

CAP 5510: Introduction to Bioinformatics
CGS 5166: Bioinformatics Tools

Giri Narasimhan

ECS 254; Phone: x3748

giri@cis.fiu.edu

www.cis.fiu.edu/~giri/teach/BioinfS11.html

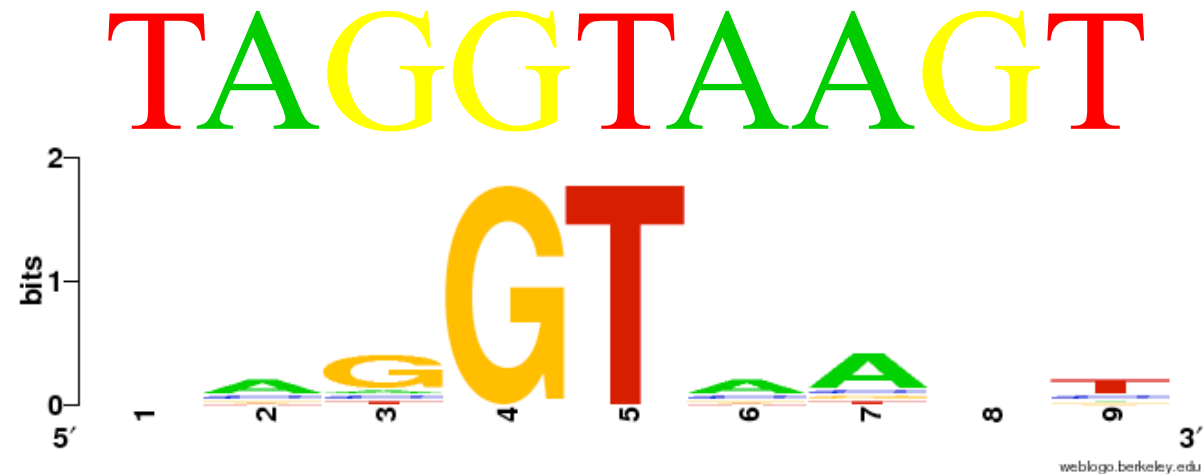
Describing & Modeling Patterns

Pattern: Representations

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

TAGGTAAGT

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



Profiles

GAGGTA AAC
TCCGTA AGT
CAGGTTG GA
ACAGTCAGT
TAGGTCATT
TAGGTACTG
ATGGTAACT
CAGGTATAC
TGTGTGAGT
AAGGTAAGT

	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

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	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

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

<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

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

-	A	C	G	T
A	0.300	0.205	0.285	0.210
C	0.322	0.298	0.078	0.302
G	0.248	0.246	0.298	0.208
T	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:

