Genes & Genome studies along with Applications in Biomolecular Diseases

BY

APARNA PATIL KOSE

LECTURER

DEPT OF BIOINFORMATICS

CONTENTS

- **Genome & Its Databases:**

1. Genes
2. Introns & Prediction Tools
3. Exons & Prediction Tools
4. ORF & Prediction Tools
5. Promoters & Prediction Tools (In Prokaryotes & Eukaryotes)
6. Splice Sites & Prediction Tools (In Prokaryotes & Eukaryotes)
7. Regulatory Regions & Prediction Tools (In Prokaryotes & Eukaryotes)
8. Prediction algorithm In Prokaryotes & Eukaryotes Genomes
9. Synteny & Gene Order

INTRODUCTION

GENES:

Gene is the basic physical and functional unit of heredity.

Genes are made up of DNA.

Types of Gene: A,T/U,G, C

Sequence within the nucleic acid ➔ represents a single protein.

Gene may exist in alternative forms ➔ alleles.

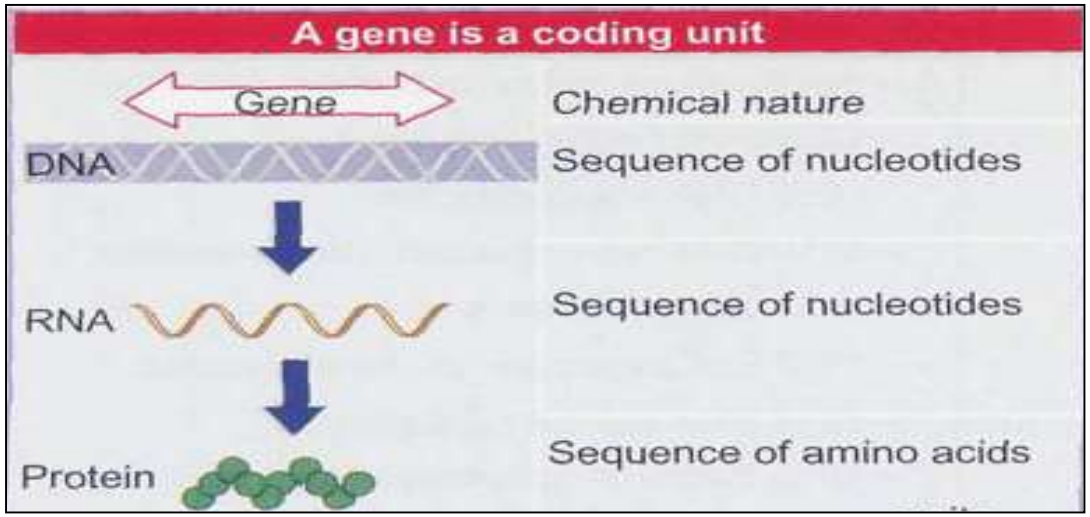
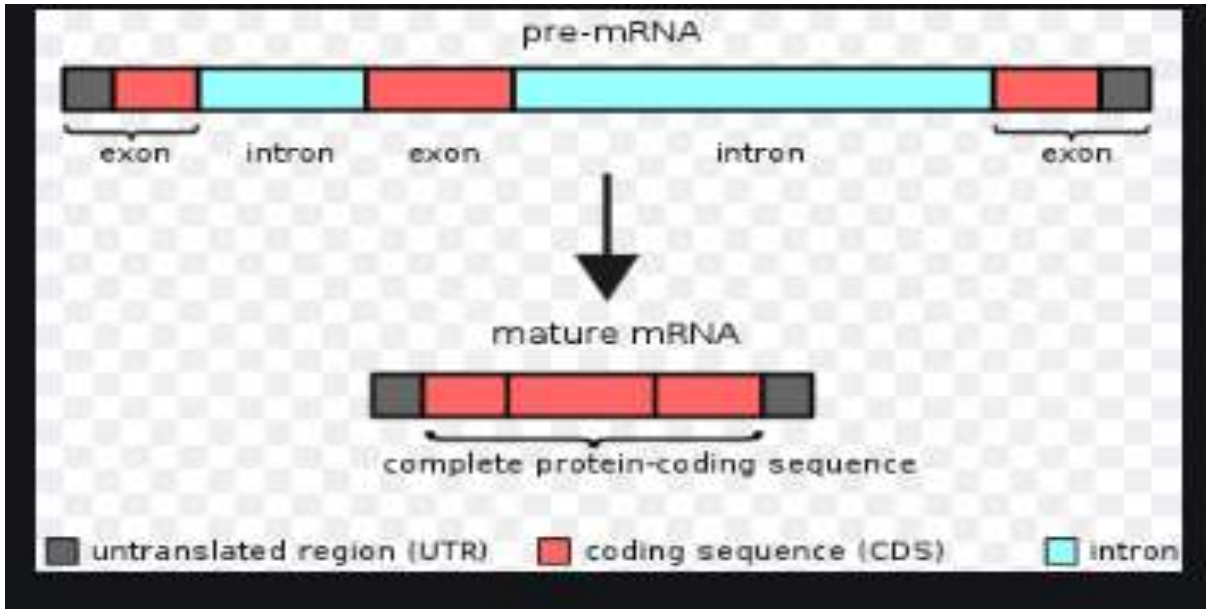


Figure 1: A gene codes for RNA which may codes for protein

INTRONS

An **intron** (for intragenic region) is **any nucleotide sequence** within a gene that is **removed by RNA splicing** during **maturation** of the **final RNA product.**

In other words, **introns** are **non-coding regions** of an RNA transcript, or the DNA encoding it, that are eliminated by **splicing before translation.**



CONT'D

- **Functions:** Transcription termination, Genome organization, Transcription initiation, etc.

- **Advantage:** Evolutionary **advantages** of **introns** include the possibility to **create new genes** by **cutting and pasting exons** from **existing genes** or to **diversify the protein output** of a single gene by **splicing the exons** together in different ways.

- What happens if introns are not removed?

Not only do the **introns not** carry information to build a protein, they actually have to be **removed** in order for the mRNA to encode a protein with the right sequence.

If the spliceosome fails to **remove** an **intron**, an mRNA with extra "junk" in it will be made, **and a wrong protein will get produced during translation.**

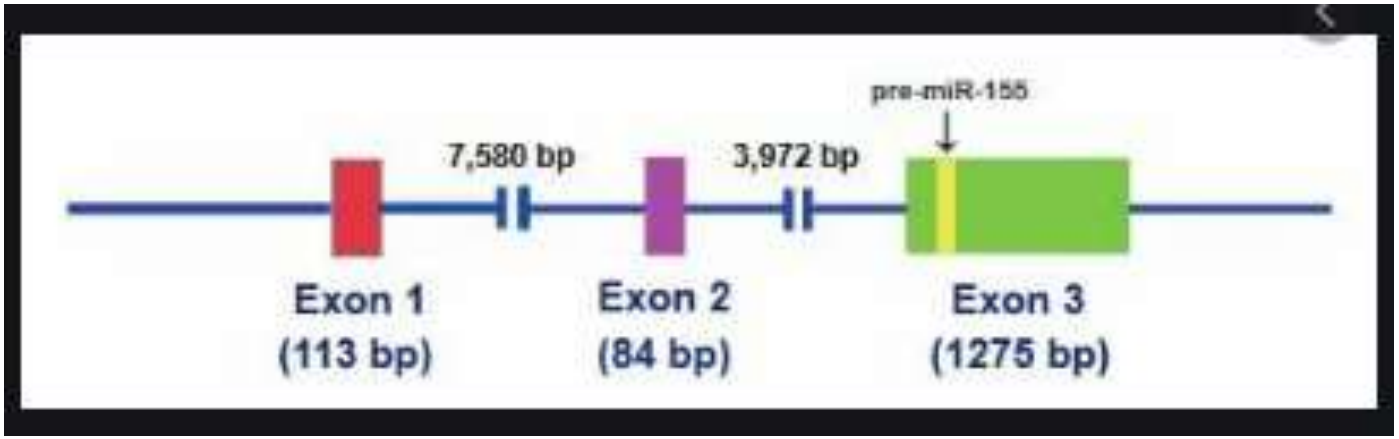
- **TOOLS:** When you search for a protein-coding exons, the software also gives to you the introns sequences. Eg: Genomescan

EXONS

- An **exon** is any part of a **gene** that will encode a part of the **final mature RNA** produced by that gene after introns have been removed by RNA splicing.
- The term **exon** refers to both the DNA sequence within a gene and to the corresponding sequence in RNA transcripts.
- **How many exons does a gene have?** **8.8 exons**

On average, there are **8.8 exons** and **7.8** introns per gene. About 80% of the exons on each chromosome are < 200 bp in length.

- **Function: exon** is a coding region of a gene that contains the information required to encode a protein.



➢TOOLS:

1. Genomescan
2. Geneid
3. GrailEXP

ETC……..

