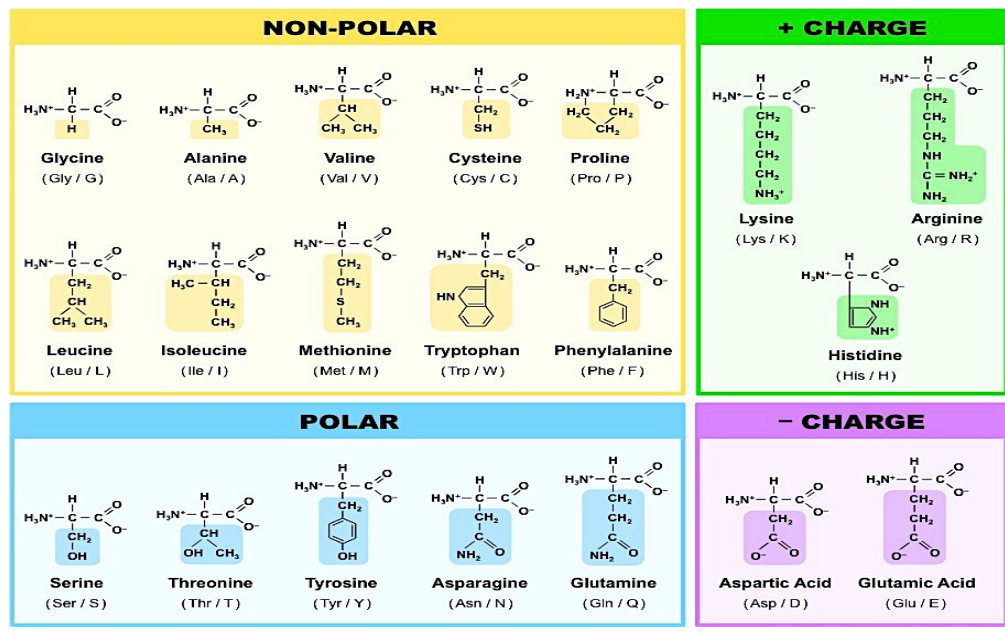
- They contain an amino group (-NH₂) and a carboxylic acid group (-COOH) .
- For example, by a) structural formula, b) ball-and-stick model, or c) space- filling model.



11

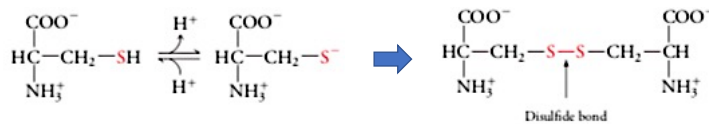
The 20 amino acids have different chemical properties

Note:-
Name:Glycine.
Symbol: G.
3-letters:Gly



12

- **Amino Acids with Hydrophobic Side Chains:** Sidechains interact very weakly or not with water. e.g. alanine (Ala).
- **Amino Acids with Polar Side Chains:** The sides interact with water because they contain hydrogen-bonding groups. E.g. Serine (Ser). Cysteine (Cys) has a thiol group: can form a disulphide bond:

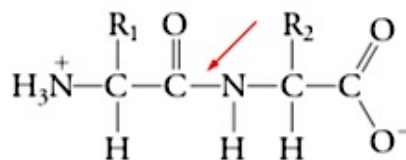


- **Amino Acids with Charged Side Chains** (4 amino acids only): Side chains are always charged under physiological conditions.
- (Asp) and (Glu), are acidic (COOH)
- (Lys) and (Arg) are positively charged (NH₃)

13

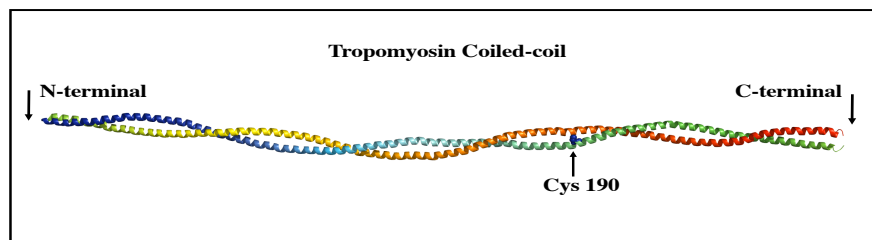
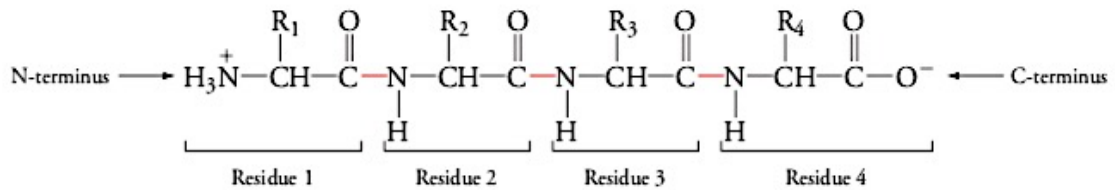
Proteins:

- Polymers of amino acids are called **polypeptides** or **proteins**.
- There are 20 different amino acids make building blocks for proteins.
- proteins may contain many hundreds of amino acid residues.
- The amino acid residues are linked to each other by amide bonds called peptide bonds



14

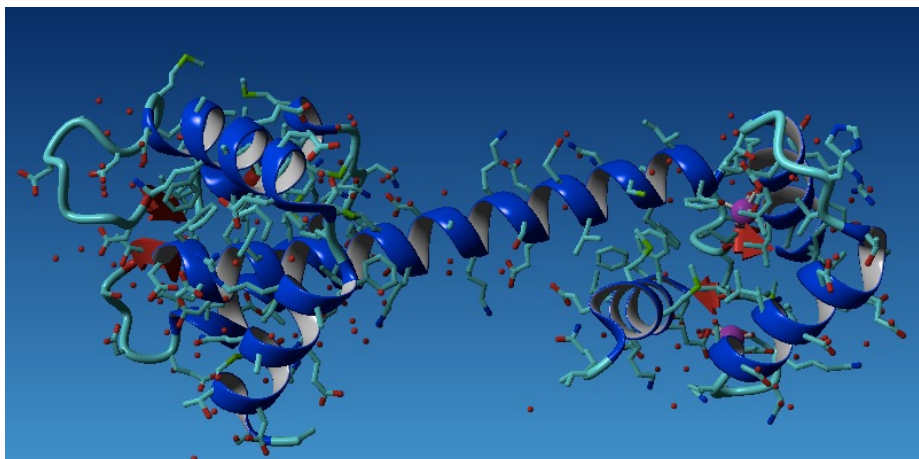
The peptide bond in the “backbone” starts with N-terminus and ends with C-terminus.



For example: Tropomyosin structure is a protein (with N- & C-terminus).

15

How many atoms are there? How many amino acids? How many polypeptides?

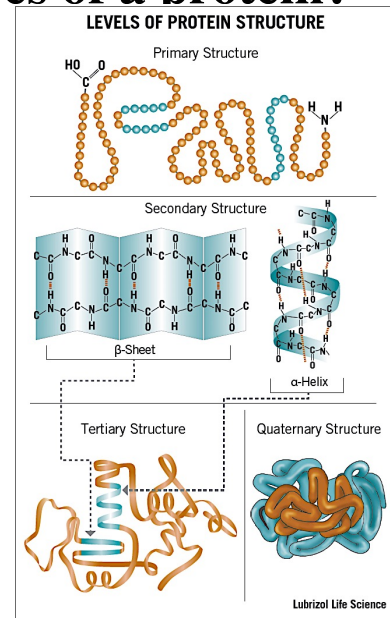


16

What are the different structures of a protein?

1. Primary structures (simple form)
2. Secondary structures (alpha helix and beta sheets) (3D)
3. Tertiary structures
4. Quaternary structures.

They differ by the types of bonds that connect each others.



17

The Origin and Evolution of Life

- Modern prokaryotic and eukaryotic cells apparently evolved from simpler non-living systems.
- What are the three domains of life? They are **bacteria, archaea, and eukaryote.**
- What is the ability of the cells? Cells have ability to **replicate** (make a replica or copy of itself) is one of the universal characteristics of living organisms.
- There are two main types of cells:

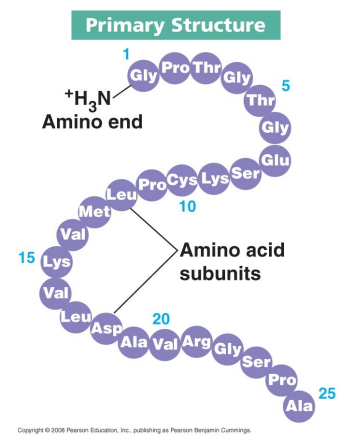
1- Prokaryotes.

2- Eukaryotic.

18

What is the Primary structure of a protein?

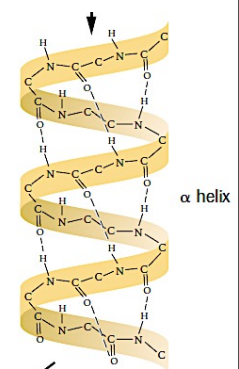
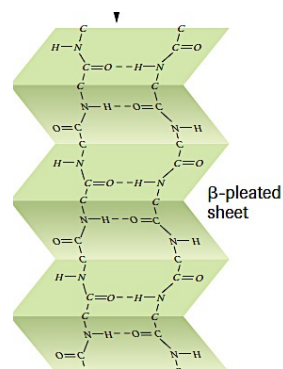
- The Primary structures of protein are:
- Linear sequence of amino acids
- This linear sequence is referred to as a polypeptide chain. The amino acids in the **primary structure** are held together by covalent bonds.



19

What is the secondary structure of protein?

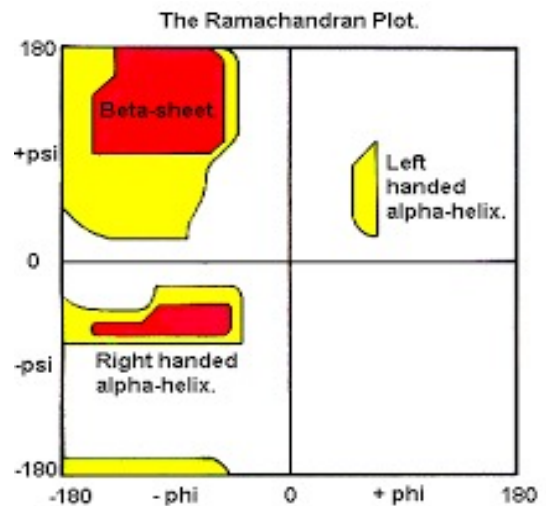
- The Secondary structures of protein are:
- Mostly consists of 2 types which are the α helix and the β pleated sheet.
- Other 2nd str:
- Both **structures** are held in shape by hydrogen bonds, which form between the carbonyl O of one amino acid and the amino H of another



20

The chain conformation of a polypeptide can be determined by?

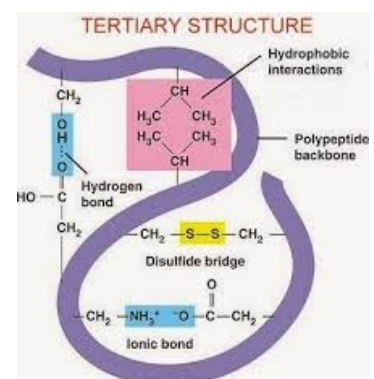
- It can be determined by **the torsion angles** around the $C\alpha-N$ binding (ϕ) and the $C\alpha-C$ binding (ψ) of the constituent amino acid residues.
- A **Ramachandran** plot is a conformation chart of those values that are sterically possible for ϕ and ψ .
- It determines the alpha helices and the beta sheets contents of a protein.



21

What are the tertiary & quaternary structures of protein?

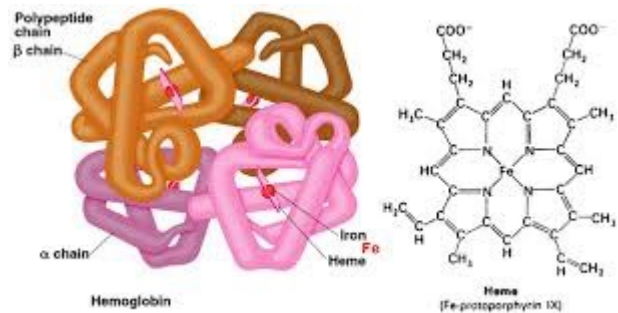
- Tertiary: It is 3D structures shape of protein.
- It has a single polypeptide chain "backbone" with one or more protein secondary structures that form the protein domain.
- Bonds?



22

What is the Quaternary structure of proteins?

- Quaternary:
- It is an arrangement of multiple folded protein subunits in a multi-subunit complex.
- It involves at least 2 polypeptides (domains).
- It can be a dimer, tetramer, homo or hetero protein.



23

Next lecture,

- We will continue explaining biomolecules, the bases bioinformatics study....
- We will talk about DNA and RNA molecules...

24

Bioinformatics I

Lecture 2: From genes to proteins

Dr Manaf A Guma

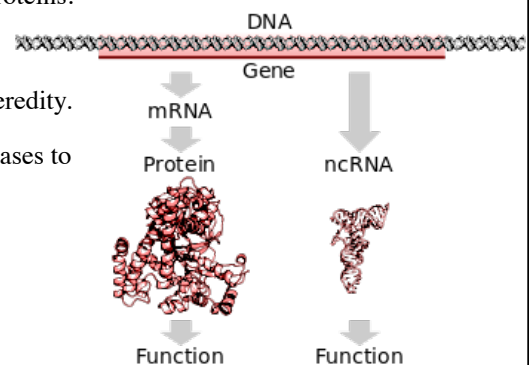
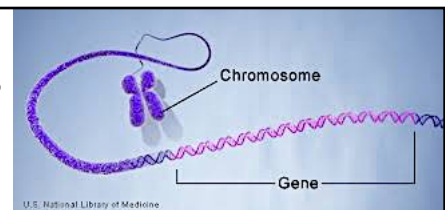
University Of Anbar- College Of Applied Sciences-hit
Department Of Applied Chemistry



1

What is a gene?

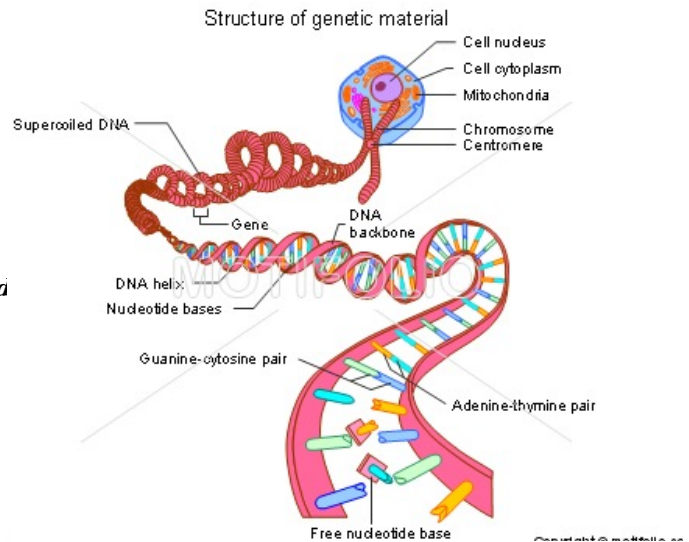
- **Genes are made up of DNA.**
- In biology, a **gene is a sequence of nucleotides in DNA or RNA that encodes the synthesis of a gene product, either RNA or protein**
- Some **genes** act as instructions to make molecules called proteins. However, many **genes** do not code for proteins.
- Gene: A **gene** is the basic physical and functional unit of heredity.
- In humans, **genes** differ in size from a few hundred DNA bases to more than 2 million bases.



2

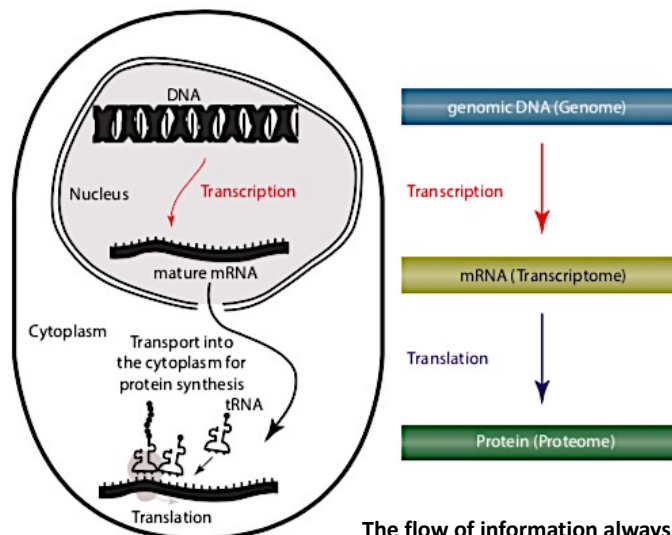
DNA is the Genetic Material

- ***DNA is a part of a Gene.***
- ***It is an (inheritance material) that is transcribed into RNA (transcription), translated to a protein (translation in the ribosome).***
- ***In chromatin, DNA is tightly coiled around histones in the nucleus.***
- ***The process is called “Central dogma of protein synthesis”.***



3

The central dogma of molecular biology



The flow of information always proceeds from the genome to the proteome, not vice versa.

4

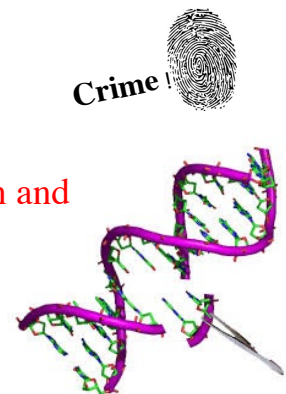
What does DNA / RNA mean ?

- **DNA is Deoxy ribo Nucleic Acid.**
- **RNA is RiboNucleic Acid.**
- **The name is based on the type of sugar molecules attached to the nucleic acid.**

5

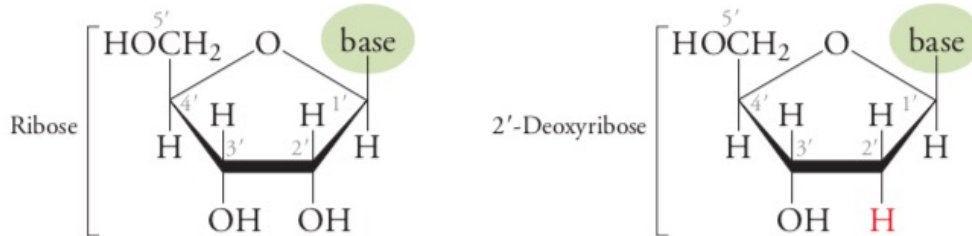
What is the units of the DNA ?

