Programme: M.Sc., Zoology
Course: HC 4.1 Advanced Genetics and Computational Biology

Gene annotation

By
Dr. Nijagal B.S.
Assistant professor
P.G. Dept of Zoology
J.S.S. College of Arts, Commerce and Science.
Ooty Road, Mysore-25

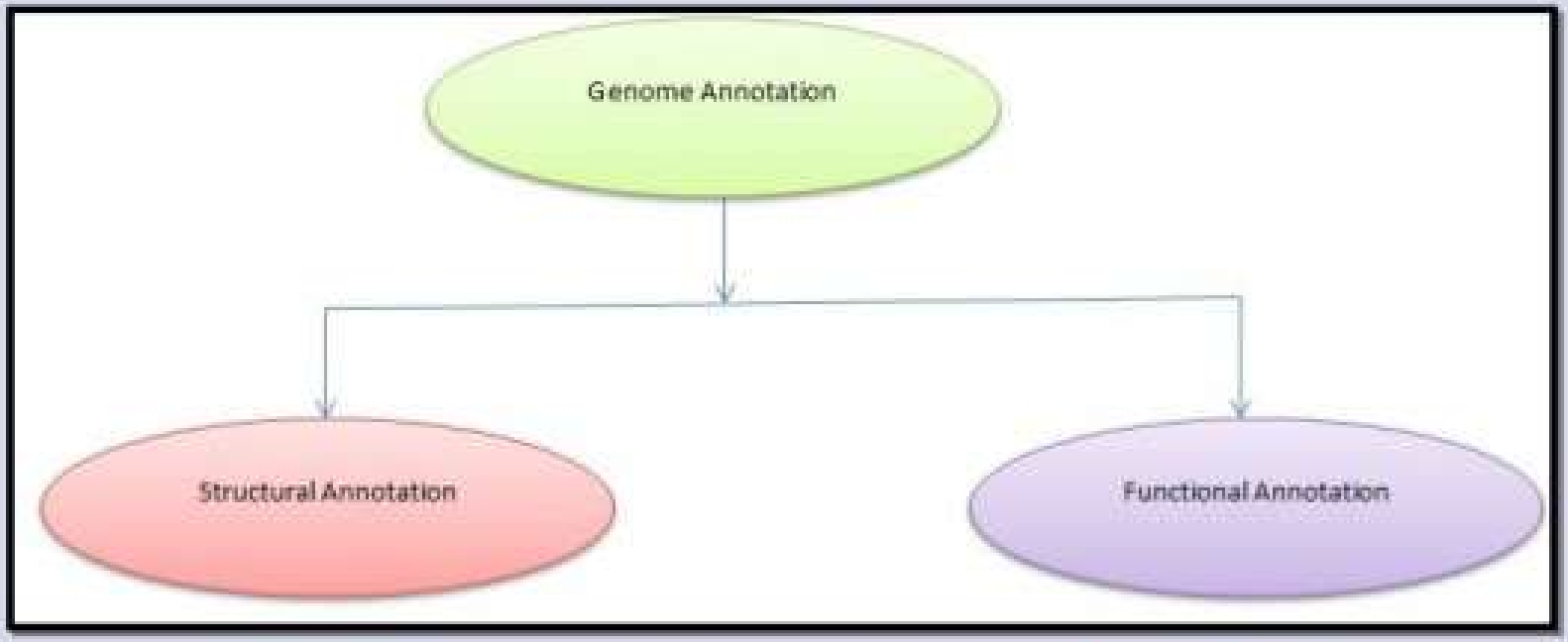
The Genome

- The genome contains all the biological information required to build and maintain any given living organism
- The genome contains the organisms molecular history
- Decoding the biological information encoded in these molecules will have enormous impact in our understanding of biology

Genome annotation

- The process of identifying the locations of genes and the coding sequences in a genome to determine what genes do
- Finding and attaching the structural elements and its related function to each genome locations

