

BSC 4934: Q'BIC Capstone Workshop

Giri Narasimhan

ECS 254A; Phone: x3748

giri@cs.fiu.edu

http://www.cs.fiu.edu/~giri/teach/BSC4934_Su10.html

July 2010

Global Alignment: An example

V: G A A T T C A G T T A
W: G G A T C G A

	G	A	A	T	T	C	A	G	T	T	A
G	0										
G	0										
A	0										
T	0										
C	0										
G	0										
A	0										

Given

$\delta[I, J]$ = Score of Matching
the I^{th} character of sequence V &
the J^{th} character of sequence W

Compute

$S[I, J]$ = Score of Matching
First I characters of sequence V &
First J characters of sequence W

Recurrence Relation

$$S[I, J] = \text{MAXIMUM} \left\{ \begin{array}{l} S[I-1, J-1] + \delta(V[I], W[J]), \\ S[I-1, J] + \delta(V[I], -), \\ S[I, J-1] + \delta(-, W[J]) \end{array} \right\}$$

Global Alignment: An example

$$S[I, J] = \text{MAXIMUM} \{ \\ S[I-1, J-1] + \delta(V[I], W[J]), \\ S[I-1, J] + \delta(V[I], -), \\ S[I, J-1] + \delta(-, W[J]) \}$$

V: G A A T T C A G T T A
W: G G A T C G A

	G	A	A	T	T	C	A	G	T	T	A
G	0										
G	0										
A	0										
T	0										
C	0										
G	0										
A	0										

	G	A	A	T	T	C	A	G	T	T	A
G	0	0									
G	0	1									
A	0										
T	0										
C	0										
G	0										
A	0										

	G	A	A	T	T	T	C	G	T	T	A
G	0	1	1	1	1	1	1	1	1	1	1
G	0	1	1	1	1	1	1	1	1	1	1
A	0	1	1	1	1	1	1	1	1	1	1
T	0	1	1	1	1	1	1	1	1	1	1
C	0	1	1	1	1	1	1	1	1	1	1
G	0	1	1	1	1	1	1	1	1	1	1
A	0	1	1	1	1	1	1	1	1	1	1

	G	A	A	T	T	C	A	G	T	T	A
G	0	1	1	1	1	1	1	1	1	1	1
G	0	1	1	1	1	1	1	1	1	1	1
A	0	1	2	2	2	2	2	2	2	2	2
T	0	1	2	2	2	2	2	2	2	2	2
C	0	1	2	2	2	2	2	2	2	2	2
G	0	1	2	2	2	2	2	2	2	2	2
A	0	1	2	2	2	2	2	2	2	2	2

	G	A	A	T	T	C	A	G	T	T	A
G	0	1	1	1	1	1	1	1	1	1	1
G	0	1	1	1	1	1	1	1	1	1	1
A	0	1	2	2	2	2	2	2	2	2	2
T	0	1	2	2	2	2	2	2	2	2	2
C	0	1	2	2	2	2	2	2	2	2	2
G	0	1	2	2	2	2	2	2	2	2	2
A	0	1	2	2	2	2	2	2	2	2	2

	G	A	A	T	T	C	A	G	T	T	A
G	0	1	1	1	1	1	1	1	1	1	1
G	0	1	1	1	1	1	1	1	1	1	1
A	0	1	2	2	2	2	2	2	2	2	3
T	0	1	2	2	3	3	3	3	3	3	3
C	0	1	2	2	3	3	3	4	4	4	4
G	0	1	2	2	3	3	3	4	4	5	5
A	0	1	2	3	3	3	3	4	5	5	6

Traceback

	G	A	A	T	T	C	A	G	T	T	A
G	0	0	0	0	0	0	0	0	0	0	0
G	0	1	1	1	1	1	1	1	1	1	1
G	0	1	1	1	1	1	1	2	2	2	2
A	0	1	1	2	2	2	2	2	2	2	3
T	0	1	2	2	3	3	3	3	3	3	3
C	0	1	2	2	3	3	4	4	4	4	4
G	0	1	2	2	3	3	4	4	5	5	5
A	0	1	2	3	3	3	4	5	5	5	6

