Gene Structure & Gene Finding

Hosein Mohimani

GHC7717

hoseinm@andrew.cmu.edu

Nucleic Acids Research, 1998, Vol. 26, No. 4 1107–1115

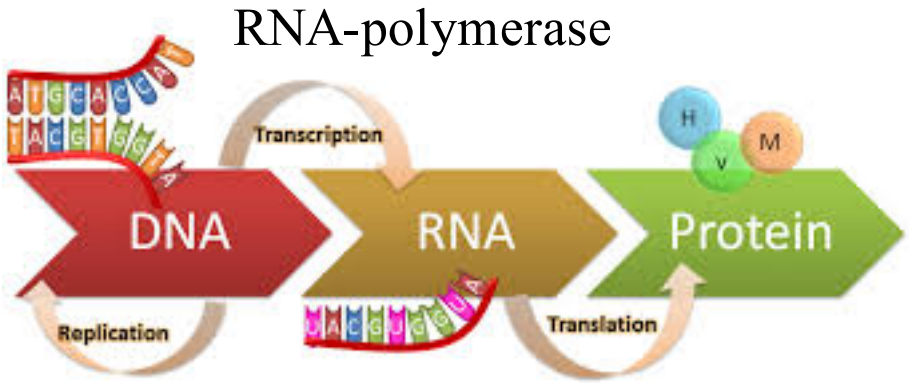
GeneMark.hmm: new solutions for gene finding

Alexander V. Lukashin and Mark Borodovsky[1,*]

School of Biology and [1]Schools of Biology and Mathematics, Georgia Institute of Technology, Atlanta, GA 30332-0230, USA

Received August 14, 1997; Revised and Accepted December 30, 1997

Gene finding



Genes are the protein encoding parts of DNA

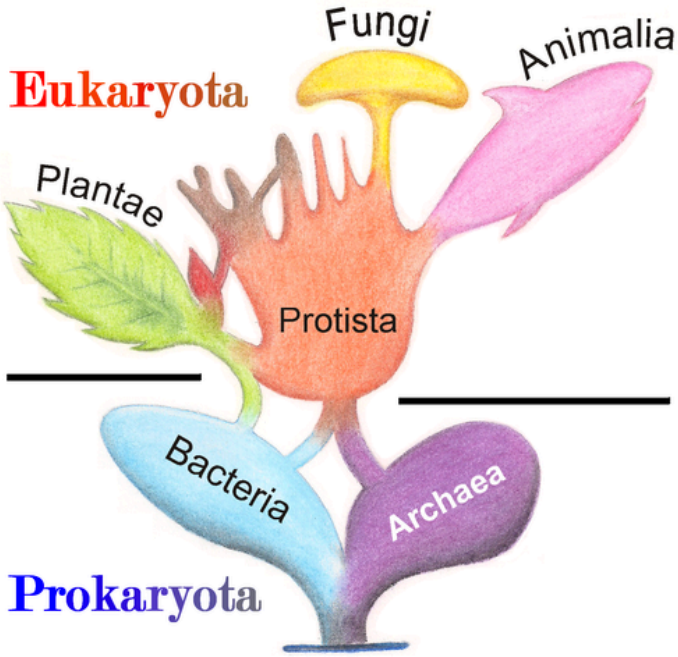
Gene prediction is key step for genome annotation

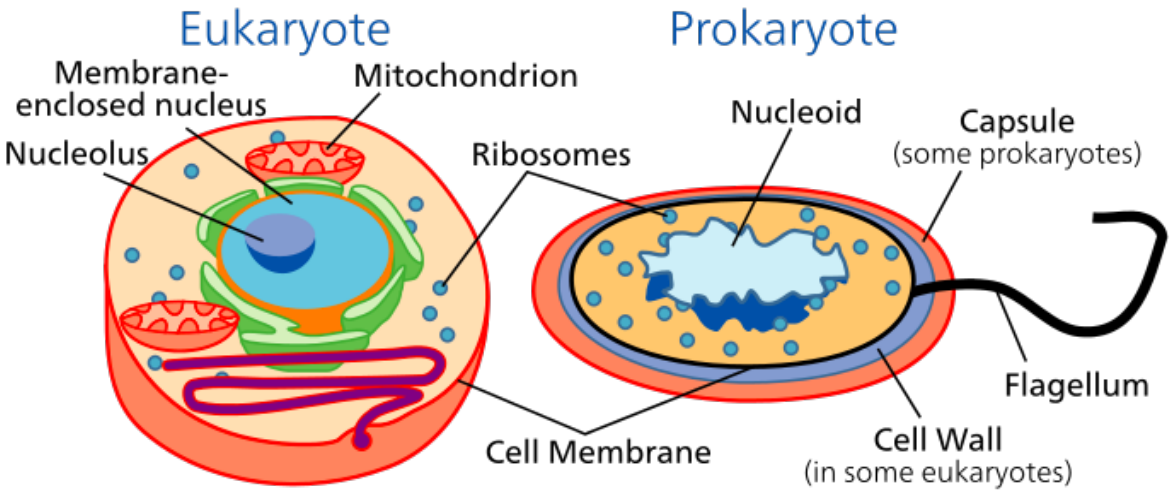
How can we go from DNA to genes ?

Eukaryotes versus Prokaryotes

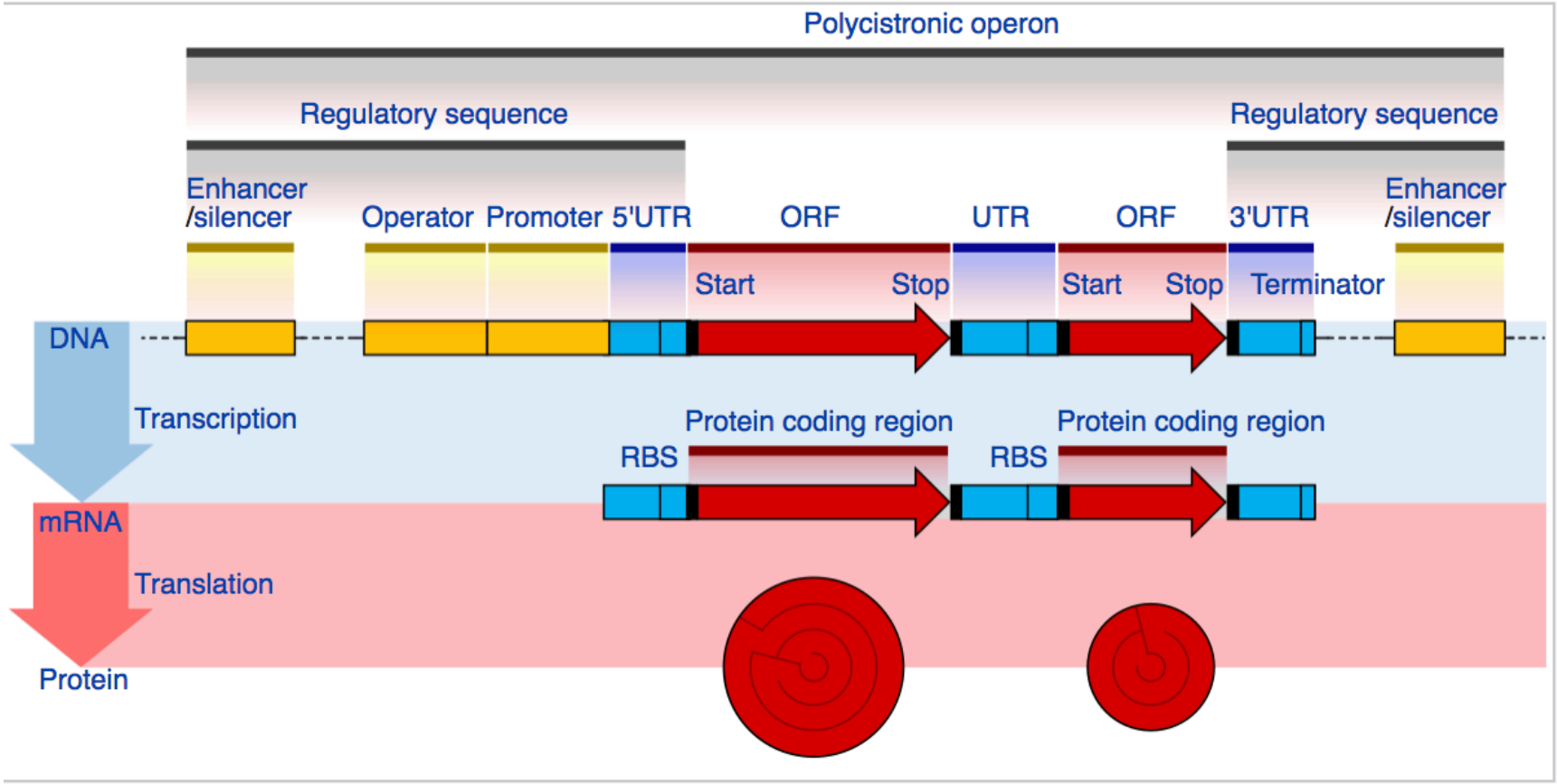
Prokaryotes lack a cell nucleus (karyon), or any other membrane-bound organelles

