3. Genome Annotation: Gene Prediction

Gene Prediction: Computational Challenge

- Gene: A sequence of nucleotides coding for protein

- Gene Prediction Problem: Determine the beginning and end positions of genes in a genome

Gene Prediction: Computational Challenge

aatgcatgcggctatgctaatgcatgcggctatgctaagctgggatccgatgacaa
tgcatgcggctatgctaatgcatgcggctatgcaagctgggatccgatgactatgc
taagctgggatccgatgacaatgcatgcggctatgctaatgaatggtcttgggatt
taccttggaatgctaagctgggatccgatgacaatgcatgcggctatgctaatgaa
tggtcttgggatttaccttggaatatgctaatgcatgcggctatgctaagctggga
tccgatgacaatgcatgcggctatgctaatgcatgcggctatgcaagctgggatcc
gatgactatgctaagctgcggctatgctaatgcatgcggctatgctaagctgggat
ccgatgacaatgcatgcggctatgctaatgcatgcggctatgcaagctgggatcct
gcggctatgctaatgaatggtcttgggatttaccttggaatgctaagctgggatcc
gatgacaatgcatgcggctatgctaatgaatggtcttgggatttaccttggaatat
gctaatgcatgcggctatgctaagctgggaatgcatgcggctatgctaagctggga
tccgatgacaatgcatgcggctatgctaatgcatgcggctatgcaagctgggatcc
gatgactatgctaagctgcggctatgctaatgcatgcggctatgctaagctcatgc
ggctatgctaagctgggaatgcatgcggctatgctaagctgggatccgatgacaat
gcatgcggctatgctaatgcatgcggctatgcaagctgggatccgatgactatgct
aagctgcggctatgctaatgcatgcggctatgctaagctcggctatgctaatgaat
ggtcttgggatttaccttggaatgctaagctgggatccgatgacaatgcatgcggc
tatgctaatgaatggtcttgggatttaccttggaatatgctaatgcatgcggctat
gctaagctgggaatgcatgcggctatgctaagctgggatccgatgacaatgcatgc
ggctatgctaatgcatgcggctatgcaagctgggatccgatgactatgctaagctg
cggctatgctaatgcatgcggctatgctaagctcatgcgg

Gene Prediction: Computational Challenge

aatgcatgcggctatgctaatgcatgcggctatgctaagctgggatccgatgacaa
tgcatgcggctatgctaatgcatgcggctatgcaagctgggatccgatgactatgc
taagctgggatccgatgacaatgcatgcggctatgctaatgaatggtcttgggatt
taccttggaatgctaagctgggatccgatgacaatgcatgcggctatgctaatgaa
tggtcttgggatttaccttggaatatgctaatgcatgcggctatgctaagctggga
tccgatgacaatgcatgcggctatgctaatgcatgcggctatgcaagctgggatcc
gatgactatgctaagctgcggctatgctaatgcatgcggctatgctaagctgggat
ccgatgacaatgcatgcggctatgctaatgcatgcggctatgcaagctgggatcct
gcggctatgctaatgaatggtcttgggatttaccttggaatgctaagctgggatcc
gatgacaatgcatgcggctatgctaatgaatggtcttgggatttaccttggaatat
gctaatgcatgcggctatgctaagctgggaatgcatgcggctatgctaagctggga
tccgatgacaatgcatgcggctatgctaatgcatgcggctatgcaagctgggatcc
gatgactatgctaagctgcggctatgctaatgcatgcggctatgctaagctcatgc
ggctatgctaagctgggaatgcatgcggctatgctaagctgggatccgatgacaat
gcatgcggctatgctaatgcatgcggctatgcaagctgggatccgatgactatgct
aagctgcggctatgctaatgcatgcggctatgctaagctcggctatgctaatgaat
ggtcttgggatttaccttggaatgctaagctgggatccgatgacaatgcatgcggc
tatgctaatgaatggtcttgggatttaccttggaatatgctaatgcatgcggctat
gctaagctgggaatgcatgcggctatgctaagctgggatccgatgacaatgcatgc
ggctatgctaatgcatgcggctatgcaagctgggatccgatgactatgctaagctg
cggctatgctaatgcatgcggctatgctaagctcatgcgg

Gene Prediction: Computational Challenge

aatgcatgcggctatgctaatgcatgcggctatgctaagctgggatccgatgacaa
tgcatgcggctatgctaatgcatgcggctatgcaagctgggatccgatgactatgc
taagctgggatccgatgacaatgcatgcggctatgctaatgaatggtcttgggatt
taccttggaatgctaagctgggatccgatgacaatgcatgcggctatgctaatgaa
tggtcttgggatttaccttggaatatgctaatgcatgcggctatgctaagctggga
tccgatgacaatgcatgcggctatgctaatgcatgcggctatgcaagctgggatcc
gatgactatgctaagctgcggctatgctaatgcatgcggctatgctaagctgggat
ccgatgacaatgcatgcggctatgctaatgcatgcggctatgcaagctgggatcct
gcggctatgctaatgaatggtcttgggatttaccttggaatgctaagctgggatcc
gatgacaatgcatgcggctatgctaatgaatggtcttgggatttaccttggaatat
gctaatgcatgcggctatgctaagctgggaatgcatgcggctatgctaagctggga
tccgatgacaatgcatgcggctatgctaatgcatgcggctatgcaagctgggatcc
gatgactatgctaagctgcggctatgctaatgcatgcggctatgctaagctcatgc
ggctatgctaagctgggaatgcatgcggctatgctaagctgggatccgatgacaat
gcatgcggctatgctaatgcatgcggctatgcaagctgggatccgatgactatgct
aagctgcggctatgctaatgcatgcggctatgctaagctcggctatgctaatgaat
ggtcttgggatttaccttggaatgctaagctgggatccgatgacaatgcatgcggc
tatgctaatgaatggtcttgggatttaccttggaatatgctaatgcatgcggctat
gctaagctgggaatgcatgcggctatgctaagctgggatccgatgacaatgcatgc
ggctatgctaatgcatgcggctatgcaagctgggatccgatgactatgctaagctg
cggctatgctaatgcatgcggctatgctaagctcatgcgg

Gene!