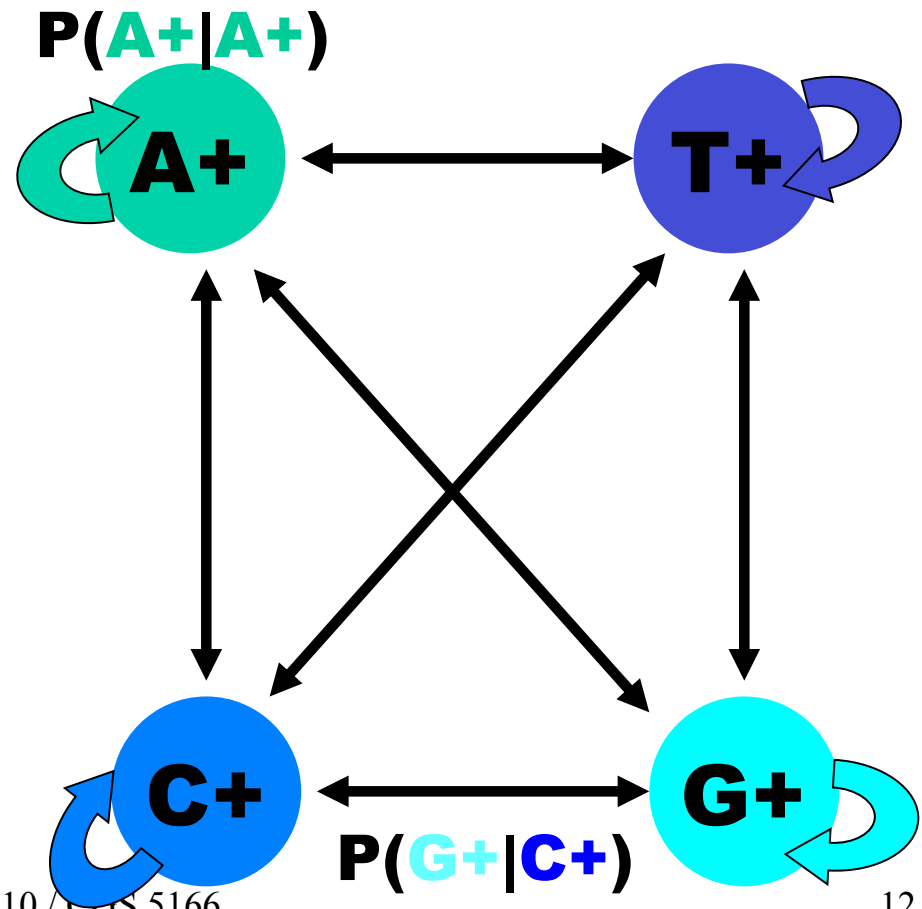
- **Input:** Small sequence **S**
- **Output:** Is **S** from a CpG island?
 - Build Markov Models: M^+ & M^-
 - Then compare

Problem 2:

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

Markov Models

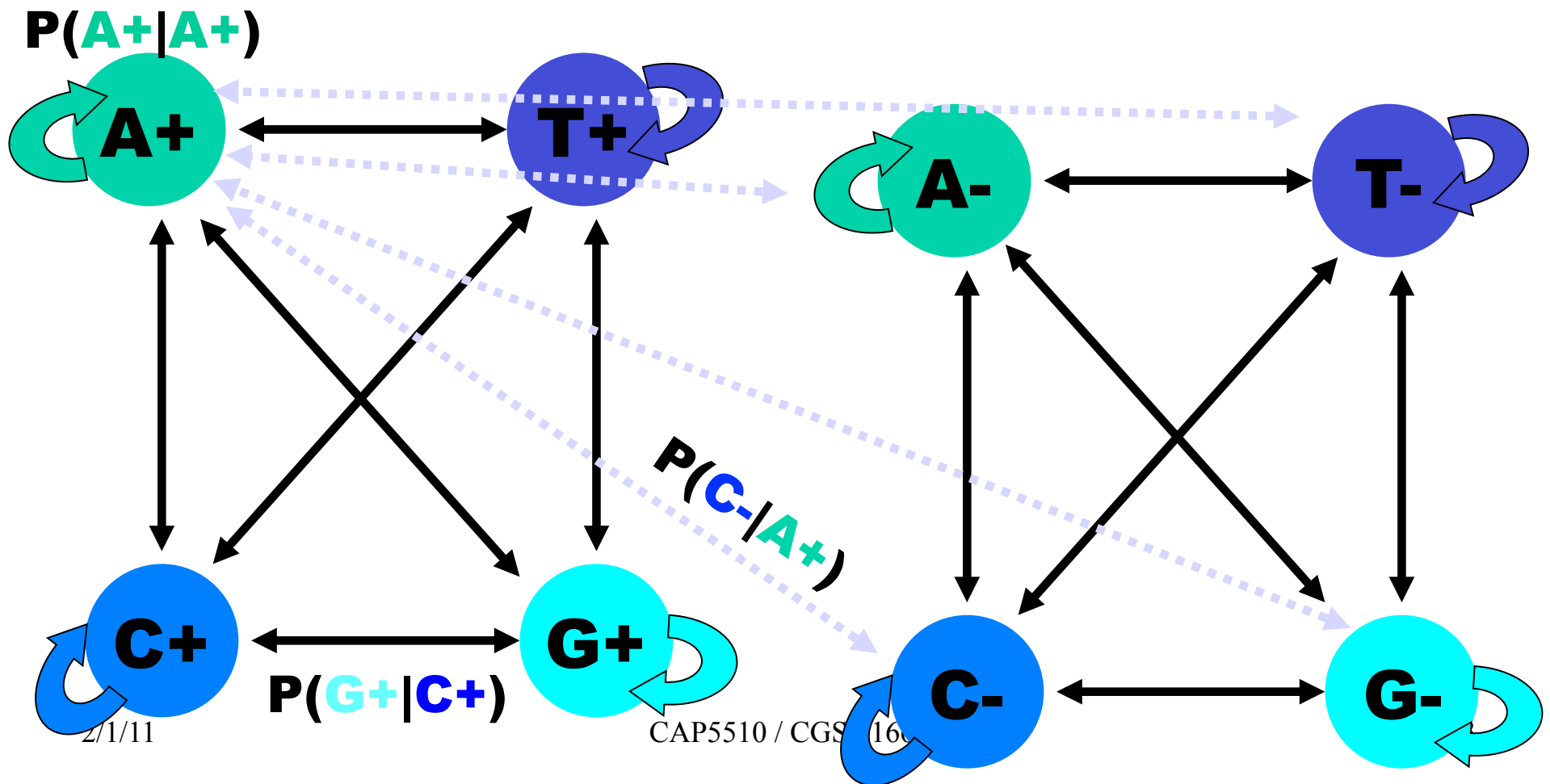
+	A	C	G	T
A	0.180	0.274	0.426	0.120
C	0.171	0.368	0.274	0.188
G	0.161	0.339	0.375	0.125
T	0.079	0.355	0.384	0.182