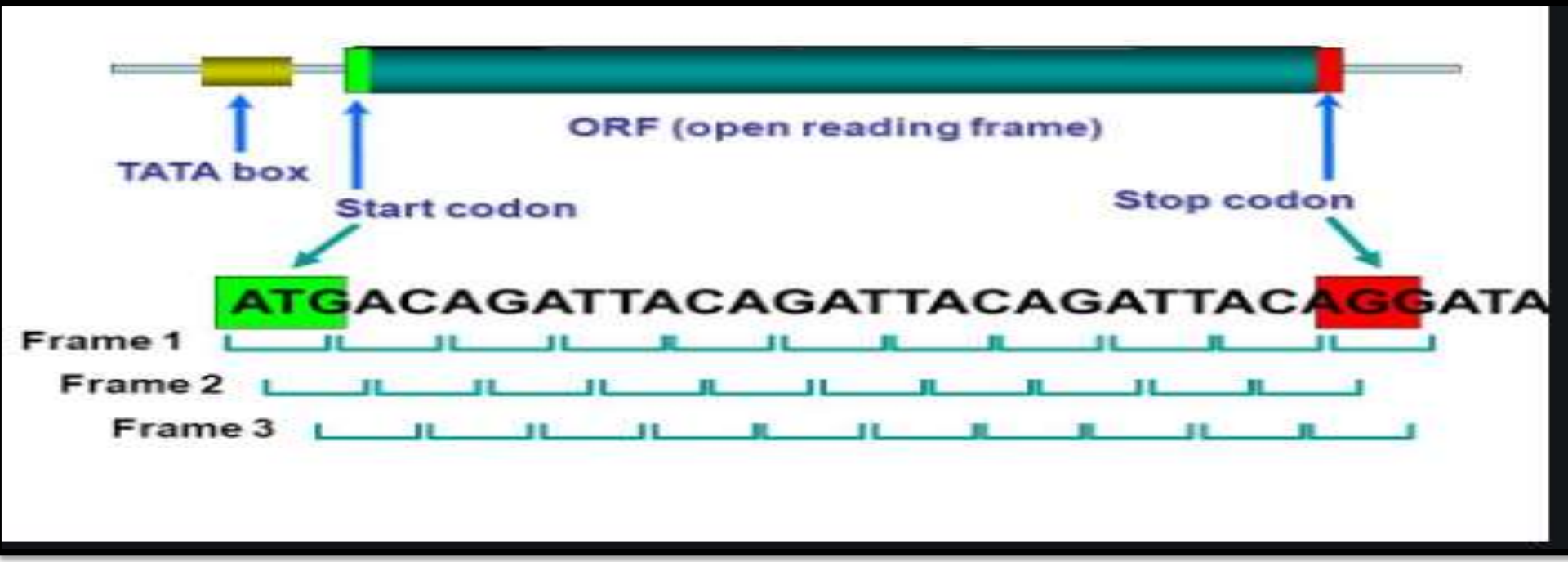
OPEN READING FRAME (ORF)

- An **open reading frame** is a portion of a DNA molecule that, when translated into amino acids, contains no stop codons.

- A long **open reading frame** is likely part of a **reading frame** that has the ability to be translated.
- An ORF is a continuous stretch of codons that may begin with a **start codon (usually AUG)** and ends at a **stop codon (usually UAA, UAG or UGA)**
- Why are open reading frames important?

ORFs is as one piece of evidence to assist in gene prediction.
Long ORFs are often used, along with other evidence, to initially identify candidate protein-coding regions or functional RNA-coding regions in a DNA sequence.

Tools: NCBI- ORF FINDER (https://www.ncbi.nlm.nih.gov/orffinder/)



IDENTIFYING ORFS

- Simple 1st step in gene findings.
- Translate genomic sequence in six frames.
- Identify stop codon in each frame.
- Regions without stop codons are called "open reading frames" or ORFs.
- Locate and tag all of the likely ORFs in a sequence.
- The longest ORF from a methionine codon is a good prediction of a protein encoding sequence.

PROMOTERS

- The DNA opens up in the **promoter** region so that RNA polymerase can begin transcription.
- Each gene (or, in bacteria, each group of genes transcribed together) has its own **promoter**.
- A **promoter** contains DNA sequences that let RNA polymerase or its helper proteins attach to the DNA

i.e. A DNA sequence that the transcription apparatus recognizes and binds and determines the transcription start site, the first nucleotide that will be transcribed into RNA.

- **Functions**:
  A **promoter** is a sequence of DNA needed to turn a gene on or off. The process of transcription is initiated at the **promoter**.

Usually found near the beginning of a gene, the **promoter** has a binding site for the enzyme used to make a messenger RNA (mRNA) molecule.

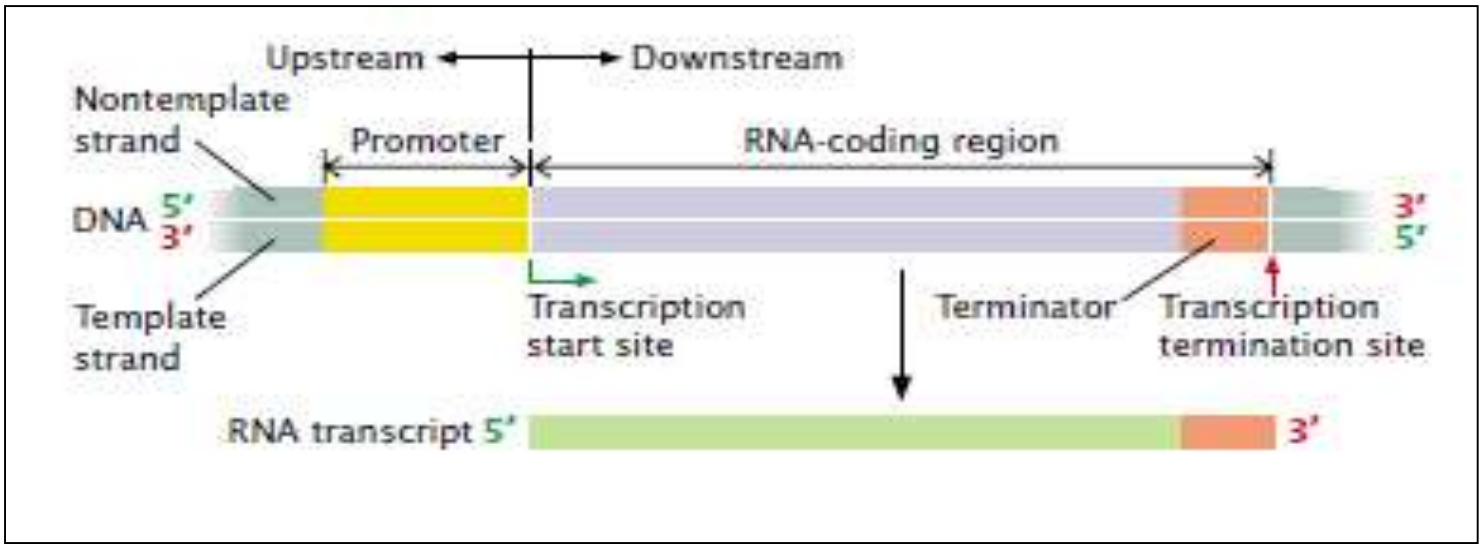


Figure 2: A transcription unit includes a promoter, an RNA-coding region, and a terminator.

➢TOOLS: (Eukaryotic as well as Prokaryotic Tools)

1. FPROM
2. TSSP
3. TSSW
4. TSSG
5. BpROM

ETC……..

SPLICE SITES

- A genetic alteration in the **DNA** sequence that occurs at the boundary of an exon and an intron (**splice site**).
- Splice sites are the sequences immediately surrounding the exon intron boundaries.
- Found: These **sites** are found at the 5′ and 3′ ends of introns. Most commonly, the RNA sequence that is removed begins with the **dinucleotide GU at its 5′ end**, and ends with **AG at its 3′ end.**

i.e. The GU-AG rule (originally called the GT-AG rule in terms of DNA sequence).

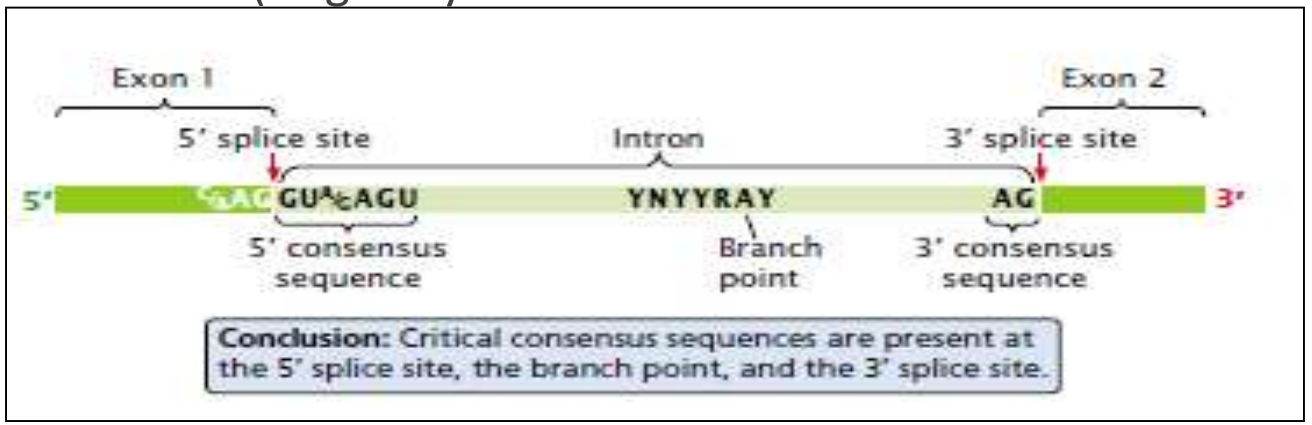
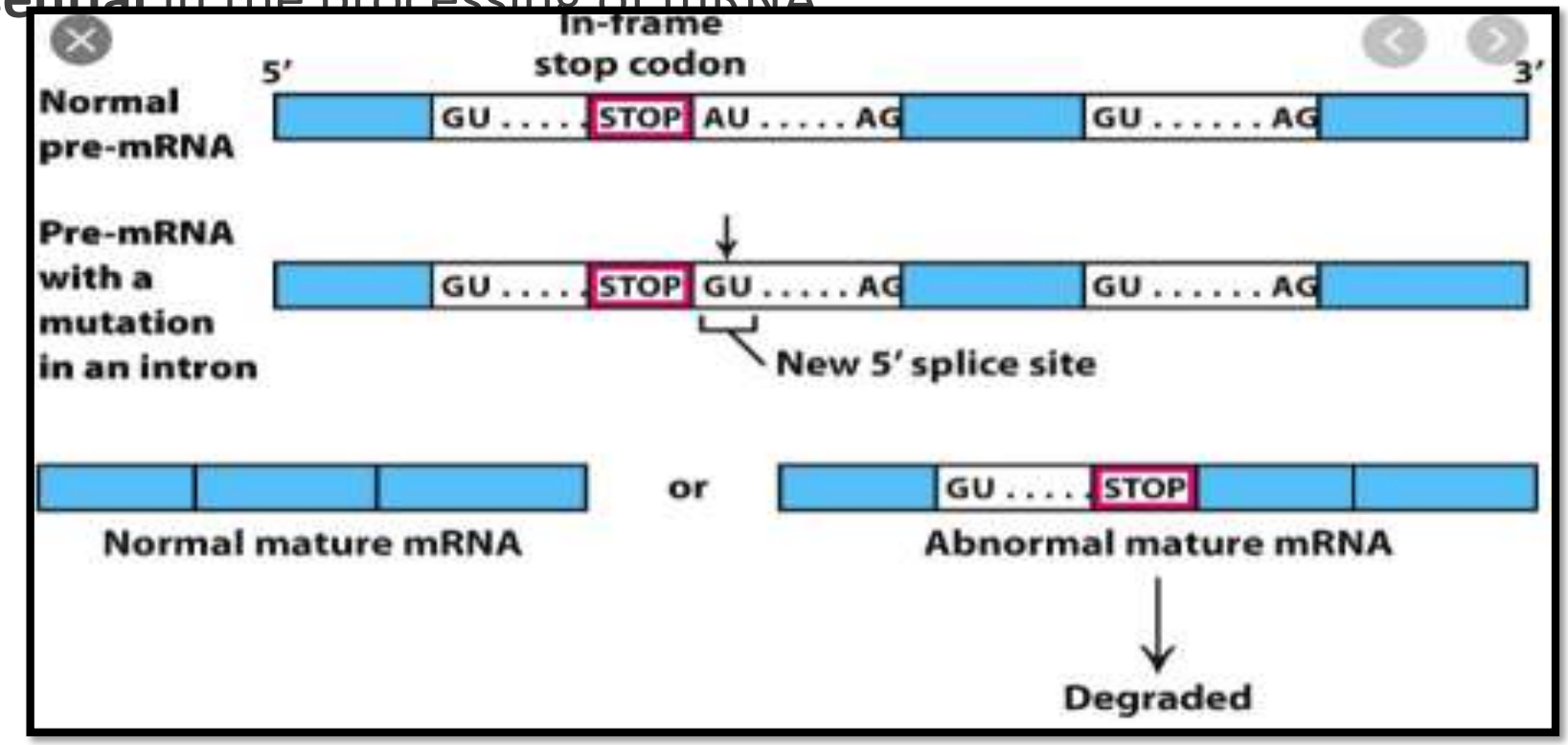