Translating Nucleotides into Amino Acids

- Codon: 3 consecutive nucleotides
- $4^3 = 64$ possible codons
- Genetic code is degenerative and redundant
  - Includes start and stop codons
  - An amino acid may be coded by more than one codon

Codons

- In 1961 Sydney Brenner and Francis Crick discovered frameshift mutations
- Systematically deleted nucleotides from DNA
  - Single and double deletions dramatically altered protein product
  - Effects of triple deletions were minor
  - Conclusion: every triplet of nucleotides, each ***codon***, codes for exactly one amino acid in a protein

Triplet Phrase

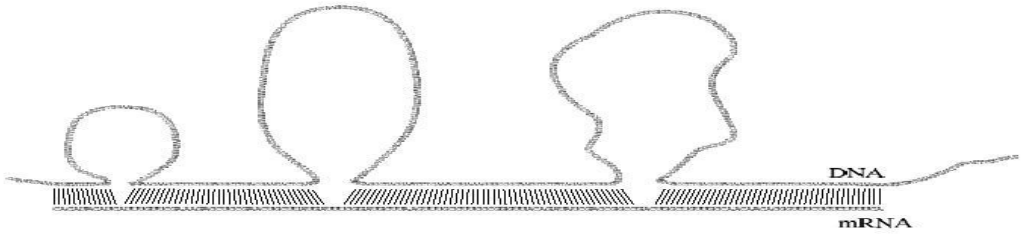
- THE SLY FOX AND THE SHY DOG
- THE SYF OXA NDT HES HYD OG
- THE SFO XAN DTH ESH HDO G

- But makes some sense after three deletions
- THE (SLY) FOX AND THE SHY DOG

Gene and Collinear Protein Products

- Inspired by Crick-Brenner experiment Yanofsky proved that that a gene and its protein product are collinear, that is, the first codon in the gene code for the first amino acid in the protein, second codon codes for the second amino acid etc.
- As a result, it was incorrectly assumed that the triplets encoding for amino acid sequences form contiguous strips of information.
- In 1977(Phillip Sharp and Richard Roberts), the discovery of split human genes proved that genes could be a collection of substrings.
- This also raised the computational problem of predicting the locations of genes in a genome which is just a string of DNA.

Discovery of Split Genes



- In 1977, Phillip Sharp and Richard Roberts experimented with mRNA encodes a viral protein called *hexon*,.
  - Map the hexon mRNA in viral genome, mRNA was hybridized to adenovirus DNA and the hybrid molecules were analyzed electron microscopy.
  - (Adenoviruses are double-stranded DNA viruses. They have icosahedral capsids with twelve vertices and seven surface proteins. The virion is non-enveloped, spherical and about seventy to ninety nm in size. The genome encodes about thirty proteins. Both strands of adenovirus DNA encode genes. Transcription occurs in three stages -- immediate early, early and late. It causes Acute Respiratory Disease , Pneumonia, Gasteroenteritis in human)
  - mRNA-DNA hybrids formed three curious loop structures instead of contiguous duplex segments

Discovery of Split Genes (cont'd)

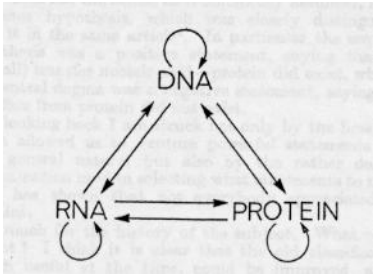
- "Adenovirus Amazes at Cold Spring Harbor" (1977, Nature 268) documented "mosaic molecules consisting of sequences complementary to several non-contiguous segments of the viral genome".
- In 1978 Walter Gilbert coined the term **intron** in the Nature paper "Why Genes in Pieces?"
- Genome size of many eukaryotes does not seem to be related to the organism's complexity (although genome size of prokaryotes like bacteria are much smaller in size). For example, the genome size of Salamandar fish is ten times larger than human genome. This is because the gene is intervened by large amount of "junk" DNA or intron.
- Further more, there are jumps between different parts of split genes and the corresponding gene segments may be organized in different ways, as for example in human and mouse genomes.

Exons and Introns

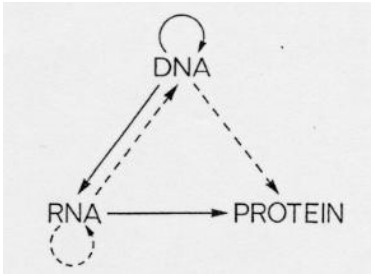
- In eukaryotes, the gene is a combination of coding segments (**exons**) that are interrupted by non-coding segments (**introns**)
- Prokaryotes don't have introns - Genes in prokaryotes are continuous. As a result, the gene detection algorithm is somewhat simpler than detecting genes in eukaryotes.
- Human genes constitute only less than 5% of the human genome. The rest is introns, the so-called "junk" DNA.
- This makes computational gene prediction in eukaryotes even more difficult

Central Dogma: Doubts

- Central Dogma was proposed in 1958 by Francis Crick
- Crick had very little supporting evidence in late 1950s
- Before Crick's seminal paper all possible information transfers were considered viable

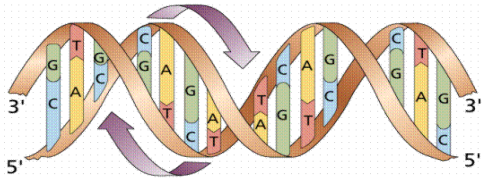


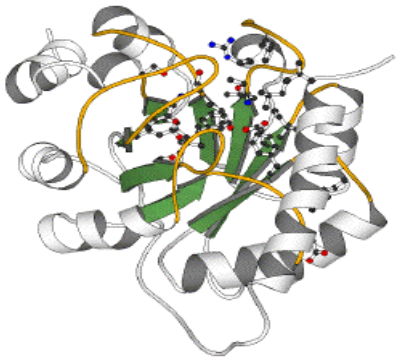
- Crick postulated that some of them are not viable (missing arrows)



- In 1970 Crick published a paper defending the Central Dogma.

