Computational Gene Prediction - I

A gene codes for a protein

DNA

transcription

mRNA

translation

Protein

CCTGAGCCAACTATTGATGAA

CCU GAG CCAACU AUU GAU GAA

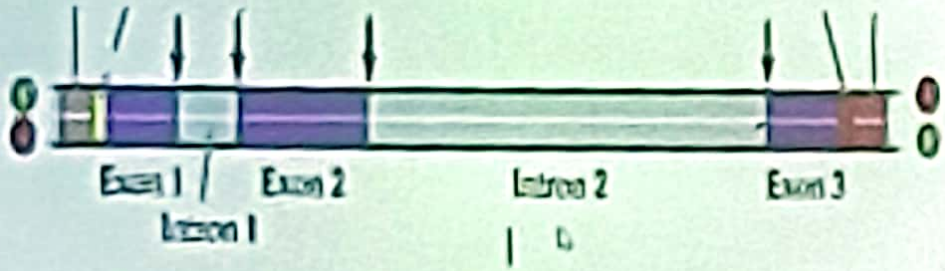
PEPTIDE

Importance of Gene Identification

- First step towards getting at the function of a protein.
  - Functional annotation
- Helps accelerate the annotation of genomes.
  - Structural annotation

Structural annotation consists in the identification of genomic elements:

* coding regions & their localisation
* gene structure
* location of regulatory motifs



Functional annotation consists in attaching biological information to genomic elements:

* biochemical/biological function
* regulation & molecular interactions
* tissue-specific expression

involves both biological expts and in silico analysis

Genome annotation: - location & function of genes

- Automatic annotation - computer analysis
- Manual annotation - human expertise

Ideally, these approaches co-exist and complement each other in the same annotation pipeline.

An active area of investigation, some of the on-going projects:

- Ensembl
- Gene Ontology Consortium
- RefSeq
- Uniprot
- Vertebrate & Genome Annotation Project (Vega)

Basic level of annotation is using BLAST for finding similarities.

What is Computational Gene Finding?

Given an uncharacterized DNA sequence, find out:

- Which region codes for a protein?
- Which DNA strand is used to encode the gene?
- Which reading frame is used in that strand?
- Where does the gene start and end?
- Where are the exon-intron boundaries in eukaryotes?
- (optionally) Where are the regulatory sequences, polyA signal for that gene?

Search space - 2-5% of Genomic DNA (~ 100Mbp)

Approaches

- Finding Open Reading Frames (ORFs)
- Homology Search
- Signal-based methods:
  - CpG islands
  - Finding promoter regions, poly adenylation sites, intron/exon splice sites
- Content-based methods:
  - Coding statistics, *viz.*, codon usage bias, periodicity in base occurrence, etc.
- Integration of these methods

Open Reading Frames (ORFs)

A long sequence between two stop codons devoid of stop codons in-between is called an ORF.

Finding ORFs is very reliable in the case of bacterial genomes compared to eukaryotic genomes, due to the structure and density of the coding regions.

Open Reading Frames (ORFs)

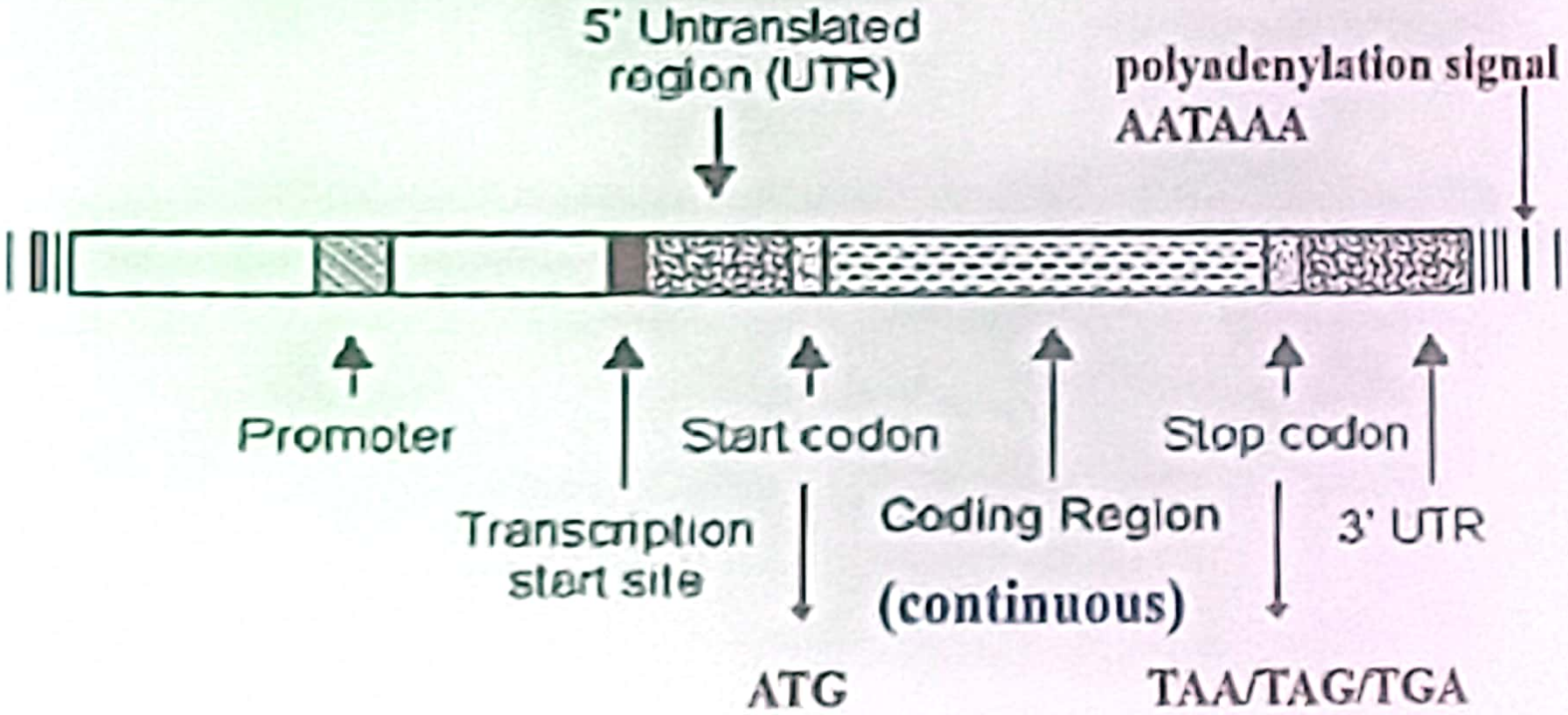
An ORF may code for a gene if it:

- Contains a homolog in the database
- Contains gene-specific features, *viz.*, 3-base periodicity, high G+C content, bias in codon usage, etc.
- Is the codon usage similar to other genes of the same organism?
- Contains signal sequence patterns for translation initiation

All these conditions may be satisfied by a pseudogene!