	G	A	A	T	T	C	A	G	T	T	A
G	0	0	0	0	0	0	0	0	0	0	0
G	0	1	1	1	1	1	1	1	1	1	1
G	0	1	1	1	1	1	1	2	2	2	2
A	0	1	2	2	2	2	2	2	2	2	2
T	0	1	2	2	3	3	3	3	3	3	3
C	0	1	2	2	3	3	4	4	4	4	4
G	0	1	2	2	3	3	4	4	5	5	5
A											6

	G	A	A	T	T	C	A	G	T	T	A
G	0	0	0	0	0	0	0	0	0	0	0
G	0	1	1	1	1	1	1	1	1	1	1
G	0	1	1	1	1	1	1	1	2	2	2
A	0	1	2	2	2	2	2	2	2	2	2
T	0	1	2	2	3	3	3	3	3	3	3
C	0	1	2	2	3	3	4	4	4	4	4
G	0	1	2	2	3	3	4	4	5	5	5
A											6

	G	A	A	T	T	C	A	G	T	T	A
G	0										
G		1									
A			1								
T				2	2						
C					3						
G						4	4				
A							5	5	5		
A											6

V: G A A T T C A G T T A
 | | | | | | |
 W: G G A - T C - G - - A

Alternative Traceback

	G	A	A	T	T	C	A	G	T	T	A
G	0	0	0	0	0	0	0	0	0	0	0
G	0	1	1	1	1	1	1	1	1	1	1
G	0	1	1	1	1	1	1	2	2	2	2
A	0	1	1	2	2	2	2	2	2	2	3
T	0	1	2	2	3	3	3	3	3	3	3
C	0	1	2	2	3	3	4	4	4	4	4
G	0	1	2	2	3	3	4	4	5	5	5
A	0	1	2	3	3	3	4	5	5	5	6

	G	A	A	T	T	C	A	G	T	T	A
G	0	0	0	0	0	0	0	0	0	0	0
G	0	1	1	1	1	1	1	1	1	1	1
G	0	1	1	1	1	1	1	2	2	2	2
A	0	1	2	2	2	2	2	2	2	2	2
T	0	1	2	2	3	3	3	3	3	3	3
C	0	1	2	2	3	3	4	4	4	4	4
G	0	1	2	2	3	3	4	4	5	5	5
A											

	G	A	A	T	T	C	A	G	T	T	A
G	0	0	0	0	0	0	0	0	0	0	0
G	0	1	1	1	1	1	1	1	1	1	1
G	0	1	1	1	1	1	1	1	2	2	2
A	0	1	2	2	2	2	2	2	2	2	2
T	0	1	2	2	3	3	3	3	3	3	3
C	0	1	2	2	3	3	4	4	4	4	4
G	0	1	2	2	3	3	4	4	5	5	5
A											

	G	A	A	T	T	C	A	G	T	T	A	
G	0											
G		1										
A		1	1									
T			2	2								
C				3								
G					4	4						
A						5	5	5				
											6	

V: G - A A T T C A G T T A
 | | | | |
 W: G G - A - T C - G - - A

V: G A A T T C A G T T A
 | | | | | |
 W: G G A - T C - G - - A

Previous

Improved Traceback

G A A T T C A G T T A

	0	0	0	0	0	0	0	0	0	0	0	0
G	0	x1	←1	←1	←1	←1	←1	←1	x1	←1	←1	←1
G	0	x1	↑1	↑1	↑1	↑1	↑1	↑1	x2	←2	←2	←2
A	0	↑1	↑1	x2	←2	←2	←2	x2	↑2	↑2	↑2	x3
T	0	↑1	←2	↑2	x3	x3	←3	←3	←3	x3	x3	↑3
C	0	↑1	↑2	↑2	↑3	↑3	x4	←4	←4	←4	←4	←4
G	0	↑1	↑2	↑2	↑3	↑3	↑4	↑4	x5	←5	←5	←5
A	0	↑1	↑2	x3	↑3	↑3	↑4	x5	↑5	↑5	↑5	x6

