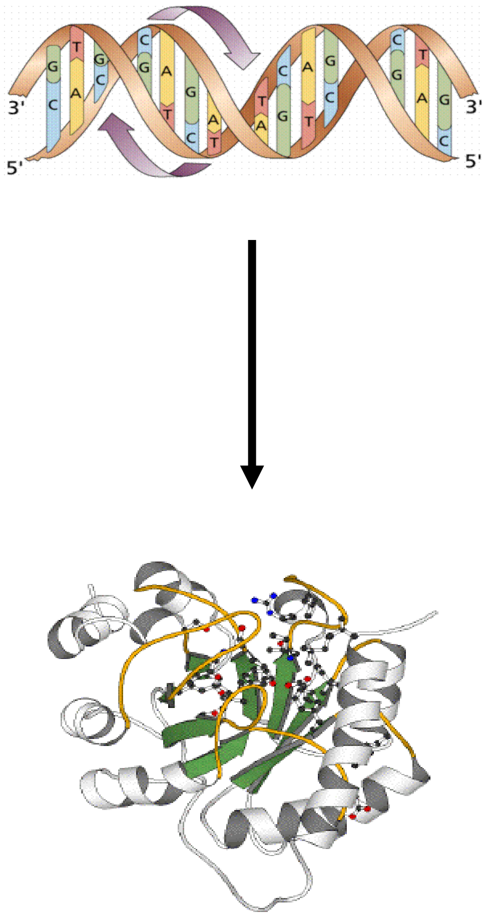
EECS730: Introduction to Bioinformatics

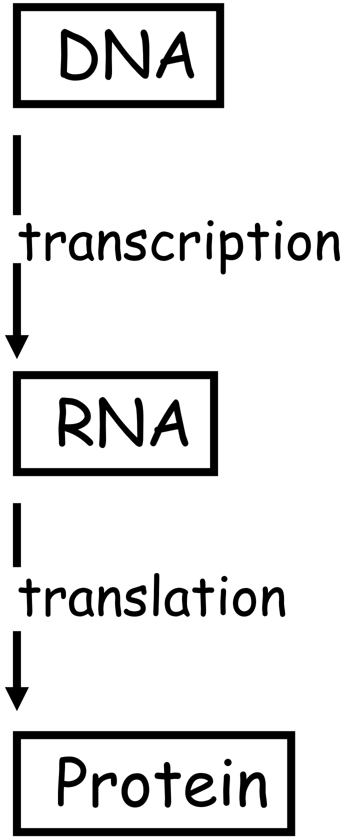
Lecture 08: Gene finding

```
aatgcatgcggctatgctaatgcatgcggctatgctaagctgggatccgatgacaatgcatgcggctatgctaatgcatgcggc
tatgcaagctgggatccgatgactatgctaagctgggatccgatgacaatgcatgcggctatgctaatgaatggtcttgggatt
taccttggaatgctaagctgggatccgatgacaatgcatgcggctatgctaatgaatggtcttgggatttaccttggaatatgc
taatgcatgcggctatgctaagctgggatccgatgacaatgcatgcggctatgctaatgcatgcggctatgcaagctgggatcc
gatgactatgctaagctgcggctatgctaatgcatgcggctatgctaagctgggatccgatgacaatgcatgcggctatgctaa
tgcatgcggctatgcaagctgggatcctgcggctatgctaatgaatggtcttgggatttaccttggaatgctaagctgggatcc
gatgacaatgcatgcggctatgctaatgaatggtcttgggatttaccttggaatatgctaatgcatgcggctatgctaagctgg
gaatgcatgcggctatgctaagctgggatccgatgacaatgcatgcggctatgctaatgcatgcggctatgcaagctgggatcc
gatgactatgctaagctgcggctatgctaatgcatgcggctatgctaagctcatgcggctatgctaagctgggaatgcatgcgg
ctatgctaagctgggatccgatgacaatgcatgcggctatgctaatgcatgcggctatgcaagctgggatccgatgactatgct
aagctgcggctatgctaatgcatgcggctatgctaagctcggctatgctaatgaatggtcttgggatttaccttggaatgctaa
gctgggatccgatgacaatgcatgcggctatgctaatgaatggtcttgggatttaccttggaatatgctaatgcatgcggctat
gctaagctgggaatgcatgcggctatgctaagctgggatccgatgacaatgcatgcggctatgctaatgcatgcggctatgcaa
gctgggatccgatgactatgctaagctgcggctatgctaatgcatgcggctatgctaagctcatgcgg
```

Slides adapted from Dr. Shaojie Zhang (University of Central Florida)

Central Dogma: DNA -> RNA -> Protein




DNA

transcription

RNA

translation

Protein

CCTGAGCCAACTATTGATGAA

CCUGAGCCAACUAUUGAUGAA

PEPTIDE

Translating Nucleotides into Amino Acids

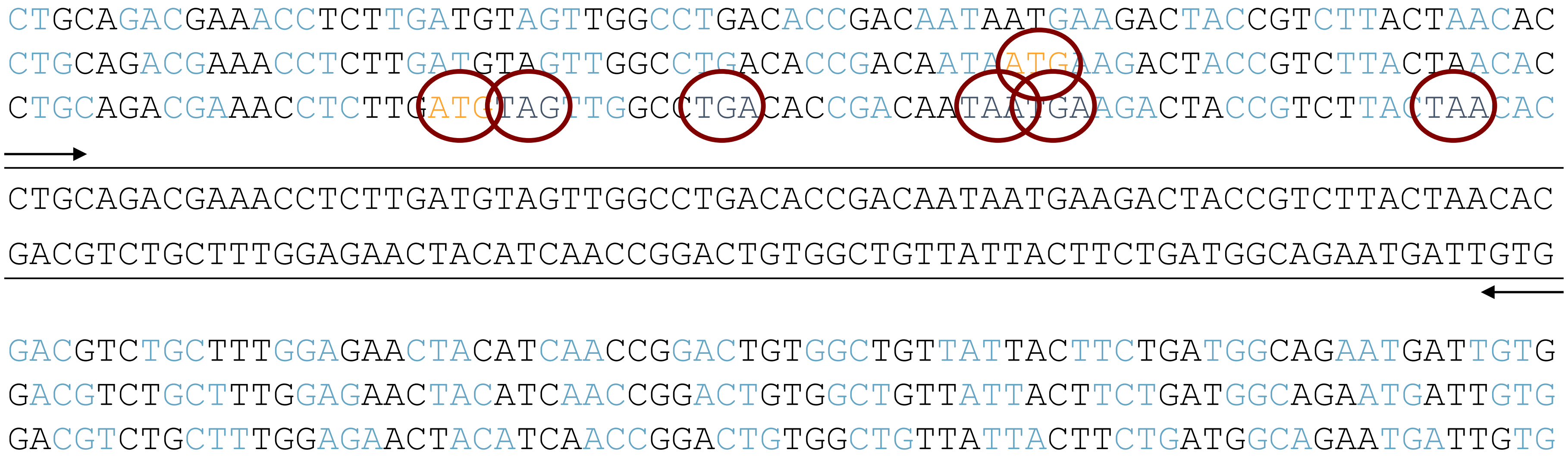
- Codon: 3 consecutive nucleotides
- $4^3 = 64$ possible codons
- Genetic code is degenerative and redundant
  - Includes start and stop codons
  - An amino acid may be coded by more than one codon