CpG Island + in an ocean of -

First order Hidden Markov Model

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

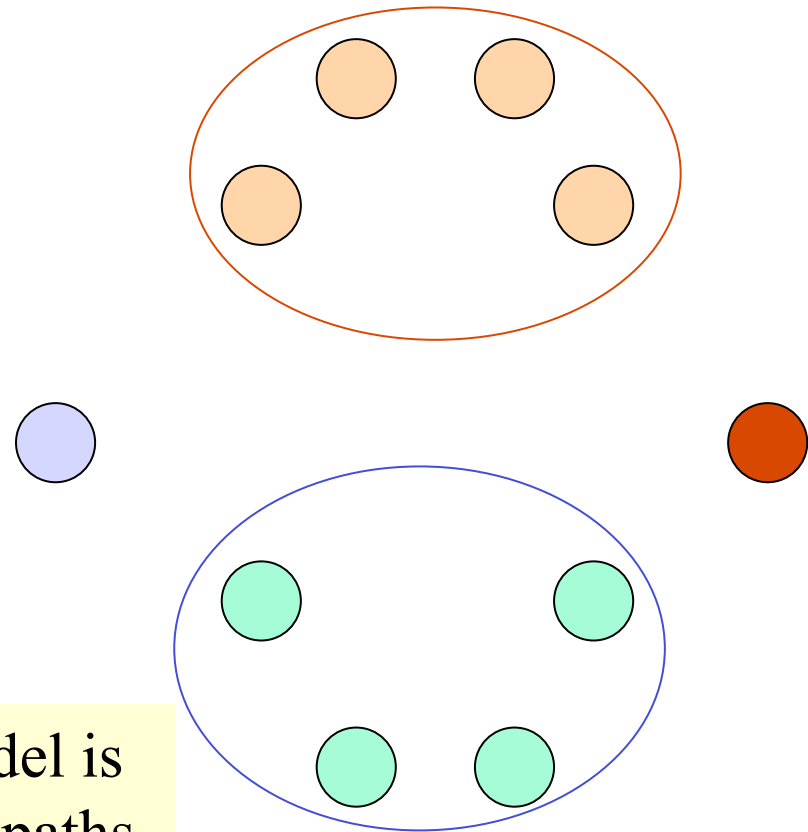


Hidden Markov Model (HMM)

- 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 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 $P[i,j]$.

