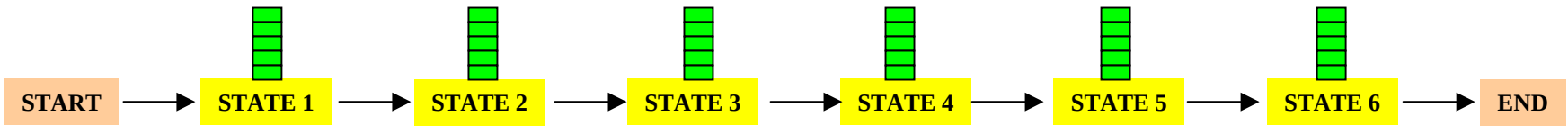
Profile HMMs

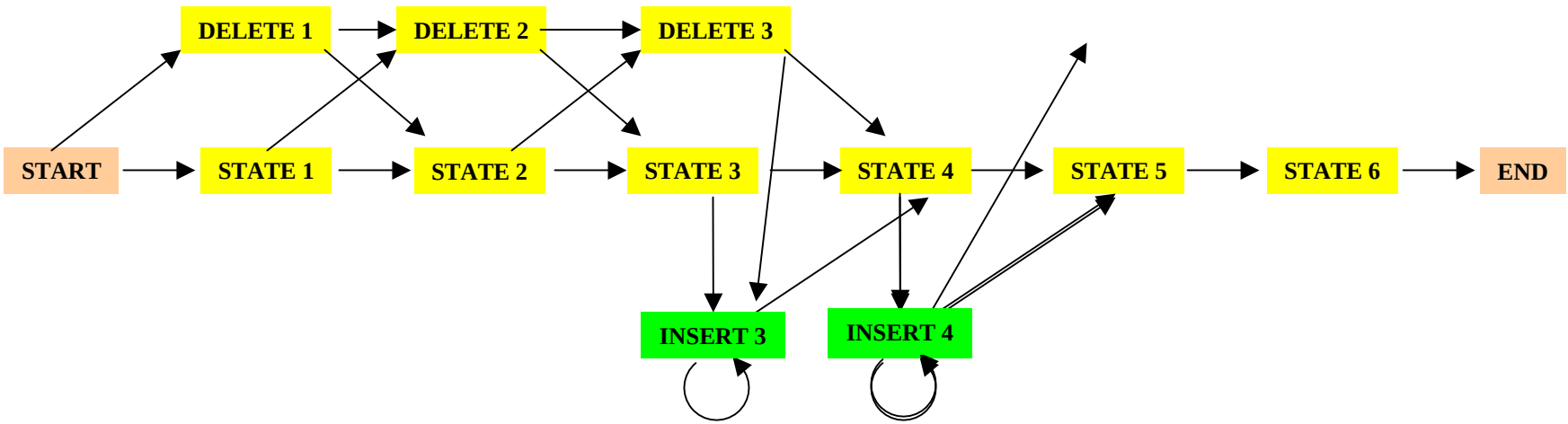
PROFILE METHOD, [M. Gribskov et al., '90]

Location in Seq.	Sequence 1	2	3	4	5	6		Protein Name
14	G	V	S	A	S	A		Ka RbtR
32	G	V	S	E	M	T		Ec DeoR
33	G	V	S	P	G	T		Ec RpoD
76	G	A	G	I	A	T		Ec TrpR
178	G	C	S	R	E	T		Ec CAP
205	C	L	S	P	S	R		Ec AraC
210	C	L	S	P	S	R		St AraC
13	G	V	N	K	E	T		Br MerR

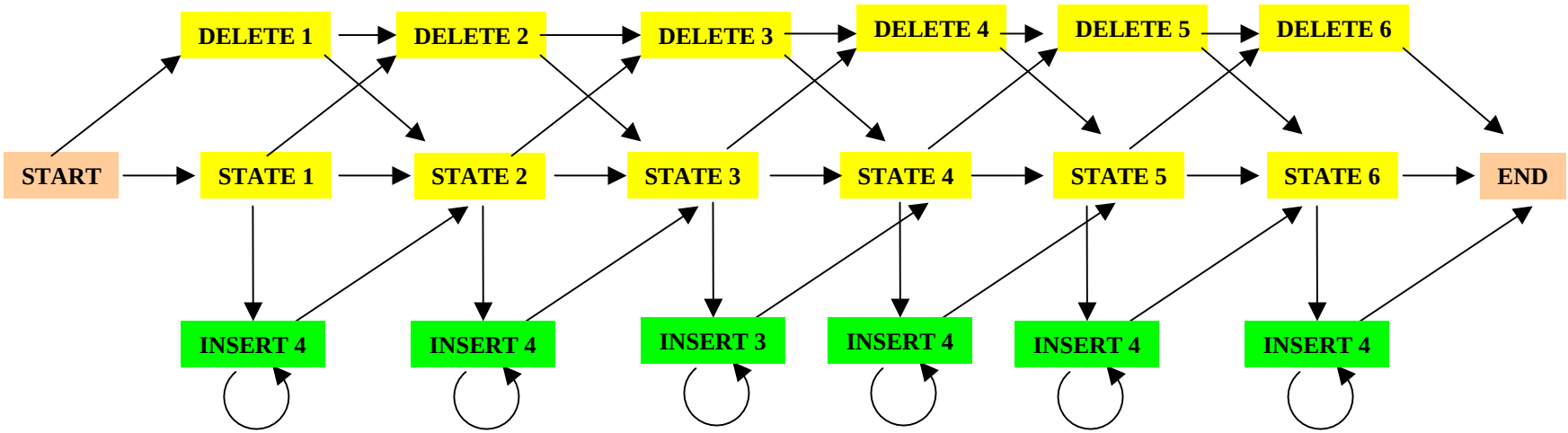


Profile HMMs with InDels

- Insertions
- Deletions
- Insertions & Deletions

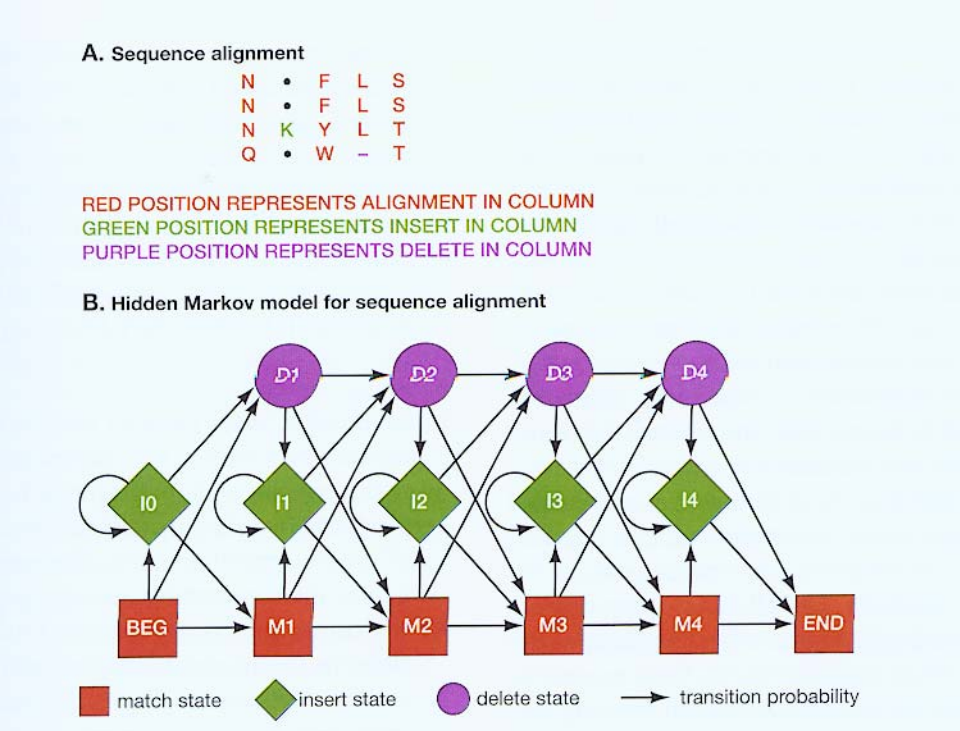


Profile HMMs with InDels



Missing transitions from DELETE j to INSERT j and from INSERT j to DELETE j+1.

Profile HMMs for MSA



How to Solve Problem 2?