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

Six frames of DNA translation



CTGCAGACGAAACCTCTTGATGTAGTTGGCCTGACACCGACAATAATGAAGACTACCGTCTTACTAACAC

CTGCAGACGAAACCTCTTGATGTAGTTGGCCTGACACCGACAATAATGAAGACTACCGTCTTACTAACAC

CTGCAGACGAAACCTCTTGATGTAGTTGGCCTGACACCGACAATAATGAAGACTACCGTCTTACTAACAC

CTGCAGACGAAACCTCTTGATGTAGTTGGCCTGACACCGACAATAATGAAGACTACCGTCTTACTAACAC

GACGTCTGCTTTGGAGAACTACATCAACCGGACTGTGGCTGTTATTACTTCTGATGGCAGAATGATTGTG

GACGTCTGCTTTGGAGAACTACATCAACCGGACTGTGGCTGTTATTACTTCTGATGGCAGAATGATTGTG

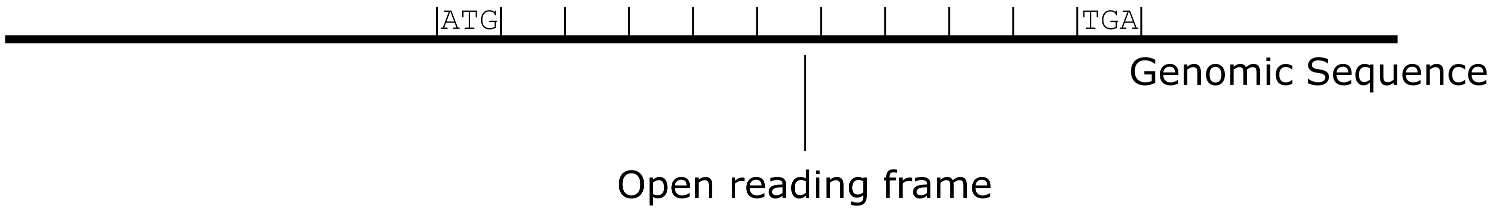
GACGTCTGCTTTGGAGAACTACATCAACCGGACTGTGGCTGTTATTACTTCTGATGGCAGAATGATTGTG

GACGTCTGCTTTGGAGAACTACATCAACCGGACTGTGGCTGTTATTACTTCTGATGGCAGAATGATTGTG

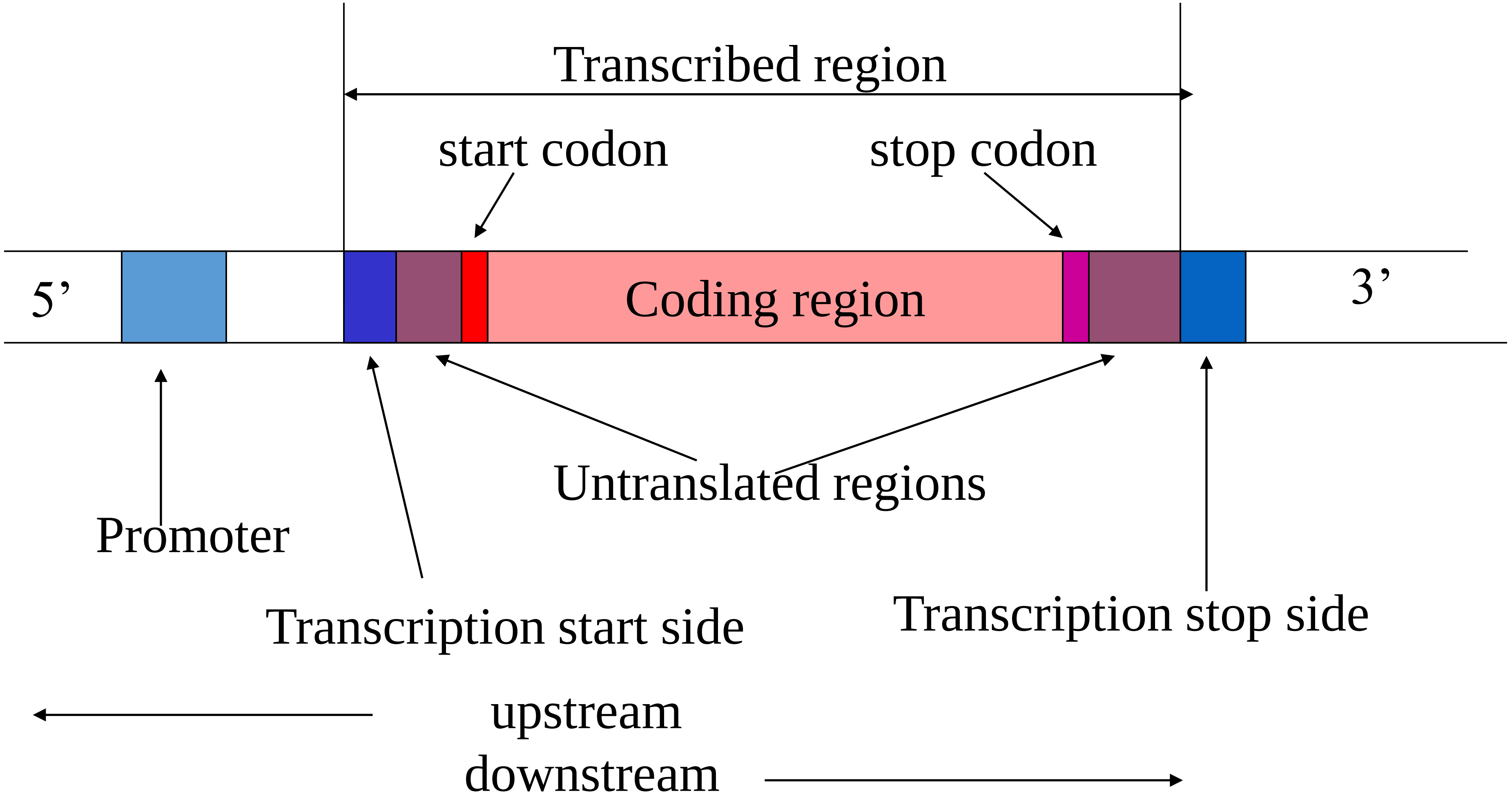
- stop codons – TAA, TAG, TGA
- start codons - ATG

Open reading frame (ORF)

