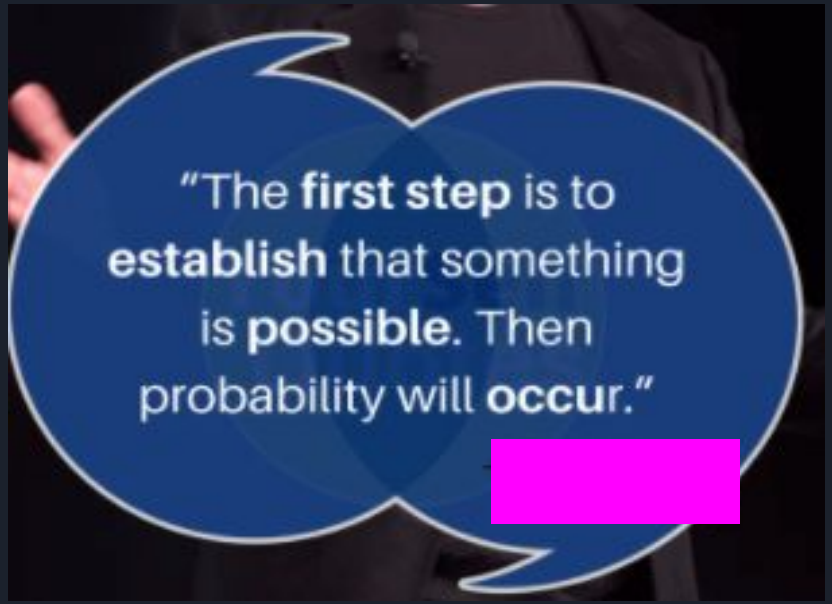


Jyoti Kataria
Bioinformatics Analyst, Medgenome Labs Ltd.
M.Sc. Biotechnology (IIT Bombay)

Evaluation

- Continuous
- In class presentations/assignments
- Group work

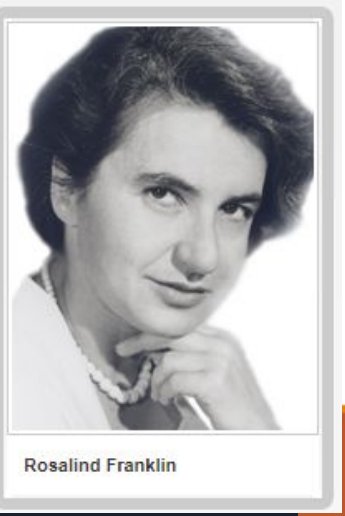
Agenda

What is biological data & Where is it coming from?

What does it mean?

Why to analyze?

How to analyze?



Rosalind Franklin

James Watson and Francis Crick

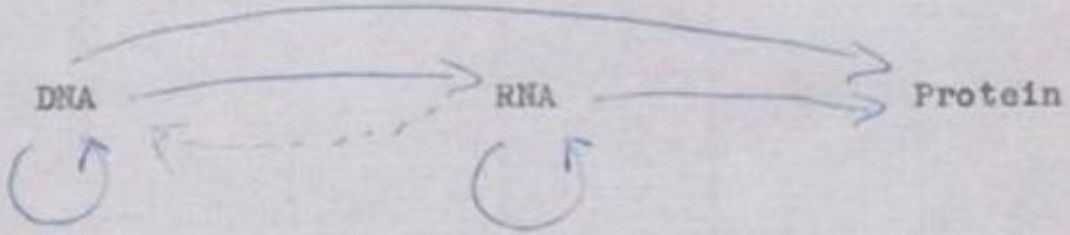
1953 – Structure of the DNA
1962 – Nobel Prize in Medicine
1965 – The first tRNA
1986 – Robert Holley Nobel Prize

Without minimizing the pleasure of receiving awards and prizes, I think it is true that the greatest satisfaction for a scientist comes from carrying a major piece of research to a successful conclusion" (Holley, 1968)

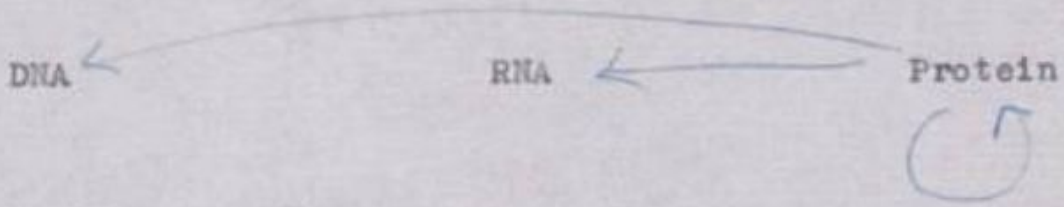
Let's get back in time!

