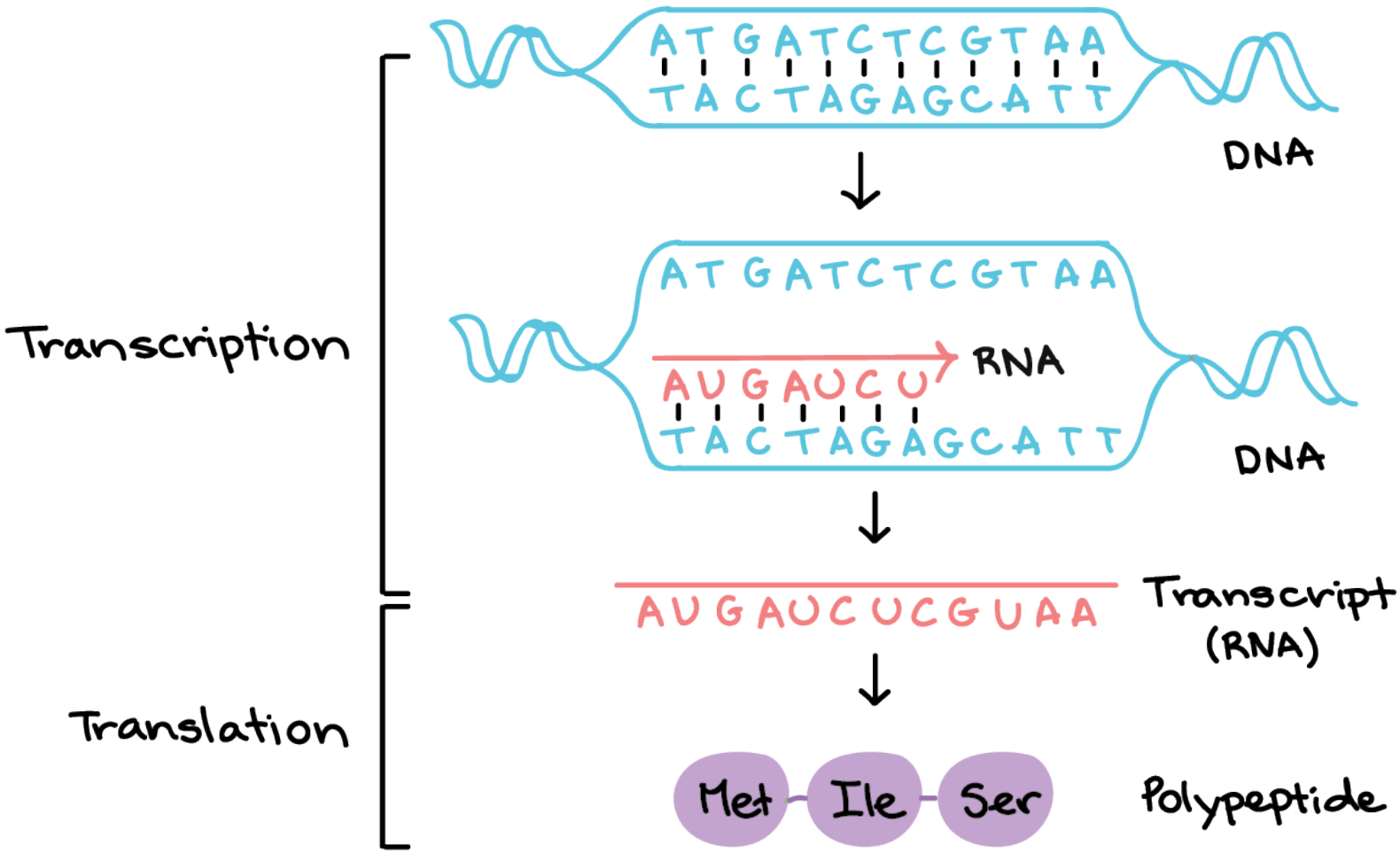
Reading Frame, Types of Codons

Manu Madhavan

Lecture 3

Recap

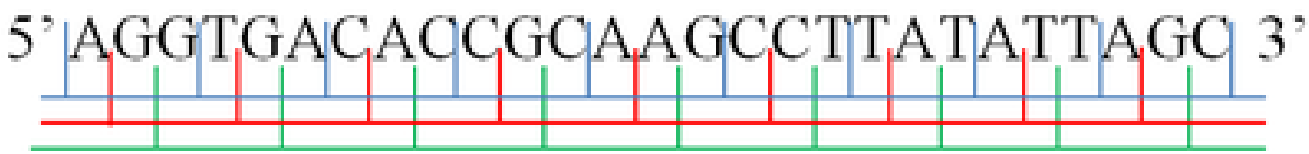

Transcription
Translation
ATGATCTCGTAA
TACTAGAGCATT
DNA
ATGATCTCGTAA
AUGAUCU
RNA
TACTAGAGCATT
DNA
AUGAUCUCGUAA
Transcript
(RNA)
Met
Ile
Ser
Polypeptide

Outline

- Open Reading frame
- Types of codons
- Biological Databases: Quick Introduction

Reading Frame

- A **Reading frame** is a way of dividing the sequence of nucleotides in a nucleic acid (DNA or RNA) molecule into a set of consecutive, non-overlapping triplets
- These triplets are called **Codons**, specifies an Amino Acids
- The codon **AUG** has a special role, serving as the start codon where translation begins