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

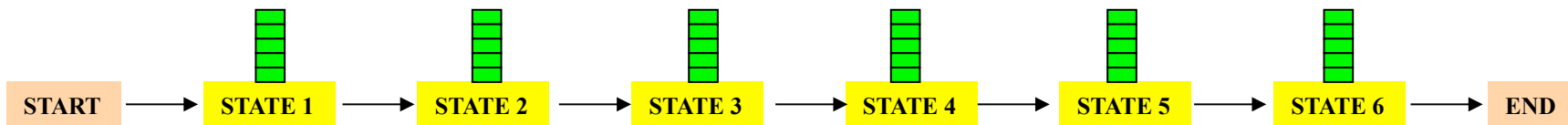
Profiles; Position Weight Matrix (PWM);
Position-Specific Scoring Matrix (PSSM)

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

Profile HMMs

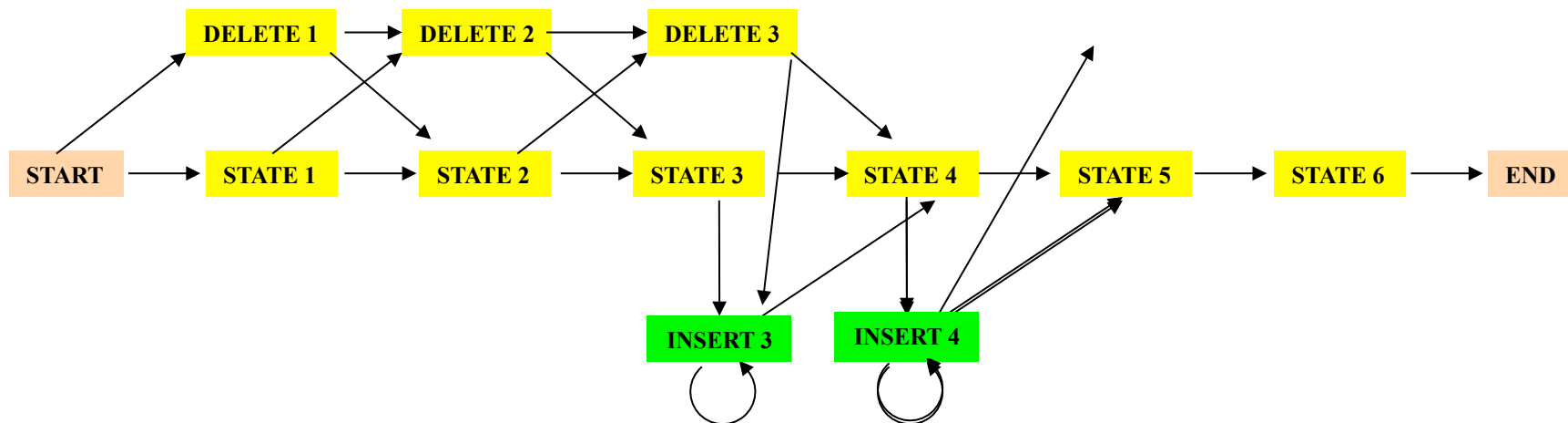
PROFILE METHOD, [M. Gribskov et al., '90]

Location in Seq.	Sequence						Protein Name
	1	2	3	4	5	6	
14	G	V	S	A	S	A	Ka 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