Genome Annotation

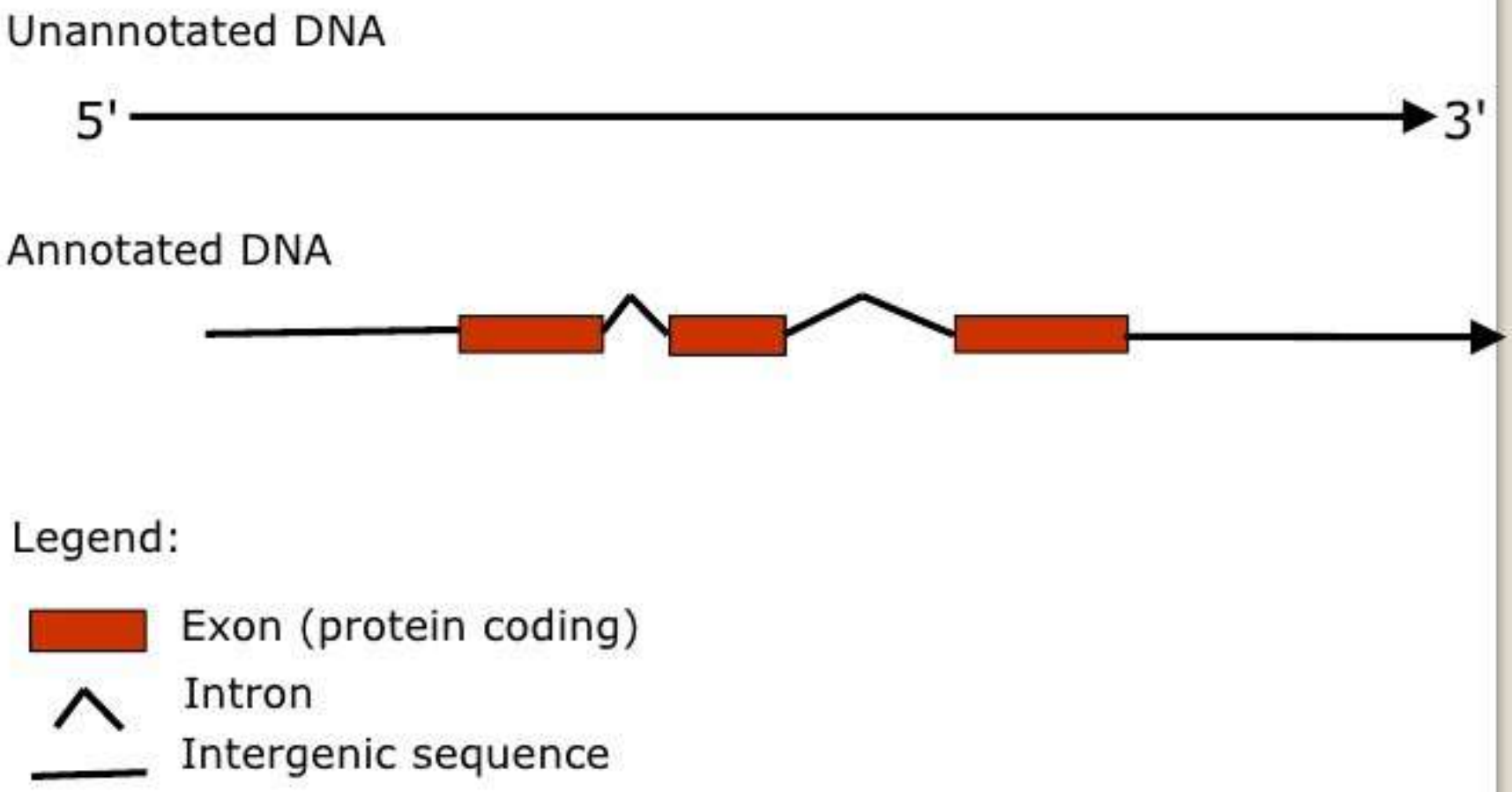


gene structure prediction

Identifying elements (Introns/exons,CDS,stop,start) in the genome

gene function prediction

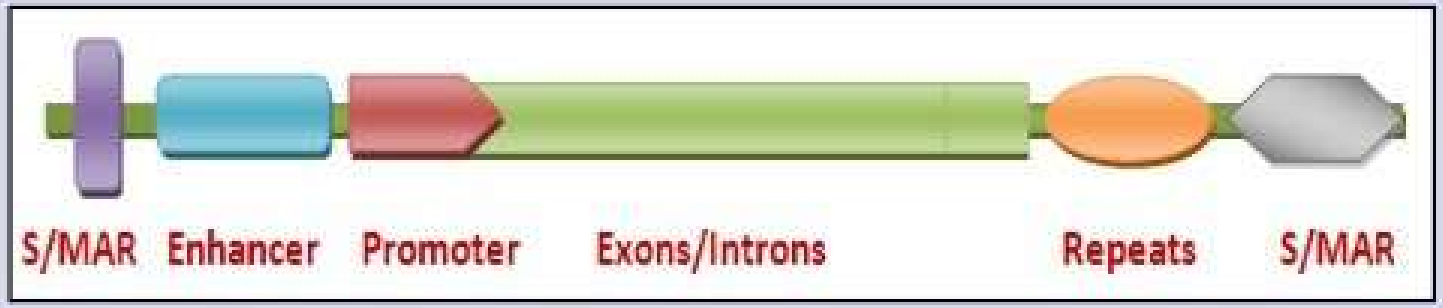
Attaching biological information to these elements- eg: for which protein exon will code for


Unannotated DNA
5'
3'
Annotated DNA
Legend:
Exon (protein coding)
Intron
Intergenic sequence

Structural annotation

Structural annotation - identification of genomic elements

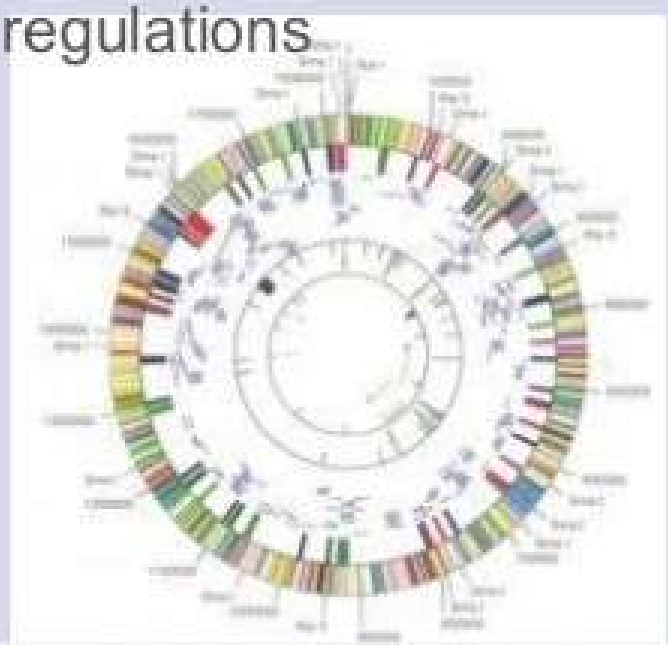
- Open reading frame and their localisation
- gene structure
- coding regions
- location of regulatory motifs