Central Dogma: DNA -> RNA -> Protein





DNA

transcription

RNA

translation

Protein

CCTGAGCCAACTATTGATGAA

CCUGAGCCAACUAUUGAUGAA

PEPTIDE

Prokaryotic and eukaryotic organisms

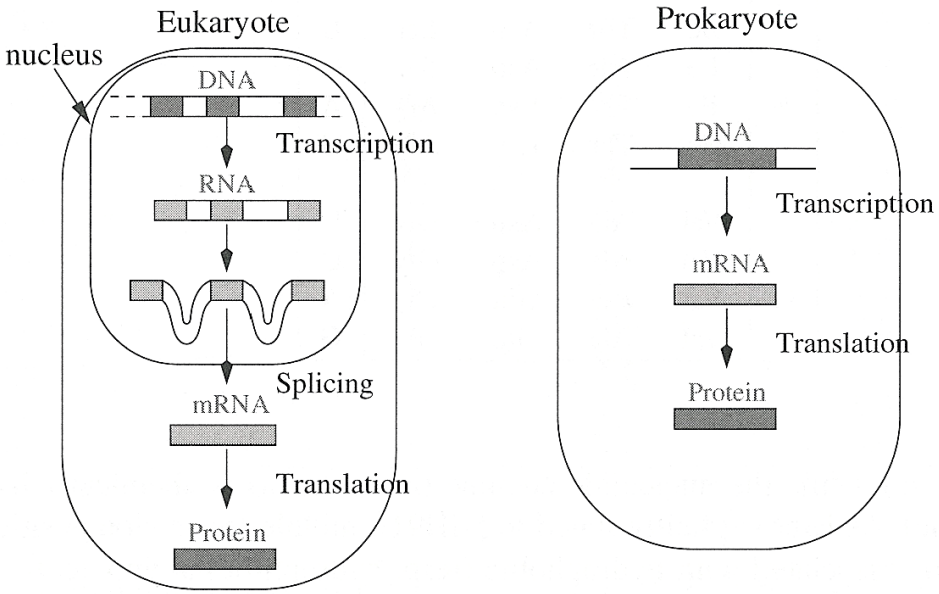
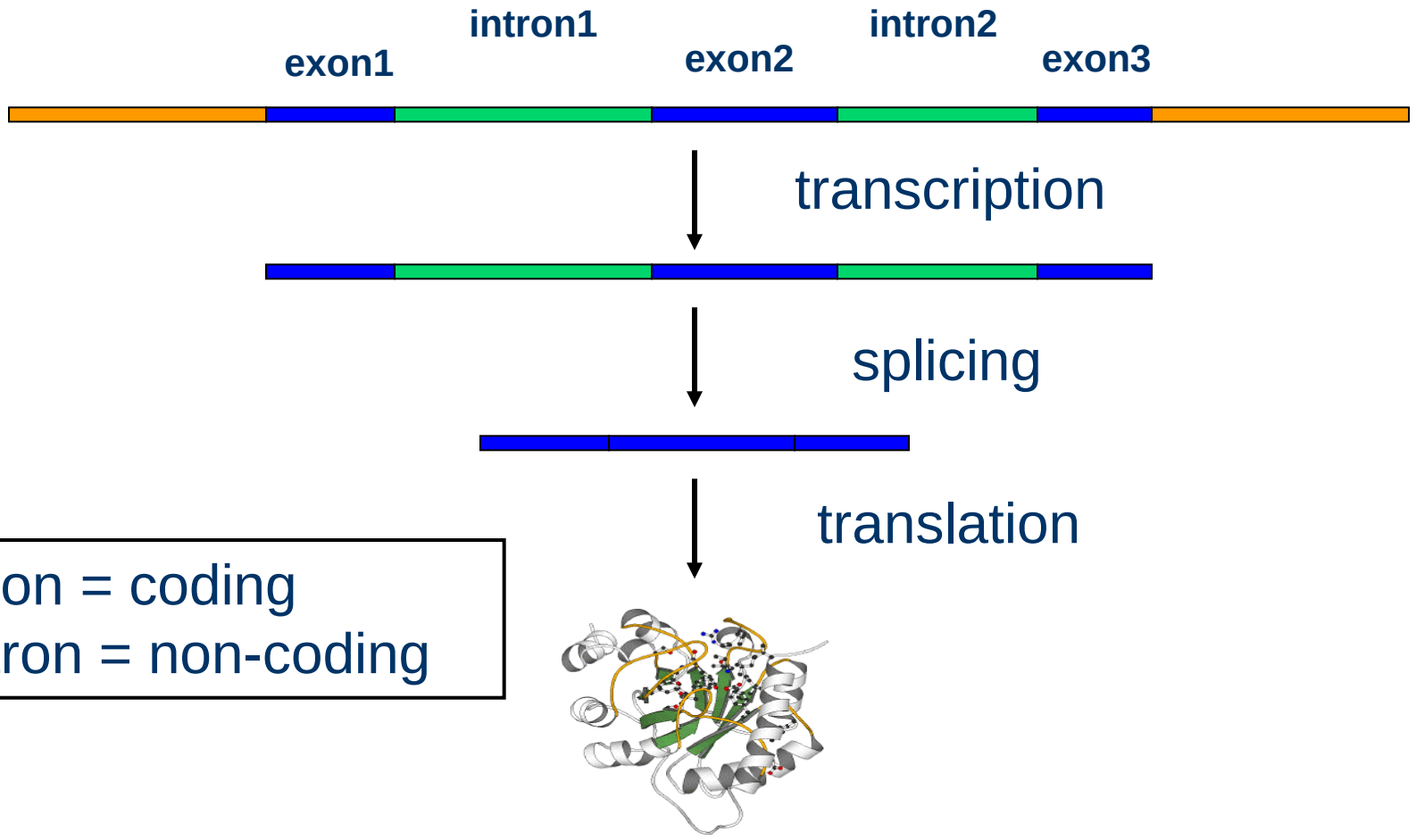


Figure 1.15 The different models of transcription and translation in prokaryotes and eukaryotes.

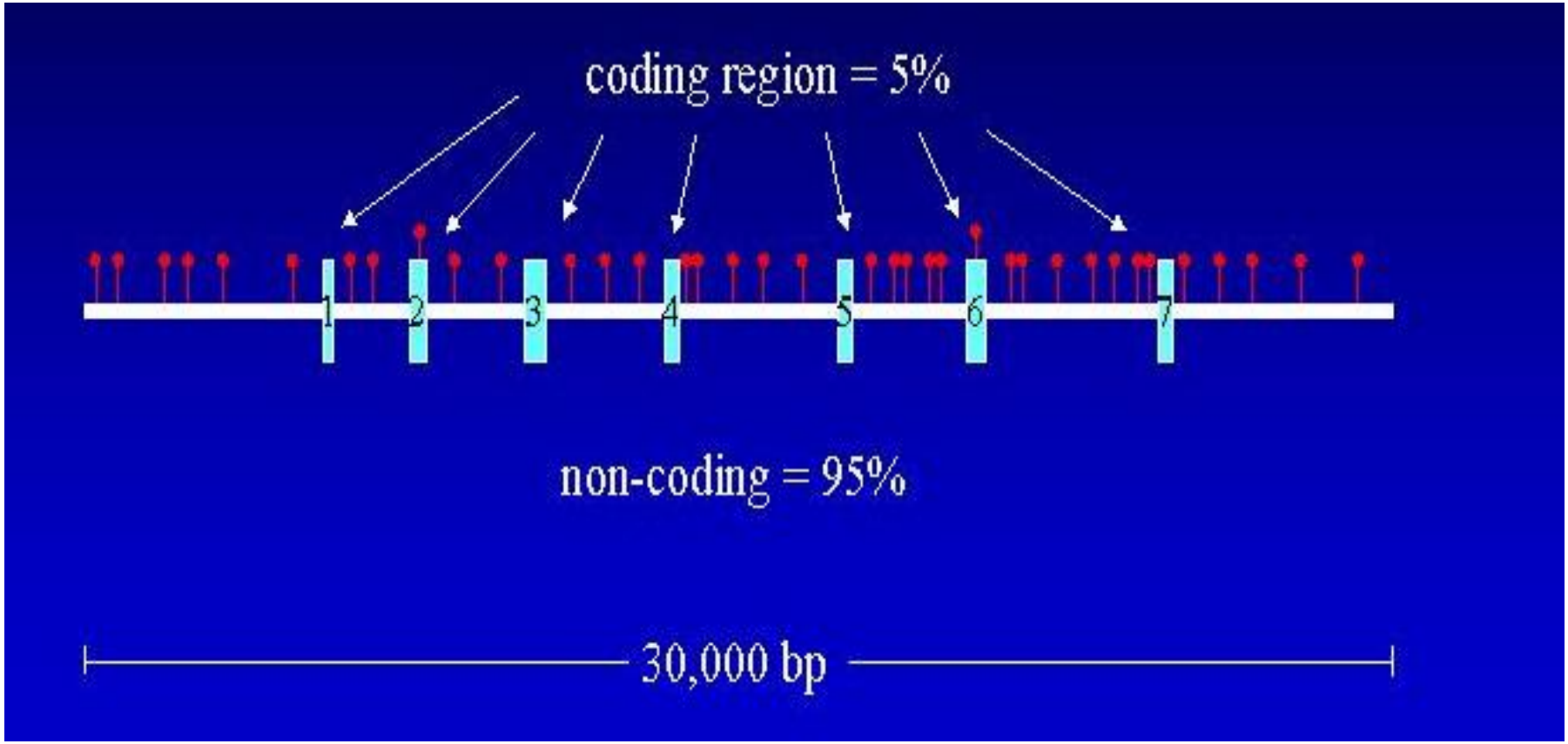
The genes themselves are structured in coding bits, that is the stuff that becomes amino acids, called exons, and non-coding stretches of sequence in between, called introns. When the gene is transcribed the whole thing becomes an RNA molecule, including the garbage in between the exons, and then these introns are cut out in a process called splicing.The resulting bits are glued together and translated into a protein