Gene Structure in Prokaryotes

- most of the DNA sequence codes for protein (e.g., 70% of *H. influenzea* bacterium genome is coding)
- each gene is one continuous stretch of bases



Finding genes in Prokaryotes

Gene prediction in prokaryotes involves:

?

- identifying long open reading frames,
- using codon frequencies

Codon Usage being species-specific, few representative set of genes of the organism is required

Finding genes in Prokaryotes

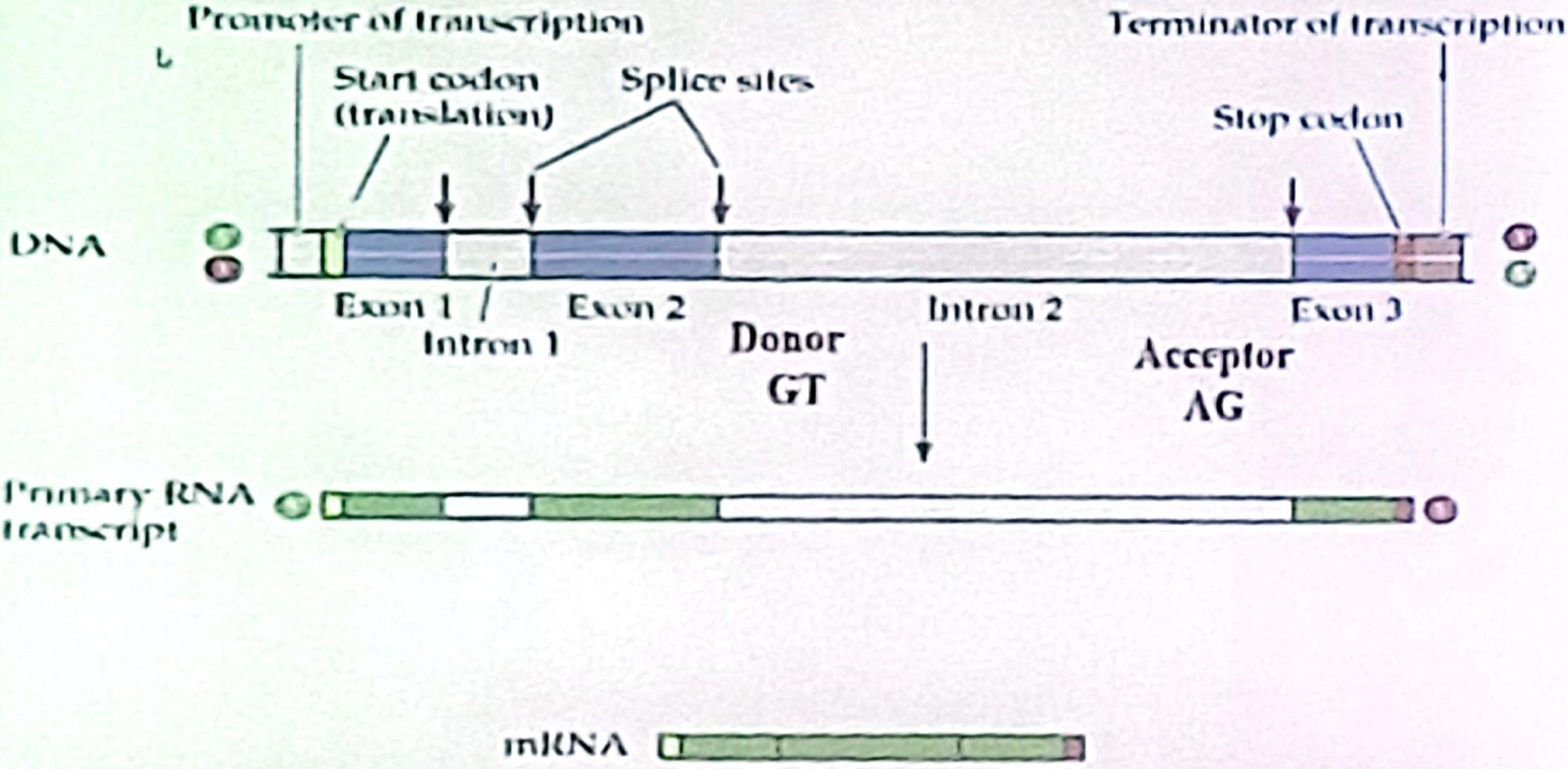
Gene prediction in prokaryotes involves:

- identifying long open reading frames,
- using codon frequencies

Codon Usage being species-specific, few representative set of genes of the organism is required

