

CAP 5510: Introduction to Bioinformatics  
CGS 5166: Bioinformatics Tools

**Giri Narasimhan**

ECS 254; Phone: x3748

[giri@cis.fiu.edu](mailto:giri@cis.fiu.edu)

[www.cis.fiu.edu/~giri/teach/BioinfF18.html](http://www.cis.fiu.edu/~giri/teach/BioinfF18.html)

---

# Profiles and HMMs



# Pattern: Representations

GAGGTA AAC  
TCCGTA AGT  
CAGGTTGGA  
ACAGTCAGT  
TAGGTCATT  
TAGGTACTG  
ATGGTAACT  
CAGGTATAC  
TGTGTGAGT  
AAGGTAAGT

- Alignments
- Consensus Sequences
- Logo Formats
- ...

TAGGTAAGT



TAGGTAAGT

# Profiles

GAGGTA AAC  
 TCCGTA AGT  
 CAGGTT GGA  
 ACAGTC AGT  
 TAGGTC ATT  
 TAGGTA CTG  
 ATGGTA ACT  
 CAGGTAT AC  
 TGTGTG AGT  
 AAGGTA AGT

	1	2	3	4	5	6	7	8	9
A	3	6	1	0	0	6	7	2	1
C	2	2	1	0	0	2	1	1	2
G	1	1	7	10	0	1	1	5	1
T	4	1	1	0	10	1	1	2	6

Frequency  
Matrix

	1	2	3	4	5	6	7	8	9
A	.3	.6	.1	0	0	.6	.7	.2	.1
C	.2	.2	.1	0	0	.2	.1	.1	.2
G	.1	.1	.7	1	0	.1	.1	.5	.1
T	.4	.1	.1	0	1	.1	.1	.2	.6

Relative  
Frequencies

# Profiles

GAGGTA AAC

TCCGTA AGT

CAGGTT GGA

ACAGTC AGT

TAGGTC ATT

TAGGTA CTG

ATGGTA ACT

CAGGTATA C

TGTGTG AGT

AAGGTA AGT

	1	2	3	4	5	6	7	8	9
A	.3	.6	.1	0	0	.6	.7	.2	.1
C	.2	.2	.1	0	0	.2	.1	.1	.2
G	.1	.1	.7	1	0	.1	.1	.5	.1
T	.4	.1	.1	0	1	.1	.1	.2	.6

Relative  
Frequencies

	1	2	3	4	5	6	7	8	9
A	0.14	0.72	-0.61	-1.43	-1.43	0.72	0.86	-0.16	-0.61
C	-0.16	-0.16	-0.61	-1.43	-1.43	-0.16	-0.61	-0.61	-0.16
G	-0.61	-0.61	0.86	-0.61	-1.43	-0.61	-0.61	0.57	-0.61
T	0.38	-0.61	-0.61	-1.43	1.19	-0.61	-0.61	-0.16	0.72

# Profiles

Profile entries:

$$P_{ij} = \ln (f_{ij}/b_i)$$

Zero counts:

$$f_{ij} = (c_{ij} + \alpha b_i) / (n + \alpha)$$

	1	2	3	4	5	6	7	8	9
A	.3	.6	.1	0	0	.6	.7	.2	.1
C	.2	.2	.1	0	0	.2	.1	.1	.2
G	.1	.1	.7	1	0	.1	.1	.5	.1
T	.4	.1	.1	0	1	.1	.1	.2	.6

Relative  
Frequencies

	1	2	3	4	5	6	7	8	9
A	0.14	0.72	-0.61	-1.43	-1.43	0.72	0.86	-0.16	-0.61
C	-0.16	-0.16	-0.61	-1.43	-1.43	-0.16	-0.61	-0.61	-0.16
G	-0.61	-0.61	0.86	1.19	-1.43	-0.61	-0.61	0.57	-0.61
T	0.38	-0.61	-0.61	-1.43	1.19	-0.61	-0.61	-0.16	0.72

<http://coding.plantpath.ksu.edu/profile/>

# CpG Islands

- ❑ Regions in DNA sequences with increased occurrences of substring "CG"
- ❑ Rare: typically C gets methylated and then mutated into a T.
- ❑ Often around promoter or "start" regions of genes
- ❑ Few hundred to a few thousand bases long

## Problem 1:

- **Input:** Small sequence **S**
- **Output:** Is **S** from a CpG island?
  - Build Markov models:  $M^+$  and  $M^-$
  - Then compare



# Markov Models

<b>+</b>	<b>A</b>	<b>C</b>	<b>G</b>	<b>T</b>
<b>A</b>	0.180	0.274	0.426	0.120
<b>C</b>	0.171	0.368	0.274	0.188
<b>G</b>	0.161	0.339	0.375	0.125
<b>T</b>	0.079	0.355	0.384	0.182

<b>-</b>	<b>A</b>	<b>C</b>	<b>G</b>	<b>T</b>
<b>A</b>	0.300	0.205	0.285	0.210
<b>C</b>	0.322	0.298	0.078	0.302
<b>G</b>	0.248	0.246	0.298	0.208
<b>T</b>	0.177	0.239	0.292	0.292

# How to distinguish?

## □ Compute

$$S(x) = \log\left(\frac{P(x|M+)}{P(x|M-)}\right) = \sum_{i=1}^L \log\left(\frac{p_{x(i-1)x_i}}{m_{x(i-1)x_i}}\right) = \sum_{i=1}^L r_{x(i-1)x_i}$$

r=p/m	A	C	G	T
A	-0.740	0.419	0.580	-0.803
C	-0.913	0.302	1.812	-0.685
G	-0.624	0.461	0.331	-0.730
T	-1.169	0.573	0.393	-0.679