Central Dogma and Splicing



exon = coding
intron = non-coding

Gene Structure



Exons and Introns

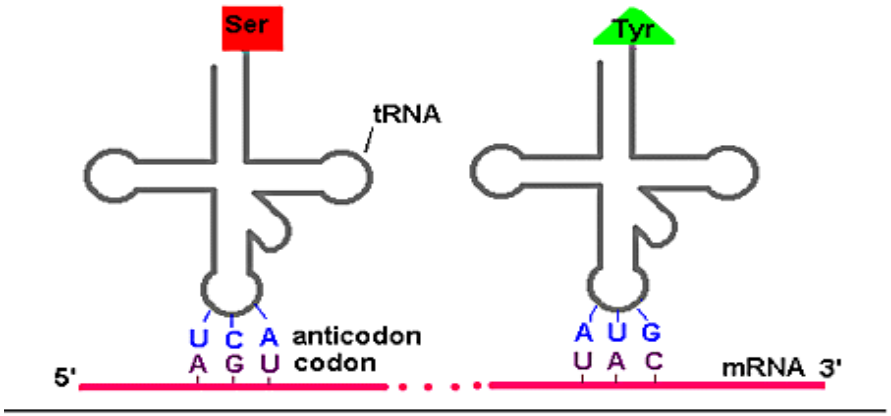
- In eukaryotes, the gene is a combination of coding segments (**exons**) that are interrupted by non-coding segments (**introns**)
- This makes computational gene prediction in eukaryotes even more difficult
- Prokaryotes don’t have introns - Genes in prokaryotes are continuous

Gene Finding Methods

There are two categories of approaches for predicting gene locations: The Statistical Approach and the Similarity Based Approach.

In statistical approach, splicing locations are detected based on *splicing signals* that appear frequently at the intron-exon boundaries such as the dinucleotides AG and GT , which have been highly preserved along with other less preserved signals. Thus, it should be possible to determine the “profiles” of the nucleotides near the boundaries. Such an approach is not very accurate because such profiles are weak and may appear in non-splice sites. So, a better approach based on “codon” frequencies” have been developed.

Genetic Code and Stop Codons



2nd base in codon

1st base in codon	U	C	A	G	3rd base in codon
U	Phe	Ser	Tyr	Cys	U
	Phe	Ser	Tyr	Cys	C
	Leu	Ser	STOP	STOP	A
	Leu	Ser	STOP	Trp	G
C	Leu	Pro	His	Arg	U
	Leu	Pro	His	Arg	C
	Leu	Pro	Gln	Arg	A
	Leu	Pro	Gln	Arg	G
A	Ile	Thr	Asn	Ser	U
	Ile	Thr	Asn	Ser	C
	Ile	Thr	Lys	Arg	A
	Met	Thr	Lys	Arg	G
G	Val	Ala	Asp	Gly	U
	Val	Ala	Asp	Gly	C
	Val	Ala	Glu	Gly	A
	Val	Ala	Glu	Gly	G

The Genetic Code

UAA, UAG and UGA correspond to 3 Stop codons that (together with Start codon ATG) delineate Open Reading Frames

Six Frames in a DNA Sequence

- stop codons – TAA, TAG, TGA
- start codons - ATG

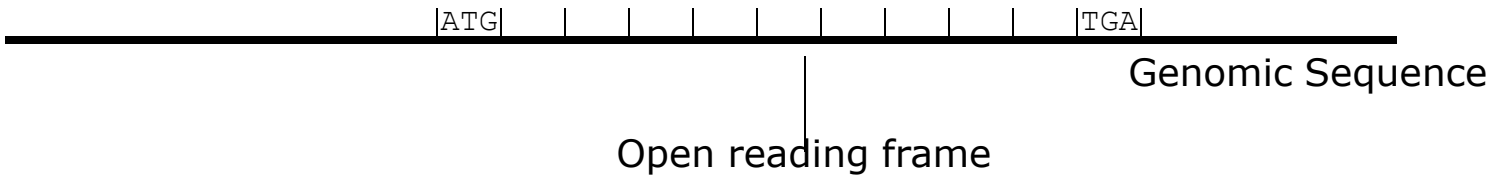
$\frac{n}{3}$

Open Reading Frames (ORFs)

Open Reading Frame (ORF) is a sequence of codons which starts with start codon, ends with an end codon and has no end codons in-between.

Searching for ORFs – consider all 6 possible reading frames: 3 forward and 3 reverse

- Detect potential coding regions by looking at **ORFs**
  - A genome of length *n* is comprised of (*n*/3) codons
  - Stop codons break genome into segments between consecutive Stop codons
  - The subsegments of these that start from the Start codon (ATG) are ORFs
    - ORFs in different frames may overlap



Long vs.Short ORFs

- Long open reading frames may be a gene
  - At random, we should expect one stop codon every (64/3) ~= 21 codons
  - **However**, genes are usually much longer than this
- A basic approach is to scan for ORFs whose length exceeds certain threshold
  - This is naïve because some genes (e.g. some neural and immune system genes) are relatively short

Testing ORFs: Codon Usage

- Create a 64-element hash table and count the frequencies of codons in an ORF
- Amino acids typically have more than one codon, but in nature certain codons are more in use
- Uneven use of the codons may characterize a real gene
- This compensate for pitfalls of the ORF length test

Codon Usage in Human Genome

	U		C		A		G	
U	UUU Phe	57	UCU Ser	16	UAU Tyr	58	UGU Cys	45
	UUC Phe	43	UCC Ser	15	UAC Tyr	42	UGC Cys	55
	UUA Leu	13	UCA Ser	13	UAA Stp	62	UGA Stp	30
	UUG Leu	13	UCG Ser	15	UAG Stp	8	UGG Trp	100
C	CUU Leu	11	CCU Pro	17	CAU His	57	CGU Arg	37
	CUC Leu	10	CCC Pro	17	CAC His	43	CGC Arg	38
	CUA Leu	4	CCA Pro	20	CAA Gln	45	CGA Arg	7
	CUG Leu	49	CCG Pro	51	CAG Gln	66	CGG Arg	10
A	AUU Ile	50	ACU Thr	18	AAU Asn	46	AGU Ser	15
	AUC Ile	41	ACC Thr	42	AAC Asn	54	AGC Ser	26
	AUA Ile	9	ACA Thr	15	AAA Lys	75	AGA Arg	5
	AUG Met	100	ACG Thr	26	AAG Lys	25	AGG Arg	3
G	GUU Val	27	GCU Ala	17	GAU Asp	63	GGU Gly	34
	GUC Val	21	GCC Ala	27	GAC Asp	37	GGC Gly	39
	GUA Val	16	GCA Ala	22	GAA Glu	68	GGA Gly	12
	GUG Val	36	GCG Ala	34	GAG Glu	32	GGG Gly	15