Figure 3: The ends of nuclear introns are defined by the GU-AG rule. In the consensus sequence surrounding the branch point (YNYYRAY) Y is any pyrimidine, R is any purine, A is adenine, and N is any base.

Why Splice sites are important

Mutations in these sequences may lead to retention of large segments of intronic DNA by the mRNA, or to entire exons being **spliced** out of the mRNA. These changes could result in production of a nonfunctional protein. ... These donor **sites**, or recognition **sites, are essential** in the processing of mRNA



Why Splice sites prediction are important

- **Prediction** of **splice sites** where accurate localization of **splice sites** can substantially help explore the structure of genes.

- Accurate **prediction** of **splice sites** can setup the boundaries of exons which is critical in alternative **splicing prediction**

➢TOOLS: (Eukaryotic as well as Prokaryotic Tools. ALSO BASED ON VARIOUS APPROACH/ ALGORITHM)

1. Human Splicing Finder
2. GeneSplicer
3. FGENES
4. Fgenesh-M
5. FGENESH_GC

ETC……..

REGULATORY REGIONS

- A regulatory sequence is a segment of a nucleic acid molecule which is capable of increasing or decreasing the expression of specific genes within an organism.

- An enhancer activates the nearest promoter to it.

- A UAS (upstream activator sequence) in yeast behaves like an enhancer but works only upstream of the promoter.

- Form complexes of activators that interact directly or with the promoter

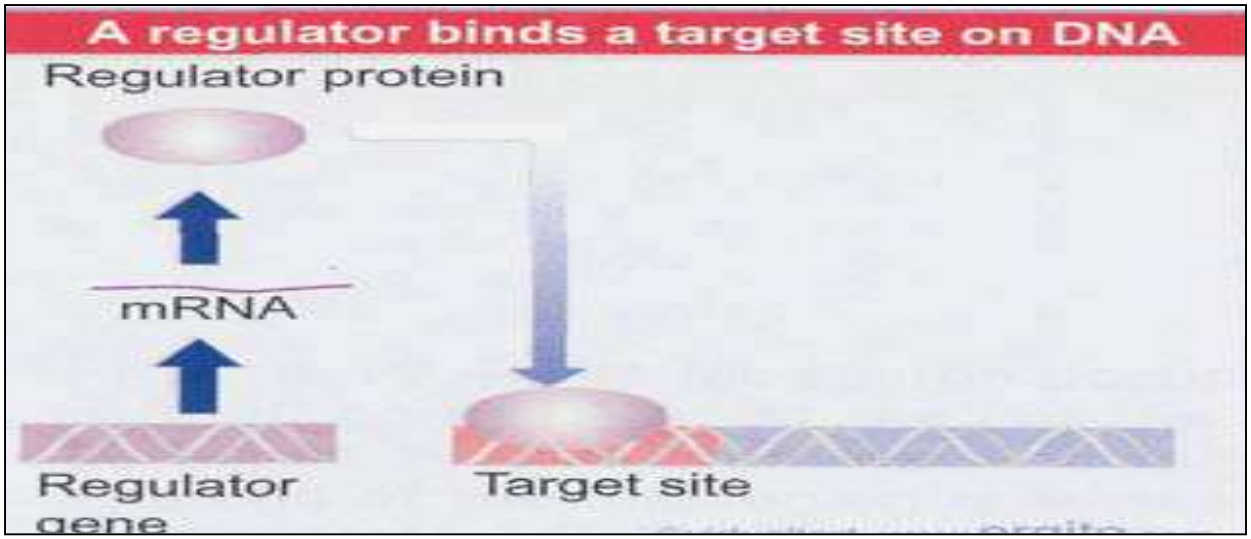


Figure 4: Regulatory model

What is the role of a regulatory region?

- **Regulatory** sequence controls when expression occurs for the multiple protein coding **regions** (red).
- Promoter, operator and enhancer **regions** (yellow) regulate the transcription of the gene into an mRNA.
- The mRNA untranslated **regions** (blue) regulate translation into the final protein products.

- **TOOLS: (Eukaryotic as well as Prokaryotic Tools. ALSO BASED ON VARIOUS APPROACH/ ALGORITHM)**

1. **TRANSFAC**
2. **RSAT Fungi/ prokaryotes/bacteria**
3. **CRÈME**
4. **RSA Tools**

ETC……..

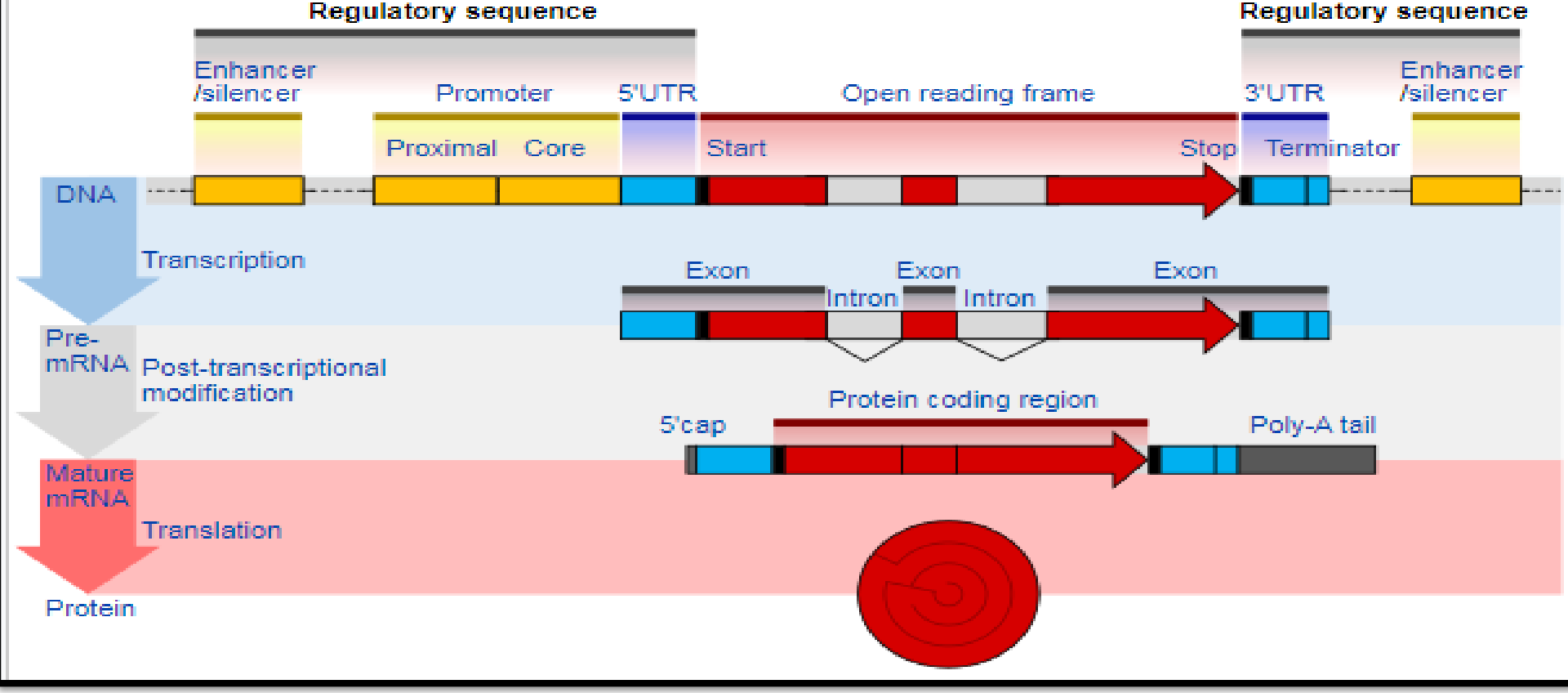


Figure 5: The structure of a eukaryotic protein-coding gene. Regulatory sequence controls when and where expression occurs for the protein coding region (red).
Promoter and enhancer regions (yellow) regulate the transcription of the gene into a pre-mRNA which is modified to remove introns (light grey) and add a 5' cap and poly-A tail (dark grey).
The mRNA 5' and 3' untranslated regions (blue) regulate translation into the final protein product

PREDICTION FOR VARIOUS SIGNALS IN GENOME IS IMPORTANT ???