**Score(GCAC)**

$$= .461 - .913 + .419 < 0.$$

**GCAC** not from CpG island.

**Score(GCTC)**

$$= .461 - .685 + .573 > 0.$$

**GCTC** from CpG island.

## Problem 1:

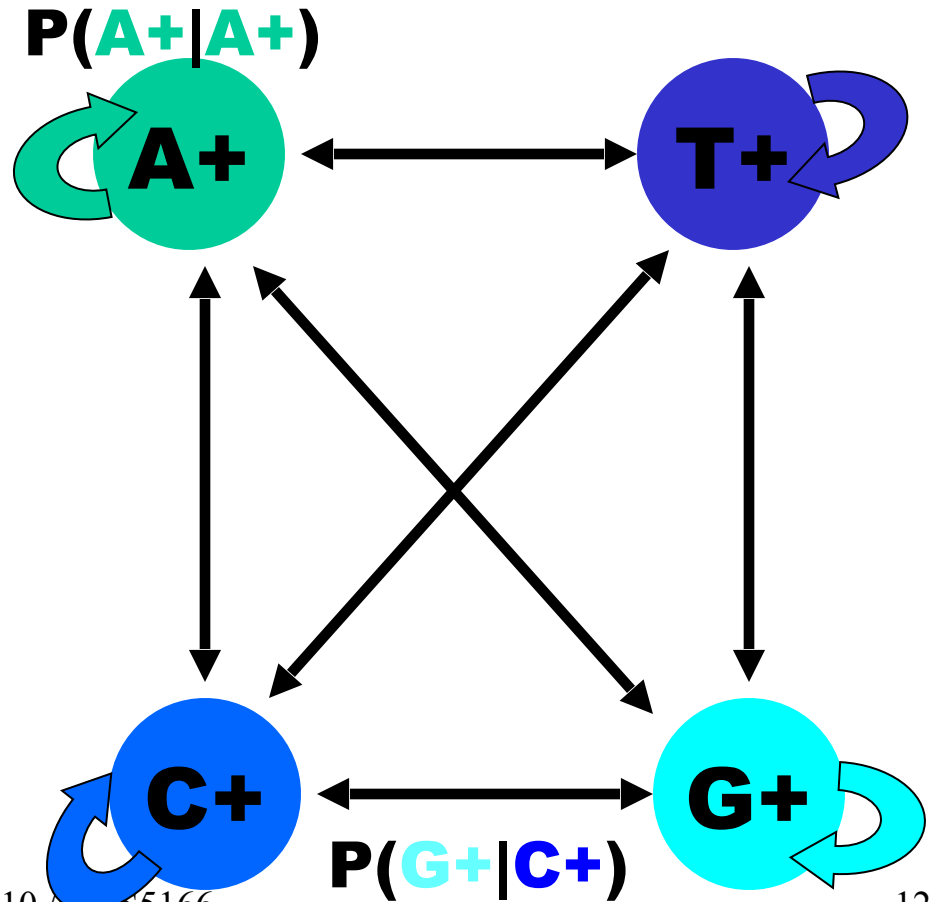
- **Input:** Small sequence **S**
- **Output:** Is **S** from a CpG island?
  - Build Markov Models:  $M^+$  &  $M^-$
  - Then compare

## Problem 2:

- **Input:** Long sequence **S**
- **Output:** Identify the CpG islands in **S**.
  - Markov models are inadequate.
  - Need Hidden Markov Models.

# Markov Models

<b>+</b>	<b>A</b>	<b>C</b>	<b>G</b>	<b>T</b>
<b>A</b>	0.180	0.274	0.426	0.120
<b>C</b>	0.171	0.368	0.274	0.188
<b>G</b>	0.161	0.339	0.375	0.125
<b>T</b>	0.079	0.355	0.384	0.182

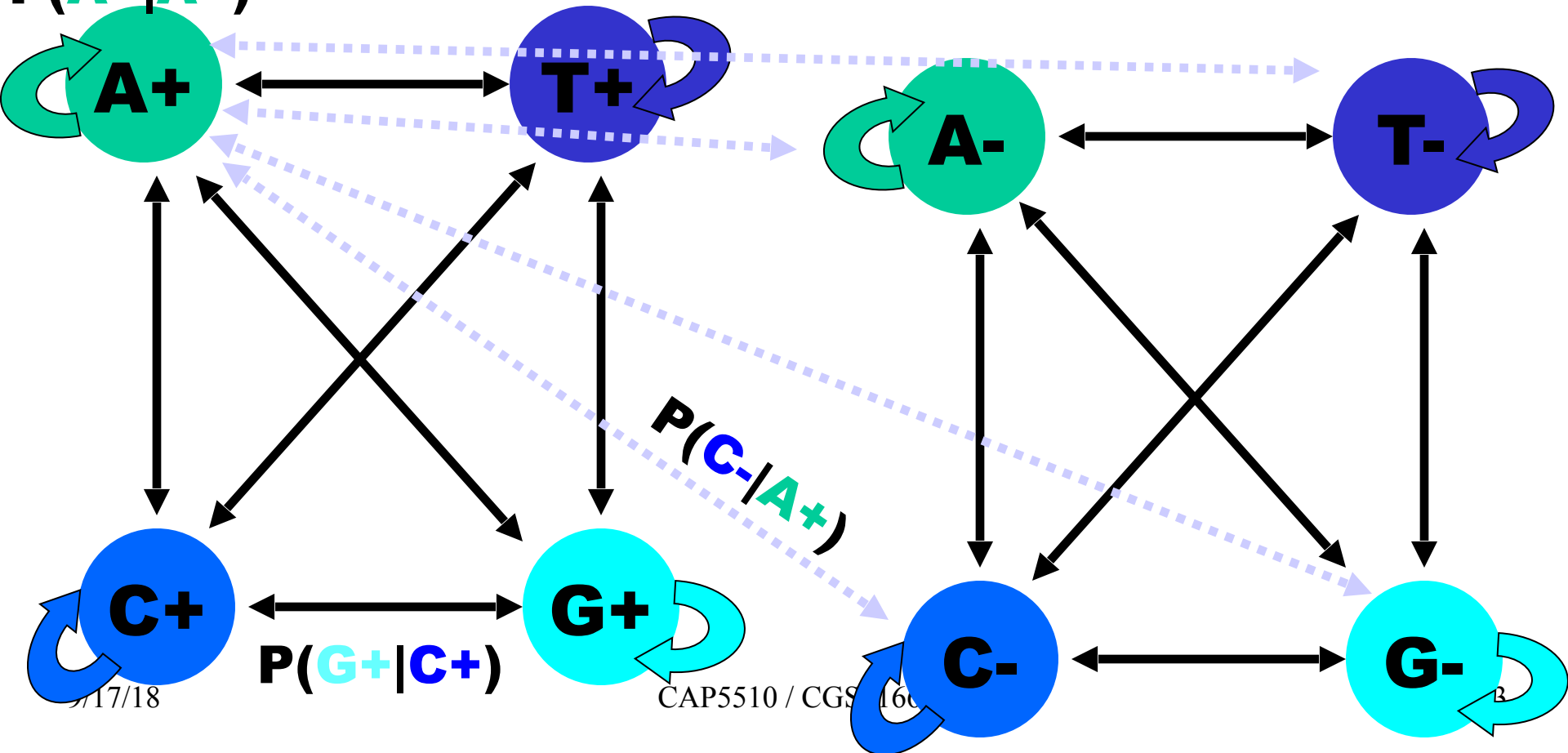


# CpG Island + in an ocean of -

## First order Hidden Markov Model

MM=16, HMM= 64 transition probabilities (adjacent bp)

