
Gene Prediction: Statistical Approaches

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

aatgcatgctggctatgctaagcatgctggctatgctaagctgggatccgatgacaatgcatgctggctatgct
aatgcatgctggctatgcaagctgggatccgatgactatgctaagctgggatccgatgacaatgcatgctggc
tatgctaataatggtcttgggatttaccttggaaatgctaagctgggatccgatgacaatgcatgctggctatgc
taatgaatggtcttgggatttaccttggaaatgctaagcatgctggctatgctaagctgggatccgatgacaa
tgcatgctggctatgctaagcatgctggctatgcaagctgggatccgatgactatgctaagctgctggctatgc
taatgcatgctggctatgctaagctgggatccgatgacaatgcatgctggctatgctaagcatgctggctatgc
aagctgggatccgctgctggctatgctaataatggtcttgggatttaccttggaaatgctaagctgggatccgatg
acaatgcatgctggctatgctaataatggtcttgggatttaccttggaaatgctaagcatgctggctatgcta
agctgggaatgcatgctggctatgctaagctgggatccgatgacaatgcatgctggctatgctaagcatgctg
gctatgcaagctgggatccgatgactatgctaagctgctggctatgctaagcatgctggctatgctaagctcat
gctggctatgctaagctgggaatgcatgctggctatgctaagctgggatccgatgacaatgcatgctggctatg
ctaagcatgctggctatgcaagctgggatccgatgactatgctaagctgctggctatgctaagcatgctggct
atgctaagctgctggctatgctaataatggtcttgggatttaccttggaaatgctaagctgggatccgatgacaat
gcatgctggctatgctaataatggtcttgggatttaccttggaaatgctaagcatgctggctatgctaagctg
ggaatgcatgctggctatgctaagctgggatccgatgacaatgcatgctggctatgctaagcatgctggctat
gcaagctgggatccgatgactatgctaagctgctggctatgctaagcatgctggctatgctaagctcatgctgg

Gene Prediction: Computational Challenge

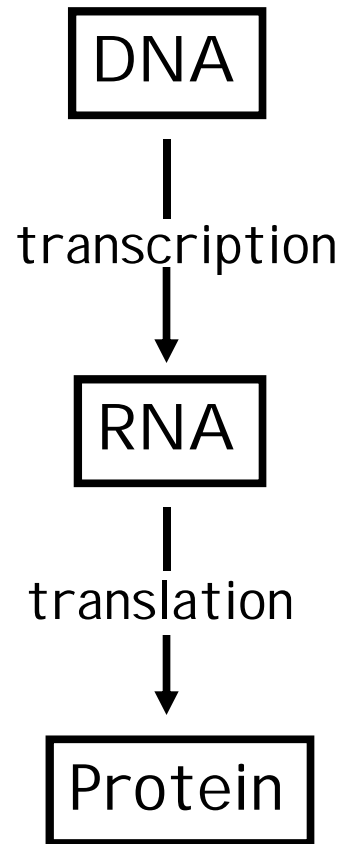
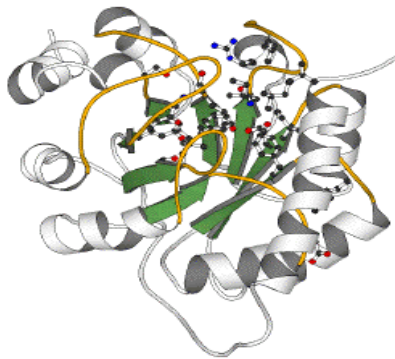
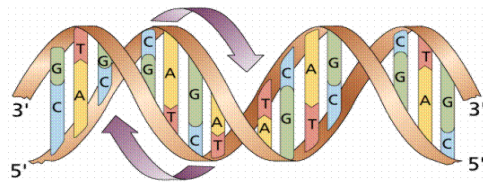
aatgcatgctggctatgctaagcatgctggctatgctaagctgggatccgatgacaatgcatgctggctatgct
aatgcatgctggctatgcaagctgggatccgatgactatgctaagctgggatccgatgacaatgcatgctggc
tatgctaataatggtcttgggatttaccttgggaatgctaagctgggatccgatgacaatgcatgctggctatgc
taatgaatggtcttgggatttaccttgggaatgctaagcatgctggctatgctaagctgggatccgatgacaa
tgcatgctggctatgctaagcatgctggctatgcaagctgggatccgatgactatgctaagctgctggctatgc
taatgcatgctggctatgctaagctgggatccgatgacaatgcatgctggctatgctaagcatgctggctatgc
aagctgggatccgctggctatgctaataatggtcttgggatttaccttgggaatgctaagctgggatccgatg
acaatgcatgctggctatgctaataatggtcttgggatttaccttgggaatgctaagcatgctggctatgcta
agctgggaatgcatgctggctatgctaagctgggatccgatgacaatgcatgctggctatgctaagcatgctg
gctatgcaagctgggatccgatgactatgctaagctgctggctatgctaagcatgctggctatgctaagctcat
gctggctatgctaagctgggaatgcatgctggctatgctaagctgggatccgatgacaatgcatgctggctatg
ctaagcatgctggctatgcaagctgggatccgatgactatgctaagctgctggctatgctaagcatgctggct
atgctaagctgctggctatgctaataatggtcttgggatttaccttgggaatgctaagctgggatccgatgacaat
gcatgctggctatgctaataatggtcttgggatttaccttgggaatgctaagcatgctggctatgctaagctg
ggaatgcatgctggctatgctaagctgggatccgatgacaatgcatgctggctatgctaagcatgctggctat
gcaagctgggatccgatgactatgctaagctgctggctatgctaagcatgctggctatgctaagctcatgctgg

