Gene Prediction: Statistical Approaches

Outline

- Codons
- Discovery of Split Genes
- Exons and Introns
- Splicing
- Open Reading Frames
- Codon Usage
- Splicing Signals
- TestCode

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

aatgcatgcggctatgctaatgcatgcggctatgctaagctgggatccgatgacaatgcatgcggctatgctaa
tgcatgcggctatgctaagctgggatccgatgactatgctaagctgggatccgatgacaatgcatgcggctatg
ctaatgaatggtcttgggatttaccttggaatgctaagctgggatccgatgacaatgcatgcggctatgctaatg
aatggtcttgggatttaccttggaatatgctaatgcatgcggctatgctaagctgggatccgatgacaatgcatg
cggctatgctaatgcatgcggctatgcaagctgggatccgatgactatgctaagctgcggctatgctaatgcat
gcggctatgctaagctgggatccgatgacaatgcatgcggctatgctaatgcatgcggctatgcaagctggga
tcctgcggctatgctaatgaatggtcttgggatttaccttggaatgctaagctgggatccgatgacaatgcatgc
ggctatgctaatgaatggtcttgggatttaccttggaatatgctaatgcatgcggctatgctaagctgggaatgca
tgcggctatgctaagctgggatccgatgacaatgcatgcggctatgctaatgcatgcggctatgcaagctggg
atccgatgactatgctaagctgcggctatgctaatgcatgcggctatgctaagctcatgcggctatgctaagctg
ggaatgcatgcggctatgctaagctgggatccgatgacaatgcatgcggctatgctaatgcatgcggctatgc
aagctgggatccgatgactatgctaagctgcggctatgctaatgcatgcggctatgctaagctcggctatgcta
atgaatggtcttgggatttaccttggaatgctaagctgggatccgatgacaatgcatgcggctatgctaatgaat
ggtcttgggatttaccttggaatatgctaatgcatgcggctatgctaagctgggaatgcatgcggctatgctaag
ctgggatccgatgacaatgcatgcggctatgctaatgcatgcggctatgcaagctgggatccgatgactatgct
aagctgcggctatgctaatgcatgcggctatgctaagctcatgcgg

Gene Prediction: Computational Challenge

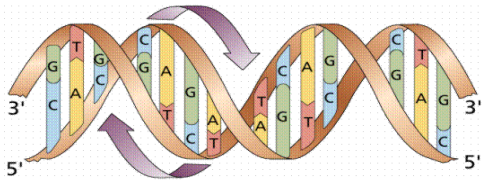
aatgcatgcggctatgctaatgcatgcggctatgctaagctgggatccgatgacaatgcatgcggctatgctaa
tgcatgcggctatgctaagctgggatccgatgactatgctaagctgggatccgatgacaatgcatgcggctatg
ctaatgaatggtcttgggatttaccttggaatgctaagctgggatccgatgacaatgcatgcggctatgctaatg
aatggtcttgggatttaccttggaatatgctaatgcatgcggctatgctaagctgggatccgatgacaatgcatg
cggctatgctaatgcatgcggctatgcaagctgggatccgatgactatgctaagctgcggctatgctaatgcat
gcggctatgctaagctgggatccgatgacaatgcatgcggctatgctaatgcatgcggctatgcaagctggga
tcctgcggctatgctaatgaatggtcttgggatttaccttggaatgctaagctgggatccgatgacaatgcatgc
ggctatgctaatgaatggtcttgggatttaccttggaatatgctaatgcatgcggctatgctaagctgggaatgca
tgcggctatgctaagctgggatccgatgacaatgcatgcggctatgctaatgcatgcggctatgcaagctggg
atccgatgactatgctaagctgcggctatgctaatgcatgcggctatgctaagctcatgcggctatgctaagctg
ggaatgcatgcggctatgctaagctgggatccgatgacaatgcatgcggctatgctaatgcatgcggctatgc
aagctgggatccgatgactatgctaagctgcggctatgctaatgcatgcggctatgctaagctcggctatgcta
atgaatggtcttgggatttaccttggaatgctaagctgggatccgatgacaatgcatgcggctatgctaatgaat
ggtcttgggatttaccttggaatatgctaatgcatgcggctatgctaagctgggaatgcatgcggctatgctaag
ctgggatccgatgacaatgcatgcggctatgctaatgcatgcggctatgcaagctgggatccgatgactatgct
aagctgcggctatgctaatgcatgcggctatgctaagctcatgcgg

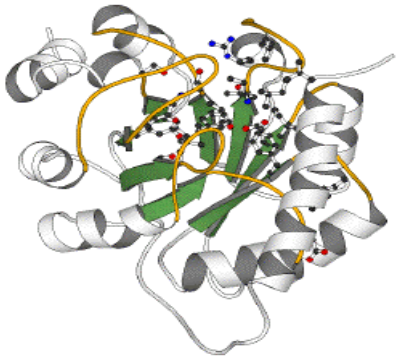
Gene Prediction: Computational Challenge

aatgcatgcggctatgctaatgcatgcggctatgctaagctgggatccgatgacaatgcatgcggctatgctaa
tgcatgcggctatgctaagctgggatccgatgactatgctaagctgggatccgatgacaatgcatgcggctatg
ctaatgaatggtcttgggatttaccttggaatgctaagctgggatccgatgacaatgcatgcggctatgctaatg
aatggtcttgggatttaccttggaatatgctaatgcatgcggctatgctaagctgggatccgatgacaatgcatg
cggctatgctaatgcatgcggctatgcaagctgggatccgatgactatgctaagctgcggctatgctaatgcat
gcggctatgctaagctgggatccgatgacaatgcatgcggctatgctaatgcatgcggctatgcaagctggga
tcctgcggctatgctaatgaatggtcttgggatttaccttggaatgctaagctgggatccgatgacaatgcatgc
ggctatgctaatgaatggtcttgggatttaccttggaatatgctaatgcatgcggctatgctaagctgggaatgca
tgcggctatgctaagctgggatccgatgacaatgcatgcggctatgctaatgcatgcggctatgcaagctggg
atccgatgactatgctaagctgcggctatgctaatgcatgcggctatgctaagctcatgcggctatgctaagctg
ggaatgcatgcggctatgctaagctgggatccgatgacaatgcatgcggctatgctaatgcatgcggctatgc
aagctgggatccgatgactatgctaagctgcggctatgctaatgcatgcggctatgctaagctcggctatgcta
atgaatggtcttgggatttaccttggaatgctaagctgggatccgatgacaatgcatgcggctatgctaatgaat
ggtcttgggatttaccttggaatatgctaatgcatgcggctatgctaagctgggaatgcatgcggctatgctaag
ctgggatccgatgacaatgcatgcggctatgctaatgcatgcggctatgcaagctgggatccgatgactatgct
aagctgcggctatgctaatgcatgcggctatgctaagctcatgcgg

Gene!

Central Dogma: DNA -> RNA -> Protein





DNA

CCTGAGCCAACTATTGATGAA

transcription

RNA

CCUGAGCCAACUAUUGAUGAA

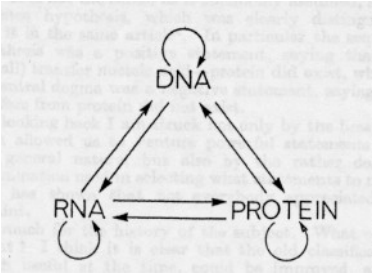
translation

Protein

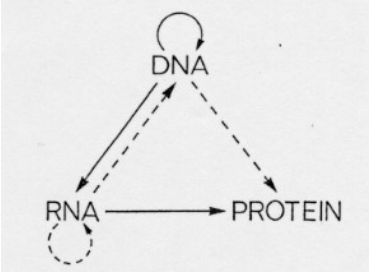
PEPTIDE

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 nucleotifes 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?