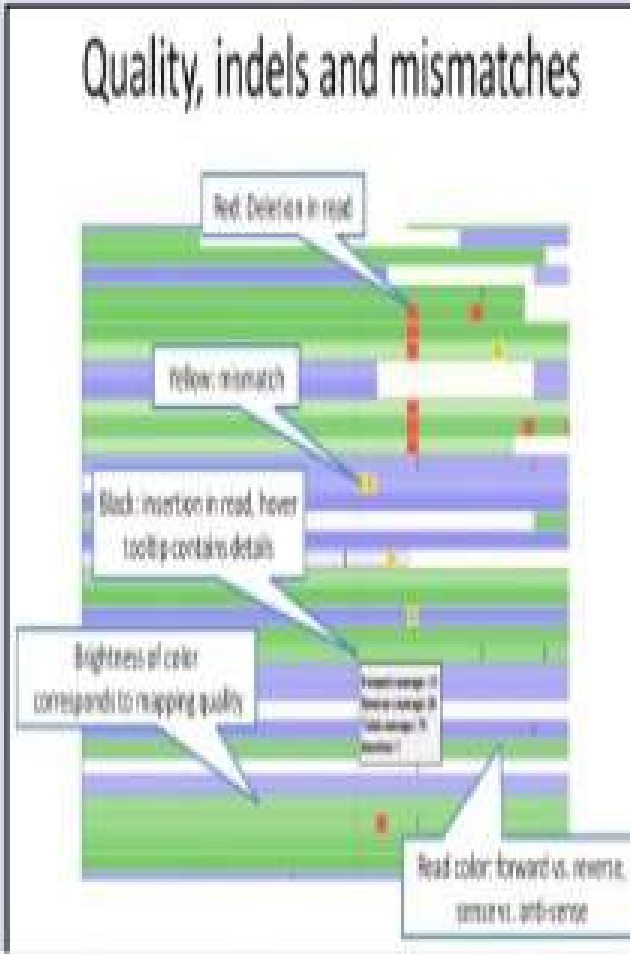


Functional annotation

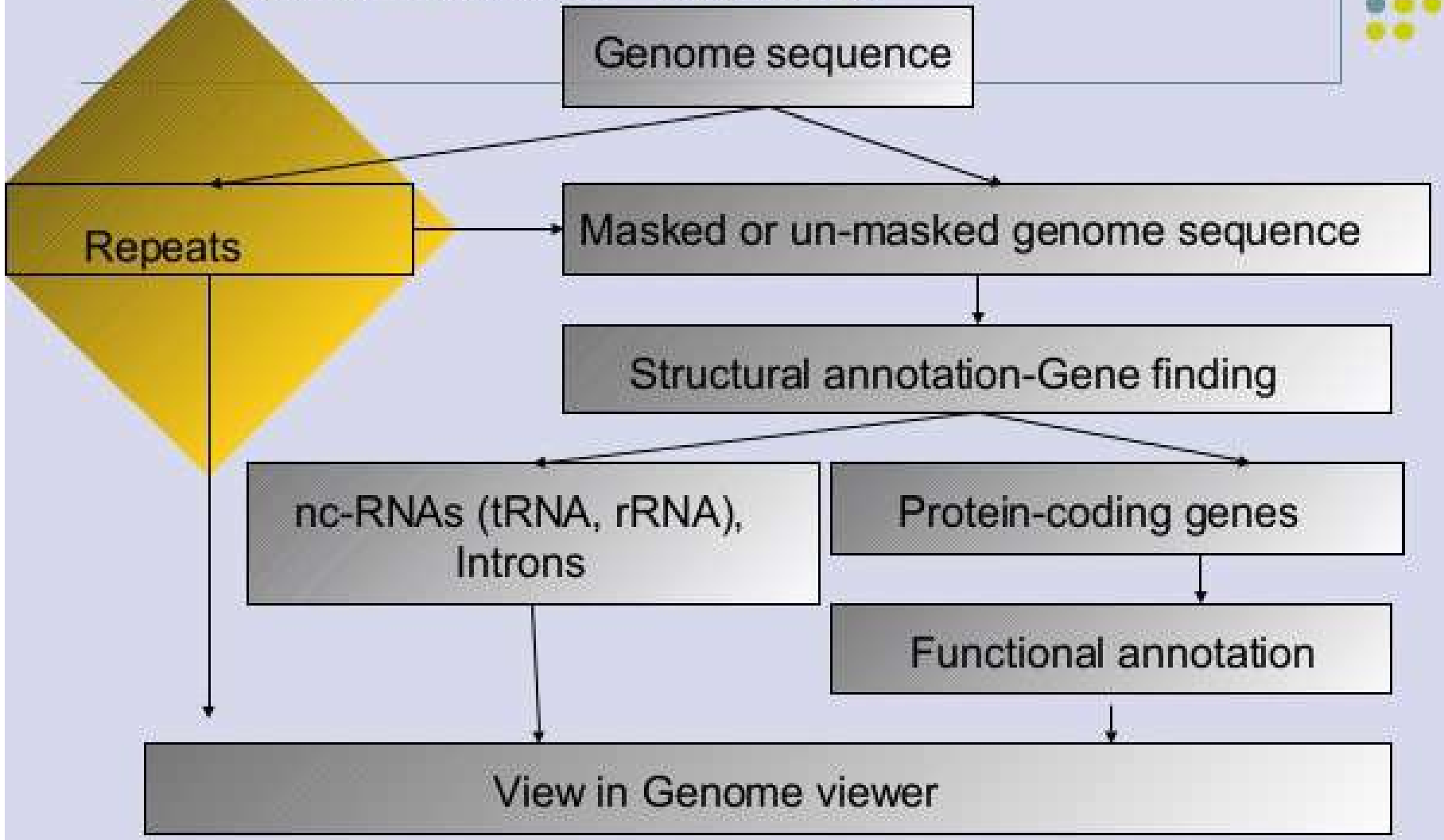
Functional annotation- attaching biological information to genomic elements