Gene structure is different between prokaryotes & Eukaryotes



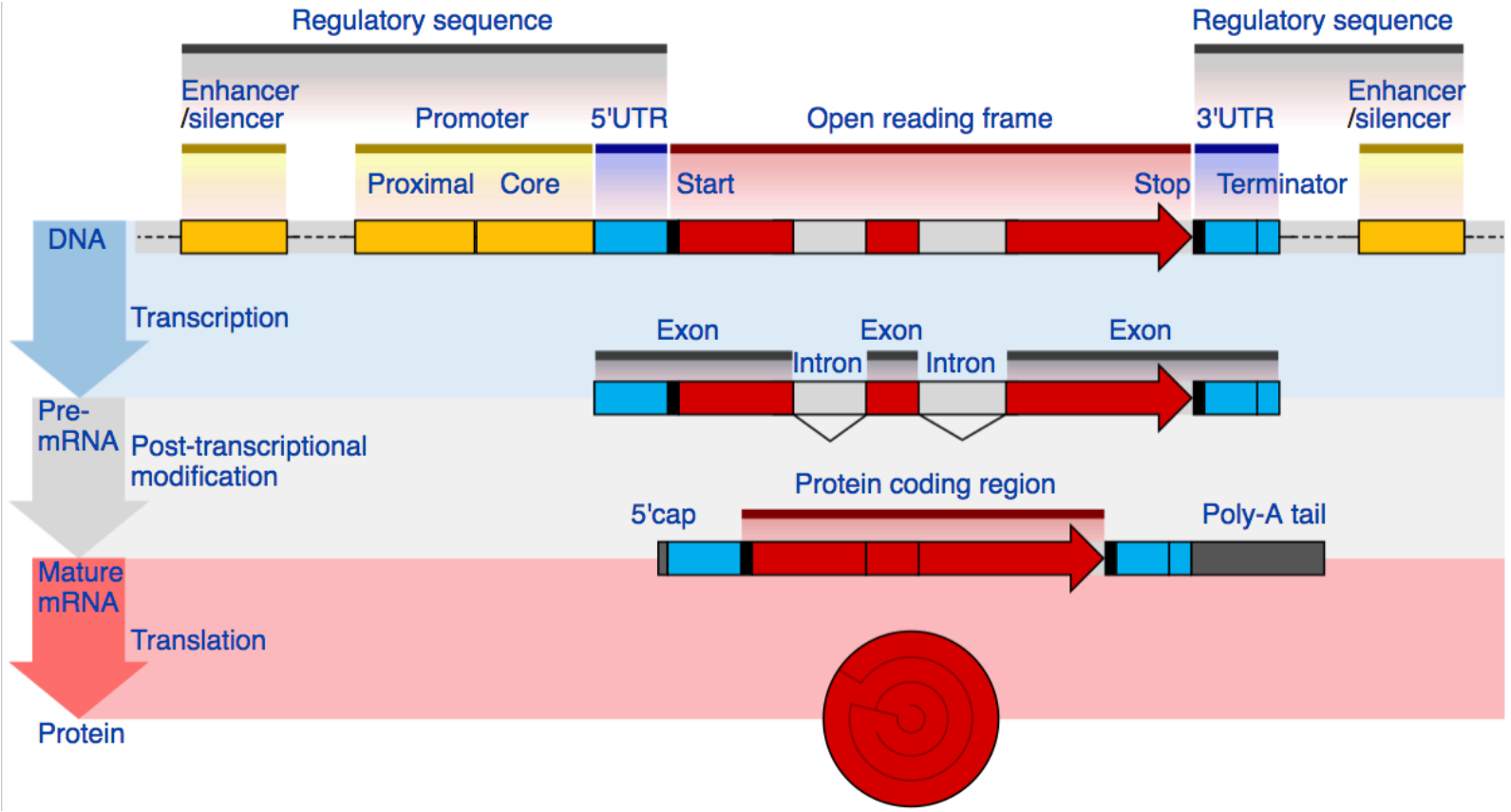


Gene Structure in prokaryotes



- **Regulatory Sequences** controls expression of the genes
- **Transcription factor** binds to regulatory sequence
- RBS : Ribosome binding site

Gene Structure in Eukaryotes



- Main difference : mRNA goes through post-translational modifications, and introns are cut out

Genetic Code- Table

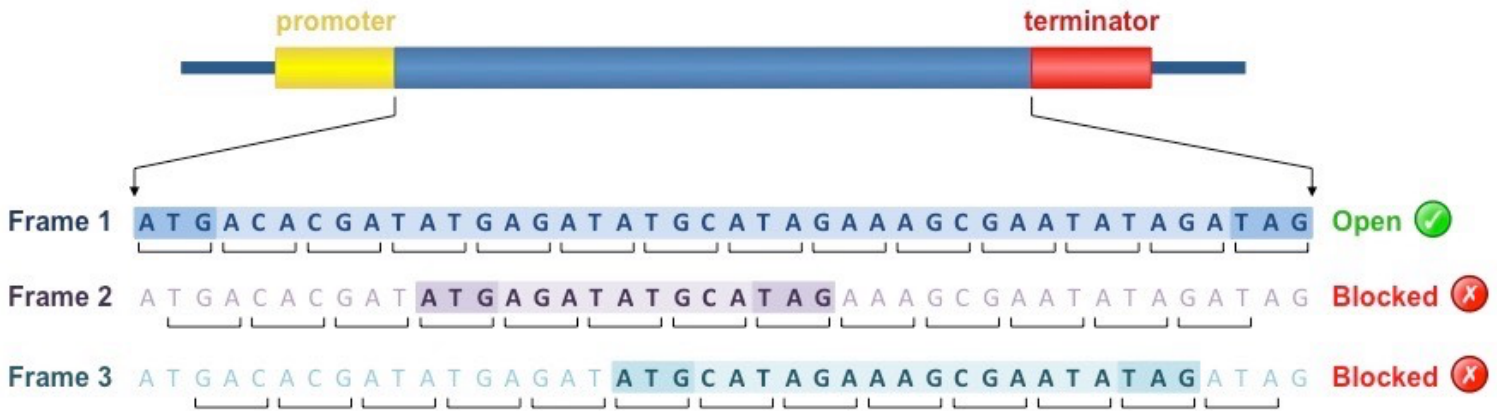
Stop Codon

1st letter	Second Letter: U		Second Letter: C		Second Letter: A		Second Letter: G		3rd letter
U	UUU	Phe	UCU	Ser	UAU	Tyr	UGU	Cys	U
	UUC		UCC		UAC		UGC		C
	UUA	Leu	UCA		UAA	Stop	UGA	Stop	A
	UUG		UCG		UAG	Stop	UGG	Trp	G
C	CUU	Leu	CCU	Pro	CAU	His	CGU	Arg	U
	CUC		CCC		CAC		CGC		C
	CUA		CCA		CAA	Gln	CGA		A
	CUG		CCG		CAG		CGG		G
A	AUU	Ile	ACU	Thr	AAU	Asn	AGU	Ser	U
	AUC		ACC		AAC		AGC		C
	AUA		ACA		AAA	Lys	AGA	Arg	A
	AUG	Met	ACG		AAG		AGG		G
G	GUU	Val	GCU	Ala	GAU	Asp	GGU	Gly	U
	GUC		GCC		GAC		GGC		C
	GUA		GCA		GAA	Glu	GGA		A
	GUG		GCG		GAG		GGG		G

Start Codon

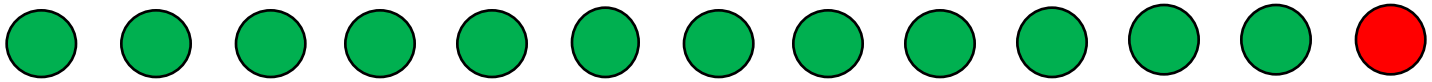
Gene Finding in Prokaryotes

- In Prokaryotes, there is no Introns
- Prokaryotes have small genomes
- Genes are the same as ORFs
- Majority of the genome is protein coding (efficient)
- Genes can overlap between different frames
- Some genes are short