- **Nucleic acids** only four different types of structural units of DNA.
- DNA was known to contain chains of polymerized nucleotides—abbreviated A, C, G, and T—but these were thought to occur as simple repeating tetranucleotides.
- For example—ACGT-ACGT-ACGT-ACGT—
- **The DNA structure ultimately elucidated by James Watson and Francis Crick in 1953 !**
- Now: it is a Fingerprinting for each “species”



6

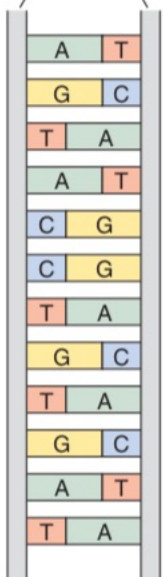
- So, DNA contains the bases A, C, G, and T, whereas RNA contains A, C, G, and U.
- Linking a purine or a pyrimidine to a five-carbon sugar forms a **nucleoside**.
- *In DNA, the sugar is deoxyribose; in RNA, the sugar is ribose .*



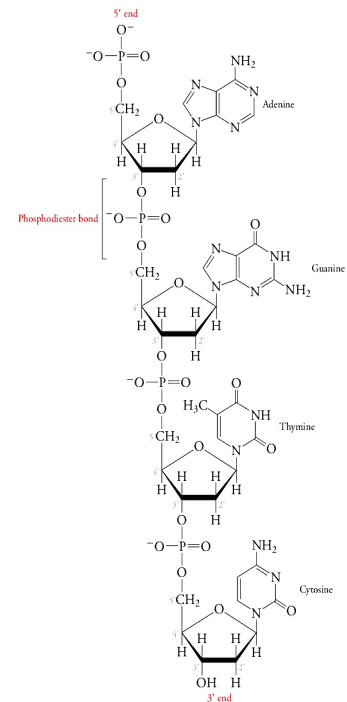
7

Sugar-phosphate backbones

DNA is a double helix

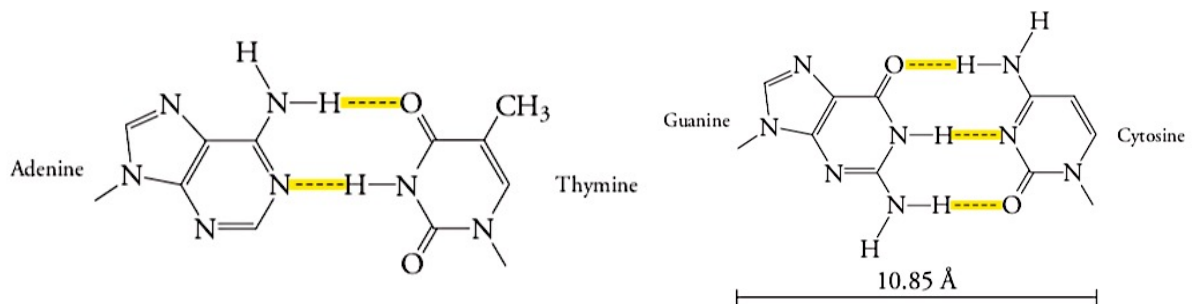


- Nucleotides are linked by **phosphodiester bonds** form a polymer.
- There are hydrophobic interactions between stacked bases.
- The end of the polymer that carries a phosphate group attached to C5' is known as the **5' end**, and the end that carries a free OH group at C3' is the **3' end**.
- **The sequence is red from 5' end to 3' end.**



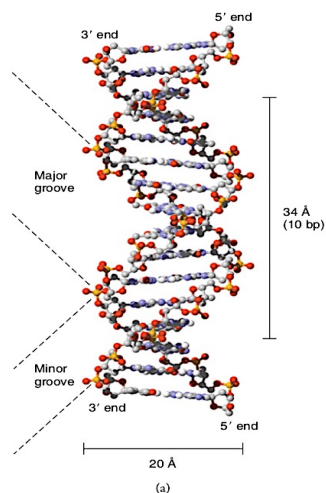
8

- DNA contains two polynucleotide strands whose bases pair through **hydrogen bonding**.
- Two hydrogen bonds link adenine and thymine, and three hydrogen bonds link guanine and cytosine:



9

- The diffraction (scattering) of an X-ray beam by a DNA fibre suggested a (helix)'' spiral'' with a repeating spacing of 3.4 Å.

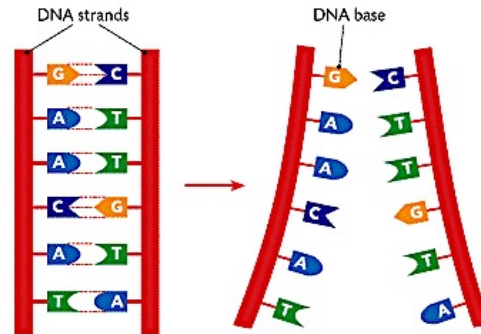


- *The size of a DNA is expressed in units of base pairs (**bp**) or kilo- base pairs (1000 bp, (**kb**)).*

Model of DNA. (a) Ball-and-stick model with atoms colored: C gray, O red, N blue, and P gold (H atoms are not shown). (b) Space-filling model with the sugar-phosphate backbone in gray and the bases color-coded: A green, C blue, G yellow, and T red.

10

Each DNA strand has a complementary strand.



Forward = Reverse Complement (Forward)

```

AGCTTCTAGTCGACTAGAAGCT
|| ||||| || ||||| || ||||| ||||| ||||| ||||| ||||| ||||| ||||| ||||| |||||
AGCTTCTAGTCGACTAGAAGCT
    
```

https://www.bioinformatics.org/sms/rev_comp.html
https://www.genscript.com/sms2/rev_comp.html

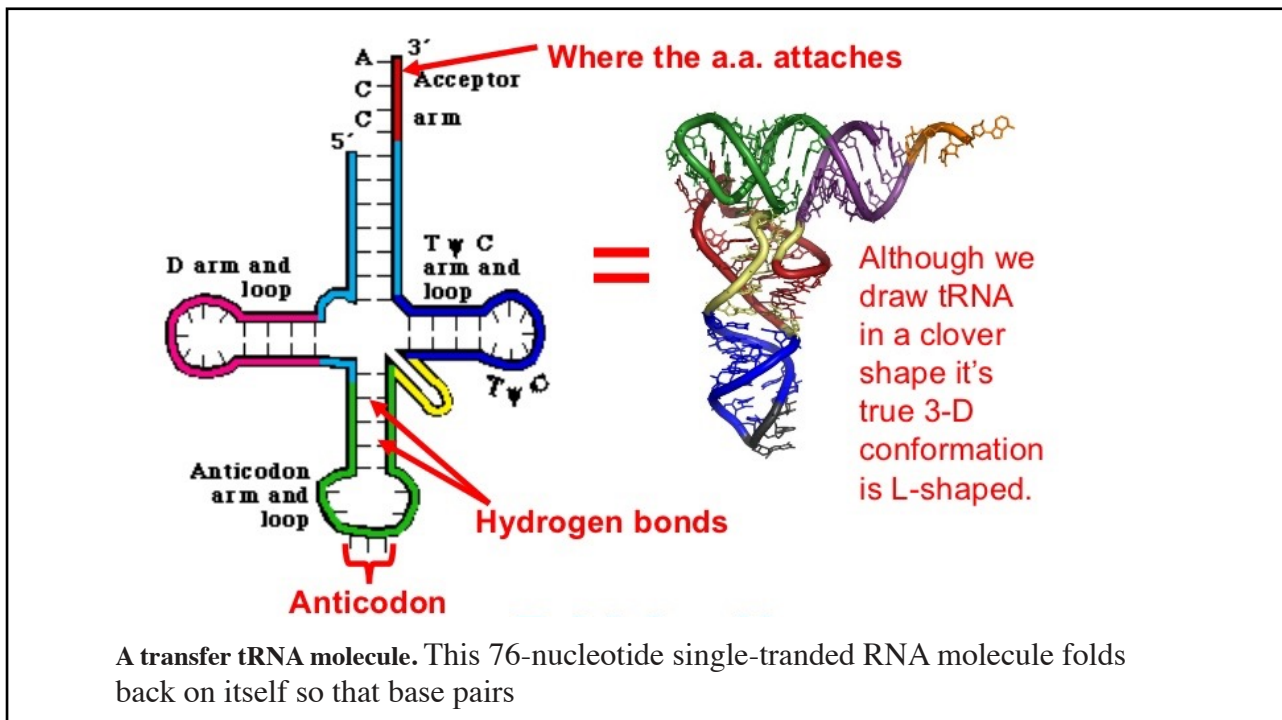
From Thomas Krahn's presentation.

11

RNA is a single-stranded

- RNA, which is a single-stranded polynucleotide, has greater conformational freedom than DNA.
- What are the types of RNA?
 1. The transcribed RNA is known as **messenger RNA (mRNA)** because it carries the same genetic message as the gene.
 2. The mRNA is translated in the **ribosome**, a cellular particle consisting of protein and **ribosomal RNA (rRNA)** (discussed later !).
 3. At the ribosome, small molecules called **transfer RNA (tRNA)**, which carry amino acids, recognize sequential sets of three bases (known as **codons**)

12



13

So far, what did we understand?

- There are two types of sequences which are constructed by a biomolecules:
- Nucleotides sequences:
 1. RNA seq: such as AUUGCCGGCUUUA
 2. DNA seq: such as ATTGCCGGCTTTA
 3. Proteins sequences: : such as ASMDAIKKKMQLKLDKENALD
- These two sequences are the base of bioinformatics that we are going to deal with.

14

Review ?

- Exercise 1.1
What is the difference between the two polynucleotides DNA and RNA?
- Exercise 1.2
DNA consists of two complementary nucleotide strands. Which base pairings are observed between these two nucleotide strands?
- Exercise 1.3
What is the meaning of the terms genome, transcriptome, and proteome?
- Exercise 1.4
The 20 naturally occurring amino acids are encoded by base triplets (codons) of the genetic code. Which consideration led to the discovery of the triplet codon organization of the genetic code?
- Exercise 1.5
Build the genetic code of your name. If this is not possible, use the name CRICK.
- Exercise 1.6
What is meant by the central dogma of molecular biology?
- Exercise 1.7
What is meant by the term splicing, and how does this process contribute to the discrepancy between the relatively low number of genes in the human genome but the larger number of proteins actually produced?

15

Review ?

- Exercise 1.8
Which amino acids show the following properties: (A) hydrophobic, polar, and small and (B) hydrophobic and aliphatic?
- Exercise 1.9
In which direction is the primary structure of proteins read?
- Exercise 1.10
Which structural elements can be found in the secondary structure of proteins?
- What is meant by the central dogma of molecular biology?
- Describe the differences between prokaryotes and eukaryotes cells.
- What are the functions of each particles in each in the cell ?
- Such as mitochondria? Cytoplasm? ...etc

16

REVIEW ?

1. How does DNA encode genetic information and how is this information expressed?
2. What is the relationship between the nucleotide sequence in a gene and the amino acid sequence of a protein?
3. List some reasons why knowing a gene's sequence might be useful.
4. How has the DNA sequence changed and how does this
5. affect the encoded protein?

Bioinformatics I

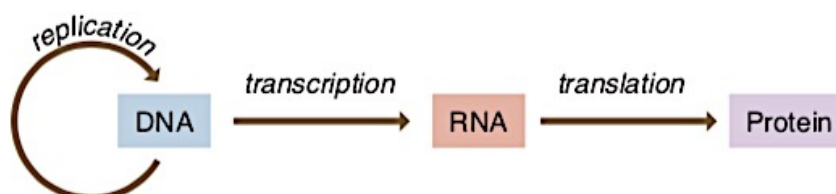
Lecture 3: Introduction to genetic materials

Dr Manaf A Guma
University of Anbar- college of applied sciences-Hit
Department of chemistry

1

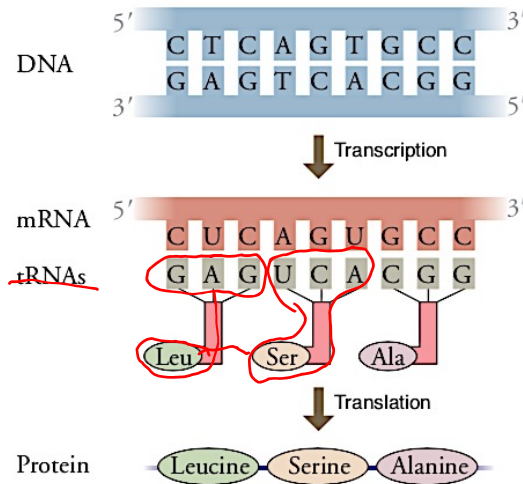
Protein synthesis: Genes Encode Proteins !

- DNA is essential for inheriting the genetic information.
- The complete set of genetic information of an organism is called its **genome**
- *A part of the DNA, a **gene**, is **transcribed** to produce a complementary strand of RNA; then the RNA is translated into protein.*



2

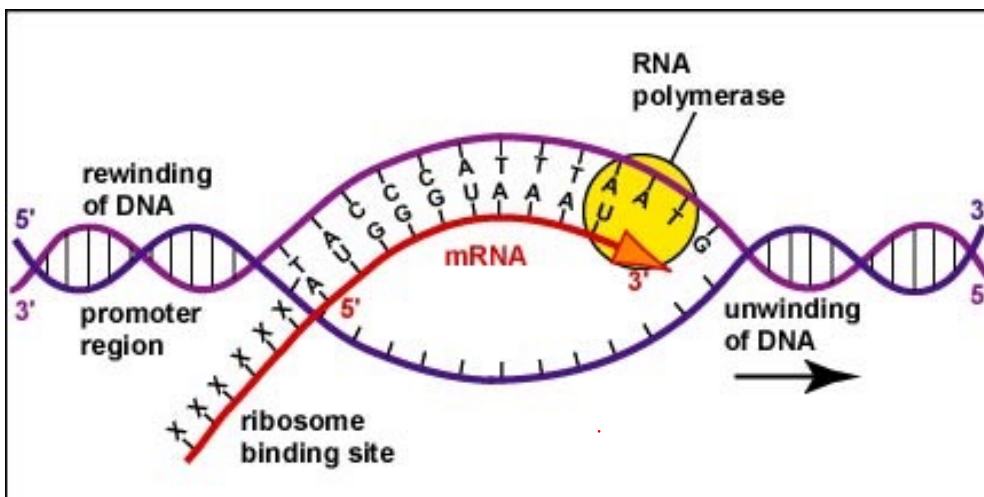
The protein's amino acid sequence therefore depends on the nucleotide sequence of the DNA.



The correspondence between amino acids and mRNA codons is known as the **genetic code**.

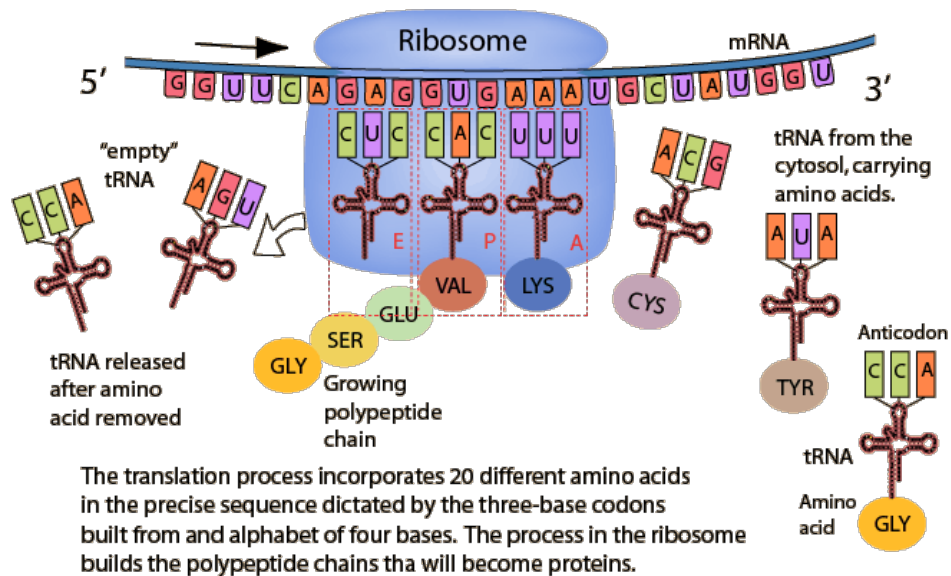
3

From DNA to RNA: Transcription (making a copy)



4

The translation to a protein



5

So, How is protein synthesized:

- Gene expression begins with the process called **transcription**, which is the synthesis of a strand of mRNA that is complementary to the gene of interest.
- A region of DNA un-winds and then the two strands separate.
- However, only that small part of the DNA will be split apart.
- The triplets within the gene on this section of the DNA molecule are used as the template to transcribe the complementary strand of RNA.