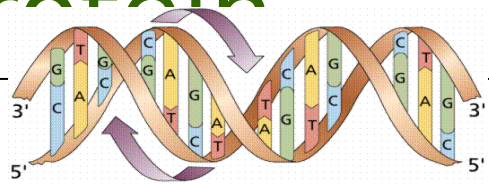
Translating Nucleotides into Amino Acids

- 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



DNA

CCTGAGCCAACTATTGATGAA

transcription

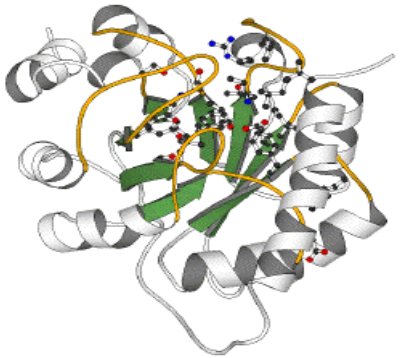
RNA

CCUGAGCCAACUAUUGAUGAA

translation

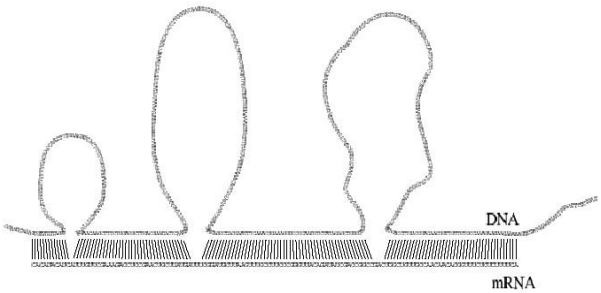
Protein

PEPTIDE



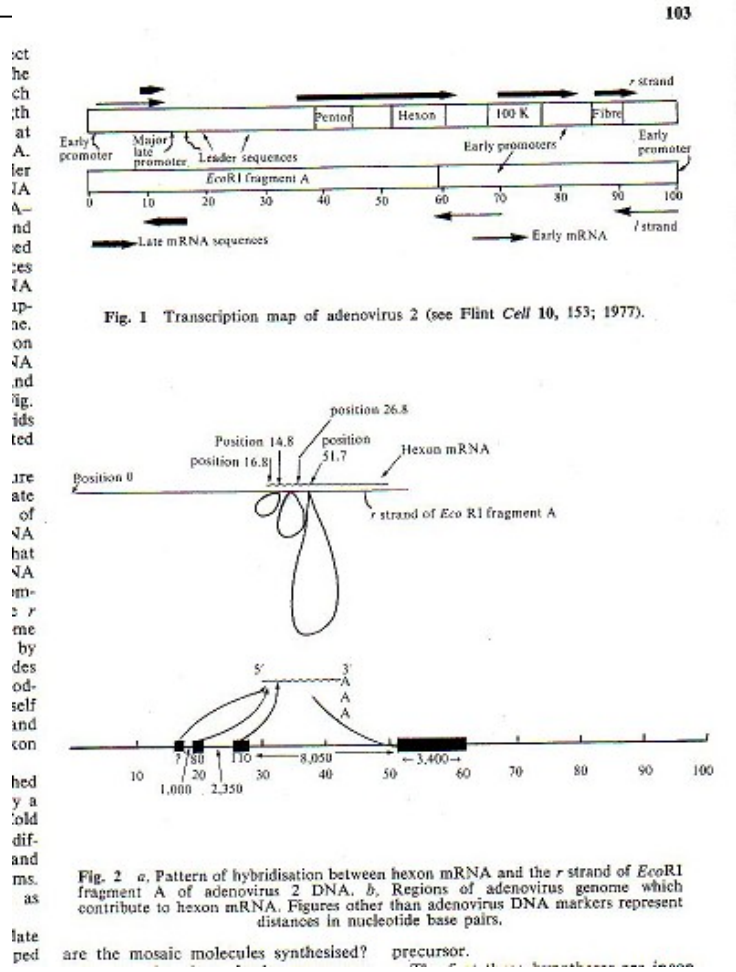
Discovery of Split Genes

- 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)

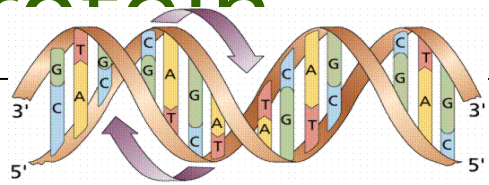
- "Adenovirus Amazes at Cold Spring Harbor" (1977, Nature 268) documented "mosaic molecules consisting of sequences complementary to several non-contiguous segments of the viral genome".
- In 1978 Walter Gilbert coined the term **intron** in the Nature paper "Why Genes in Pieces?"