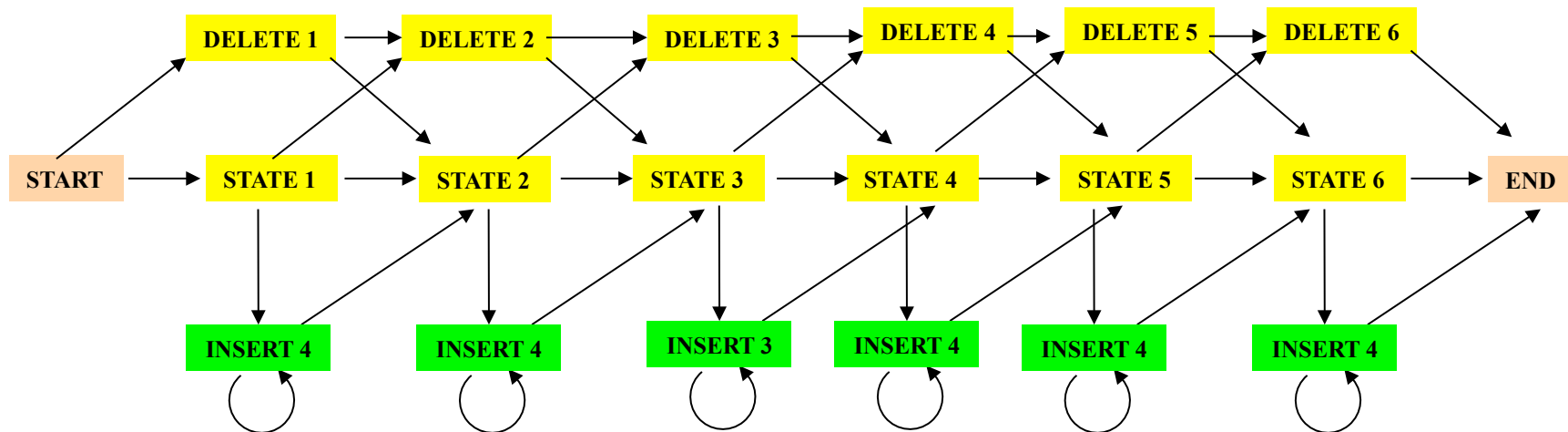


Profile HMMs with InDels

- Insertions
- Deletions
- Insertions & Deletions



Profile HMMs with InDels

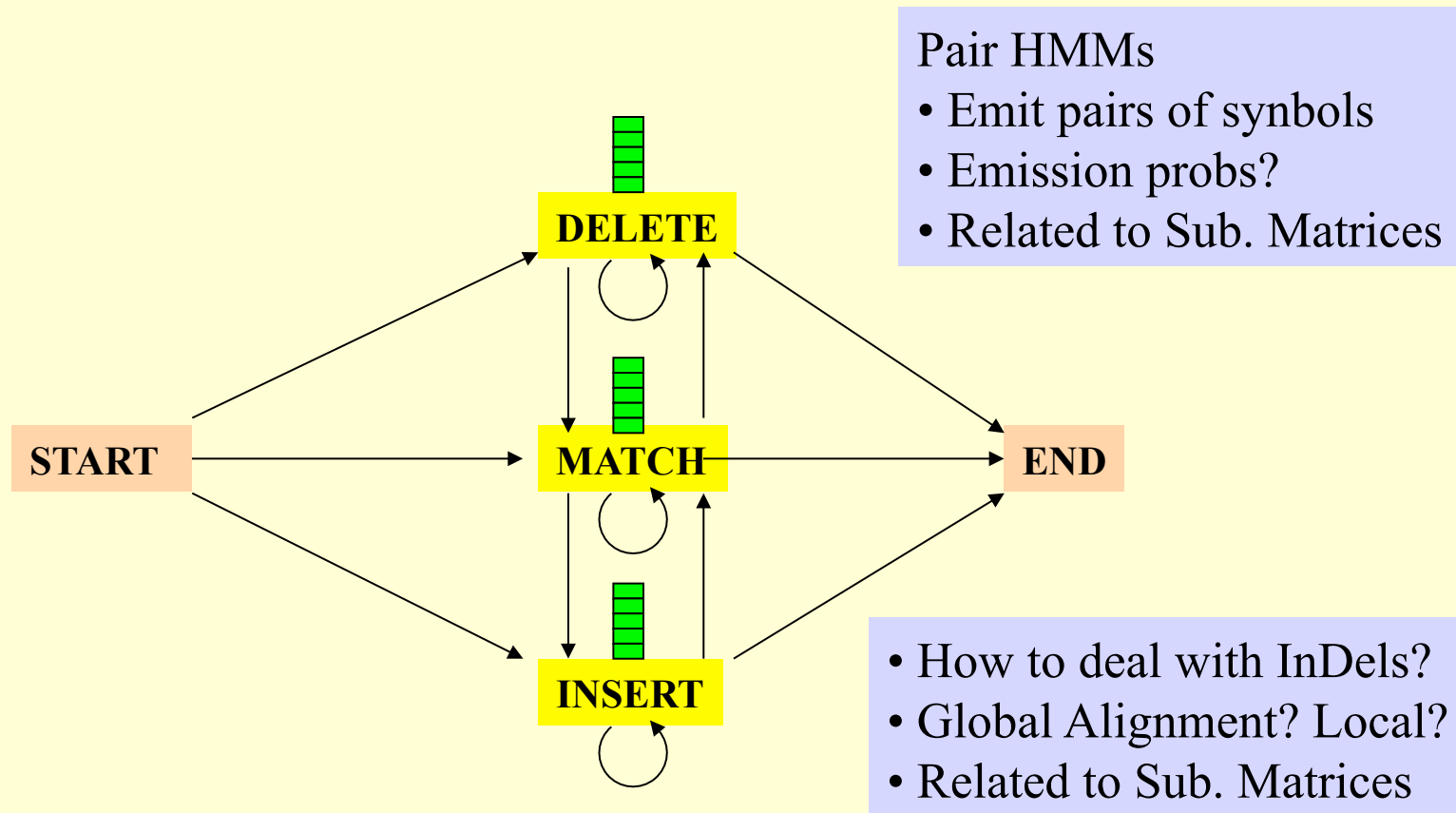


Missing transitions from **DELETE j** to **INSERT j** and