- 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



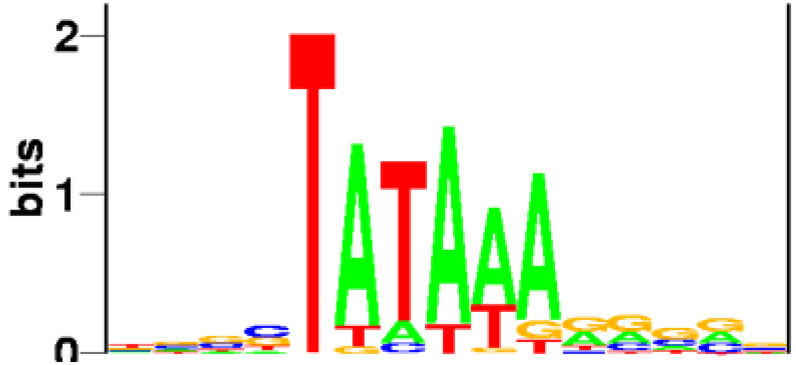
Prokaryotes gene structure

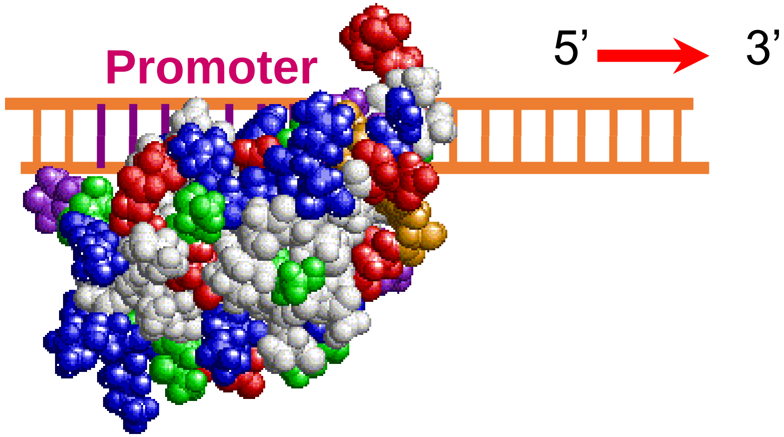


-k denotes kth base before transcription, +k denotes kth transcribed base

Promoter

- Promoters are DNA segments upstream of transcripts that initiate transcription

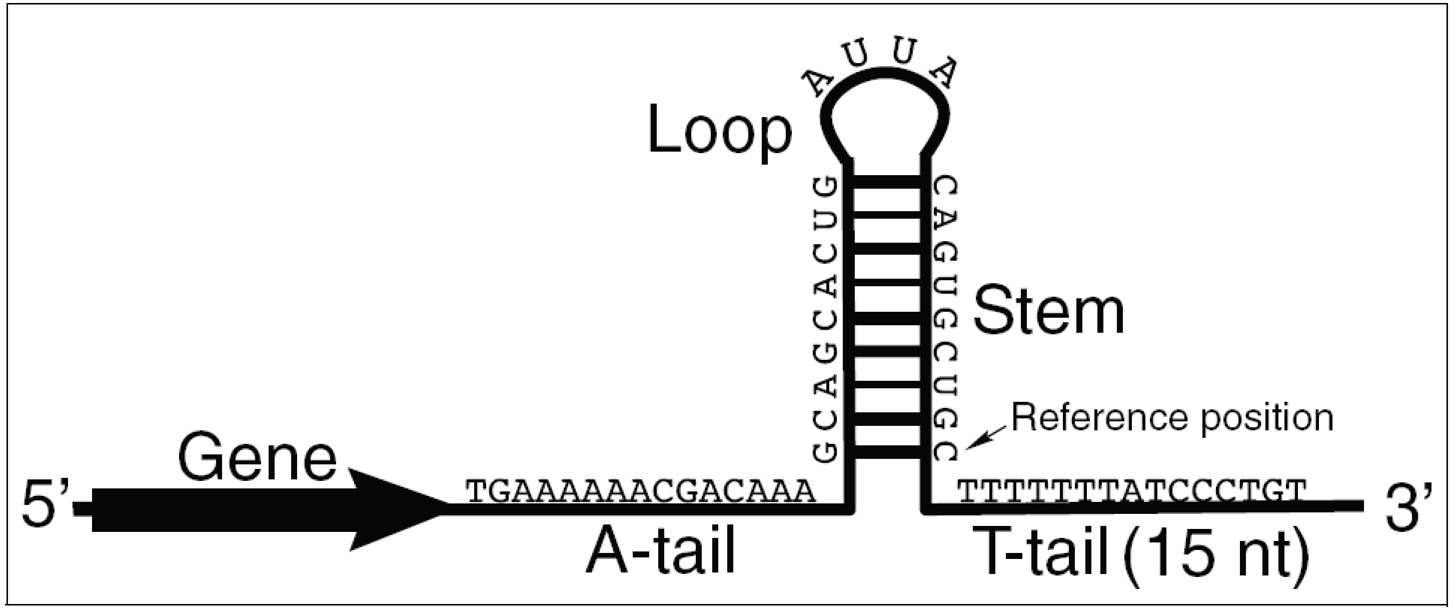




- Promoter *attracts* RNA Polymerase to the transcription start site

Other signals

- Terminator in prokaryotes: Rho-independent (intrinsic) transcription termination – G-C reach inverted repeat.
- Poly-A signal in eukaryotes



Long vs short genes

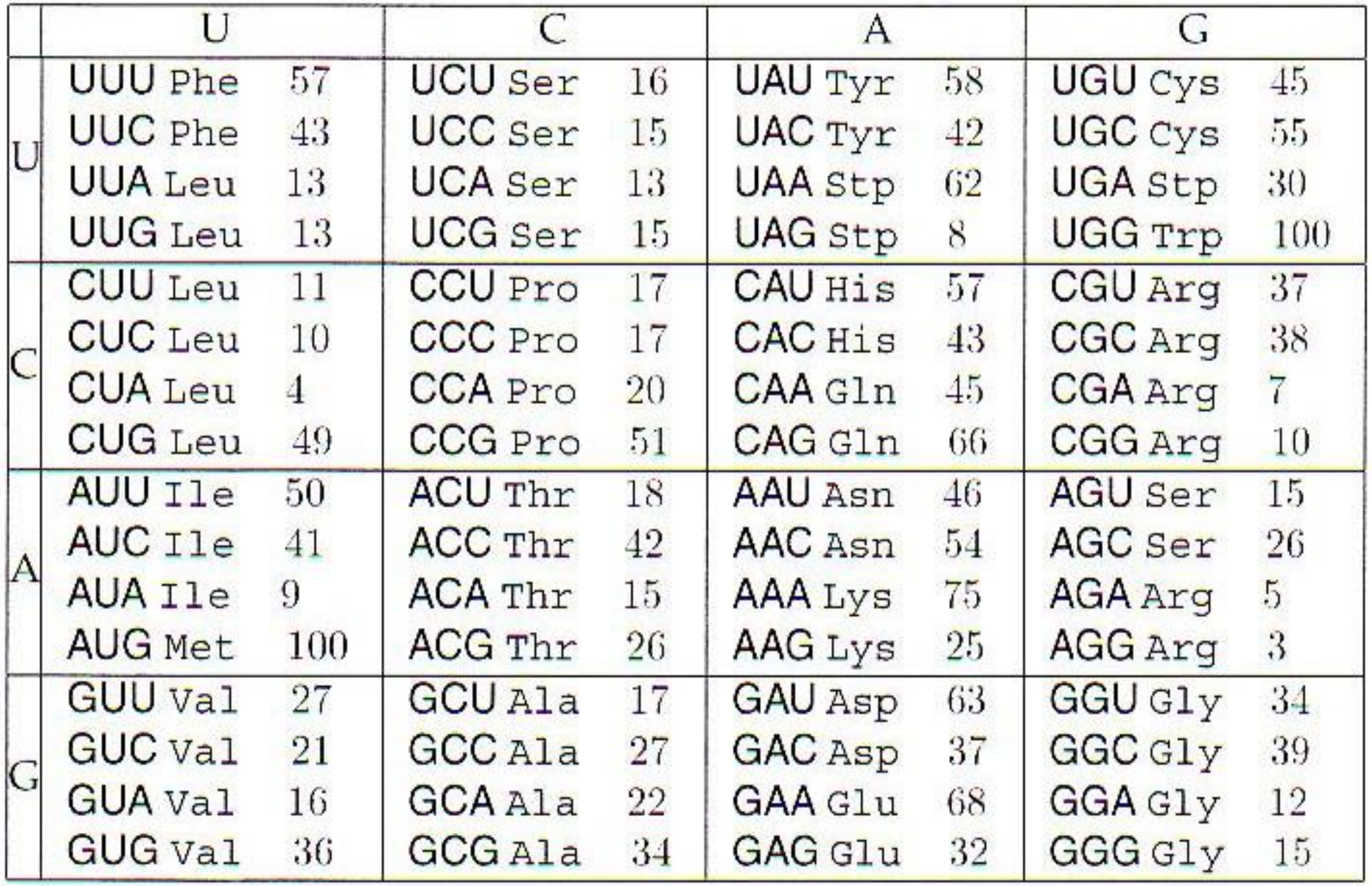
- Long open reading frames may be a gene
  - At random, we should expect one stop codon every (64/3) ~= 21 codons
  - **However**, genes are usually much longer than this

- 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

Codon usage

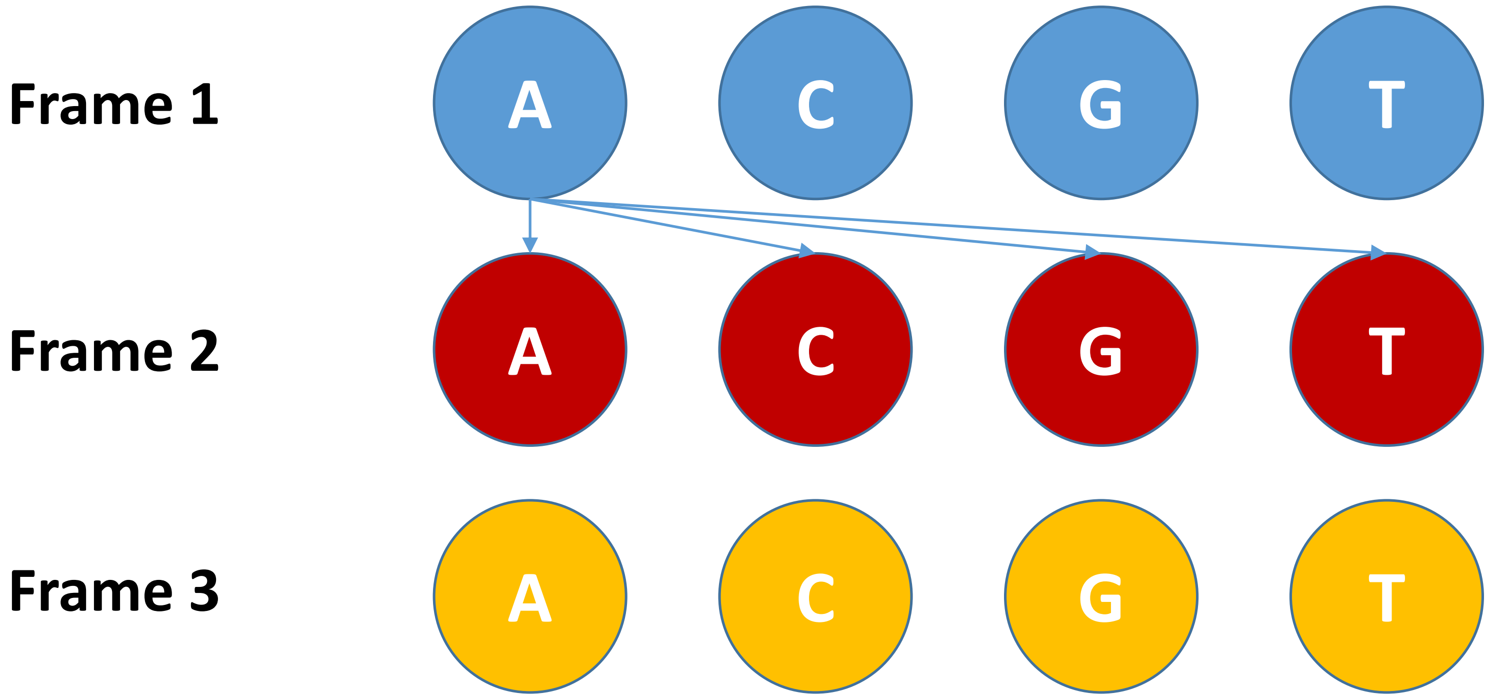
- Create a 64-element hash table and count the frequencies of codons in an ORF
- Amino acids typically have more than one codon, but in nature certain codons are more in use
- Uneven use of the codons may characterize a real gene
- This compensate for pitfalls of the ORF length test