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



DNA

CCTGAGCCAACTATTGATGAA

transcription

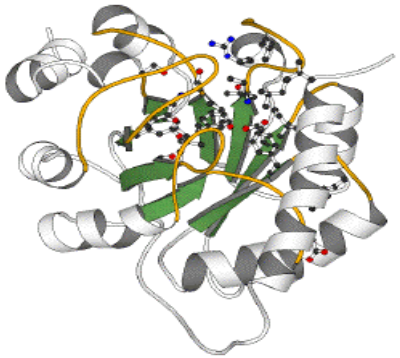
RNA

CCUGAGCCAACUAUUGAUGAA

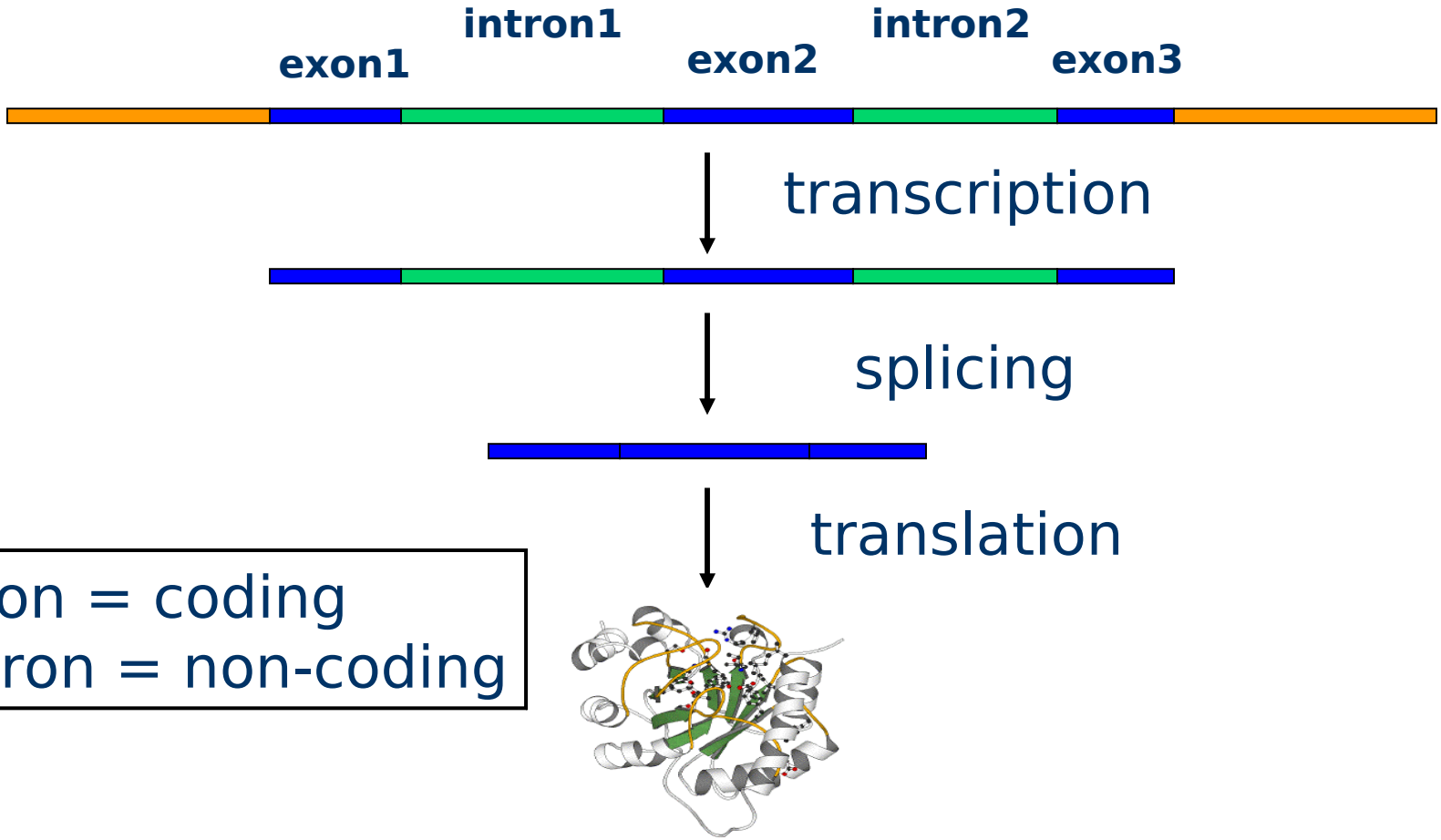
translation

Protein

PEPTIDE

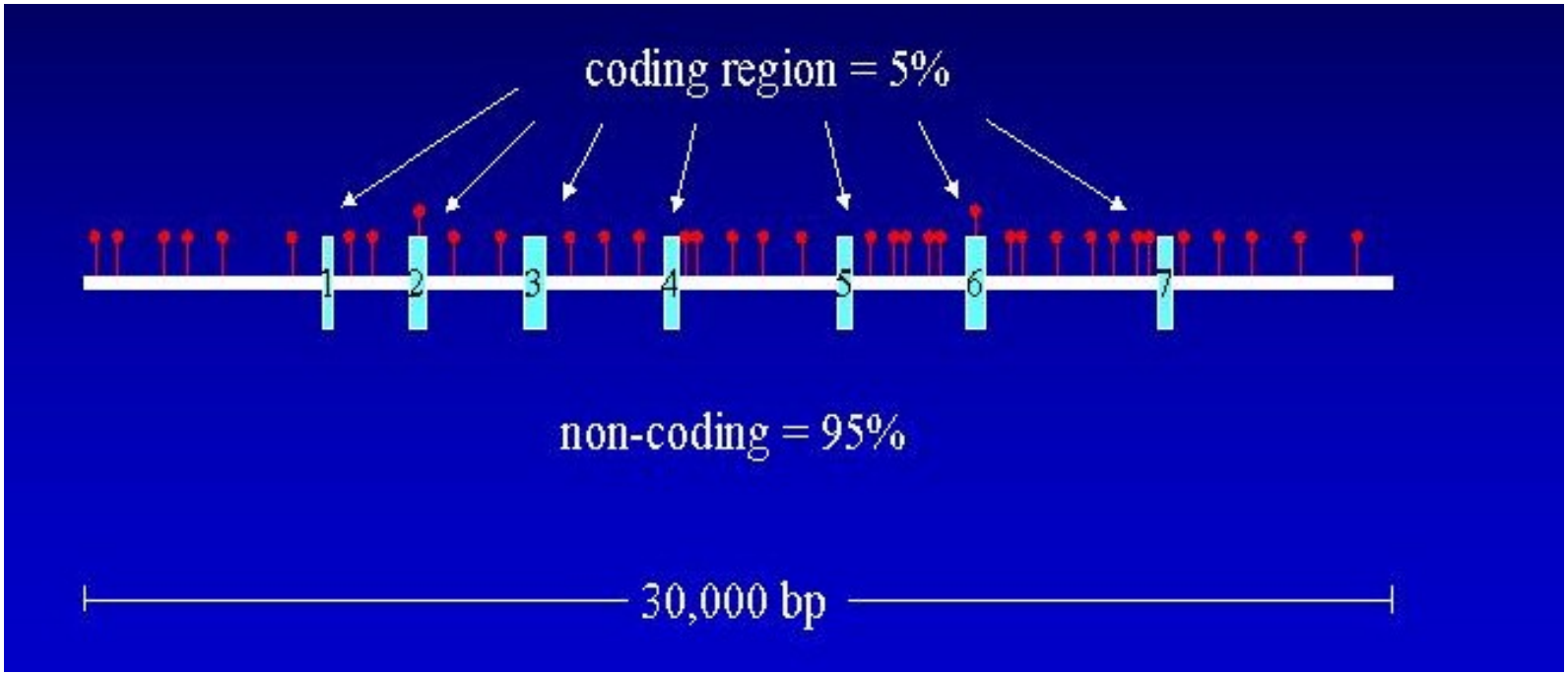


Central Dogma and Splicing

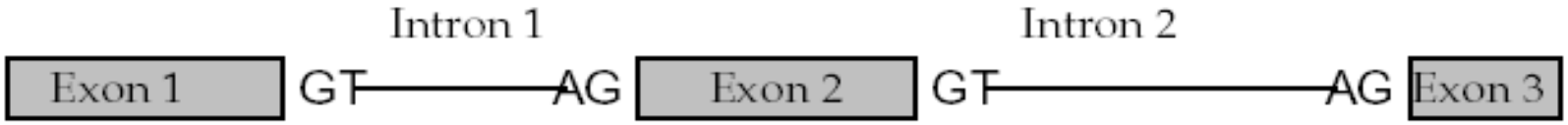


exon = coding
intron = non-coding

Gene Structure

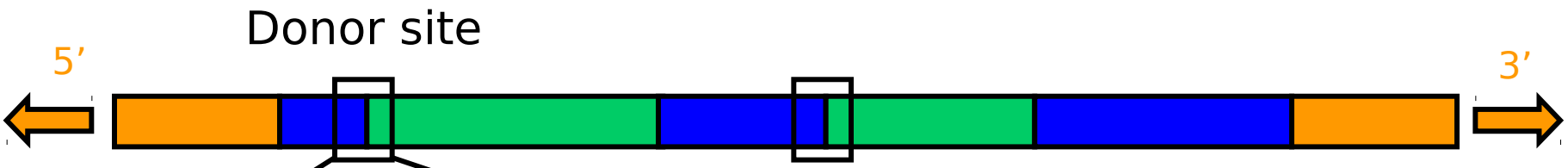


Splicing Signals



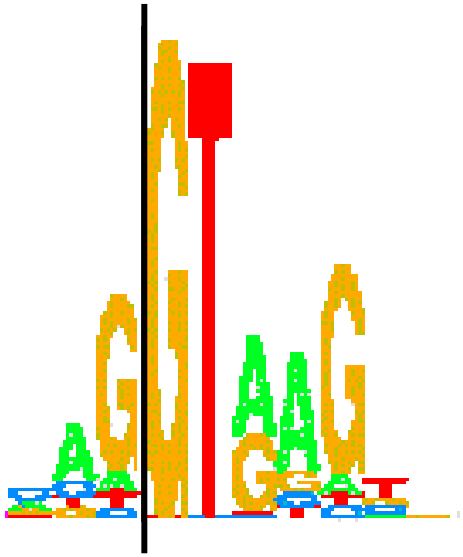
Exons are interspersed with introns and typically flanked by GT and AG