6

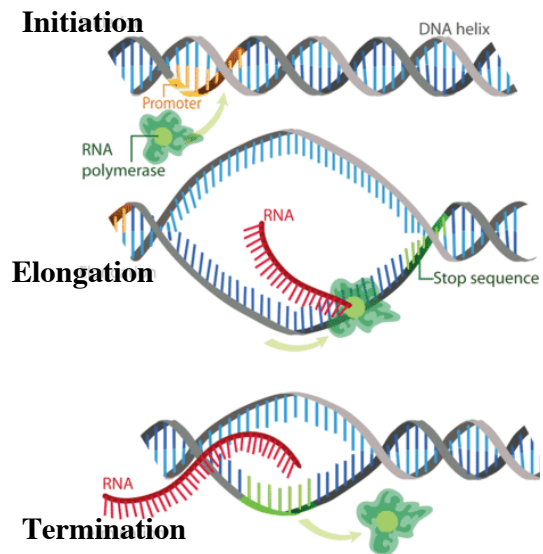
Three stages of transcription

a. *Stage 1: Initiation:* by a promoter.

b. *Stage 2 : Elongation:* RNA

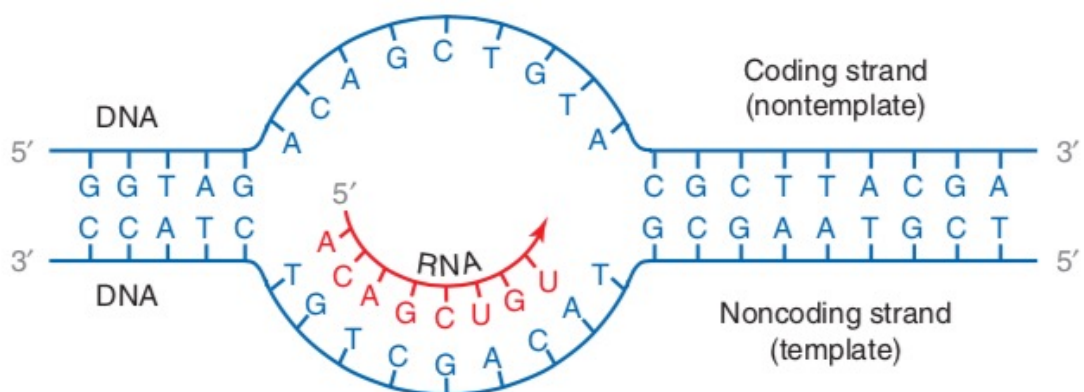
polymerase is an enzyme that adds new nucleotides to a growing strand of RNA.

c. *Stage 3: Termination:* (UAA, UAG, or UGA) codes a “stop codons”.



7

Transcription of RNA from the DNA



8

- There are a total of 64 codons: 3 of these are “stop” signals that terminate translation, and the remaining 61 represent, with some redundancy, the 20 standard amino acids found in proteins

First Position (5' end)	Second Position				Third Position (3' end)				
	U	C	A	G					
U	UUU	Phe	UCU	Ser	UAU	Tyr	UGU	Cys	U
	UUC	Phe	UCC	Ser	UAC	Tyr	UGC	Cys	C
	UUA	Leu	UCA	Ser	UAA	Stop	UGA	Stop	A
	UUG	Leu	UCG	Ser	UAG	Stop	UGG	Trp	G
C	CUU	Leu	CCU	Pro	CAU	His	CGU	Arg	U
	CUC	Leu	CCC	Pro	CAC	His	CGC	Arg	C
	CUA	Leu	CCA	Pro	CAA	Gln	CGA	Arg	A
	CUG	Leu	CCG	Pro	CAG	Gln	CGG	Arg	G
A	AUU	Ile	ACU	Thr	AAU	Asn	AGU	Ser	U
	AUC	Ile	ACC	Thr	AAC	Asn	AGC	Ser	C
	AUA	Ile	ACA	Thr	AAA	Lys	AGA	Arg	A
	AUG	Met	ACG	Thr	AAG	Lys	AGG	Arg	G
G	GUU	Val	GCU	Ala	GAU	Asp	GGU	Gly	U
	GUC	Val	GCC	Ala	GAC	Asp	GGC	Gly	C
	GUA	Val	GCA	Ala	GAA	Glu	GGA	Gly	A
	GUG	Val	GCG	Ala	GAG	Glu	GGG	Gly	G

- U=Uracil which is a nucleotide found in the RNA only.
- We read it in the DNA seq. as T=Thymine.

A **codon** is a three-base sequence of mRNA, so-called because they directly encode amino acids.

9

Translation DNA to a protein ?! Is it possible?

- You can also find the DNA sequence for a specific protein using the following web site:
- <https://www.ncbi.nlm.nih.gov/gene>
- This is the opposite to the translation of DNA to proteins ?why?
- What are the possibilities for converting a protein to DNA based on the codon chart.
- Is it possible? What do you know about **codon optimization?**

10

Translation DNA to protein using Expasy Translate website

- We can use the website: <https://web.expasy.org/translate/>
- You will have 6 translation for your query DNA sequence which are based on the frameshift.
- Be aware of the translation:
- You could have a problem while you copy and paste the sequence.
- You could the wrong frameshift. **ACTGCAGTGCAA**

11

Tutorial

- Translate the following DNA sequence to protein: how many possibilities you can get?
- CGAGCATGGACGCGATCAAGAAGAAGATGCAAATGCTGAAACTGGAC
AAAGAAAATGCGCTGGACCGTGCGGAACAGGCGGAGGCGGACAAGAA
AGCGGCGGAGGATCGTAGCAAGCAGCTGGAAGACGAGCTGGTGAGCC
- Can you do the opposite? Protein to DNA ?

12

Expasy tool param database to find the chemical properties for a peptide of a protein

- An engine can calculate the following parameters for the protein.

- <https://web.expasy.org/protparam/>

- You can calculate the the chemical properties for the following seq:

MDDIYKAAVEQLTEEQNEFKAAFDIFVLGAEDGCISTKELGKVMRMLGQNPTPEELQEM

IDEVDEGSGTVDFDEFVMMVRCMKDDSKGKSEELSDLFRMPDKNADGYIDLDELKIM

LQATGETITEDDIEELMKDGDKNNDGRIDYDEFLEFMKGVE

1. The umber of amino acids: Amino acid composition and the negative and the positives amino acids.
2. Molecular weight:
3. Theoretical pI ?
4. Formula. Like $C_{46}H_{61}N_{11}O_{18}$
5. Total number of atoms.
6. Extinction coefficient ? What is it ? Next page

13

How to calculate extinction coefficient for a peptide or a protein?

- The **extinction coefficient** is the absorbance divided by the concentration and the pathlength, according to Beer's Law: $A=a*b*c$ (a= molar absorptivity, b= length of light bath, c=concentration)
- ϵ (epsilon = absorbance/concentration/pathlength).
- The **units** of **extinction coefficients** are usually $M^{-1}cm^{-1}$, but for **proteins** it is often more convenient to use $(mg/ml)^{-1}cm^{-1}$.
- **a= Molar Extinction Coefficient "molar absorptivity"** = (Number of Tryptophan residues X 5500) + (Number of Tyrosine residues X 1490)= $gm/l = A0.1\% mg/ml$
- And then divide to check?

14

Bioinformatics I

Lecture 4: Gene Expression

Dr Manaf A Guma
University Of Anbar- college of Applied sciences-Hit
Department of applied chemistry



1

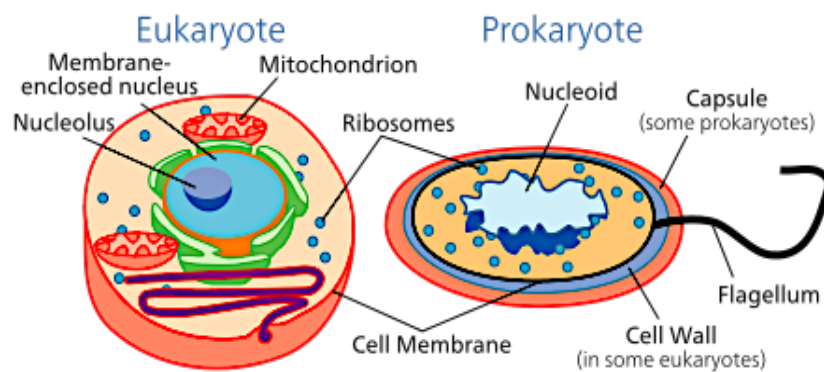
What is a gene expression?

- **Gene expression** is the process by which the information encoded in a **gene** is used to direct the assembly of a **protein** molecule.
- The cell reads the sequence of the **gene** in groups of three bases (as explained in the last lecture).
- In prokaryotic and **eukaryotes**, gene expression is regulated differently.
- **In prokaryotic**, Gene expression is regulated primarily at the transcriptional level.
- **In eukaryotes**, Gene expression is regulated at many levels (epigenetic, transcriptional, nuclear shuttling, post-transcriptional, translational, and post-translational).

2

Eukaryotic and Prokaryote

- *They are usually larger than prokaryotic cells.*
- *It contain a nucleus and other membrane-bounded cellular compartments (such as mitochondria, chloroplasts, and endoplasmic reticulum).*



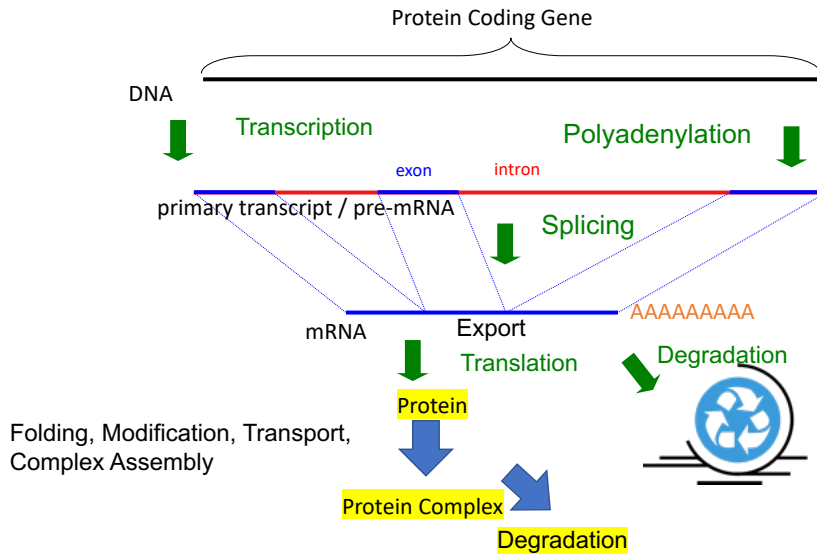
3

The most differences between prokaryotic and eukaryotes genes ?

- The most striking difference is that prokaryotic gene information is encoded on a continuous DNA stretch.
- In eukaryotes, coding exons are interrupted **by noncoding introns.**
- Eukaryotic transcription of DNA to mature mRNA (containing information derived only from exons) requires several steps.
- The introns are removed during the process of splicing.
- **What is meant by the term splicing?**
- Through alternative splicing (removing and joining different introns and exons), different mRNAs and, consequently, different proteins can result from one gene

4

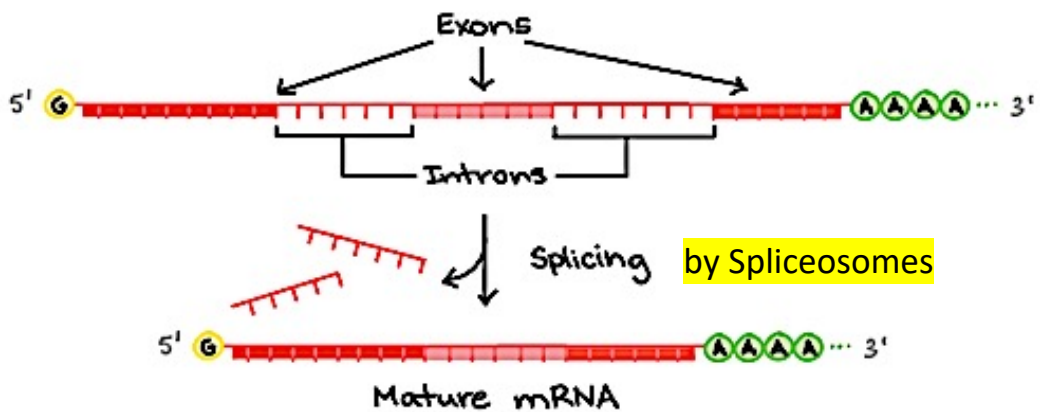
Describe the Expression of a Typical Eukaryotic Gene?



5

5

Draw the process of RNA splicing?



6

How does RNA splice?

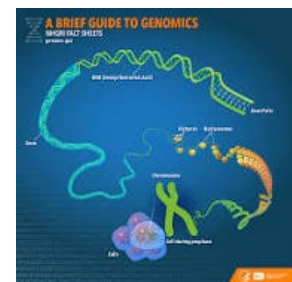
- For nuclear-encoded genes, **splicing** takes place within the nucleus either during or immediately after transcription. For those eukaryotic genes that contain introns, **splicing** is usually required in order to create an mRNA molecule that can be translated into protein.
- In **splicing**, some sections of the **RNA** transcript (introns) are removed, and the remaining sections (exons) are stuck back together.
- Some genes can be alternatively **spliced**, leading to the production of different mature mRNA molecules from the same initial transcript.
- Regulation of splicing therefore represents a critical step of gene expression.

<https://dnalc.cshl.edu/resources/3d/rna-splicing.html>

7

What is a genome ?

- In the fields of molecular biology and genetics:
- A **genome** is the genetic material of an organism. It consists of DNA.
- **The genome includes both the genes and the noncoding DNA.**
- It also includes mitochondrial DNA and chloroplast DNA.
- The study of the genome is called **genomics**.
- So, **Genome** is the entire DNA sequence of an organism.
- Meaning: the genes and all of non-coding DNA is in between.



8

Genomics

- **KEY CONCEPTS**
- **The genomes of different species vary "differ" in size and number of genes.**
- **Genes can be identified by their nucleotide sequences.**
- **Analysis of genetic data can provide information about gene function and risk of disease.**

NCBI Nucleotide search interface showing results for Staphylococcus aureus DNA polymerase I (polA) gene.

Search: Nucleotide for [] Go Clear

Display: GenBank Show 5 Send to Hide: sequence all but gene, CDS and mRNA features

Range: from begin to end Reverse complemented strand Features: + Refresh

1: AF193842. Reports Staphylococcus au...[gi:6110604] Links

Features Sequence

LOCUS AF193842 2631 bp DNA linear BCT 19-JAN-2000

DEFINITION Staphylococcus aureus DNA polymerase I (polA) gene, complete cds.

9

How Does Evolution Work?

It works by analysing the sequences of nucleotides in certain genes that are present in all species.

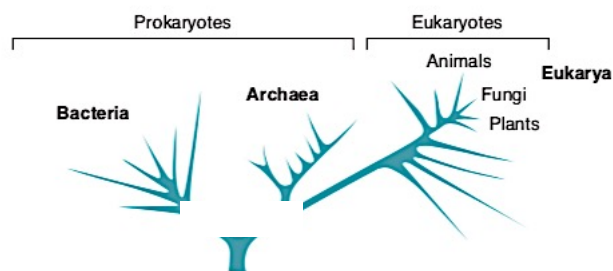


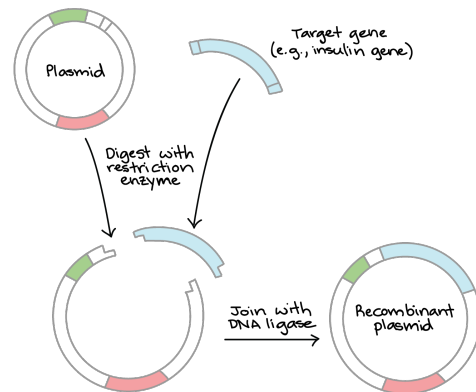
Figure 1-15 Evolutionary tree based on nucleotide sequences. This diagram reveals that the bacteria separated before the archaea and eukarya diverged. Note that the closely spaced fungi, plants, and animals are actually more similar to each other than are many groups of prokaryotes. [After Wheelis, M. L., Kandler, O., and Woese, C. R., *Proc. Natl. Acad. Sci. USA* 89, 2930-2934 (1992).]

It is possible to construct a diagram that indicates how the **bacteria, archaea, and eukaryote** are related.

10

Bioinformatics involves Biotechnology

- **What is a DNA cloning:**
- It is making an identical copy for an organism.
- It refers to the process of isolating a DNA sequence of interest for the purpose of making multiple.
- In labs, vectors are used as a host to make an identical copy for a specific gene.
- Then, this gene is hosted in *E. coli* produce protein (outside of the living body).
- In vivo or in vitro?



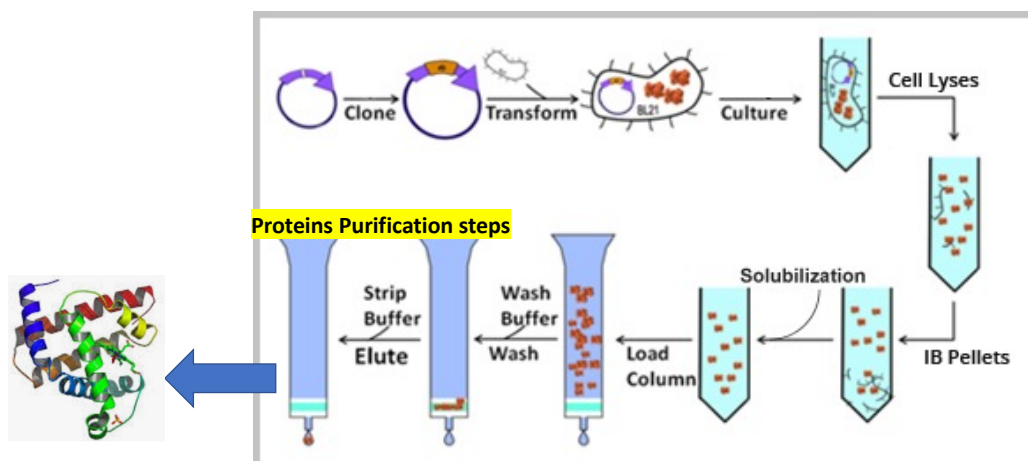
11

We can do cloning in the Prokaryotic

- They are small unicellular organisms. No real nucleus. (usually just called **bacteria**)
Exemplified by *E. coli*, and the **archaea**.



E. coli



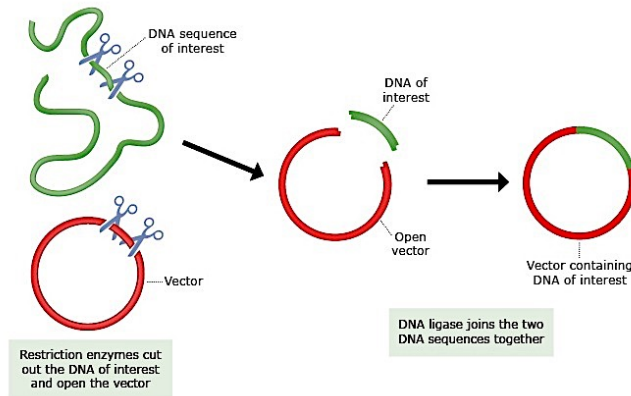
Study the Protein structure and function

12

What is a genetic engineering?



- It is a removal of genes from one organism and insertion into another.



© The University of Waikato Te Whare Wānanga o Waikato | www.sciencelearn.org.nz

13

What is recombinant DNA?

- It is DNA has been mixed with one of another species.