Codon usage of the 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

GeneMark MM model



Frequencies for first order MM

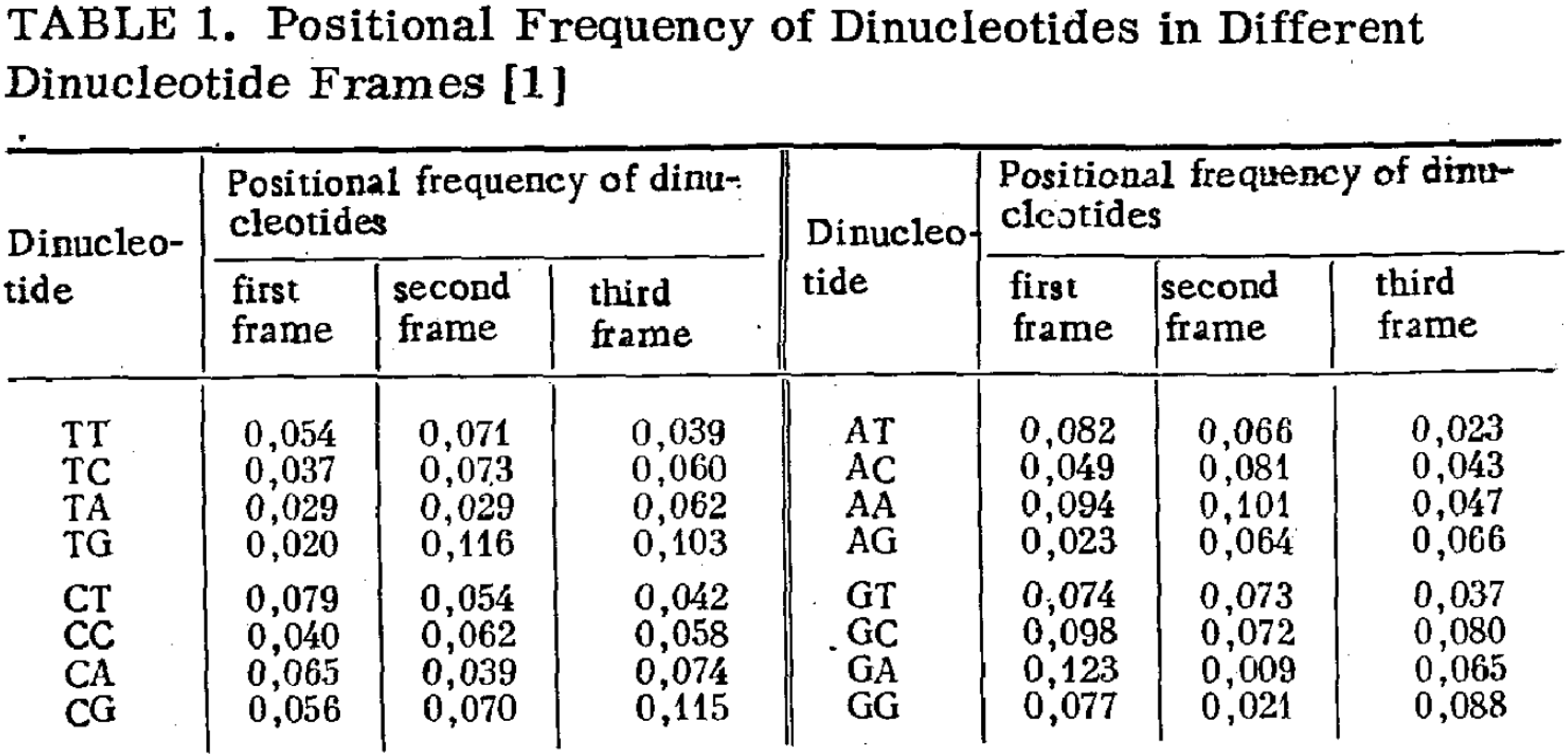


TABLE 1. Positional Frequency of Dinucleotides in Different Dinucleotide Frames [1]

Dinucleotide	Positional frequency of dinucleotides			Dinucleotide	Positional frequency of dinucleotides		
	first frame	second frame	third frame		first frame	second frame	third frame
TT	0,054	0,071	0,039	AT	0,082	0,066	0,023
TC	0,037	0,073	0,060	AC	0,049	0,081	0,043
TA	0,029	0,029	0,062	AA	0,094	0,101	0,047
TG	0,020	0,116	0,103	AG	0,023	0,064	0,066
CT	0,079	0,054	0,042	GT	0,074	0,073	0,037
CC	0,040	0,062	0,058	GC	0,098	0,072	0,080
CA	0,065	0,039	0,074	GA	0,123	0,009	0,065
CG	0,056	0,070	0,115	GG	0,077	0,021	0,088

Borodovskii et al. 1987

Frequencies for second order MM

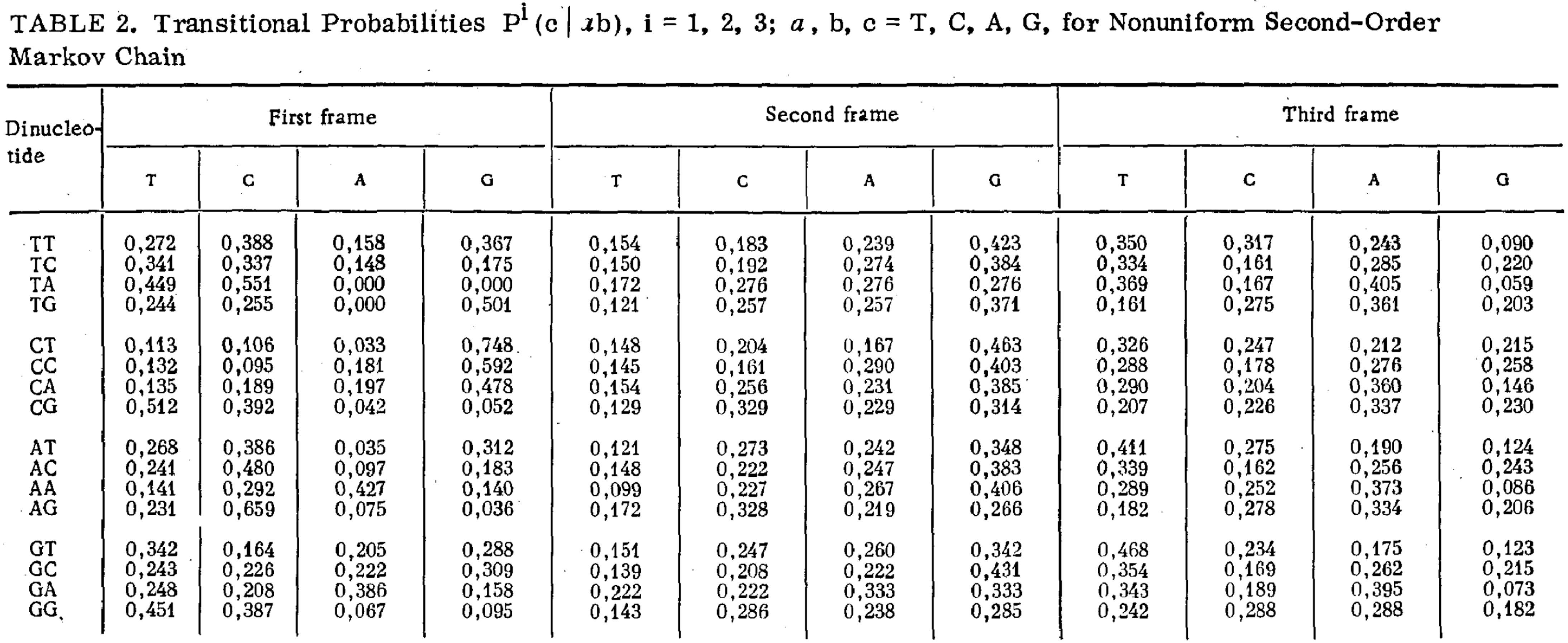
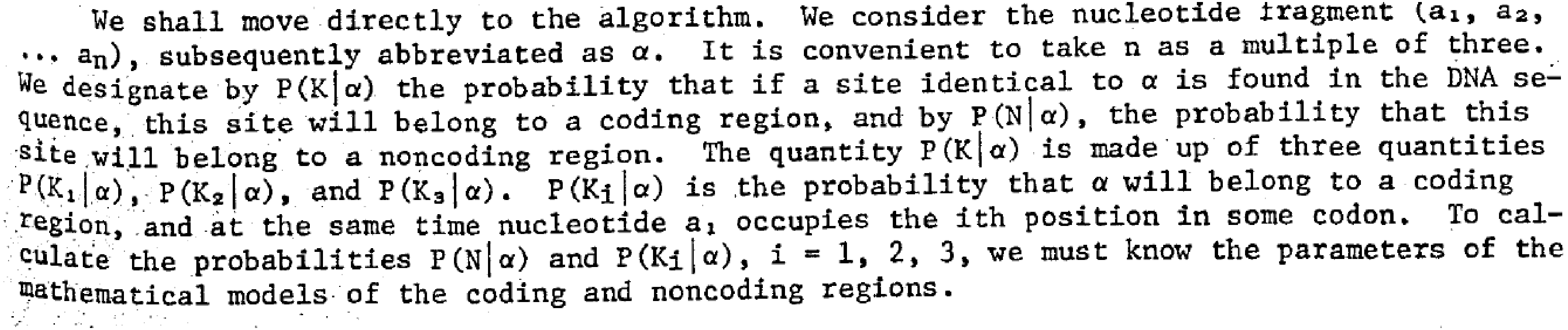