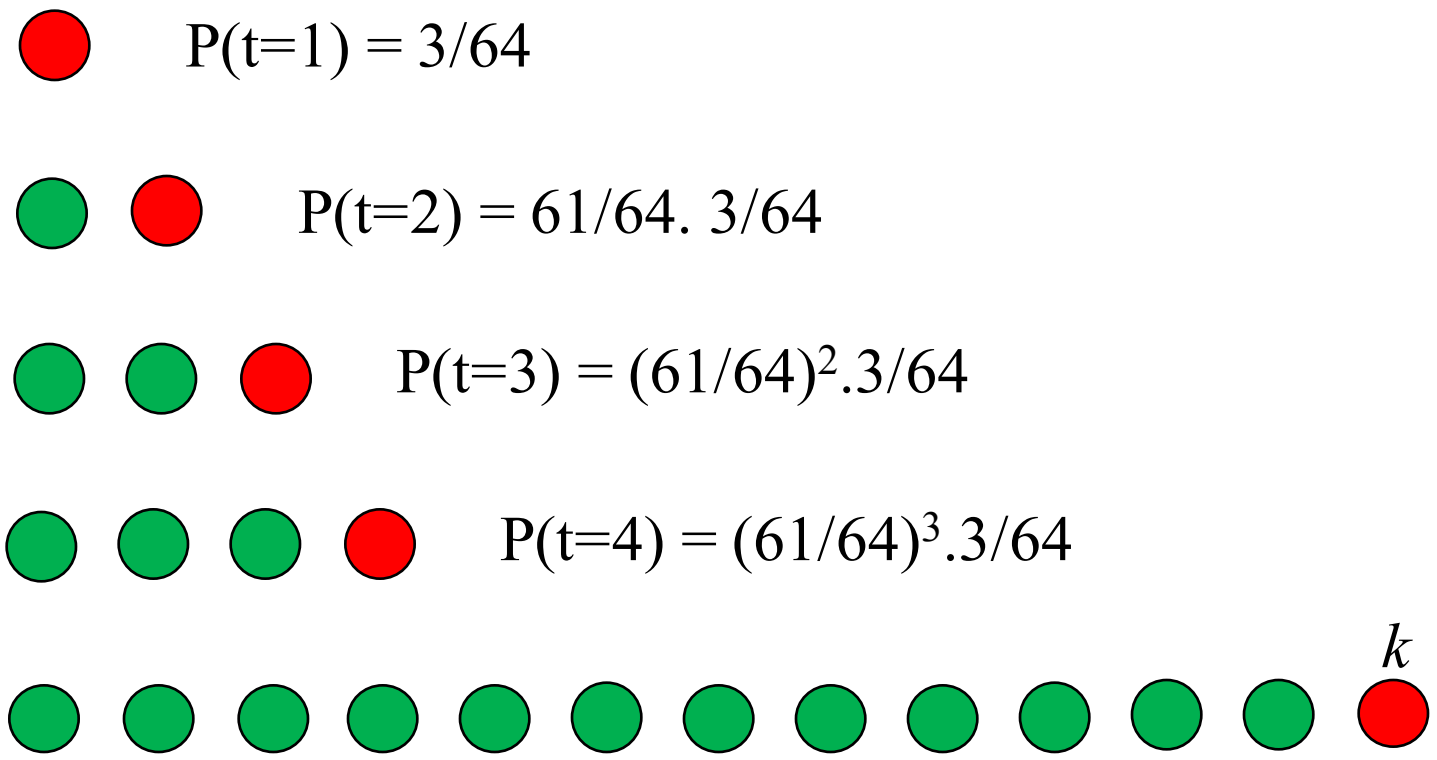
A simple strategy for gene finding

- Lets compute the average length of an ORF, if nucleotides where drawn by random chance
- We have a coin with 61/64 probability of green (non-stop-codon), and 3/64 probability of red (stop-codon).
- We toss this coin till there is a red.
- How many tosses, in average, does it take to reach red?



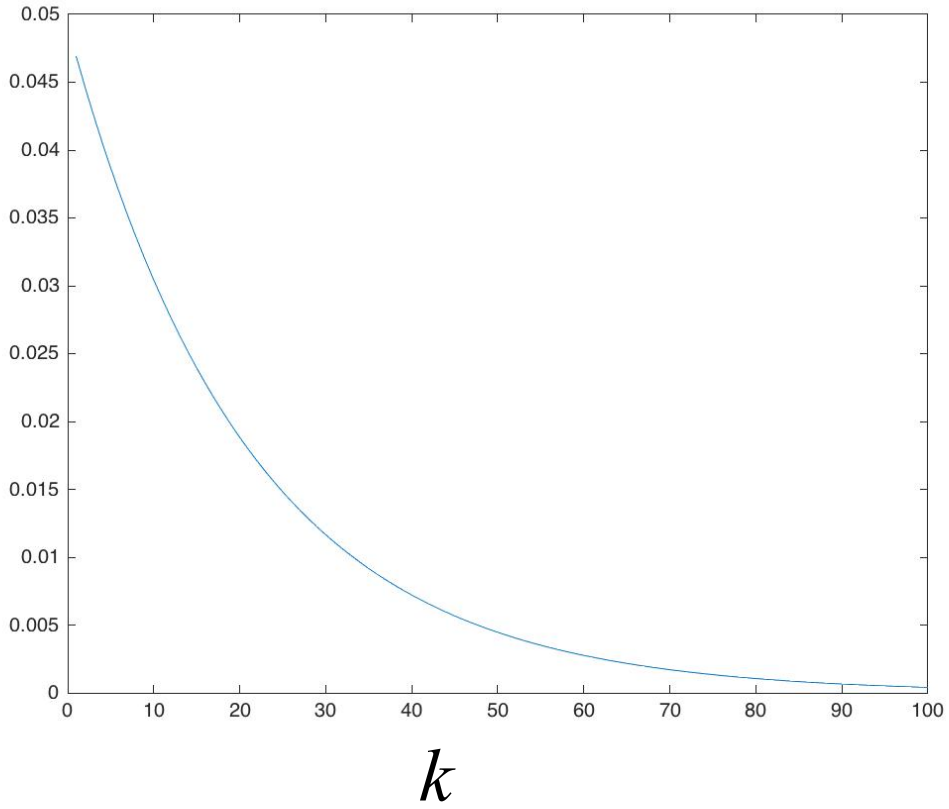
Average ORF length

- Lets *t* be a random variable that account for the number of tossed it takes to reach a red. What is the probability of *t=1* ?



Average ORF length

$P(t=k)$



What is the expected value of the ORF length ?

$$\mathrm{E(t)} = \sum_{k=0}^{\infty} k.P(t=k) = \sum_{k=0}^{\infty} k.\left(\frac{3}{64}\right)\left(\frac{61}{64}\right)^{k-1} = \frac{1}{1-61/64} = 21.33$$