Why gene prediction?

- With the rapid accumulation of genomic sequence information, there is a pressing need to use computational approaches to accurately predict gene structure.
- Computational gene prediction is a **prerequisite** for detailed functional annotation of genes and genomes.
- The process includes detection of :
  - the location of open reading frames (ORFs) and
  - Predict the structures of introns as well as exons if the genes of interest are of eukaryotic origin.

THE ULTIMATE GOAL

- To describe all the genes computationally with near 100% accuracy.
- The ability to accurately predict genes can significantly reduce the amount of experimental verification work required.

DISADVANTAGE

- For eukaryotes, because many problems in computational gene prediction are still largely unsolved.
- This is because coding regions normally do not have conserved motifs.
- The elements are diverse and not clearly defined.
- Detecting coding potential of a genomic region has to rely on subtle features associated with genes that may be very difficult to detect.
- Normally elements are short (6 to 8 nucleotide) and found in any sequence by random chance, thus high rate of false positive results because of which sensitivity drops and specificity is hampered.

Solution

For preliminary identification of these elements , we need to combine a multitude of features and use sophisticated algorithms that give either

1. ab initio–based predictions OR

2. Predictions based on evolutionary information (Homology based) OR

3. Experimental data

Types of approaches

The current gene prediction methods can be classified into two major categories,

1. AB INITIO–BASED: (prediction based on given sequence only)

- It does so by relying on two major features associated with genes:

The first feature is the existence of gene signals, which include start and stop codons, intron splice signals, transcription factor binding sites, ribosomal binding sites, and polyadenylation (poly-A) sites, the triplet codon ETC.

The second feature used by ab initio algorithms is gene content, which is statistical description of coding regions.

It has been observed that nucleotide composition and statistical patterns of the coding regions tend to vary significantly from those of the noncoding regions.

- Thus unique features can be detected by applying probabilistic models such as **Markov models** or **hidden Markov models** , **MLT** to help distinguish coding from noncoding regions.

2. HOMOLOGY BASED APPROACHES:

- Predictions based on significant matches of the query sequence with sequences of known genes.

Eg: if a **translated DNA sequence** is found to be **similar** to a **known protein or protein family** from a database search, this can be strong evidence that the **region codes for a protein**

- **Also if** possible **exons** of a **genomic DNA region** match a **sequenced cDNA**, this also provides **experimental evidence** for the **existence of a coding region.**

Therefore,

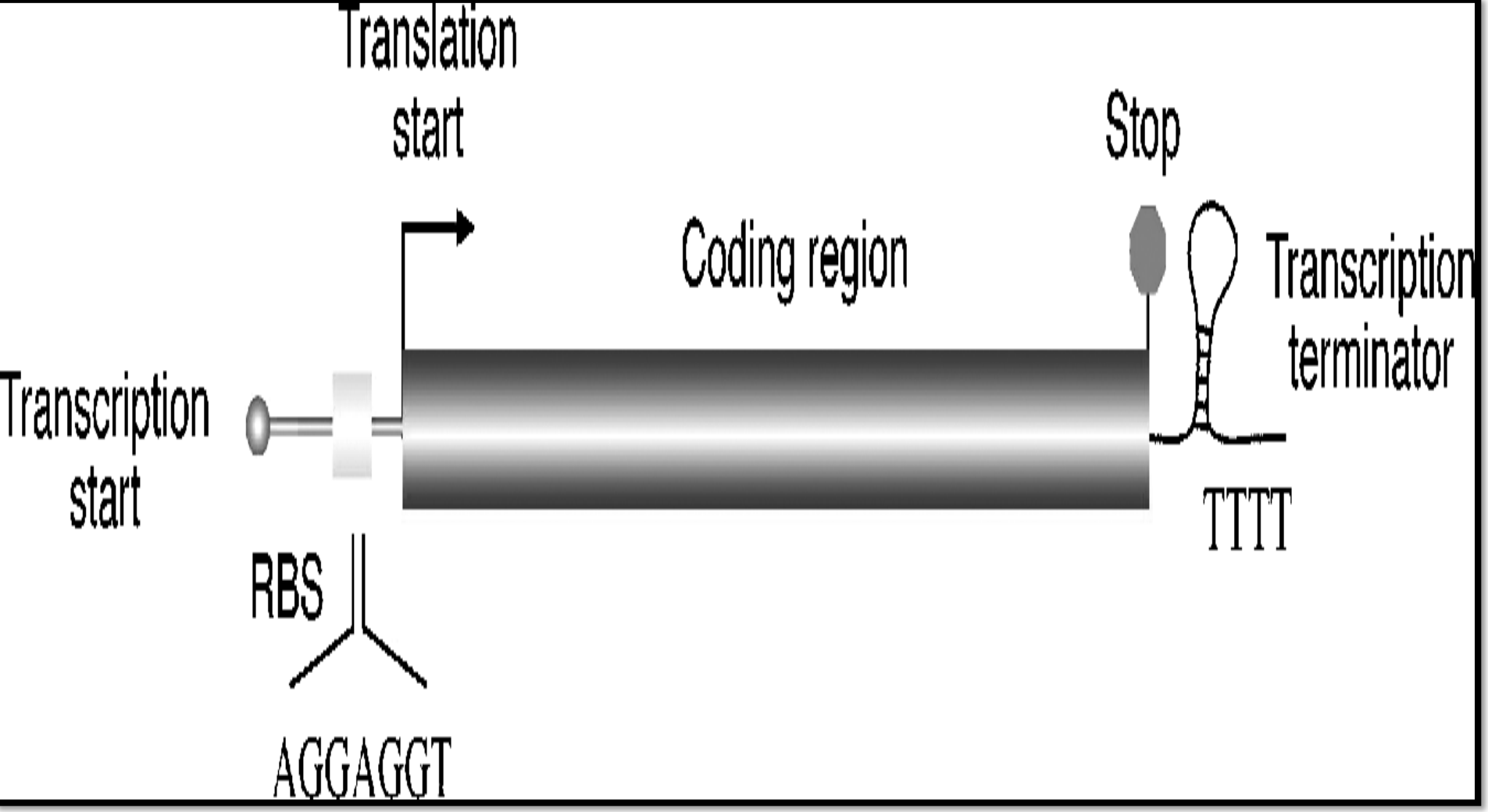
There are also a number of programs that actually combine prediction results from multiple individual programs to derive a consensus prediction.

This type of algorithms can therefore be considered as consensus based.

GENE PREDICTION IN PROKARYOTES

- Prokaryotes, which include bacteria and Archaea, have relatively small genomes with sizes ranging from 0.5 to 10Mbp (**1 Mbp = 10^6 bp**)
- The gene density in the genomes is high, with more than 90% of a genome sequence containing coding sequence.
- There are very few repetitive sequences.
- Each prokaryotic gene is composed of a single contiguous stretch of ORF coding for a single protein or RNA with no interruptions within a gene.
- In bacteria, the majority of genes have a start codon ATG (or AUG in mRNA; because prediction is done at the DNA level, T is used in place of U), which codes for methionine.
- Occasionally, GTG and TTG are used as alternative start codons, but methionine is still the actual amino acid inserted at the first position.

- There may be **multiple ATG, GTG, or TGT codons in a frame,**
- But presence of these codons at the beginning of the frame does not **necessarily give a clear indication of the translation initiation site.**
- Instead, to help identify this **initiation codon**, other **features associated** with **translation are used**
- And one such feature is the **ribosomal binding site**, also called the ***Shine-Delgarno sequence,*** *which is a stretch of* ***purine-rich sequence complementary*** **to 16S rRNA in the ribosome .**
- It is located immediately **downstream** of the **transcription initiation site** and slightly **upstream** of the **translation start codon**.
- In many bacteria, it has a consensus motif of **AGGAGGT**.
- Identification of the ribosome binding site can help locate the start codon.