$P(A+|A+)$

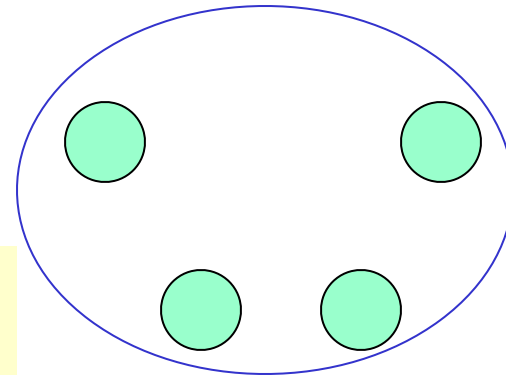
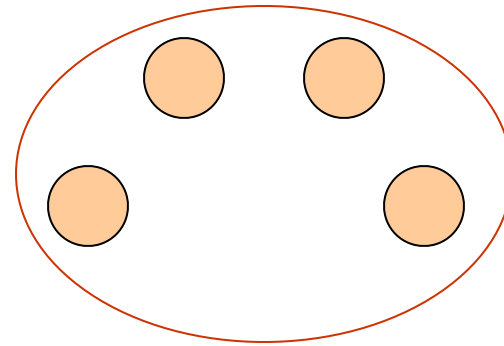


# Hidden Markov Model (HMM)

- 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.



# 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  $\Pi[i,j]$  = 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  $P[i,j]$ .

# Profiles

Profile entries:

$$P_{ij} = \ln (f_{ij}/b_i)$$

Zero counts:

$$f_{ij} = (c_{ij} + \alpha b_i) / (n + \alpha)$$

	1	2	3	4	5	6	7	8	9
A	.3	.6	.1	0	0	.6	.7	.2	.1
C	.2	.2	.1	0	0	.2	.1	.1	.2
G	.1	.1	.7	1	0	.1	.1	.5	.1
T	.4	.1	.1	0	1	.1	.1	.2	.6

Relative  
Frequencies

	1	2	3	4	5	6	7	8	9
A	0.14	0.72	-0.61	-1.43	-1.43	0.72	0.86	-0.16	-0.61
C	-0.16	-0.16	-0.61	-1.43	-1.43	-0.16	-0.61	-0.61	-0.16
G	-0.61	-0.61	0.86	1.19	-1.43	-0.61	-0.61	0.57	-0.61
T	0.38	0.61	0.61	1.43	1.19	0.61	0.61	0.16	0.72

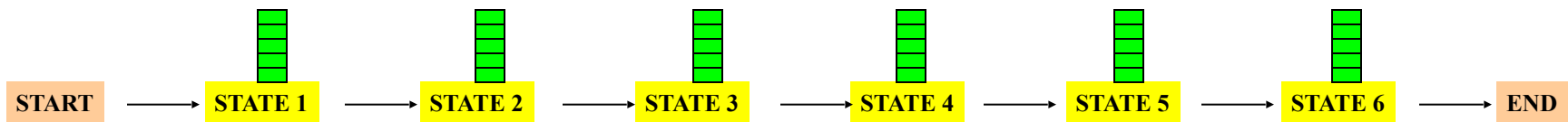
Profiles; Position Weight Matrix (PWM);  
Position-Specific Scoring Matrix (PSSM)



# Profile HMMs

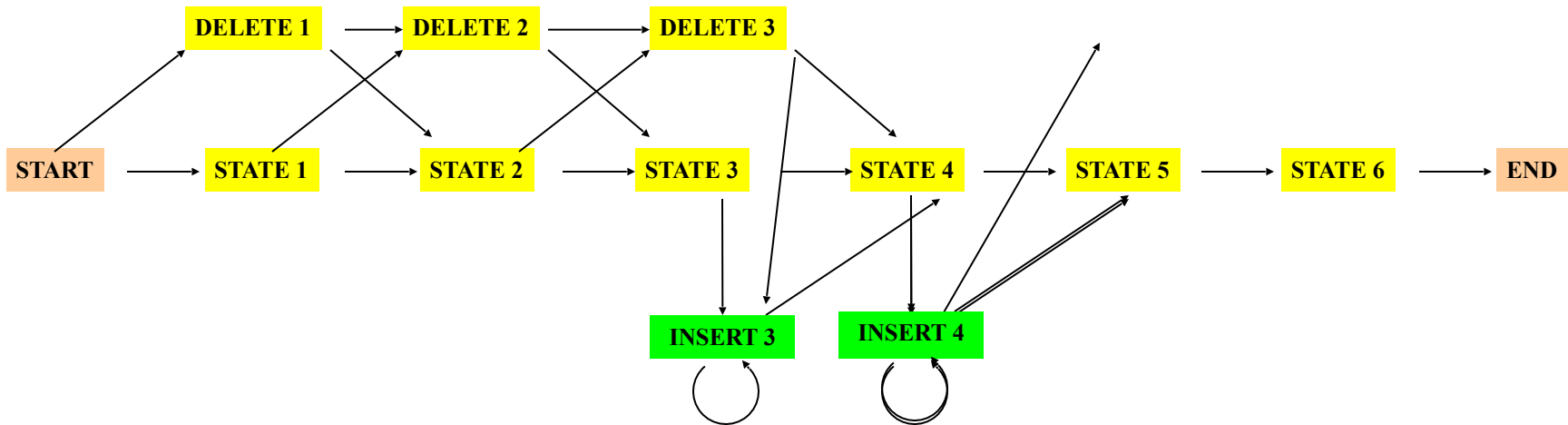
PROFILE METHOD, [M. Gribskov et al., '90]

Location in Seq.	Sequence						Protein Name
	1	2	3	4	5	6	
14	G	V	S	A	S	A	Ka RotR
32	G	V	S	E	M	T	Ec DsoR
33	G	V	S	P	G	T	Ec RpoD
76	G	A	G	I	A	T	Ec TrpR
178	G	C	S	R	R	T	Fc 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