- There are three stop codons in the genetic code, UAA, UAG, and UGA.
- The complete set of correspondences between codons and amino acids (or stop signals) is known as the **genetic code**

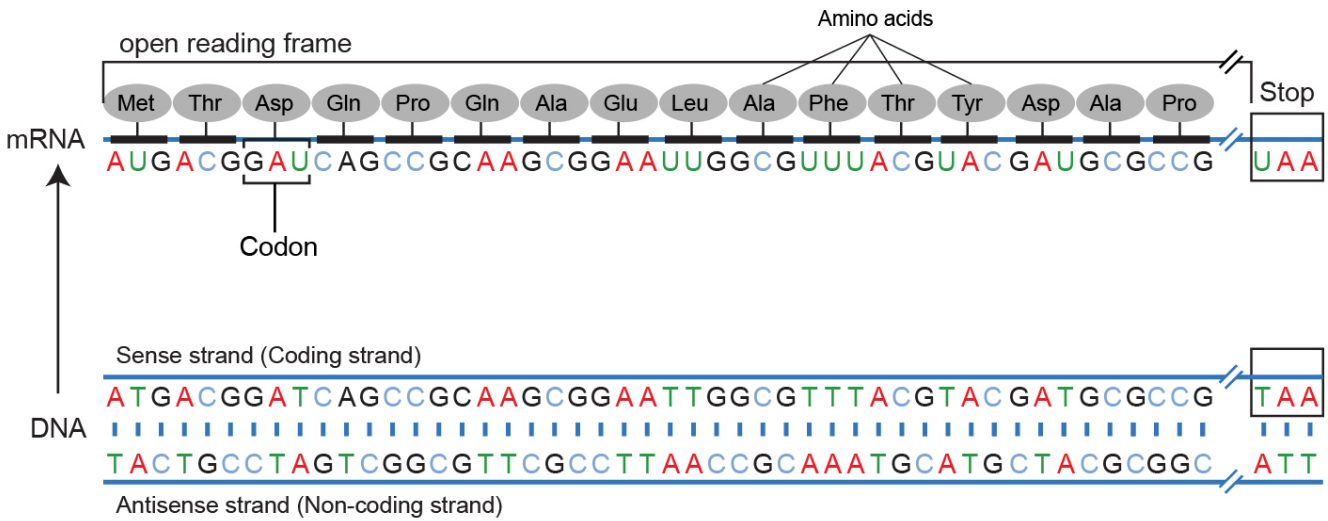


Open Reading Frame

- The region of the nucleotide sequences from the start codon (ATG) to the stop codon is called the **Open Reading frame**
- "Reading" $\rightarrow$ the RNA code is being read by the ribosomes in order to make a protein
- "Open" $\rightarrow$the road is open to keep reading, and the ribosome will be able to keep reading the RNA code and add another amino acid one after another
- The longer an ORF is, the longer you get before you get to a stop codon, the more likely it is to be part of a gene which is coding for a protein.