- Solve the following problem:

Input: Hidden Markov Model M,
parameters Θ, emitted sequence S

Output: Most Probable Path Π

How: Viterbi's Algorithm (Dynamic Programming)

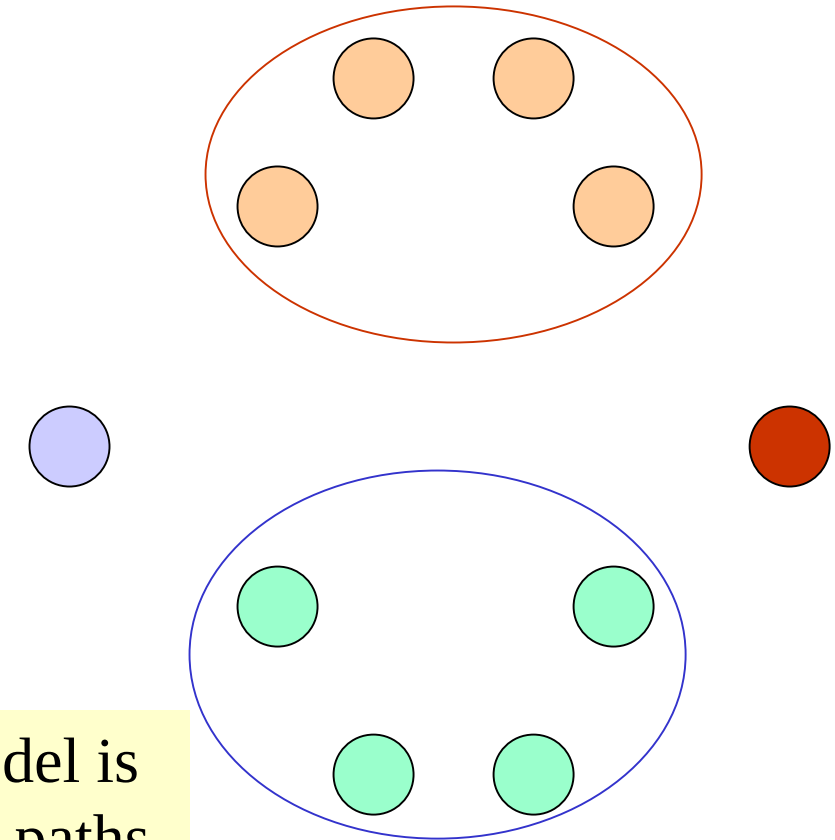
Define $\Pi[i,j]$ = MPP for first j characters of S ending in state i

Define $P[i,j]$ = Probability of $\Pi[i,j]$

- Compute state i with largest $P[i,j]$.

Hidden Markov Model (HMM)

- States
- Transitions
- Transition Probabilities
- Emissions
- Emission Probabilities



- What is hidden about HMMs?

Answer: The path through the model is hidden since there are many valid paths.

Problem 5: LEARNING QUESTION

- Input: model structure M, Training Sequence S
- Output: Compute the parameters Θ
- Criteria: ML criterion
    - maximize $P(S \mid M, \Theta)$ HOW???

Problem 6: DESIGN QUESTION

- Input: Training Sequence S
- Output: Choose model structure M, and compute the parameters Θ
    - No reasonable solution
    - Standard models to pick from

Iterative Solution to the LEARNING QUESTION (Problem 5)

- Pick initial values for parameters Θ_0
- Repeat

  Run training set *S* on model M

  Count # of times transition i $\Rightarrow$ j is made

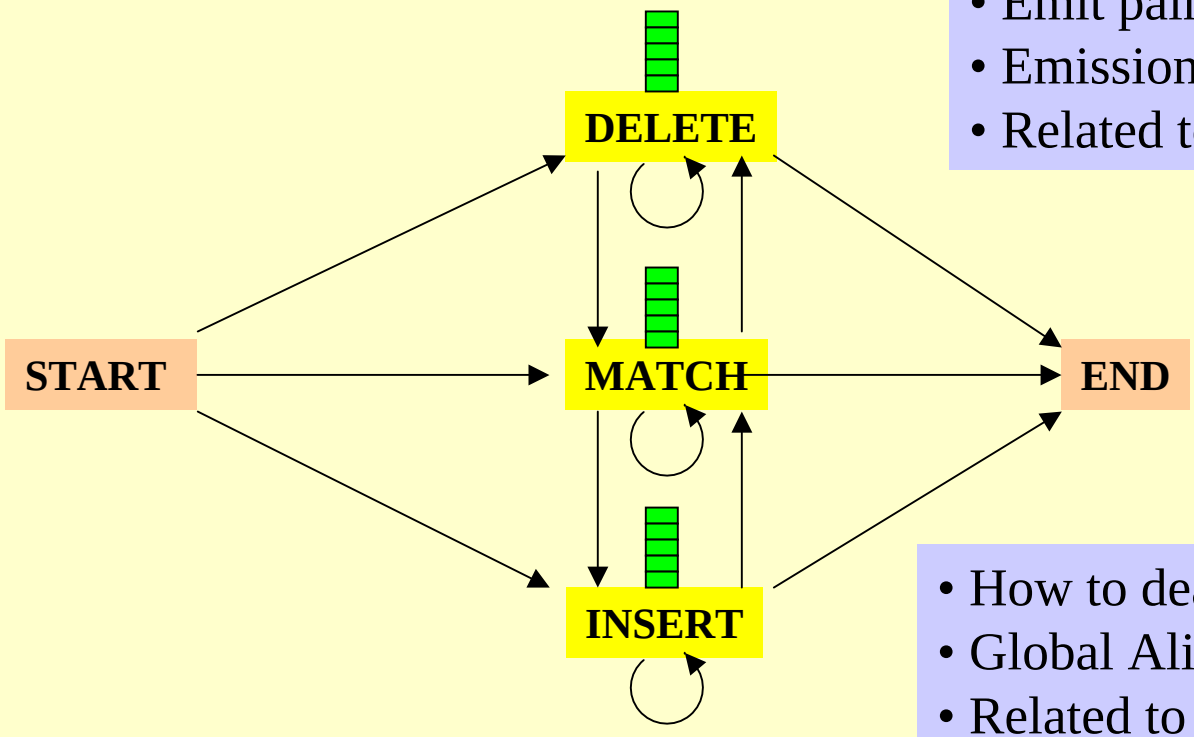
  Count # of times letter x is emitted from state i

  Update parameters Θ
- Until (some stopping condition)

How to model Pairwise Sequence Alignment

LEAPVE

LAPVIE

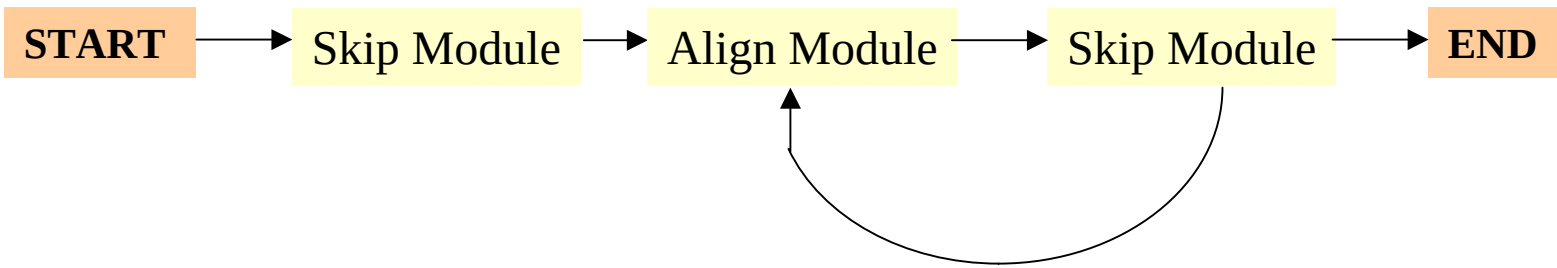


Pair HMMs
- Emit pairs of synbols
- Emission probs?
- Related to Sub. Matrices

- How to deal with InDels?
- Global Alignment? Local?
- Related to Sub. Matrices

How to model Pairwise Local Alignments?

How to model Pairwise Local Alignments with gaps?



Entropy

- Entropy measures the variability observed in given data.

$$E = -\sum_{c} p_c \log p_c$$

- Entropy is useful in multiple alignments & profiles.
- Entropy is max when uncertainty is max.