Central Dogma of Life

The Central Dogma: "Once information has got into a protein it can't get out again". Information here means the sequence of the amino acid residues, or other sequences related to it. That is, we may be able to have



but never



where the arrows show the transfer of information.

Sequencing Breakthrough : The first generation

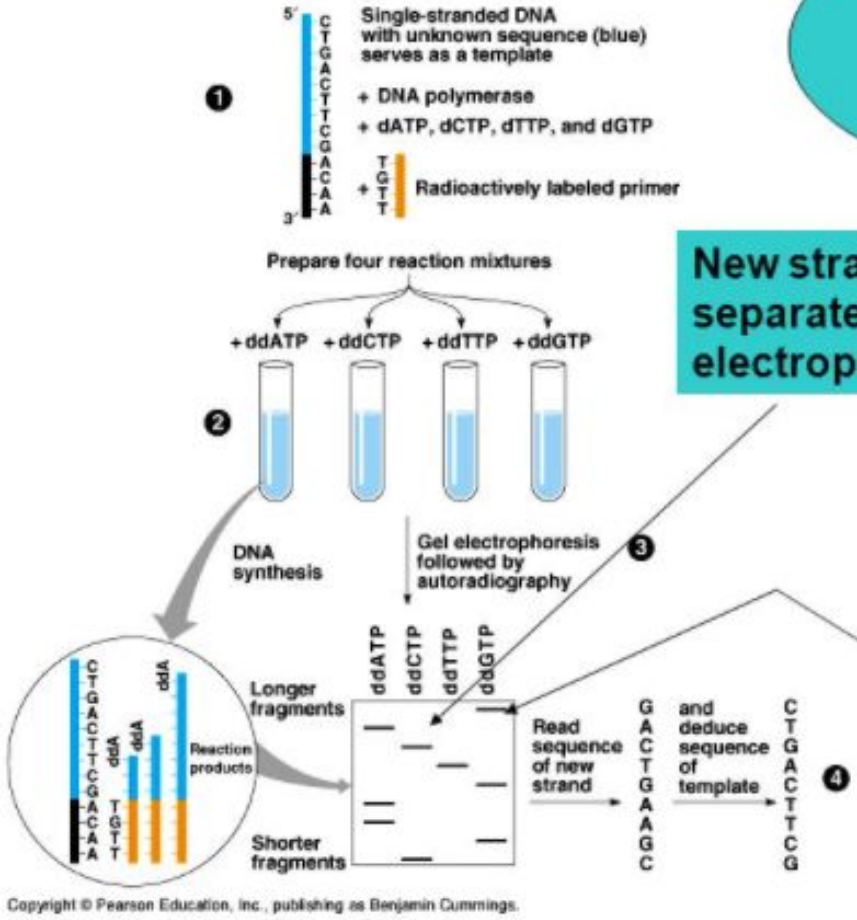
1. Chain termination method or Sanger's dideoxy method (1977)
2. Chemical degradation method or Maxam and Gilbert's method (1977)

1958 – Insulin
1980 – Sequencing

Nobel Prize in Chemistry

Scientific research is one of the most exciting and rewarding of occupations. It is like a voyage of discovery into unknown lands, seeking not for new territory but for new knowledge. It should appeal to those with a good sense of adventure.
Frederick Sanger's speech at the Nobel Banquet, December 10, 1980

SANGER METHOD



New strands separated by electrophoresis

Sequence can be read from bands on autoradiograph and original template sequence deduced. Longest fragment ends with a ddG, so G must be the last base in the sequence

The Human Genome Project (1990-2003)