Open Reading Frame

- The region of the nucleotide sequences from the start codon (ATG) to the stop codon is called the **Open Reading frame**



Open Reading Frame

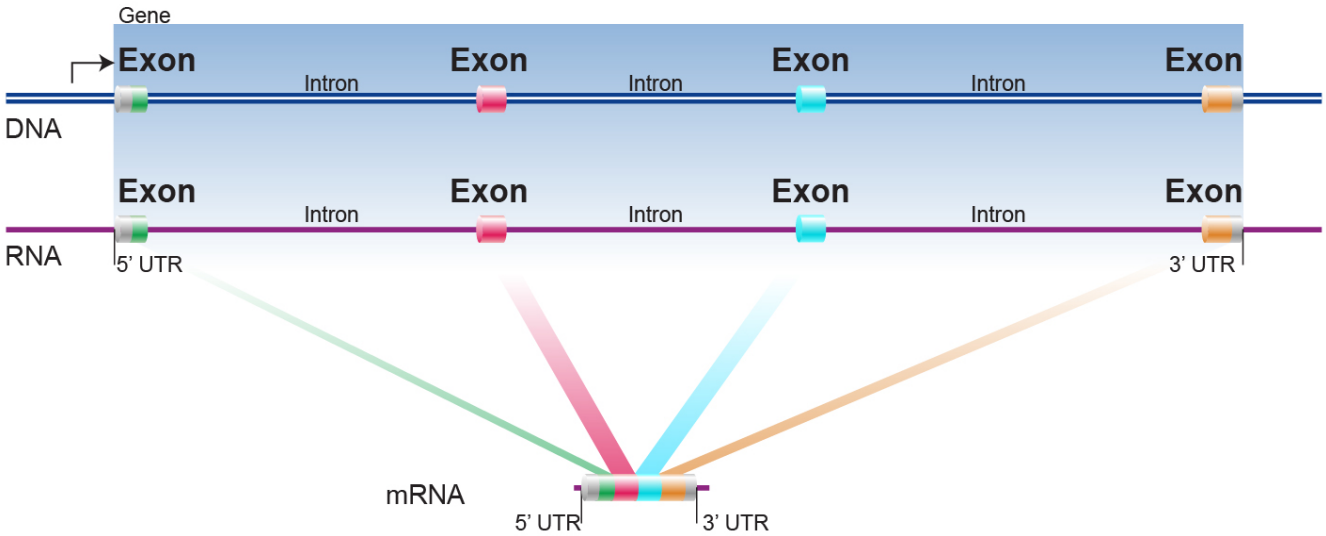
- **Six reading frames**: three on the positive strand, and three (which are read in the reverse direction) on the negative strand

```
reading frame:                  123
                                |||
                                acttacccgggacta

first reading frame               T  Y  P  G  L
second reading frame               L  T  R  D
third reading frame                 L  P  G  T
```