14

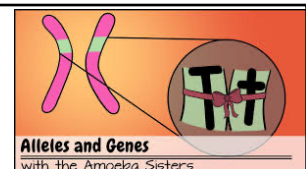
Bioinformatics I

Lecture 5: Wild-types and mutants

Dr Manaf A Guma
University Of Anbar- College Of Applied Sciences-hit
Department Of Applied Chemistry

1

What is a genotype?

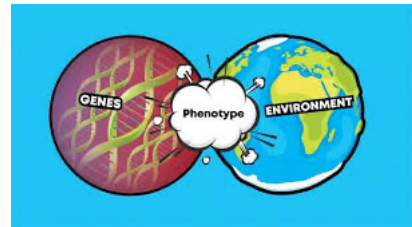
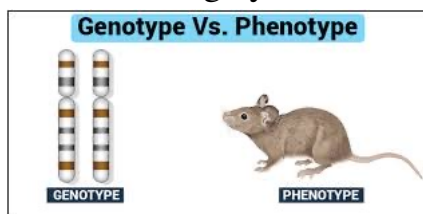


- The term "genotype" refers to the genetic makeup of an organism; in other words, it describes an organism's complete set of genes.
- It can be used to refer to the alleles.
- humans are diploid organisms, which means that they have two alleles at each genetic position, or locus, with one allele inherited from each parent. Each pair of alleles represents the genotype of a specific gene.

2

What is a phenotype?

- A **phenotype** is an individual's observable traits, such as height, eye colour, and blood type.
- The genetic contribution to the **phenotype** is called the genotype.
Some traits are largely determined by the genotype, while other traits are largely determined by environmental factors.

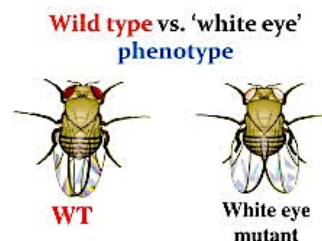


3

Wild-Type WT and mutations



- **Wild-Type WT** (the natural **phenotype** was expressed from a genotype)
- **Wild type (WT)** refers to the phenotype of the typical form of a species as it occurs in nature.
- **Species** a group of living organisms consisting of similar individuals.
- Such as Human, rat, rabbits, pigs....
- A **mutation** is a replacement or substitution in any gene that can be expressed into a different amino acid which can change the function or the structure of the proteins. It is named as (**mutant**).
- It can be appeared **naturally**.
- The **mutagenesis** can be done in the lab.
- Note: not all the mutations are Harm !



4

A mutated gene can cause disease !

- A **mutation** is a change that occurs in our DNA sequence, either due to mistakes when the DNA is copied or as the result of environmental factors such as UV light and cigarette smoke etc.

<i>Normal gene</i>	... ACT CCT	GAG	GAG	AAG ...
Protein	... Thr – Pro	Glu	Glu	– Lys ...
<i>Mutated gene</i>	... ACT CCT	GTG	GAG	AAG ...
Protein	... Thr – Pro	Val	Glu	– Lys ...

In the gene for that protein chain, the normal GAG codon has been **mutated** (altered) to GTG.

5

How do you name the substituted mutation ?

- From the previous example:
- Substitution is the main cause of mutations.
- To give a name:
- Glutamic acid (site =3) Valine
- OR: **Glu3Val**, OR: **E3V** (very common name to a mutant).



	1	2	3	4	5
Protein	... Thr – Pro	Glu	Glu	– Lys ...	
Mutation	... Thr – Pro	Val	Glu	– Lys ...	

6

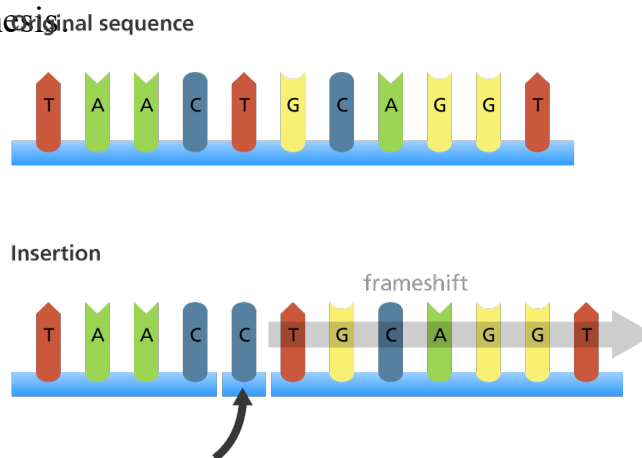
How do you check the mutation

WT	L	E	D	E	L	V	S	L	Q	K	K	L	K	G	T	E	D	E		
	CTG	GAA	GAC	GAG	CTG	GTG	AGC	CTG	CAA	AAG	AAA	CTG	AAG	GGC	GGC	GGC	GGC	GGC		
	GAC CTT CTG CTC GAC CAC TCG GAC GTT CTC TTT GAC TTC																			
Mutant	L	E	D	E	L	V	S	L	Q	E	K	L	K	G	T	E	D	E		
	CTG	GAA	GAC	GAG	CTG	GTG	AGC	CTG	CAA	GAG	AAA	CTG	AAG	GGC	GGC	GGC	GGC	GGC		
										AGC	CTG	CAA	GAG	AAA	CTG	AAG	GGC	GGC	GGC	GGC

7

Types of Mutations that could appear during the synthesis of proteins; such as insertion

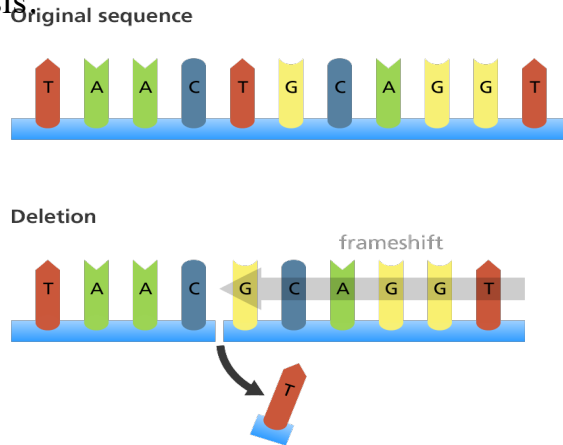
- **Insertion:** when a single nucleotide or more are inserted during the protein synthesis



8

Other factors could lead to Mutations; such as deletion

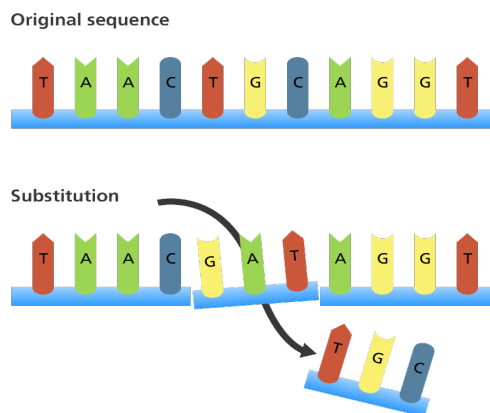
Deletion 1: when a single nucleotide or more are deleted during the protein synthesis.



9

Other factors could lead to Mutations; such as deletion

Deletion 2: when a single nucleotide or more are substituted during the protein synthesis.



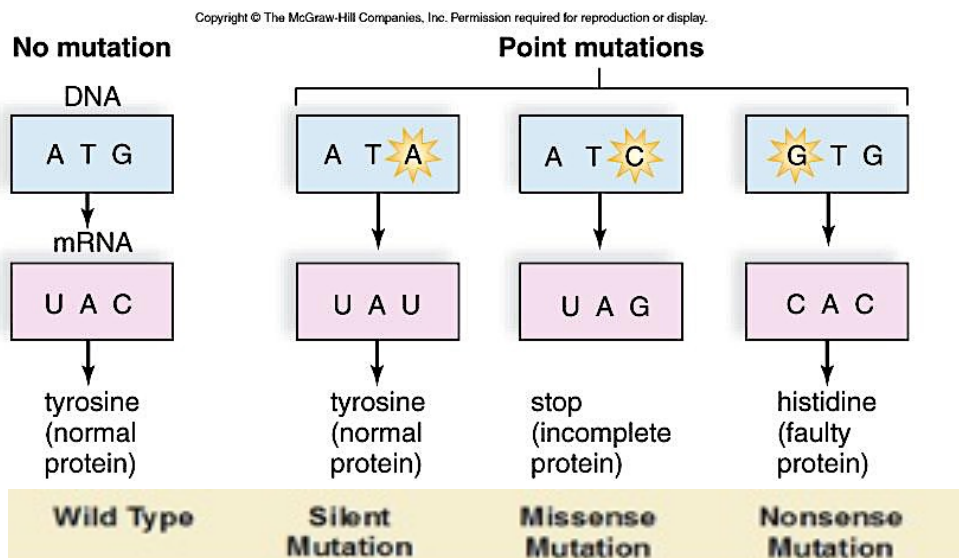
10

What is the difference between point mutation and frameshift mutation?

- Point mutation is an alteration of **a single nucleotide** in a gene whereas frameshift mutation involves **one or more nucleotide changes** of a particular gene.
- Point mutations are mainly nucleotide substitutions, which lead to silent, missense or nonsense mutations. Frameshift mutations occur by insertion or deletion of nucleotides.

11

What are the types of Mutations?



12

Define?

- **Nonsense Mutations:** the alteration of a nucleotide in a particular codon may introduce a stop codon to the gene. This stops the translation of the protein at halfway of the complete protein.
- **Silent mutations,** a single base pair has changed in a particular **codon**, the same **amino acid** is coded by the altered codon as well.
- **Missense mutations,** once the alteration occurs in a particular codon by a nucleotide substitution, the codon is altered in such a way to code a different amino acid.

13

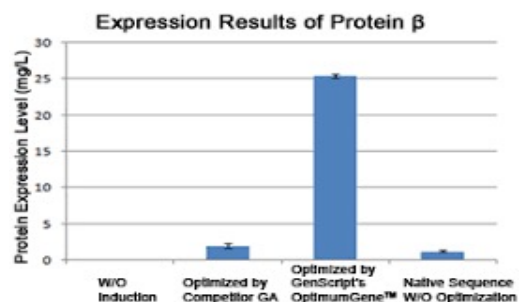
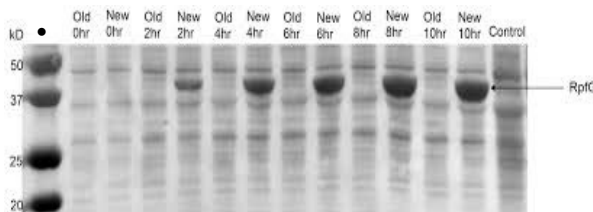
What other factors could cause mutation?

- **There are also environmental causes for mutations:**
- **Substitution one or more bp (less problematic than others).**
- **Errors in DNA Replication.**
- **Errors in DNA Recombination.**
- **Chemical Damage to DNA.**
- **Radiation.**

14

What can silent mutation be used for? What is a codon optimization?

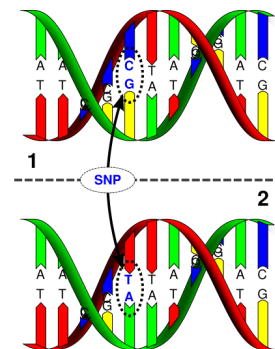
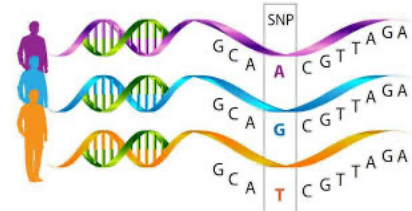
- It is a designing for a DNA sequence that could result in a same protein sequence by a chemical synthesis but with high protein expression. So, the propose of that:
- It is to increase the level of protein expression. For example:
- M D A I K
- ATG GAC GCG ATT AAG
- ATG GAC GCT ATC AAA



15

What is a SNP? Single nucleotide polymorphisms

- SNPs: It is a variation in a genetic sequence that affects only one of the basic building block adenine (A), Guanine (G), Thymine (T) or Cytosine (C) in a segment of a DNA molecule and that occurs in more than 1% in the population.
- For example, at a specific base position in the human genome, the C nucleotide may appear in most individuals, but in a minority of individuals, the position is occupied by an A. This means that there is a SNP at this specific position.



16

Bioinformatics I

Lecture 6: Biological Databases

Dr Manaf A Guma

University Of Anbar- college of Applied sciences-Hit

Department of applied chemistry

1

Biological Knowledge is Stored in Global Databases

- The most important basis for applied bioinformatics **is the collection of sequence data** and its associated to biological information.
- For example, with genome sequencing projects such data are generated daily in very large quantities worldwide.
- Furthermore, for a number of databases, original articles describe their functions.

2

What are the data bases that used for bioinformatics ?

1. Primary databases contain primary sequence information (nucleotide or protein).
 2. Secondary biological databases, however, summarize the results from analyses of primary protein sequence databases.
- The aim of these analyses is to derive common features for sequence classes, which in turn can be used for the classification of **unknown sequences (annotation)**.

3

What are the Primary Databases?

- **First: Nucleotide Sequence Databases.**
- 1- GenBank <https://www.ncbi.nlm.nih.gov/genbank/>
- The GenBank database [genbank] is perhaps the best-known nucleotide sequence database available at the U.S. National Center for Biotechnology Information (NCBI) [ncbi].
- It is associated with other databases, for example the European Nucleotide Archive (ENA) or the DNA Database of Japan (DDBJ).

4

How to find the primary sequence of a gene using ncbi?

- Open the link: <https://www.ncbi.nlm.nih.gov/genbank/>
- Search for a specific gene Troponin C, for example.
- What do you find?
- describe the results..?

5

The figure shows a GenBank entry

GenBank: AJ419175.1
[FASTA](#) [Graphics](#)

Go to:

LOCUS AJ419175 470 bp mRNA linear INV 26-JUL-2016
 DEFINITION Ostertagia ostertagi partial mRNA for troponin (trp gene).
 ACCESSION AJ419175
 VERSION AJ419175.1
 KEYWORDS troponin; trp gene.
 SOURCE Ostertagia ostertagi
 ORGANISM *Ostertagia ostertagi*
 Eukaryota; Metazoa; Ecdysozoa; Nematoda; Chromadorea; Strongylida;
 Trichostrongyloidea; Haemonchidae; Ostertagia.

REFERENCE 1
 AUTHORS Geldhof,P., Vercauteren,I., Knox,D., Demaere,V., Van Zeveren,A.,
 Berx,G., Verduyts,J. and Claerebout,E.
 TITLE Protein disulphide isomerase of Ostertagia ostertagi: an
 excretory-secretory product of L4 and adult worms?
 JOURNAL Int. J. Parasitol. 33 (2), 129-136 (2003)
 PUBMED [12633650](#)

REFERENCE 2 (bases 1 to 470)
 AUTHORS Geldhof,P.B.
 TITLE Direct Submission
 JOURNAL Submitted (05-NOV-2001) Geldhof P.B., Parasitology, Ghent
 University, Salisburylaan 133, Merelbeke, BELGIUM

FEATURES
 Location/Qualifiers
 source 1..470
 /organism="Ostertagia ostertagi"
 /mol_type="mRNA"
 /db_xref="taxon:6317"
 /dev_stage="adult"
 gene <1..>470
 /gene="trp"
 CDS <1..>470
 /gene="trp"
 /codon_start=2
 /product="troponin"
 /protein_id="CAD11862.1"
 /db_xref="InterPro:IPR001978"
 /db_xref="UniProtKB/TrEMBL:Q95PN9"
 /translation="NFKINSKGEQAQFGLAGVVDGQTEQEEAKAFLAAVCR
 SVDISLLPNDLKERIKLHNRIKLEADKYDLEKRHERQYDMKELHERQVARNK
 ALKGLDPEEAASSQHPKITTASKFRQIDRRSYGDRRELFHPVKKPPTIA"

ORIGIN

 1 caatttcaag atcaattcca aaggcgagca ggcggcgag ttcgccaatc tggcacaagg
 61 agtaaaacaa gatggacaaa cgaagaaca gcaagaaga gccaggcag cgtttcttgc
 121 agccgcttgc cgttcagtg atctctcgtc gctgcttccg aacgatctga aggagcgaat
 181 caaaacgttg cataaccgaa tctgtaaatt ggaggccgat aagtatgat tggagaagcg
 241 ccatagagct caggaatatg acatgaaaga gctgcacgaa cgtcaacgcc aagttgccag
 301 gaacaagcgc ctcaaaaagg gactcgtacc tgaggagacc gcttcattct aacatctccc
 361 aaaaactcact accgcttcca agtttgatcg tcagattgac agaaggtctt atggagatcg
 421 acgagagctg tttagcacc cagtcaccaa gaagccacc accattgccc
 //

ORIGIN

6

Describe the results..? For tutorial !

- Each entry starts with the keyword LOCUS followed by a locus name.
- Like the AN, the locus name is also unique. Unlike the AN, it may change after revisions of the database.
- The locus name consists of eight characters, including the first letter of the genus and species names, in addition to a six-digit AN.
- A sequence must have at least 50 base pairs to be entered into GenBank.
- Every GenBank entry must contain coherent sequence information of a single molecule type, that is, an entry cannot contain sequence information of both genomic DNA and RNA.
- The last column in the LOCUS line gives the date of the last entry modification. The end of the database record starts with the keyword ORIGIN.

7

What are the Primary Databases?

- **2- Entrez:** <http://www.ncbi.nlm.nih.gov/Entrez/>.
- Query of the GenBank database is carried out via the NCBI Entrez system [entrez],
- [entrez] is used to query all NCBI-associated databases.
- Entrez is an important and effective tool for the execution of both simple and complicated searches for genes.
- To use this search, follow the link beneath the Entrez search field.

8

Field IDs to restrict research terms to certain database fields in the Entrez system ?

Field ID	Database field
ACC	Accession number
AU	Author name
DP	Publication date
GENES	Gene name
ORGN	Scientific and common name of the organism
PT	Publication type, e.g., review, letter, technical publication
TA	Journal name, official abbreviation, or ISSN number

9

What are other Primary Databases?

- **3- EMBL and DDBJ** dbgap. <http://www.ncbi.nlm.nih.gov/gap> and ddbj. <http://www.ddbj.nig.ac.jp/>
- The European counterpart to GenBank is the ENA [ena], located at the European Bioinformatics Institute (EBI) [ebi].
- The three database operators, NCBI, EBI, and NIG, compose the International Nucleotide Sequence Database Collaboration and synchronize their databases every 24 h.