from **INSERT j** to **DELETE $j+1$** .

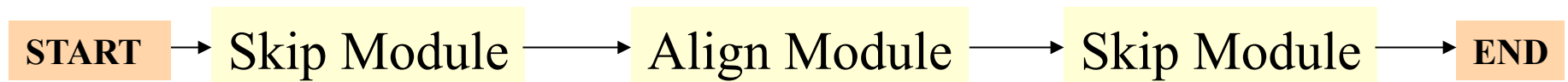
How to model Pairwise Sequence Alignment

LEAPVE

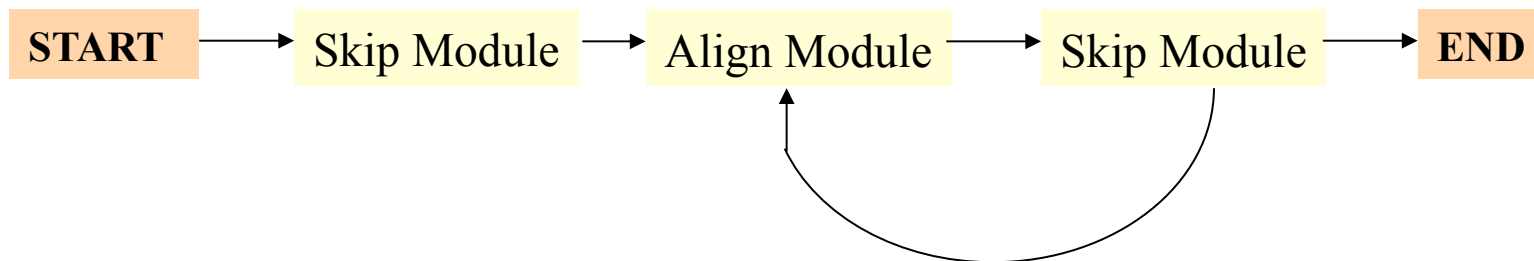
LAPVIE



How to model Pairwise Local Alignments?

