Profile HMMs

**Profile Method**, [M. Gribskov et al., ’90]

<table>
<thead>
<tr>
<th>Location in Seq</th>
<th>Sequence</th>
<th>Protein Name</th>
</tr>
</thead>
<tbody>
<tr>
<td>14</td>
<td>G V S A S A</td>
<td>Ka RbtR</td>
</tr>
<tr>
<td>32</td>
<td>G V S E M T</td>
<td>Ec DeoR</td>
</tr>
<tr>
<td>33</td>
<td>G V S P G T</td>
<td>Ec RpoD</td>
</tr>
<tr>
<td>76</td>
<td>G A G I A T</td>
<td>Ec TrpR</td>
</tr>
<tr>
<td>178</td>
<td>G C S R E T</td>
<td>Ec CAP</td>
</tr>
<tr>
<td>205</td>
<td>C L S P S R</td>
<td>Ec AraC</td>
</tr>
<tr>
<td>210</td>
<td>C L S P S R</td>
<td>St AraC</td>
</tr>
<tr>
<td>13</td>
<td>G V N K E T</td>
<td>Br MerR</td>
</tr>
</tbody>
</table>
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

A. Sequence alignment

```
N • F L S
N • F L S
N K Y L T
Q • W - T
```

RED POSITION REPRESENTS ALIGNMENT IN COLUMN
GREEN POSITION REPRESENTS INSERT IN COLUMN
PURPLE POSITION REPRESENTS DELETE IN COLUMN

B. Hidden Markov model for sequence alignment

[Diagram showing a Hidden Markov Model with states and transitions]
How to Solve Problem 2?

• Solve the following problem:

**Input**: Hidden Markov Model $M$, parameters $\Theta$, emitted sequence $S$

**Output**: Most Probable Path $\Pi$

**How**: Viterbi’s Algorithm (Dynamic Programming)

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

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

- 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 $\Theta$
- **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 $\Theta$
  - No reasonable solution
  - Standard models to pick from
Iterative Solution to the LEARNING QUESTION (Problem 5)

- Pick initial values for parameters $\Theta_0$
- Repeat
  - Run training set $\mathcal{S}$ on model $M$
  - Count # of times transition $i \Rightarrow j$ is made
  - Count # of times letter $x$ is emitted from state $i$
  - Update parameters $\Theta$
- Until (some stopping condition)
How to model Pairwise Sequence Alignment

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

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

START → Skip Module → Align Module → Skip Module → END

How to model Pairwise Local Alignments with gaps?

START → Skip Module → Align Module → Skip Module → END
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 \( \pi \)

\[ \text{Score}(S) = -\log(P(\pi \mid S, M)) \]
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 the actualized 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 \textit{start codon ATG} and \textit{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’

Template Strand

Nontemplate or Coding Strand

Transcription unit

RNA-coding region

Promoter

RNA starts

Terminator

Upstream

Downstream

Slide courtesy Prof. Mathee
Transcriptional unit and single gene mature mRNA

Transcriptional unit

5' 3'

-35 -10 +1

Promoter

Transcription start site

ORF

RNA-coding region

Terminator

mRNA

5' 3'