- biochemical function
- biological function
- involved regulations





Genome annotation - workflow



Things we are looking to annotate?

- CDS
- mRNA
- Promoter and Poly-A Signal
- Pseudogenes
- ncRNA

Tools

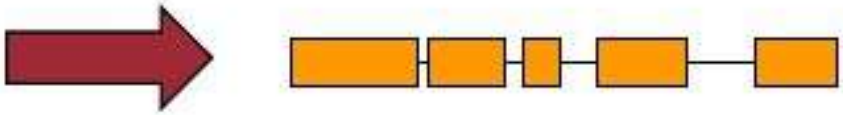
- ORF detectors
  - NCBI: http://www.ncbi.nih.gov/gorf/gorf.html
- Promoter predictors
  - CSHL: http://rulai.cshl.org/software/index1.htm
  - BDGP: fruitfly.org/seq_tools/promoter.html
  - ICG: TATA-Box predictor
- PolyA signal predictors
  - CSHL: argon.cshl.org/tabaska/polyadq_form.html
- Splice site predictors
  - BDGP: http://www.fruitfly.org/seq_tools/splice.html
- Start-/stop-codon identifiers
  - DNALC: Translator/ORF-Finder
  - BCM: Searchlauncher

What is gene prediction?

Detecting meaningful signals in uncharacterised DNA sequences.

Knowledge of the interesting information in DNA.

```
GATCGGTCGAGCGTAAGCTAGCTAG
ATCGATGATCGATCGGCCATATATC
ACTAGAGCTAGAATCGATAATCGAT
CGATATAGCTATAGCTATAGCCTAT
```



- **Gene prediction is 'recognising protein-coding regions in genomic sequence'**

Basic Gene Prediction Flow Chart

Obtain new genomic DNA sequence

↓

1. Translate in all six reading frames and compare to protein sequence databases
2. Perform database similarity search of expressed sequence tag Sites (EST) database of same organism, or cDNA sequences if available