Gene Prediction: Computational Challenge

aatgcatgCGGctatgctaataatgcatgCGGctatgctaagctgggatccgatgacaataatgcatgCGGctatgct
aatgcatgCGGctatgcaagctgggatccgatgactatgctaagctgggatccgatgacaataatgcatgCGGc
tatgctaataatgaatggtcttgggatttaccttgggaatgctaagctgggatccgatgacaataatgcatgCGGctatgc
taatgaatggtcttgggatttaccttgggaatgctaataatgcatgCGGctatgctaagctgggatccgatgacaa
tgcatgCGGctatgctaataatgcatgCGGctatgcaagctgggatccgatgactatgctaagctgCGGctatgc
taatgcatgCGGctatgctaagctgggatccgatgacaataatgcatgCGGctatgctaataatgcatgCGGctatgc
aagctgggatccgctgCGGctatgctaataatgaatggtcttgggatttaccttgggaatgctaagctgggatccgatg
acaataatgcatgCGGctatgctaataatgaatggtcttgggatttaccttgggaatgctaataatgcatgCGGctatgcta
agctgggaatgcatgCGGctatgctaagctgggatccgatgacaataatgcatgCGGctatgctaataatgcatgCG
gctatgcaagctgggatccgatgactatgctaagctgCGGctatgctaataatgcatgCGGctatgctaagctcat
gCGGctatgctaagctgggaatgcatgCGGctatgctaagctgggatccgatgacaataatgcatgCGGctatg
ctaataatgcatgCGGctatgcaagctgggatccgatgactatgctaagctgCGGctatgctaataatgcatgCGGct
atgctaagctcggctatgctaataatgaatggtcttgggatttaccttgggaatgctaagctgggatccgatgacaat
gcatgCGGctatgctaataatgaatggtcttgggatttaccttgggaatgctaataatgcatgCGGctatgctaagctg
ggaatgcatgCGGctatgctaagctgggatccgatgacaataatgcatgCGGctatgctaataatgcatgCGGctat
gcaagctgggatccgatgactatgctaagctgCGGctatgctaataatgcatgCGGctatgctaagctcatgCGG

Gene!

Central Dogma: DNA -> RNA -> Protein



CCTGAGCCAACTATTGATGAA



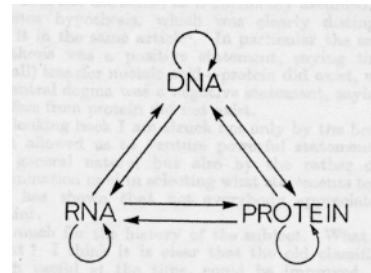
CCUGAGCCAAACUAUUGAUGAA



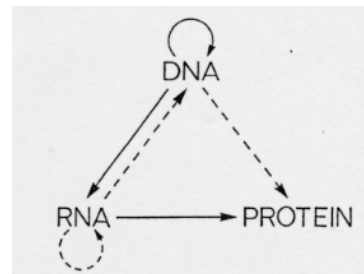
PEPTIDE

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.

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

The Sly Fox

- In the following string

THE SLY FOX AND THE SHY DOG

- Delete 1, 2, and 3 nucleotides after the first 'S':

THE SYF OXA NDT HES HYD OG

THE SFO XAN DTH ESH YDO G

THE SOX AND THE SHY DOG

- Which of the above makes the most sense?