- Identify all the approximately 20,000-25,000 genes in human DNA
- Determine the sequences of the 3 billion chemical base pairs that make up human DNA
- Store this information in databases
- Improve tools for data analysis

https://web.ornl.gov/sci/techresources/Human_Genome/project/index.shtml

U.S. Human Genome Project Funding ($Millions)			
FY	DOE	NIH*	U.S. Total
1988	10.7	17.2	27.9
1989	18.5	28.2	46.7
1990	27.2	59.5	86.7
1991	47.4	87.4	134.8
1992	59.4	104.8	164.2
1993	63.0	106.1	169.1
1994	63.3	127.0	190.3
1995	68.7	153.8	222.5
1996	73.9	169.3	243.2
1997	77.9	188.9	266.8
1998	85.5	218.3	303.8
1999	89.9	225.7	315.6
2000	88.9	271.7	360.6
2001	86.4	308.4	394.8
2002	90.1	346.7	434.3
2003	64.2	372.8	437

The Human Genome Project (1990-2003)

"The genome could be thought of in terms of a book with multiple uses: "It's a history book - a narrative of the journey of our species through time. It's a shop manual, with an incredibly detailed blueprint for building every human cell. And it's a transformative textbook of medicine, with insights that will give health care providers immense new powers to treat, prevent and cure disease."

Francis Collins
ForMemRS

16th Director of the National Institutes of Health

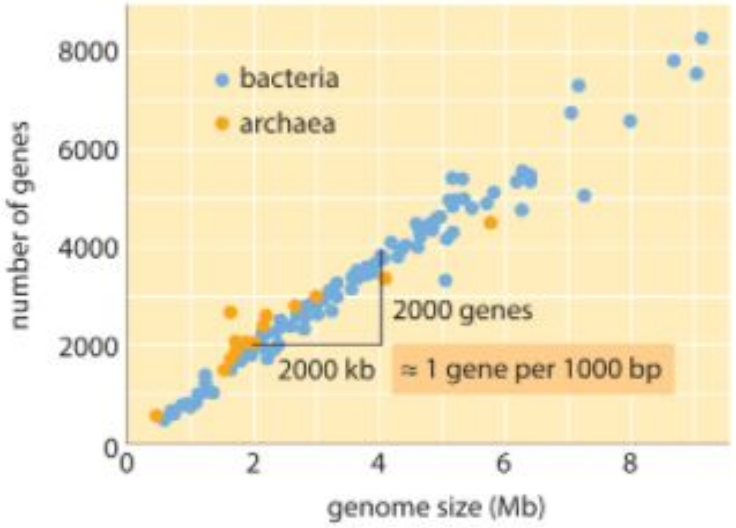
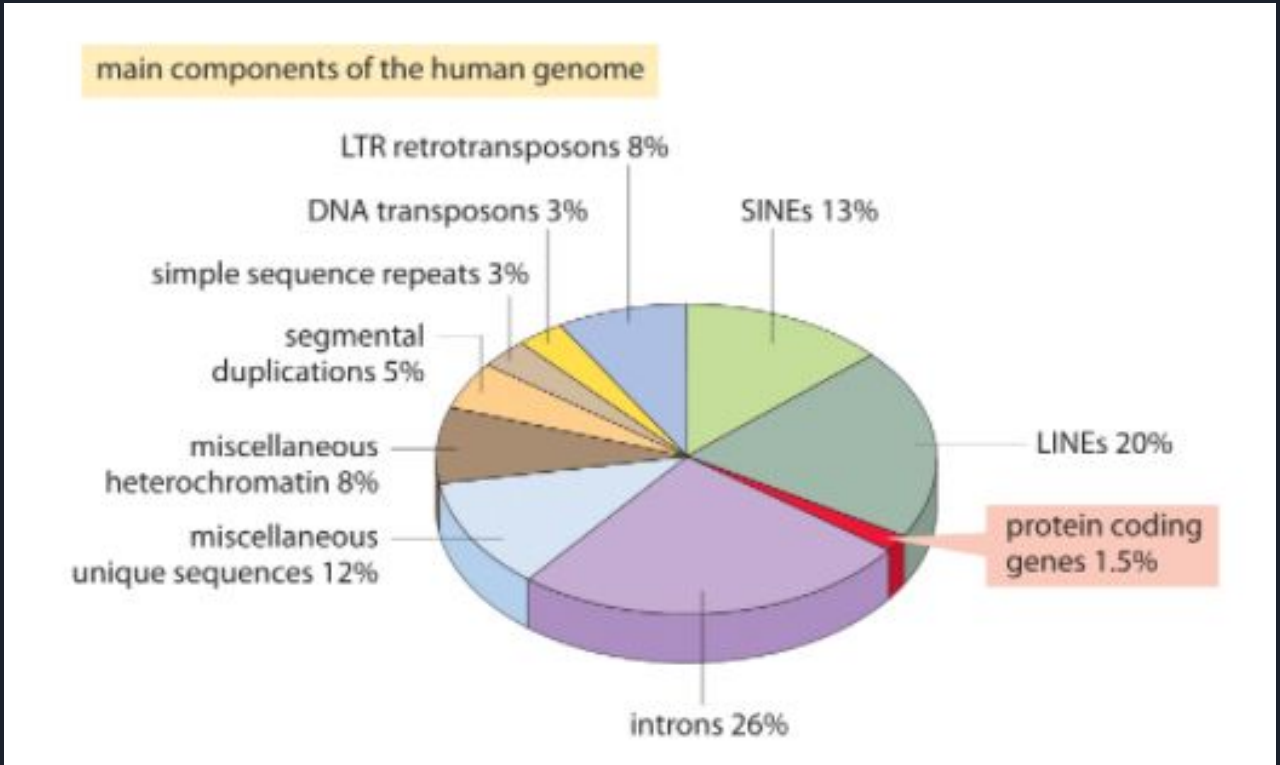