Translation
start
Stop
Coding region
Transcription
terminator
Transcription
start
TTTT
RBS
AGGAGGT

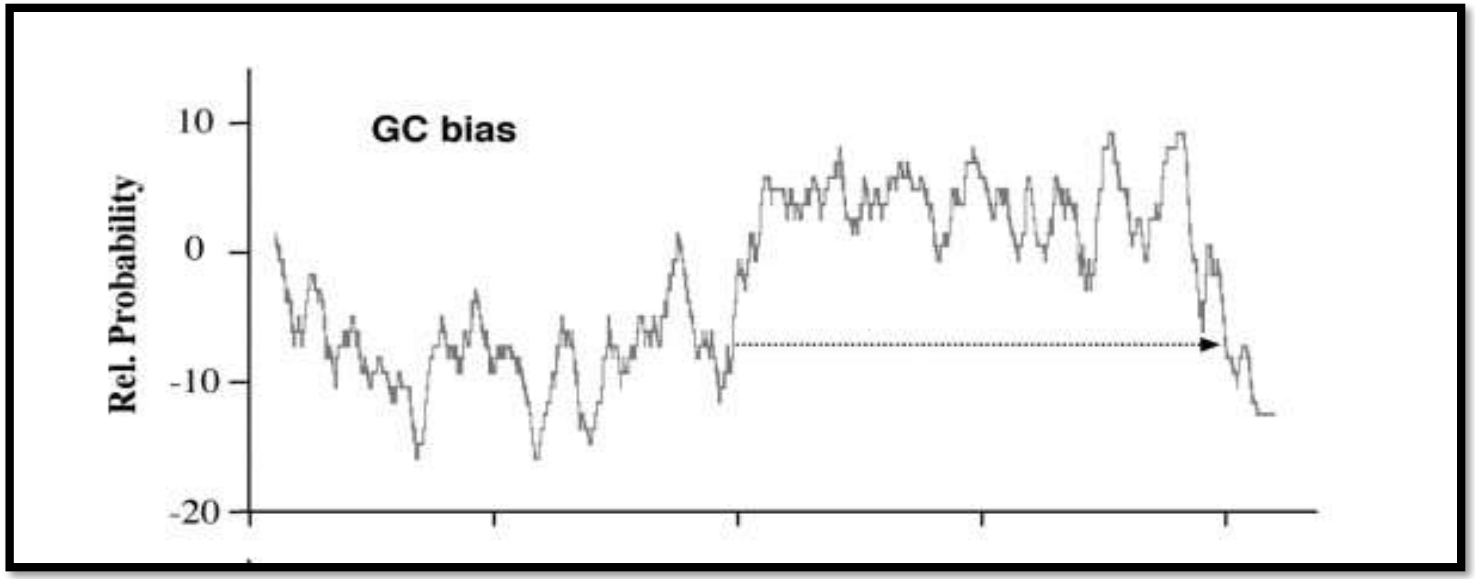
- At the end of the protein coding region is a stop codon that causes translation to stop.

- Many prokaryotic genes are transcribed together as one operon.
- The end of the operon is characterized by a transcription termination signal called *ρ-independent terminator.*
- *The terminator sequence has a distinct stem-loop secondary structure* followed by a string of Ts.
- Identification of the terminator site, in conjunction with promoter site identification can sometimes help in gene prediction.

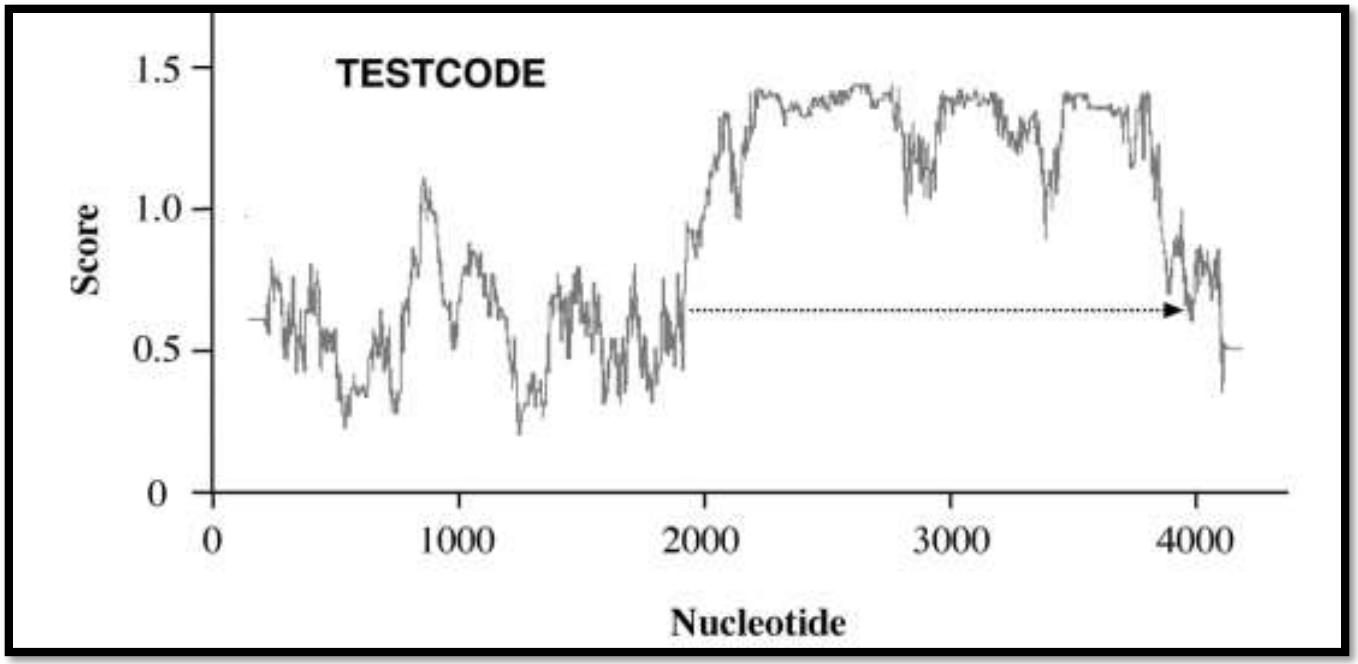
Conventional Determination method of ORF

- Without use of specialized programs, **prokaryotic gene identification** can rely on **manual determination of ORFs** and **major signals** related to **prokaryotic genes.**
- Prokaryotic DNA is first subject to **conceptual translation** in all **six possible frames**, **three frames forward** and **three frames reverse**.
- Because a **stop codon** occurs in about every **20 codons** by **chance** in a noncoding region, a **frame longer than thirty codons without interruption by stop codons** is suggestive of a **gene coding region**.
- Although the **threshold for an ORF** is **normally** set even **higher at fifty or sixty codons**.
- The **putative frame** is further manually confirmed by the presence of other signals such as a **start codon and Shine–Delgarno sequence**.
- Furthermore, the **putative ORF** can be **translated** into a **protein sequence**, which is then used to **search against a protein database**.
- Detection of **homolog**s from this search is probably the **strongest indicator of a protein-coding frame.**

- In the **early stages** of development of gene prediction algorithms, **genes** were predicted by examining the **non randomness of nucleotide distribution**.
- One method is based on the **nucleotide composition** of the **third position** of a codon. Thus , in coding sequence, it has been observed that this position has a **preference to use G or C over A or T**.
- By plotting the **GC composition at this position**, regions with values **significantly above the random level** can be identified, which are **indicative of the presence of ORFs.**
- As genes can be in any of the **six frames**, the **statistical patterns** are computed for **all possible frames**.



- In addition to **codon bias**, there is a similar method called **TESTCODE** that exploits the fact that the **third codon nucleotides** in a coding region tend to **repeat themselves**.
- By plotting the **repeating patterns of the nucleotides** at this position, **coding and noncoding regions can be differentiated.**
- The results of the two methods are often **consistent**.
- The two methods are often used in **conjunction** to confirm the results of each other.



- Thus, **statistical methods**, which are based on **empirical rules**, examine the **statistics** of a **single nucleotide** (either G or C).
- They identify only **typical genes** and **tend to miss atypical genes** in which the **rule of codon bias is not strictly followed**.
- To improve the prediction accuracies, the **new generation of prediction algorithms use more sophisticated statistical models.**

Gene Prediction Using Markov Models and Hidden Markov Models

- Markov models and HMMs can be very helpful in providing finer statistical description of a gene.

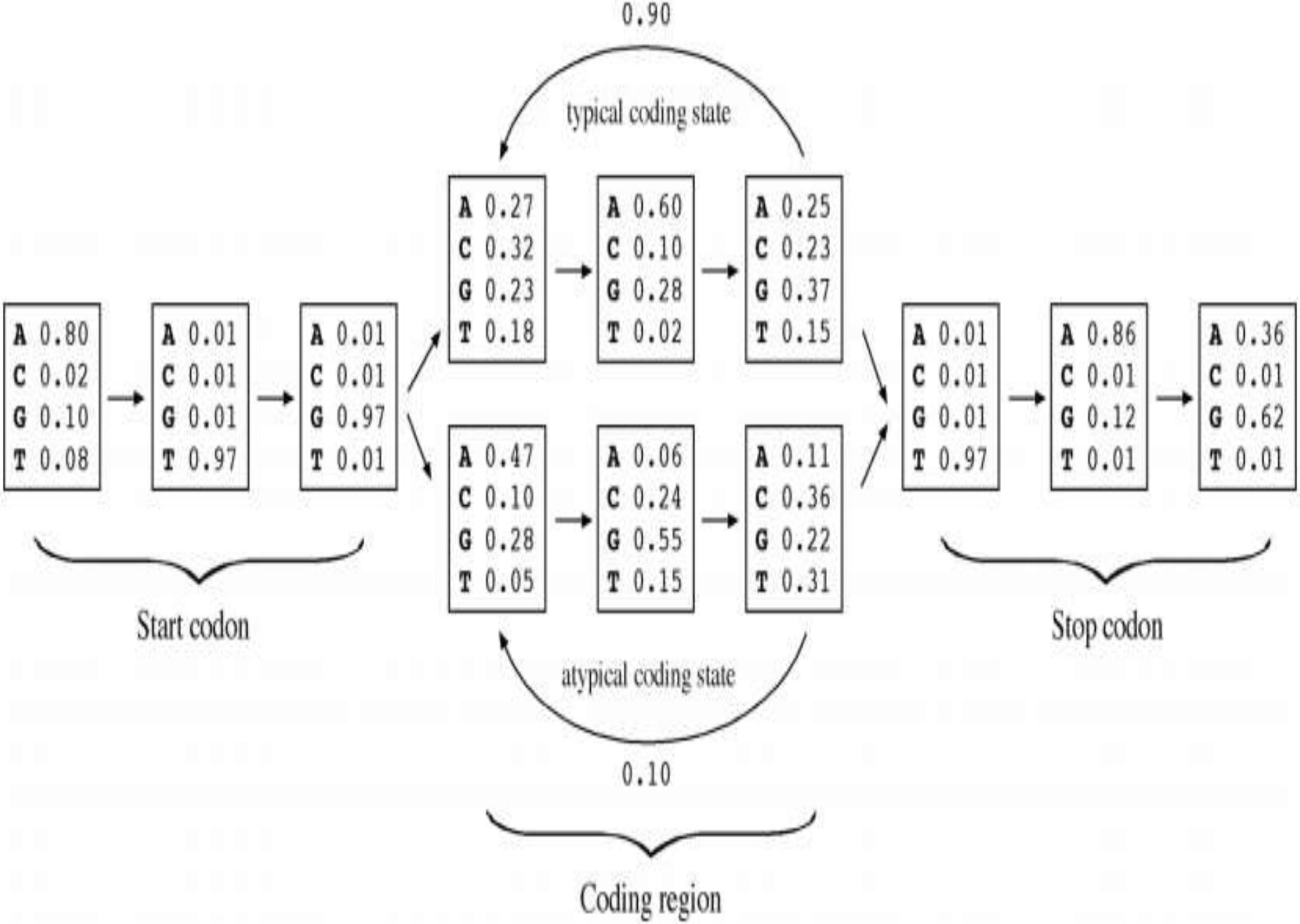
- A Markov model describes the **probability** of the **distribution of nucleotides** in a DNA sequence, in which the **conditional probability** of a **particular sequence position depends on k previous positions**.
- In this case, **k is the order of a Markov model**.
- A **zero-order Markov model** assumes **each base** occurs **independently** with a **given probability**.This is often the case for noncoding sequences.
- A **first-order Markov model** assumes that the **occurrence of a base** depends on the **base preceding it**.
- A **second-order model** looks at the preceding **two bases** to determine which base follows, which is more **characteristic of codons in a coding sequence**.