Splice site detection



Position

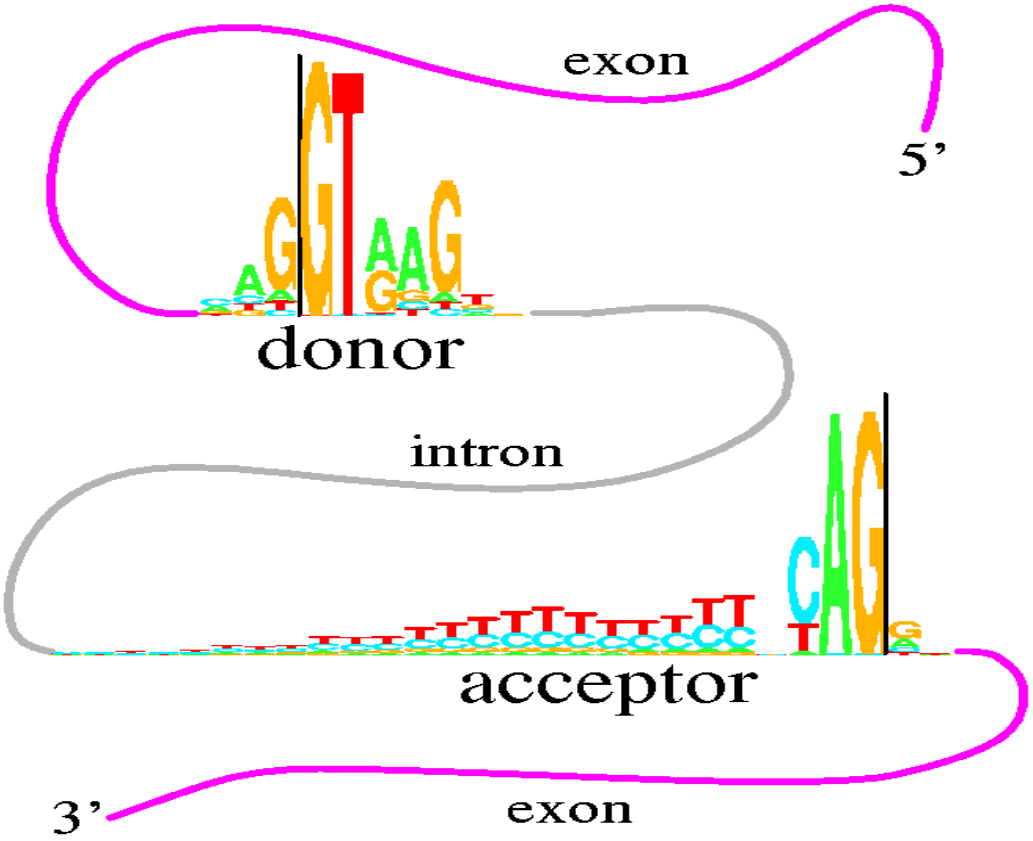
%	-8	…	-2	-1	0	1	2	…	17
A	26	…	60	9	0	1	54	…	21
C	26	…	15	5	0	1	2	…	27
G	25	…	12	78	99	0	41	…	27
T	23	…	13	8	1	98	3	…	25



Consensus splice sites

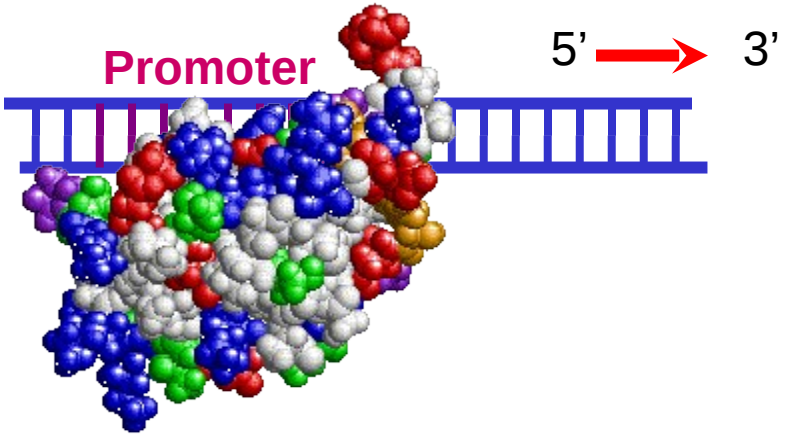
This figure shows two "sequence logos" which represent sequence conservation at the 5' (donor) and 3' (acceptor) ends of human introns. The region between the black vertical bars is removed during mRNA splicing. The logos graphically demonstrate that most of the pattern for locating the intron ends resides on the intron. This allows more codon choices in the protein-coding exons. The logos also show a common pattern "CAG|GT", which suggests that the mechanisms that recognize the two ends of the intron had a common ancestor. See R. M. Stephens and T. D. Schneider, "Features of spliceosome evolution and function inferred from an analysis of the information at human splice sites", J. Mol. Biol., 228, 1124-1136, (1992)