Gene Finding In Prokaryotes : A simple approach

ACGAACGATATTGGGACGATTGACGGTAC

ACGAACGATATTGGGACGATTGACGGTAC
Val Pro Ser Ile Val Pro Ile Ser Phe

Frame 1

ACGAACGATATTGGGACGATTGACGGTAC
Arg Thr Ile Leu Gly Arg Leu Thr Val

Frame 2

Glu Arg Tyr Trp Asp Asp * Arg Tyr

Frame 3

Gene Finding In Prokaryotes : A simple approach

ACGAACGATATTGGGACGATTGACGGTAC sequence

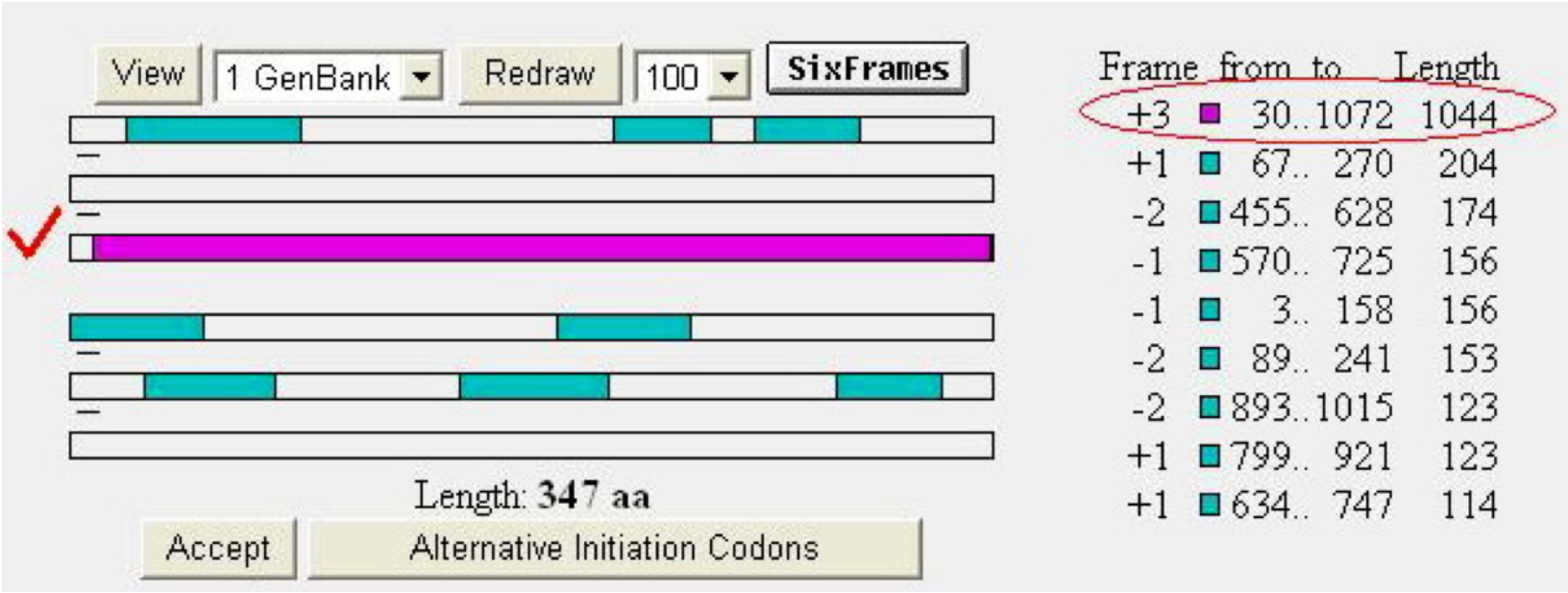
GCACCGTCAATCGTCCCAATATCGTTCGT reverse-complement

Val Pro Ser Ile Val Pro Ile Ser Phe Frame -1

Tyr Arg Gln Ser Ser Gln Tyr Arg Ser Frame -2

Thr Val Asn Arg Pro Asn Ile Val Arg Frame -3

Gene Finding In Prokaryotes : A simple approach



- Looks for long ORFs (between start and top codon, no stop codon in between)
- By random chance, one in very 20 codon is stop codon
- Long ORFs are statistically significant

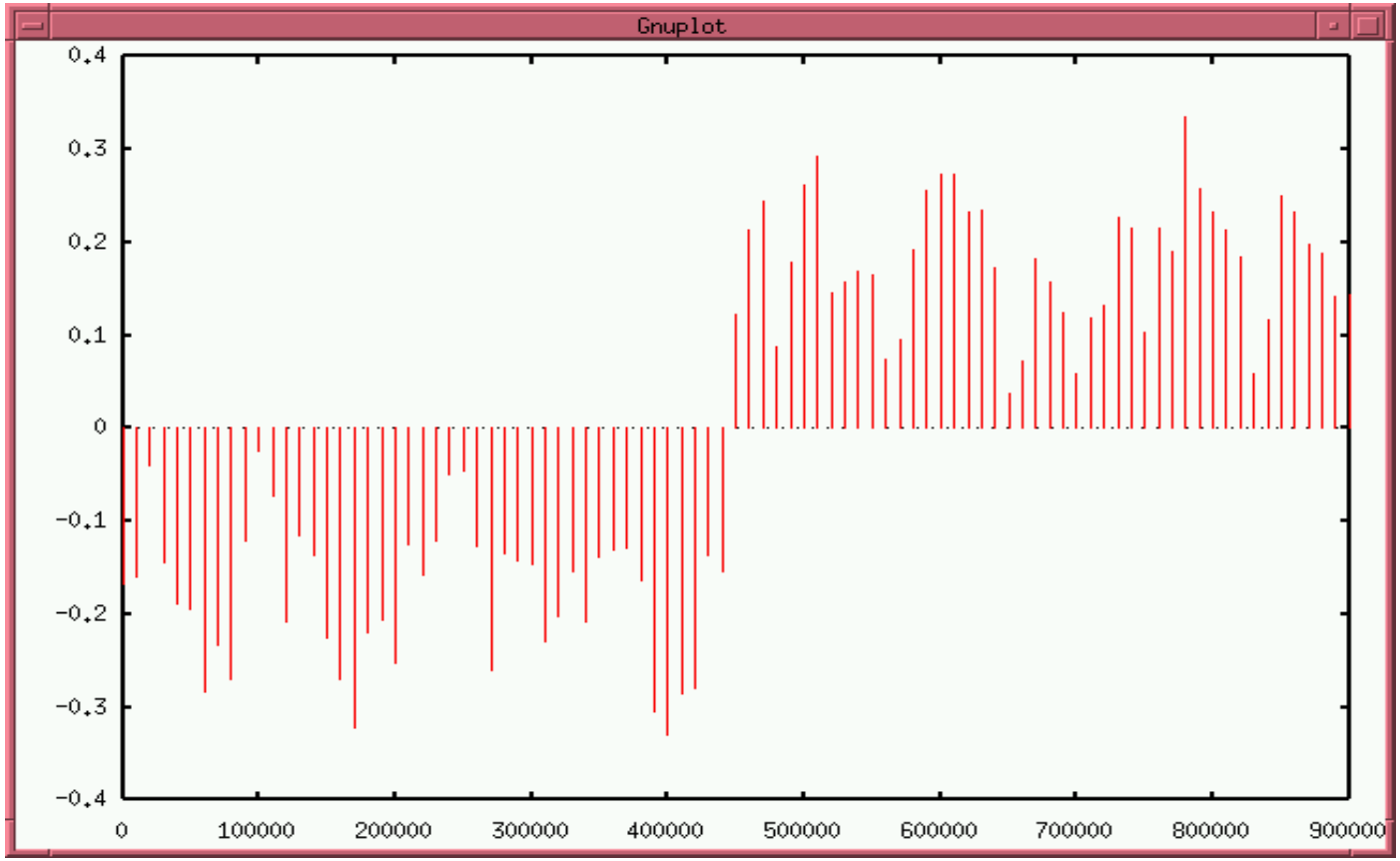
Cons

- Miss small genes, over-predict long ones

- Can we use other signals ?
  - Promoters sequence, transcription factor binding sites
  - Ribosome binding sites
  - Periodicities in protein encoding DNA
  - K-mer statistics (e.g. high GC content)