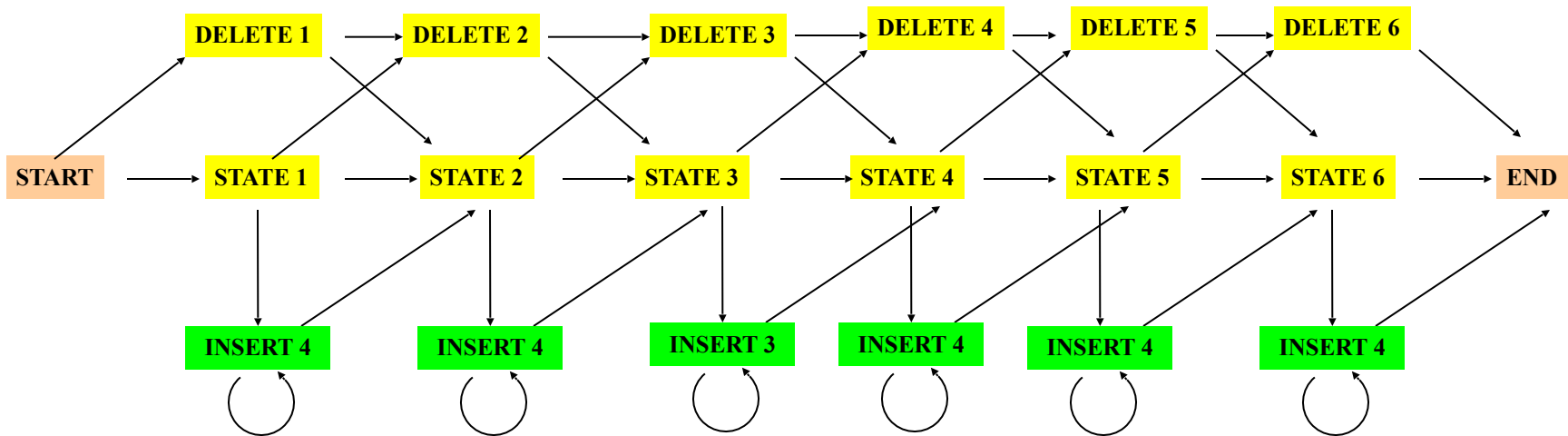


# Profile HMMs with InDels

- Insertions
- Deletions
- Insertions & Deletions



# Profile HMMs with InDels



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

# HMM for Sequence Alignment

## 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

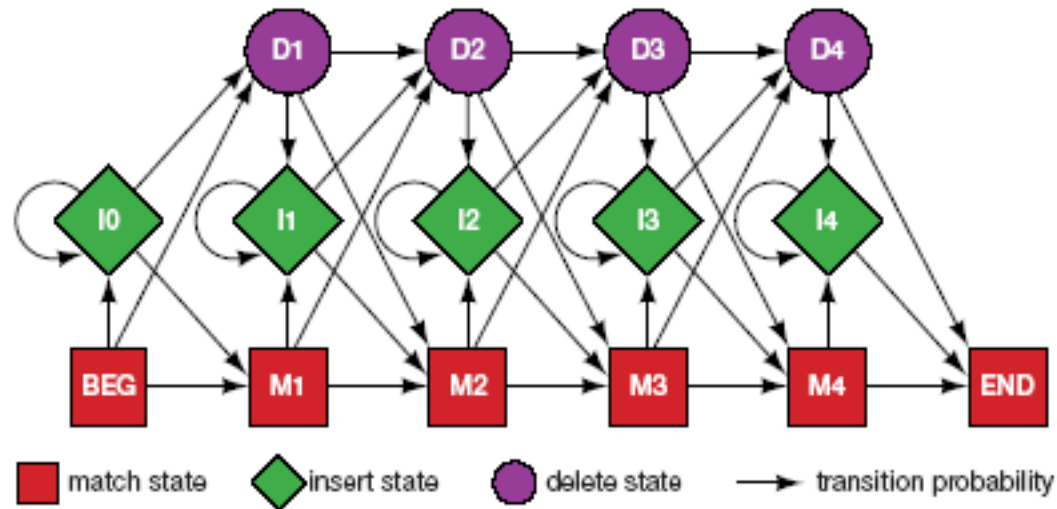



FIGURE 5.16. Relationship between the sequence alignment and the hidden Markov model of the alignment (Krogh et al. 1994). This particular form for the HMM was chosen to represent the sequence, structural, and functional variation expected in proteins. The model accommodates the identities, mismatches, insertions, and deletions expected in a group of related proteins. (A) A section of an msa. The illustration shows the columns generated in an msa. Each column may include matches and mismatches (*red* 9/17/18), insertions (*green* positions), and deletions (*purple* positions). (B) The HMM. Each column in the model represents the possibility of a match, insert, or delete in each column of the alignment in A. The HMM is a probabilistic representation of a section of the msa. Sequences can be generated from the HMM by starting at the beginning state labeled BEG and then by following