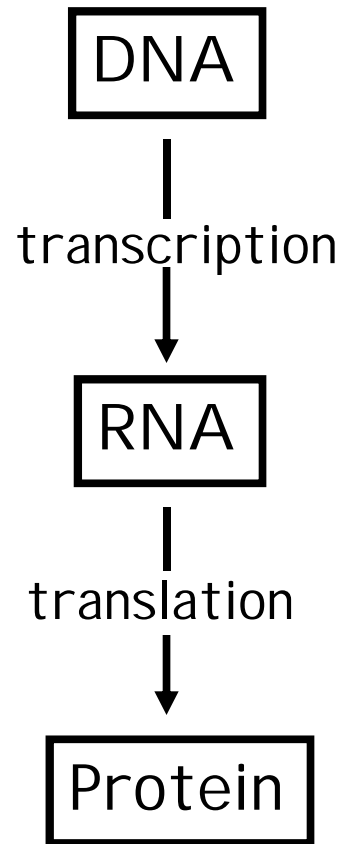
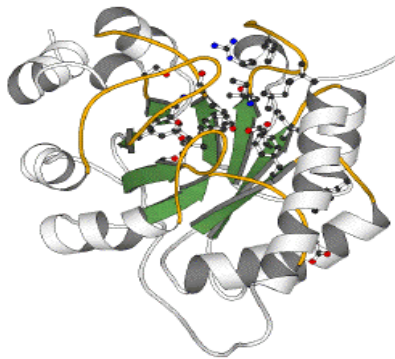
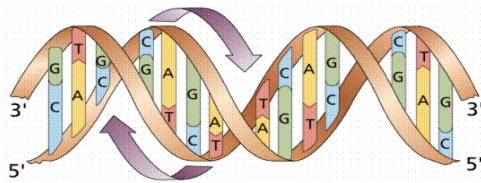
Translating Nucleotides into Amino Acids

- 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

Great Discovery Provoking Wrong Assumption

- In 1964, Charles Yanofsky and Sydney Brenner proved colinearity in the order of codons with respect to amino acids in proteins
- In 1967, Yanofsky and colleagues further proved that the sequence of codons in a gene determines the sequence of amino acids in a protein
- As a result, it was **incorrectly** assumed that the triplets encoding for amino acid sequences form contiguous strips of information.

Central Dogma: DNA -> RNA -> Protein



CCTGAGCCAACTATTGATGAA



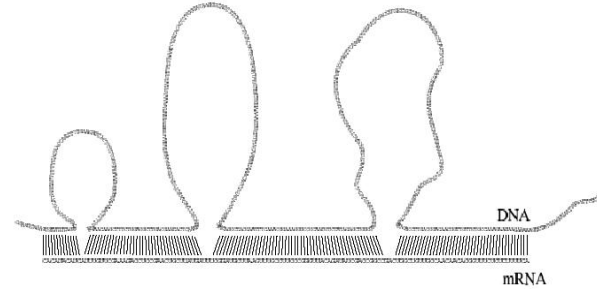
CCUGAGCCAAUUAUUGAUGAA



PEPTIDE

Discovery of Split Genes

- In 1977, Phillip Sharp and Richard Roberts experimented with mRNA of *hexon*, a viral protein.
 - Map **hexon mRNA** in viral genome by hybridization to **adenovirus DNA** and electron microscopy
 - mRNA-DNA hybrids formed three **curious loop structures** instead of contiguous duplex segments



Discovery of Split Genes (cont'd)

103

- “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?”

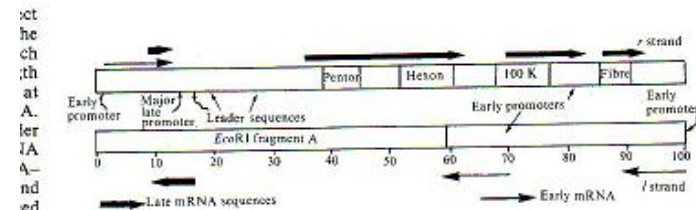


Fig. 1 Transcription map of adenovirus 2 (see Flint Cell 10, 153; 1977).