G-Protein Couple Receptors

- Transmembrane proteins with 7 α-helices and 6 loops; many subfamilies
- Highly variable: 200-1200 aa in length, some have only 20% identity.
- [Baldi & Chauvin, '94] HMM for GPCRs
- HMM constructed with 430 match states (avg length of sequences) ; Training: with 142 sequences, 12 iterations

GPCR - Analysis

- Compute main state entropy values

$$H_i = -\sum_a e_{ia} \log e_{ia}$$

- For every sequence from test set (142) & random set (1600) & all SWISS-PROT proteins
  - Compute the negative log of probability of the most probable path π

$$Score(S) = -\log\left(P(\pi \mid S, M)\right)$$

Entropy

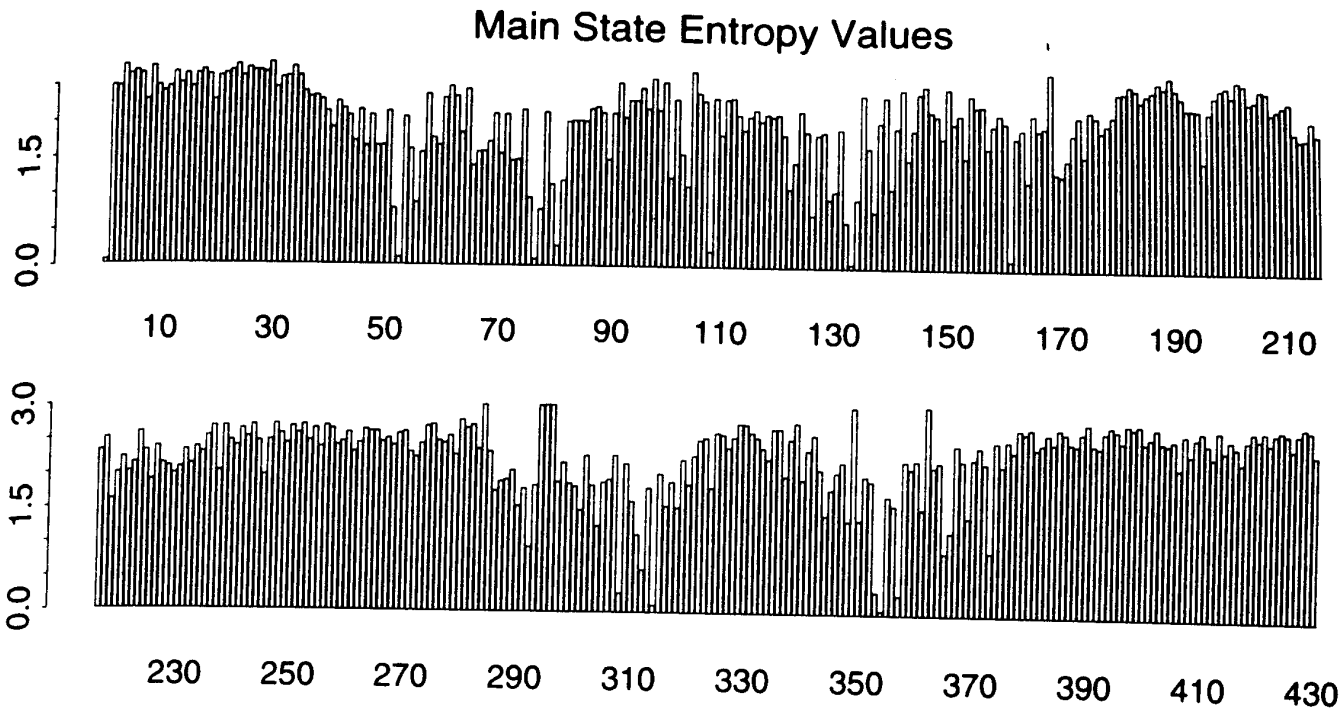
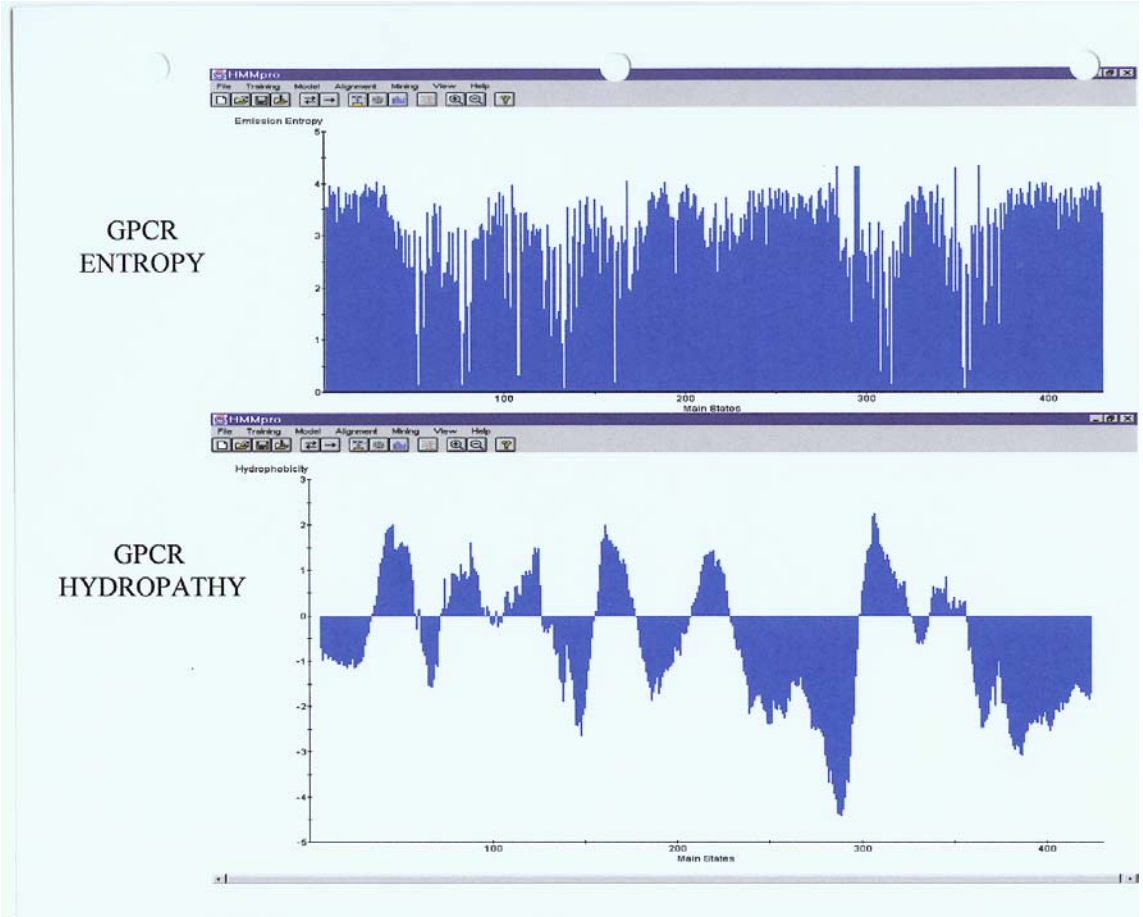


Figure 8.1: Entropy Profile of the Emission Probability Distributions Associated with the Main States of the HMM After 12 Cycles of Training.

GPCR Analysis



GPCR Analysis (Cont'd)