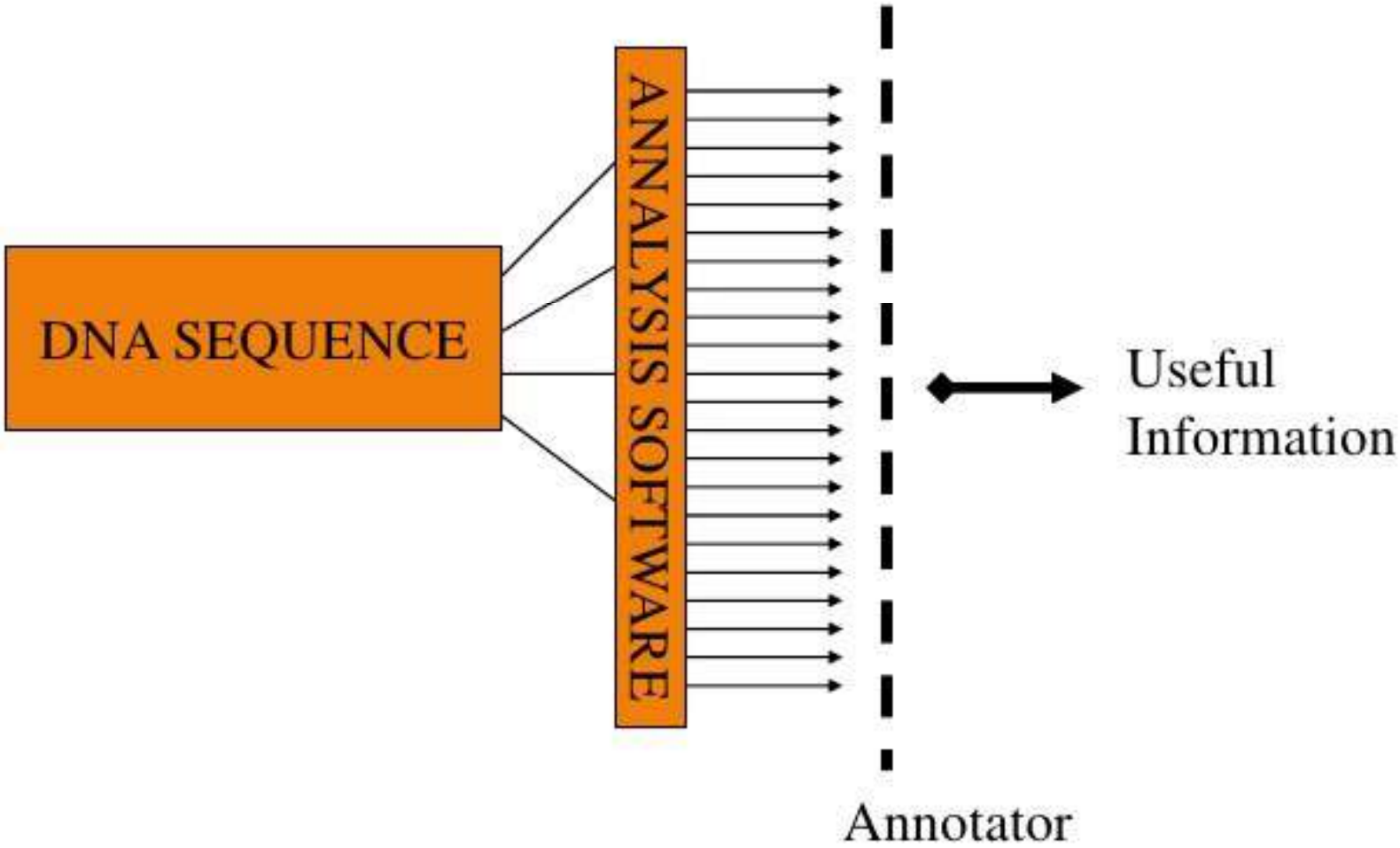
↓

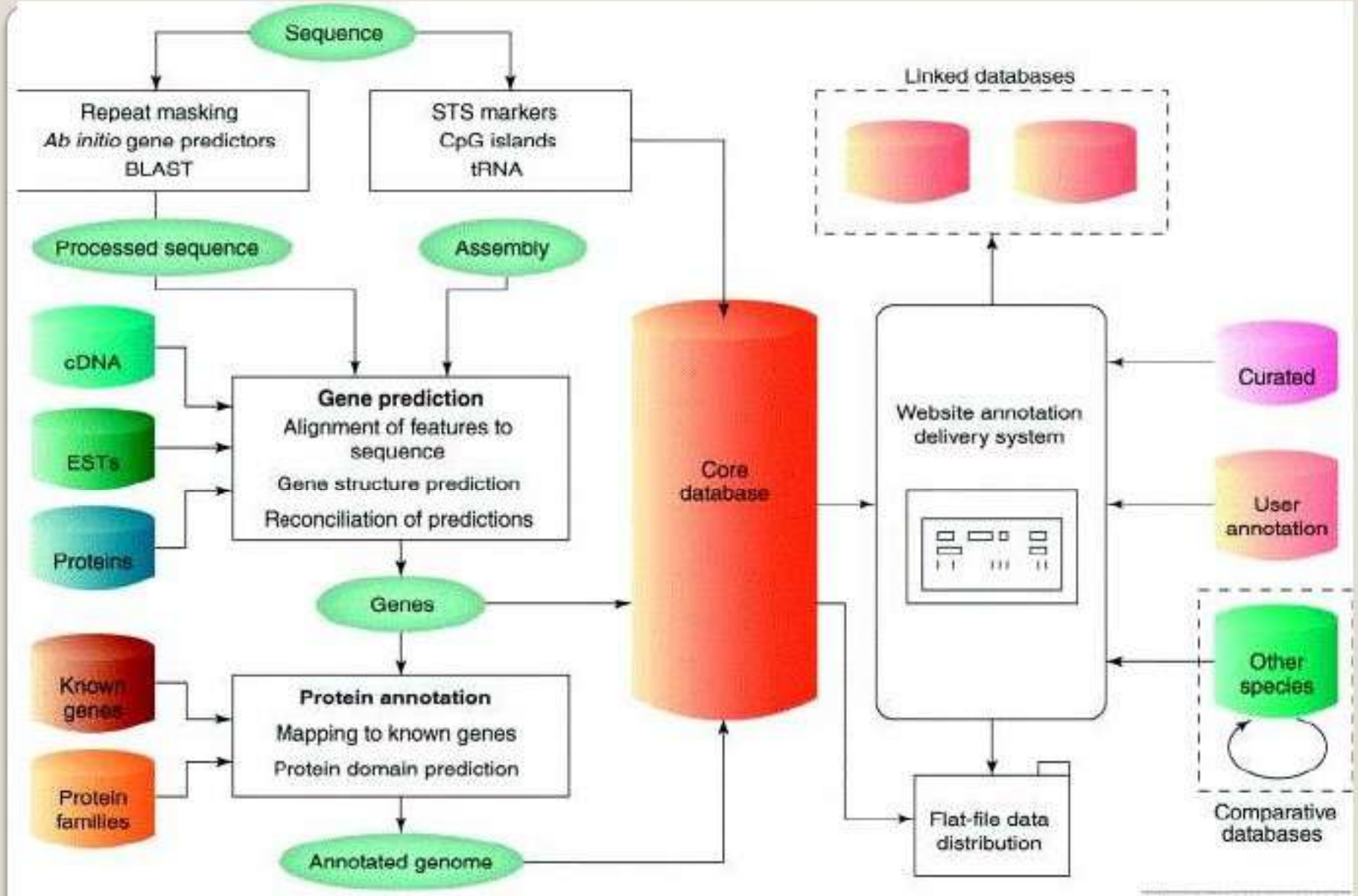
Use gene prediction program to locate genes