TABLE 2. Transitional Probabilities $P^{i}(c \mid ab)$, i = 1, 2, 3; a, b, c = T, C, A, G, for Nonuniform Second-Order Markov Chain

Dinucleotide	First frame				Second frame				Third frame			
	T	C	A	G	T	C	A	G	T	C	A	G
TT	0,272	0,388	0,158	0,367	0,154	0,183	0,239	0,423	0,350	0,317	0,243	0,090
TC	0,341	0,337	0,148	0,175	0,150	0,192	0,274	0,384	0,334	0,161	0,285	0,220
TA	0,449	0,551	0,000	0,000	0,172	0,276	0,276	0,276	0,369	0,167	0,405	0,059
TG	0,244	0,255	0,000	0,501	0,121	0,257	0,257	0,371	0,161	0,275	0,361	0,203
CT	0,113	0,106	0,033	0,748	0,148	0,204	0,167	0,463	0,326	0,247	0,212	0,215
CC	0,132	0,095	0,181	0,592	0,145	0,161	0,290	0,403	0,288	0,178	0,276	0,258
CA	0,135	0,189	0,197	0,478	0,154	0,256	0,231	0,385	0,290	0,204	0,360	0,146
CG	0,512	0,392	0,042	0,052	0,129	0,329	0,229	0,314	0,207	0,226	0,337	0,230
AT	0,268	0,386	0,035	0,312	0,121	0,273	0,242	0,348	0,411	0,275	0,190	0,124
AC	0,241	0,480	0,097	0,183	0,148	0,222	0,247	0,383	0,339	0,162	0,256	0,243
AA	0,141	0,292	0,427	0,140	0,099	0,227	0,267	0,406	0,289	0,252	0,373	0,086
AG	0,231	0,659	0,075	0,036	0,172	0,328	0,219	0,266	0,182	0,278	0,334	0,206
GT	0,342	0,164	0,205	0,288	0,151	0,247	0,260	0,342	0,468	0,234	0,175	0,123
GC	0,243	0,226	0,222	0,309	0,139	0,208	0,222	0,431	0,354	0,169	0,262	0,215
GA	0,248	0,208	0,386	0,158	0,222	0,222	0,333	0,333	0,343	0,189	0,395	0,073
GG	0,451	0,387	0,067	0,095	0,143	0,286	0,238	0,285	0,242	0,288	0,288	0,182

Borodovskii et al. 1987

Computing the likelihood of a nucleotide fragment



We shall move directly to the algorithm. We consider the nucleotide fragment (a_1, a_2, ... a_n), subsequently abbreviated as α. It is convenient to take n as a multiple of three. We designate by $P(K|\alpha)$ the probability that if a site identical to α is found in the DNA sequence, this site will belong to a coding region, and by $P(N|\alpha)$, the probability that this site will belong to a noncoding region. The quantity $P(K|\alpha)$ is made up of three quantities $P(K_1|\alpha)$, $P(K_2|\alpha)$, and $P(K_3|\alpha)$. $P(K_i|\alpha)$ is the probability that α will belong to a coding region, and at the same time nucleotide a_1 occupies the ith position in some codon. To calculate the probabilities $P(N|\alpha)$ and $P(K_i|\alpha)$, $i = 1, 2, 3$, we must know the parameters of the mathematical models of the coding and noncoding regions.

Borodovskii et al. 1987

Computing the likelihood

non-protein-coding region

$$P(\alpha|N) = P_0(a_1)P(a_2|a_1)\cdot\ldots\cdot P(a_n|a_{n-1}).$$

protein-coding region

$$P(\alpha|K_1) = P_0^1(a_1)P^1(a_2|a_1)P^2(a_3|a_2)\cdot\ \ldots\ \cdot P^2(a_n|a_{n-1}),$$

$$P(\alpha|K_2) = P_0^2(a_1)P^2(a_2|a_1)P^3(a_3|a_2)\cdot\ \ldots\ \cdot P^3(a_n|a_{n-1}),$$

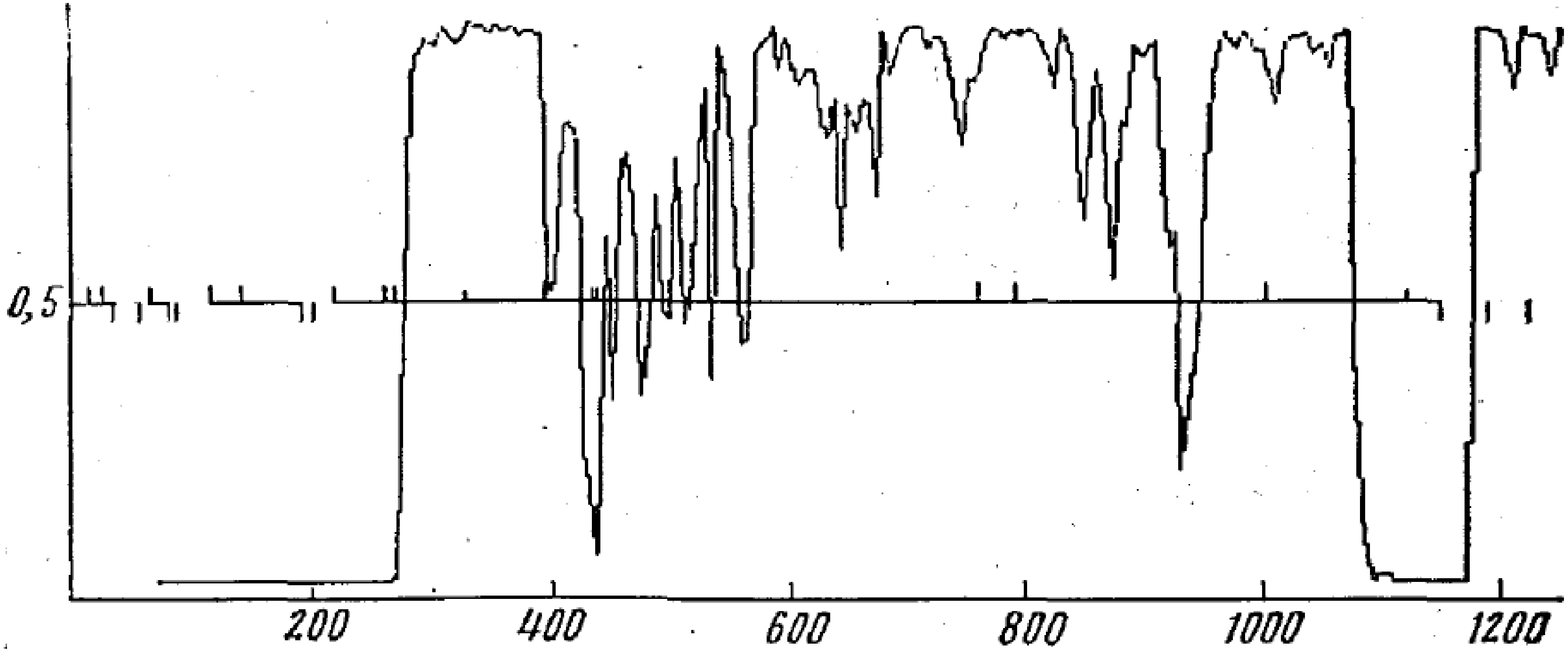
$$P(\alpha|K_3) = P_0^3(a_1)P^3(a_2|a_1)P^1(a_3|a_2)\cdot\ \ldots\ \cdot P^1(a_n|a_{n-1}).$$