Improving gene detection

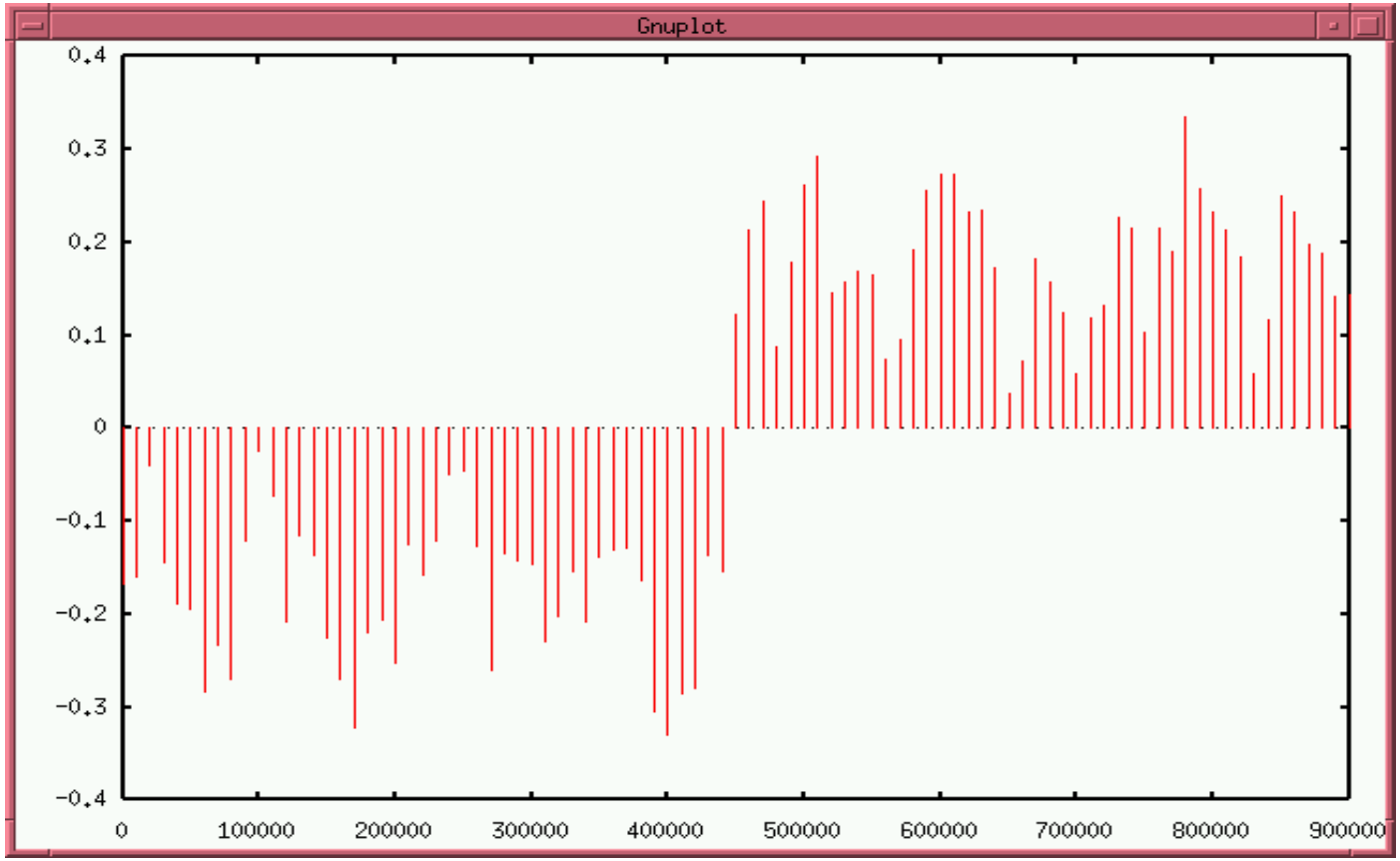
• In bacteria, genes have a higher ratio of G & C, in compare to the non-gene regions.



GC Content

Improving gene detection

• In bacteria, genes have a higher ratio of G & C, in compare to the non-gene regions.



GC bias

Improving gene detection

- Different codons show up with different probabilities

		Non-coding	coding			Non-coding	coding			Non-coding	coding			Non-coding	coding
Ala	GCG	7.27	0.10	Thr	ACG	6.80	0.12	Ser	TCG	4.38	0.06	Pro	CCG	7.02	0.11
Ala	GCA	15.50	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 bias

Gene prediction from GC statistics

- In coding regions : we have A/C/G/T appearing with equal probabilities

- In non-coding region : C/G have probability 0.4 and A/T have probability

```
CGCCGCGGCCAACGGTCCACTGCCTGCGGGCACGTGCGAC

ATGAGCTCATGCCATCCTTGAGGGATTCACACTGCGGTCA

CTGCTGGCCCATTAGAAAATTGTAGGTATTCTCCAATTTC

TGAGGGACCGCGTCTTGCCTGGCCCACTTCCAGGCCCGCT
```

Which part are coding / non-coding

```
CGCCGCGGCCAACGGTCCACTGCCTGCGGGCACGTGCGACATGAGCTCATGC
CATCCTTGAGGGATTCACACTGCGGTCATGAGGGACCGCGTCTTGCCTGGCC
CACTTCCAGGCCCGCT
```

Back to fair bet casino ?

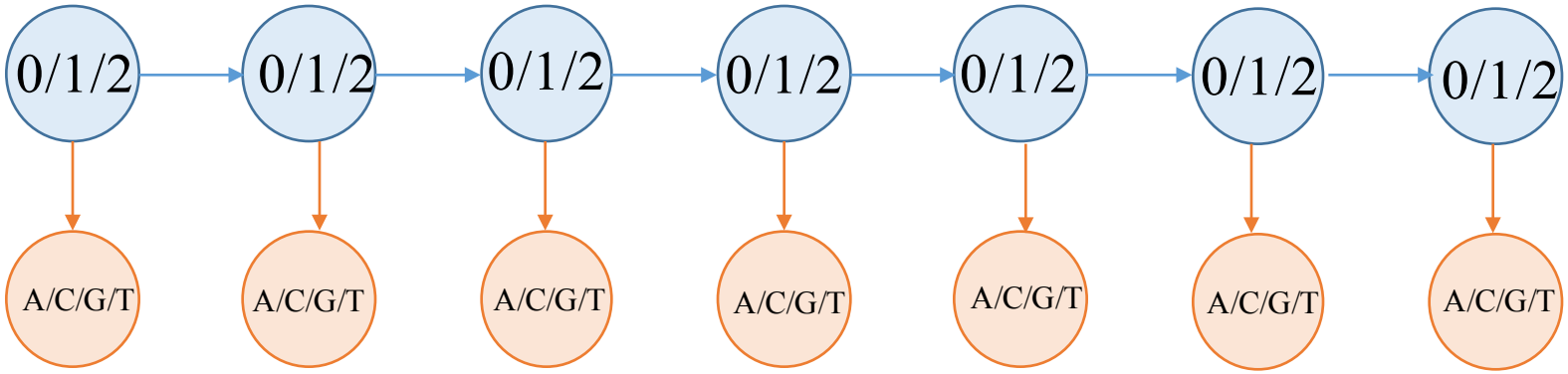
CGCCGCGGCCAACGGTCCACTGCCTGCGGGCACGTGCGACATGAGCTCATGC
CATCCTTGAGGGATTCACACTGCGGTCATGAGGGACCGCGTCTTGCCTGGCC
CACTTCCAGGCCCGCT

ccccccccccccccccccccccccccccccccccccccccnnnnnnnnnnnn
nnnnnnnnnnnnnnnnnnnnnnnnnnnnnnccccccccccccccccccccccc
ccccccccccccccc

Hidden Markov Model

- Our hidden states have three cases {0,1,2} :
  - a_i=*0* if non-coding
  - a_i=*1* coding in forward strand
  - a_i=*2* if coding in reverse strand

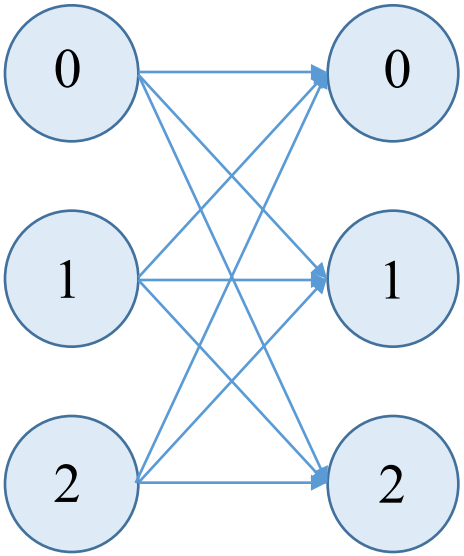
- Observed states are from {A,C,G,T} alphabet

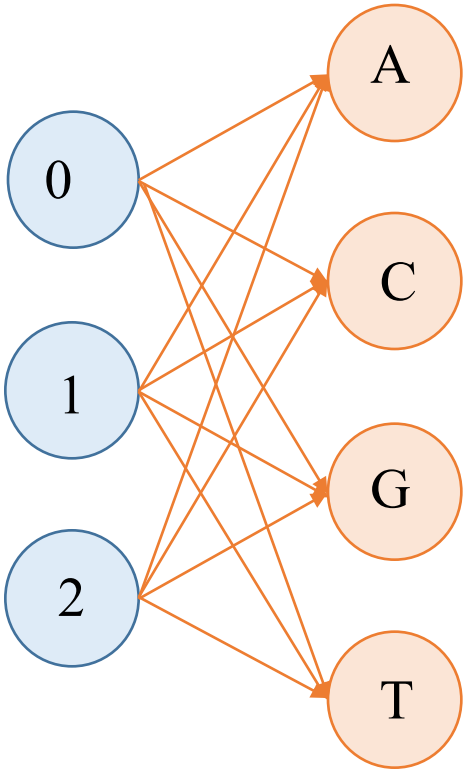


- Given the observations, how can we predict the hidden states ?