Improved Traceback

G A A T T C A G T T A

	0	0	0	0	0	0	0	0	0	0	0	0
G	0	x1	←1	←1	←1	←1	←1	←1	x1	←1	←1	←1
G	0	x1	↑1	↑1	↑1	↑1	↑1	↑1	x2	←2	←2	←2
A	0	↑1	↑1	x2	←2	←2	←2	x2	↑2	↑2	↑2	x3
T	0	↑1	←2	↑2	x3	x3	←3	←3	←3	x3	x3	↑3
C	0	↑1	↑2	↑2	↑3	↑3	x4	←4	←4	←4	←4	←4
G	0	↑1	↑2	↑2	↑3	↑3	↑4	↑4	x5	←5	←5	←5
A	0	↑1	↑2	x3	↑3	↑3	↑4	x5	↑5	↑5	↑5	x6

Improved Traceback

	G	A	A	T	T	C	A	G	T	T	A
	0	0	0	0	0	0	0	0	0	0	0
G	0	x1	←1	←1	←1	←1	←1	←1	x1	←1	←1
G	0	x1	↑1	↑1	↑1	↑1	↑1	↑1	x2	←2	←2
A	0	↑1	↑1	x2	←2	←2	←2	x2	↑2	↑2	↑2
T	0	↑1	←2	↑2	x3	x3	←3	←3	←3	x3	x3
C	0	↑1	↑2	↑2	↑3	↑3	x4	←4	←4	←4	←4
G	0	↑1	↑2	↑2	↑3	↑3	↑4	↑4	x5	←5	←5
A	0	↑1	↑2	x3	↑3	↑3	↑4	x5	↑5	↑5	↑5

V: G A - A T T C A G T T A

| | | | | | |

W: G - G A - T C - G - - A

Generalizations of Similarity Function

- ❑ Mismatch Penalty = α
- ❑ Spaces (Insertions/Deletions, **InDels**) = β
- ❑ Affine Gap Penalties:
(Gap open, Gap extension) = (γ, δ)
- ❑ Weighted Mismatch = $\Phi(a, b)$
- ❑ Weighted Matches = $\Omega(a)$

Alternative Scoring Schemes

	G	A	A	T	T	C	A	G	T	T	A	
0	0	-2	-3	-4	-5	-6	-7	-8	-9	-10	-11	-12
G	-2	x 1	← -1	← -2	← -3	← -4	← -5	← -6	← -7	← -8	← -9	← -10
G	-3	↑ -1	x -1	← -3	← -4	← -5	← -6	← -7	x -5	← -7	← -8	← -9
A	-4	↑ -2	x 0	x 0	← -2	← -3	← -4	← -5	← -6	← -7	← -8	x -7
T	-5	↑ -3	↑ -2	↑ -2	x 1	← -1	← -2	← -3	← -4	← -5	← -6	← -7
C	-6	↑ -4	↑ -3	↑ -3	↑ -1	x -1	x 0	← -2	← -3	← -4	← -5	← -6
G	-7	↑ -5	↑ -4	↑ -4	↑ -2	↑ -3	↑ -2	x -2	x -1	← -3	← -4	← -5
A	-8	↑ -6	↑ -5	↑ -5	↑ -3	↑ -4	↑ -3	x -1	↑ -3	x -3	x -5	x -3

Match +1
Mismatch -2
Gap (-2, -1)

V: G A A T T C A G T T A
| | | | | |
W: G G A T - C - G - - A

Local Sequence Alignment

- **Example:** comparing long stretches of anonymous DNA; aligning proteins that share only some motifs or domains.
- **Smith-Waterman** Algorithm

Recurrence Relations (Global vs Local Alignments)

□ $S[I, J] = \text{MAXIMUM} \{$
 $S[I-1, J-1] + \delta(V[I], W[J]),$
 $S[I-1, J] + \delta(V[I], \text{—}),$
 $S[I, J-1] + \delta(\text{—}, W[J]) \}$

Global
Alignment

□ $S[I, J] = \text{MAXIMUM} \{ 0,$
 $S[I-1, J-1] + \delta(V[I], W[J]),$
 $S[I-1, J] + \delta(V[I], \text{—}),$
 $S[I, J-1] + \delta(\text{—}, W[J]) \}$

Local
Alignment

Local Alignment: Example

	G	A	A	T	T	C	A	G	T	T	A
G	0	0	0	0	0	0	0	0	0	0	0
G	0	×1	←0	0	0	0	0	0	0	0	0
A	0	0	×2	×1	0	0	0	×1	0	0	×1
T	0	0	↑0	×1	×2	←1	0	0	×1	×1	0
C	0	0	0	0	↑0	×0	×2	0	0	0	0
G	0	0	0	0	0	0	0	×1	0	0	0
A	0	0	×1	×1	0	0	0	×1	0	0	×1