Codon Usage in Mouse Genome

AA	codon	/1000	frac
Ser	TCG	4.31	0.05
Ser	TCA	11.44	0.14
Ser	TCT	15.70	0.19
Ser	TCC	17.92	0.22
Ser	AGT	12.25	0.15
Ser	AGC	19.54	0.24
Pro	CCG	6.33	0.11
Pro	CCA	17.10	0.28
Pro	CCT	18.31	0.30
Pro	CCC	18.42	0.31

AA	codon	/1000	frac
Leu	CTG	39.95	0.40
Leu	CTA	7.89	0.08
Leu	CTT	12.97	0.13
Leu	CTC	20.04	0.20
Ala	GCG	6.72	0.10
Ala	GCA	15.80	0.23
Ala	GCT	20.12	0.29
Ala	GCC	26.51	0.38
Gln	CAG	34.18	0.75
Gln	CAA	11.51	0.25

Codon Usage and Likelihood Ratio

- An ORF is more “believable” than another if it has more “likely” codons
- Do sliding window calculations to find ORFs that have the “likely” codon usage
- Allows for higher precision in identifying true ORFs; much better than merely testing for length.
- However, average vertebrate exon length is 130 nucleotides, which is often too small to produce reliable peaks in the likelihood ratio
- Further improvement: **in-frame hexamer count** (frequencies of pairs of consecutive codons)

Microbial gene finding

- Microbial genome tends to be gene rich (80%-90% of the sequence is coding)
- The most reliable method – homology searches (e.g. using BLAST and/or FASTA)
- Major problem – finding genes without known homologue.

Open Reading Frame

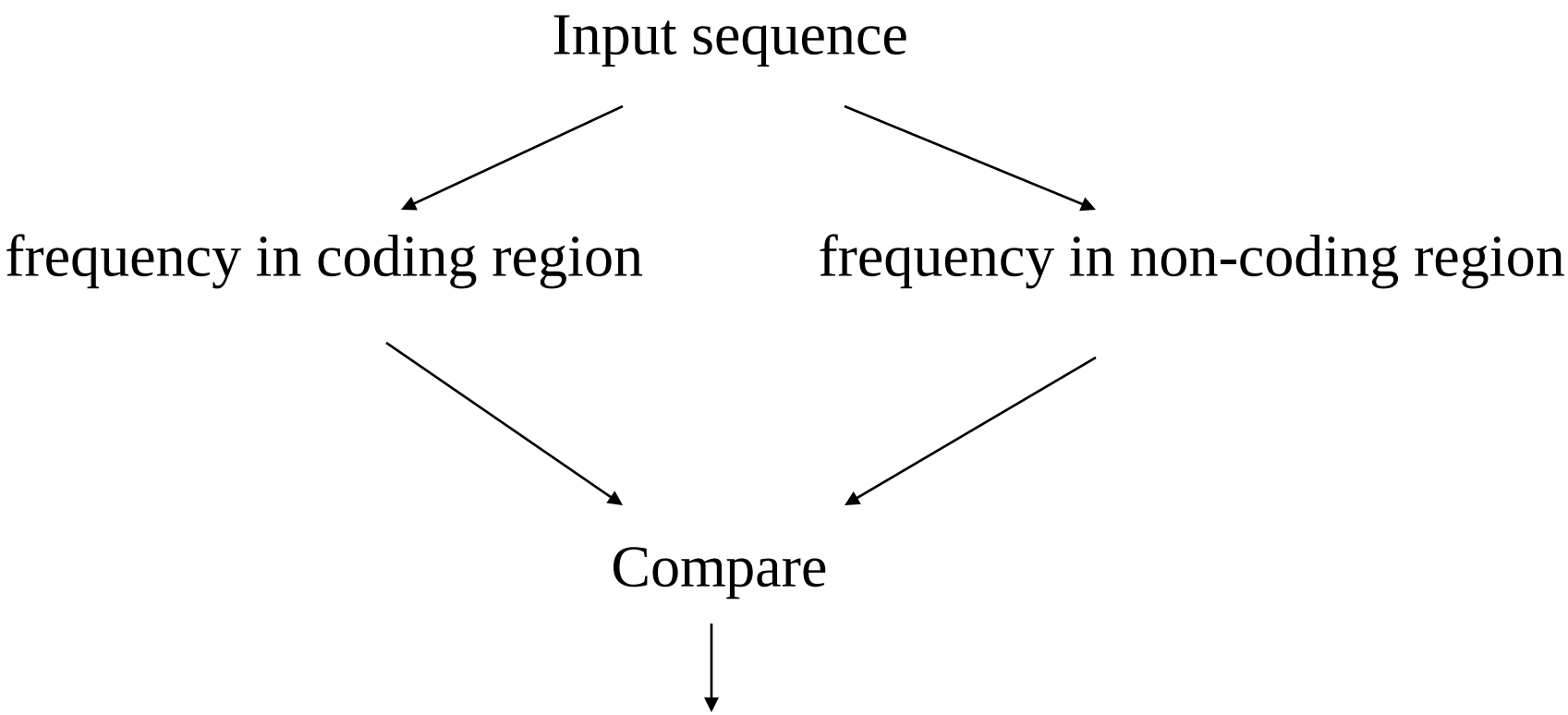
Open Reading Frame (ORF) is a sequence of codons which starts with start codon, ends with an end codon and has no end codons in-between.

Searching for ORFs – consider all 6 possible reading frames: 3 forward and 3 reverse

Is the ORF a coding sequence?