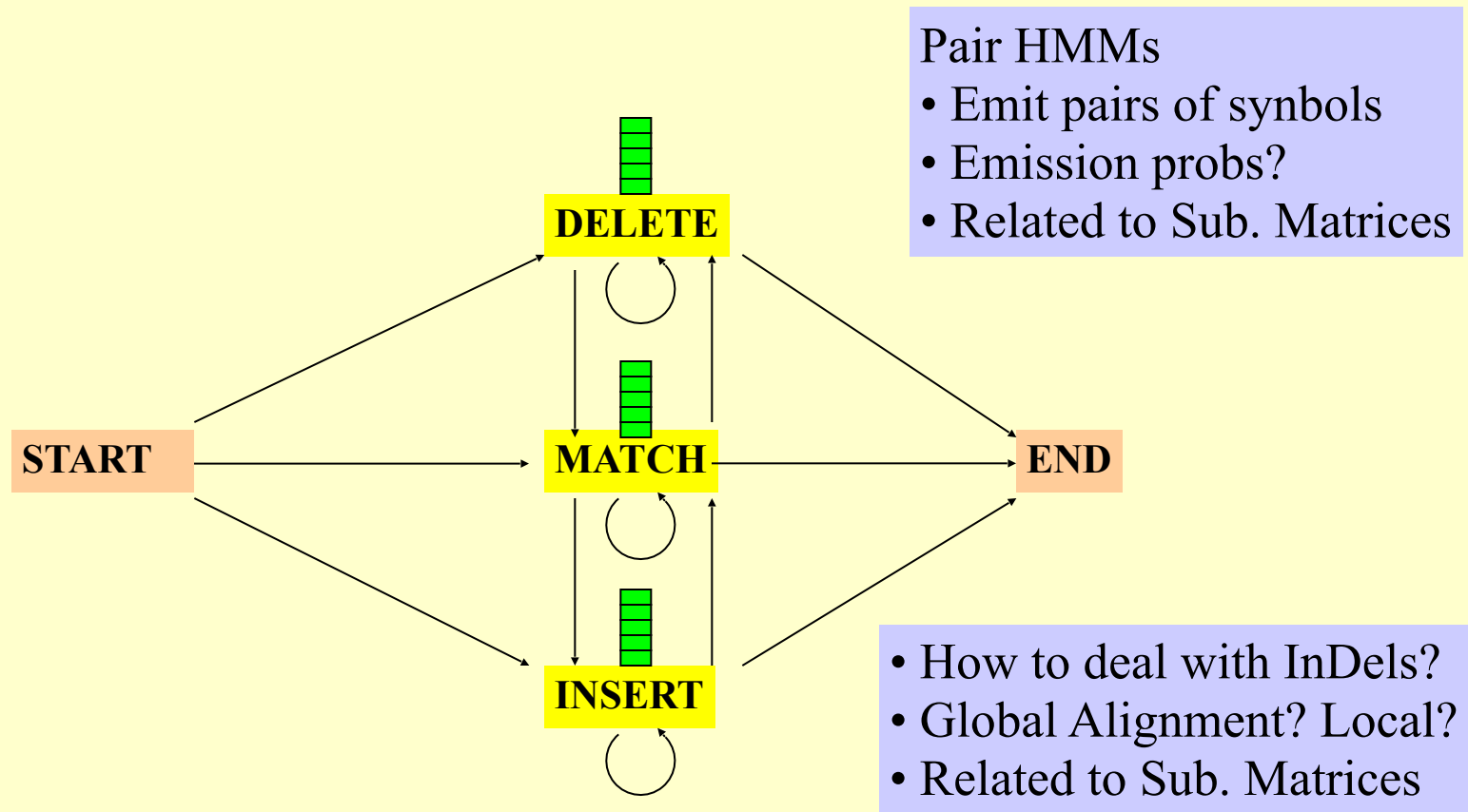
# Profile HMM Software

- ❑ **HMMER** <http://hmmer.wustl.edu/>
- ❑ **SAM** <http://www.cse.ucsc.edu/research/compbio/sam.html>
- ❑ **PFTOOLS** <http://www.isrec.isb-sib.ch/ftp-server/pftools/>
- ❑ **HMMpro** <http://www.netid.com/html/hmmpro.html>
- ❑ **GENEWISE** <http://www.ebi.ac.uk/Wise2/>
- ❑ **PROBE** <ftp://ftp.ncbi.nih.gov/pub/neuwald/probe1.0/>
- ❑ **META-MEME** <http://metameme.sdsc.edu/>
- ❑ **BLOCKS** <http://www.blocks.fhcrc.org/>
- ❑ **PSI-BLAST** <http://www.ncbi.nlm.nih.gov/BLAST/newblast.html>
  
- ❑ Read more about Profile HMMs at  <http://www.csb.yale.edu/userguides/seq/hmmer/docs/node9.html>

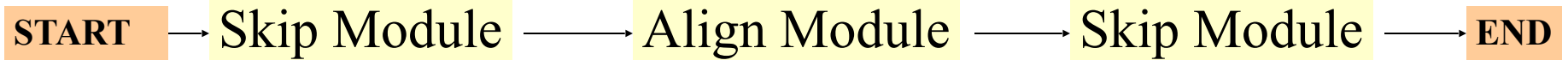
# How to model Pairwise Sequence Alignment