Figure 1: Number of genes as a function of genome size. The figure shows data for a variety of bacteria and archaea, with the slope of the data line confirming the simple rule of thumb relating genome size and gene number. (Adapted from M. Lynch, The Origins of Genome Architecture.)

Genome size & no. of genes

What all is there inside this huge DNA?

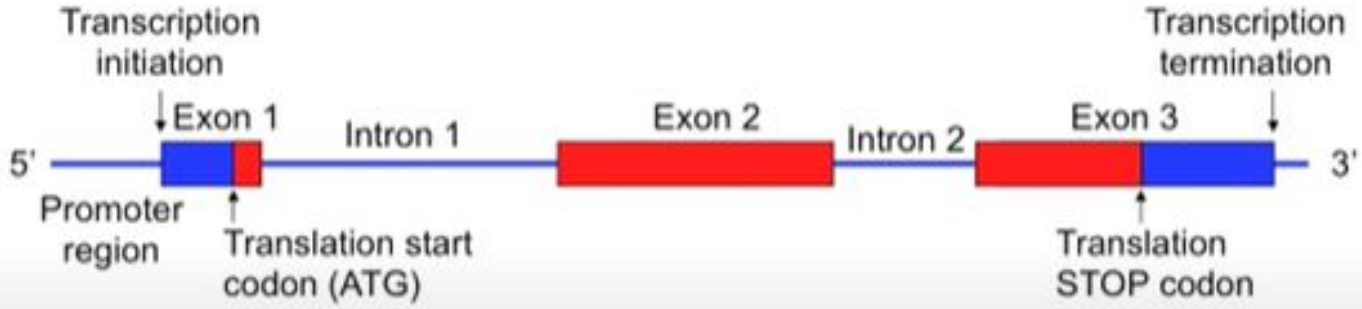


Genome Size for some species

Genome Size and Number of Protein-Coding Genes for a Select Handful of Species

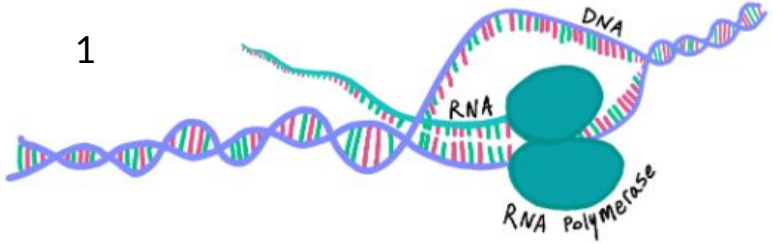
Species and Common Name	Estimated Total Size of Genome (bp)*	Estimated Number of Protein-Encoding Genes*
Saccharomyces cerevisiae (unicellular budding yeast)	12 million	6,000
Plasmodium falciparum (unicellular **malaria parasite**)	23 million	5,000
Drosophila melanogaster (**fruit fly**)	170 million	14,000
Arabidopsis thaliana (mustard; thale cress)	125 million	25,000
Oryza sativa (rice)	470 million	51,000
Gallus gallus (chicken)	1 billion	20,000-23,000
Canis familiaris (domestic dog)	2.4 billion	19,000
Mus musculus (laboratory mouse)	2.5 billion	30,000
Homo sapiens (human)	2.9 billion	20,000-25,000