Gene Structure in Eukaryotes

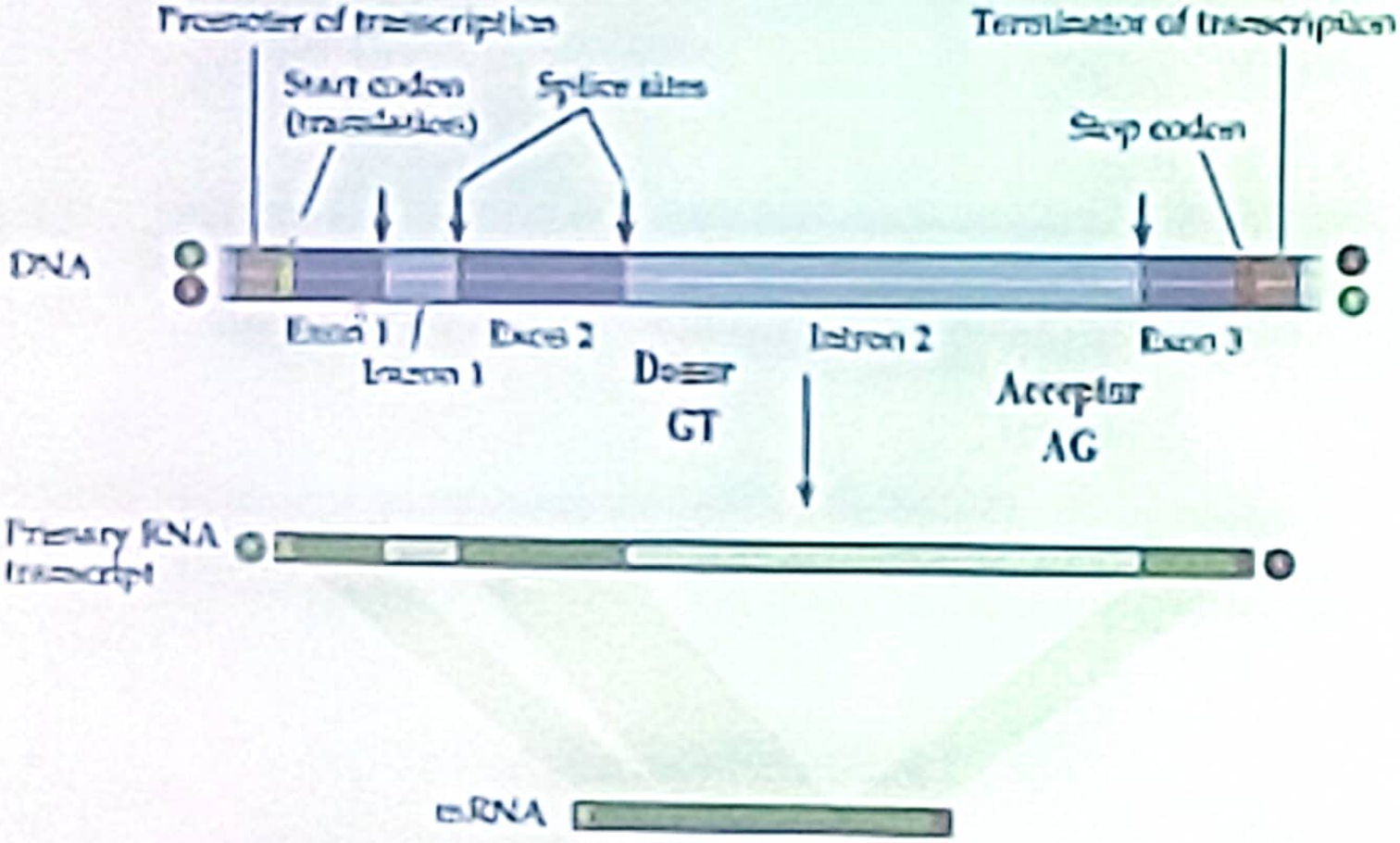
- coding region is usually discontinuous
- composed of alternating stretches of exons & introns



Note: Some eukaryotic genes are single-exon, or, intron-less genes

Gene Structure in Eukaryotes

- coding region is usually discontinuous
- composed of alternating stretches of exons & introns



Note: Some eukaryotic genes are single-exon, or, intron-less genes

Finding genes in Eukaryotes

Gene finding problem complicates:

- due to the existence of interweaving exons and introns – stop codons may exist in intronic regions making it difficult to identify the correct ORF
- a gene region may encode many proteins – due to alternative splicing / alternative translation initiation
- Exon length need not be multiple of three – resulting in frameshift between exons
- Gene may be intron-less (single-exon genes)
- Relatively low gene density - only 2 - 3% of the human genome codes for proteins

Finding Open Reading Frames (ORF)

5′ TCAATGTAACGCGCTACCCGGAGCTCTGGGCCCAA
ATTTCATCCACT 3′

Three reading frames in the forward direction:

Start codon ↘ ↙ Stop codon

1. TCA ATG TAA CGC GCT ACC CGG AGC TCT GGG
CCC AAA TTT CAT CCA CT

2. CAA TGT AAC GCG CTA CCC GGA GCT CTG GGC
CCA AAT TTC ATC CAC T

3. AAT GTA ACG CGC TAC CCG GAG CTC TGG GCC
CAA ATT TCA TCC ACT

Note: set of codons obtained in each of the three cases are different – resulting in a different amino acid sequence.

Finding Open Reading Frames (ORF)

3′ AGTTACATTGCGCGATGGGCCTCGAGACCCGGGTTT
AAAGTAGGTGA 5′

Three reading frames in the reverse direction:

1. AG TTA CAT TGC GCG ATG GGC CTC GAG ACC CGG
GTT TAA AGT AGG TGA

2. A GTT ACA TTG CGC GAT GGG CCT CGA GAC CCG
GGT TTA AAG TAG GTG

3. AGT TAC ATT GCG CGA TGG GCC TCG AGA CCC
GGG TTT AAA GTA GGT

Stop codon

Start codon

Finding Open Reading Frames (ORF)

In this case the longest open reading frame (ORF) is the 3rd reading frame on the complementary strand :

AGT TAC ATT GCG CGA TGG GCC TCG AGA CCC GGG TTT AAA GTA

When read 5′ to 3′, the longest ORF is:

ATG AAA TTT GGG CCC AGA GCT CCG GGT AGC GCG TTA CAT TGA

Is the longest ORF always the coding DNA?

If not, what's the solution?

Finding Long ORFs

- One way to distinguish between a coding and a non-coding region is to look at the frequency of stop codons
- Sequence similarity search
- When no sequence similarity is found, an ORF can still be considered gene-like according to some statistical features:
  - three-base periodicity,
  - higher G+C content in the coding regions
  - signal sequence patterns for translation initiation

An ORF having statistical features of a gene may be a psuedogene – if it lacks signals associated with transcription/translation initiation.

Finding genes in Prokaryotes

Once a long ORF/ all ORFs above a certain threshold are identified,

- these ORF sequences are called putative coding sequences
- translate each ORF using the Universal Genetic Code to obtain amino acid sequence
- search against the protein database for homologs

Gene discovery in prokaryotic genomes is a simple problem, owing to the higher gene density and the absence of introns in their protein coding regions.