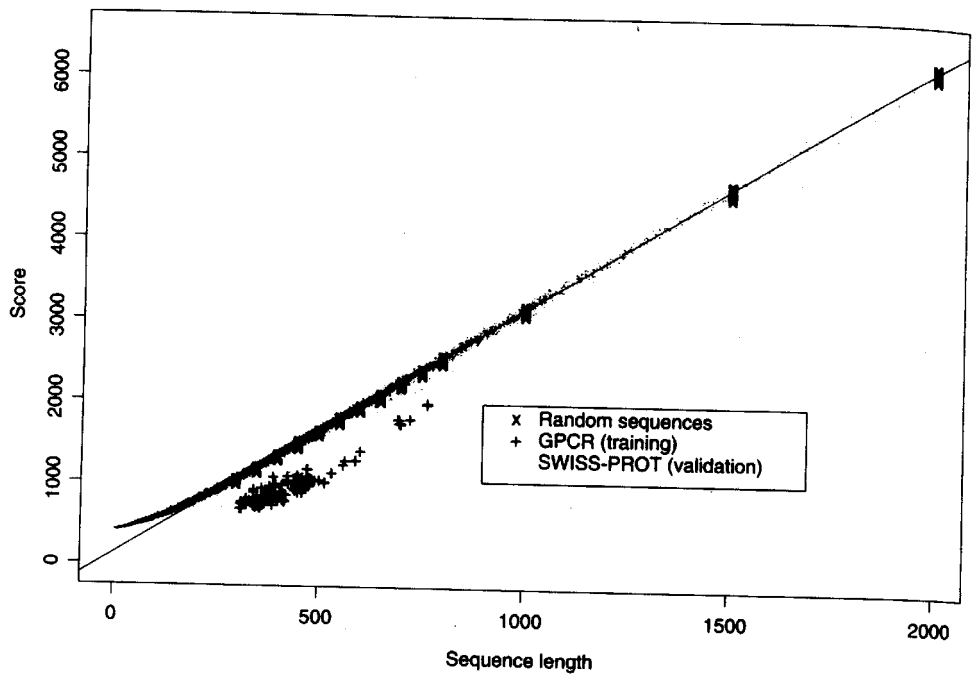


Figure 8.2: Scores (Negative Log-likelihoods of Optimal Viterbi Paths). Represented sequences consist of 142 GPCR training sequences, all sequences from the SWISS-PROT database of length less than or equal to 2000, and 220 randomly generated sequences with same average composition as the GPCRs of length 300, 350, 400, 450, 500, 550, 600, 650, 700, 750, 800 (20 at each length). The regression line was obtained from the 220 random sequences. The horizontal distances in the histogram correspond to normalized scores (6).

Applications of HMM for GPCR

- Bacteriorhodopsin
  - Transmembrane protein with 7 domains
  - But it is not a GPCR
  - Compute score and discover that it is close to the regression line. Hence not a GPCR.
- Thyrotropin receptor precursors
  - All have long initial loop on INSERT STATE 20.
  - Also clustering possible based on distance to regression line.

HMMs – Advantages

- Sound statistical foundations
- Efficient learning algorithms
- Consistent treatment for insert/delete penalties for alignments in the form of locally learnable probabilities
- Capable of handling inputs of variable length
- Can be built in a modular & hierarchical fashion; can be combined into libraries.
- Wide variety of applications: Multiple Alignment, Data mining & classification, Structural Analysis, Pattern discovery, Gene prediction.

HMMs – Disadvantages

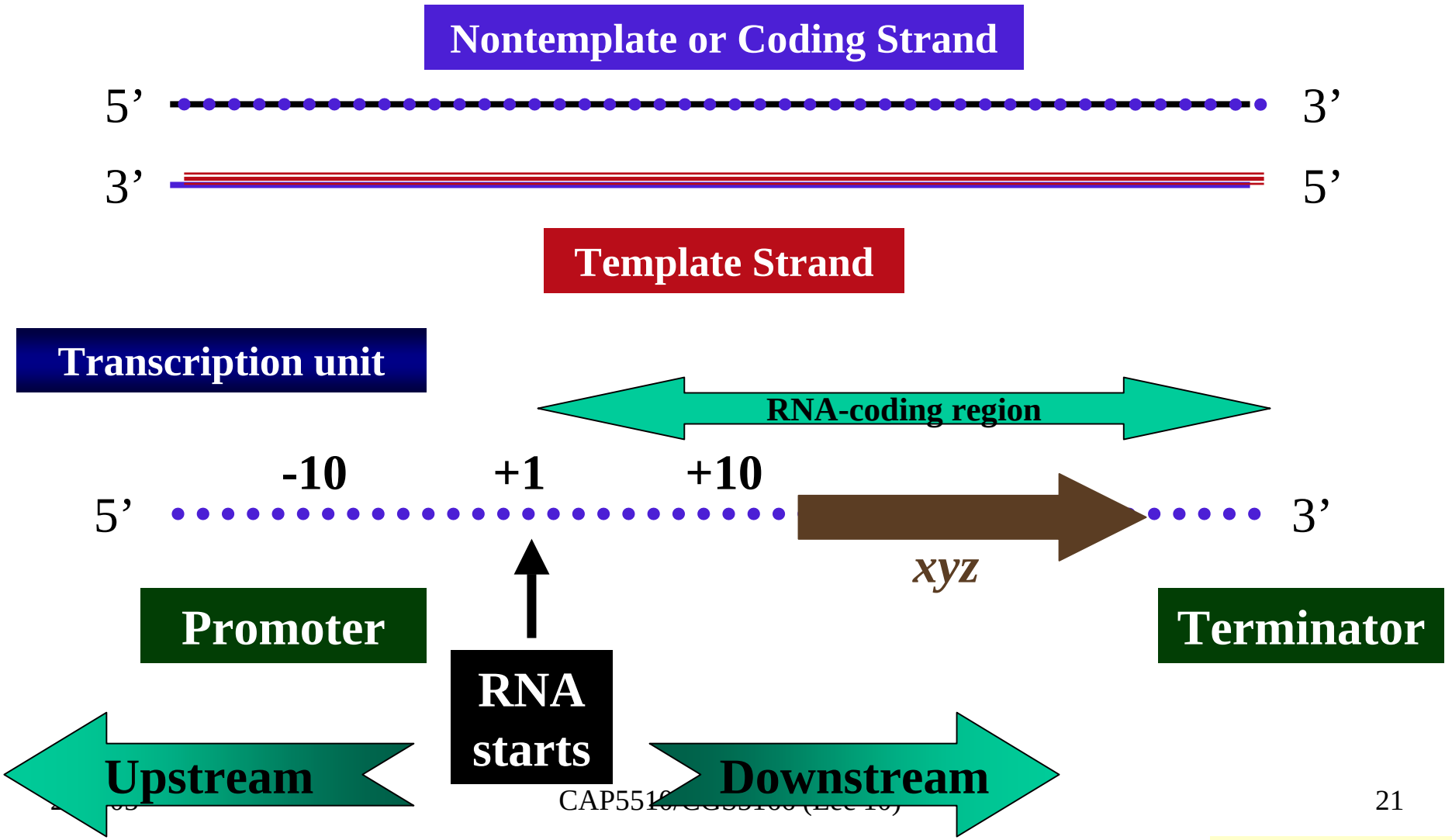
- Large # of parameters.
- Cannot express dependencies & correlations between hidden states.

Prokaryotic Gene Prediction

- Genes: region between *start codon* ATG and *stop codon* (TAA, TAG, or TGA). Absence of introns.
- Codon Bias
- Locate Promoter region
- Ribosome Binding site
- Terminator site

Nomenclature

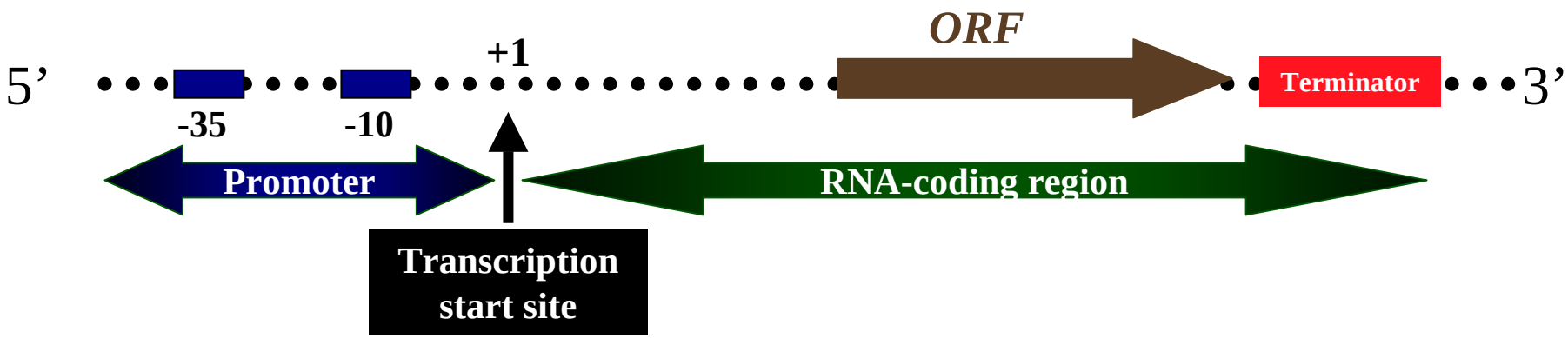
RNA Polymerization occurs 5’ to 3’