- The use of **Markov models** in gene finding exploits the fact that **oligonucleotide distributions in the coding regions are different from those for the noncoding regions**.
- These can be represented with various orders of **Markov models**.

- Since a **fixed-order Markov chain** describes the **probability of a particular nucleotide** that depends on **previous k nucleotides**, the **longer the oligomer unit, the more non randomness** can be described for the **coding region.**
- Therefore, the **higher the order of a Markov model**, the **more accurately** it can predict a gene. Because a protein-encoding gene is composed of **nucleotides in triplets as codons**, more effective **Markov models** are built in sets of **three nucleotides**, describing nonrandom distributions of trimers or hexamers, and so on.
- The **parameters of a Markov model** have to be **trained using a set of sequences with known gene locations**.
- Once the **parameters of the model are established**, it can be used to **compute the nonrandom** distributions of **trimers or hexamers** in a new sequence to find regions that are **compatible with the statistical profiles in the learning set.**

- The **frequency of six unique nucleotides** appearing together in a **coding region** is much **higher** than by **random chance**.
- Therefore, a **fifth-order Markov model**, which calculates **the probability of hexamer bases**, can **detect nucleotide correlations found in coding regions more accurately**.

- **Problem of fifth-order Markov chain:** is that if there are **not enough hexamers**, which happens in **short gene sequences**, the **method's efficacy may be limited**.
- **Solution:** A **variable-length Markov model**, called an **interpolated Markov model(IMM),** has been developed where the IMM method **samples the largest number of sequence patterns with k ranging from 1 to 8** (dimers to ninemers) and uses a **weighting scheme**, placing **less weight on rare k-mers** and **more weight on more frequent k-mers.**
- The **probability of the final model** is the **sum of probabilities** of **all weighted k-mers**. i.e. this method has **more flexibility** in using **Markov models** depending on the amount of data available.
- Sometimes, genes tend to escape detection using the typical gene model. Thus, to make the algorithm capable of fully describing all genes in a genome, more than one Markov model is needed and therefore HMM prediction algorithm are implemented.


0.90
typical coding state
A 0.80
C 0.02
G 0.10
T 0.08
A 0.01
C 0.01
G 0.01
T 0.97
A 0.01
C 0.01
G 0.97
T 0.01
A 0.27
C 0.32
G 0.23
T 0.18
A 0.60
C 0.10
G 0.28
T 0.02
A 0.25
C 0.23
G 0.37
T 0.15
A 0.47
C 0.10
G 0.28
T 0.05
A 0.06
C 0.24
G 0.55
T 0.15
A 0.11
C 0.36
G 0.22
T 0.31
A 0.01
C 0.01
G 0.01
T 0.97
A 0.86
C 0.01
G 0.12
T 0.01
A 0.36
C 0.01
G 0.62
T 0.01
Start codon
Stop codon
atypical coding state
0.10
Coding region

Prediction via HMM/IMM-based algorithm

1. GeneMark
2. Glimmer
3. FGENESB
4. RBSfinder

ETC……..

Performance Evaluation

- Accuracy of a prediction program : **Sensitivity and Specificity**.
- Therefore, sensitivity and specificity are calculated on four features accurately which are

1. **true positive (TP),** which is a correctly predicted feature;
2. **false positive (FP),** which is an incorrectly predicted feature;
3. **false negative (FN),** which is a missed feature;
4. **true negative (TN),** which is the correctly predicted absence of a feature

Formula used is:

1. Sensitivity (Sn) = TP/(TP + FN) (**Proportion** of **true signals** predicted among all **possible true signals,i.e ability to include correct predictions)**
2. Specificity (Sp) = TP/(TP + FP) (**Proportion** of **true signals** among all signals **that are predicted**. i.e. an **ability to exclude incorrect predictions)**

- A **program** is considered **accurate** if **both sensitivity and specificity are simultaneously high and approach a value of 1**.
- If **sensitivity is high** but **specificity is low**, the program is said to have a **tendency to over predict**.

- If the **sensitivity is low** but **specificity high**, the program is **too conservative and lacks predictive power.**
- Because neither **sensitivity nor specificity alone can fully describe accuracy**, it is desirable to use a single value to summarize both of them.
- In the field of gene finding, a single parameter known as the **correlation coefficient (CC)** is often used, which is defined by the following formula:

$$CC = TP \bullet TN - FP \bullet FN / \sqrt{(TP + FP)(TN + FN)(FP + TN)}$$

CC provides an overall measure of accuracy, which ranges from −1 to +1, with +1 meaning always correct prediction and −1 meaning always incorrect prediction.

Gene Prediction in Eukaryotes

- Eukaryotic nuclear genomes are much larger than prokaryotic ones
- Sizes ranging from: 10 Mbp to 670 Gbp (**1 Gbp = 10^9 bp**).

- Have a very **low gene density**.
- In humans, for instance, **only 3% of the genome codes for genes** (1 gene/100 kbp on average)
- The **space between genes** is often **very large** and **rich in repetitive sequences** and **transposable elements**.
- Eukaryotic genomes are characterized by a **mosaic organization** in which a gene is **split into pieces (exons**) by intervening noncoding sequences **(introns).**
- A eukaryotic gene is modified in **three different ways** before becoming a mature mRNA for protein translation.

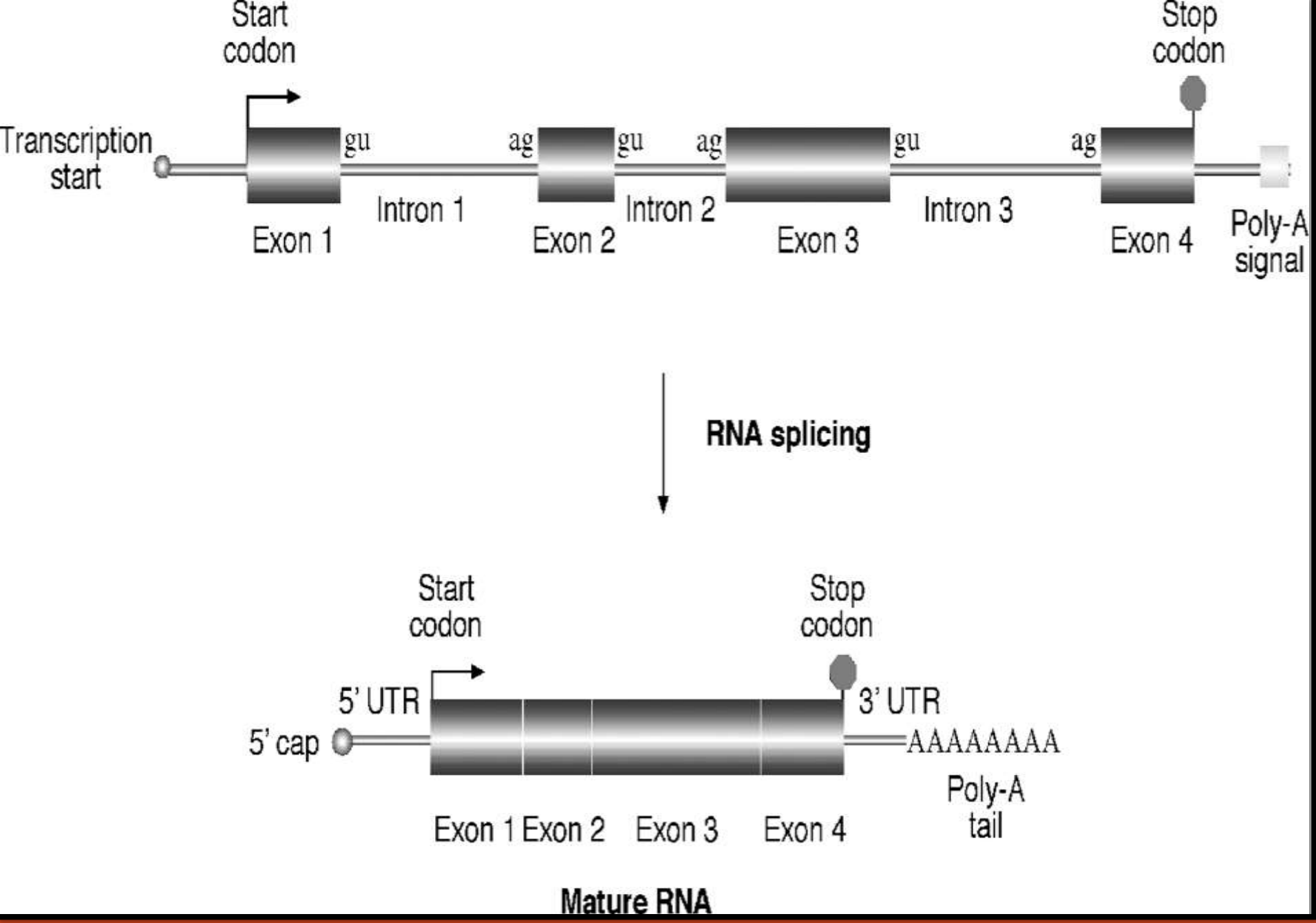
a. capping at the 5 end of the transcript (methylation at the initial residue of the RNA)

b. splicing (which is the process of removing introns and joining exons)

c. polyadenylation (which is the addition of a stretch of As (~250) at the 3 end of the RNA, which is controlled by a poly-A signal, a conserved motif slightly downstream of a coding region with a consensus CAATAAA(T/C)

- **Problem in prediction:** identification of exons, introns, and splicing sites.
- **Computationally very demanding** because of the presence of **split gene structures, alternative splicing**, and very **low gene densities**.
- **Solution:** some conserved sequence features in eukaryotic genes that allow the computational prediction.