Finding genes in Prokaryotes

Drawbacks:

- Addition or deletion of one or two bases will cause all the codons scanned to be different

  ⇒ sensitive to *frame shift errors*

- Fails to identify very small coding regions
- In general, the largest ORF is the one that codes for proteins – need not be always true.
- Fails to identify the occurrence of overlapping long ORFs on opposite DNA strands (genes and 'shadow genes')

Overlapping genes on the same strand observed in bacterial genomes – an overlap of 2-3 bases in an operon

Web-based tools

ORF Finder (NCBI)

https://www.ncbi.nlm.nih.gov/orffinder

EMBOSS Programs

- getorf - Finds and extracts open reading frames
- plotorf - Plot potential open reading frames
- showorf - Pretty output of DNA translations
- Sixpack - Display a DNA sequence with 6-frame translation and ORFs

http://www.ebi.ac.uk/Tools/emboss/

Homology Search

Involves sequence-based database search:

- **DNA Database search**
- **Protein Database search**

Homology Search

Use protein for database similarity searches whenever possible

Homology Search

Reasons for Protein vs. DNA Database Search:

- Very different DNA sequences may code for similar protein sequences – we wouldn't want to miss such hits
- Translating to a protein sequence corrects for different codon usage, base composition, and other organism specific DNA sequence variations.
- When comparing DNA sequences, we get significantly more random hits than with proteins. There are several reasons for these:

Homology Search

- DNA being composed of 4 characters: two unrelated DNA sequences are expected to have 25% similarity
- In contrast, an amino acid sequence being composed of 20 characters, the sensitivity of the comparison is improved
- It is expected that convergence of proteins is rare, implies that high similarity between proteins implies homology
- DNA databases are much larger, and grow faster than proteins databases. Bigger databases ⇒ more random hits

Homology Search

- For protein similarity searches more sensitive scoring matrices like PAM and BLOSUM are used – resulting in a more sensitive search.
- Because of less mutations in proteins during evolution, searching protein databases may reveal remote evolutionary relationships.

Homology Search

Three main search tools used for database search:

- **FastA - algorithm by Pearson & Lipman**
  http://www.ebi.ac.uk/fasta33/
- **BLAST - algorithm by Karlin & Altschul**
  http://www.ncbi.nlm.nih.gov/BLAST/
- **SSearch - Smith-Waterman (SW) algorithm**

SSearch - can be very specific when identifying long regions of low similarity especially for highly diverged sequences

Homology Search

Tips for Databases Searches

- Always use the latest database version
- Run BLAST first, then depending on the results, run a finer tool (FastA, SW, etc.)
- Whenever possible, use the translated sequence.
- $E < 0.05$ is statistically significant, usually biologically interesting. However, also check $0.05 < E < 10$, as one might find interesting hits.

Homology Search

- Pay attention to abnormal composition of the query sequence, it usually causes biased scoring
- If the query has repeated segments, mask them and then perform the search.

Can one find novel genes by this approach?

Homology Search

Data used for gene identification by homology search:

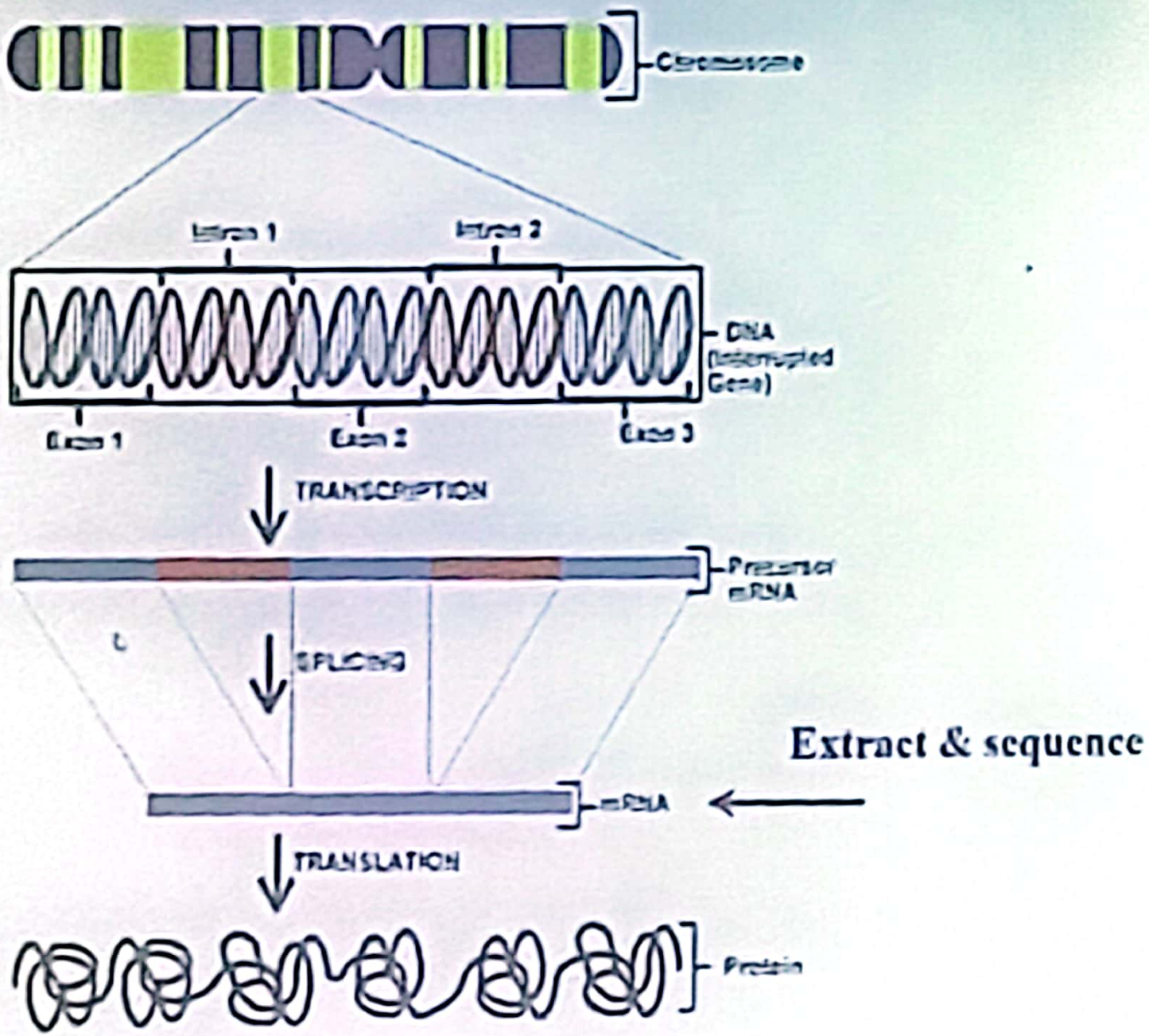
- **Putative ORFs**
- **EST sequences**

What Are ESTs?