1. **Must be long enough (roughly 300 bp or more)**
2. **Should have average amino-acid composition specific for a give organism.**
3. **Should have codon use specific for the given organism.**

Codon frequency



Example

codon position	A	C	T	G
1	28%	33%	18%	21%
2	32%	16%	20%	32%
3	33%	15%	14%	38%
freqency in genome	31%	21.3 %	17.3 %	30.3 %

Assume: bases making codon are independent

$$\frac{P(x|\text{in coding})}{P(x|\text{random})} = \prod_i \frac{P(Ai \text{ at ith position})}{P(Ai \text{ in the sequence})}$$

Score of AAAGAT:

$$\frac{.28*.32*.33*.21*.\quad 32*.14}{.31*.31*.31*.30.3*.31*.17.3}$$

Using codon frequency to find correct reading frame

Consider sequence $x_1\, x_2\, x_3\, x_4\, x_5\, x_6\, x_7\, x_8\, x_9 \ldots$
where x_i is a nucleotide

let $p_1 = p_{x1\, x2\, x3}\, p_{x3\, x4\, x5} \cdots$
$p_2 = p_{x2\, x3\, x4}\, p_{x5\, x6\, x7} \cdots$
$p_3 = p_{x3\, x4\, x5}\, p_{x6\, x7\, x8} \cdots$

Similar analysis can be done for **in-frame hexamer count** (frequencies of pairs of consecutive codons)

then probability that ith reading frame is the coding frame is:

$$P_i = \frac{p_i}{p_1 + p_2 + p_3}$$

Algorithm:

- slide a window along the sequence and compute P_i
- Plot the results

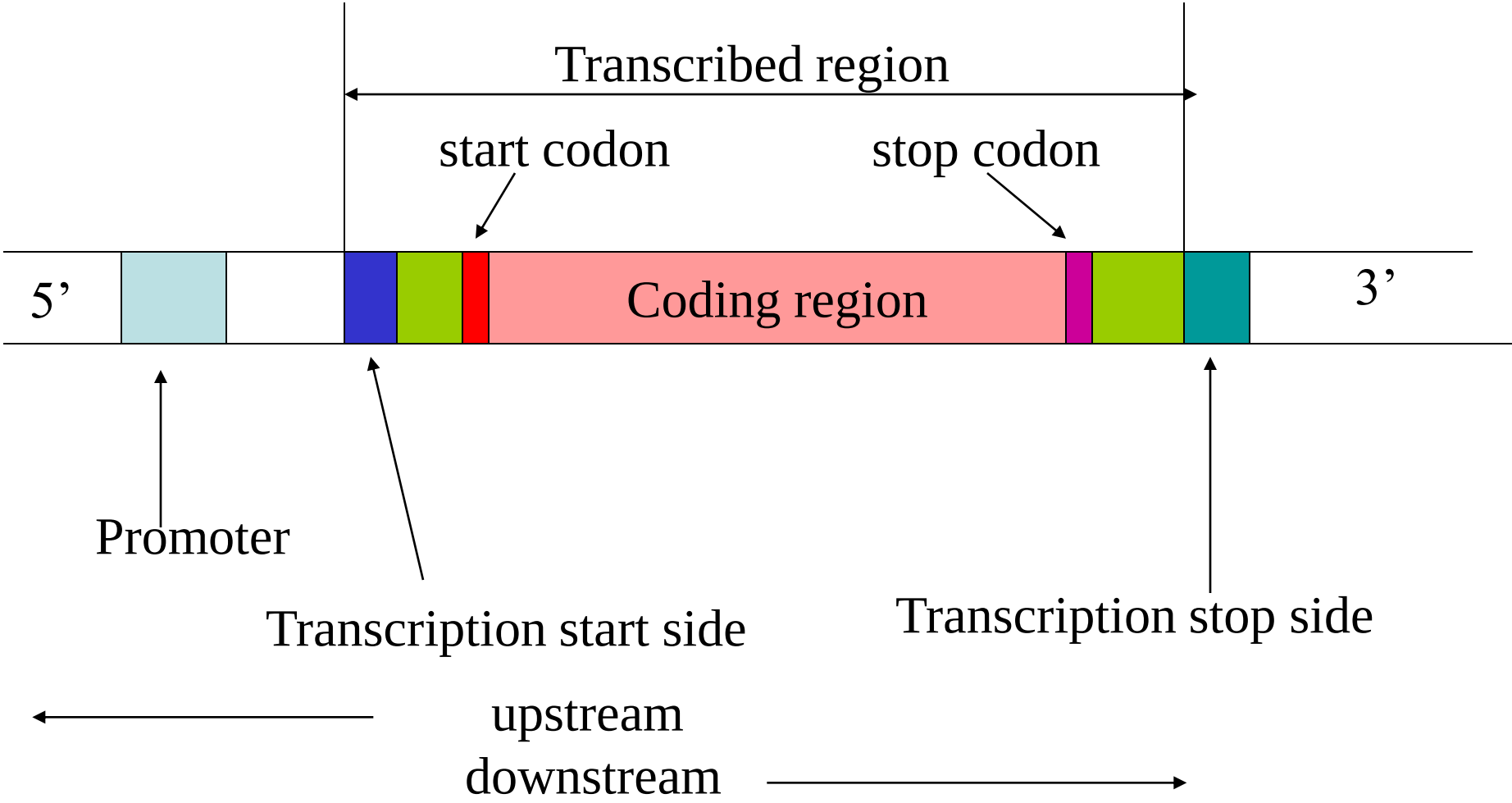
Bacterial Genome

Bacterial genomes have several conserved motifs often found near the start of the transcription regions. But , such motifs are not usually easy to find for eukaryotes.

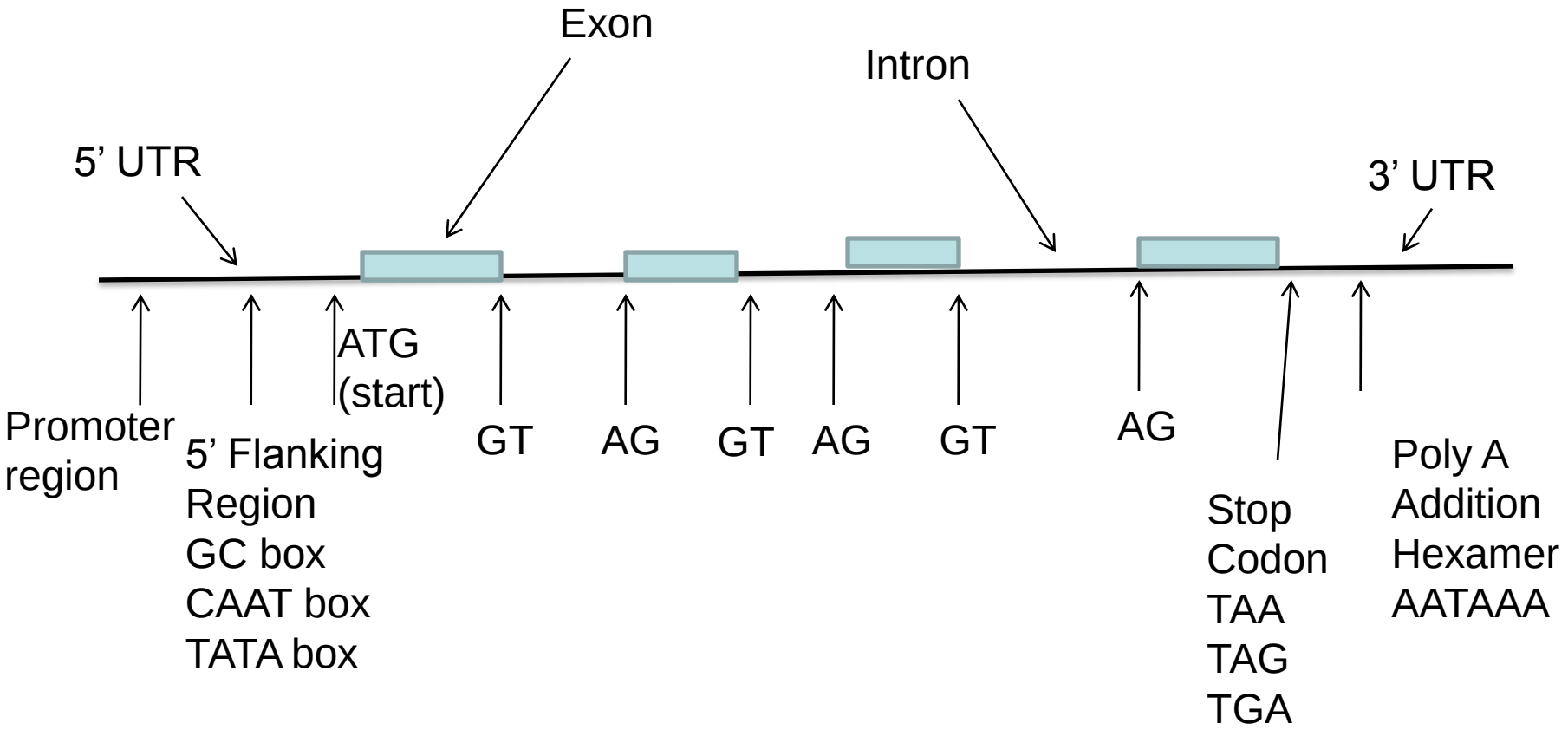
Also for eukaryotes , the average length of exons in vertebrates
Is 130 nucleotides. Exons of this length does not give reliable peaks in likelihood ratio.