Eg: the **splice junctions of introns and exons** follow the **GT–AG** rule in which an **intron at the 5 splice junction** has a consensus motif of **GTAAGT**; and at the **3 splice junction is a consensus motif** of **(Py)12NCAG**


Start codon
Stop codon
Transcription start
gu
ag
gu
ag
gu
ag
Exon 1
Intron 1
Exon 2
Intron 2
Exon 3
Intron 3
Exon 4
Poly-A signal
RNA splicing
Start codon
Stop codon
5' UTR
3' UTR
5' cap
AAAAAAAA
Poly-A tail
Exon 1 Exon 2
Exon 3
Exon 4
Mature RNA

- **Some prokaryotic program can be applied in eukaryotes studies such as TESTCODE and GC bias to check the Hexamer frequencies** in coding regions as they are also higher than in the noncoding regions**.**
- Most **vertebrate genes use ATG** as the **translation start codon** and have a **uniquely conserved flanking sequence** call a ***Kozak sequence*** (CCGCC**ATG**G).
- In addition, most of these genes have a **high density of CG dinucleotides** near the **transcription start site**.
- And this region is referred to as a **CpG island** (*p* refers to the phosphodiester bond connecting the two nucleotides), which helps to identify the **transcription initiation site** of a **eukaryotic gene**.
- The **poly-A signal** can also help locate the final **coding sequence**.

Gene Prediction Programs

- 3 categories of algorithms:

1. Ab initio based:

a. Prediction using Neural Network

b. Using Discriminant Analysis

c. Using HMM's

2. Homology based

3. Consensus based

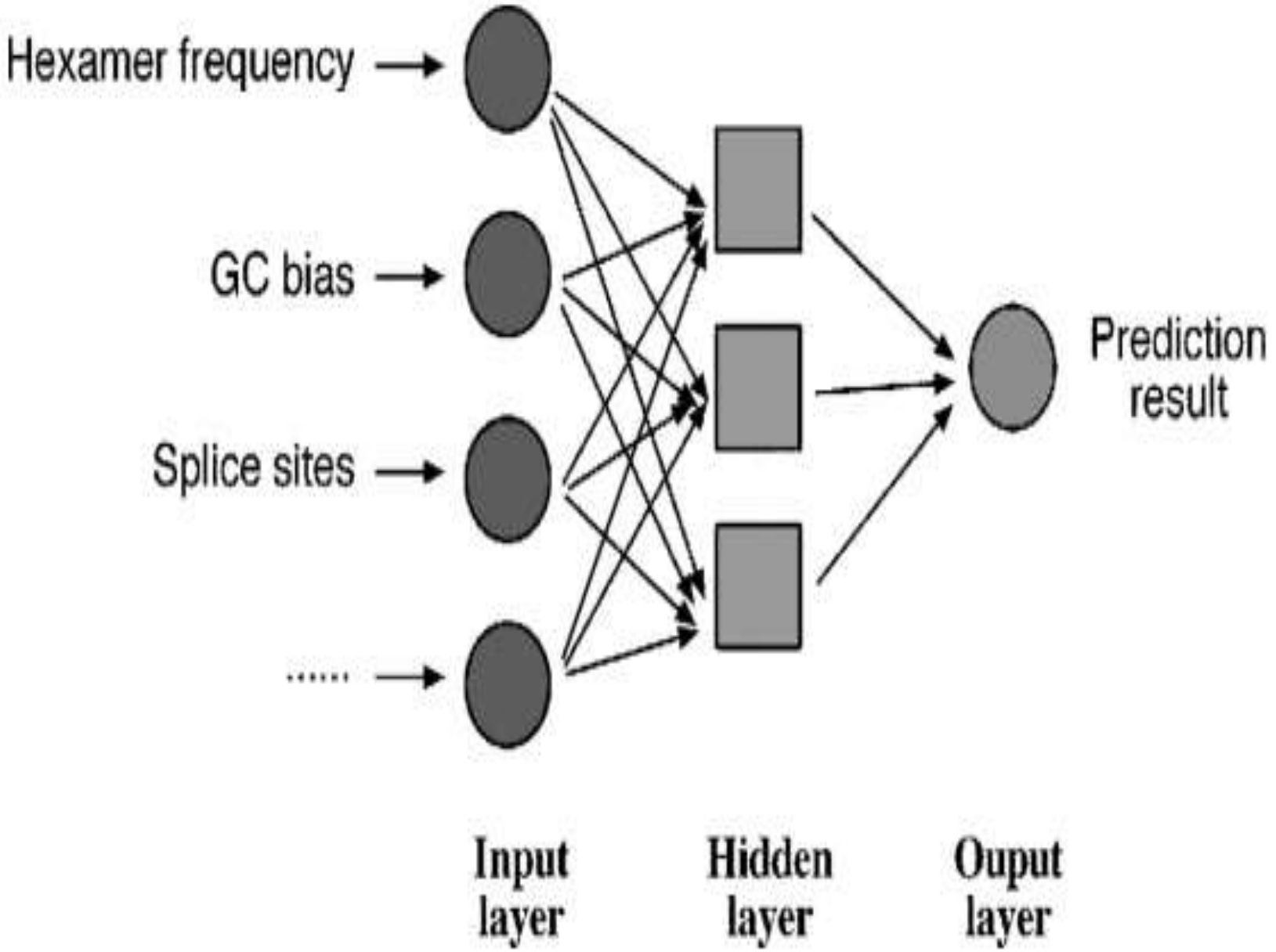
Ab Initio–Based Programs

- **Ab initio gene prediction programs** is to **discriminate exons from introns** and subsequently join the **exons together** in the **correct order.**
- **Difficulty: To identify correct exons.**
- **Therefore, to predict exon :** algorithms rely on two features,

a. **Gene signals:** which includes signals such as gene start and stop sites and putative splice sites, recognizable consensus sequences such as poly-A sites

b. **Gene content:** which includes coding statistics, such as non random nucleotide distribution, amino acid distribution, synonymous codon usage, and **hexamer frequencies** (most discriminative for coding potentials)

- **Assessment:** performed by HMM, neural network

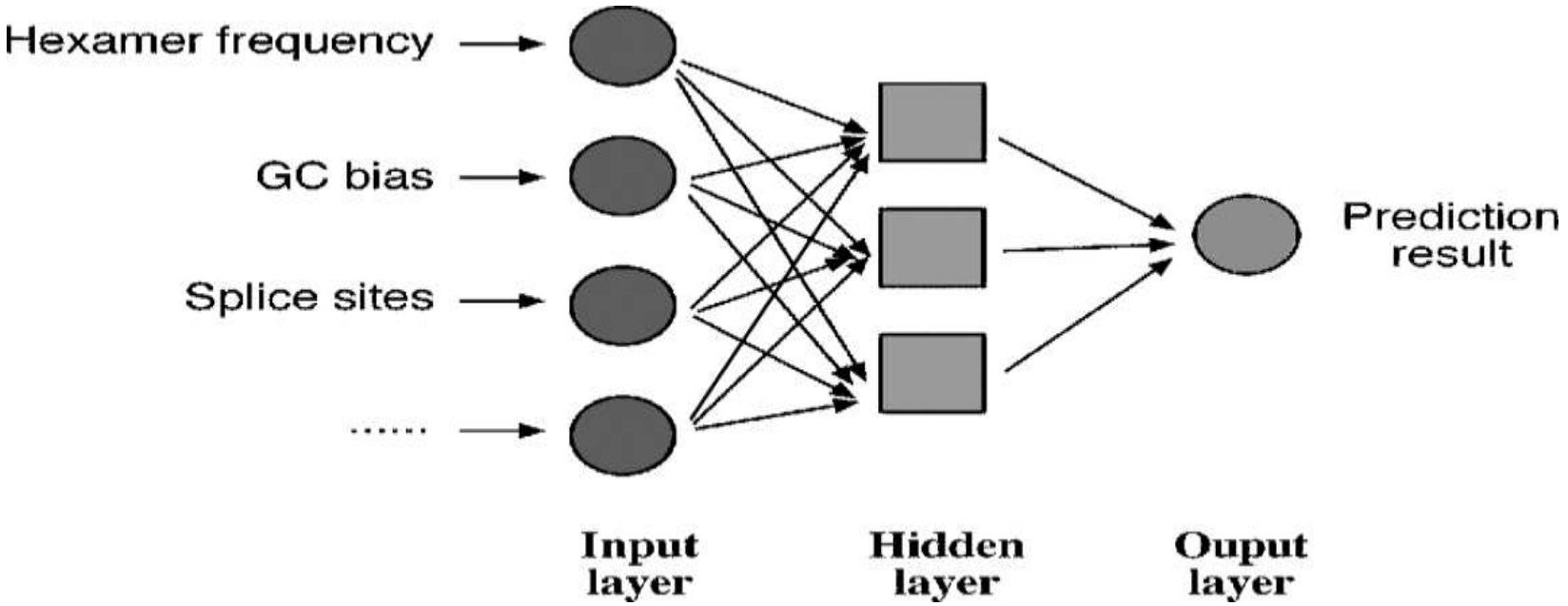

Hexamer frequency
GC bias
Splice sites
.....
Prediction result
Input layer
Hidden layer
Ouput layer