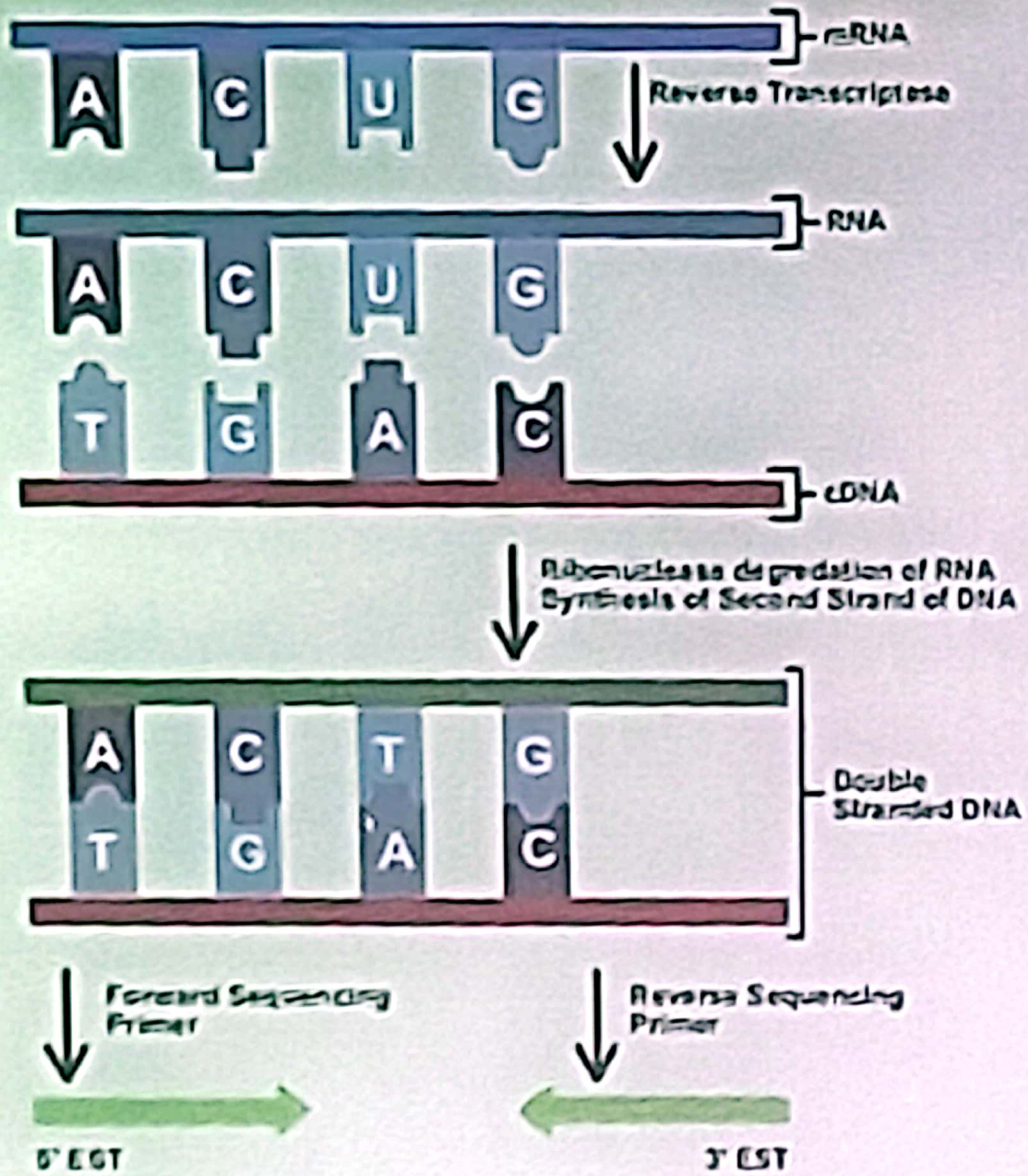
- small pieces of DNA sequence (~ 200 - 500 bp), generated by sequencing either one or both ends of an expressed gene

- represent a snapshot of genes expressed in a certain tissue or at certain developmental stage.

An overview of the process of protein synthesis



An overview of how ESTs are generated



Using ESTs for Gene Finding

Gene identification by homology-based approach can also be done using EST sequences.

How reliable is this approach compared to *in silico* gene prediction approaches?

Using ESTs for Gene Finding

As the number of ESTs runs into millions now and approaches saturation,

Blastn (est) has become a far more better method for finding gene features.

Unlike *in silico* gene predictions, EST matches are experimentaliy grounded.

One issue: since non-coding RNA genes are also expressed, we need to confirm whether the EST finds a hit to protein sequence

Using ESTs for Gene Finding

Things to note on Blast matches:

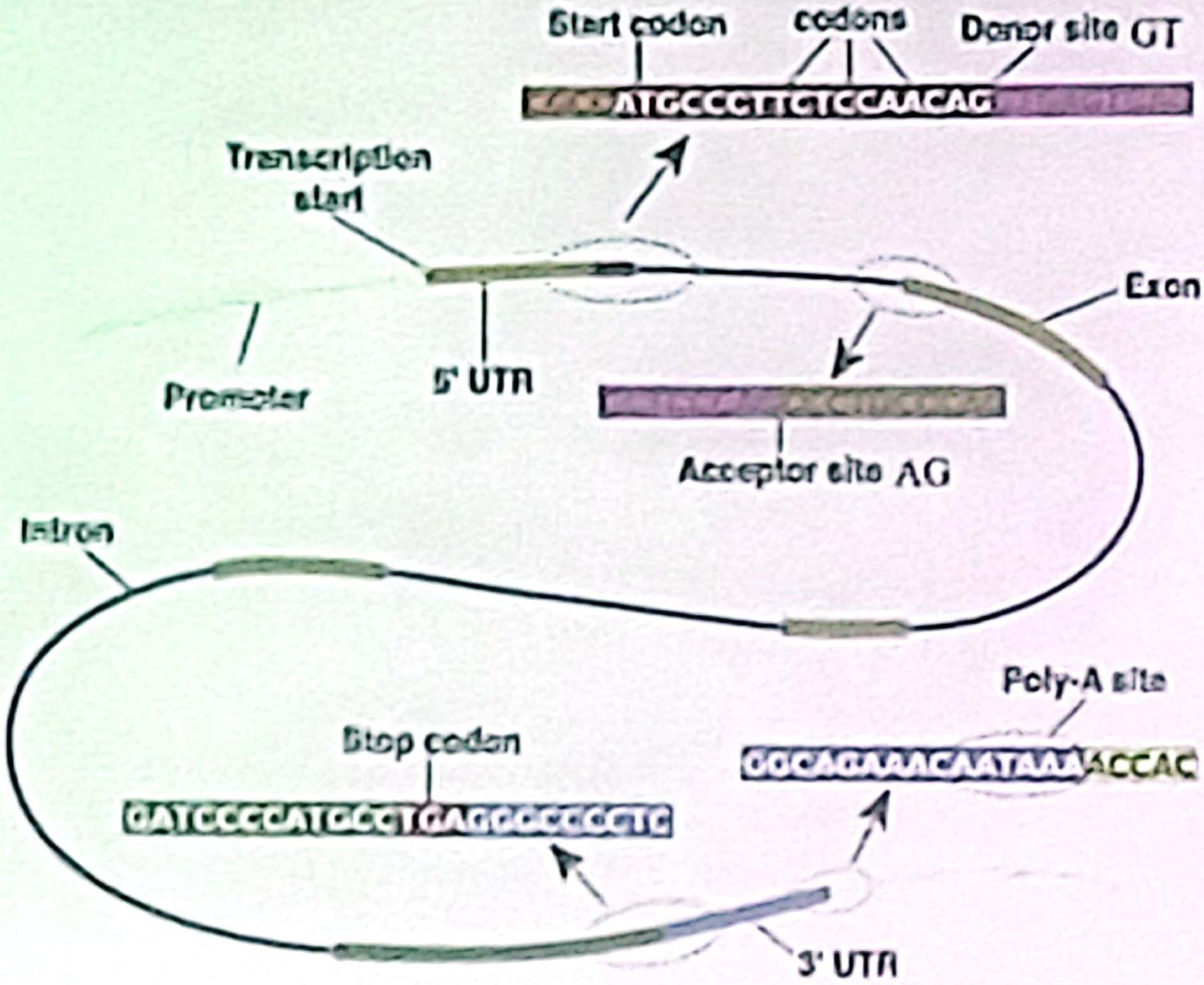
- whether the match is near-perfect (e.g., 99%) or just good
- whether a few or a no. of similar matches are found.
- Eliminate repeats - the best hit should be immediately checked by RepeatMasker.
- Watch for gapped EST matches indicating a spliced-out intron

Using ESTs for Gene Finding

- For experimental reasons, EST data is strongly biased towards the 3' end of genes;
- Some pseudogenes are also transcribed.
- Some genes have functional anti-sense transcripts.
- Many proteins have internal duplications
- Tandem genomic duplication is also common, often with one copy becoming a pseudogene.

Signal-based Methods

Signal – a string of DNA recognized by the cellular machinery



Signals for gene identification

Many signals are associated with genes, each of which suggests but does not prove the existence of a gene

- CpG Islands - helps to identify the coding regions (mainly housekeeping genes ~ 50% in humans)
- Start / Stop Codons – signifies start /end of coding regions
- Transcription Start Site - identifies start of coding regions
- Donor / Acceptor Sites – signifies start / end of intronic regions

Signals for gene identification:

- Promoters – short sequences that regulate gene expression (found in 5′ UTR region)
- Enhancers – regulates gene expression, (found in 5′ or 3′ UTR regions, intronic regions, or up to few Kb away from the gene)
- Motifs – short DNA sequences where proteins bind to initiate transcription / translation process
- Poly-A Site – identifies the end of coding region (found in 3 UTR region)

Most of these signals can be modeled by Position Specific Scoring Matrices (PSSM), or Hidden Markov models (HMM)

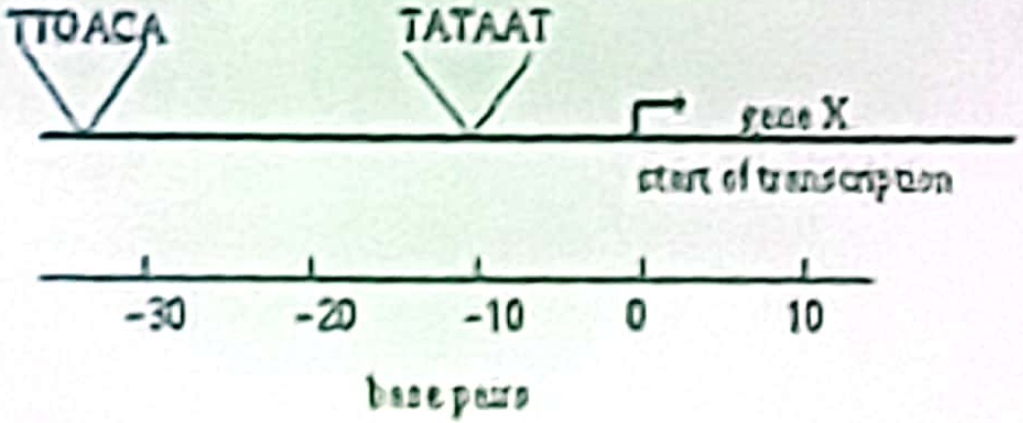
Promoter Detection

Not all ORFs are genes

True coding regions have specific sequences upstream of the start site known as promoters where RNA polymerase binds to initiate transcription, e.g., in *E. coli*:

- **TATA box around -10bp: TATAAT**
- **Around –35 bp: TTGACA**

Consensus patterns



Promoter Detection

Signals are short and variable

- Simplest approach to identify signals is by using positional frequencies
- Compute frequency of nucleotide b at position i, $f(b,i)$
- Probability of sequence S to be a promoter is:
  $\Pi_i f(b,i)$ ($i = 1,\ldots,6$ for TATA box)
- Probability S is not a promoter $\Pi f(b)$, where $f(b)$ is the expected frequency of b
- Find odds ratio of S being a promoter to not being a promoter

Positional Weight Matrix

	1	2	3	4	5	6
A	2	**95**	26	**59**	**51**	1
C	9	2	14	13	20	3
G	10	1	16	15	12	0
T	**79**	2	**44**	13	17	**96**

Divide the frequencies by the total no. to obtain weights.
Similar PSSM approach is used for splice-site recognition and regulatory motifs.

Splice-Site Detection

Can PSSM approach be used for splice-site recognition?