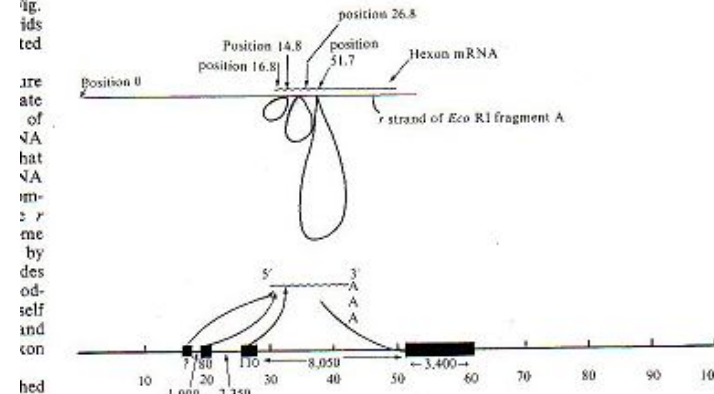


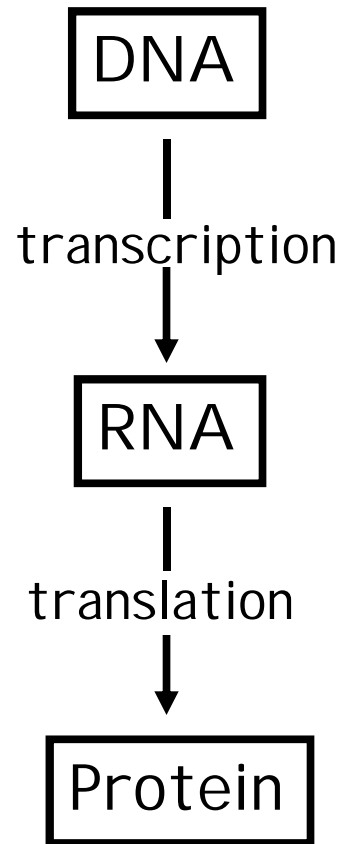
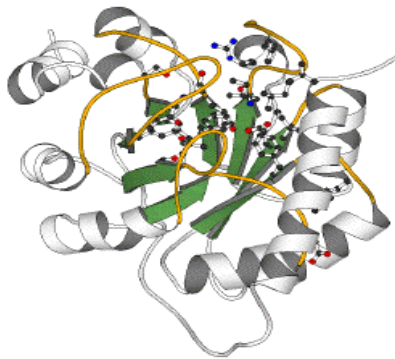
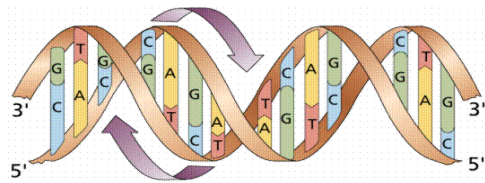
Fig. 2 a. Pattern of hybridisation between hexon mRNA and the r strand of EcoRI fragment A of adenovirus 2 DNA. b. Regions of adenovirus genome which contribute to hexon mRNA. Figures other than adenovirus DNA markers represent distances in nucleotide base pairs.

late precursor.