Match +1
Mismatch -1
Gap (-1, -1)

V: - G A A T T C A G T T A
 | | | |
 W: G G - A T - C - G - - A

Properties of Smith-Waterman Algorithm

- How to find all regions of "high similarity"?
 - Find **all** entries above a threshold score and traceback.
- What if: Matches = 1 & Mismatches/spaces = 0?
 - Longest Common Subsequence Problem
- What if: Matches = 1 & Mismatches/spaces = $-\infty$?
 - Longest Common Substring Problem
- What if the average entry is positive?
 - Global Alignment

How to score mismatches?

	A	C	D	E	F	G	H	
A	4	0	-2	-1	-2	0	-2	
C	0	9	-3	-4	-2	-3	-3	
D	-2	-3	6	2	-3	-1	-1	
E	-1	-4	2	5	-3	-2	0	
F	-2	-2	-3	-3	6	-3	-1	
G	0	-3	-1	-2	-3	6	-1	
H	-2	-3	-1	0	-1	-1	6	

BLOSUM 62

BLOSUM n Substitution Matrices

- For each amino acid pair a, b
 - For each BLOCK
 - Align all proteins in the BLOCK
 - Eliminate proteins that are more than $n\%$ identical
 - Count $F(a), F(b), F(a,b)$
 - Compute **Log-odds Ratio**

$$\log\left(\frac{F(a,b)}{F(a)F(b)}\right)$$

Multiple Alignments

- Global
 - ClustalW, ClustalX
 - MSA
 - T-Coffee
- Local
 - BLOCKS
 - eMOTIF
 - GIBBS
 - HMMER
 - MACAW
 - MEME
- Other
 - Profile Analysis from msa (UCSD)
 - SAM HMM (from msa)

Multiple Alignments: CLUSTALW

- * identical
- : conserved substitutions
- . semi-conserved substitutions

```

gi|2213819          CDN-ELKSEAIIEHLCASEFALR-----MKIKEVKKENGDKK 223
gi|12656123        -----ELKSEAIIEHLCASEFALR-----MKIKEVKKENGD-   31
gi|7512442         CKNKNDDDNDIMETLCKNDFALK-----IKVKEITYINRDTK  211
gi|1344282         QDECKFDYVEVYETSSSGAFSLLGRFCGAEPPPHLVSSHHELAVLFRTDH 400
  
```

```

: . : * . . *:* . :*
  
```

- Red: AVFPMLW (Small & hydrophobic)
- Blue: DE (Acidic)
- Magenta: RHK (Basic)
- Green: STYHCNGQ (Hydroxyl, Amine, Basic)
- Gray: Others

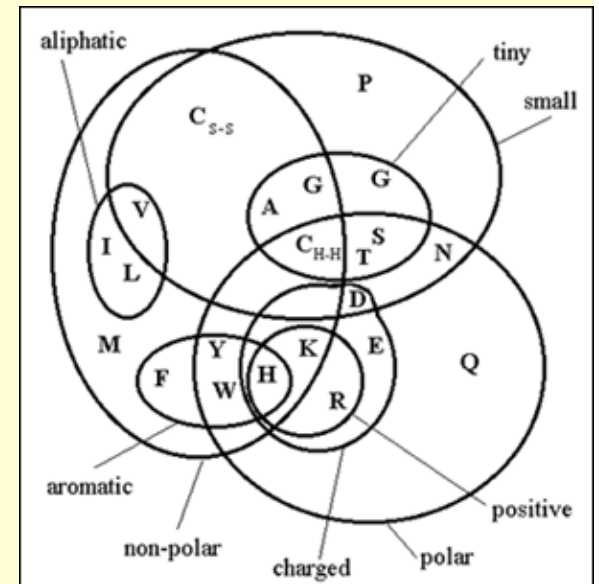


Figure 1. A Venn diagram showing the relationship of the 20 naturally occurring amino acids to a selection of physio-chemical properties thought to be important in the determination of protein structure.