Start Codon

Stop Codon

RBS

Protein-coding region

5' untranslated region

5' UTR

Leader

3' UTR

Trailer

RBS

Ribosome binding site
Prokaryotic Gene Characteristics

DNA PATTERNS IN THE E. coli lexA GENE

<table>
<thead>
<tr>
<th>GENE SEQUENCE</th>
<th>PATTERN</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. GAATTCGATGTAATTCCTGTGTTTATTGACAGTTTATTGTTT</td>
<td>CTGNNNNNRRRNCAG</td>
</tr>
<tr>
<td>41. CCARATCAGCTTTTNGCTTATACCGACATACCTG</td>
<td>CTCGAC</td>
</tr>
<tr>
<td>CCAA -35</td>
<td>CTGNNNNNRRRNCAG</td>
</tr>
<tr>
<td>81. TATTAACACCCAGGCCGAGGACTAAGCGGTGTTACGGCCA</td>
<td>TATAT, &gt; mRNA start</td>
</tr>
<tr>
<td>+10. GGGG Ribosomal binding site</td>
<td>CTGNNNNNRRRNCAG</td>
</tr>
<tr>
<td>121. GCCGACAGGCGTGTGCTGATCAGATCAGACGAGAGAGCA</td>
<td>GGAGG</td>
</tr>
<tr>
<td>161. CCAGACAGCTGCGCGCAGCGCGGCGGAAATCATCGAGAGAGCA</td>
<td>ANG</td>
</tr>
<tr>
<td>201. CGTTGCGGTGCCTCGTCGACCAGCGCGGTGAAGACAGATCG</td>
<td>OPEN READING FRAME</td>
</tr>
<tr>
<td>241. TGAAGGCGCTGCGACGAGGACTAAGCGGTGTTACGGCCA</td>
<td>CTCGAC</td>
</tr>
<tr>
<td>281. CGCGCATGACGCGGCTGCTGATCAGATCAGACGAGAGAGCA</td>
<td>GGAGG</td>
</tr>
<tr>
<td>321. GAAGGCGTGCCTCGTCGACCAGCGCGGTGAAGACAGATCG</td>
<td>OPEN READING FRAME</td>
</tr>
<tr>
<td>361. CACTTTGCGCTGCTGATCAGATCAGACGAGAGAGCA</td>
<td>CTCGAC</td>
</tr>
<tr>
<td>401. CGAGTCTCCGCTGCTGATCAGATCAGACGAGAGAGCA</td>
<td>GGAGG</td>
</tr>
<tr>
<td>441. CCAGACAGCTGCGCGCAGCGCGGCGGAAATCATCGAGAGAGCA</td>
<td>ANG</td>
</tr>
<tr>
<td>481. ACGAGGCGCTGCGACGAGGACTAAGCGGTGTTACGGCCA</td>
<td>OPEN READING FRAME</td>
</tr>
<tr>
<td>521. TAAGGCGTGCCTCGTCGACCAGCGCGGTGAAGACAGATCG</td>
<td>CTCGAC</td>
</tr>
<tr>
<td>561. CCAGACAGCTGCGCGCAGCGCGGCGGAAATCATCGAGAGAGCA</td>
<td>GGAGG</td>
</tr>
<tr>
<td>601. CACTTTGCGCTGCTGATCAGATCAGACGAGAGAGCA</td>
<td>CTCGAC</td>
</tr>
<tr>
<td>641. CGAGTCTCCGCTGCTGATCAGATCAGACGAGAGAGCA</td>
<td>GGAGG</td>
</tr>
<tr>
<td>681. CGTTGCGGTGCCTCGTCGACCAGCGCGGTGAAGACAGATCG</td>
<td>OPEN READING FRAME</td>
</tr>
<tr>
<td>721. ACGAGGCGCTGCGACGAGGACTAAGCGGTGTTACGGCCA</td>
<td>CTCGAC</td>
</tr>
<tr>
<td>761. CGAGTCTCCGCTGCTGATCAGATCAGACGAGAGAGCA</td>
<td>GGAGG</td>
</tr>
<tr>
<td>801. CCAGACAGCTGCGCGCAGCGCGGCGGAAATCATCGAGAGAGCA</td>
<td>ANG</td>
</tr>
<tr>
<td>841. CACTTTGCGCTGCTGATCAGATCAGACGAGAGAGCA</td>
<td>CTCGAC</td>
</tr>
<tr>
<td>881. CGTTGCGGTGCCTCGTCGACCAGCGCGGTGAAGACAGATCG</td>
<td>OPEN READING FRAME</td>
</tr>
<tr>
<td>921. GAAGGCGCTGCGACGAGGACTAAGCGGTGTTACGGCCA</td>
<td>CTCGAC</td>
</tr>
</tbody>
</table>

Shown are matches to approximate consensus binding sites for LexA repressor (CTGNNNNNRRRNCAG), the -10 and -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 (ANG...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

**Initiation Codon**

- **AUG** Methionine

**Termination Codons**

- **UAA** Ochre
- **UAG** Amber
- **UGA** Opal

**Others:**

- **GUG** Valine
- **UUG** Leucine
- **AUU** Isoleucine

**Untranslated leader**

**Intracistronic Distance**

1-40 bp Trailer

**Coding region**

Open Reading Frame (ORF)

- **Start**
- **Stop**

**RBS**

Ribosome Binding Site

Shine-Dalgarno Sequence

**5’--AGGAGG--3’**

**7 bp upstream of start codon**

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

Slide courtesy Prof. Mathee
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

![Genetic Code Table](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*

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

Start → UTR → State1 → OneExon → State4 → End

UTR

1st Exon → Exon

State2

State3 → Last Exon

Intron
Transcriptional machinery: RNA Polymerase and DNA

RNAP Holoenzyme

Basal Promoter:

-35  
-10  
4-8 bp  
Transcription Starts

E. coli consensus for $\sigma^{70}$

TTGACA TATAAT

Spacer Region $16\sim18$ bp

Stronger Promoter:

-35  
-10  
Transcription Starts

UP Element

α-CTD makes the contact

AT-rich

2/10/05

CAP5510/CGS5166 (Lec 10)

Slide courtesy Prof. Mathee