Neural network/ANN for gene prediction

- A NN/ANN is a statistical model with a **special architecture** for **pattern recognition and classification.**
- It is composed of a **network of mathematical variables** that resemble the biological **nervous system**, with **variables or nodes** connected by **weighted functions** that are analogous to **synapses.**
- **Another feature that it resembles to nervous system is :** is its ability to **"learn" and then make predictions after being trained.**
- The **network** is able to **process information** and **modify parameters** of the **weight functions** between v**ariables** during **the training stage**. Once it is **trained**, it is able to **make automatic predictions about the unknown**.
- A NN is constructed with multiple layers; the input, output, and hidden layers.
- The i**nput** is the **gene sequence** with **intron and exon signals.**

- The output is the probability of an exon structure.
- Between input and output, there may be one or several hidden layers where the machine learning takes place.
- The machine learning process starts by feeding the model with a sequence of known gene structure.
- The gene structure information is separated into several classes of features such as hexamer frequencies, splice sites, and GC composition during training.
- The weight functions in the hidden layers are adjusted during this process to recognize the nucleotide patterns and their relationship with known structures.
- When the algorithm predicts an unknown sequence after training, it applies the same rules learned in training to look for patterns associated with the gene structures.

Neural network



Prediction TOOL via NN/HMM-based/ Discriminant algorithm

GRAIL: (Gene Recognition and Assembly Internet Link; http://compbio.ornl.gov/ public/tools/)

A. NN network algorithm

B. traning giving on various features such as splice junctions, start and stop codons, poly-A sites, promoters, and CpG islands.

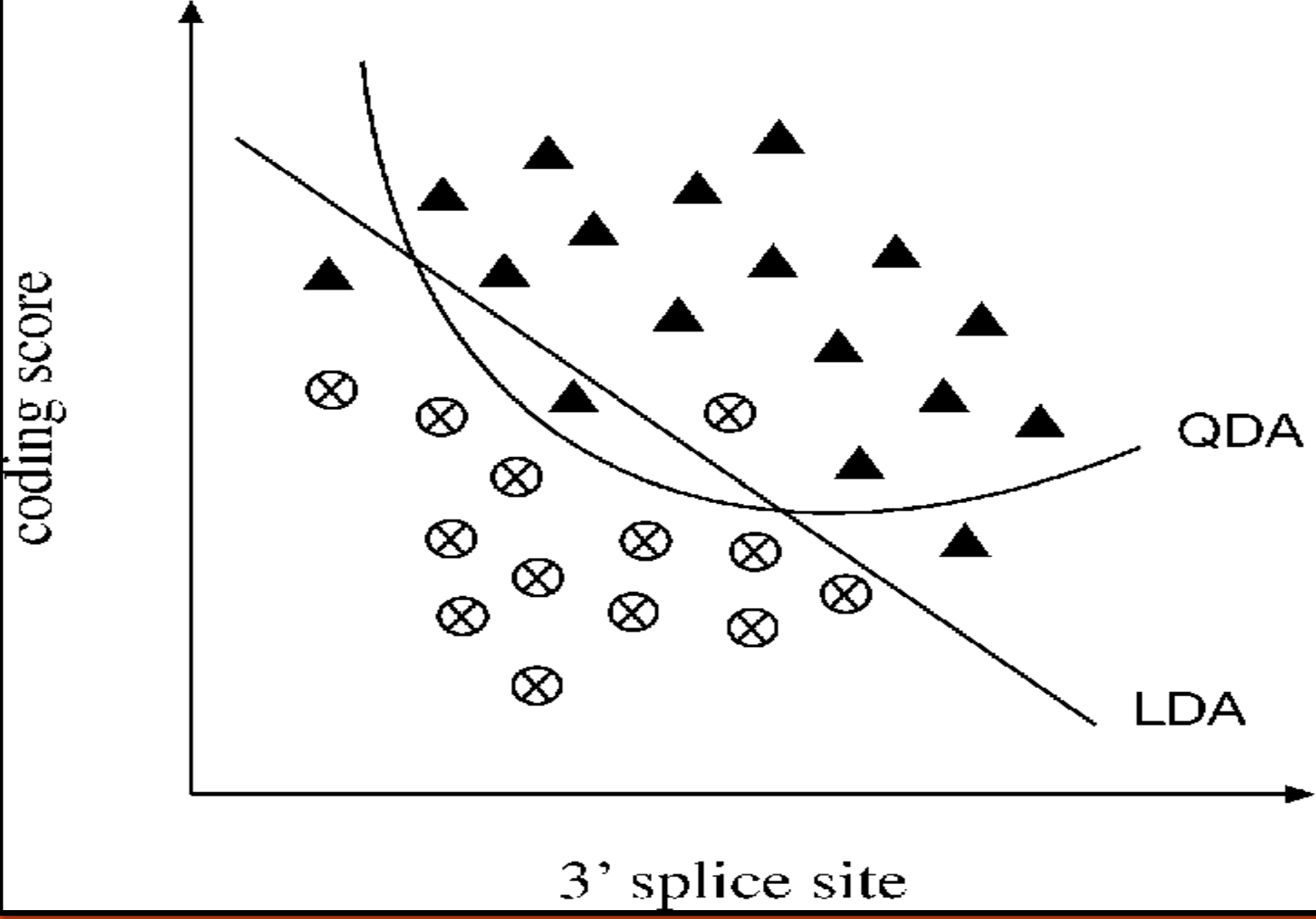
C. program scans the **query sequence** with **windows of variable lengths** and **scores for coding potentials** and finally produces an output that is the result of **exon candidates.**

Prediction Using Discriminant Analysis

- Some gene prediction algorithms rely on **discriminant analysis**, either **Linear Discriminant analysis (LDA)** or **quadratic discriminant analysis (QDA)**, to improve accuracy.

- **LDA** works by **plotting a 2D graph** of coding signals versus all potential 3 splice site positions and drawing a **diagonal line** that **best separates coding signals from noncoding signals** based on **knowledge learned** from **training data sets** of known gene structures .

- **QDA** draws a **curved line** based on a **quadratic function** instead of drawing a straight line to **separate coding and noncoding features**. This strategy is designed to be more **flexible and provide a more optimal separation between the data points.**


coding score
QDA
LDA
3' splice site

Prediction TOOL via Discriminant algorithm

- **FGENES** (FindGenes;www.softberry.com/) is a web-based program that uses **LDA to determine whether a signal is an exon**.

- In addition to FGENES, there are many variants of the program. Some programs, such as **FGENESH, make use of HMMs**.

- **Other program/ tool are:**

a. FGENESH C

b. FGENESH+ (combine both ab initio and similarity-based approaches)

c. MZEF (Michael Zhang's Exon Finder; http://argon.cshl.org/genefinder/) is a webbased program that uses QDA for exon prediction.

Prediction TOOL via HMMs

GENSCAN (http://genes.mit.edu/GENSCAN.html)

a. is a web based program that makes predictions based on **fifth-order HMMs.**

b. It combines **hexamer frequencies** with **coding signals** (initiation codons, TATA box, cap site, polyA, etc.) in prediction.

c. **Putative exons** are assigned a **probability score (P)** of being a **true exon**.

d. Only predictions with **P > 0.5** are **deemed reliable**.

Other Tool: HMMgene

The homology-based method

- It makes predictions based on significant matches of the query sequence with sequences of known genes.

- Homology-based programs are based on the fact that exon structures and exon sequences of related species are highly conserved.
- When potential coding frames in a query sequence are translated and used to align with closest protein homologs found in databases, near perfectly matched regions can be used to reveal the exon boundaries in the query.
- This approach assumes that the database sequences are correct.
- The drawback of this approach is its reliance on the presence of homologs in databases.
- If the homologs are not available in the database, the method cannot be used.
- Novel genes in a new species cannot be discovered without matches in the database.

Prediction TOOL via homology-based method

GenomeScan :

- web-based server that combines **GENSCAN prediction results with BLASTX similarity searches.**
- Input : **genomic DNA and protein sequences** from **related species.**
- The **genomic DNA** is **translated in all six frames** to cover all possible exons.
- The **translated exons** are then used to **compare** with the **user-supplied protein sequences**.
- **Translated genomic regions** having **high similarity** at the **protein level** receive **higher scores**.
- The **same sequence** is also predicted with a **GENSCAN** algorithm, which gives **exons probability scores.**
- **Final exons** are assigned based on **combined score information** from both analyses
- **OTHER: SGP-1, EST2Genome,Twinscan,etc..**

Consensus-Based Programs

- As different prediction programs have different levels of sensitivity and specificity, it makes sense to **combine results** of **multiple programs based on consensus.**
- These programs work by **retaining common predictions** agreed by most programs and **removing inconsistent predictions**.
- Such an **integrated approach** may improve the **specificity** by correcting the false positives and the problem of over prediction.
- However, since this procedure punishes **novel predictions**, it may lead to **lowered sensitivity** and **missed predictions**. Two examples of consensus-based programs are:

Prediction TOOL via Consensus-Based Programs

GeneComber (www.bioinformatics.ubc.ca/genecomber/index.php)

- is a web server that combines **HMMgene and GenScan** prediction results.
- The **consistency of both prediction** methods is **calculated**.
- If the **two predictions match**, the **exon score is reinforced**.
- **If not**, exons are proposed based on separate **threshold scores**.

Other Tool: DIGIT (FGENESH, GENSCAN, and HMMgene.**), etc..**

THANKYOU