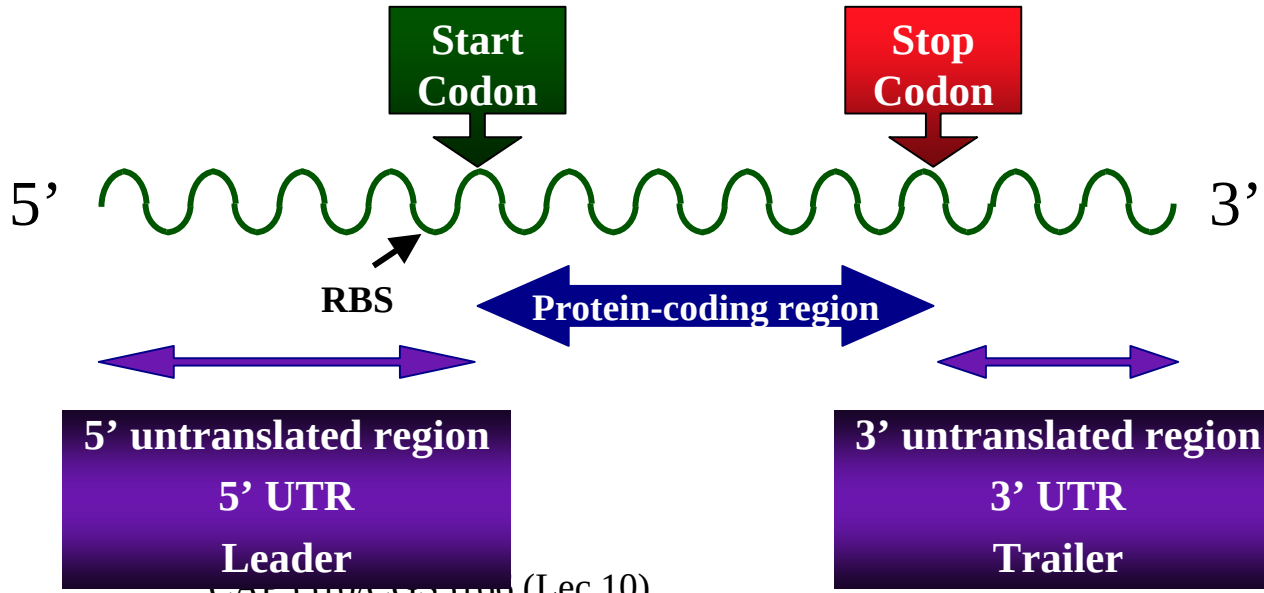
Slide courtesy Prof. Mathee

Transcriptional unit and single gene mature mRNA

Transcriptional unit



mRNA



RBS

Ribosome binding site

Slide courtesy Prof. Mathee

Prokaryotic Gene Characteristics

DNA PATTERNS IN THE *E. coli lexA* GENE

	GENE SEQUENCE	PATTERN
1	GAATTCGATAAATCTCTGGTTTATTGTGCAGTTTATGGTT	CTGNNNNNNNNNNCAG
	TT	TTGACA
41	CCAAAATCGCCTTTTGCTGTATATACTCACAGCATAACTG	CTGNNNNNNNNNNCAG
	CCAA -35 -10 TATACT >	TATAAT, > mRNA start
81	TATATACACCCAGGGGGCGGAATGAAAGCGTTAACGGCCA	CTGNNNNNNNNNNCAG
	+10 GGGGG Ribosomal binding site	GGAGG
121	GGCAACAAGAGGTGTTTGATCTCATCCGTGATCACATCAG	
161	CCAGACAGGTATGCCGCCGACGCGTGCGGAAATCGCGCAG	ATG
201	CGTTTGGGGTTCCGTTCCCCAAACGCGGCTGAAGAACATC	
241	TGAAGGCGCTGGCACGCAAAGGCGTTATTGAAATTGTTTC	
281	CGGCGCATCACGCGGGATTCGTCTGTTGCAGGAAGAGGAA	
321	GAAGGGTTGCCGCTGGTAGGTCGTGTGGCTGCCGGTGAAC	
361	CACTTCTGGCGCAACAGCATATTGAAGGTCATTATCAGGT	OPEN READING FRAME
401	CGATCCTTCCTTATTCAAGCCGAATGCTGATTTCCTGCTG	
441	CGCGTCAGCGGGATGTCGATGAAAGATATCGGCATTATGG	
481	ATGGTGACTTGCTGGCAGTGCATAAAACTCAGGATGTACG	
521	TAACGGTCAGGTCGTTGTCGCACGTATTGATGACGAAGTT	
561	[illegible]	
601	TGTTGCCAGAAAATAGCGAGTTTAAACCAATTGTCGTTGA	
641	CCTTCGTCAGCAGAGCTTCACCATTGAAGGGCTGGCGGTT	
681	GGGGTTATTCGCAACGGCGACTGGCTGTAACATATCTCTG	TAA
721	AGACCGCGATGCCGCCTGGCGTCGCGGTTTGTTTTTCATC	
761	TCTCTTCATCAGGCTTGTCTGCATGGCATTCCTCACTTCA	
801	TCTGATAAAGCACTCTGGCATCTCGCCTTACCCATGATTT	
841	TCTCCAATATCACCGTTCCGTTGCTGGGACTGGTCGATAC	
881	GGCGGTAATTGGTCATCTTGATAGCCCGGTTTATTTGGGC	
921	GGCGTGGCGGTTGGCGCAACGGCGGACCAGCT	

Shown are matches to approximate consensus binding sites for LexA repressor (CTGNNNNNNNNNNCAG), the -10 amd -35 promoter regions relative to the start of the mRNA (TTGACA and TATAAT), the ribosomal binding site on the mRNA (GGAGG), and the open reading frame (ATG...TAA). Only the second two of the predicted LexA binding sites actually bind the repressor.

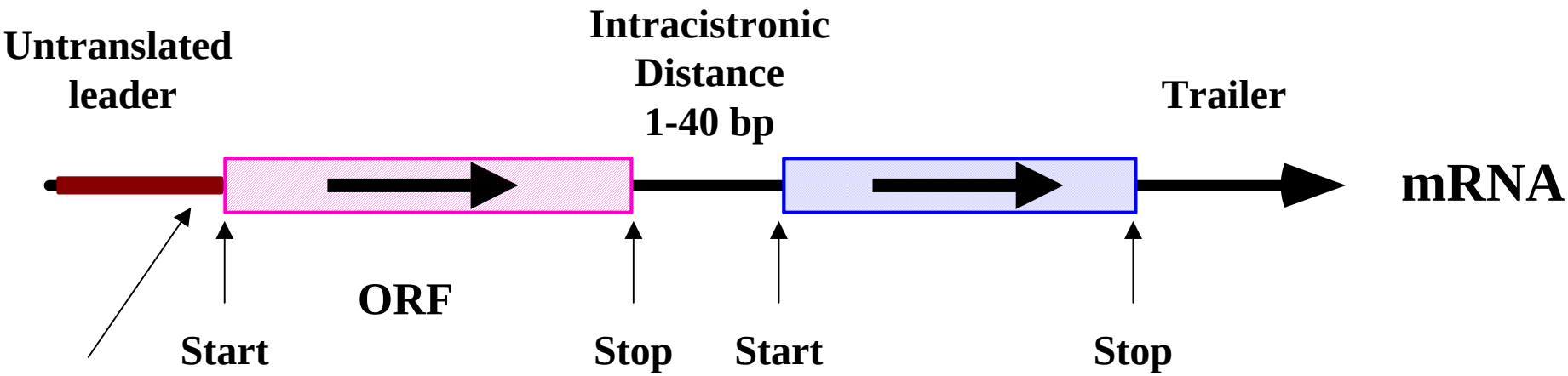
FIGURE 9.6. The promoter and open reading frame of the *E. coli lexA* gene.

Messenger RNA or mRNA

Initation Codon	AUG	Methionine

Termination Codons	
UAA	Ochre
UAG	Amber
UGA	Opal

Others:		
	GUG	Valine
	UUG	Leucine
	AUU	Isoleucine



RBS
Ribosome Binding Site
Shine-Dalgarno Sequence

7 bp upstream of start codon
5'--AGGAGG--3'

Coding region
Open Reading Frame (ORF)

Reading frame is one of three possible ways of reading a nucleotide sequence as a series of triplets.