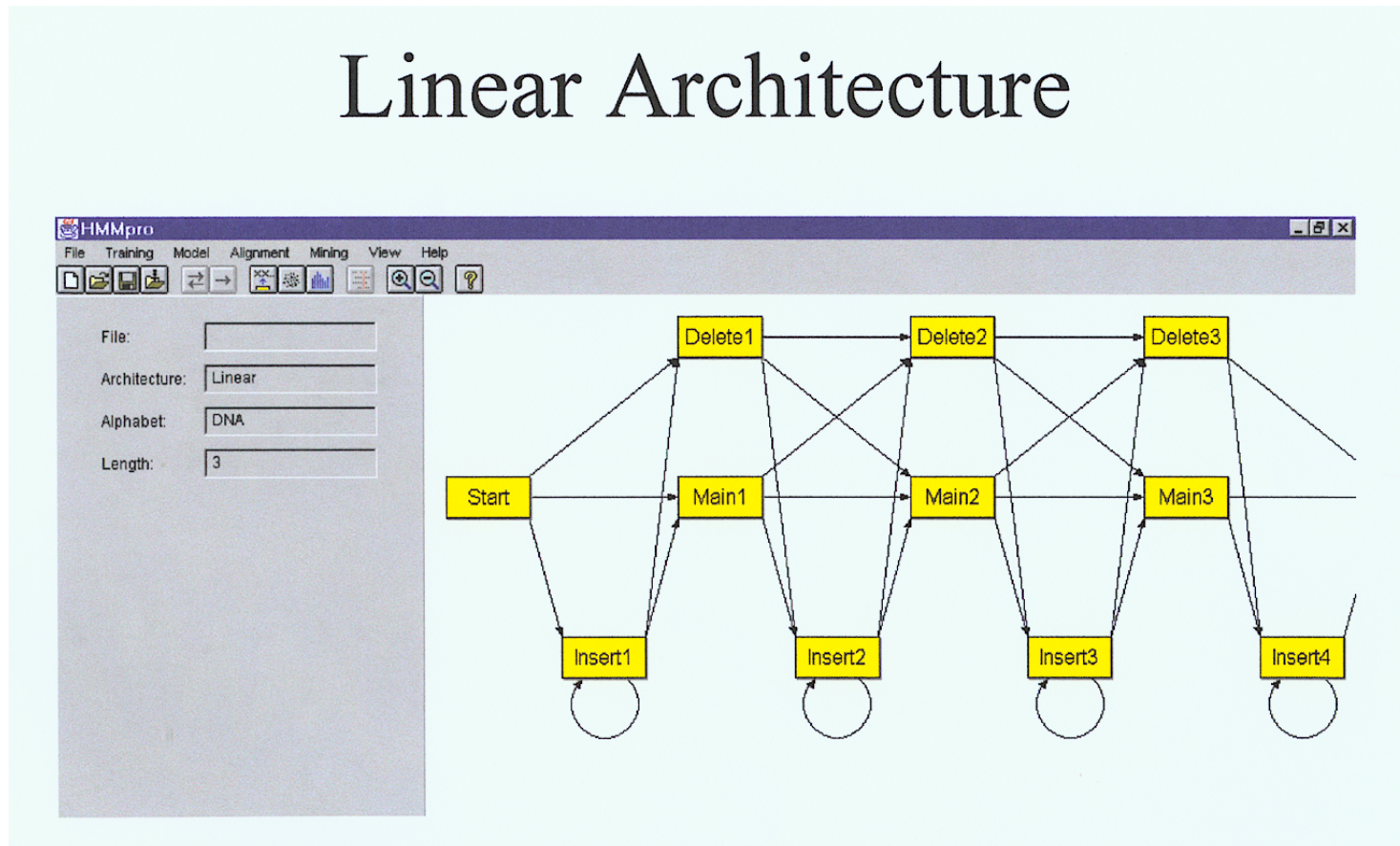


How to model Pairwise Local Alignments with gaps?