Multiple Alignments

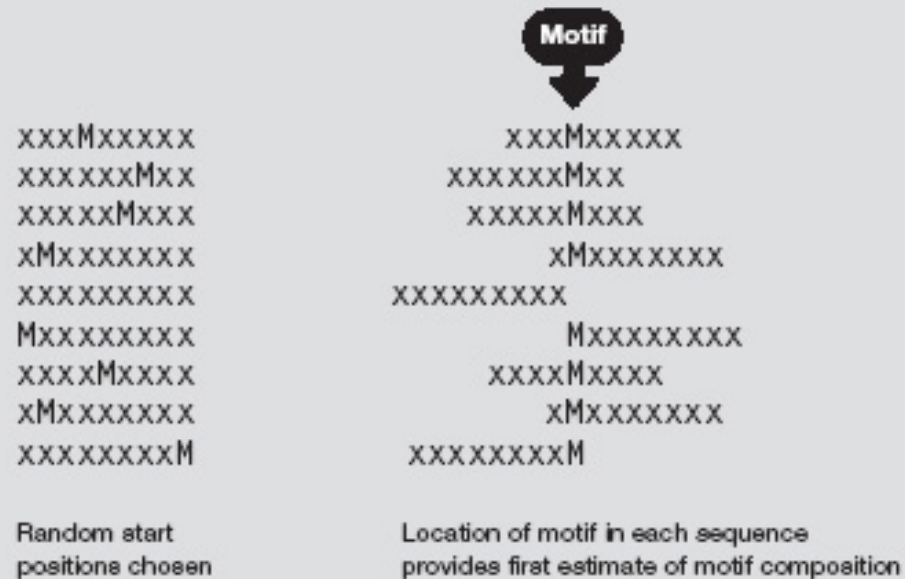
- Family alignment for the ITAM domain (Immunoreceptor tyrosine-based activation motif)

- | | | | |
|----------------|-------------|------------|----|
| CD3D_MOUSE/1-2 | EQLYQPLRDR | EDTQ-YSRLG | GN |
| Q90768/1-21 | DQLYQPLGER | NDGQ-YSQLA | TA |
| CD3G_SHEEP/1-2 | DQLYQPLKER | EDDQ-YSHLR | KK |
| P79951/1-21 | NDLYQPLGQR | SEDT-YSHLN | SR |
| FCEG_CAVPO/1-2 | DGIYTGLSTR | NQET-YETLK | HE |
| CD3Z_HUMAN/3-0 | DGLYQGLSTA | TKDT-YDALH | MQ |
| C79A_BOVIN/1-2 | ENLYEGLNLD | DCSM-YEDIS | RG |
| C79B_MOUSE/1-2 | DHTYEGLNID | QTAT-YEDIV | TL |
| CD3H_MOUSE/1-2 | NQLYNELNLG | RREE-YDVLE | KK |
| CD3Z_SHEEP/1-2 | NPVYNELNVG | RREE-YAVLD | RR |
| CD3E_HUMAN/1-2 | NPDYEPIRKG | QRDL-YSGLN | QR |
| CD3H_MOUSE/2-0 | EGVYNALQKD | KMAEAYSEIG | TK |
| Consensus/60% | - .1YpsLspc | pcsp.YspLs | pp |

Simple
Modular
Architecture
Research
Tool

Multiple Alignment

A. Estimate the amino acid frequencies in the motif columns of all but one sequence. Also obtain background.



How to Score Multiple Alignments?

□ Sum of Pairs Score (SP)

- Optimal alignment: $O(d^N)$ [Dynamic Prog]
- Approximate Algorithm: **Approx Ratio 2**
 - Locate Center: $O(d^2N^2)$
 - Locate Consensus: $O(d^2N^2)$

Consensus char: char with min distance sum

Consensus string: string of consensus char

Center: input string with min distance sum

Multiple Alignment Methods

- Phylogenetic Tree Alignment (NP-Complete)
 - Given tree, task is to label leaves with strings
- Iterative Method(s)
 - Build a MST using the distance function
- Clustering Methods
 - Hierarchical Clustering
 - K-Means Clustering

Multiple Alignment Methods (Cont'd)