Slide courtesy Prof. Mathee

Start and Stop Codon Distribution

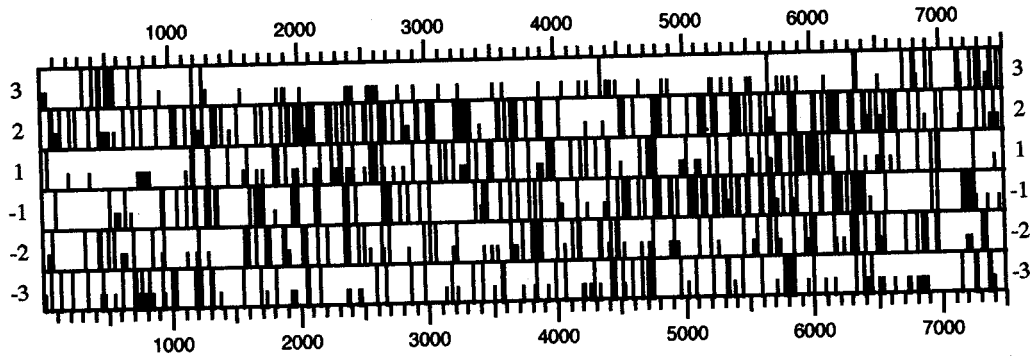
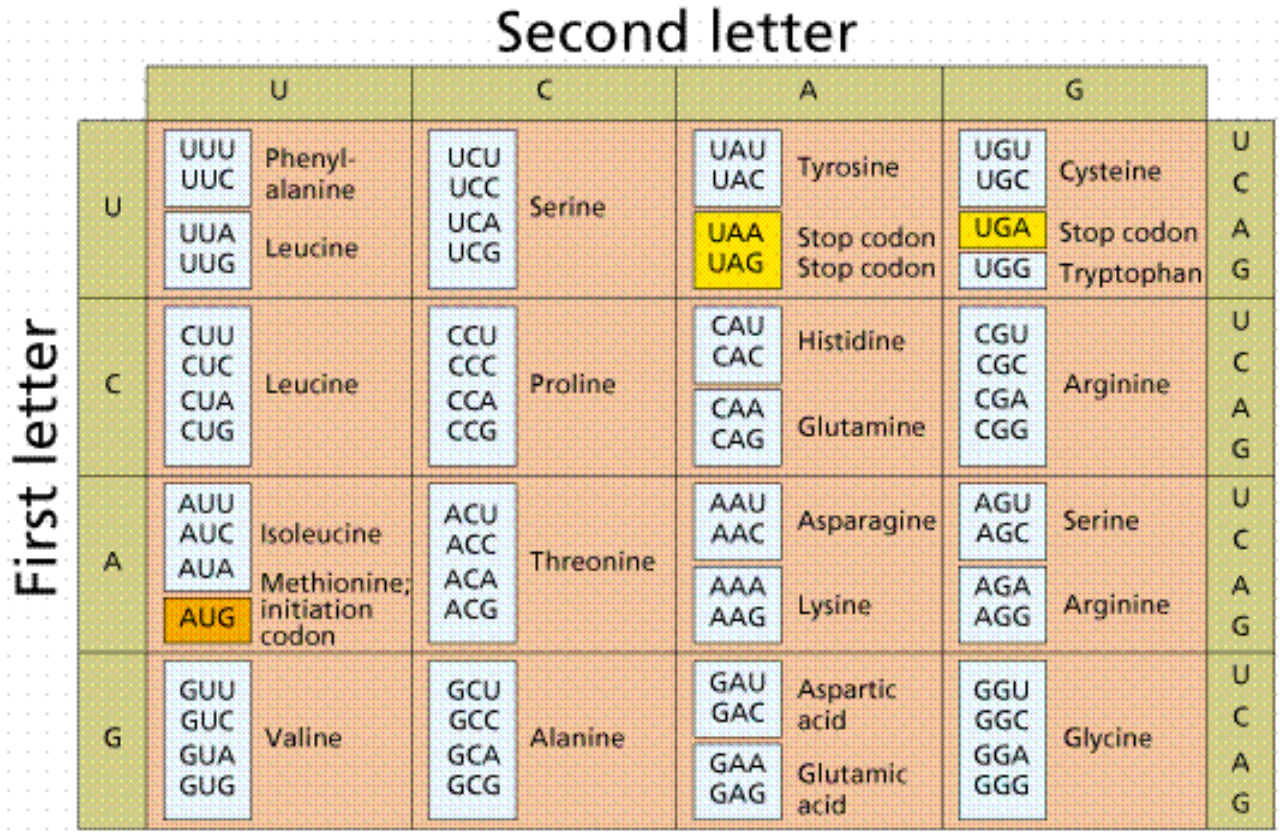


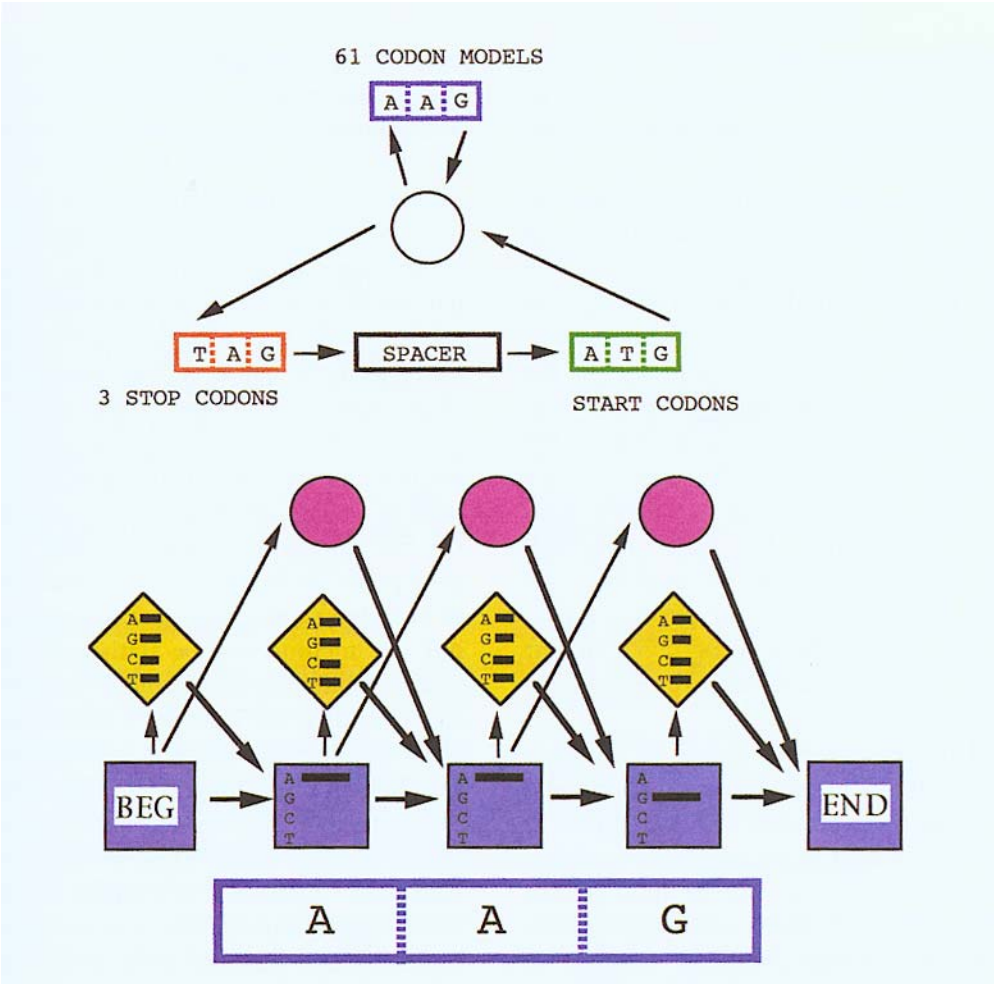
FIGURE 9.1. ORF map of a portion of the *E. coli lac* operon using the DNA STRIDER program (Marck 1988). Shown are AUG and termination codons as one-half and full vertical bars, respectively, in all six possible reading frames. The *lacZ* gene is visible as an ORF that runs from positions 1284 to 4355 in frame 3.