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

Central Dogma: DNA -> RNA -> Protein



CCTGAGCCAACTATTGATGAA

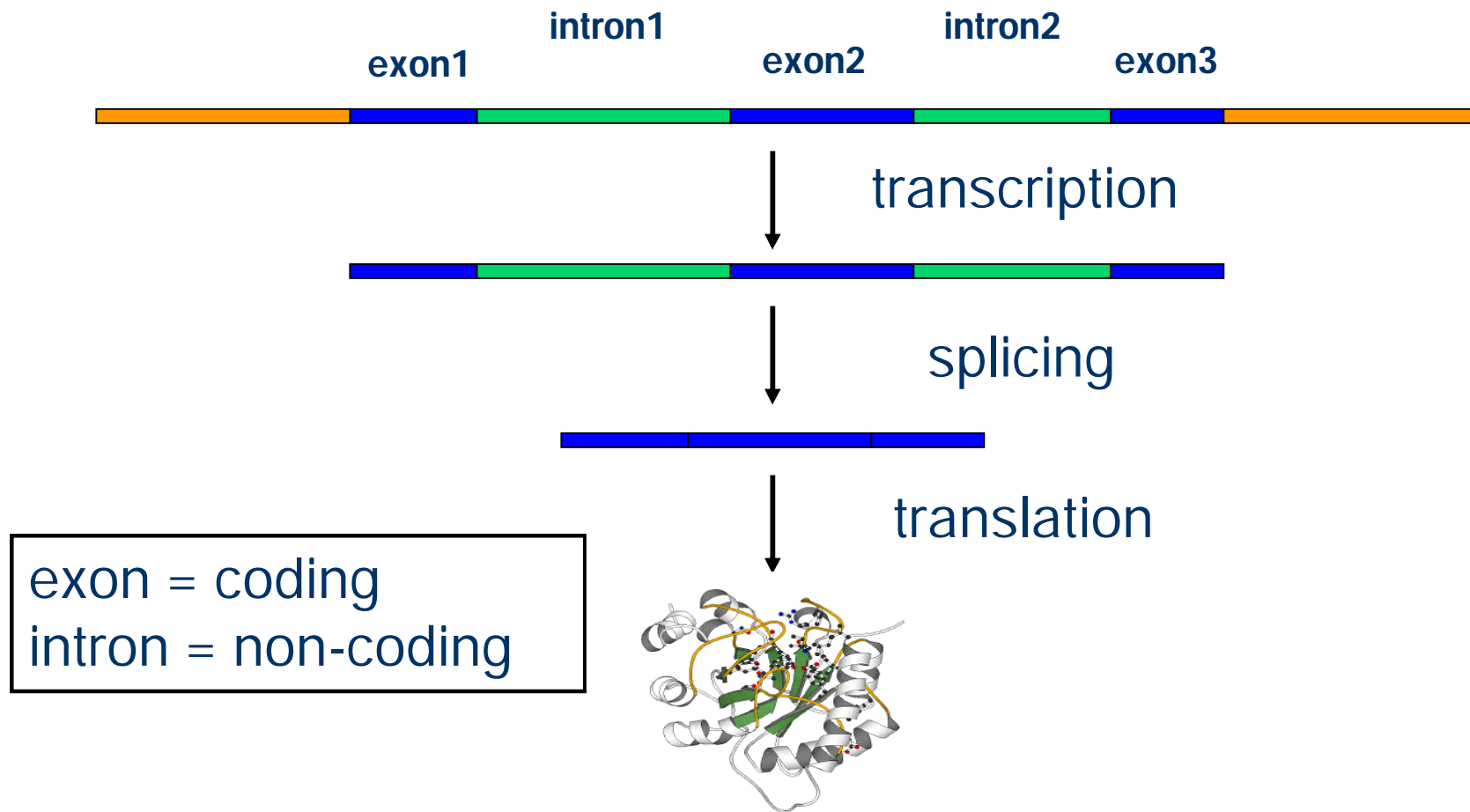


CCUGAGCCAAUUAUUGAUGAA

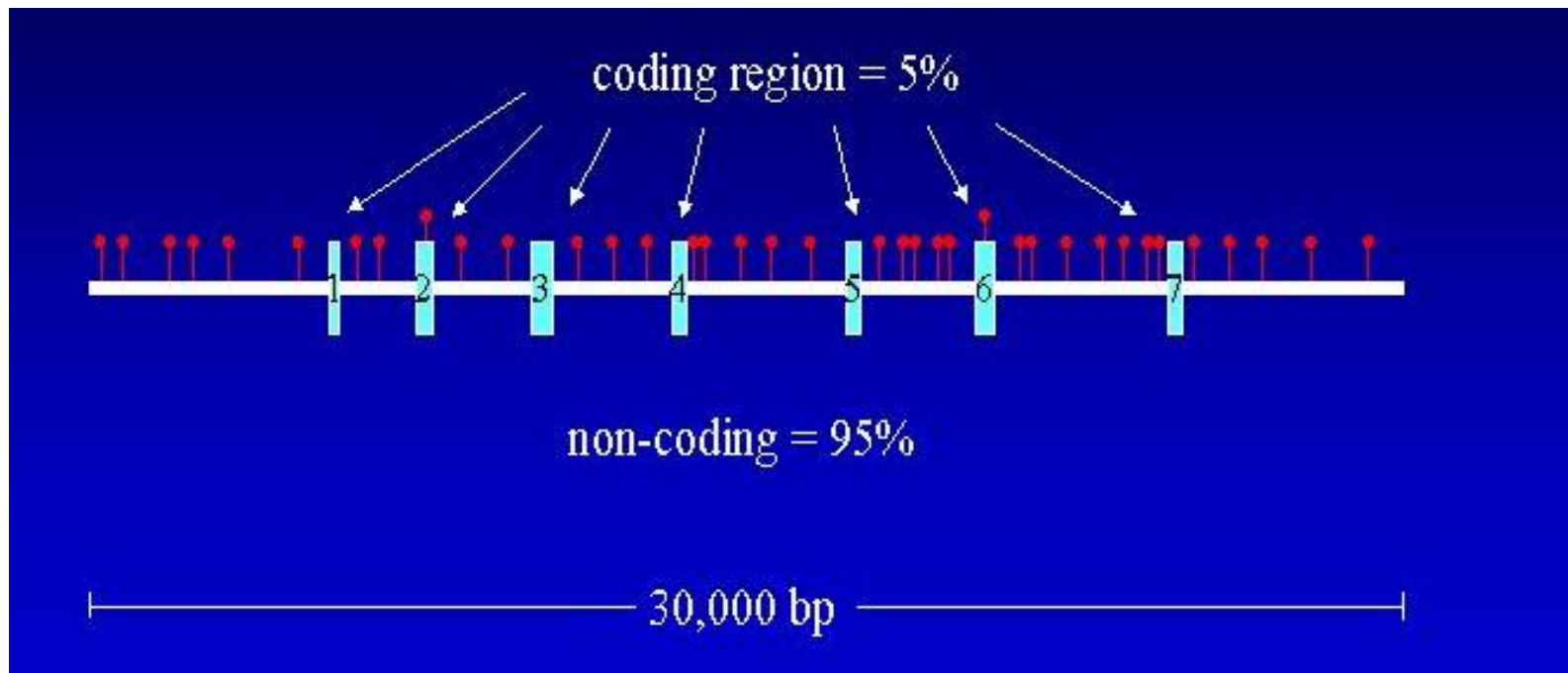


PEPTIDE

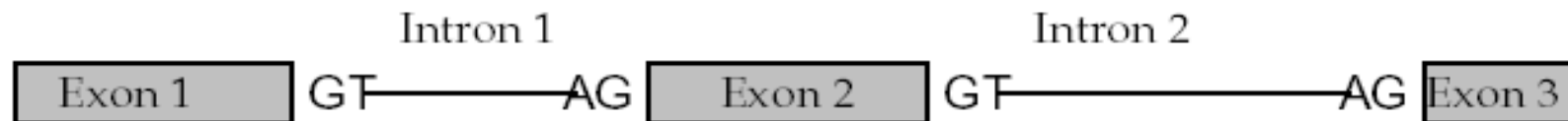
Central Dogma and Splicing



Gene Structure

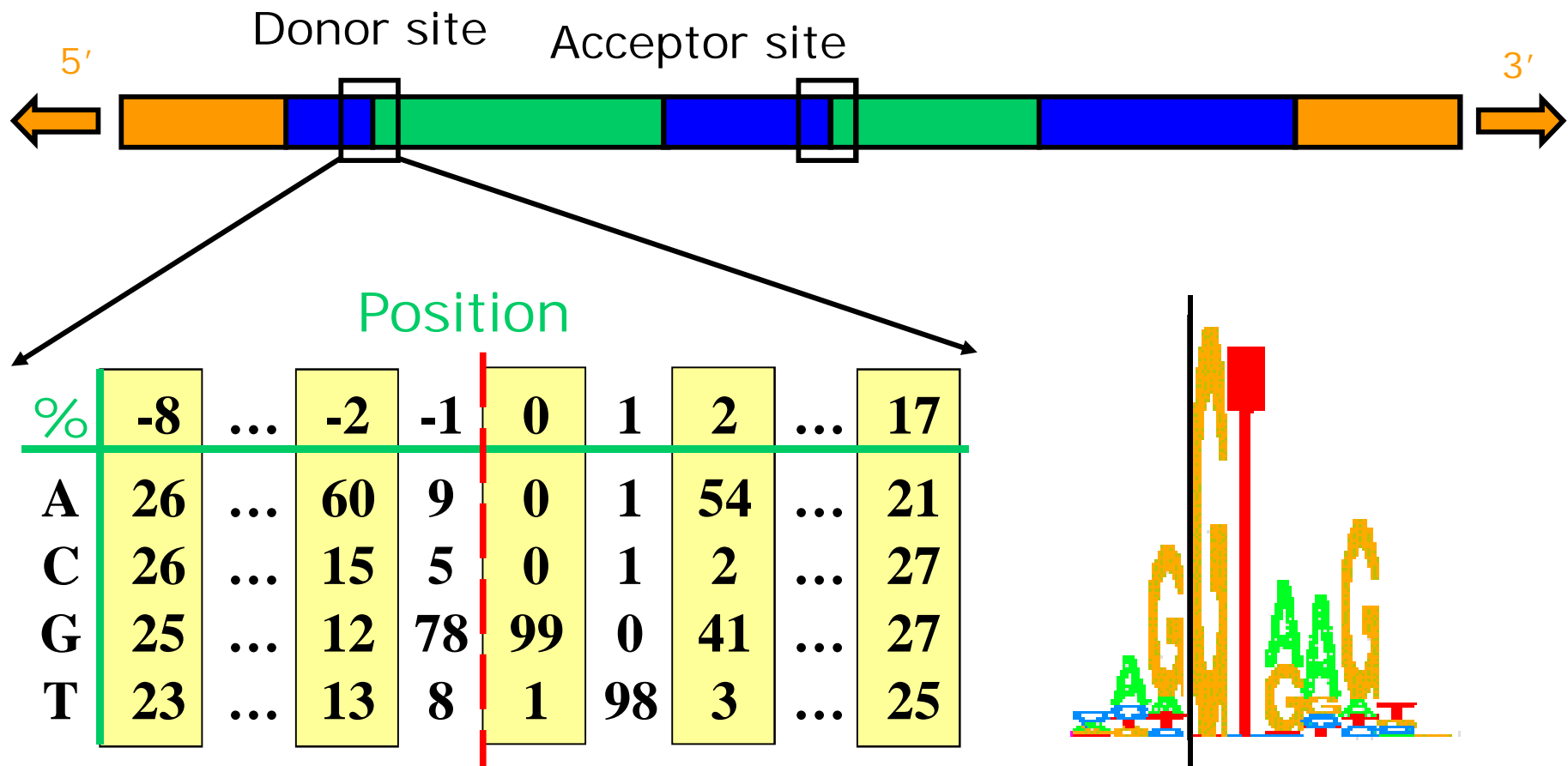


Splicing Signals



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

Splice site detection



Gene Prediction Analogy

- Newspaper written in unknown language
 - Certain pages contain encoded message, say 99 letters on page 7, 30 on page 12 and 63 on page 15.
- How do you recognize the message? You could probably distinguish between the ads and the story (ads contain the “\$” sign often)