Posterior likelihood

$$P(K_i|\alpha) = \frac{P(\alpha|K_i)P(K_i)}{\sum_i P(\alpha|K_i)P(K_i) + P(\alpha|N)P(N)}; \quad i = 1, 2, 3.$$

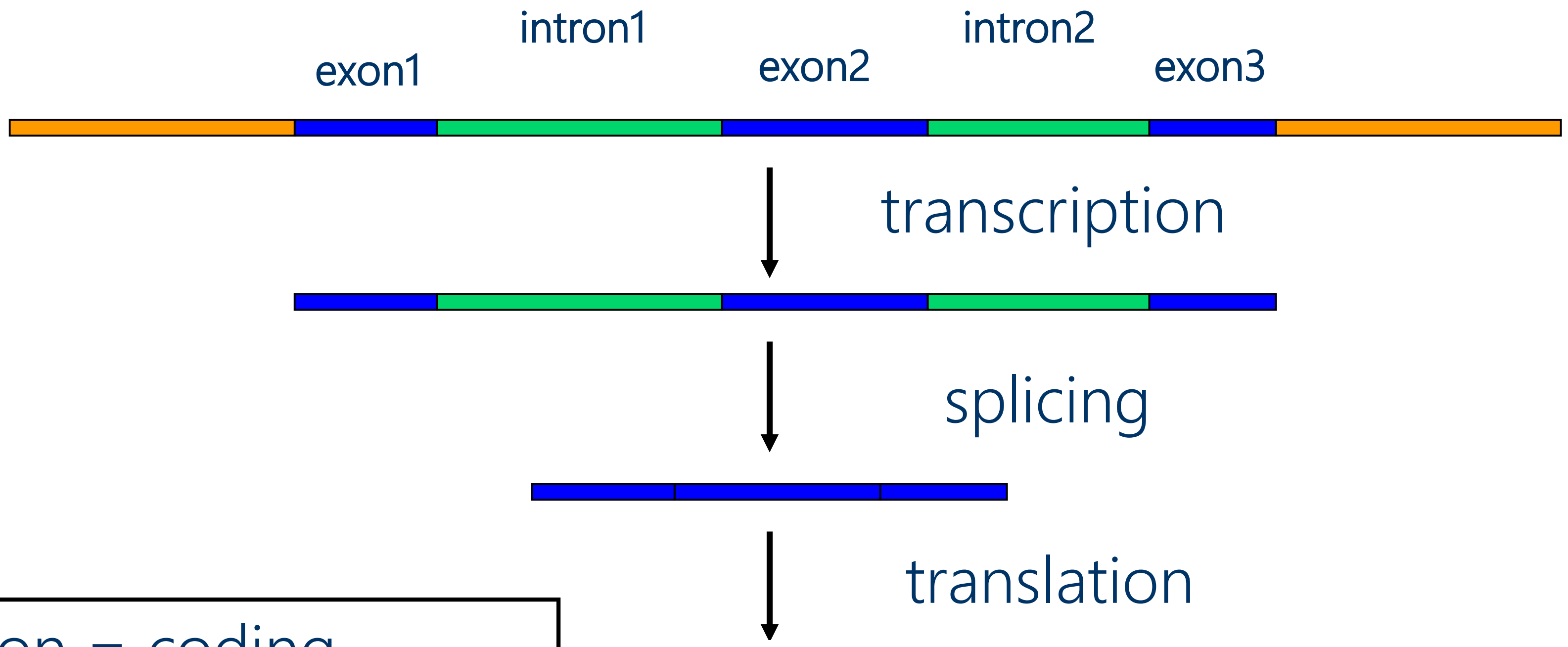
Borodovskii et al. 1987

Scan the genome



Borodovskii et al. 1987

Eukaryotes gene prediction



intron1

intron2

exon1

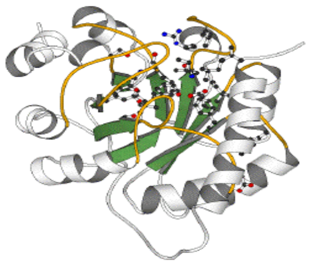
exon2

exon3

transcription

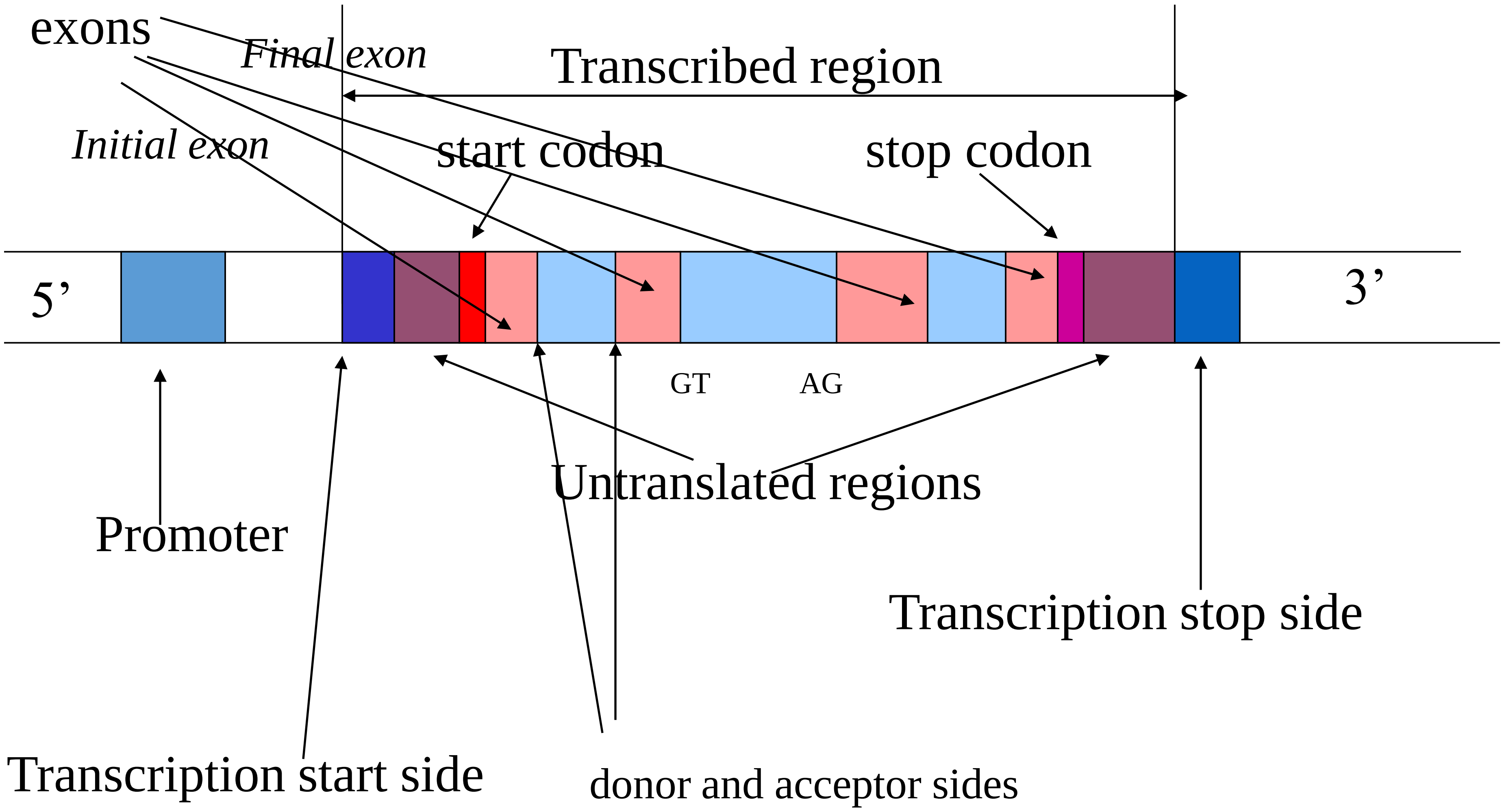