Gene Structure

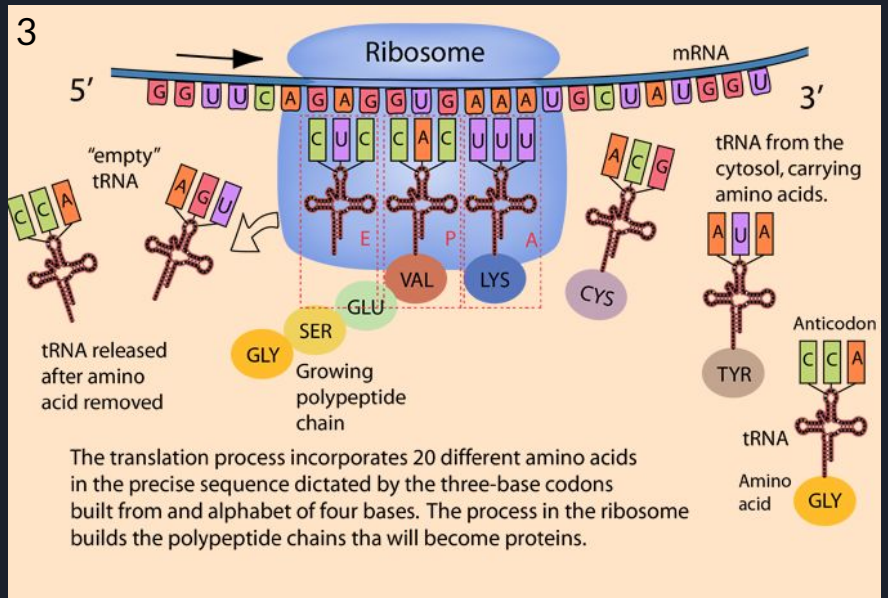


Central dogma again!! DNA > RNA > protein

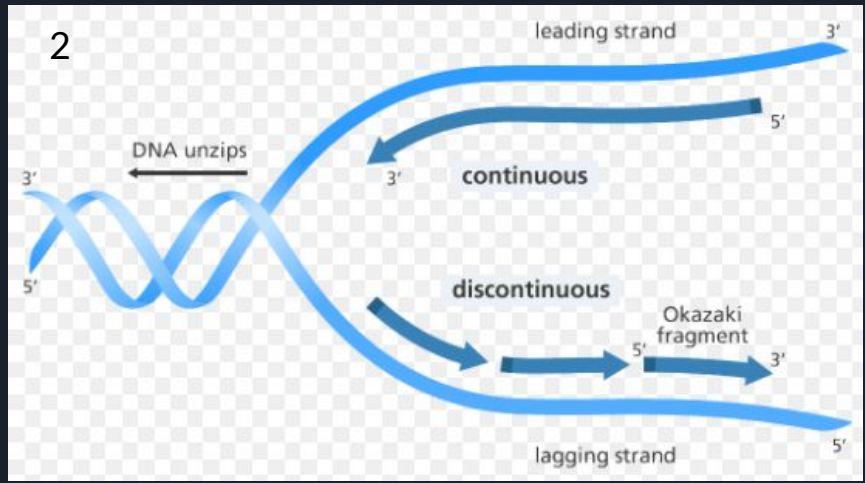
1

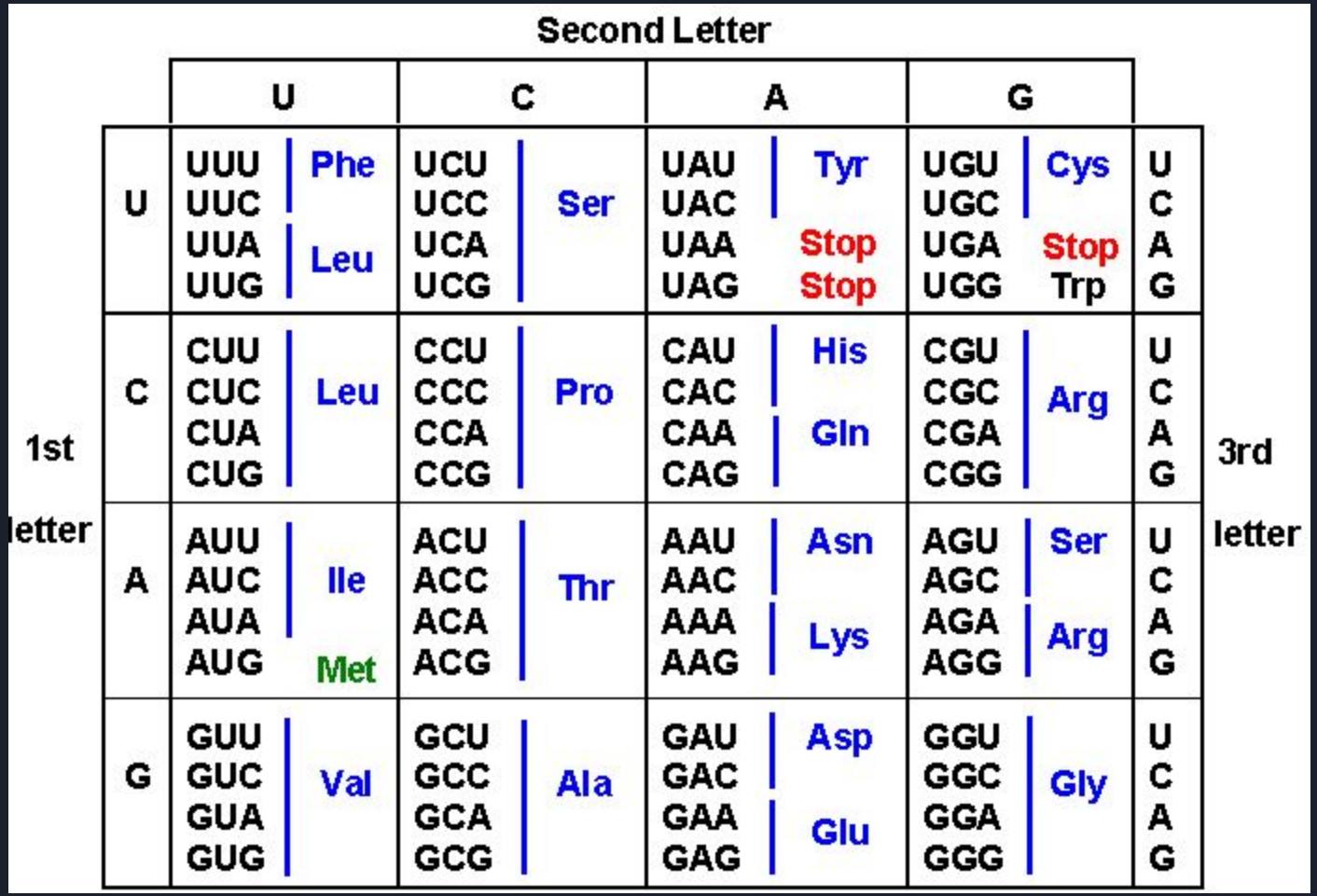


3



2




Second Letter
U
C
A
G
1st letter
3rd letter
UUU Phe
UUC
UUA Leu
UUG
UCU
UCC Ser
UCA
UCG
UAU Tyr
UAC
UAA Stop
UAG Stop
UGU Cys
UGC
UGA Stop
UGG Trp
CUU
CUC Leu
CUA
CUG
CCU
CCC Pro
CCA
CCG
CAU His
CAC
CAA Gln
CAG
CGU
CGC Arg
CGA
CGG
AUU
AUC Ile
AUA
AUG Met
ACU
ACC Thr
ACA
ACG
AAU Asn
AAC
AAA Lys
AAG
AGU Ser
AGC
AGA Arg
AGG
GUU
GUC Val
GUA
GUG
GCU
GCC Ala
GCA
GCG
GAU Asp
GAC
GAA Glu
GAG
GGU
GGC Gly
GGA
GGG

Inclass Activity 1 :
Biological databases trip (10-15 mins)

Take the assigned databases from the google sheet, find out the answers to the following three questions :-

1. What does it have?
2. How is the data stored inside it?
3. Why this database is important or Who all would be the potential users for the same?

Quick Recap

DNA & RNA

Central Dogma

Gene

Introns & Exons

Dna replication

Transcription

Translation

Codons

Databases