Donor: 7.9 bits
Acceptor: 9.4 bits



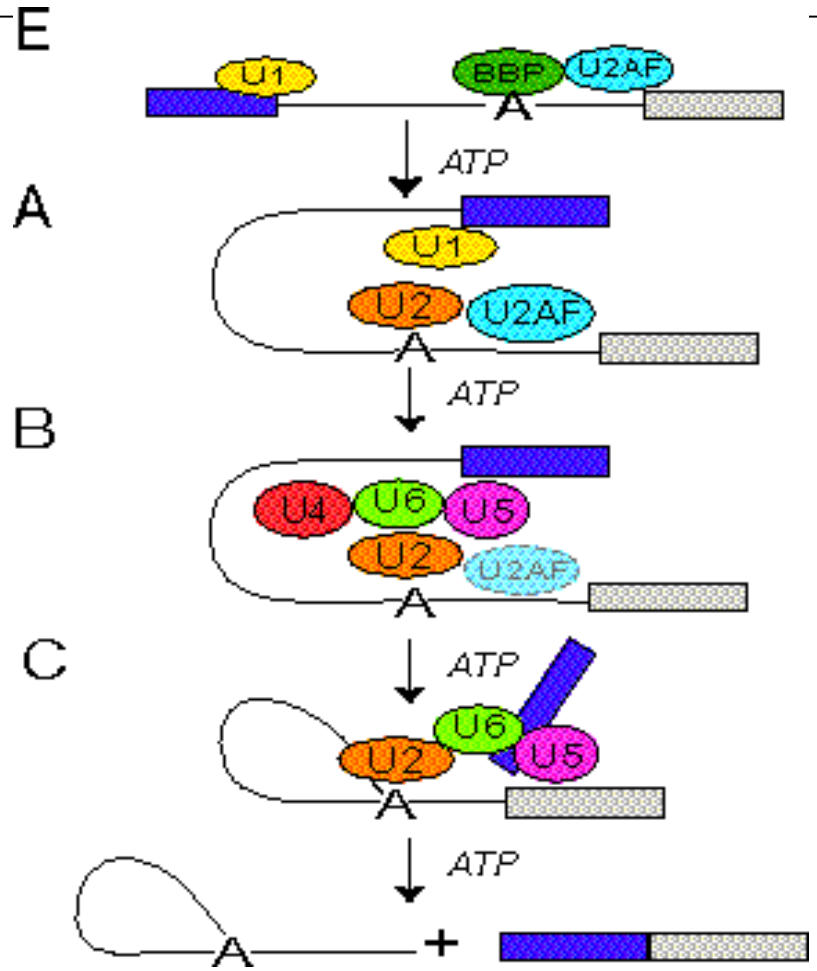
Promoters

- Promoters are DNA segments upstream of transcripts that initiate transcription



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

Splicing mechanism

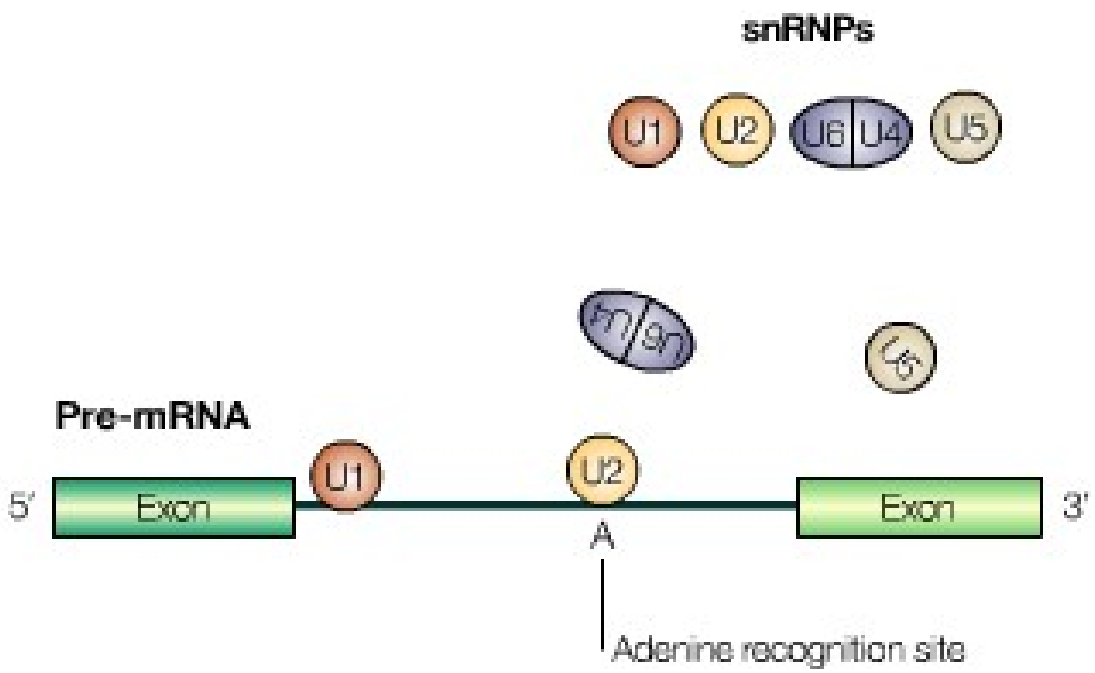


(http://genes.mit.edu/chris/)

Splicing mechanism

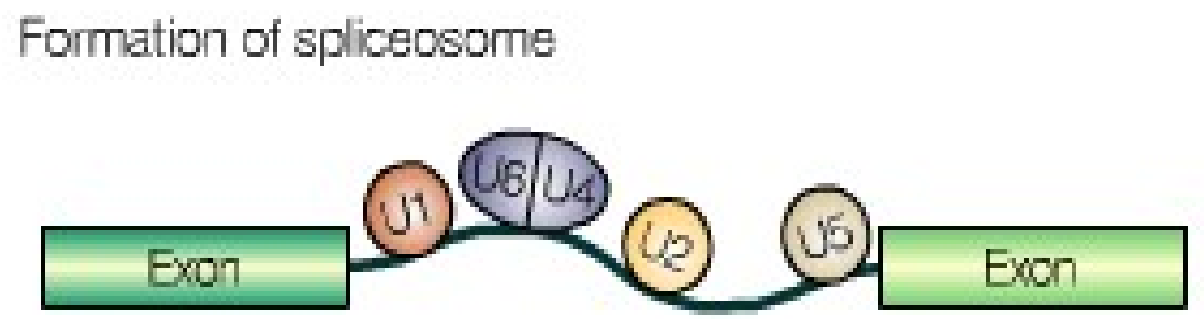
- Adenine recognition site marks intron
- snRNPs bind around adenine recognition site
- The *spliceosome* thus forms
- Spliceosome excises introns in the mRNA

Activating the snRNPs



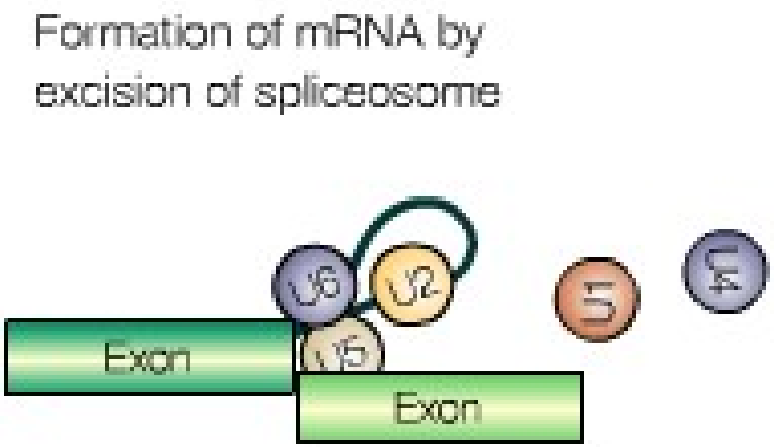
From lectures by Chris Burge (MIT)

Spliceosome Facilitation



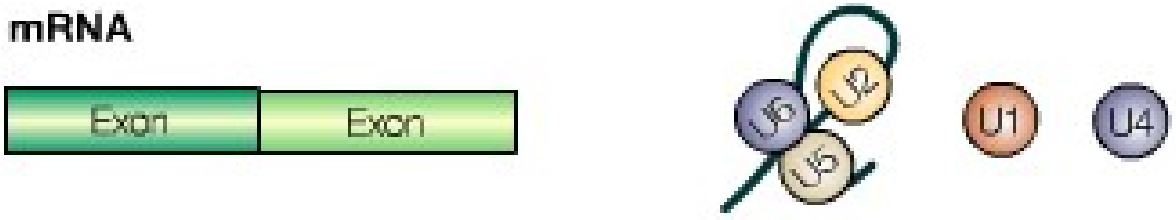
From lectures by Chris Burge (MIT)

Intron Excision



From lectures by Chris Burge (MIT)

mRNA is now Ready

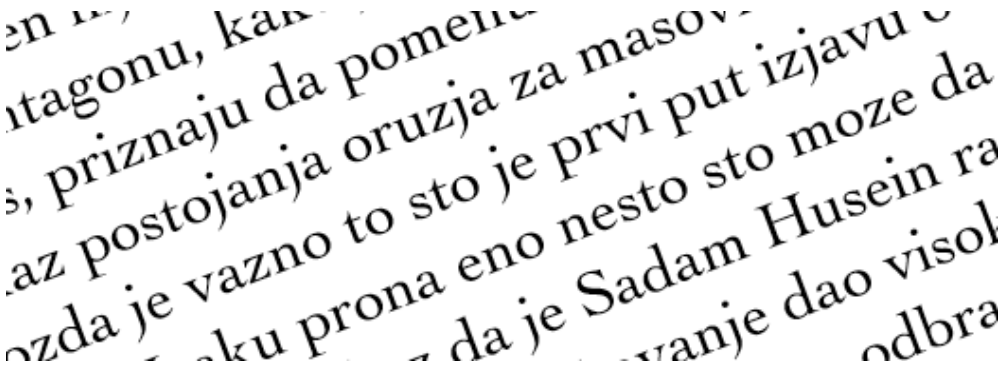


From lectures by Chris Burge (MIT)

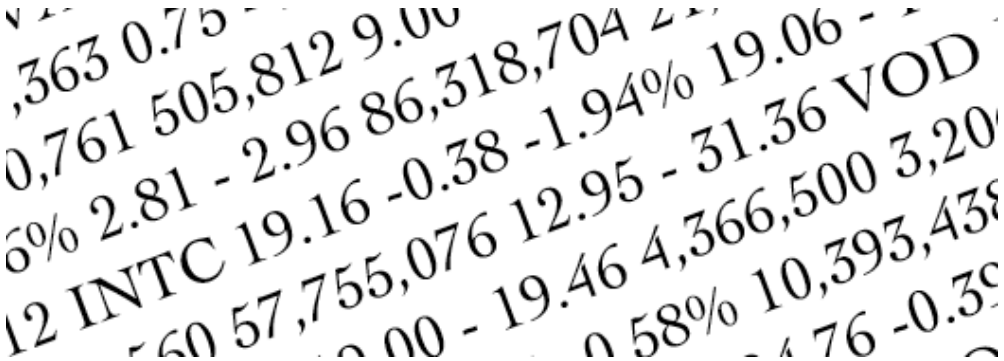
Gene Prediction Analogy

- 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



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?