□ Gibbs Sampling Method

- Lawrence, Altschul, Boguski, Liu, Neuwald, Winton, *Science*, 1993

□ Hidden Markov Model

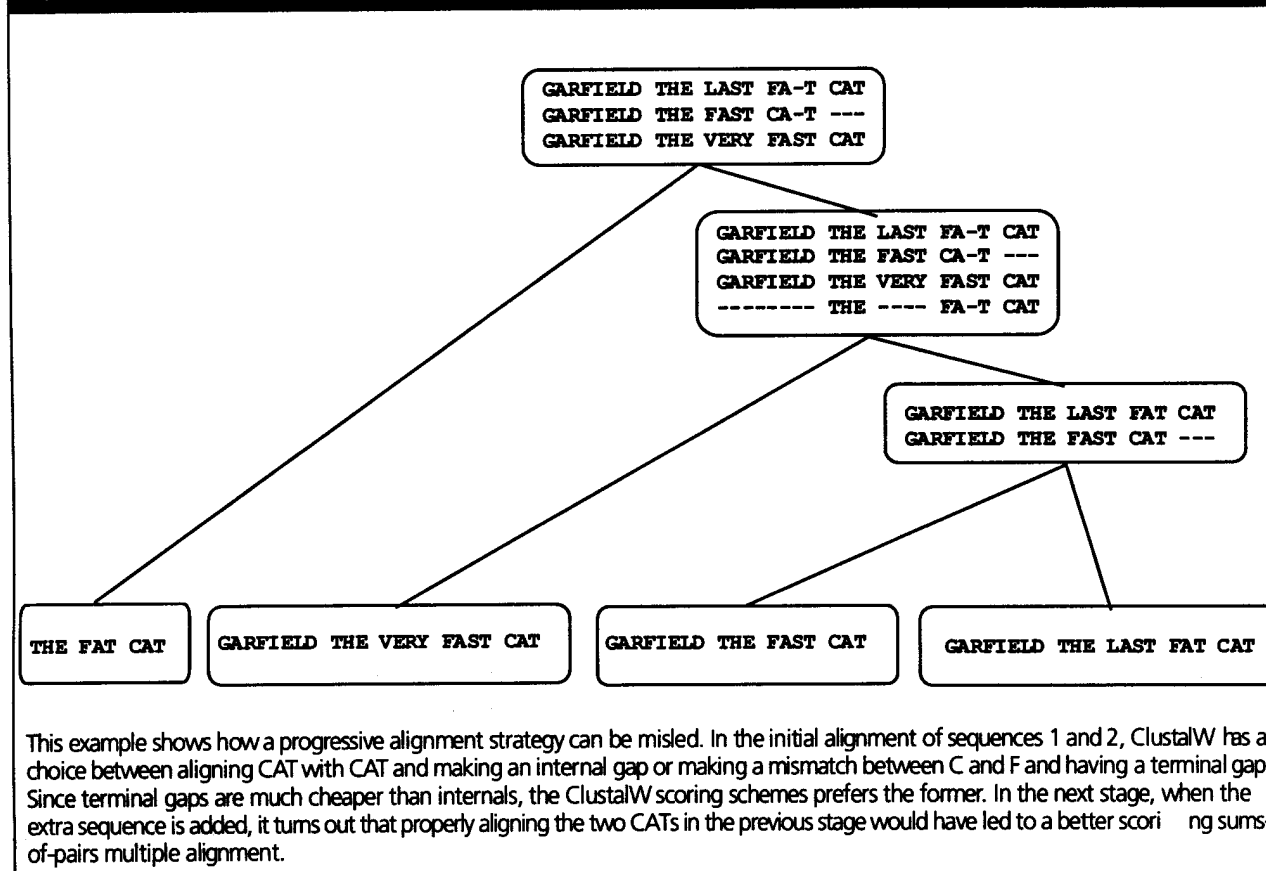
- Krogh, Brown, Mian, Sjolander, Haussler, *JMB*, 1994

Multiple Sequence Alignments (MSA)

- Choice of Scoring Function
 - Global vs local
 - Gap penalties
 - Substitution matrices
 - Incorporating other information
 - Statistical Significance
- Computational Issues
 - Exact/heuristic/approximate algorithms for optimal MSA
 - Progressive/Iterative/DP
 - Iterative: Stochastic/Non-stochastic/Consistency-based
- Evaluating MSAs
 - Choice of good test sets or benchmarks (BALiBASE)
 - How to decide thresholds for good/bad alignments

Progressive MSA: CLUSTALW

Figure 1. Limits of the progressive strategy.