↓

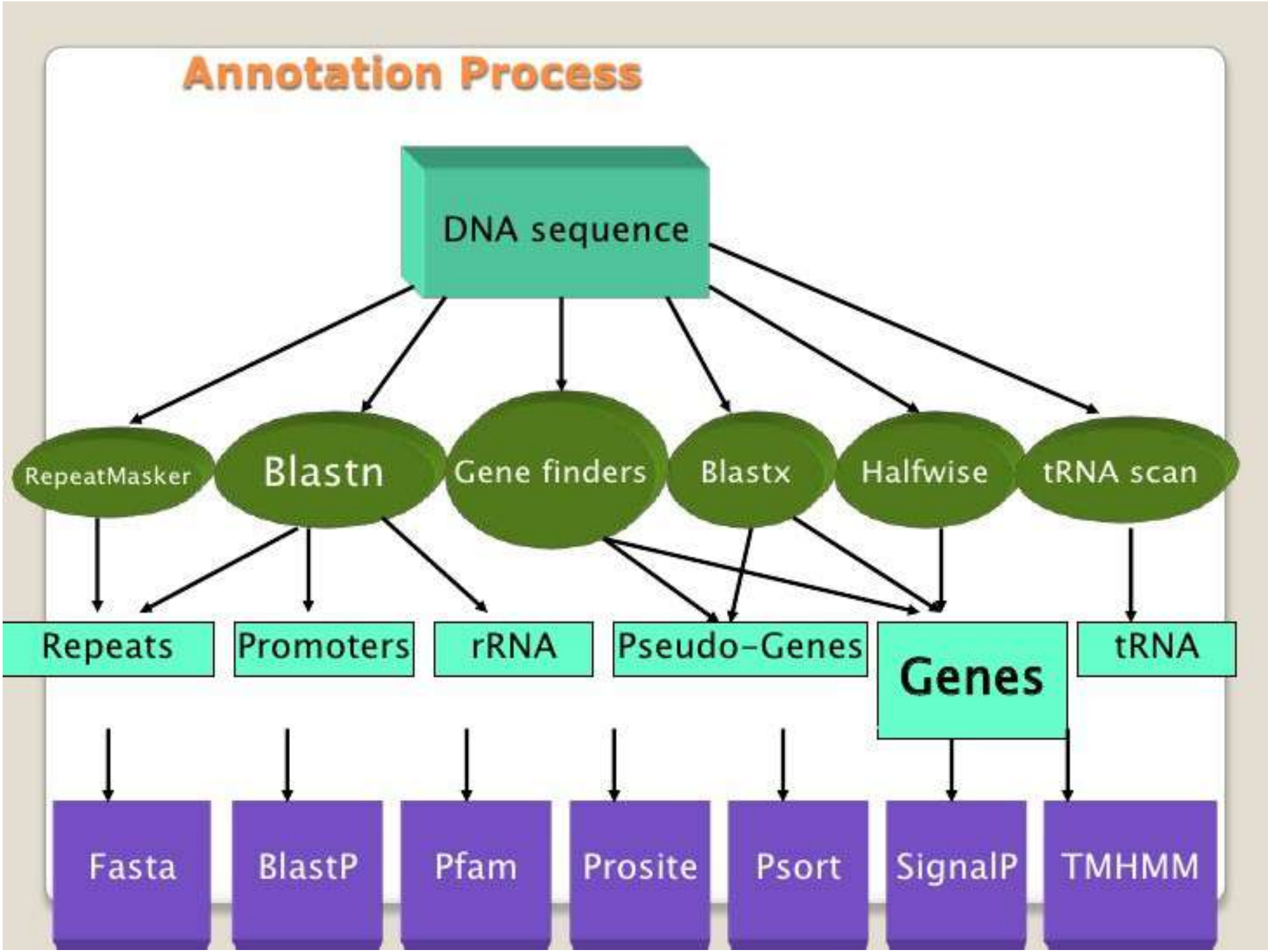
Analyze regulatory sequences in the gene

The Annotation Process





The generic structure of an automatic genome annotation pipeline and delivery system


Annotation Process
DNA sequence
RepeatMasker
Blastn
Gene finders
Blastx
Halfwise
tRNA scan
Repeats
Promoters
rRNA
Pseudo-Genes
Genes
tRNA
Fasta
BlastP
Pfam
Prosite
Psort
SignalP
TMHMM

Programme: M.Sc., Zoology
Course: HC 4.1 Advanced Genetics and Computational Biology

Gene families and clusters

By
Dr. Nijagal B.S.
Assistant professor
P.G. Dept of Zoology
J.S.S. College of Arts, Commerce and Science.
Ooty Road, Mysore-25

GENE FAMILY

- *A gene family is a set of several similar genes, formed by duplication of a single original gene, and generally with similar biochemical functions.*
- *A gene family is a set of homologous genes within one organism.*

- *When a gene is present in two or more copies per genome, the condition is known as " redundancy".*
- *The members of a gene family may be either clustered together , dispersed on different chromosomes or present in a combination of both.*

- *If the genes of a gene family encode proteins, the term "protein family" is often used in an analogous manner to gene family.*
- *One example for such family are the genes for Human haemoglobin subunits.*

GENE CLUSTER

- *A gene cluster is part of a gene family.*
- *A gene cluster is a group of two or more genes found within an organism's DNA that encode for similar polypeptides or proteins, which collectively share a generalised function and are located within a few thousand base pairs of each other.*

- *The size of gene clusters can vary significantly, from a few genes to several hundred genes.*
- *Genes found in a gene cluster may be observed near one another on the same chromosome or on different, but homologous chromosomes.*

- *Extensive tanden repetition of a gene normally occurs when the gene product is needed in unusually large amounts. E.g., genes for rRNA, histone genes, etc.*
- *Sometimes all the members of a gene family are functional, but often some members are nonfunctional pseudogenes.*

TANDEM REPEAT

- In a tandem repeat, the nucleotide sequence is repeated in the same orientation.
- For example, the trinucleotide sequence GAA is repeated two times in the DNA segment –GAAGAA-.