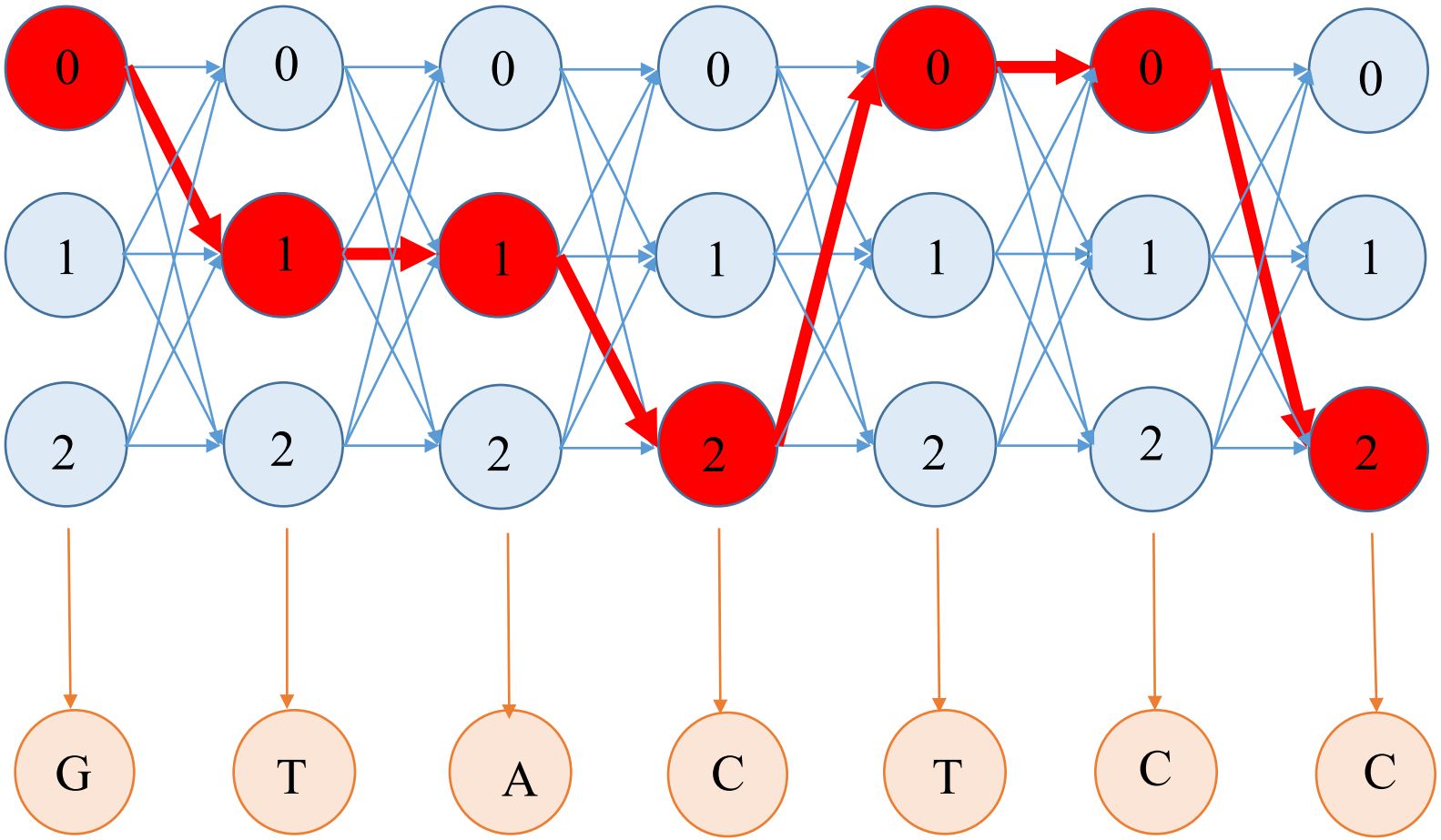
Learning transition and emission probabilities

Transition & Emission probabilities can be learned from training data (e.g. the list of all genes & non-genes in Ecoli)