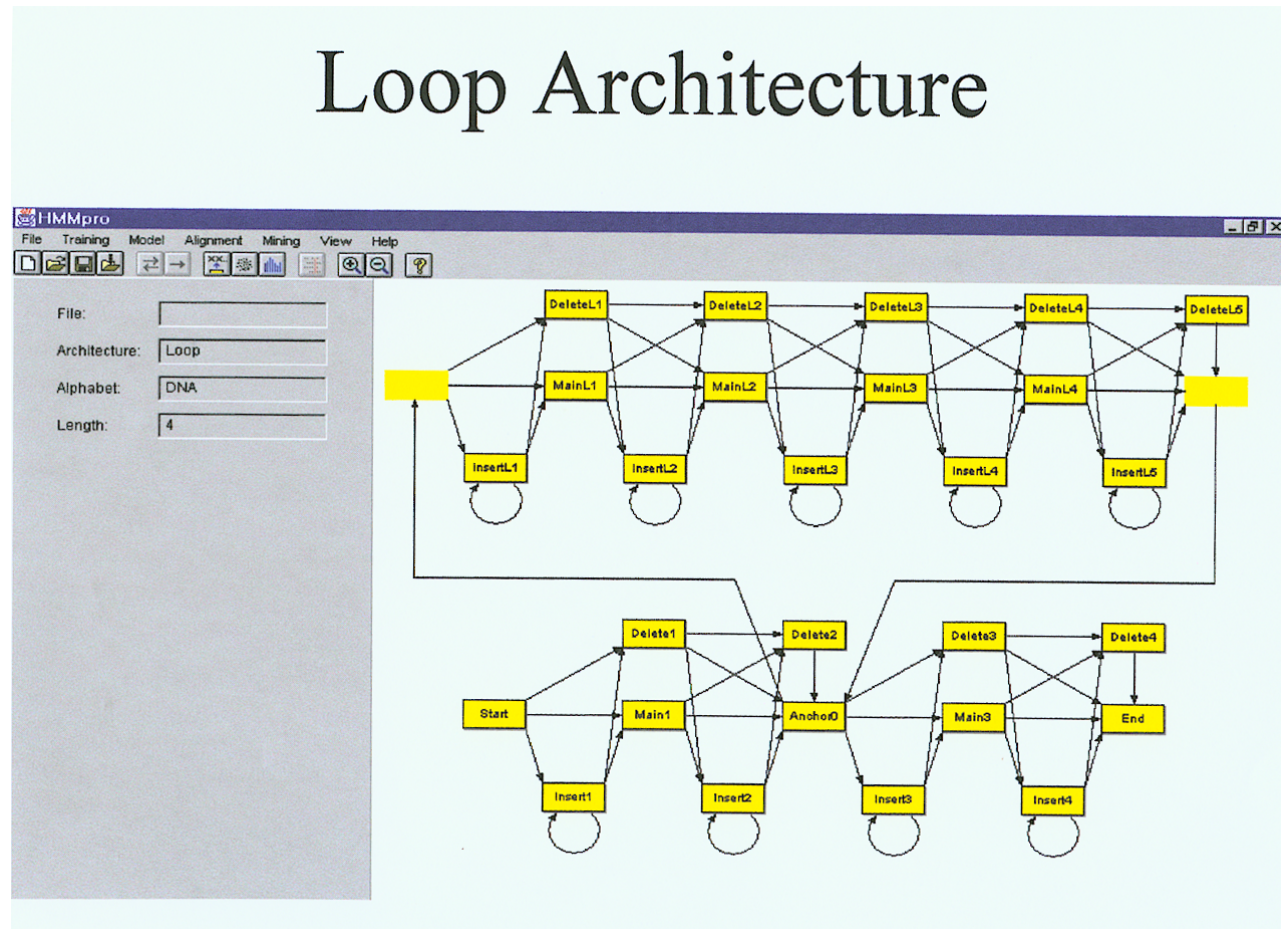
Standard HMM architectures

Linear Architecture



Standard HMM architectures

Loop Architecture



Standard HMM architectures

Wheel Architecture