LEAPVE

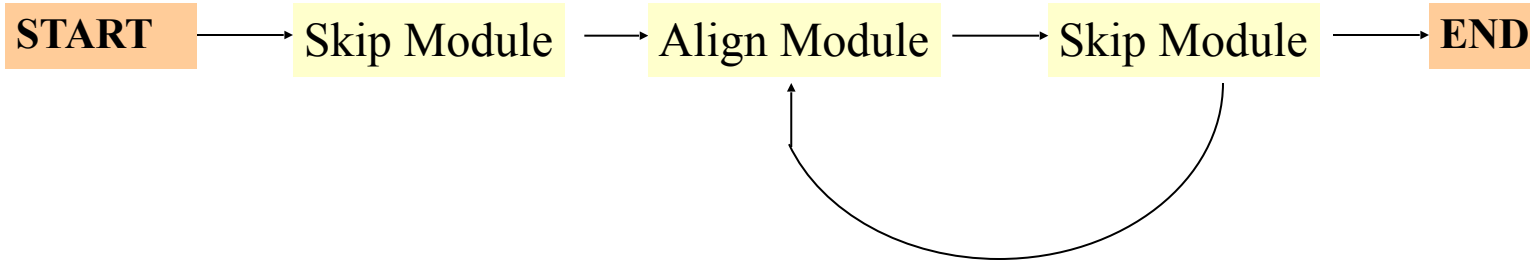
LAPVIE



# How to model Pairwise Local Alignments?

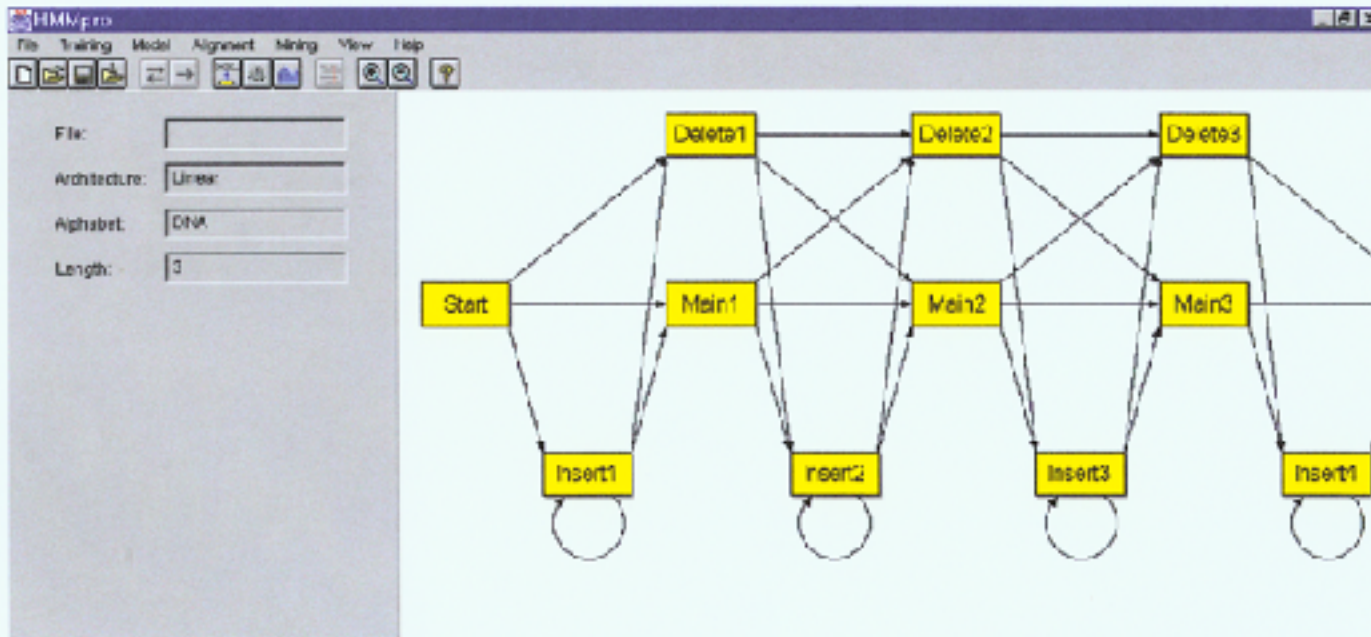


## How to model Pairwise Local Alignments with gaps?



# Standard HMM architectures

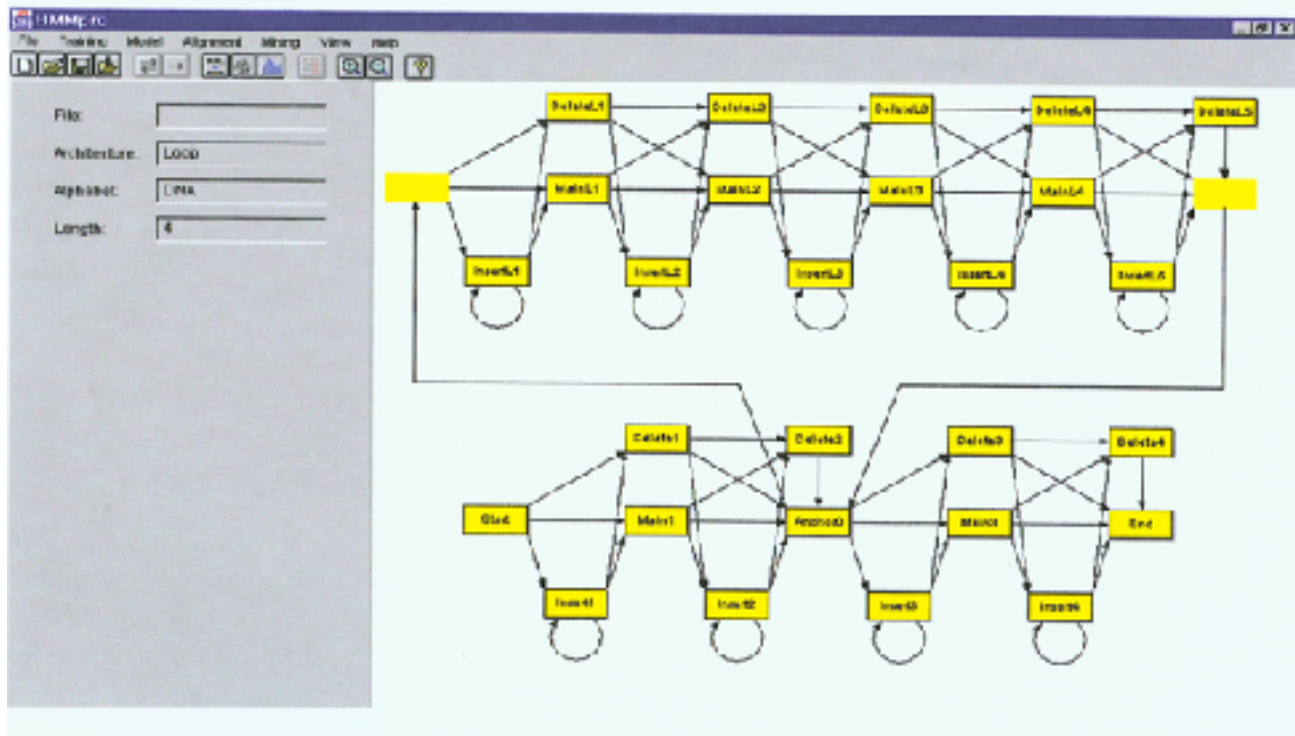
## Linear Architecture





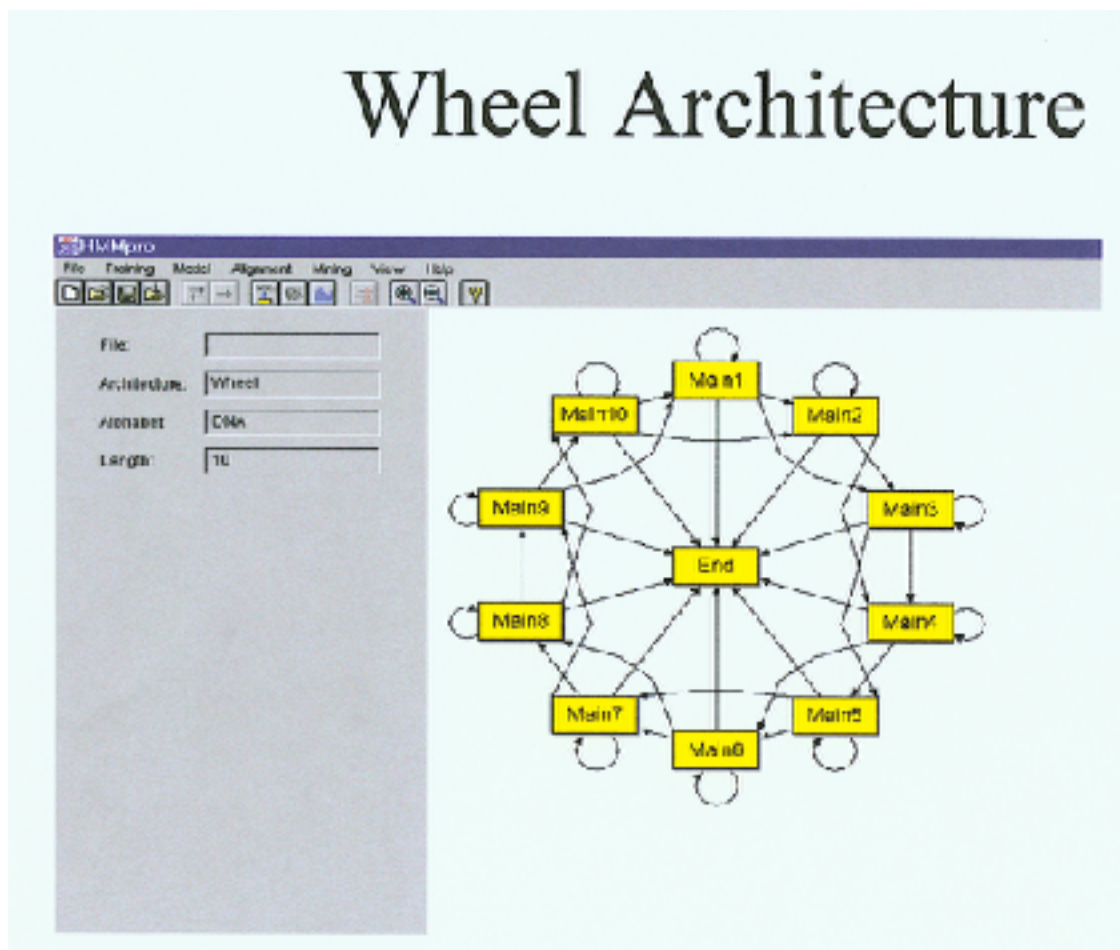
# Standard HMM architectures

## Loop Architecture



# Standard HMM architectures

## Wheel Architecture



### Problem 3: LIKELIHOOD QUESTION

- **Input:** Sequence  $S$ , model  $M$ , state  $i$
- **Output:** Compute the probability of reaching state  $i$  with sequence  $S$  using model  $M$ 