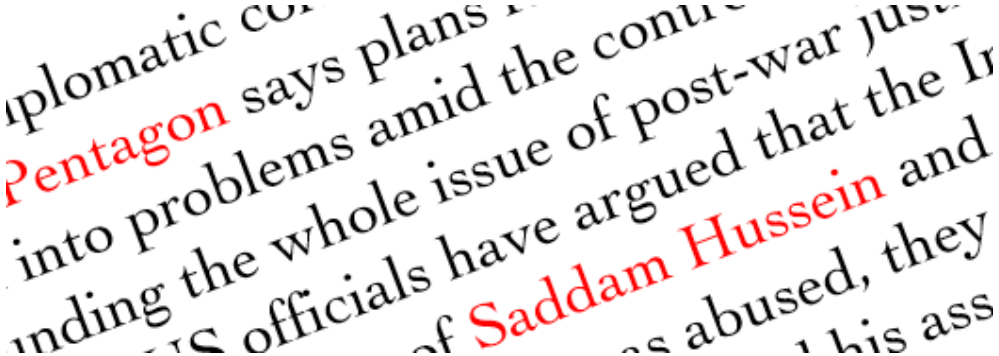


Two Approaches to Gene Prediction

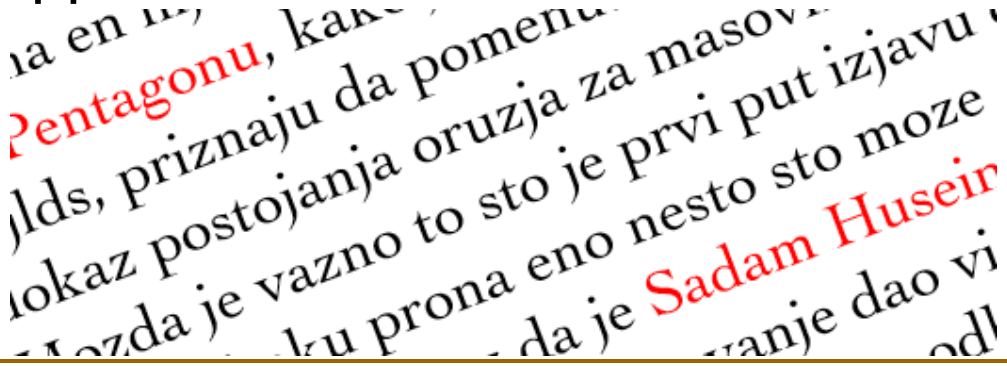
- **Statistical**: coding segments (exons) have typical sequences on either end and use different subwords than non-coding segments (introns).

- **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.

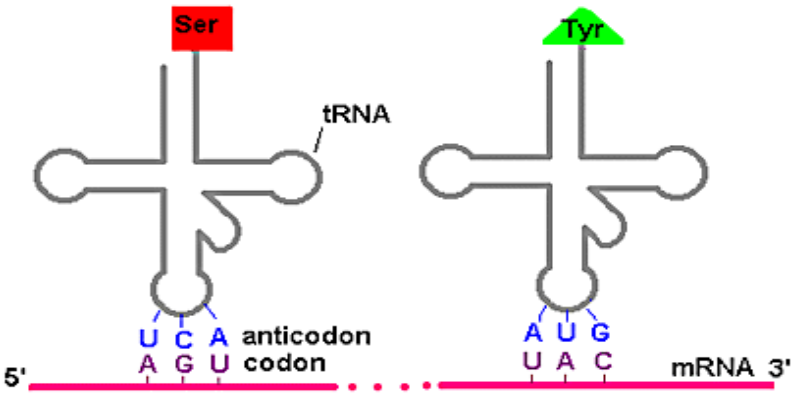
Similarity-Based Approach: Metaphor in Different Languages



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



Genetic Code and Stop Codons



2nd base in codon

1st base in codon	U	C	A	G	3rd 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

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

CTGCAGACGAAACCTCTTGATGTAGTTGGCCTGACACCGACAATAATGAAGACTACCGTCTTACTAACAC
CTGCAGACGAAACCTCTTGATGTAGTTGGCCTGACACCGACAATAATGAAGACTACCGTCTTACTAACAC
CTGCAGACGAAACCTCTTGATGTAGTTGGCCTGACACCGACAATAATGAAGACTACCGTCTTACTAACAC

CTGCAGACGAAACCTCTTGATGTAGTTGGCCTGACACCGACAATAATGAAGACTACCGTCTTACTAACAC

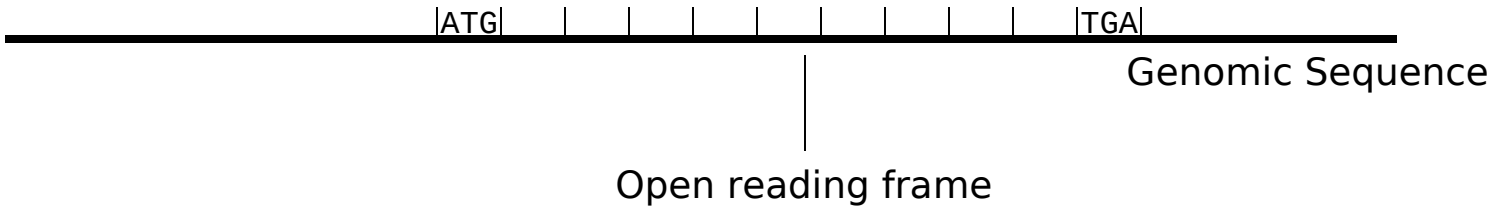
GACGTCTGCTTTGGAGAACTACATCAACCGGACTGTGGCTGTTATTACTTCTGATGGCAGAATGATTGTG

GACGTCTGCTTTGGAGAACTACATCAACCGGACTGTGGCTGTTATTACTTCTGATGGCAGAATGATTGTG
GACGTCTGCTTTGGAGAACTACATCAACCGGACTGTGGCTGTTATTACTTCTGATGGCAGAATGATTGTG
GACGTCTGCTTTGGAGAACTACATCAACCGGACTGTGGCTGTTATTACTTCTGATGGCAGAATGATTGTG

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

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) ~= 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

- 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

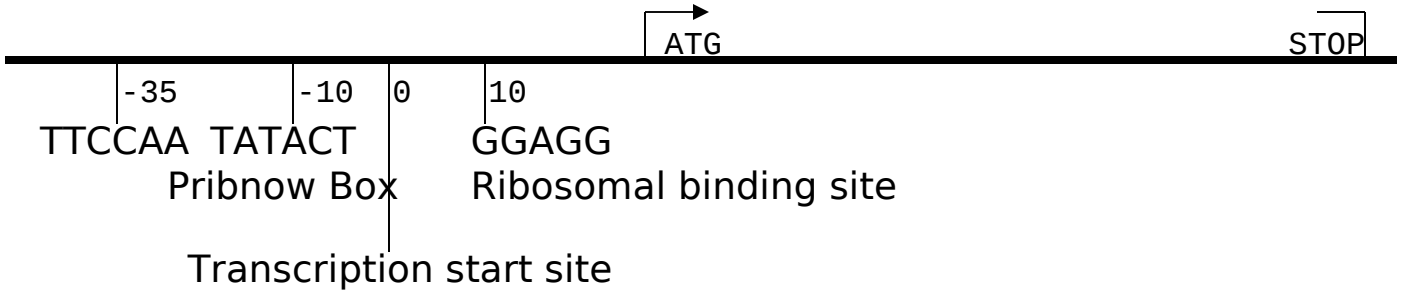
AA	codon	/1000	frac
Leu	CTG	39.95	0.40
Leu	CTA	7.89	0.08
Leu	CTT	12.97	0.13
Leu	CTC	20.04	0.20
Ala	GCG	6.72	0.10
Ala	GCA	15.80	0.23
Ala	GCT	20.12	0.29
Ala	GCC	26.51	0.38
Gln	CAG	34.18	0.75
Gln	CAA	11.51	0.25