Since a dinulceotide AG/GT may occur a number of times in a gene sequence,

- to identify AG/GT that corresponds to a true exon-intron boundary, typically about 6 bases flanking the known exon-intron boundary is considered and a PSSM constructed.

CpG islands

Detection of regions of genomic sequences that are rich in CpG pattern is important because such regions are resistant to methylation and tend to be associated with genes that are frequently switched on.

- 5' ends of all housekeeping genes and many tissue-specific genes, and 3' ends of some tissue-specific genes

Unmethylated regions of the genome rich in the CpG pattern are known as CpG islands

Methylation – addition of a methyl group to C of the CG dinucleotide – plays an important role in the suppression of downstream genes.

CpG islands

Methyl-C tends to mutate to T, and so CpG dinucleotides tend to decay to TpG / CpA.

- this results in CpG dinucleotides occurring about five times less frequently than expected

Absence of methylation slows CpG decay, and so CpG islands can be detected in DNA sequence as regions in which CG pairs occur at close to the expected frequency.

About 56% of human genes and 47% of mouse genes are associated with CpG islands

CpG islands

- Often CpG islands overlap promoter and extend about 1000 base pairs downstream into transcription unit and are generally found in the same position relative to the transcription unit of equivalent genes in different species

  - helps to define the extreme 5' ends of genes

- Probably because they are associated with genes, CpG islands tend to be unique sequences - presence of G/C boxes, GGGCGG, or its reverse complement, CCGCCC

  - and are very useful in genome mapping projects

Web-based Tools

EMBOSS

- **cpgreport: Reports CpG rich regions**
- **newcpgseek: Reports CpG rich regions**
- **cpgplot: Plot CpG rich areas**

http://www.ebi.ac.uk/emboss/cpgplot/

CpG Islands by Gardiner-Garden and Frommer

http://bioinformatics.org/sms/

Content-based Methods

At the core of all gene identification programs

- there exist one or more coding measures

- sequence-based measures indicative of protein-coding regions in a DNA sequence.

A coding statistic - a function that computes the likelihood that the sequence is coding for a protein.

A good knowledge of core coding statistics is important to understand how gene identification programs work.

Classification of Coding Measures

Coding statistics measure

- base compositional bias
- periodicity in base occurrence
- codon usage bias

Main distinction is between

- measures *dependent* on a model of coding DNA
- measures *independent* of such a model.

Measures dependent on a Model of Coding DNA

Measures may be based on

- **Unequal usage of codons in the coding regions - a universal feature of the genomes.**
- **Dependencies between nucleotide positions**
- **Base compositional bias between codon positions**

The Human Codon Usage Table

Gly	GGG	17.02	0.23	Arg	AGG	12.09	0.22	Trp	TGG	14.74	1.00	Arg	CGG	10.40	0.19
Gly	GGA	19.31	0.26	Arg	AGA	11.73	0.21	End	TGA	2.64	0.61	Arg	CGA	5.63	0.10
Gly	GGT	13.66	0.18	Ser	AGT	10.18	0.14	Cys	TGT	9.99	0.42	Arg	CGT	5.16	0.09
Gly	GGC	24.54	0.33	Ser	AGC	18.54	0.25	Cys	TGC	13.85	0.58	Arg	CGC	10.82	0.19
Glu	GAG	39.22	0.59	Lys	AAG	33.79	0.60	End	TAG	0.73	0.17	Gln	CAG	32.95	0.73
Glu	GAA	27.51	0.41	Lys	AAA	22.32	0.40	End	TAA	0.95	0.22	Gln	CAA	11.94	0.27
Asp	GAT	21.45	0.44	Asn	AAT	16.43	0.44	Tyr	TAT	11.80	0.42	His	CAT	9.56	0.41
Asp	GAC	27.05	0.56	Asn	AAC	21.30	0.56	Tyr	TAC	16.43	0.58	His	CAC	14.00	0.59
Val	GTG	22.60	0.48	Met	ATG	21.86	1.00	Leu	TTG	11.43	0.12	Leu	CTG	39.93	0.43
Val	GTA	6.09	0.10	Ile	ATA	6.05	0.14	Leu	TTA	5.55	0.06	Leu	CTA	6.42	0.07
Val	GTT	10.30	0.17	Ile	ATT	15.03	0.35	Phe	TTT	15.36	0.43	Leu	CTT	11.24	0.12
Val	GTC	15.01	0.25	Ile	ATC	22.47	0.52	Phe	TTC	20.72	0.57	Leu	CTC	19.14	0.20
Ala	GCG	7.27	0.10	Thr	ACG	6.20	0.12	Ser	TCG	4.38	0.06	Pro	CCG	7.02	0.11
Ala	GCA	15.90	0.22	Thr	ACA	15.04	0.27	Ser	TCA	10.96	0.15	Pro	CCA	17.11	0.27
Ala	GCT	20.23	0.28	Thr	ACT	13.24	0.23	Ser	TCT	13.51	0.18	Pro	CCT	18.03	0.29
Ala	GCC	28.43	0.40	Thr	ACC	21.52	0.38	Ser	TCC	17.37	0.23	Pro	CCC	20.51	0.33

Codon usage table is used to compute the coding potential of a nucleotide sequence, given by the log-likelihood ratio:

$$LP^{i}(S) = \log \frac{P^{i}(S)}{P_0(S)}$$

$P^i(S)$ - prob. of sequence S, given that S is coding in frame i (i=1,2,3)

$P^0(S)$ – prob. of S given a model of non-coding DNA. If

$$LP^{i}(S) > 0$$

$\Rightarrow$ prob. that S is coding in frame i is higher than S being non-coding.

Compute log-likelihood ratio in the 3 frames:

$$LP^{i}(S) = \log \frac{P^{i}(S)}{P_0(S)}$$

If the sequence is coding, $LP^{i}(S)$ will be larger for one of the frames.

Non-coding DNA - random DNA sequence with nucleotide equiprobability & independence between positions.