  - **Backward Algorithm (DP)**

### Problem 4: LIKELIHOOD QUESTION

- **Input:** Sequence  $S$ , model  $M$
- **Output:** Compute the probability that  $S$  was emitted by model  $M$ 
  - **Forward Algorithm (DP)**

## Problem 5: LEARNING QUESTION

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

# 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  $\alpha$ -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

$$H_i = - \sum_a \theta_a \log \theta_a$$

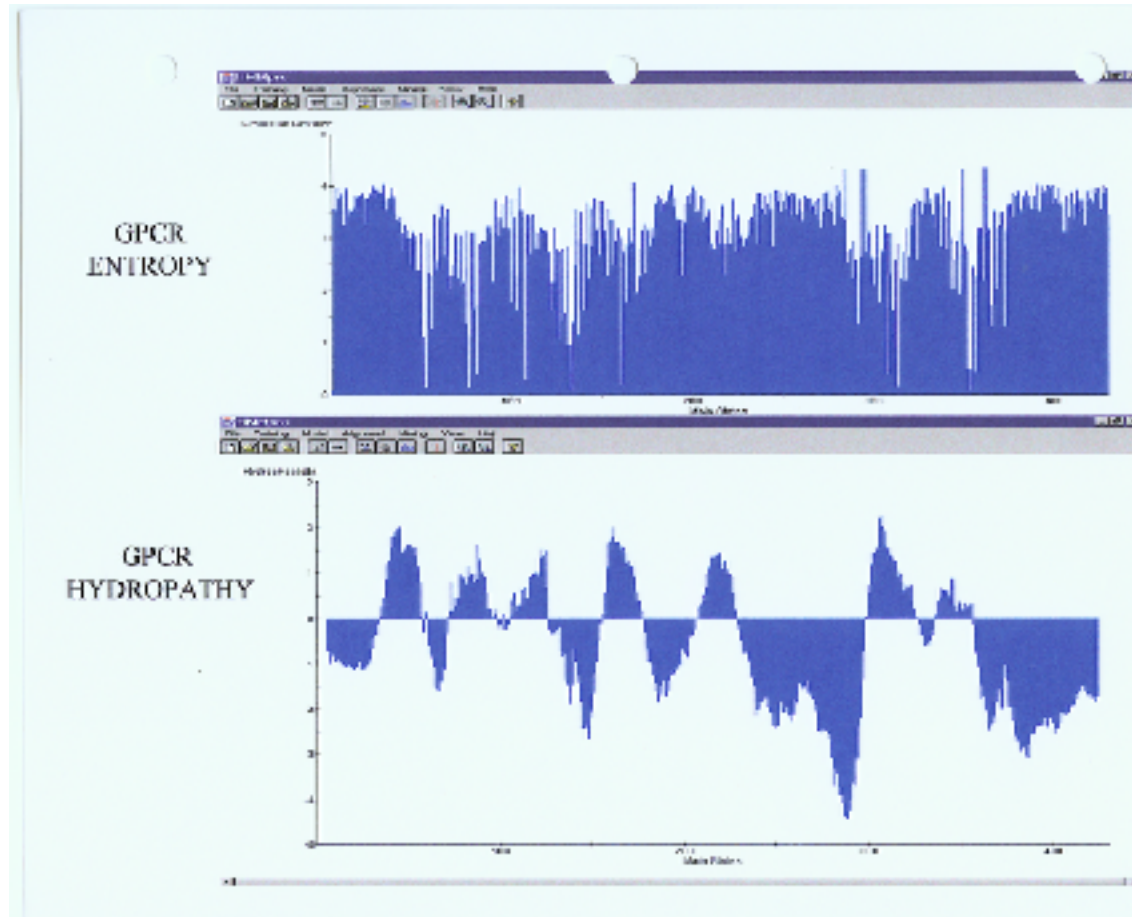
- For every sequence from test set (142) & random set (1600) & all SWISS-PROT proteins