splicing

translation

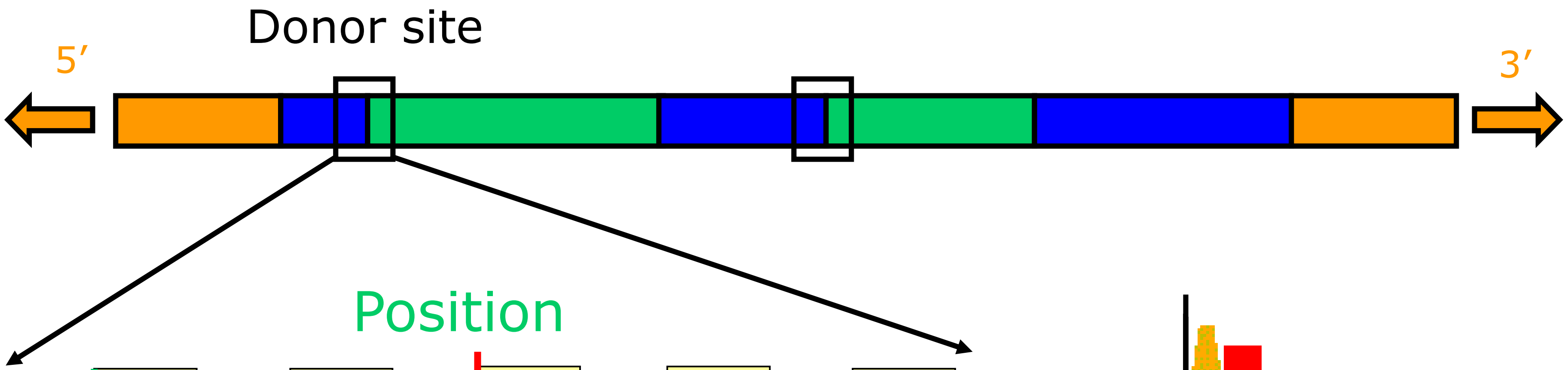


exon = coding
intron = non-coding

Eukaryotes gene structure

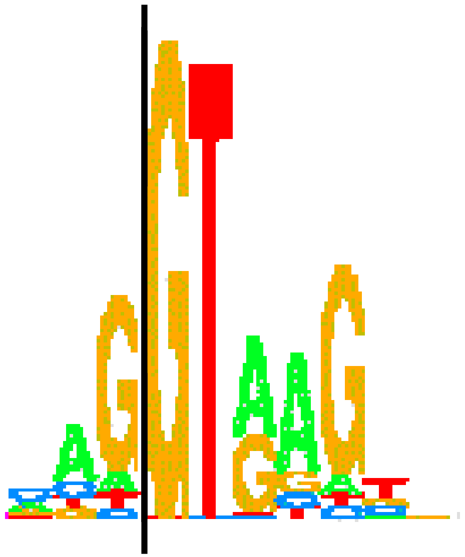


Splicing Signals for eukaryotes



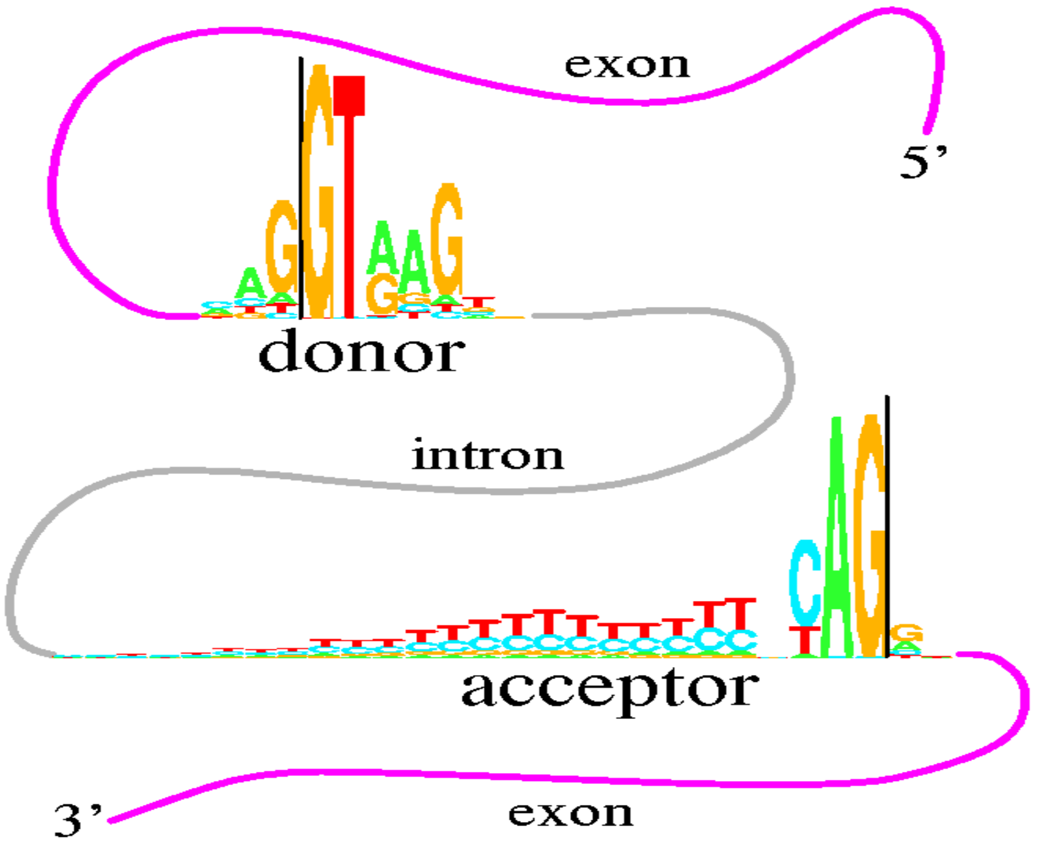
Position

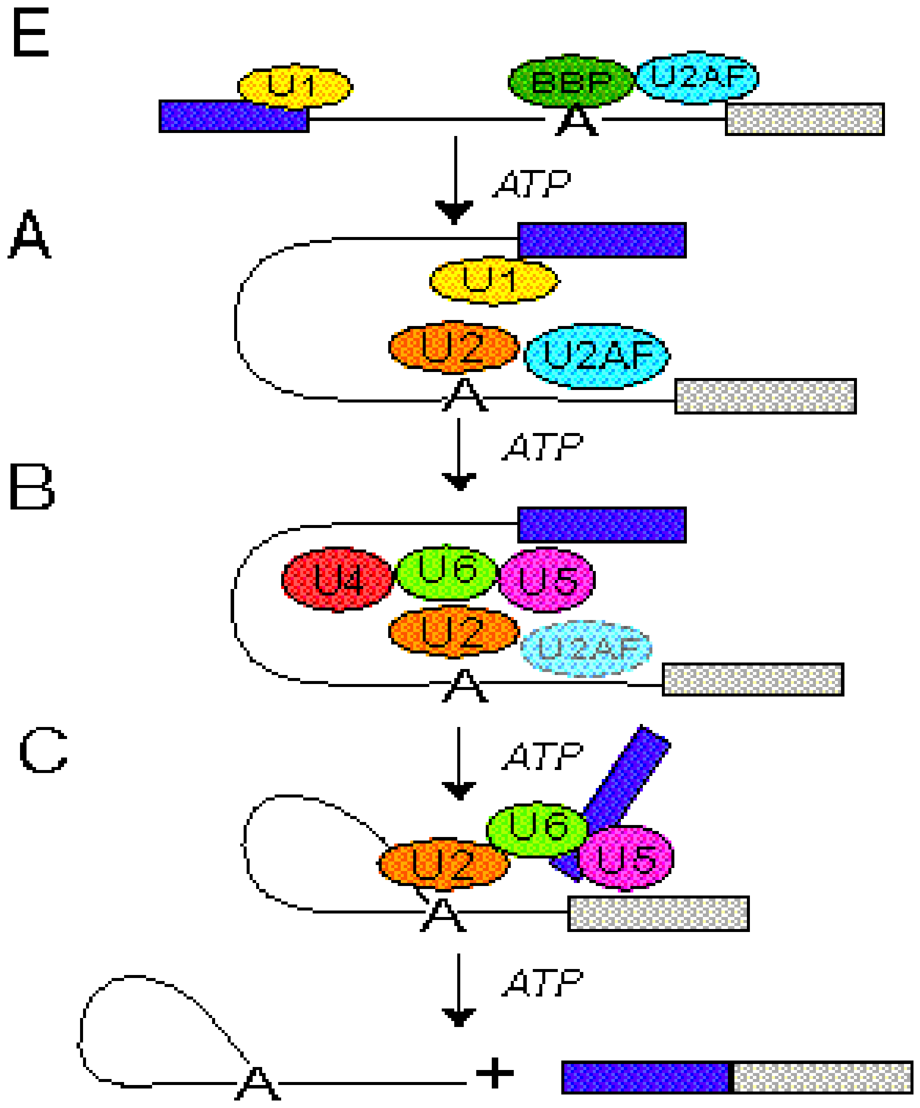
%	-8	…	-2	-1	0	1	2	…	17
A	26	…	60	9	0	1	54	…	21
C	26	…	15	5	0	1	2	…	27
G	25	…	12	78	99	0	41	…	27
T	23	…	13	8	1	98	3	…	25