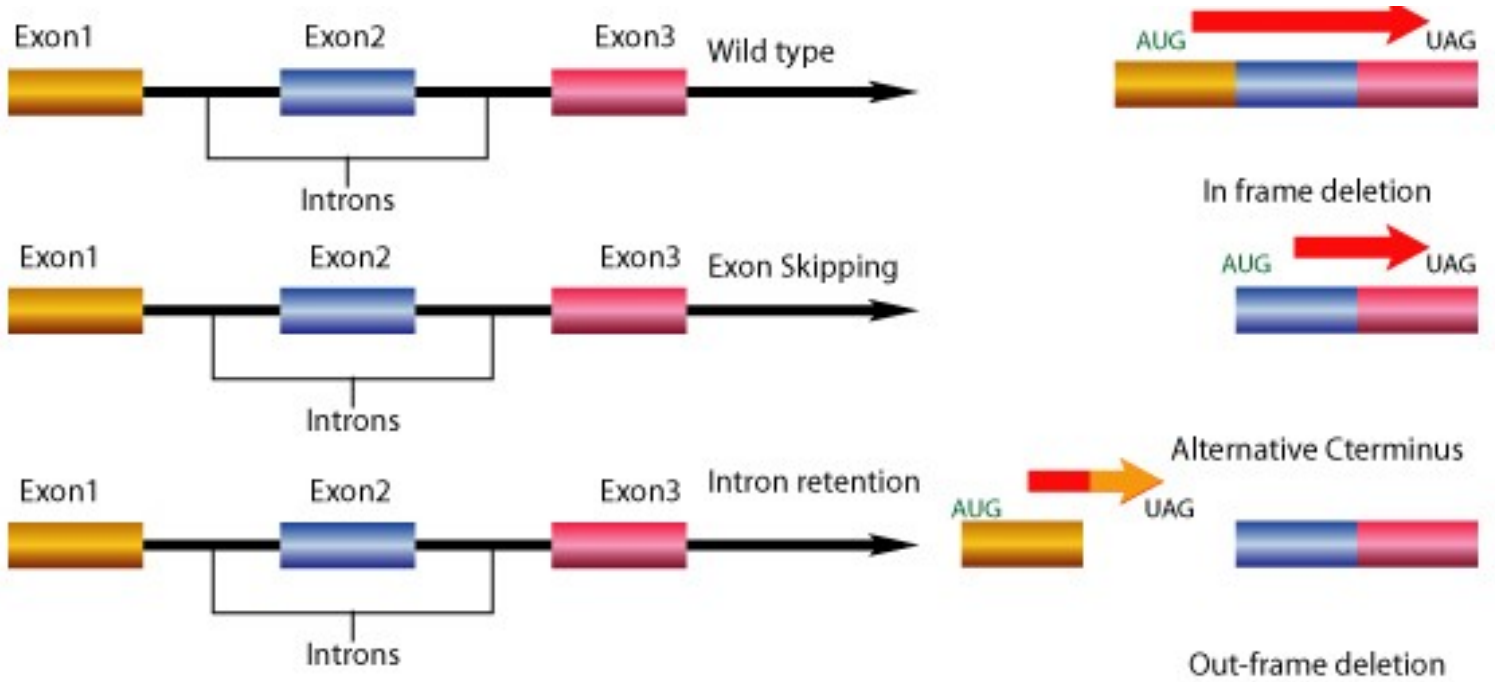
Exons and Introns

- In Prokaryotes, the ORF is continuous (exactly same as the RNA transcribed from DNA)
- But, in Eukaryotes, most gene sequences are broken up by one or more DNA sequences called **introns**
- An **exon** is the portion of a gene that codes for amino acids