● Compute the negative log of probability of the most probable state

$$\text{Score}(S) = -\log(P(\pi | S, M))$$



# GPCR Analysis



# Entropy

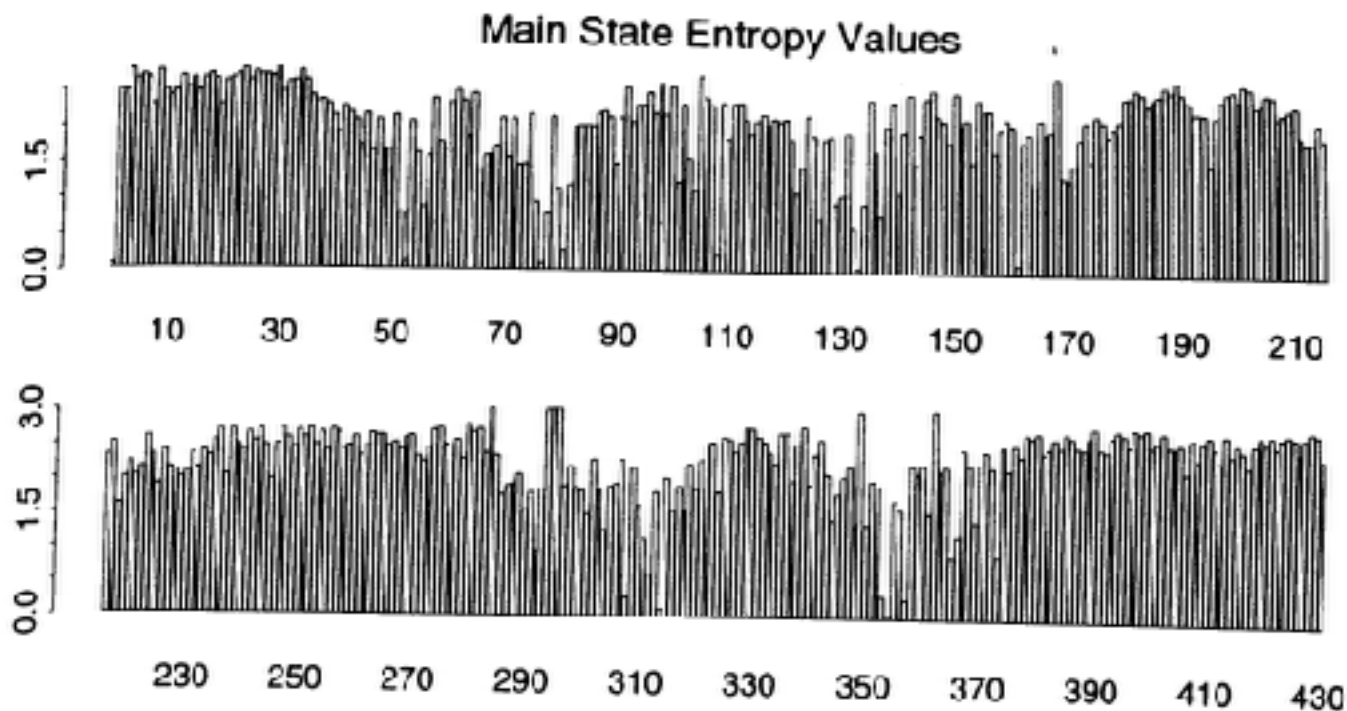


Figure 3.1: Entropy Profile of the Emission Probability Distributions Associated with the Main States of the HMM After 12 Cycles of Training.

# GPCR Analysis (Cont'd)

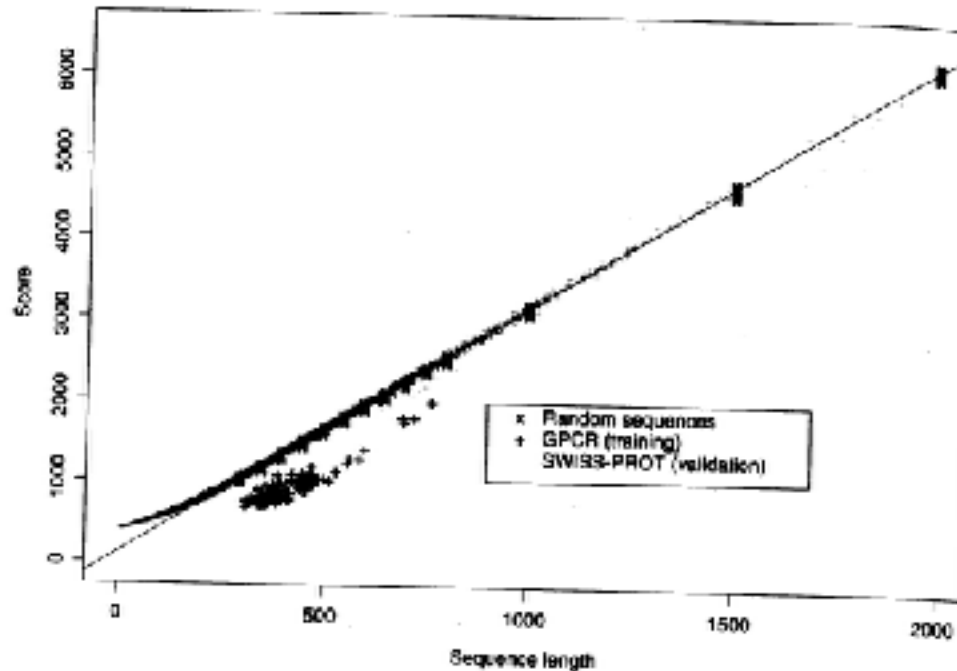


Figure 8.2: Scores (Negative Log-likelihoods of Optimal Viterbi Paths). Represented sequences consist of 142 GPCR training sequences, all sequences from the SWISS-PROT database of length less than or equal to 2000, and 220 randomly generated sequences with same average composition as the GPCRs of length 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.

# References

- ❑ Krogh, Brown, Mian, Sjolander, Haussler, J. Mol. Biol. 235:1501-1531, 1994
- ❑ Gribskov, Luthy, Eisenberg, Meth. Enzymol. 183:146-159, 1995
- ❑ Gribskov, McLachlan, Eisenberg, Proc Natl. Acad. Sci. 84:4355-4358, 1996.