Splice site signals

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)

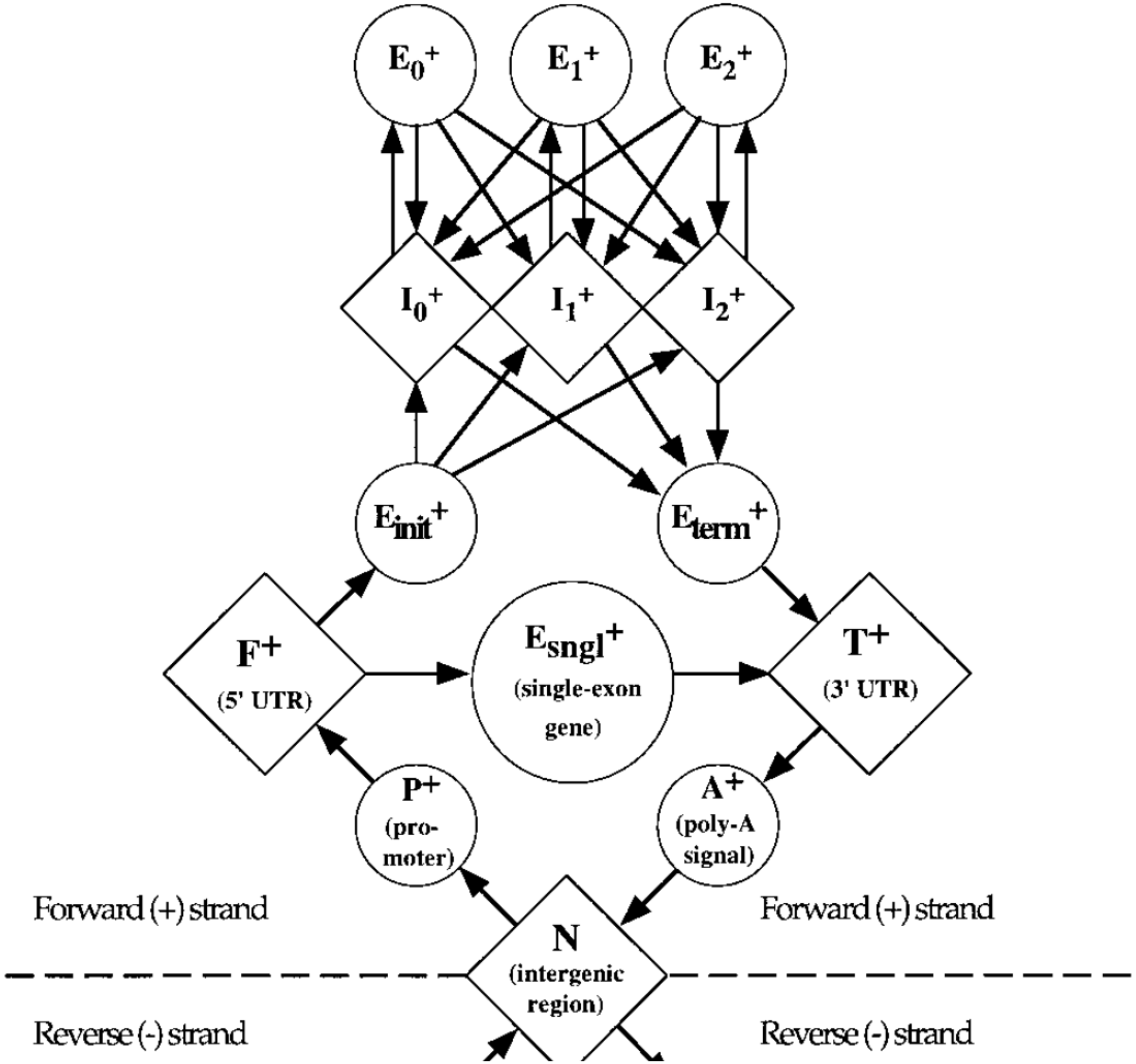




(http://genes.mit.edu/chris/)

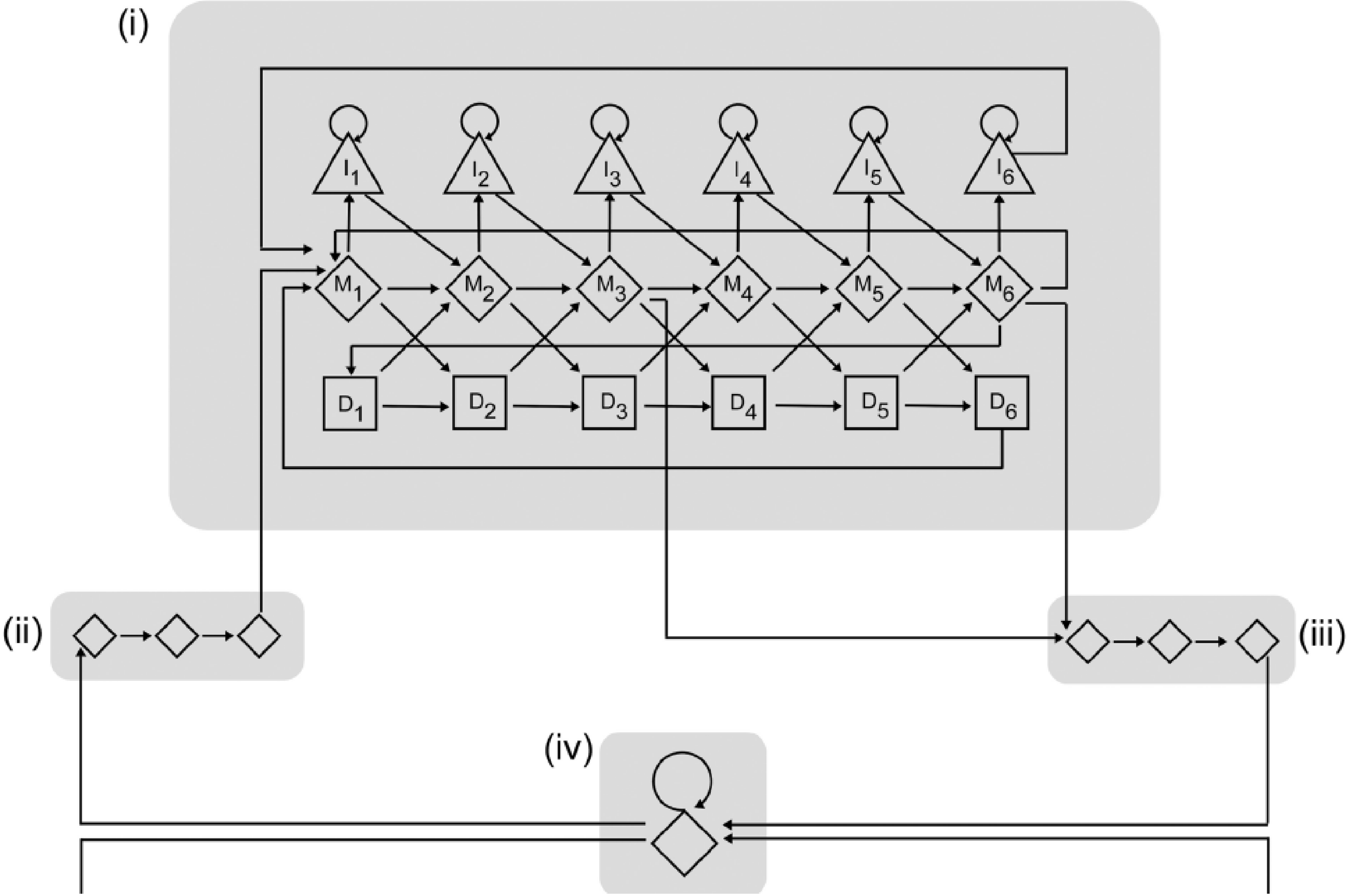
GeneScan model

- States- correspond to different functional units of a genome (promoter region, intron, exon,….)
- The states for introns and exons are subdivided according to “phase” three frames.
- There are two symmetric sub modules for forward and backward strands.



Burge and Karlin, J. Molecular Biology, 1997

FragGeneScan model



Rho et al. *NAR* 2010