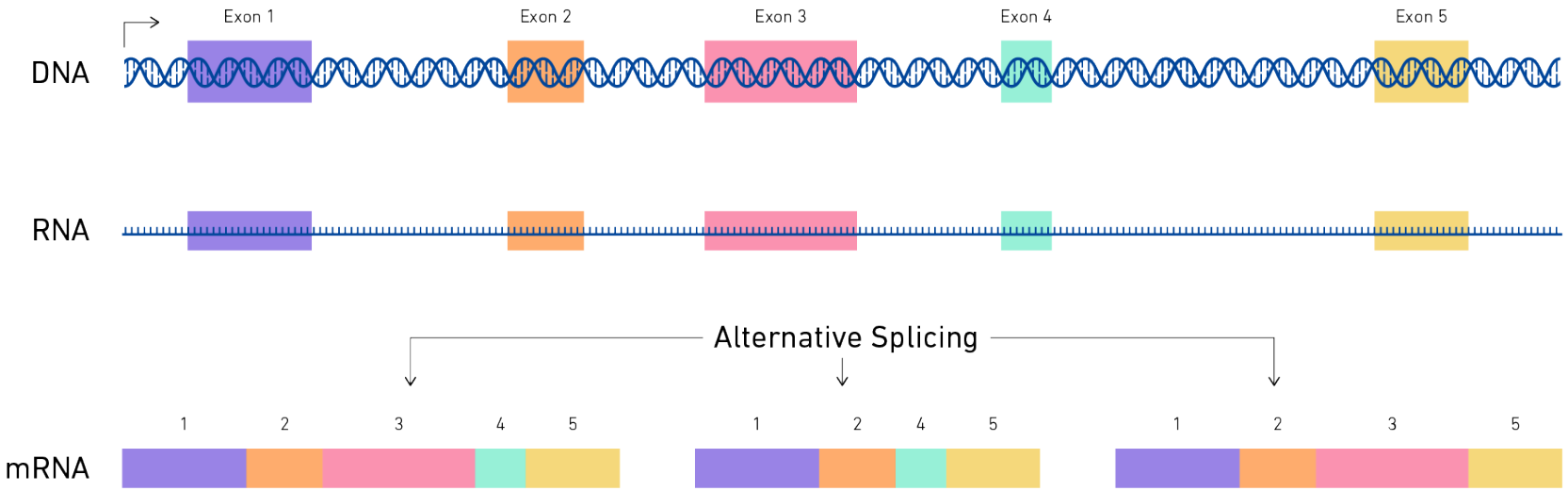
Gene Splicing

- **Splicing**: the precise excision of internal sequences known as introns and the rejoining of the exons that flank them
- The enzyme complexes responsible for splicing in eukaryotes, **spliceosomes**



Gene Splicing

- Same gene can code for multiple proteins by **differential inclusion and exclusion** of exons
- This property is called Alternative splicing



ORF Finding Algorithm

Objective:

- Given a DNA sequence, identify all the possible open reading frames in the sequence

Steps:

- Divide the sequence into 6 different reading frames(+1, +2, +3, -1, -2 and -3)
- Mark the start codon and stop codons in the reading frames
- Identify the open reading frame (ORF) - sequence stretch begining with a start codon and ending in a stop codon
- Based on the amino acid table the peptide sequence is found

ORF Finding Algorithm

Input: CGCTACGTCTTACGCTGGAGCTCTCATGGATCGGTTCGG-TAGGGCTCGATCACATCGCTAGCCAT

- Divide the sequence into 6 different reading frames(+1, +2, +3, -1, -2 and -3)

FRAME +1: CGC TAC GTC TTA CGC TGG AGC TCT CAT GGA TCG GTT CGG TAG GGC TCG ATC ACA TCG CTA GCC AT

The second reading frame is formed after leaving the first nucleotide and then grouping the sequence into words of 3 nucleotides

FRAME +2: C GCT ACG TCT TAC GCT GGA GCT CTC ATG GAT CGG TTC GGT AGG GCT CGA TCA CAT CGC TAG CCA T

The third reading frame is formed after leaving the first 2 nucleotides and then grouping the sequence into words of 3 nucleotides

FRAME +3: CG CTA CGT CTT ACG CTG GAG CTC TCA TGG ATC GGT TCG GTA GGG CTC GAT CAC ATC GCT AGC CAT

The other 3 reading frames can be found only after finding the reverse complement.
Complement : **GCGATGCAGAATGCGACCTCGAGAGTACCTAGCCAAGCCATCCCGAGCTAGTGTAGCGATCGGTA**
Reverse complement: **ATGGCTAGCGATGTGATCGAGCCCTACCGAACCGATCCATGAGAGCTCCAGCGTAAGACGTAGCG**

Now same process as that of +1, +2 and +3 strands is repeated for -1, -2 and -3 strands with reverse complement sequence
FRAME -1: ATG GCT AGC GAT GTG ATC GAG CCC TAC CGA ACC GAT CCA TGA GAG CTC CAG CGT AAG ACG TAG CG
FRAME -2: A TGG CTA GCG ATG TGA TCG AGC CCT ACC GAA CCG ATC CAT GAG AGC TCC AGC GTA AGA CGT AGC G
FRAME -3: AT GGC TAG CGA TGT GAT CGA GCC CTA CCG AAC CGA TCC ATG AGA GCT CCA GCG TAA GAC GTA GCG