GAA | GAA
CTT | CTT

tandem repeat

GLOBIN GENES

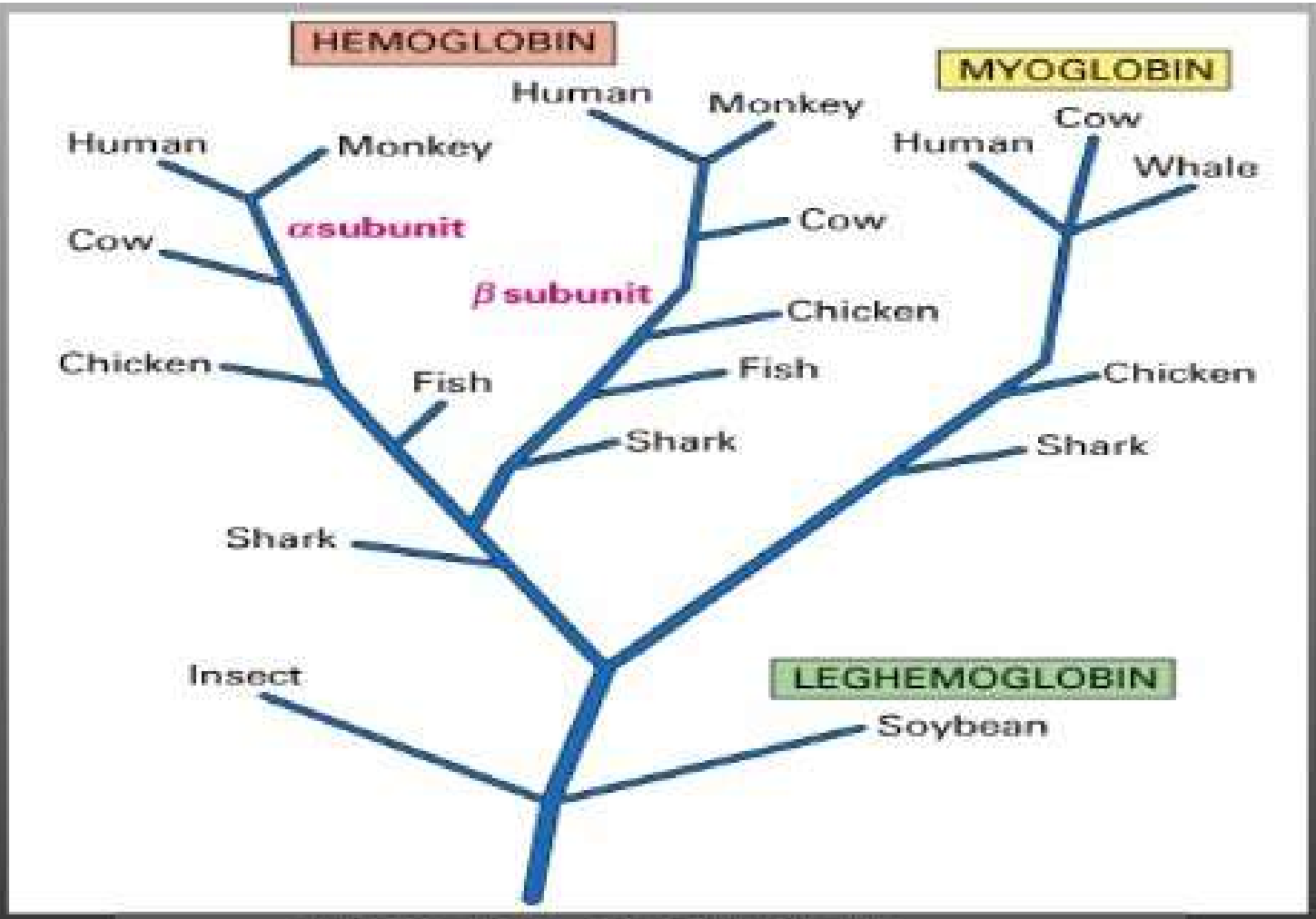
- *Genes encoding the various globin proteins evolved from one common ancestral globin gene, which duplicated and diverged about 450-500 million years ago.*
- *After the duplication events, differences between the genes in globin family arose from the accumulation of mutations.*

ANCESTRAL GLOBIN GENE

Haemoglobin genes

Myoglobin genes

Plant globin genes



ANCESTRAL GLOBIN GENE

REPRESENTATION OF EVOLUTION OF ANCESTRAL GLOBIN GENE

HAEMOGLOBIN GENES

- The haemoglobin molecule is a tetramer and is composed of two similar polypeptides, the alpha and beta chains, encoded by two distinct genes.
- Each polypeptide incorporates a Haeme group, that reversibly binds oxygen.