10

The results of the DDBJ

The screenshot shows the DDBJ database interface for the study '1 NHLBI Framingham SABRe CVD'. The interface includes a navigation bar with tabs for 'Studies (1)', 'Phenotype Datasets (0)', 'Variables (0)', 'Molecular Datasets (0)', 'Analyses (0)', and 'Documents (19)'. Below the navigation bar are buttons for 'Save Results' and 'Save Query'. The main content area displays the following details:

Accession	phs000363.v18.p12
Parent study	Framingham Cohort (phs000007.v31.p12)
Study Disease/Focus	Cardiovascular Diseases
Study Design	Prospective Longitudinal Cohort
Study Markerset	HuEx-1_0-st, custom_probe_set
Study Molecular Data Type	miRNA Expression (Array), mRNA Expression (Array)
Study Content	21 phenotype datasets, 1404 variables, 73 documents, 3 molecular datasets, 7554 subjects, 11323 samples
NIH Institute	NHLBI
Study Consent	HMB-IRB-MDS --- Health/medical/biomedical (irb, mds) , HMB-IRB-NPU-MDS --- Health/medical/biomedical (irb, npu, mds)
Release Date	2020-03-30
Embargo Release Date	2020-03-30
Related Terms	Body System, Cardiovascular; Cardiovascular Body System; Cardiovascular Organ System; Cardiovascular System; Cardiovascular Systems ...

Below the table, there is a summary paragraph: 'This substudy phs000363 Framingham SABRe contains immunoassays, gene expression profiling, and microRNA data. Summary level phenotypes for the Framingham Cohort study participants can be viewed at the top-level study ... CT (data available in 3500 people), b) aortic plaque burden by MRI (n 2000), c) carotid intimal-medial thickness by ultrasound (n 3500), d) clinical...'. At the bottom of the summary, there are links for 'FileSelector', 'PubMed', 'PMC', 'MeSH', and 'BioProject'.

While the database format of the DDBJ is identical to that of the NCBI, that of the ENA differs somewhat

11

What are the Primary Databases?

- **4- EMBL database:**
- <https://www.ebi.ac.uk>
- The most obvious difference is the use of two-letter codes instead of full keywords.
- Furthermore, there are small changes in the organization of the individual data fields.
- For example, the date of the last modification is not listed in the field ID (corresponding to the LOCUS field in GenBank) but appears in the field DT (database field).

12

What is the ENA web for?

- **5- ENA Online Retrieval** <https://www.ebi.ac.uk>

The ENA offers several search forms.

- First is a simple search, which allows for text searches as well as for sequence retrieval .
- For text search, it is possible to search for accession numbers and for simple free text.
- ENA also allows for sequence searches using sequence comparisons.

13

What are the Primary Databases?

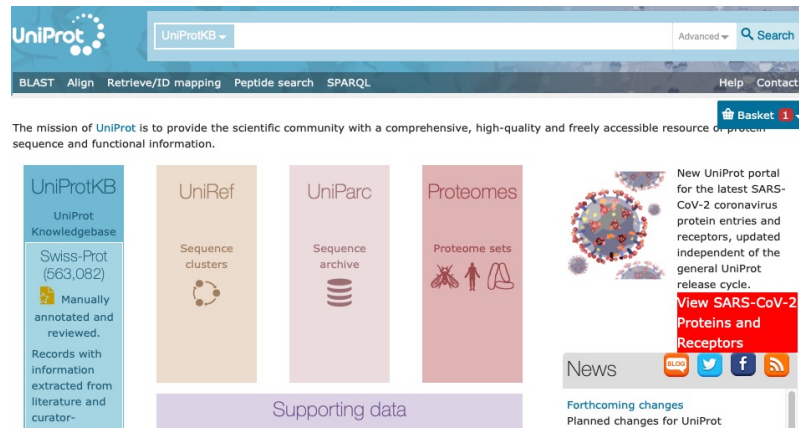
- **6- UniProt Protein Sequence Databases** <https://www.uniprot.org> (very important)

1. The information available for proteins continues to grow rapidly.
2. UniProt consists of three parts, the UniProt Knowledgebase (UniProtKB), the UniProt Reference Clusters Database (UniRef), and the UniProt Archive (UniPArc), a collection of protein sequences and their history.
3. Besides sequence information, expression profiles can be examined, secondary structures predicted, and biological/biochemical function(s) analysed.
4. The result is the Universal Protein Resource (UniProt) [uniprot], which unites the information in the three protein databases Swissprot, TrEMBL, and Protein Information Resource (PIR).
5. Protein sequences and their annotations are stored in the UniProt Knowledgebase (UniProtKB), which is divided into two realms.

14

The entry in the UniProtKB/SwissProt database website.

- At first glance the entry is similar to an ENA entry. Indeed, the two database formats are related.



15

What are other Primary Databases?

- **7- NCBI Protein Database** <https://www.ncbi.nlm.nih.gov> (very important)
- Another well-known protein sequence database is maintained at the NCBI.
- This database, however, is not a single database but a compilation of entries found in other protein sequence databases.
- For example, the NCBI database contains entries from Swissprot, the PIR database [pir], the Protein Data Bank (PDB) database [pdb], protein translations of the GenBank database, and several other sequence databases.
- Its format corresponds to that of GenBank, and queries are carried out analogously to those in GenBank via the Entrez system of NCBI.

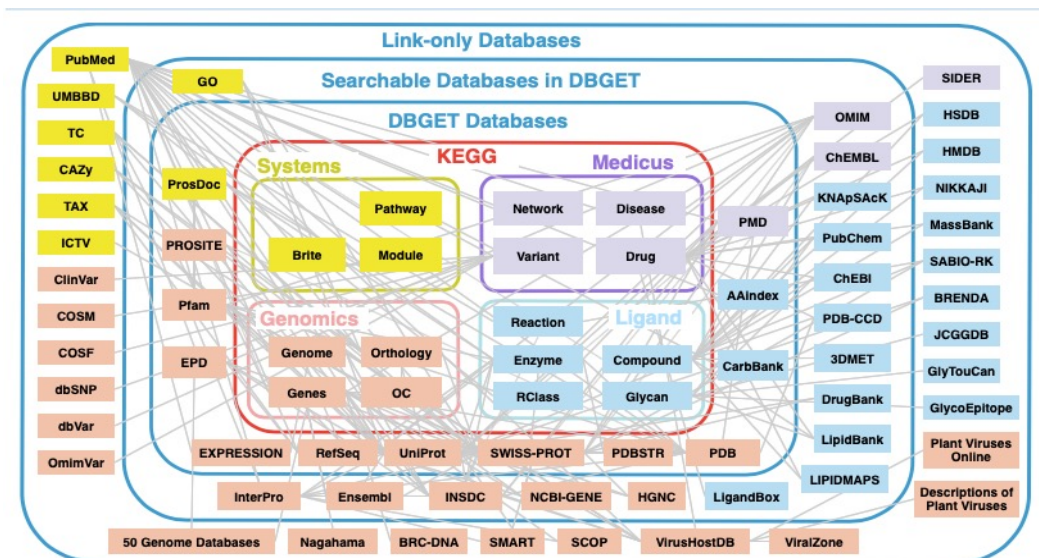
16

KEGG

- <https://www.genome.jp/kegg/>
- KEGG is a database resource for understanding high-level functions and utilities of the biological system.
- The biological system examples: the cell, the organism and the ecosystem, from molecular-level information, especially large-scale molecular datasets generated by genome sequencing and other high-throughput experimental technologies.

17

Correlation Map



18

Exercise?

- Use the website <https://www.uniprot.org> to find the sequence of Tropomyosin TPM1. Download it a Fasta format.
- Use the website <https://www.ncbi.nlm.nih.gov/protein/> to find the information about Tropomyosin TPM1. Download as much as you can and then discuss it.
- Use the website <https://www.ncbi.nlm.nih.gov/genbank/> to find information about Tropomyosin TPM1. Discuss it as you studied!

Bioinformatics I

Lecture 7: Biological Databases

Dr Manaf A Guma

University Of Anbar- college of Applied sciences-Hit

Department of applied chemistry

1

What are the secondary Databases?

- 1- Prosite database <https://prosite.expasy.org>
- Prosite [prosite] is an important secondary biological databas.
- It resides at the SIB [expasy].
- Classification of proteins in Prosite is determined using single conserved motifs, i.e., short sequence regions (10–20 amino acids) that are conserved in related proteins and usually have a key role in the protein's function.
- A typical regular expression in Prosite would have the following form:
[GSTNE]-[GSTQCR]- [FYW]-{ANW}-x(2)-P.

2

What are the secondary Databases?

- 2- PRINTS <http://130.88.97.239/PRINTS/index.php>
- The PRINTS database [prints] uses fingerprints to classify sequences.
- Fingerprints consist of several sequence motifs, represented in the PRINTS database by short, local, un-gapped alignments (talk about it later).
- The PRINTS database takes advantage of the fact that proteins usually contain functional regions that result in several sequence motifs per protein.

3

What are the secondary Databases?

- 3- Pfam database <https://pfam.xfam.org/>
- The Pfam database [pfam] classifies protein families according to profiles.
- A profile is a pattern that evaluates the probability of the appearance of a given amino acid, an insertion, or a deletion at every position in a protein sequence.
- Pfam is based on sequence alignments.

4

What are the secondary Databases?

- **4- Interpro database** <https://www.ebi.ac.uk/interpro/>
- The Integrated Resource of Protein Families, Domains and Sites (Interpro) [interpro] integrates important secondary databases into a comprehensive signature database.
- Interpro merges the databases Swissprot, TrEMBL, Prosite, Pfam, PRINTS, ProDom, Smart, and TIGRFAMs [tigr] and thereby allows a simple and simultaneous query of these databases.
- The result page combines the output of the individual queries.

5

What are the secondary Databases?

- **5-Genotype-Phenotype Databases** <https://www.omim.org> and <https://www.ncbi.nlm.nih.gov/gap/phegeni>
- A number of **genotype-phenotype databases** have been established that record relationships between genes and the biological properties of organisms.
- The Online Mendelian Inheritance in Man (OMIM) <https://www.omim.org> database of the NCBI [omim] is perhaps the best-known genotype- phenotype database.
- It searches for genes related to diseases and drug discovery.
- The **Online Mendelian Inheritance in Animals (OMIA)** database [omia] at the NCBI also contains genotype-phenotype relationships of various animals, except mice and humans.

6

What are the secondary Databases?

- **6- PhenomicDB**

<https://academic.oup.com/bioinformatics/article/21/3/418/237882>

- The PhenomicDB database is a multi-organism genotype-phenotype database containing data from humans and other important organisms such as the mouse, zebra fish (*Danio rerio*), fruit fly (*D. melanogaster*), nematode (*C. elegans*), baker's yeast (*S. cerevisiae*), and cress plant (*Arabidopsis thaliana*).
- A complete listing of all underlying data sources can be found on the home page [phenomicdb]

7

Molecular Structure Databases

- **Secondly: bank databases for proteins structures.**

- **1- Protein Data Bank** <https://www.rcsb.org>
- **The PDB is a database of experimentally determined crystal structures** of biological macromolecules and is coordinated by a consortium located in the USA, Europe, and Japan [wwpdb] (Berman et al. 2000).
- The PDB was founded at the Brookhaven National Laboratory in 1971, reflected in the frequent use of the name Brookhaven Protein Data Bank.
- **It includes DNA and RNA structures and protein–nucleic acid complexes. (that were solved by X-ray, NMR and Cryo EM techniques)**

8

What can you see in (RCSB) web? PDB

RCSB PDB Deposit Search Visualize Analyze Download Learn More MyPDB

SCIENTIFIC NAME OF SOURCE ORGANISM Clear

- Homo sapiens (37987)
- Mus musculus (5205)
- Escherichia coli (4206)
- synthetic construct (3767)
- Escherichia coli K-12 (2531)
- Saccharomyces cerevisiae (2119)
- Rattus norvegicus (2083)
- Bos taurus (2062)
- Gallus gallus (1590)
- Saccharomyces cerevisiae S288C (1584)
- More...

TAXONOMY Clear

- Eukaryota (65789)
- Bacteria (45626)
- Riboviria (7076)
- artificial sequences (3789)
- Archaea (3385)
- Duplodnaviria (1442)
- Varidnaviria (424)
- Monodnaviria (298)
- unidentified (155)
- metagenomes (47)
- More...

EXPERIMENTAL METHOD Clear

- X-RAY DIFFRACTION (110766)
- SOLUTION NMR (8203)
- ELECTRON MICROSCOPY (4290)
- ELECTRON CRYSTALLOGRAPHY (110)
- NEUTRON DIFFRACTION (96)
- SOLID-STATE NMR (84)
- SOLUTION SCATTERING (41)
- FIBER DIFFRACTION (27)
- SOLUTION CRYSTALLOGRAPHY (110)

Displaying 1 to 25 of 124535 Structures Page 1 of 4982 Display 25 per page

1BLQ
STRUCTURE AND INTERACTION SITE OF THE REGULATORY DOMAIN OF TROPONIN-C WHEN COMPLEXED WITH THE 96-148 REGION OF TROPONIN-I, NMR, 29 STRUCTURES
 Mckay, R.T., Pearlstone, J.R., Corson, D.C., Gagne, S.M., Smillie, L.B., Sykes, B.D.
 (1998) *Biochemistry* **37**: 12419-12430
 Released 1999-01-13
 Method SOLUTION NMR
 Organisms Gallus gallus
 Macromolecule N-TROPONIN C (protein)

1YVO
Crystal structure of skeletal muscle troponin in the Ca²⁺-free state
 Vinogradova, M.V., Stone, D.B., Malanina, G.G., Karatzaferi, C., Cooke, R., Mendelson, R.A., Fletterick, R.J.
 (2005) *Proc Natl Acad Sci U S A* **102**: 5038-5043
 Released 2005-04-12
 Method X-RAY DIFFRACTION 7 Å
 Organisms Gallus gallus
 Macromolecule Troponin C, skeletal muscle (protein)
 Troponin I, fast skeletal muscle (protein)
 Troponin T, fast skeletal muscle isoforms (protein)

9

PDB web search

RCSB PDB Deposit Search Visualize Analyze Learn More

RCSB PDB 167780 Biological Macromolecular Structures Enabling Breakthroughs in Research and Education

WORLDWIDE PROTEIN DATA BANK

EMDataResource Verified Data Resource for 2020

WORLDWIDE PROTEIN DATA BANK Foundation

Welcome Deposit Search Visualize Analyze Learn

A Structural View of Biology

This resource is powered by the Protein Data Bank archive-information about the 3D shapes of proteins, nucleic acids, and complex assemblies that helps students and researchers understand all aspects of biomedicine and agriculture, from protein synthesis to health and disease.

August Molecule of the Month

10

Molecular Structure Databases

- **2-SCOP database** <http://scop.mrc-lmb.cam.ac.uk>
- Proteins that perform **a similar biological function and are evolutionary related** must have a similar structural organization, at least in the region of their active centres.
- It should, therefore, be possible **to predict the function of an unknown protein by comparison** of its structural organization with that of known proteins.
- Two databases, SCOP and CATH, provide such predictions. SCOP (Structural Classification Of Proteins) [scop] classifies proteins of a known structure in a hierarchical manner.

11

Molecular Structure Databases

- **3- CATH database** <https://www.cathdb.info>
- The CATH database [cath] (Greene et al. 2007) classifies protein structures hierarchically into four categories: Class (C), Architecture (A), Topology (T), and Homologous Superfamily (H).
- **Four classes of proteins are distinguished:** proteins composed mainly of helices (*mainly alpha*), sheets (*mainly beta*), both helices and sheets (*alpha-beta*), and, finally, proteins with very few secondary structural elements.

12

Molecular Structure Databases

- 4-PubChem database <https://pubchem.ncbi.nlm.nih.gov>
- The PubChem database at the NCBI [pubchem] stores small chemical molecules and information about their biological activities.
- PubChem Compound contains about 91 million molecules (July 2016) together with their two-dimensional (2D) molecular structures.
- For example, with a known enzyme inhibitor it is possible to find other similar potential inhibitors.

13

Molecular Structure Databases

PubChem

About Blog Submit Contact

Explore Chemistry

Quickly find chemical information

Browse COVID-19 data available

SEARCH FOR

benzene

×



Treating this as a text search.

COMPOUND BEST MATCH



[Benzene; Benzol; Cyclohexatriene; 71-43-2; Benzole; Pyrobenzole; Benzine; Benzen;](#)

...

Compound CID: 241

MF: C_6H_6 MW: 78.11g/mol

InChIKey: UHOVQNZJYSORNB-UHFFFAOYSA-N

IUPAC Name: benzene

Create Date: 2004-09-16

[Summary](#)

[Similar Structures Search](#)

[Related Records](#)

[PubMed \(MeSH Keyword\)](#)

14

For diseases' research

- <https://www.omim.org>
- It is an Online Catalogue of Human Genes and Genetic Disorders.
- It searches for diseases based on genetics.
- very useful website.

15

How to find reported mutations?

- <http://www.hgmd.cf.ac.uk/ac/index.php>
- **HGMD** only reports published variants and only the first published account at that. If the disease phenotype reported in that publication is atypical, you'll never know. Gene specific LSDBs are likely to contain much more data, including unpublished variants. However, some LSDBs are better curated than others.
- Also, <https://www.genecards.org/cgi-bin/carddisp.pl?gene=ERBB4&keywords=tropomyosin>
- if you want to find all the germline mutation, search gene in NCBI and it will report all SNP in this gene from **dbSNP**: <http://www.ncbi.nlm.nih.gov/gene/>
- If you want to find somatic mutation especially somatic mutation in cancer, cosmic is the best choice: http://www.sanger.ac.uk/perl/genetics/CGP/cosmic?action=gene_list

16

Exercises

- **Exercise 2.1**

Search for a protein (enzyme) from the organism *Bacillus subtilis* that hydrolyzes terminal nonreducing arabinofuranoside residues. To do this, use the keyword search under Entrez (7 <http://www.ncbi.nlm.nih.gov/entrez/>). Note: hydrolysis, arabino- furanoside, hydrolases, glycosyl, terminal, nonreducing. The Advanced search link leads you to an editor and your query history, so you can modify previous searches of the same session. Possible combinations are AND, OR, NOT.