ORF Finding Algorithm

Input: CGCTACGTCTTACGCTGGAGCTCTCATGGATCGGTTCGG-TAGGGCTCGATCACATCGCTAGCCAT

- Mark the start codon and stop codons in the reading frames

```
FRAME +1:  CGC TAC GTC TTA CGC TGG AGC TCT CAT GGA TCG GTT CGG TAG GGC TCG ATC ACA TCG CTA GCC AT
FRAME +2:  C GCT ACG TCT TAC GCT GGA GCT CTC ATG GAT CGG TTC GGT AGG GCT CGA TCA CAT CGC TAG CCA T
FRAME +3:  CG CTA CGT CTT ACG CTG GAG CTC TCA TGG ATC GGT TCG GTA GGG CTC GAT CAC ATC GCT AGC CAT
FRAME -1:  ATG GCT AGC GAT GTG ATC GAG CCC TAC CGA ACC GAT CCA TGA GAG CTC CAG CGT AAG ACG TAG CG
FRAME -2:  A TGG CTA GCG ATG TGA TCG AGC CCT ACC GAA CCG ATC CAT GAG AGC TCC AGC GTA AGA CGT AGC G
FRAME -3:  AT GGC TAG CGA TGT GAT CGA GCC CTA CCG AAC CGA TCC ATG AGA GCT CCA GCG TAA GAC GTA GCG
```

ORF Finding Algorithm

Input: CGCTACGTCTTACGCTGGAGCTCTCATGGATCGGTTCGG-TAGGGCTCGATCACATCGCTAGCCAT

- Identify ORF: sequence stretch beginning with a start codon and ending in a stop codon

FRAME +2: **ATG** GAT CGG TTC GGT AGG GCT CGA TCA CAT CGC **TAG**
FRAME -1: **ATG** GCT AGC GAT GTG ATC GAG CCC TAC CGA ACC GAT CCA **TGA**
FRAME -3: **ATG** AGA GCT CCA GCG **TAA**

ORF Finding Algorithm

- Based on the amino acid table the peptide sequence is found

First Nucleotide	Second Nucleotide: U code	U Amino acid	C code	C Amino acid	A code	A Amino acid	G code	G Amino acid	Third Nucleotide
U	UUU	phe	UCU	ser	UAU	tyr	UGU	cys	U
	UUC		UCC		UAC		UGC		C
	UUA	leu	UCA		UAA	STOP	UGA	STOP	A
	UUG		UCG		UAG	STOP	UGG	trp	G
C	CUU	leu	CCU	pro	CAU	his	CGU	arg	U
	CUC		CCC		CAA		CGC		C
	CUA		CCA		CAC	gln	CGA		A
	CUG		CCG		CAG		CGG		G
A	AUU	ile	ACU	thr	AAU	asn	AGU	ser	U
	AUC		ACC		AAC		AGC		C
	AUA		ACA		AAA	lys	AGA	arg	A
	AUG	met	ACG		AAG		AGG		G
G	GUU	val	GCU	ala	GAU	asp	GGU	gly	U
	GUC		GCC		GAC		GGC		C
	GUA		GCA		GAA	glu	GGA		A
	GUG		GCG		GAG		GGG		G

FRAME +2: ATG GAT CGG TTC GGT AGG GCT CGA TCA CAT CGC TAG
met asp arg phe gly arg ala arg ser his arg stop

FRAME -1: ATG GCT AGC GAT GTG ATC GAG CCC TAC CGA ACC GAT CCA TGA
met ala ser asp val ile glu pro tyr arg thr asp pro stop

FRAME -3: ATG AGA GCT CCA GCG TAA
met arg ala pro ala stop

Try yourself

- Identify the ORFs and polypeptides for the sequence: `ATGCAATGGGGAAATGTTACCAGGTCCGAACTTATTCAGGTAAGACAGATTTAA`
- You can also explore: ORF Finder: `https://www.ncbi.nlm.nih.gov/orffinder/`
- ORFPY - python package

Biological Databases

Manu Madhavan

Lecture 4

Biological Databases

- Store and handle the staggering volume of Biological information through the establishment and use of computer databases
- Current biological databases use all three types of database structures: flat files, relational, and object oriented
- Based on their contents, biological databases can be roughly divided into three categories: primary databases, secondary databases, and specialized databases.