Codon Usage

F(C) - frequency of codon C in genes of the species under consideration (from codon usage table)

For a given sequence of codons

$$C = C_1 C_2 \ldots C_m$$

the probability of the sequence of codons C coding for a protein is given by

$$P(C) = F(C_1)F(C_2)\ldots F(C_m)$$

Codon Usage

For e.g., if *S* is the sequence *S* = AGGACG, when read in frame 1, it results in the sequence

$$C_1^1 = \mathtt{AGG},\ C_2^1 = \mathtt{ACG}.$$

$$P^1(S) = P(C^1) = F(\mathtt{AGG})F(\mathtt{ACG})$$

Substituting appropriate values from Table-I to compute $P^i(C)$, $i = 1, 2$

$$P^1(S) = P(C^1) = 0.0121 \times 0.007 = 0.0000847$$

Codon Usage

Log-likelihood ratio for *S* coding in frame 1, LP^1,

$$LP^1(S) = \log\left(\frac{P(C)}{P_1(C)}\right)$$

$$LP^1(S) = \log(0.000836/0.000244) = \log(3.43) = 0.53$$

Compute log-likelihood ratio for *S* coding in frames 2 & 3 also and compare

- Frame with the largest value of LP will be the coding frame

Complications in Gene Prediction

Problem of gene identification is further complicated in case of eukaryotes by the vast variation that is found in the structure of genes.

On an average, a vertebrate gene is 30Kb long. Of this, the coding region is only about 1Kb.

The coding region typically consists of 6 exons, each about 150bp long.

These are average statistics

Complications in Gene Prediction

Huge variations from the average are observed.

Biggest human gene, *dystrophin* is 2.4Mb long.

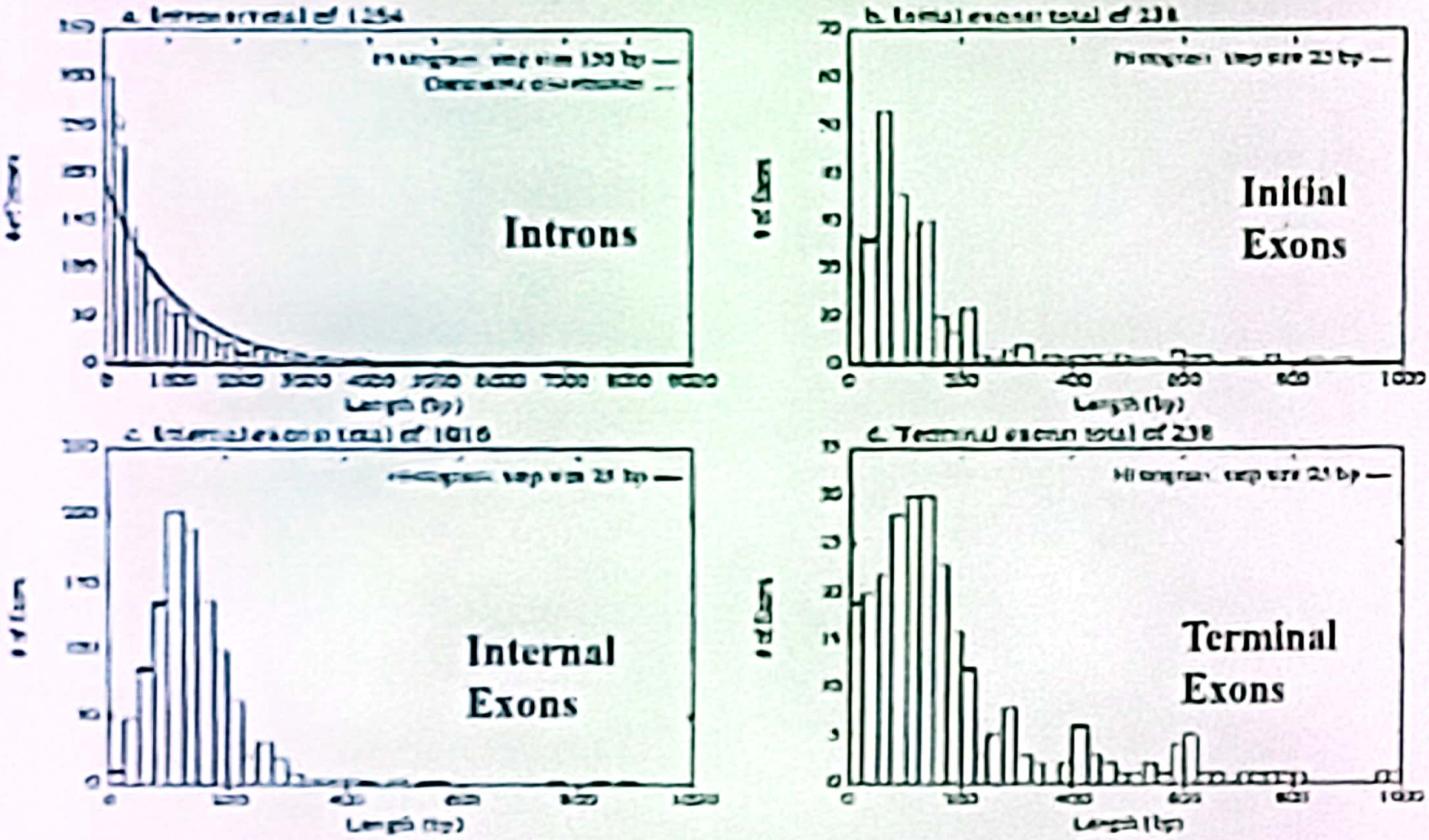
***Blood coagulation human factor VIII gene* is ~ 186Kb.**
- 26 exons with sizes varying from 69 bp to 3106 bp,
- 25 introns range in size from 207 to 32,400 bp.

On average 5′ UTR is 750bp long, but it can be longer and span several exons (e.g., in MAGE family).

On an average, 3′ UTR is ~ 450bp long, but for e.g., gene for Kallman's syndrome, the length exceeds 4Kb

Within intron 22 of human coagulation factor gene, there are two transcripts, one in the same orientation and another in the reverse.

Length distribution of human exons and introns



A large variation in the size of genes and exons observed in the eukaryotic genome – 238 multi-exon genes analysis shown

Some facts about human genes

- Comprises about 3% of the genome
- Average gene length: ~ 8,000 bp
- Average of 5 - 6 exons/gene
- Average exon length: ~ 200 bp
- Average intron length: ~ 2,000 bp
- ~ 8% genes have a single exon

Some exons can be as small as 1 or 3 bp

Complications in Gene Prediction

In higher eukaryotes gene finding becomes far more difficult because it is now necessary to combine multiple ORFs to obtain a spliced coding region.

Alternative splicing is not uncommon.

Variations in exon/intron lengths - exons can be very short, and introns can be very long.

Given the nature of genomic sequence in humans, where large introns are known to exist, there is a definite need for highly specific gene finding algorithms.