- ? **Exercise 2.2**

Locate the gene for the enzyme IABF-BACSU from 7 Exercise 3.1 in the nucleotide database. If you are unable to find it, try to develop new search strategies from the results and hints provided.

- ? **Exercise 2.3**

Search for the protein with the following accession number in Entrez: P94552.

- ? **Exercise 2.4**

Search for the same accession number on the EBI home page (7 <http://www.ebi.ac.uk/>).

Bioinformatics I

Lecture 8: Pairwise alignment of DNA & protein sequencing (Manually & Matrices)

Dr Manaf A Guma
University Of Anbar- college of Applied sciences-Hit
Department of applied chemistry

1

Before we start, lets remember!

- How many types of changes which can present between two sequence alignment?

1. A mutation that replaces one character with other.
2. An insertion that adds one or more positions.
3. A deletion that deletes one or more positions.

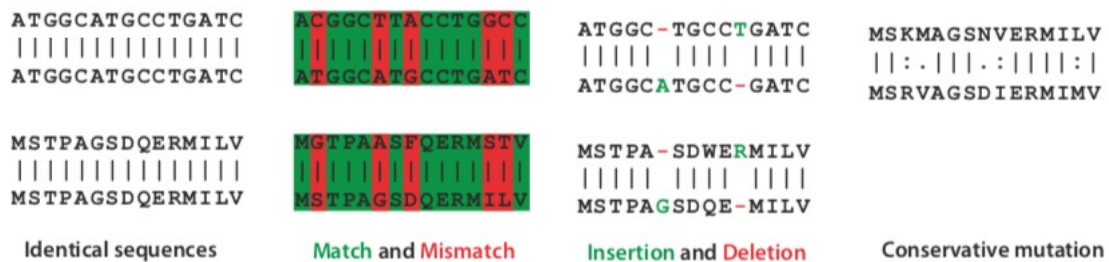
- Why do we need to do alignment? for similarity

Experimentally, to check the product of the PCR which is called subject (hint) with original sequence which is called the query (WT for e.g)

2

What is the Pair-wise sequence alignment?

1. Display one sequence above another with spaces inserted in both, to **reveal similarity**.
2. It does show how programs work.



Sequence alignments of nucleotide and amino acid sequences

3

How many types of alignments? There are 2 types of alignment

S = CTGTCGCTGCACG
T = TGCCGTG

Global alignment
dealing with big
sequence of DNA

Local alignment:
dealing with small
segment of DNA

CTGTCG-CTGCACG
-TGC-CG-TG-----

CTGTCGCTGCACG--
-----TGC-CGTG

4

4

How do you calculate Global alignment: Scoring method ?

CTGTCG-CTGCACG
 -TGC-CG-TG----

Let's assume that
 Reward for matches: α
 Mismatch penalty: β
 Space penalty: γ

$$\text{score}(A) = \alpha w - \beta x - \gamma y$$

$w = \# \text{matches}$ $x = \# \text{mismatches}$ $y = \# \text{spaces}$

5

5

How do you calculate Global alignment: Scoring method ? Global alignment: Calculate Scoring?

Reward for matches: 10
 Mismatch penalty: 2
 Space penalty: 5

C	T	G	T	C	G	-	C	T	G	C
-	T	G	C	-	C	G	-	T	G	-
-5	10	10	-2	-5	-2	-5	-5	10	10	-5

Total = 11

6

6

Optimum Alignment

- What does the score measure?
- It is a measure of its quality.
- **How does it solve the problem of the alignment?**
- Given a pair of sequences X and Y , find an alignment (global or local) with maximum score.
- What does maximum score of an alignment represent?
- The **similarity** between X and Y , denoted $sim(X,Y)$, is the maximum score of an alignment of X and Y

7

7

Who designed the Alignment algorithms?

- Global: **Needleman-Wunsch**
- Local: **Smith-Waterman**
- NW and SW use ***dynamic programming***.
- ***How do the*** algorithms express the variations of alignment?
 1. Gap penalty functions
 2. Scoring matrices

8

8

**Determine the Global alignment in two sequences using
*dynamic programming method?***

Two seq. ATCG and TCG?

Scoring ?

If you assume that :

- Match +1
- Mis-match -1
- Gap -2
- This method determined the alignment between 2 seq by drawing a table.
- Sequence 1 is then written in the y axis and Sequence2 is written in the x-axis.
- Then we need to follow the following rules to find the score.

9

**Rules to determine the alignment using a
*dynamic programming method***

1. Put the gap in the first.
2. The higher is score the best because it shows the maximum alignment.
3. The first box always zero.
4. Add the initial column and row by adding gap value -2.
5. To calculate the score, it will be the highest value in the diagonal Once we finish, we need to do trace back to find the score.
6. Trace back: we go back to find the highest value in the diagonal only.(the small score is, the best is)
7. Then to express that:
8. We start from the back to starting point which is the zero. It has to be a straight line to the zero.
9. We need to do the alignment to find why the score is +1.

10

10

Solve out this alignment? find the alignment between: **ATCG** and **-TCG**

G	-8			
C	-6			
T	-4			
A	-2	-1	-3	-5
-	0	-2	-4	-6
	-	T	C	G

1. box beside +gap (-2)
2. box bottom + gap (-2)
3. Diagonal box (match /mismatch).

Match + 1
Mis-match -1
Gap -2

The direction means the number that can come from this direction or from the other direction.

11

Go on for this alignment?

<https://www.youtube.com/watch?v=vqxc2EfPWdk&t=1113s>

G	-8	-8-2=-10	-5-2=-7	-2-2=-4
		-6-1=-7	-3-1=-4	0+1=+1
		-3-2=-5	0-2=-2	-2-2=-4
C	-6	-6-2=-8	-3-2=-5	0-2=-2
		-4-1=-5	-1+1=0	-2-1=-3
		-1-2=-3	-2-2=-4	-4-2=-6
T	-4	-4-2=-6	-1-2=-3	-2-2=-4
		-2+1=-1	-1-1=-2	-3-1=-4
		-1-2=-3	-3-2=-4	-5-2=-7
A	-2	-2-2=-4	-1-2=-3	-3-2=-5
		0-1=-1	-2-1=-3	-4-1=-5
		-2-2=-4	-4-2=-6	-2-6=-8
0	0	-2	-4	-6
	0	T	C	G

Match + 1
Mis-match -1
Gap -2

We take the highest value among the three values...

1. box beside +gap (-2)
2. box bottom + gap (-2)
3. Diagonal box (match /mismatch).

→ Score= 1

12

How to find the score ?

<https://www.youtube.com/watch?v=vqxc2EfPWdk&t=1113s>

G	-8	-5	-2	1
C	-6	-3	0	-2
T	-4	-1	-2	-4
A	-2	-1	-3	-5
0	0	-2	-4	-6
0	T	C	G	

Match + 1
Mis-match -1
Gap -2

The score value is the highest value in the diagonal !...

1. box beside +gap (-2)
2. box bottom + gap (-2)
3. Diagonal box (match /mismatch).

Score= 1

13

Find the Local Alignment using *dynamic programming method*?

- To find the local alignment between two seq:
- We follow the same rules as same of the global alignment.
- Example:
- Find the local alignment of two seq:
- ATCG and TCG.

14

14

Find the score using local alignment matrices.
 For local alignment only: place any value (-) with zero, whenever!

C	0				
C	0				
T	0				
-	0	0	0	0	0
	-	A	T	C	G

Match + 1
 Mis-match 0
 Gap 0

1. box beside +gap (0)
2. box bottom + gap (0)
3. Diagonal box (match /mismatch).

15

Go on for this alignment?

C	0	0	0	1	0
C	0	0	0	1	0
T	0	0	1	0	0
-	0	0	0	0	0
	-	A	T	C	G

Score= 1

Match + 1
 Mis-match 0
 Gap 0

- box beside +gap (-2)
- box bottom + gap (-2)
- Diagonal box (match /mismatch).

16

Local alignment: Example

S = g g t c t g a g
 T = a a a c g a

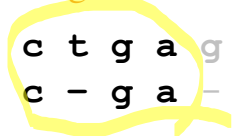
If you assume that

Match: +2

Mismatch and space penalty: -1

Best local alignment:

g g t c t g a g
 a a a c - g a -



Score = 5

17

Bioinformatics I

Lecture 9: Pairwise alignment of DNA & protein (using matrices)

Dr Manaf A Guma
University Of Anbar- college of Applied sciences-Hit
Department of applied chemistry

1

What are the purposes of pairwise alignment comparison?

- The purposes of pairwise **alignment** comparison are (using the matrices or the manual methods):
 1. To find the **score of the identity** between two sequences.
 2. To find whether two (or more) genes or proteins **are evolutionarily related to each other.**
 3. To find structurally or functionally **similar regions** within proteins

2

Common types of matrices are used for Sequence Comparison

- There are various methods available for pairwise alignment. the common methods are:
 1. Dot matrix analysis.
 2. Dynamic Programming.
 3. Formula (by hand) approaches e.g (FASTA and BLAST).

3

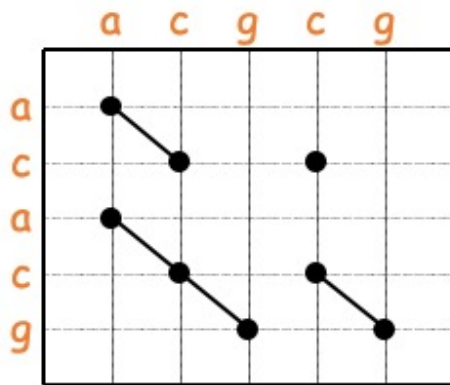
3

1- Pairwise alignment using (Dot plot)Matrices

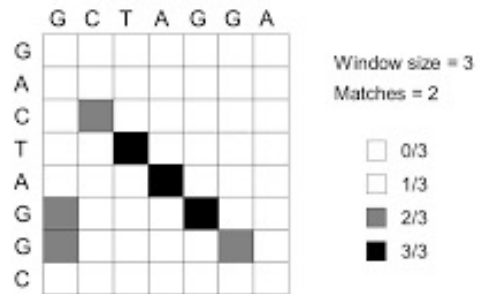
- This is one of the most popular graphical methods of aligning two sequences.
- The sequences are placed on the X- and Y-axes of the matrix and a dot is placed wherever a match is found between the two sequences.
- Diagonal runs of dots are joined to form the alignment.
- However, dot matrices give only a graphical representation and do not reveal the similarity score.

4

How do "Dot matrices" look like?



Filtering the dot plot

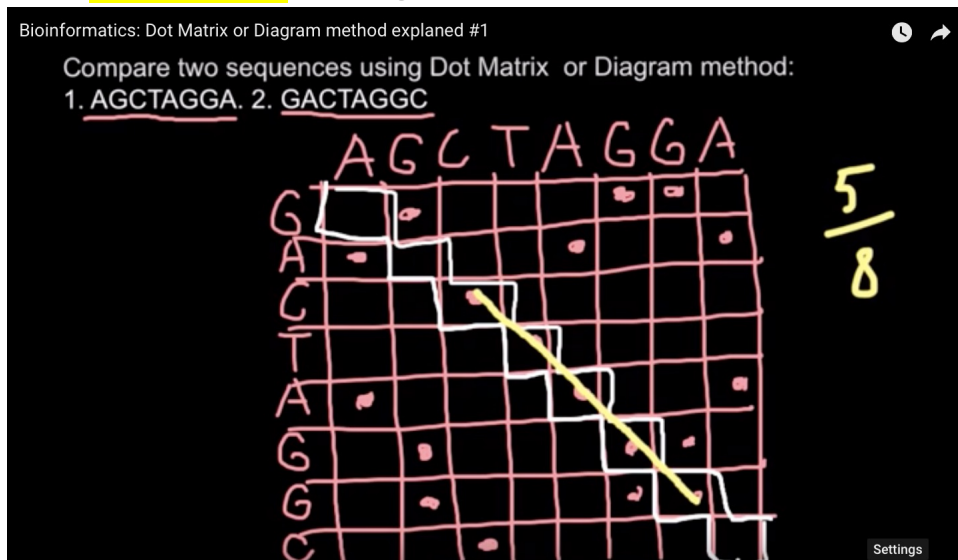


©2009 2007

33

5

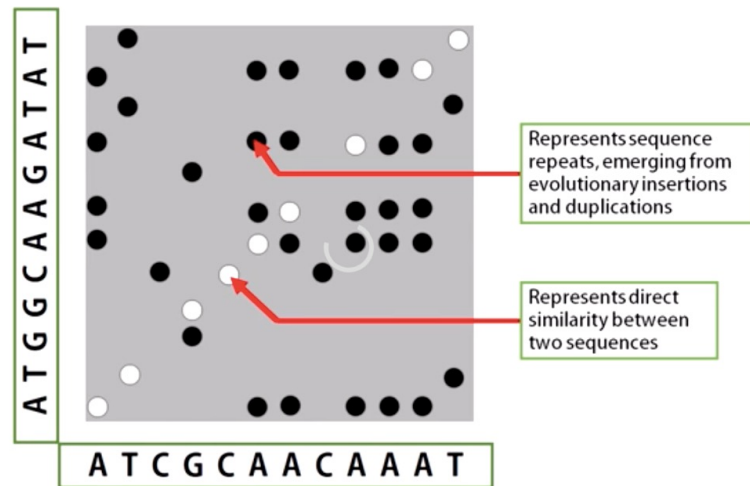
Dot matrix comparison by finding how many coincidences for alignment



6

6

Another example



7

Give an interpretation for the matrices.

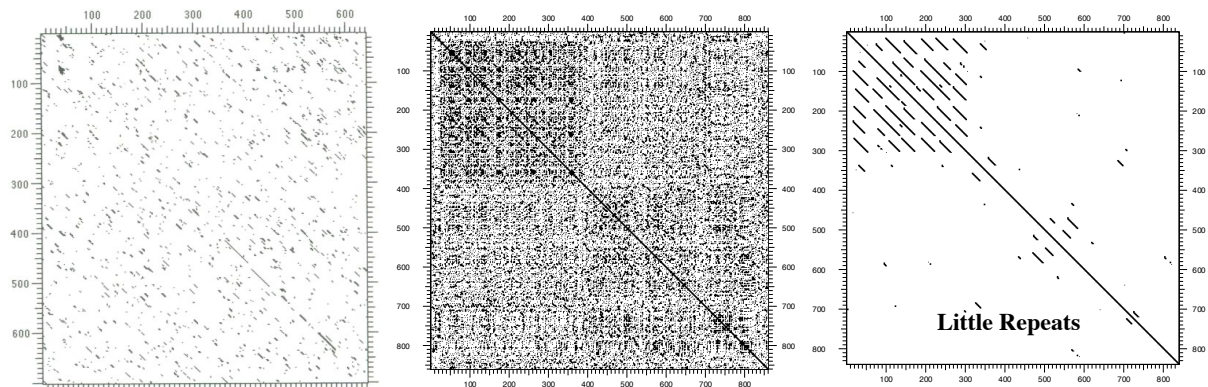
1. Regions of similarity appear as diagonal runs of dots.
2. Reverse diagonals (perpendicular to diagonal) indicate inversions
3. Reverse diagonals crossing diagonals (Xs) indicate palindromes.
4. Link can separate diagonals to form **alignment** with *gaps*; each amino acid. or base can only be used once (Can't double back)

8

8

What are the artifact of Dot matrices? By Filtering?

- Dot matrices for long sequences can be noisy due to insignificant matches.



9

What are the uses of dot matrices for?

1. Aligning two proteins or two nucleic acid sequences.
2. Finding amino acid repeats within a protein by comparing a protein sequence to itself.
3. Repeats appear as a set of diagonal runs stacked vertically and/or horizontally.

10

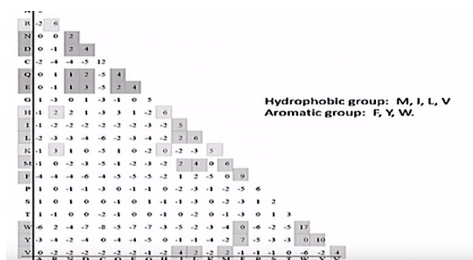
2- PAM matrices

- **Point accepted mutation matrix** known as a PAM. مصفوفة الطفرة المقبولة بنقطة
- It is also called **Percent Accepted Mutation** مصفوفة النسبة المئوية للطفرة المقبولة
- Dayhoff and colleagues defined the PAM1 matrix as that which produces 1 accepted point mutation per 100 amino acid residues.
- **PAM matrix** is designed to compare two sequences which are a specific number of PAM units apart.
- Only mutations are allowed.
- <https://www.youtube.com/watch?v=UCtP5-KtB94>,
<https://www.youtube.com/watch?v=F8WdDfpQqCM>

11

PAM matrices are calculated by BLAST websites

- PAM matrices are also used as a scoring matrix when comparing DNA sequences or protein sequences to judge the quality of the alignment.
- This form of scoring system is utilized by a wide range of alignment software including BLAST.
- PAM250 corresponds to 20% amino acid identity, represents 250 mutations per 100 residues.
- E.g:



12

What BLOSUM is based on?

- It is based on comparisons of **blocks of sequences** derived from the Blocks database.
- **The block length is 60 amino acids. (without any gaps or frequencies).**
- Blocks database refers to the alignment not to the individual sequence.
- BLOSUM matrices tell the % of matching.
- It can be 100% even if there is a substitution.
- It tells how much the sequence is conserved!

15

What does block mean in BLOSUM method?

- It means creating a block of 2 seq or multiple sequences which refers to a best alignment in order to recognize the mutations, gaps and penalties in each row.
- The block presents the same length of sequences ' about 60 letter of amino acids or nucleotide'.