Codon Usage and Likelihood

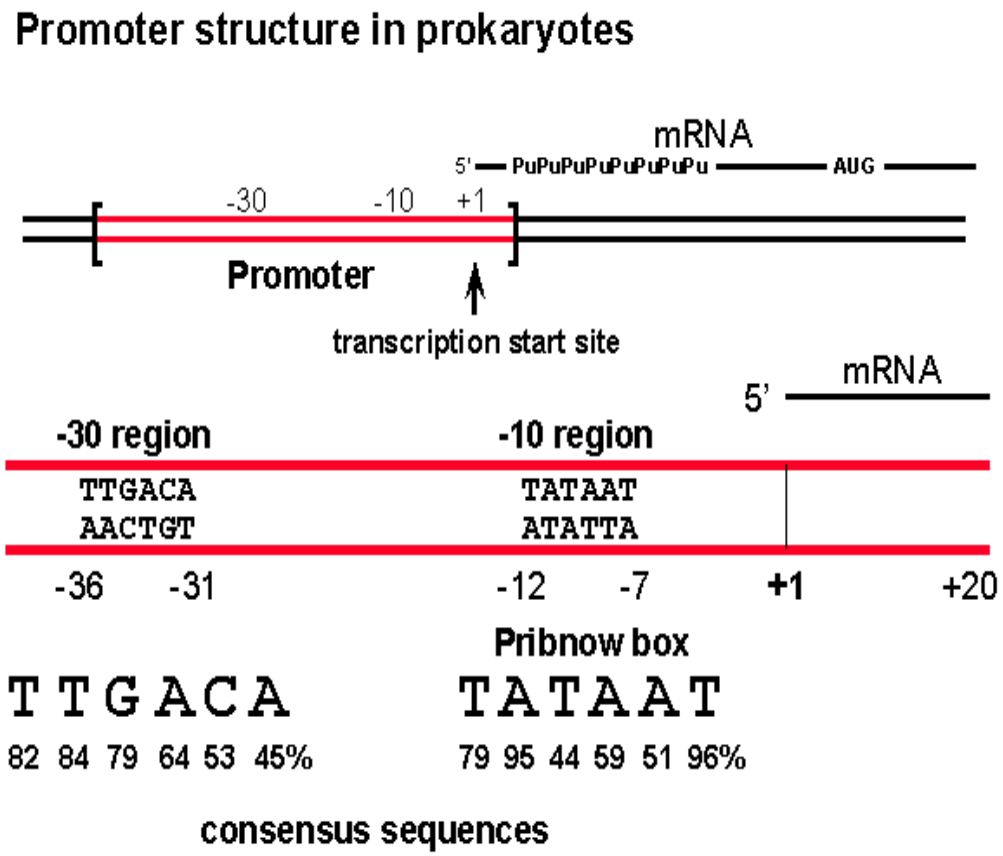
- An ORF is more “believable” than another if it has more “likely” codons
- Do sliding window calculations to find ORFs that have the “likely” codon usage
- Allows for higher precision in identifying true ORFs; much better than merely testing for length.
- However, average vertebrate exon length is 130 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



Promoter Structure in Prokaryotes (E.Coli)

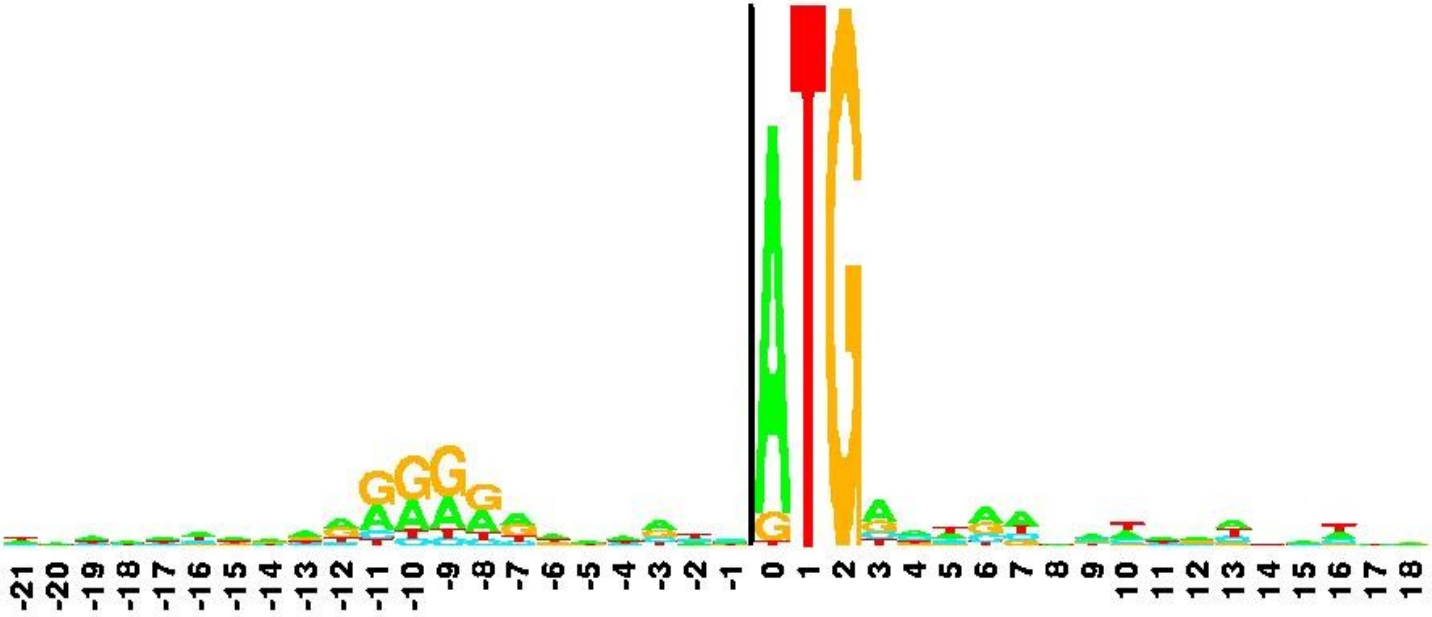


Transcription starts at offset 0.

- Pribnow Box (-10)
- Gilbert Box (-30)
- Ribosomal Binding Site (+10)