Genetic Code



Slide courtesy http://www.emc.maricopa.edu/faculty/farabee/BIOBK/code.gif

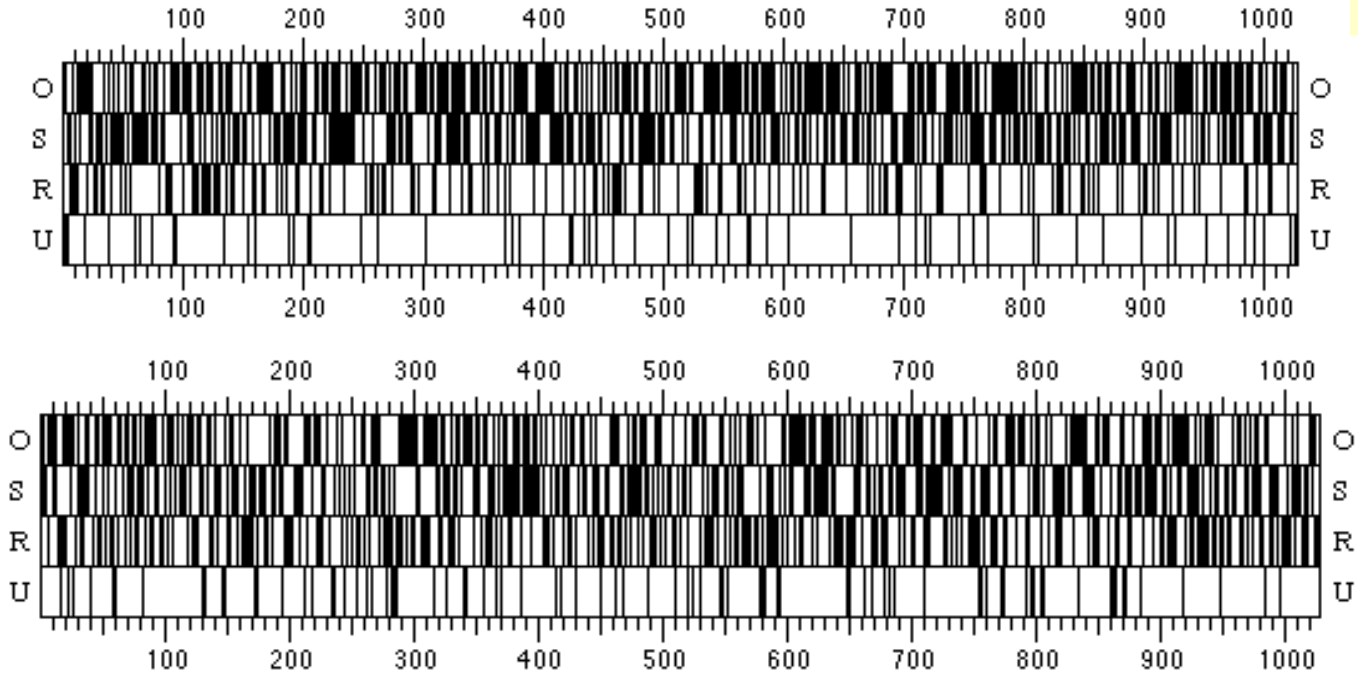
Recognizing Codons



Codon Bias

- Some codons preferred over others.

O = optimal
S = suboptimal
R = rare
U = unfavorable



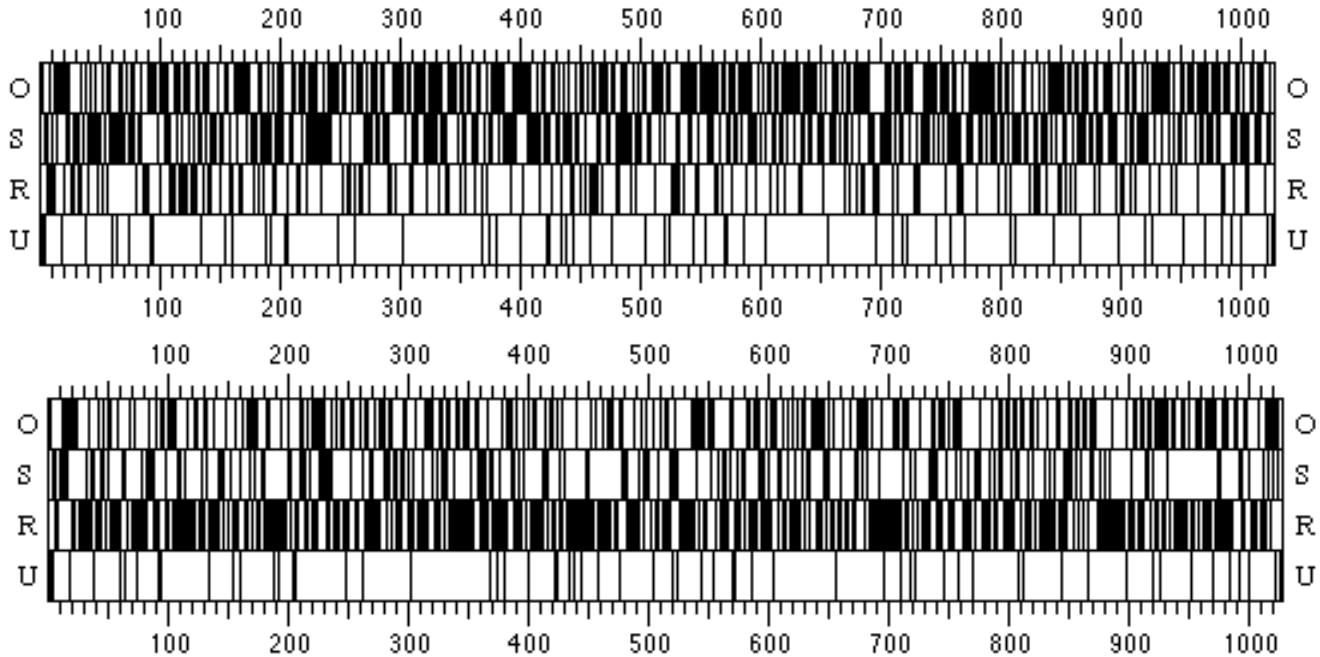
Frame Shift 1

Frame Shift 2

Codon Bias

- Codon biases specific to organisms

O = optimal
S = suboptimal
R = rare
U = unfavorable



Same Frames; Different labeling of codon types (i.e., from yeast)

Eukaryotic Gene Prediction

- Complicated by introns & alternative splicing
- Exons/introns have different GC content.
- Many other measures distinguish exons/introns
- Software:
  - GENEPARSER Snyder & Stormo (NN)
  - GENIE Kulp, Haussler, Reese, Eckman (HMM)
  - GENSCAN Burge, Karlin (Decision Trees)
  - XGRAIL Xu, Einstein, Mural, Shah, Uberbacher (NN)
  - PROCRUSTES Gelfand (Formal Languages)
  - MZEF Zhang

Introns/Exons in *C. elegans*

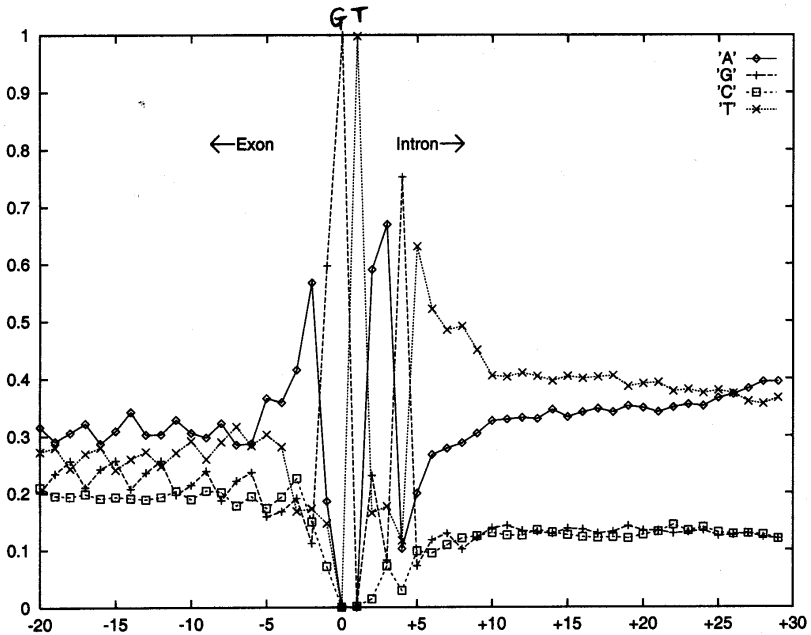


Figure 2: Profile of the same 5' collection but around a larger window.