- Statistics-based approach to Gene Prediction tries to make similar distinctions between exons and introns.

Statistical Approach: Metaphor in Unknown Language

en n,
tagonu, kan
s, priznaju da pomen
az postojanja oruzja za masov
ozda je vazno to sto je prvi put izjavu
ku prona eno nesto sto moze da
da je Saddam Husein ra
vanje dao visol
odbra

Noting the differing frequencies of symbols (e.g. '%', '.', '-') and numerical symbols could you distinguish between a story and the stock report in a foreign newspaper?

,363 0.75
0,761 505,812 9.00
6% 2.81 - 2.96 86,318,704
12 INTC 19.16 -0.38 -1.94% 19.06 -
57,755,076 12.95 - 31.36 VOD
00 - 19.46 4,366,500 3,20
58% 10,393,438
76 -0.36

Two Approaches to Gene Prediction

- **Statistical**: coding segments (exons) have typical sequences on either end and use different subwords than non-coding segments (introns).
- **Similarity-based**: many human genes are similar to genes in mice, chicken, or even bacteria. Therefore, already known mouse, chicken, and bacterial genes may help to find human genes.

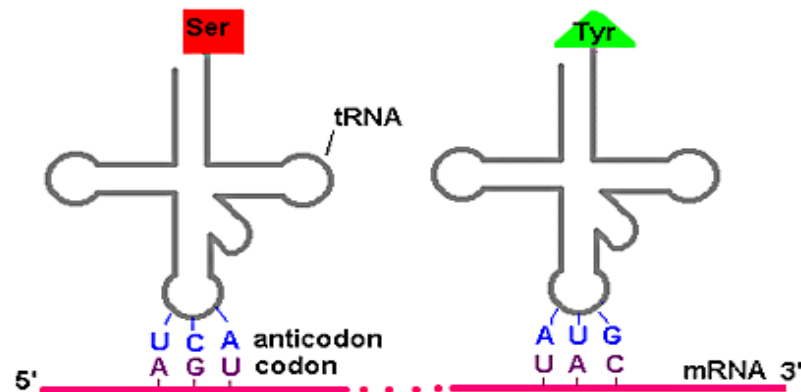
Similarity-Based Approach: Metaphor in Different Languages

diplomatic co...
Pentagon says plans...
into problems amid the conu...
ending the whole issue of post-war jus...
US officials have argued that the Ir...
of **Saddam Hussein** and...
as abused, they...
his ass