KKAS	KPKKAASKAP	T	KKPKATPVKKAKKKL	AATPKK	AKKPK	TVK
KKAA	KPKKAASKAP	S	KKPKATPVKKAKKK	PAATPKK	AKKPK	VVK
KKAA	KPKKAASKAP	S	KKPKATPVKKAKKK	PAATPKK	AKKPK	IVK
KKAA	KPKKAASKAP	S	KKPKATPVKKAKKK	PAATPKK	T	KKPKTVK
KKAS	KPKKAASKAP	T	KKPKATPVKKAKKKL	AATPKK	AKKPK	TVK

Matches = 39 columns × 6 rows = 234

Percentage of identity (234/264) = 89%

16

BLOSUM in BLAST

Range 1: 127 to 501 [GenPept](#) ▼ Next Match ▲ Previous Match

Score	Expect	Method	Identities	Positives	Gaps	Frame
431 bits(1108)	3e-147()	Compositional matrix adjust.	203/375(54%)	278/375(74%)	8/375(2%)	
Query 32	EKKRRDREERQNIWLWRQPLITLQYFSLETLVVLKEWTSKLWHRQSI VVSFLLLLAALVA			91		
	+++ R+R ER +VLWR+PL T +Y LE +L+ W+++L ++ ++ + ++L					
Sbjct 127	KQREERLERGQLVLWRRPLQPTTKYCGLELFTLLRTWSTRLLQQRLLLATLSEVLSIVFSV			186		
Query 92	TYVEGAHQYVQRIEKQFLLYAYWIGLILSSVGLGTGLHTFLLYLGPHIASVTLAAYE			151		
	Y ++G HQ ++ + + + YW+GLG+LSSVGLGTGLHTFLLYLGPHIASVTLAAYE					
Sbjct 187	IYKIDGPHQLAIEFVRRNTWFFVYWLGLGVLSSVGLGTGLHTFLLYLGPHIASVTLAAYE			246		
Query 152	CNSVNFPEPPYPDQIICPEEEGAEGAISLWSIISKVRIEACMWGIGTAIGELPPYFMARA			211		
	CNS+ FP+PPYPD IICPEE + ++WSI+SKVR+EA +WG GTA+GELPPYFMA+A					
Sbjct 247	CNSLRFPPYPDDIICPEEPYDKHVPNIWSIMSKVRLEAFLWGAGTALGELPPYFMAKA			306		

The positives (+) in the alignment indicate good high scoring mismatches. Matrix scores >0.

Mis-matching !!!!!

Bioinformatics I

Lecture 10: Pairwise Sequencing For DNA & Protein Using Dynamic Programming

Dr Manaf A Guma

University Of Anbar- college of Applied sciences-Hit

Department of applied chemistry

1

BLAST database for sequencing?

- **BLAST** Basic Local Alignment Search Tool
- A frequently used application of pairwise alignments is the search **for similar protein or nucleotide sequences in sequence databases using smart computers.**
- With older dynamic alignment algorithms such as those designed by Smith and Watermann (1981) or Needleman and Wunsch (1970), this is too slow to perform even on current computers.

2

2

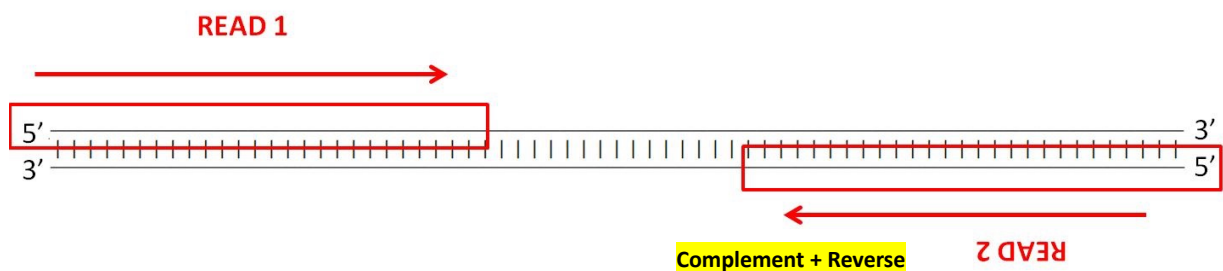
BLAST

- **What does BLAST look for?** It looks for the score of the alignment.
- **What does it indicate?** The BLAST Score indicates the **quality of the best alignment** between the **query** sequence and the **found sequence (hit)**.
- **What does high/ low score represent?** The higher the score, the better the alignment.
- **When the score of the alignment is reduced?** Scores are reduced by mismatches and gaps in the best alignment

3

How should we order the sample to check?

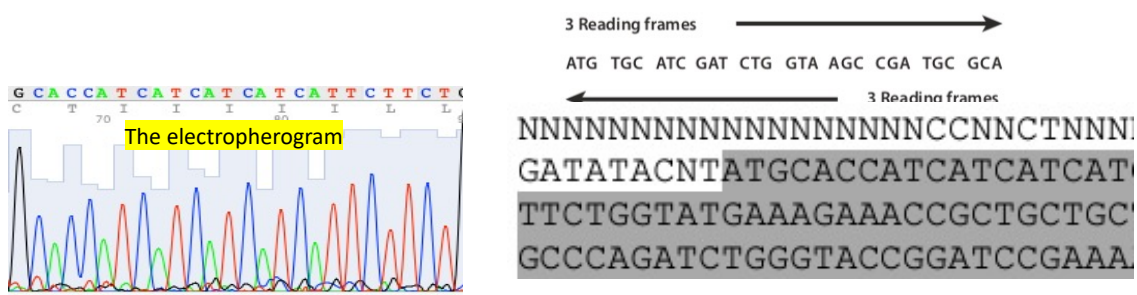
- Follow the rules of the companies, they ask for **specific concentrations (conc.)**.
- Usually, 0.375 ng/ 10 μ L for both (Forward **F** and Revers **R**)
- = 75ng/ 20 μ L of DNA or RNA
- + 10ng/ 1 μ L of each **F** or **R**.



4

Receiving the sample & starting check!

- Commonly, three files
- **name. seq.** (Forward **F** or Revers **R**) =====Microsoft WORD
- **name. ab1** (Forward **F** or Revers **R**)===== needs a Software
- We need both (Forward **F** & Revers **R**), if ... The seq is toooo long ! more than 400 bp.



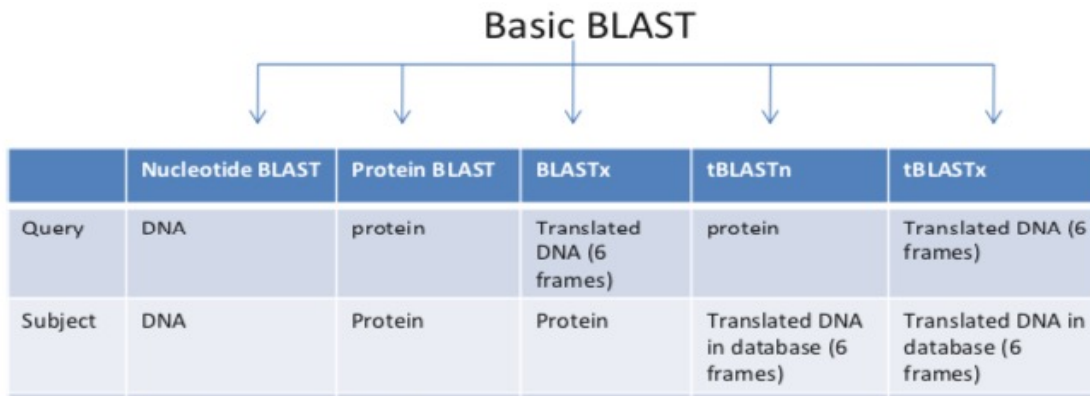
5

BLAST - general

- For a BLAST similarity search we need:
 1. input sequence (**query**)
 2. algorithm (implemented in the BLAST software)
 3. database (of protein or nucleotide sequences)
 4. Finally, we need to understand the output

6

There are different types of BLAST



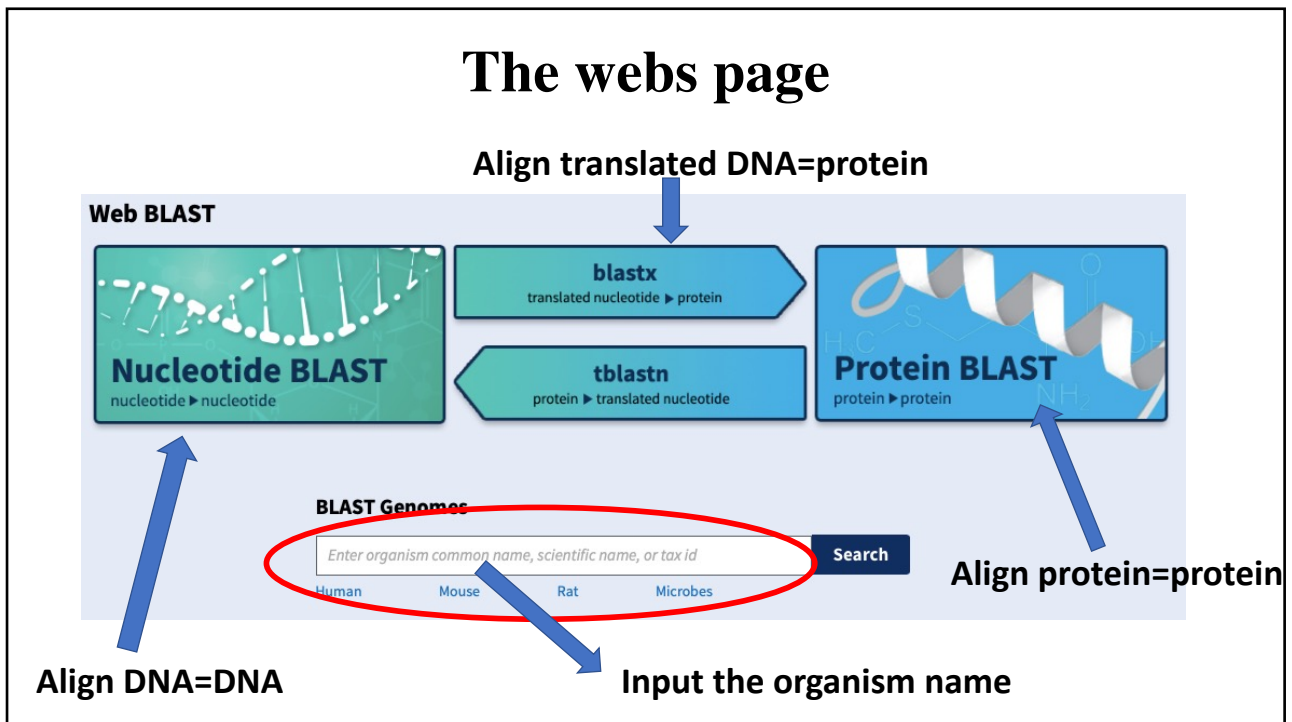
7

Translates DNA sequence into all six possible reading

- For example: if you translate the following seq
- AACCTGTATTTTCAGGGCGCCATG
- You will find:
- 6 possible proteins depends on the frameshift.



8



9

Use BLAST database (alignment) to find the similarity?

- **The query:**

```
GAMDAIKKKMQLKLDKENALDRAEQAEADKKAEDRSKQLEDELVSLQKKLKGTEDELDKYSEALKDAQEKLELAEKKATD
AEADVASLNRRIQLVEEELDRAQERLATALQKLEEAKAADESERGMKVIESRAQKDEEKMEIQEIQLKEAKHIAEDADRKY
EEVARKLVIIESDLERAERAELSEGKCAELEELKTVTNNLKSLEAQAEKYSQKEDRYEEEIKVLSDKLKEAETRAEFAER
SVTKLEKSIDDLEDELYAQKLYKAI SEELDHALNDMTSI
```

- **The subject:**

- ASMDAIKKKMQLKLDKENALDRAEQAEADKKAEDRSKQLEDELVSLQKKLKGTEDELDKYSEALKDAQEKLELAEKKATD


```
AEADVASLNRRIQLVEEELDRAQERLATALQKLEEAKAADESERGMKVIESRAQKDEEKMEIQEIQLKEAKHIAEDADRKY
EEVARKLVIIESDLERAERAELSEGKCAELEELKTVTNNLKSLEAQAEKYSQKEDRYEEEIKVLSDKLKEAETRAEFAER
SVTKLEKSIDDLEDELYAQKLYKAI SEELDHALNDMTSI
```

10

Translation the sequence to protein (make it easy)

1. Check the alignment with your original copy using:

<https://blast.ncbi.nlm.nih.gov/Blast.cgi>

2. If you do not know the original copy of your protein, just type it

here: <https://www.uniprot.org>

11

Example: Protein Sequence check up

Query	1	MDAIKKKMQMLKLDKENALDRAEQAEADKKAEDRSKQLEDELVSLQKKLKGTEELD	60
Query_182929	47E.....	106
Query	61	KTSEALKDAQEKLELAEEKATDAEADVASLNRRQLVEEELDRAQERLATALQKLEEAEK	120
Query_182929	107	166
Query	121	AADESERGMKVIESRAQKDEEKMEIQEIQLKEAKHIAEDADRKYEEVARKLVIIESDLER	180
Query_182929	167	226

Example: DNA Sequence check up

Sequence ID: **Query_35585** Length: **861** Number of Matches: **1**

Range 1: 1 to 861 [Graphics](#) [Next Match](#) [Previous](#)

Score	Expect	Identities	Gaps	Strand
1591 bits(861)	0.0	861/861(100%)	0/861(0%)	Plus/Plus
Query 1	GCCAGCATGGACGCGATCAAGAAGAAGATGCAAATGCTGAAACTGGACAAAGAAAATGCG	60		
Sbjct 1	GCCAGCATGGACGCGATCAAGAAGAAGATGCAAATGCTGAAACTGGACAAAGAAAATGCG	60		
Query 61	CTGGACCGTGC CGGAACAGGCGGAGCGGACAAAGAAAGCGGCGGAGGATCGTAGCAAGCAG	120		
Sbjct 61	CTGGACCGTGC CGGAACAGGCGGAGCGGACAAAGAAAGCGGCGGAGGATCGTAGCAAGCAG	120		
Query 121	CTGGAAGACGAGCTGGTGAGCCTGCAAAAAGAAACTGAAGGGCACCGAAGACGAGCTGGAT	180		
Sbjct 121	CTGGAAGACGAGCTGGTGAGCCTGCAAAAAGAAACTGAAGGGCACCGAAGACGAGCTGGAT	180		

Different styles of alignment



12

Tutorial

- Check the pairwise similarity between two DNA sequences (given by you tutor) .
- What did you find?
- Discuss!

Bioinformatics I

Lecture 11: Multiple Sequence Alignment MSA using Dynamic Programming

Dr Manaf A Guma
University Of Anbar- college of Applied sciences-Hit
Department of applied chemistry

1

What is the MSA?

- It is an alignment of more than 2 sequences.
- Why do we do MSA? Or what is the purposes of MSA?
 1. To **highlight conservation and variation**. How? By identifying the regions of similarity among different species.
 2. To find the relation among different species.
 3. To find the **profile** of sequence from the database.
 4. To know how to draw **phylogenetic trees**.

2

Why do we use dynamic programming in MSA?

- Because there is a huge database which makes the comparison very difficult if we run MSA by hand.
- Which software and websites are commonly used to do MSA?

1. BLAST.

FASTA format) do you remember it !

2. FASTA.

```
>AT1G09780 | 1 | training
GTGGAGTAGAAGAATTGAGAGCCTTATCAG
TTTTTGAAGAGAGGGCTGAAACTCTCTAGT
TATCTTTTGTTGCTTTTCTAATAATAAGAG
TTTACACACAG
```

Part 1

Part 2

Part 3

3. ClustalW.

3

How do you use BLAST to run MSA? (Tutorial)

1. We have to have a specific sequence for (protein or DNA for a specific species) that we need to find the similarity with it.
2. If we do not have it, then we go to <https://www.uniprot.org> and then find the Protein seq.
3. Copy the seq (in a FASTA format) do you remember it !
4. Open <https://blast.ncbi.nlm.nih.gov/Blast.cgi> and find blast protein-protein.
5. Paste the seq in the box labeled with **Enter Query Sequence:**
6. Click on BLAST to find the similarities.
7. The result will show the comparison (the identity and the scoring of the similarity) of the protein to various proteins in the database.
8. It also show you the matrices used to generate the comparison.

4

Can we get MSA form BLAST? What can we get?

- We can get only pairwise alignment using BLAST. (what is pairwise-do you remember?)
- But we can not get all of the sequences aligned together in the same screen using BLAST.
- We can get the profile of each sequence (the type of the species, the gene name and gene number etc.)

5

An example to see how BLAST works

Description	Max Score	Total Score	Query Cover	E value	Per. Ident	Accession
<input checked="" type="checkbox"/> PREDICTED: tropomyosin alpha-1 chain isoform X1 [Callithrix jacchus]	531	531	100%	0.0	99.65%	XP_002753250.2
<input checked="" type="checkbox"/> PREDICTED: tropomyosin alpha-1 chain isoform X5 [Chlorocebus sabaeus]	531	531	100%	0.0	99.65%	XP_008014544.1
<input checked="" type="checkbox"/> PREDICTED: tropomyosin alpha-1 chain isoform X1 [Macaca fascicularis]	531	531	100%	0.0	99.65%	XP_005559773.1
<input checked="" type="checkbox"/> tropomyosin alpha-1 chain isoform Tpm1.1st [Homo sapiens]	528	528	100%	0.0	100.00%	NP_001018005.1
<input checked="" type="checkbox"/> tropomyosin alpha-1 chain isoform 16 [Homo sapiens]	527	527	100%	0.0	99.65%	NP_001352708.1
<input checked="" type="checkbox"/> tropomyosin alpha-1 chain [Oryctolagus cuniculus]	526	526	100%	0.0	99.65%	NP_001099158.1
<input checked="" type="checkbox"/> tropomyosin alpha-1 chain isoform X2 [Lagenorhynchus obliquidens]	526	526	100%	0.0	99.65%	XP_026979007.1
<input checked="" type="checkbox"/> tropomyosin alpha striated muscle isoform [Homo sapiens]	526	526	100%	0.0	99.65%	AAT88285.1
<input checked="" type="checkbox"/> Chain A_Tropomyosin [Oryctolagus cuniculus]	526	526	100%	0.0	99.30%	2TMA_A
<input checked="" type="checkbox"/> tropomyosin alpha-1 chain isoform X2 [Heterocephalus glaber]	525	525	100%	0.0	99.30%	XP_004855748.1
<input checked="" type="checkbox"/> PREDICTED: tropomyosin alpha-4 chain isoform X6 [Chrysochloris asiatica]	525	525	100%	0.0	99.30%	XP_006831632.1
<input checked="" type="checkbox"/> tropomyosin alpha-1 chain isoform X1 [Balaenoptera acutorostrata scammonii]	525	525	100%	0.0	99.30%	XP_007166029.2
<input checked="" type="checkbox"/> PREDICTED: tropomyosin alpha-1 chain isoform X7 [Sorex araneus]	524	524	100%	0.0	99.30%	XP_004616749.1
<input checked="" type="checkbox"/> tropomyosin alpha-1 chain isoform X4 [Otolemur garnettii]	523	523	100%	0.0	99.30%	XP_003784447.1

6

How do you use FASTA to run MSA?