Primary Databases

- Contain original biological data. They are archives of raw sequence or structural data submitted by the scientific community
- GenBank, the European Molecular Biology Laboratory (EMBL) database, Protein Data Bank (PDB) and the DNA Data Bank of Japan (DDBJ)

Secondary Databases

- Secondary databases contain computationally processed or manually curated information, based on original information from primary databases.
- Translated protein sequence databases containing functional annotation belong to this category
- SWISS-PROT,

Specialized Databases

- Specialized databases normally serve a specific research community or focus on a particular organism
- The content of these databases may be sequences or other types of information
- Examples include Flybase, WormBase, AceDB, and TAIR

Information Retrieval from Biological Databases

- The most popular retrieval systems for biological databases are **Entrez** and **Sequence Retrieval Systems (SRS)**
- Join a series of keywords using logical terms such as AND, OR, and NOT to indicate relationships between the keywords used in a search
- Entrez, a biological database retrieval system by NCBI
- For a complex search, a user can use the Boolean operators
- Online Mendelian Inheritance in Man (OMIM) accessible from Entrez, which is a non-sequence-based database of human disease genes and human genetic disorders

GenBank

- GenBank is the most complete collection of annotated nucleic acid sequence data for almost every organism.
- The content includes genomic DNA, mRNA, cDNA, ESTs, high throughput raw sequence data, and sequence polymorphisms
- There is also a GenPept database for protein sequences

GenBank: Sequence Format

Header

```
LOCUS       SMU25150                359 bp    DNA     linear   BCT 09-MAY-1995
DEFINITION  Serratia marcescens HU beta (hupB) gene, complete cds.
ACCESSION   U25150
VERSION     U25150.1
KEYWORDS    .
SOURCE      Serratia marcescens
  ORGANISM  Serratia marcescens
            Bacteria; Proteobacteria; Gammaproteobacteria; Enterobacterales;
            Yersiniaceae; Serratia.
REFERENCE   1  (bases 1 to 359)
  AUTHORS   Oberto,J. and Rouviere-Yaniv,J.
  TITLE     Direct Submission
  JOURNAL   Submitted (18-APR-1995) Jacques Oberto, Physiologie Bacterienne,
            IBPC, 13 rue Pierre et Marie Curie, Paris, 75005, France
```

Gene annotation

```
FEATURES             Location/Qualifiers
     source          1..359
                     /organism="Serratia marcescens"
                     /mol_type="genomic DNA"
                     /strain="SM369"
                     /db_xref="taxon:615"
     gene            79..351
                     /gene="hupB"
     CDS             79..351
                     /gene="hupB"
                     /note="histone-like protein"
                     /codon_start=1
                     /transl_table=11
                     /product="HU beta"
                     /protein_id="AAA65988.1"
                     /translation="MNKSQLIDKIAAGADISKAAAGRALDAVIASVTDSLKAGDDVAL
                     VGFGSFTVRERSARTGRNPQTGKEIKIAARKVPAFRAGKALKDAVN"
```

DNA sequence

```
ORIGIN
        1 cgctaagtta gatctctgtc ggcccgcgtt ttgtcaccca gtcggtggct tgcaaggttc
       61 gatgggattg atataacagt gaataagtca caactgatcg acaagattgc ggcaggtgct
      121 gatatttcca aagcggcagc gggacgtgct ttagacgcag taatcgcttc cgttaccgac
      181 tccctgaaag caggggatga cgtggctctg gtaggtttcg gttcctttac cgtgcgtgaa
      241 cgttcggccc gtaccggccg caacccgcag accggtaaag agatcaagat cgcggcacgc
      301 aaagtacctg ccttccgtgc agggaaagcg ctgaaagacg cggtaaacta agcggatcc
//
```

Fasta: Sequence Format

Reading Assignment

- Read more on Biological Databases: `https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4411498/`
- Explore Entrez: `https://www.ncbi.nlm.nih.gov/search/`
- Explore NCBI databases

Next

- Sequence Alignment