There are also special structures present at the exon-intron interface which biologists have recognized. These are called “splicing signals”. There are some weakly conserved sequence of 8 nucleotides at the boundary of exon –intron, called **donor site,** and a sequence of 4 nucleotides called **acceptor** site. Since these splice sites are weak, these produced only limited success. New HMM(Hidden Markov Model) based methods are now currently used, GENSCAN.

Transcription in prokaryotes



Eukaryotic Gene Structure

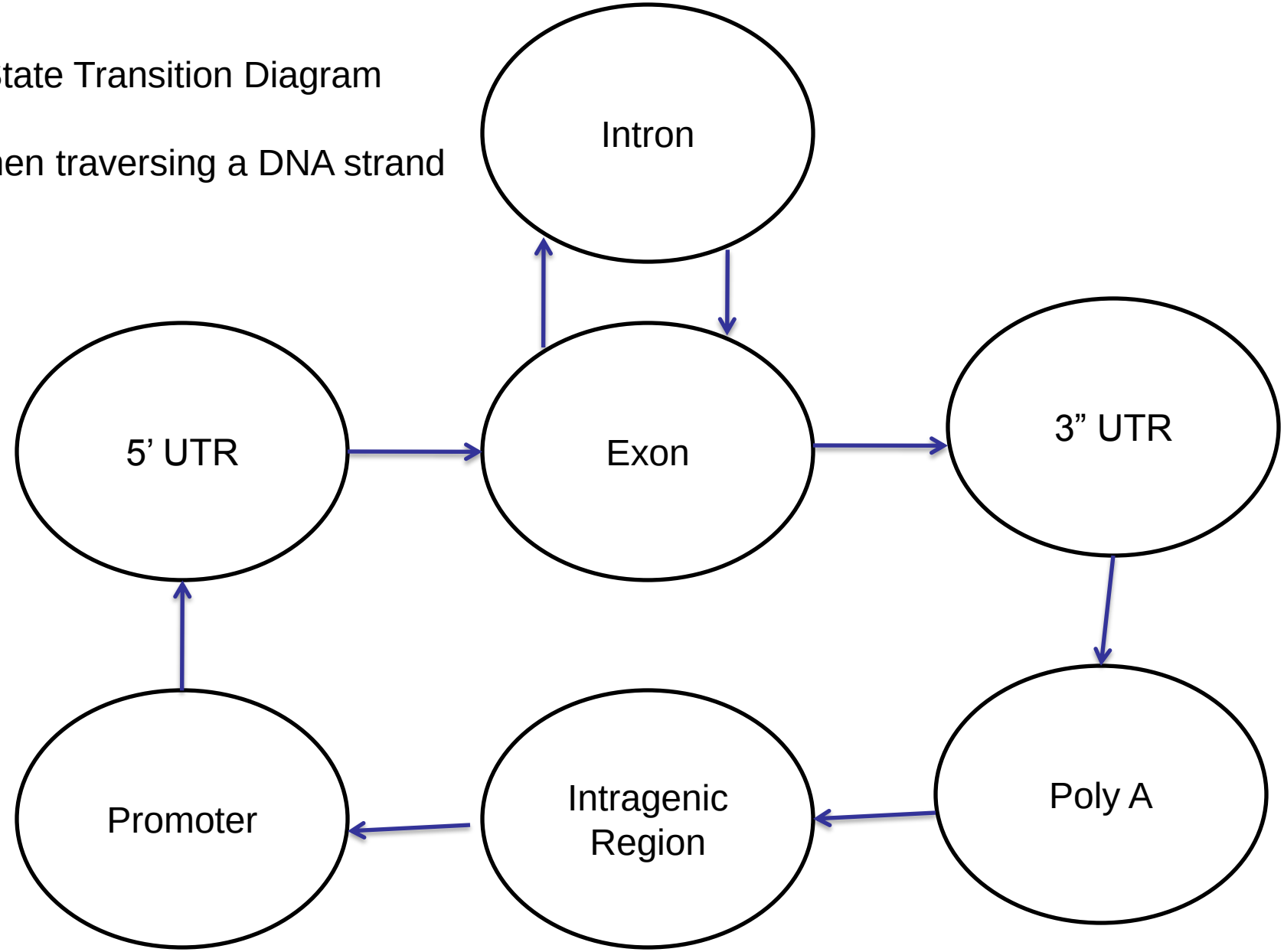


Transcription from DNA to RNA proceeds along 5’UTR to 3”UTR direction

A possible Hidden Markov Model

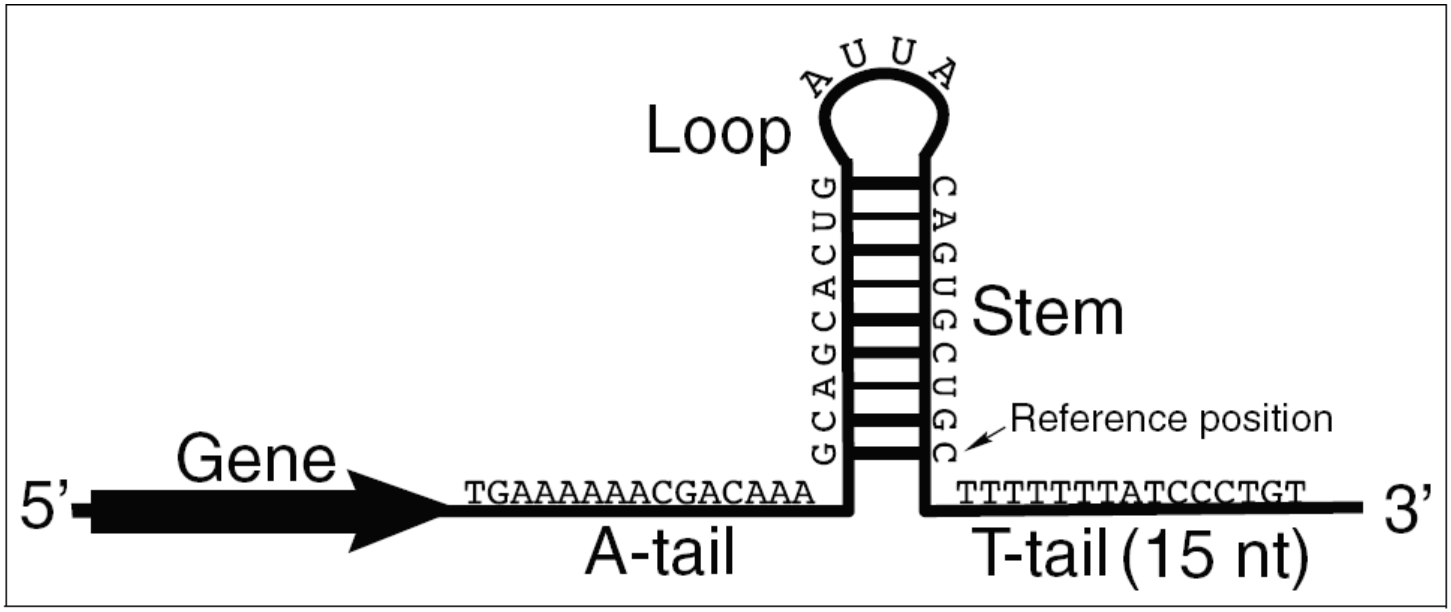
: State Transition Diagram

when traversing a DNA strand



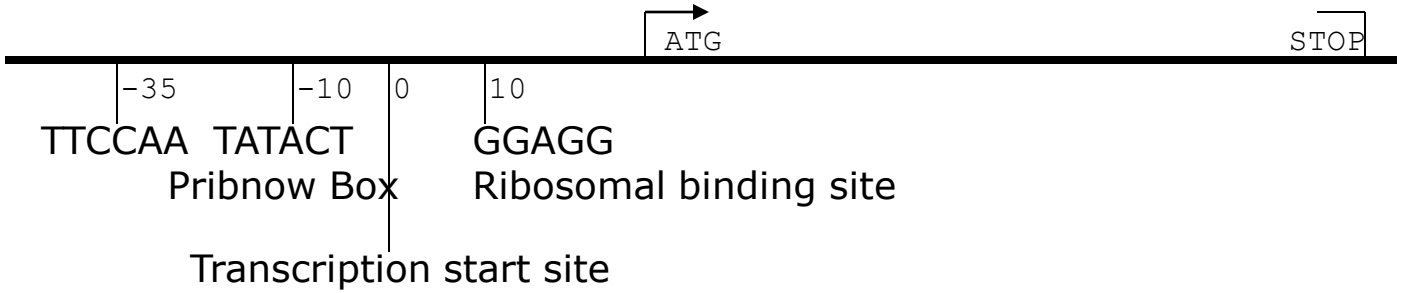
Recognition of gene related signals

- Promoter: bacteria TATAAT – Pribnow box at about –10;
- Terminator: Rho-independent (intrinsic) transcription termination – G-C reach inverted repeat.
- Other signal recognition –e.g binding motifs



Gene Prediction and Motifs

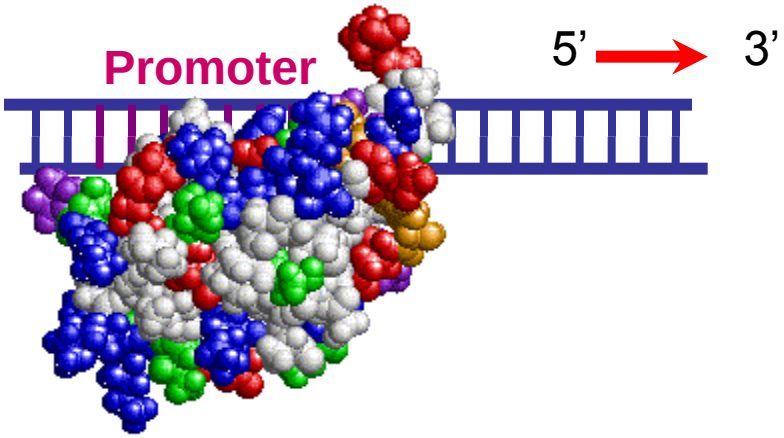
- Upstream regions of genes often contain motifs that can be used for gene prediction