A/T

G/C

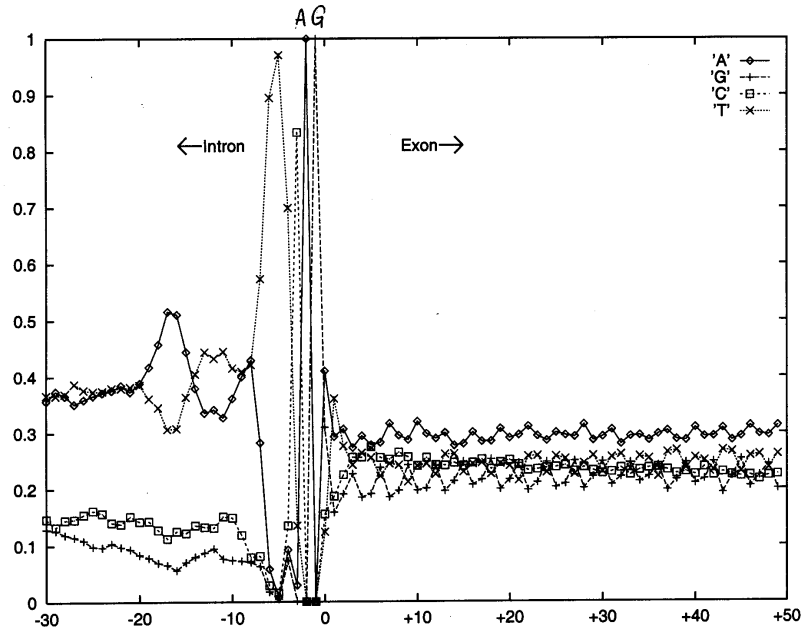
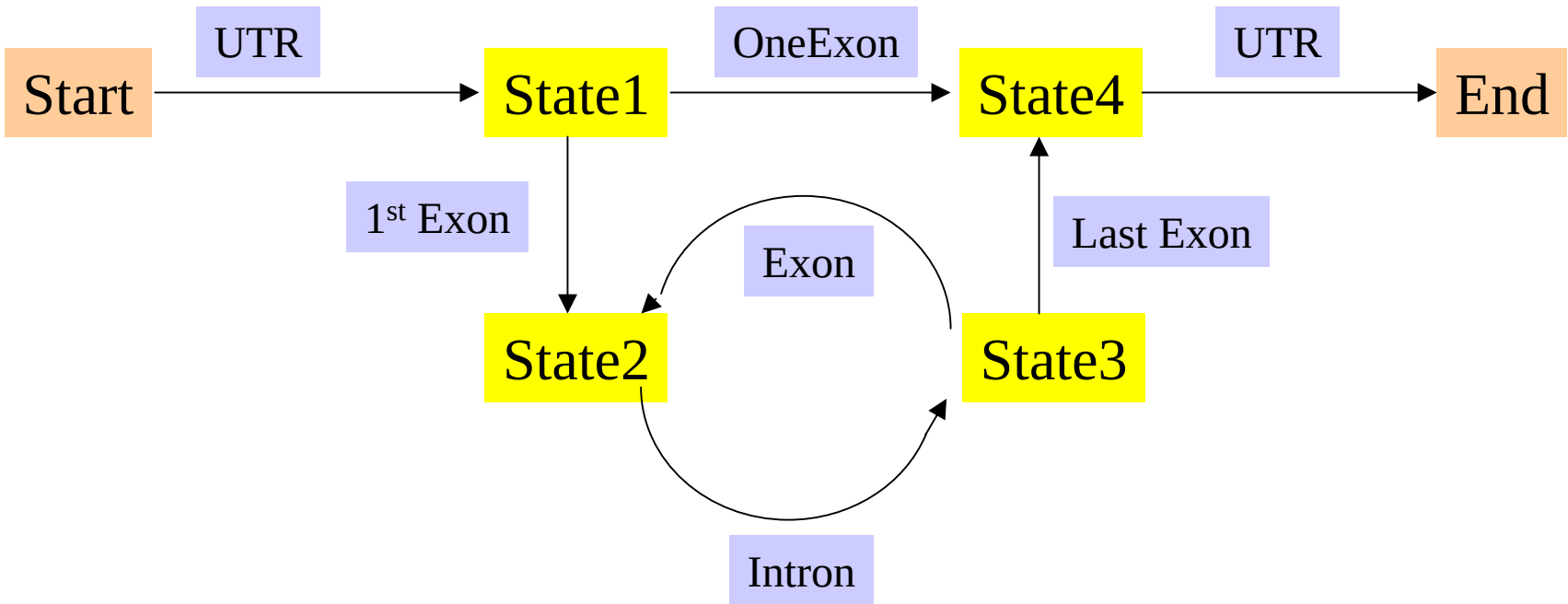


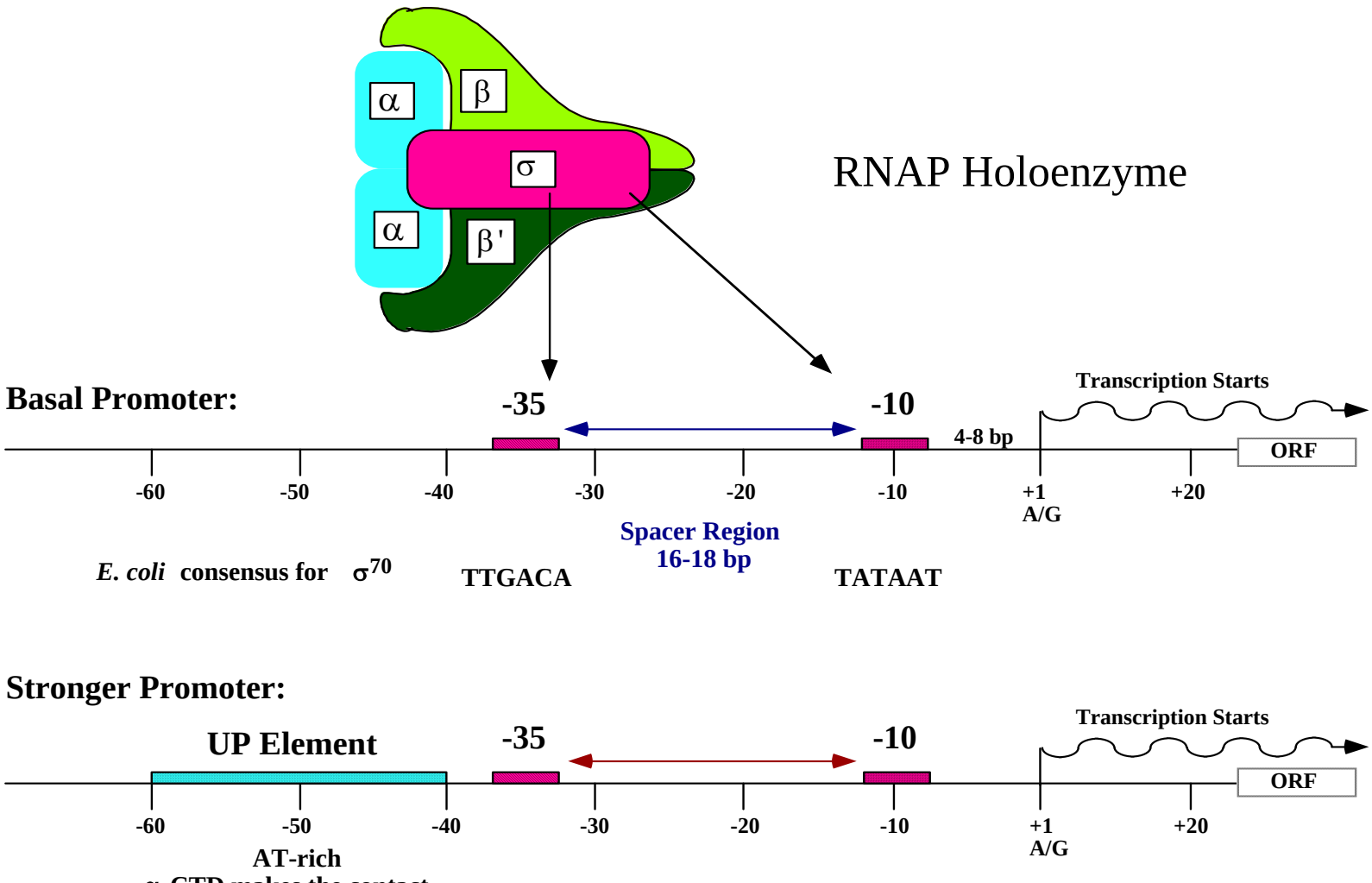
Figure 4: Profile of 8,192 sequences of length 80 around the 3' site. The first position in the exon is labeled 0.

- 8192 Introns in *C. elegans* : [**GT**...**AG**]
- Vary in lengths from 30 to over 600; Complexity varies

HMM structure for Gene Finding



Transcriptional machinery: RNA Polymerase and DNA



Slide courtesy Prof. Mathee