- *The genes are co-ordinatedly turned on and turned off during the embryonic, foetal and adult stages of development.*
- *The genes for α - globin lie in a cluster on chromosome 16, while those for β- globin are located on chromosome 11.*

- The β - cluster extends over 50 kb and has five functional genes(E , $G\gamma$,$A\gamma$,δ , β) and one pseudogene ($\Psi\beta$) .
- The α - cluster is smaller, extends over ~20 kb and has four functional genes ($\xi 2$, $\xi 1$, $\alpha 2$,$\alpha 1$, and θ) and two pseudogenes ($\Psi\alpha$,$\Psi\alpha$) .

- The two γ chains, viz., $G\gamma$ and $A\gamma$, differ for a single amino acid i.e; glycine and alanine.
- The two α genes, namely $\alpha 1, \alpha 2$, code for the same protein; such identical genes present in the same chromosome constitute "non allelic copies" of the gene.

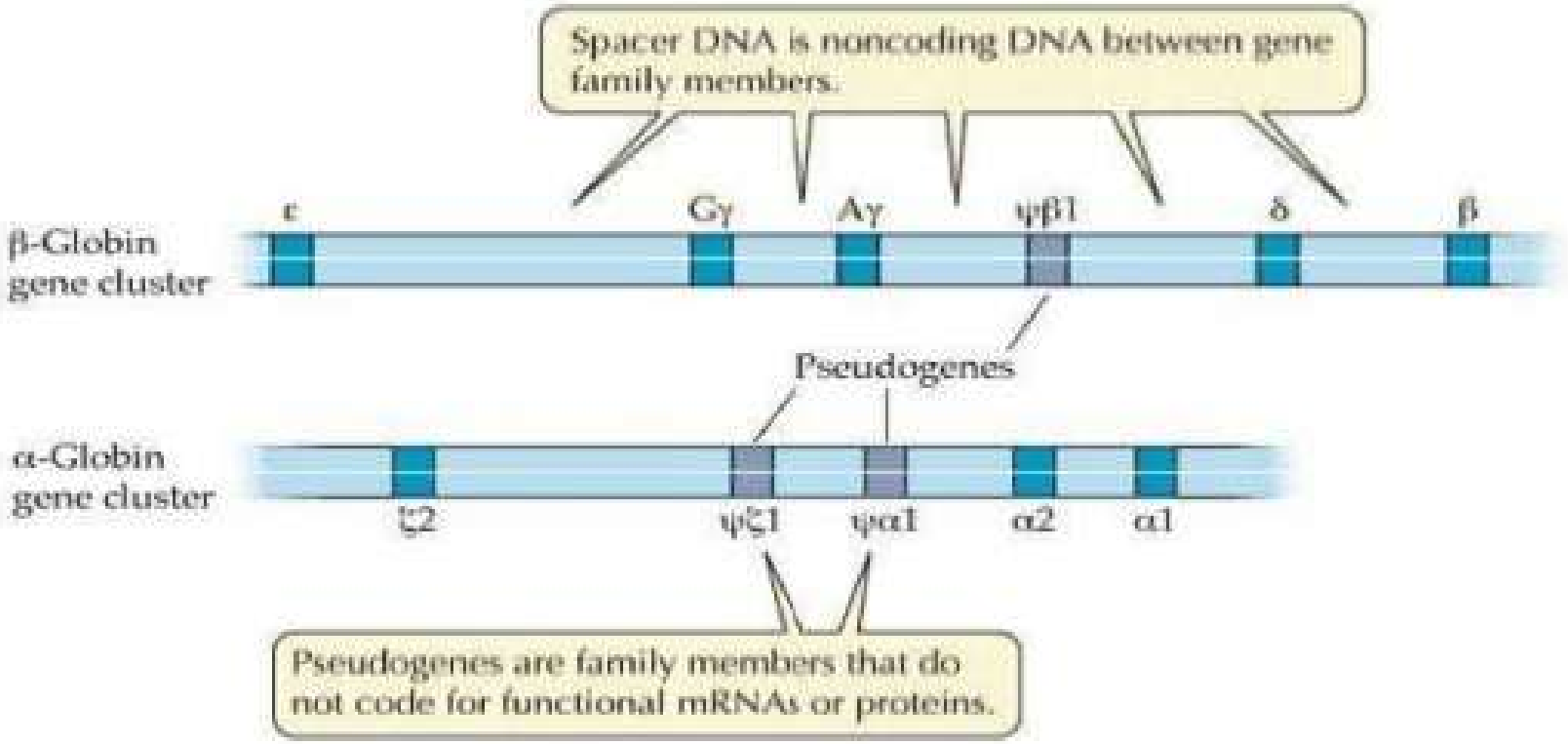
VARIOUS GLOBIN GENES EXPRESSED DURING THE EMBRYONIC, FOETAL AND ADULT STAGES OF DEVELOPMENT :-

1. α -globins

- ξ genes are expressed during the embryonic development.
- $\alpha 2, \alpha 1$ genes are expressed during the foetal and the adult stages of the development.

2. β - globins

- Epsilon genes (E) are expressed during the embryonic development.
- $G\gamma, A\gamma$ genes are expressed during the foetal development.
- δ, β genes are expressed during the adult stages of development.



HUMAN HAEMOGLOBIN GENE CLUSTERS , WITH THE ALPHA AND BETA CHAINS.