If you could compare the day's news in English, side-by-side to the same news in a foreign language, some similarities may become apparent

ia en...
Pentagonu, ka...
lds, priznaju da pomena...
okaz postojanja oruzja za masova...
tozda je vazno to sto je prvi put izjavu...
ku prona eno nesto sto moze...
da je **Sadam Husein**...
anje dao vi...
adl

Genetic Code and Stop Codons



		2nd base in codon				
		U	C	A	G	
1st base in codon	U	Phe Phe Leu Leu	Ser Ser Ser Ser	Tyr Tyr STOP STOP	Cys Cys STOP Trp	U C A G
	C	Leu Leu Leu Leu	Pro Pro Pro Pro	His His Gln Gln	Arg Arg Arg Arg	U C A G
	A	Ile Ile Ile Met	Thr Thr Thr Thr	Asn Asn Lys Lys	Ser Ser Arg Arg	U C A G
	G	Val Val Val Val	Ala Ala Ala Ala	Asp Asp Glu Glu	Gly Gly Gly Gly	U C A G

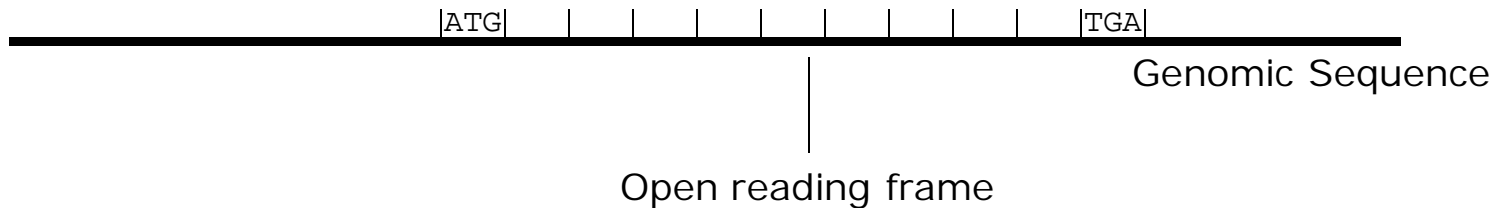
3rd base in codon

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

The Genetic Code

Open Reading Frames (ORFs)

- 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 $(64/3) \approx 21$ 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

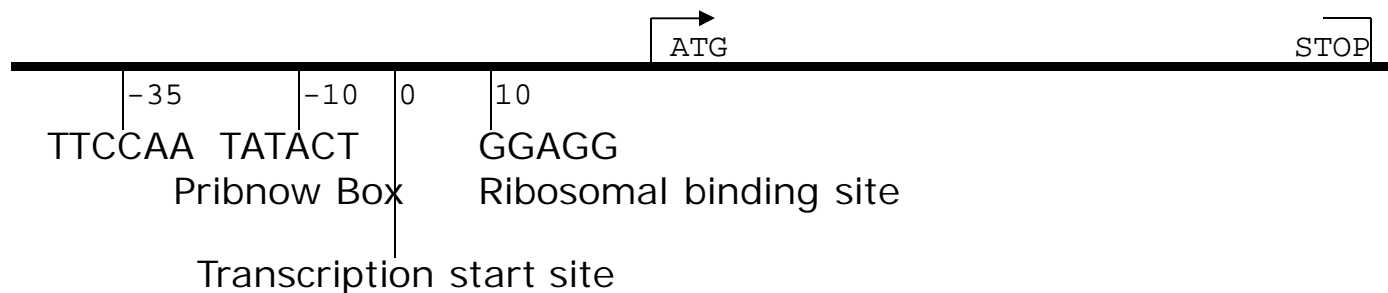
	U	C	A	G
U	UUU Phe 57	UCU Ser 16	UAU Tyr 58	UGU Cys 45
	UUC Phe 43	UCC Ser 15	UAC Tyr 42	UGC Cys 55
	UUA Leu 13	UCA Ser 13	UAA Stp 62	UGA Stp 30
	UUG Leu 13	UCG Ser 15	UAG Stp 8	UGG Trp 100
C	CUU Leu 11	CCU Pro 17	CAU His 57	CGU Arg 37
	CUC Leu 10	CCC Pro 17	CAC His 43	CGC Arg 38
	CUA Leu 4	CCA Pro 20	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 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)

Gene Prediction and Motifs

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



Gene Prediction: Similarity-Based Approaches

- Section 6.13