1. Get the protein/DNA seq from <https://www.uniprot.org>.
2. copy the seq in FSATA format.
3. Open FASTA web page <https://www.ebi.ac.uk/Tools/sss/fasta/>.
4. Paste the seq.,
5. The results will show different choses to get various bioinformatic analysis in a table.
6. You can show the MSA by clicking on **visual output**.
7. You can also download the seq by clicking on Download

7

The tables of FASTA results: an example

Tools > Sequence Similarity Searching > FASTA

Results for job fasta-l20200310-083954-0267-59723302-p2m

[Summary Table](#)
[Tool Output](#)
[Visual Output](#)
[Functional Predictions](#)
[Submission Details](#)

Selection:

Apply to selection:

Annotations:

Alignments:

Entries:

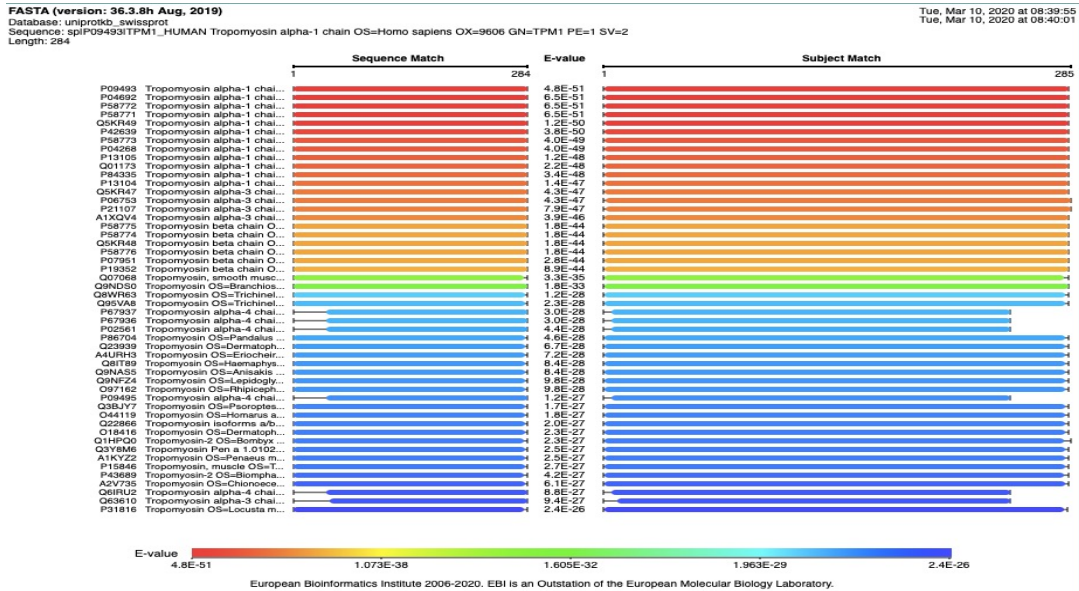
Tools:

You can download all the seq form here

Align.	DB-ID	Source	Length	Score (Bits)	Identities %	Positives %	E()
<input checked="" type="checkbox"/>	SP-P09493	Tropomyosin alpha-1 chain OS=Homo sapiens OX=9606 GN=TPM1 PE=1 SV=2 <i>Cross-references and related information in:</i> <ul style="list-style-type: none"> ▶ Gene expression ▶ Bioactive molecules ▶ Nucleotide sequences ▶ Genomes & metagenomes ▶ Literature ▶ Samples & ontologies ▶ Diseases ▶ Molecular interactions ▶ Protein families ▶ Macromolecular structures ▶ Protein expression data ▶ Reactions & pathways ▶ Protein sequences 	284	202.6	100.0	100.0	4.8E-51
<input checked="" type="checkbox"/>	SP-P04692	Tropomyosin alpha-1 chain OS=Rattus norvegicus OX=10116 GN=Tpm1 PE=1 SV=3 <i>Cross-references and related information in:</i> <ul style="list-style-type: none"> ▶ Gene expression ▶ Bioactive molecules ▶ Nucleotide sequences ▶ Genomes & metagenomes ▶ Literature ▶ Samples & ontologies ▶ Molecular interactions ▶ Protein families ▶ Macromolecular structures ▶ Protein expression data ▶ Reactions & pathways ▶ Protein sequences 	284	202.1	99.6	100.0	6.5E-51
<input checked="" type="checkbox"/>	SP-P58772	Tropomyosin alpha-1 chain OS=Oryctolagus cuniculus OX=9986 GN=TPM1 PE=1 SV=1 <i>Cross-references and related information in:</i> <ul style="list-style-type: none"> ▶ Bioactive molecules ▶ Nucleotide sequences ▶ Literature ▶ Samples & ontologies ▶ Molecular interactions ▶ Protein families ▶ Macromolecular structures ▶ Protein expression data ▶ Protein sequences 	284	202.1	99.6	100.0	6.5E-51
<input checked="" type="checkbox"/>	SP-P58771	Tropomyosin alpha-1 chain OS=Mus musculus OX=10090	284	202.1	99.6	100.0	6.5E-51

8

An example to see how FASTA works



9

What is ClustalW ?

- ClustalW is the “classic” MSA tool using C++ programming made by JD Thompson, DG Higgins, and TJ Gibson.
- The original publication describing ClustalW is one of the 100 most cited publications in ‘web of science’.
- How CLUSTAL W deals with MSA?

CLUSTAL W: deals with multiple sequence alignment through:

1. Sequence weighting.
2. position-specific gap penalties
3. weight matrix choice.

- What is the last version of ClustalW?
- ClustalW It is an old version, the version is Clustal Omega which is much faster and better tools are available.

<http://www.ebi.ac.uk/Tools/msa/>

10

How do you use ClustalW to run MSA? (very common)

1. Get the protein/DNA seq from <https://www.uniprot.org>.
2. copy the seq in FSATA to download multiple seq.
3. Open FASTA web page <https://www.ebi.ac.uk/Tools/sss/fasta/>.
4. Paste the multiple seq in the box.
5. Run the FASTA omega. You can color it.
6. You see also the phylogenetic tree as well.

11

An example of ClustalW Omega

Results for job clustalo-l20200310-104708-0114-18168141-p2m

Alignments Result Summary Guide Tree Phylogenetic Tree Results Viewers Submission Details

Download Alignment File Hide Colors

CLUSTAL O(1.2.4) multiple sequence alignment

```

UNIPROT:TPM2_BIOGL      -----MDAIKKRMLAMKMEKENAIDRAEQMEQKVRDVEETKPKLEEENLNQKFFSNLQ      54
UNIPROT:TPM1_CAEEI      -----MDAIKKRQAMKIEKDNALDRADAEEKVRQITERLERVEEELRDTPKMMTQTG      54
UNIPROT:TPM1_ANISI      -----MDAIKKRQAMKIEKDNALDRADAEEKVRQITERLERVEEELRDTPKMMQTE      54
UNIPROT:TPM1_TRICO      -----MDAIKKRQAMKIEKDNALDRADAEEKVRQITERLERVEEELRDTPKMMQTE      54
UNIPROT:TPM1_TRIPS      -----MDAIKKRQAMKIEKDNAMDRADAEEKARQQQERVEKLEELERDTPKMMQVE      54
UNIPROT:TPM1_TRISP      -----MDAIKKRQAMKIEKDNAMDRADAEEKARQQQERVEKLEELERDTPKMMQVE      54
UNIPROT:TPM2_BONMO      -----MDAIKKRQAMKLEKDNALDRAAMCEQQAKDANLRAEKABEEARLQKKIQTIE      54
UNIPROT:TPM1_LOCHI      -----MDAIKKRQAMKLEKDNALDRAAMCEQQAKDANLRAEKABEEARLQKKIQTIE      54
UNIPROT:TPM1_PANBO      -----MDAIKKRQAMKLEKDNAMDRADTLEQQNREANNRAEKSEEEVFLQKKLQGLE      54
UNIPROT:TPM1_FENHO      -----MDAIKKRQAMKLEKDNAMDRADTLEQQNREANNRAEKSEEEVHNLQKRMQGLE      54
UNIPROT:TPM1_FENAT      -----MDAIKKRQAMKLEKDNAMDRADTLEQQNREANNRAEKSEEEVHNLQKRMQGLE      54
UNIPROT:TPM1_CHIOP      -----MDAIKKRQAMKLEKDNAMDRADTLEQQNREANNRAEKSEEEVHNLQKRMQGLE      54
UNIPROT:TPM1_ERISI      -----MDAIKKRQAMKLEKDNAMDRADTLEQQNREANNRAEKSEEEVHNLQKRMQGLE      54
UNIPROT:TPM1_HOMAM      -----MDAIKKRQAMKLEKDNAMDRADTLEQQNREANNRAEKSEEEVHNLQKRMQGLE      54
UNIPROT:TPM1_LEPDS      -----MEAIKRNKQAMKLEKDNADRAEIAEQKSRDANLRAEKSEEEVRLQKKIQQIE      54
UNIPROT:TPM1_DERPT      -----MEAIKRNKQAMKLEKDNADRAEIAEQKSRDANLRAEKSEEEVRLQKKIQQIE      54
UNIPROT:TPM1_DERFA      -----MEAIKRNKQAMKLEKDNADRAEIAEQKSRDANLRAEKSEEEVRLQKKIQQIE      54
UNIPROT:TPM1_FSOOV      -----MEAIKRNKQAMKLEKDNADRAEIAEQKSRDANLRAEKSEEEVRLQKKIQQIE      54
UNIPROT:TPM1_HAELO      -----MDAIKKRQAMKLEKDNADRAEIAEQKSRDANLRAEKSEEEVRLQKKIQQIE      54
UNIPROT:TPM1_RHIMP      -----MEAIKRNKQAMKLEKDNADRAEIAEQKSRDANLRAEKSEEEVRLQKKIQQIE      54
UNIPROT:TPM3_RAT        MAGSTTIEAVKRRKIQVLQQA-----                21
UNIPROT:TPM4_RAT        MAGLNSLEAVKRRKIQALQQA-----                21
UNIPROT:TPM4_FIG        MAGLNSLEAVKRRKIQALQQA-----                21
UNIPROT:TPM4_HUMAN      MAGLNSLEAVKRRKIQALQQA-----                21
UNIPROT:TPM4_HORSE      MAGLNSLEAVKRRKIQALQQA-----                21
UNIPROT:TPM4_MOUSE      MAGLNSLEAVKRRKIQALQQA-----                21
UNIPROT:TPM1_CIOIN      -----MEAIKRNKQAMKLEKDNADRAEIAEQKSRDANLRAEKSEEEVRLQKKIQQIE      54

```

12

The old version presentation of the ClustalW

```

TBD_1265/493-734 493 TRLRQALERNELV LHYQPIVELASGRIVGGEALVRWEDPERGLVMPSAFI PA AEDTGLIVALSDWVLEACQTQLRAWQQQG573
YahA7-246          7 E A I S L A L E N H E F K P W I O P V F C A Q T G V L T G C E V L V R W E H P Q T G I I P P D Q F I P L A E S S G L I V I M T R Q L M K T A D I L M P V K H - - 85
FimK2/7-242       7 S E L V H A I Q N G O V Y P V F O P I V D I H L - H I K G I E V L S R W R K D G V - V L L P T E F L P N I Q S E A I W F S L T A F V L G E A V Q G I N R Y Q G - - 83
CKO_03715/1-236  1 R E F I H A I H S Q V F P V F O P I T D G H L - R L Q G V E I L S R W R R G D N - V L L P G E F L P Q I H A E Y A W L L L T A F V L G I A I Q N I N Q H G G - - 77
FimK7-242         7 Q E W V Q A I H D R O V F P V F O P I V D S R S - Q L Q G V E I L I R W R R G Q - V L L P Q T F L P H F R A D Y T W L L T A F V L G E A V Q N I N E Y P G - - 83
PigX/7-240        7 T L L E H T L S R G G P R L Y Q K P A I T R E G - E V H H R E L I S R I Y D G S Q - E L L A A E Y M P L V R Q L G L T A S Y D R Q L I T R S I A L T V S W P - - 82
MrkJ7-230         7 E D N I L S R N D I A V R Y V F C K M F S P Q G - T L V A V E C L S R F D - - - N L S I S P E D F F R H A T - - - - - A A V R R E R I F L E Q L A L I E K H K A - - 76

TBD_1265/493-734 574 R A A D D L T L S V N I S T R O F E G E H L T R A V D R A L A R S G L R P D C L E L E I T E N V M L V M T D E V R T C L D A L R A R G V R L A L D D F G T G Y S S 654
YahA7-246          86 L L P D N F H I G I N V S A G C F L A A G F E K E C L N L V N K I G N D K I K L V L E L T E R N P I P V T R E A R A I F D S L H O H N I T F A L D D F G T G Y A T 166
FimK2/7-242       84 E F Y F T V N I P T C I A H H H L I C L M E T A W L G L H N P L W A D - - C L V L E F A E T V D L T Q Q G N T I A N M R K I Q E R G F R I F L D D C F S Q N S V 162
CKO_03715/1-236  78 K F W F S I N I P P C I A N H E N L L R M M E T A R Q L Q Q P O W S G - - R L V L E F A E T V N L H Q Q G R T A E N M D K I Q R Q G F R I F L D D C F S H S S V 156
FimK7-242         84 T F Y F S V N I P S S L A D S D S L L R M V E A A R Q L R Q P E G V A - - R L V L E Y A E T I D F R H Q S R S A A H V A Q L Q R A G V R V M L D D C F S Q S S V 162
PigX/7-240        83 E A V L A L P I T V D S L L Q R P F L H W L R E T L L C P K K Q R Q R - - - I F F E L A E A D V Q O Y I G R L R P I L S L I S G L G C R L A V T Q A G L T L V S 160
MrkJ7-230         77 - W F L R N H I S A T I N V D D H I L N L L R Q K D I K A K V A A L T C - - - V H F E V T N A E N L L H N S L A A W Q S P Q - - - D T S L W L D D F G S G Y A G 150

TBD_1265/493-734 655 L S Y L S Q L P F H G L K I D O S F V R K I P A H P S E T Q I V T T I L A L A R G L E M E V V A E G I E T A C Q Y A F L R D R G C E F G O G N L M S T P Q A A D 734
YahA7-246          167 Y R Y L G A F P V D F I K I D K S F V Q M A S V D E I S G H I V D N I V E L A R K P G L S I V A E G V E T Q E Q A D L M I G K G V H F L O G Y L Y S P P V P G N 246
FimK2/7-242       163 I P I R L A R F C G Y K L D K S I I N D F Q R D P H A M A L M K S L I Y Y C Q L T Q S D C I A E G V D S L E K F N K L K G M G L V F F G G Y L F S P P V E L E 242
CKO_03715/1-236  157 M F P V R T I R F S G Y K L D M S I V N D F Q R D P H A L A L I K S L L Y Y C Q L T Q S R C I A E G V D S L E K F N Q L K A L G V D R F G G Y L F S P P I T H D 236
FimK7-242         163 I F P A R R L H F N A Y K L D M S I V N D A Q H D P K A L A L I K S L A Y Y C Q L S G S R C V A E G V D S L A K F T Q L K S L G I D R F O G Y L F S P P M R R E 242
PigX/7-240        161 I T Y I K S L Q I E I I K L H P G L V R S L E K R L E N Q L F V G S L E A C K G T H V K V F A V G V R T K S E W Q T L L D K G V C G G G D F F A S S E V G 240
MrkJ7-230         151 I N A I R G Y H E D Y V X I D K D F F W H L M R K E S G R Q L M D A L V T F L S R N H H N V I I E G V E S E A H K E W L O G M E W F A I Q G H Y W R E V S I E Q 230

```

13

What other programs used for MSA?

Because Often multiple sequence alignments require manual editing:

1. **Jalview** is a powerful MSA-editor for MSA. see

<http://www.jalview.org/index.html>

2. **Muscle**: <https://www.ebi.ac.uk/Tools/msa/muscle/>

3. **PRANK**: <https://www.ebi.ac.uk/research/goldman/software/prank.>

4. **MAFFT**: <https://mafft.cbrc.jp/alignment/software/>

14

What are the benefits of MSA?

1. Find out which parts “do the same thing”

Similar genes are conserved across widely divergent species, often performing similar functions.

2. Structure prediction

Use knowledge of structure of one or more members of a protein MSA to predict structure of other members

3. Create “profiles” for protein families

Allow us to search for other members of the family

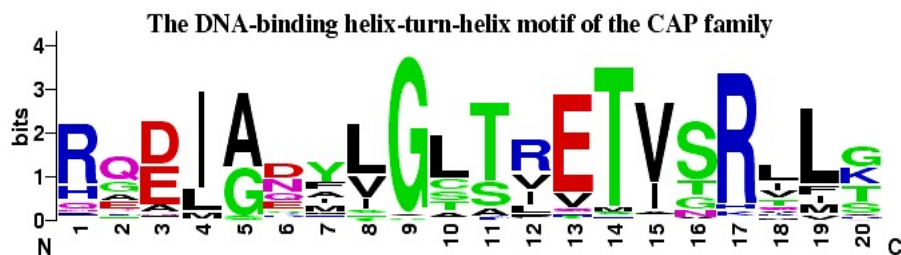
4. Genome assembly: how many gene in this genome.

5. MSA is to build a phylogenetic analysis.

15

How to find the most conservative amino acid in a seq among multiple species?

- **Sequence Logos and conservativity can be found using**
- <http://weblogo.berkeley.edu/>
- Sequence logos are based on **Multiple Sequence Alignments**
- Very useful to visualise Sequence profiles and motifs.



16

Tutorial

- Find a TPM1 (tropomyosin) gene for human by typing it in www.uniprot.com. Type the gene name
- Go inside the page do click alignment.
- The job will take time.
- Download seq, paste it in <https://www.uniprot.org/blast/uniprot/B202003208BC4D7ADE02784B0C2481C7F3DE0963A0E5076S>
- Download the whole seq, paste it in <http://weblogo.berkeley.edu/logo.cgi>