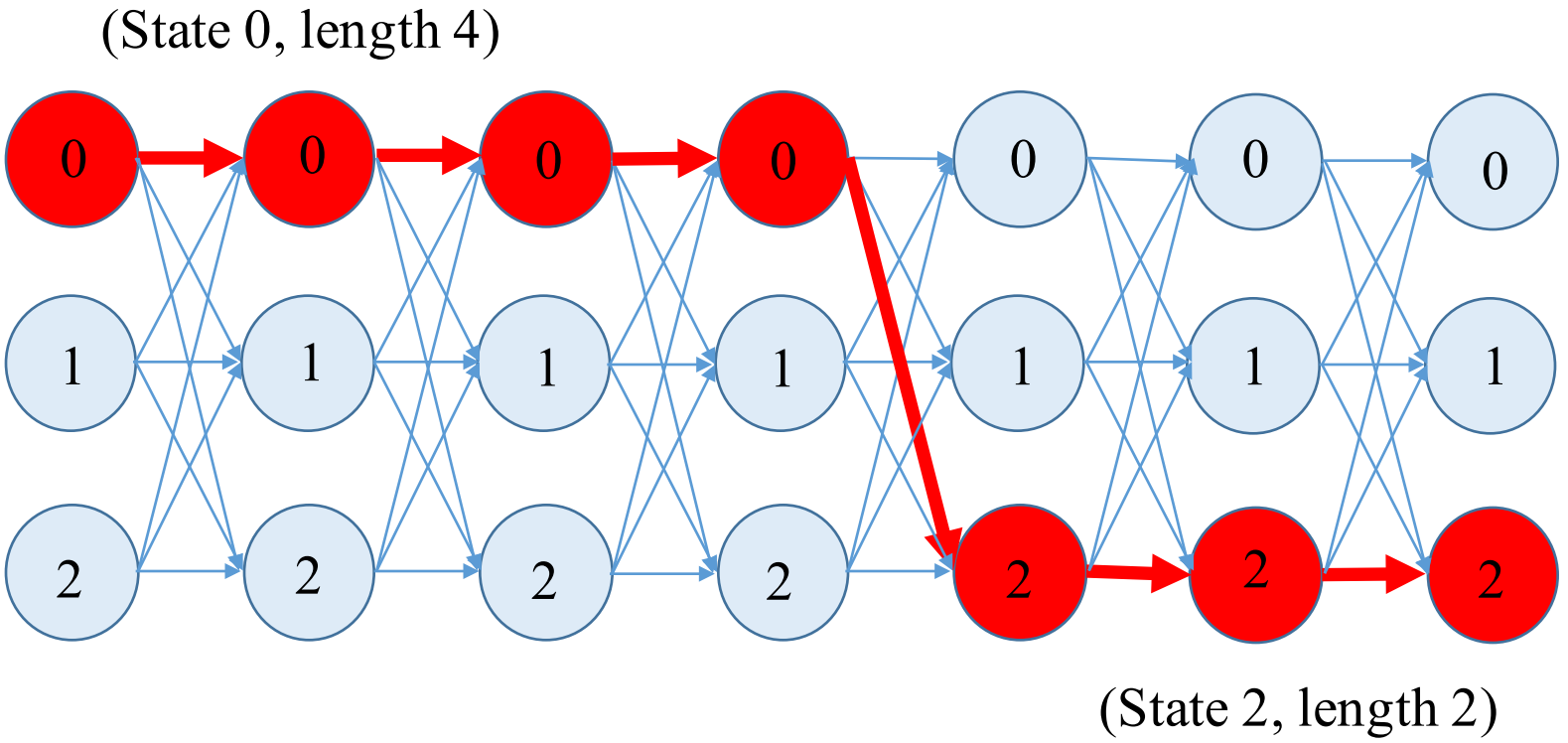


Viterbi decoding

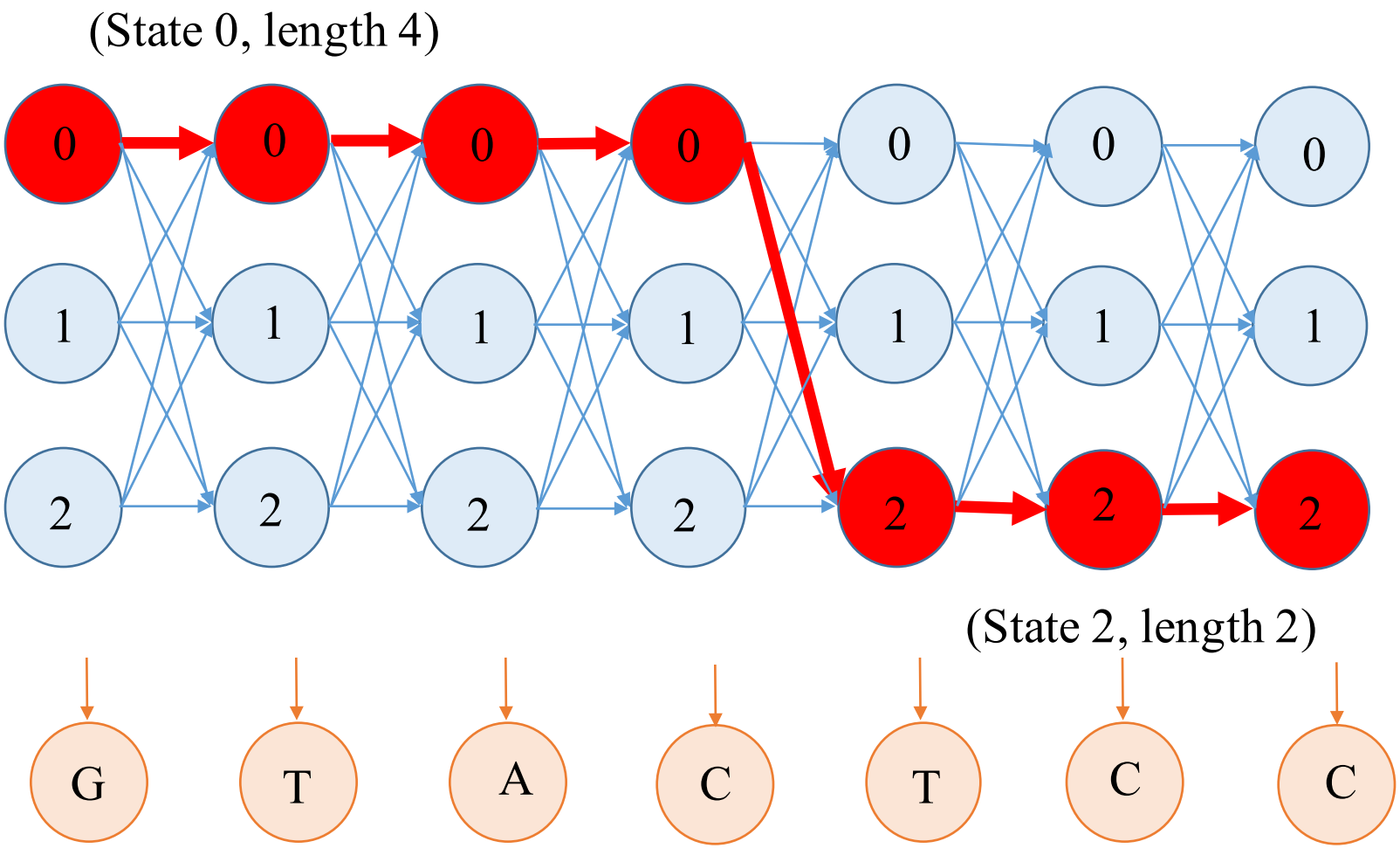


Variable length HMM

- In a lot of occasions, the HMM tends to stay in a single state for a while and then move to a new state
- In this case, we can have a more compact representation of HMM.



Variable length HMM



Variable length HMM

- Consider a gene of length 50bp, then a non-coding region of length 100bp, and a coding region of length 90bp

- Currently, we represent the states like this :

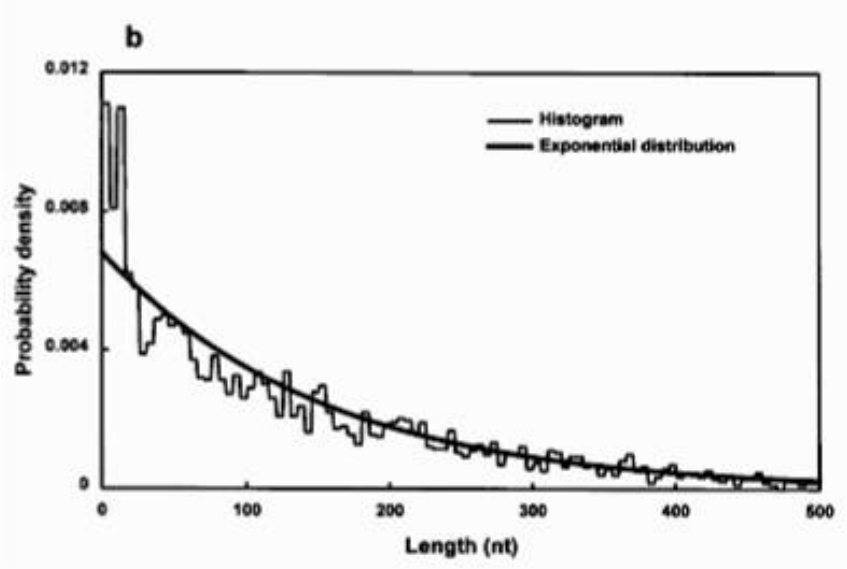
111111111111111111111111111111111111111111111111110000000000000
00000000000000000000000000000000000000000000000000000000000
0000000000000000000000000000111111111111111111111111111111111
111111111111111111111111111111111111111111111111111111111

- A more efficient representation would be :

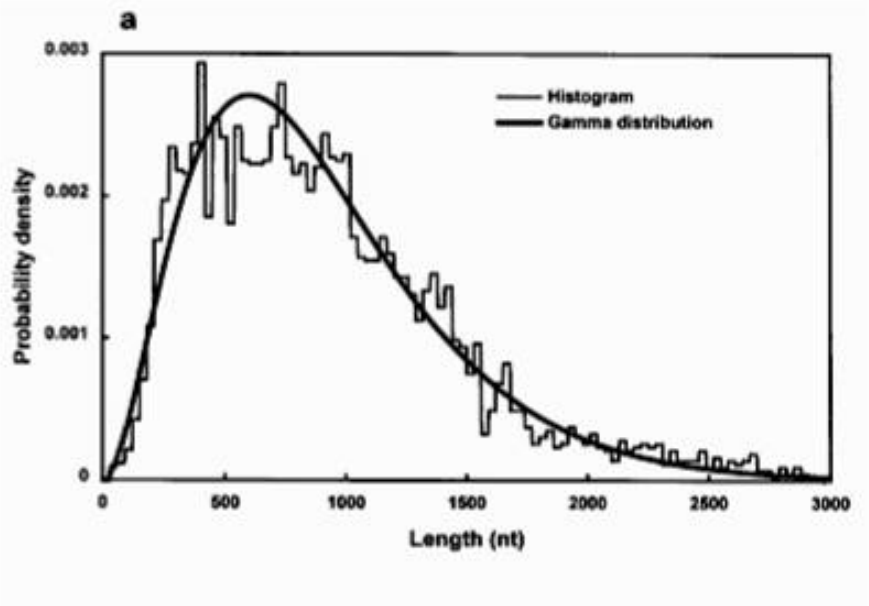
(1,50), (0,100), (1,90)

Probability density of duration

- In addition to transition and emission probabilities, we need to also know how long do we stay in each state.



Probability density of Duration in coding regions



Probability density of Duration in non- coding regions

Typical and Atypical states

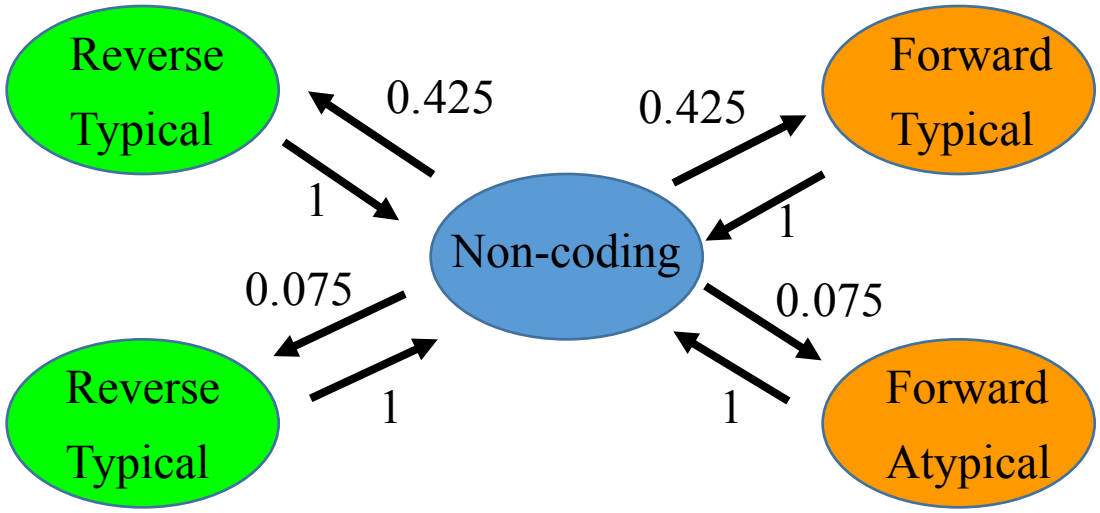
- "Typical" and "Atypical" gene states (one for each of forward and reverse strands)