- -k denotes k^{th} base before transcription, +k denotes k^{th} transcribed base

Promoters

- Promoters are DNA segments upstream of transcripts that initiate transcription

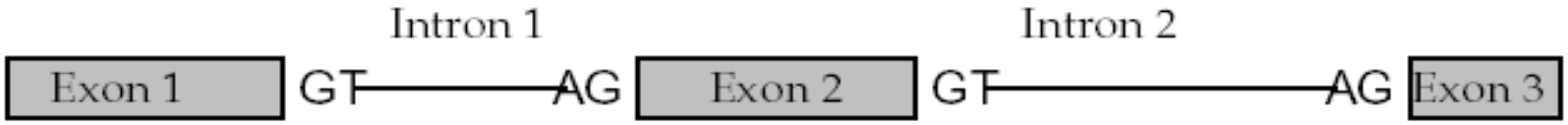


- Promoter *attracts* RNA Polymerase to the transcription start site

Splicing Signals

- Try to recognize location of splicing signals at exon-intron junctions
  - This has yielded a weakly conserved donor splice site and acceptor splice site
- Profiles for sites are still weak, and lends the problem to the Hidden Markov Model (HMM) approaches, which capture the statistical dependencies between sites

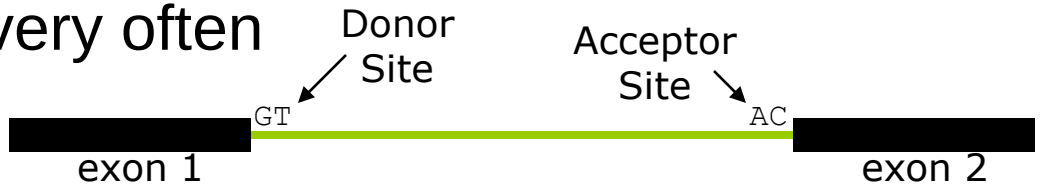
Splicing Signals



Exons are interspersed with introns and typically flanked by GT and AG

Donor and Acceptor Sites: GT and AG dinucleotides

- The beginning and end of exons are signaled by donor and acceptor sites that usually have GT and AC dinucleotides
- Detecting these sites is difficult, because GT and AC appear very often



Popular Gene Prediction Algorithms

- **GENSCAN**: uses Hidden Markov Models (HMMs)
- **TWINSCAN**
  - Uses both HMM and similarity (e.g., between human and mouse genomes)