Ribosomal Binding Site

1055 E. coli Ribosome binding sites listed in the Miller book

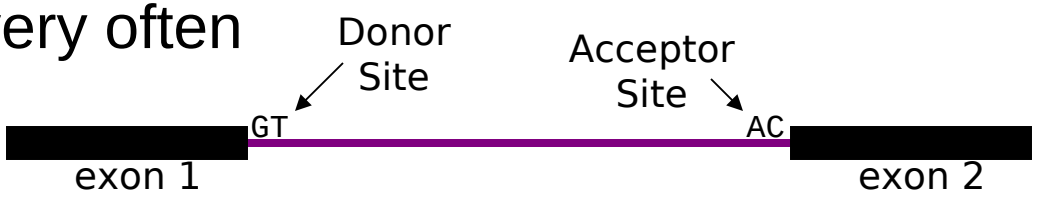


Splicing Signals

- 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

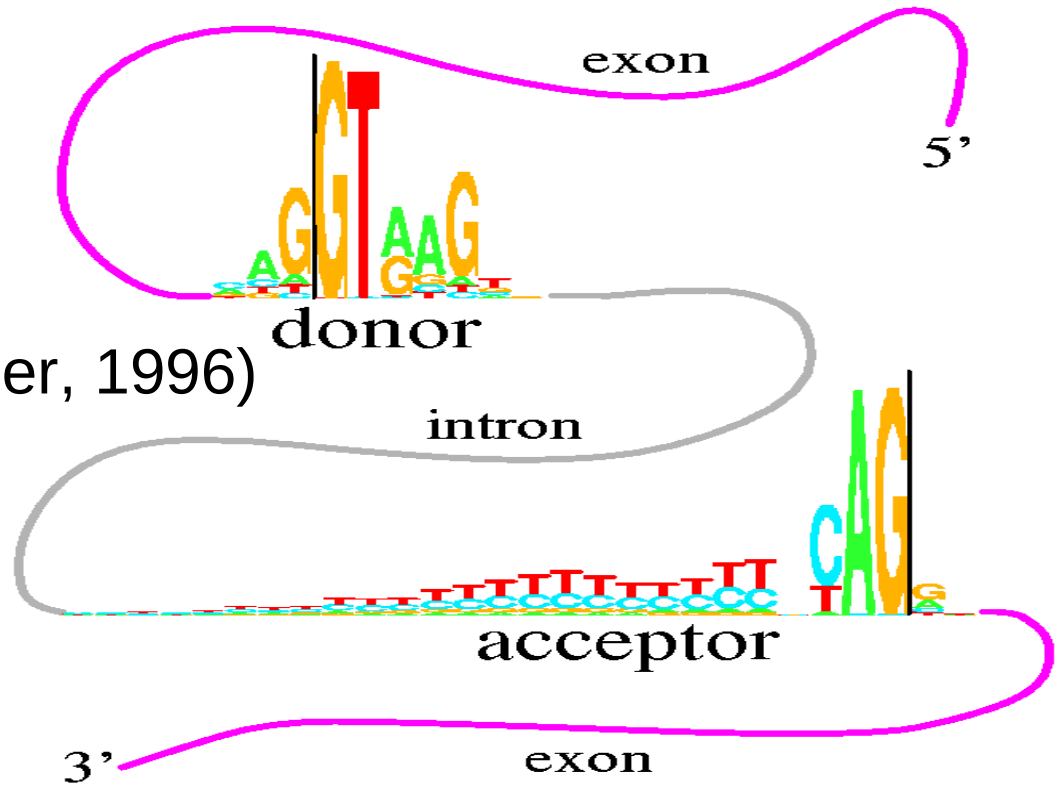
Donor and Acceptor Sites: GT and AG dinucleotides

- 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



Donor and Acceptor Sites: Motif Logos

This figure shows the two "sequence logos" which represent sequence conservation at the 5' (donor) and 3' (acceptor) ends of human introns. The region between the black vertical bars is removed during mRNA splicing. The logos graphically demonstrate that most of the pattern for locating the intron ends resides on the intron. This allows more codon choices in the protein-coding exons. The logos also show a common pattern "CAG|GT", which suggests that the mechanisms that recognize the two ends of the intron had a common ancestor. See R. M. Stephens and T. D. Schneider, "Features of spliceosome evolution and function inferred from an analysis of the information at human splice sites", J. Mol. Biol., 228, 1124-1136, (1992)



Donor: 7.9 bits
Acceptor: 9.4 bits
(Stephens & Schneider, 1996)

(http://www-lmmb.ncifcrf.gov/~toms/sequencelogo.html)

TestCode

- Statistical test described by James Fickett in 1982: tendency for nucleotides in coding regions to be repeated with periodicity of 3
  - Judges randomness instead of codon frequency
  - Finds “putative” coding regions, not introns, exons, or splice sites
- TestCode finds ORFs based on compositional bias with a periodicity of three

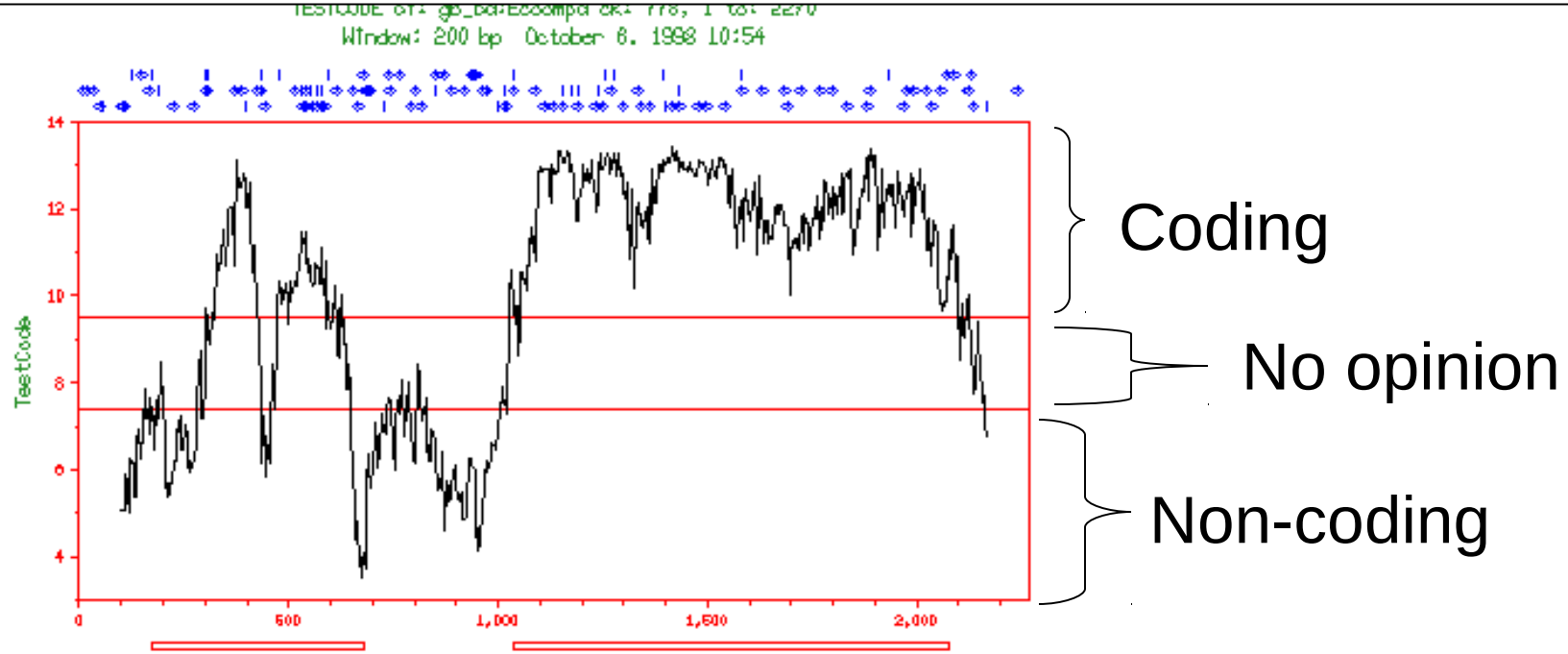
TestCode Statistics

- Define a window size no less than 200 bp, slide the window the sequence down 3 bases. In each window:
  - Calculate for each base {A, T, G, C}
    - max (n_{3k+1}, n_{3k+2}, n_{3k}) / min (n_{3k+1}, n_{3k+2}, n_{3k})
    - Use these values to obtain a probability from a lookup table (which was a previously defined and determined experimentally with known coding and noncoding sequences

TestCode Statistics (cont'd)

- Probabilities can be classified as indicative of " coding" or "noncoding" regions, or "no opinion" when it is unclear what level of randomization tolerance a sequence carries
- The resulting sequence of probabilities can be plotted

TestCode Sample Output



Popular Gene Prediction Algorithms

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