- Typical/Atypical states emit coding sequence with different codon usage patterns

- In E. coli
  - majority of genes are Typical
  - "horizontally transferred" genes are "Atypical"

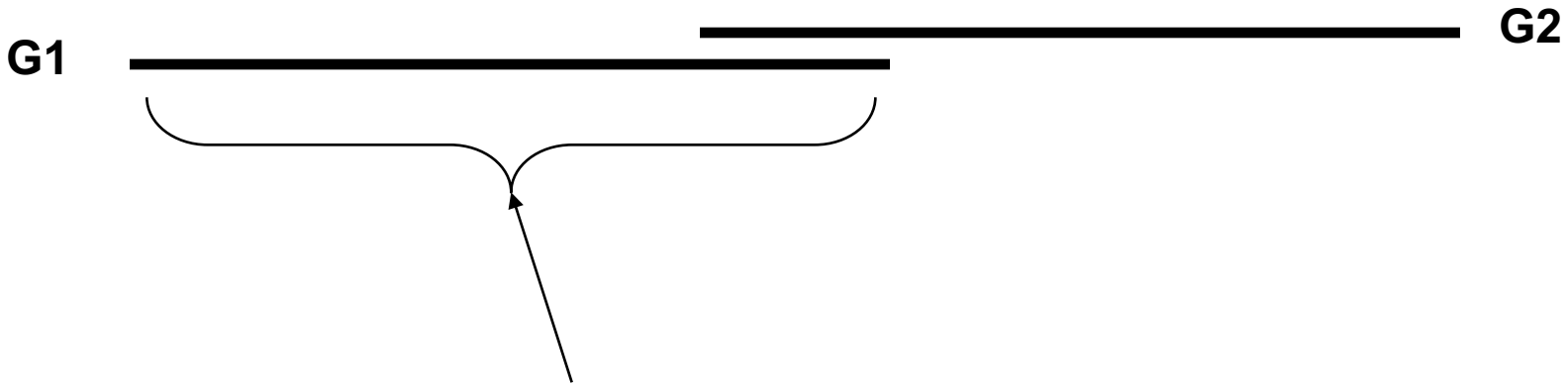
- Movement of genetic material between organisms ther than transmission of DNA from parent to offspring

Transition probabilities



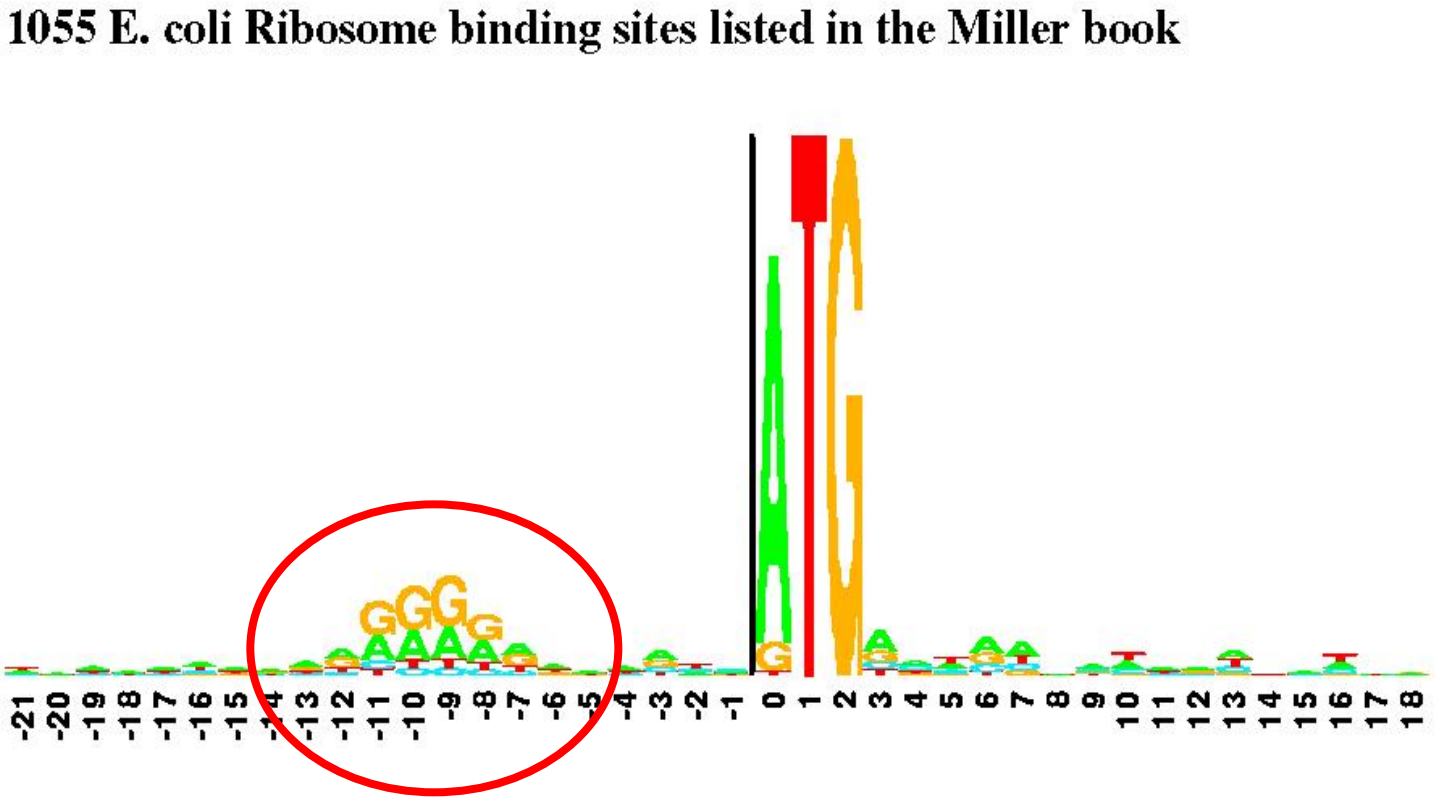
Post-processing

- HMM, assume genes cannot overlap. In reality, genes may overlap.



- Look for an RBS somewhere here.
- Take each start codon here, and find RBS -19 to -4 bp upstream of it

Ribosome binding site (RBS)



Results

The GeneMark.hmm performance

Set #	Number of genes	Prediction method	Exact prediction	Only 3'-end prediction	Missing genes
1	4288	VA	2483 (58%)	1592 (37%)	213 (5%)
1	4288	PP	3233 (75%)	842 (20%)	213 (5%)
2	2821	VA	2017 (71%)	750 (27%)	54 (2%)
2	2821	PP	2268 (80%)	499 (18%)	54 (2%)
3	325	VA	255 (78%)	64 (20%)	6 (2%)
3	325	PP	289 (89%)	30 (9%)	6 (2%)
4	204	VA	156 (76.5%)	47 (23%)	1 (0.5%)
4	204	PP	177 (87.5%)	26 (12%)	1 (0.5%)

VA: Viterbi algorithm
PP: With post-processing

- Data set #1: all annotated E. coli genes
- Data set #2: non-overlapping genes
- Data set #3: Genes with known RBS
- Data set #4: Genes with known start positions

Results

- Gene overlap is an important factor
- Performance goes up from 58% to 71% when overlapping genes are excluded from data set
- Post-processing helps a lot
  - 58% --> 75% for data set #1
- “False negatives” < 5%
- “Wrong” gene predictions: “False positives” ~8%
  - Are they really false positives, or are they unannotated genes?

Results

- Compared with other programs

Number of genes	Prediction method	Exact prediction	Only 3'-end prediction	Missing genes
148	GeneMark.hmm	105 (71%)	28 (19%)	15 (10%)
148	GeneMark	92 (62%)	37 (25%)	19 (13%)
148	ECOPARSE	79 (53%)	33 (23%)	36 (24%)

Results

- Robustness to parameter settings
- Alternative set of transition probability values used
- Little change in performance (~20% change in parameter values leads to < 5% change in performance)