C. Notredame, *Pharmacogenomics*, 3(1), 2002.

Software for MSA

REVIEW

Table 1. Some recent and less recent available methods for MSAs.

Method	Algorithm	URL	Reference
MSA	Exact	http://www.ibc.wustl.edu/ibc/msa.html	[28]
OMA	Iterative DCA	http://bibiserv.techfak.uni-bielefeld.de/oma	[61]
MultAlin	Progressive	http://www.toulouse.inra.fr/multalin.html	[41]
ComAlign	Consistency-based	http://www.daimi.au.dk/~ocaprani	[75]
Praline	Iterative/progressive	jhering@nimr.mrc.ac.uk	[48]
Prrp	Iterative/Stochastic	ftp://ftp.genome.ad.jp/pub/genome/saitama-cc/	[47]
HMMER	Iterative/Stochastic/HMM	http://hmmer.wustl.edu/	[68]
GA	Iterative/Stochastic/GA	czhang@watnow.uwaterloo.ca	[52]

C. Notredame, *Pharmacogenomics*, 3(1), 2002.

MSA: Conclusions

- ❑ Very important
 - Phylogenetic analyses
 - Identify members of a family
 - Protein structure prediction
- ❑ No perfect methods
- ❑ Popular
 - Progressive methods: **CLUSTALW**
 - Recent interesting ones: **Prrp, SAGA, DiAlign, T-Coffee**
- ❑ Review of Methods [C. Notredame, *Pharmacogenomics*, 3(1), 2002]
 - **CLUSTALW** works reasonably well, in general
 - **DiAlign** is better for sequences with long insertions & deletions (indels)
 - **T-Coffee** is best available method

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

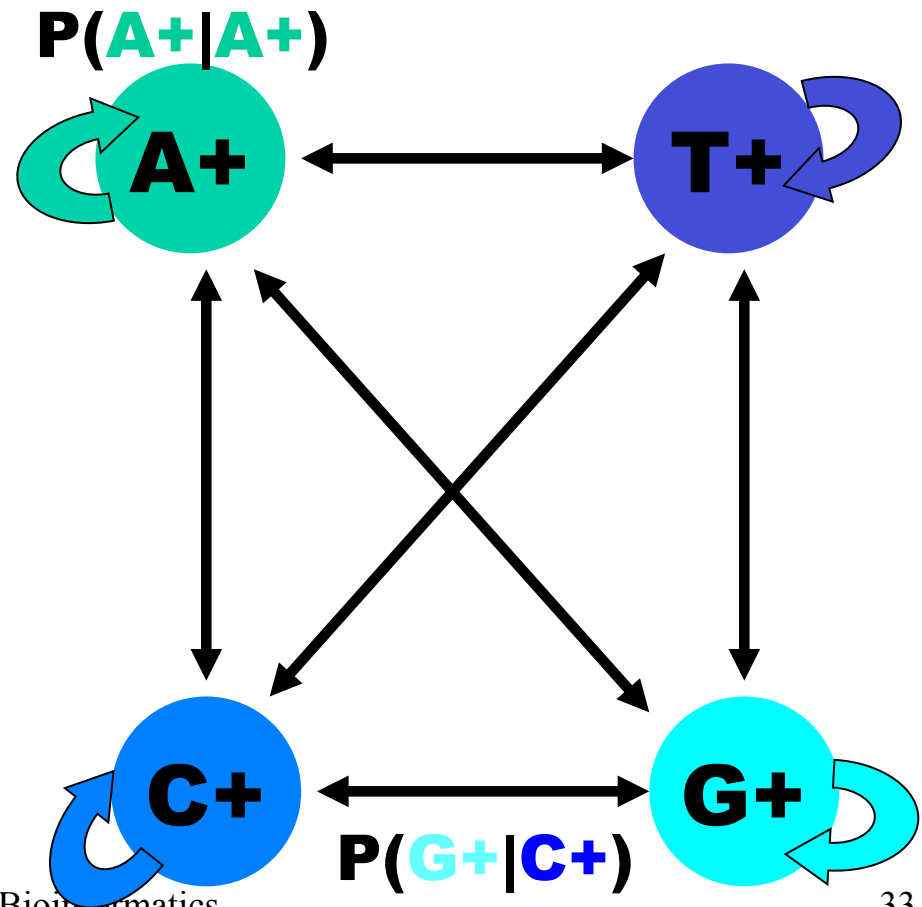
Problem 2:

- **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

7/14/10



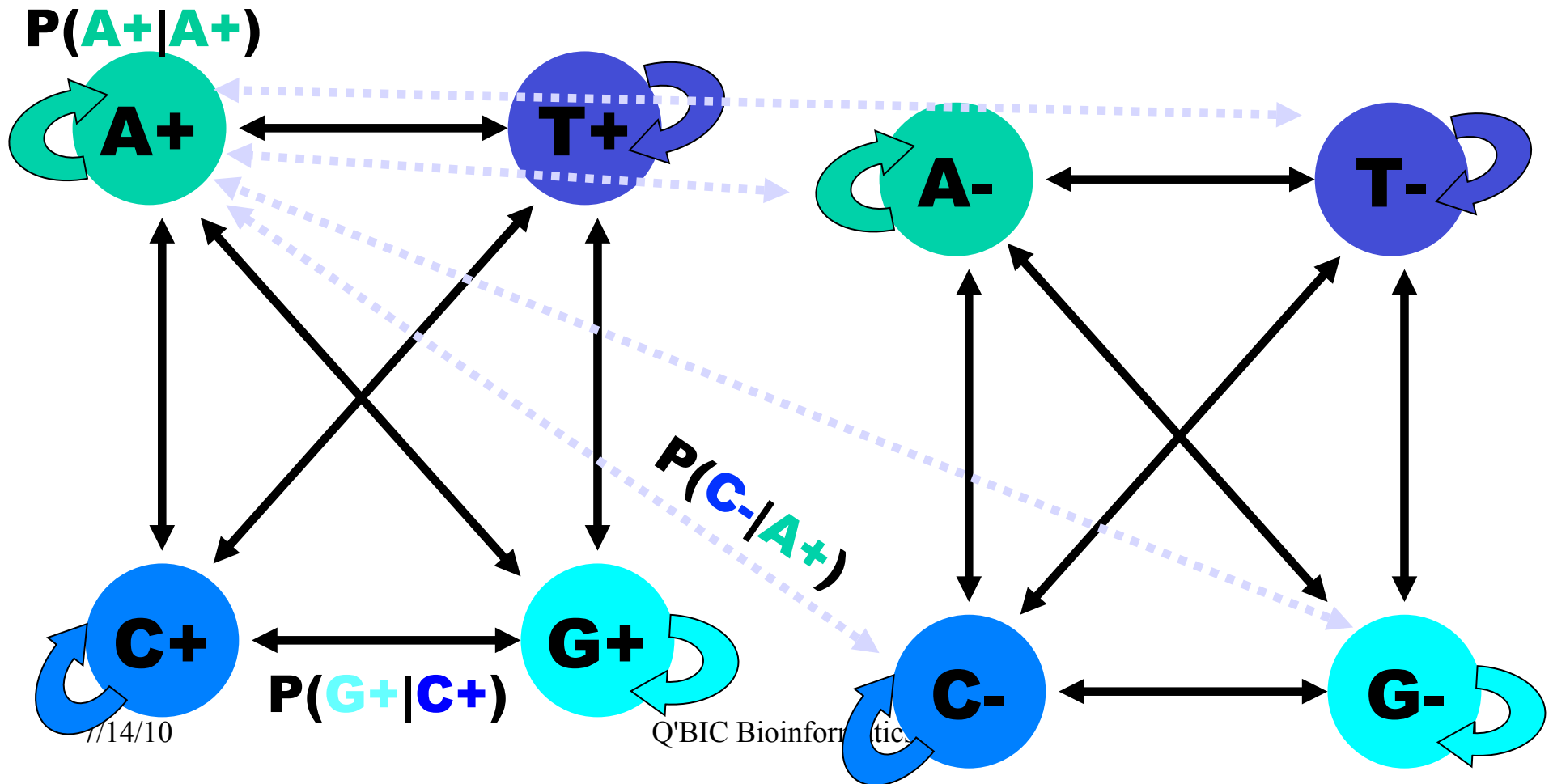
Q'BIC Bioinformatics

33

CpG Island + in an ocean of -

First order Hidden Markov Model

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

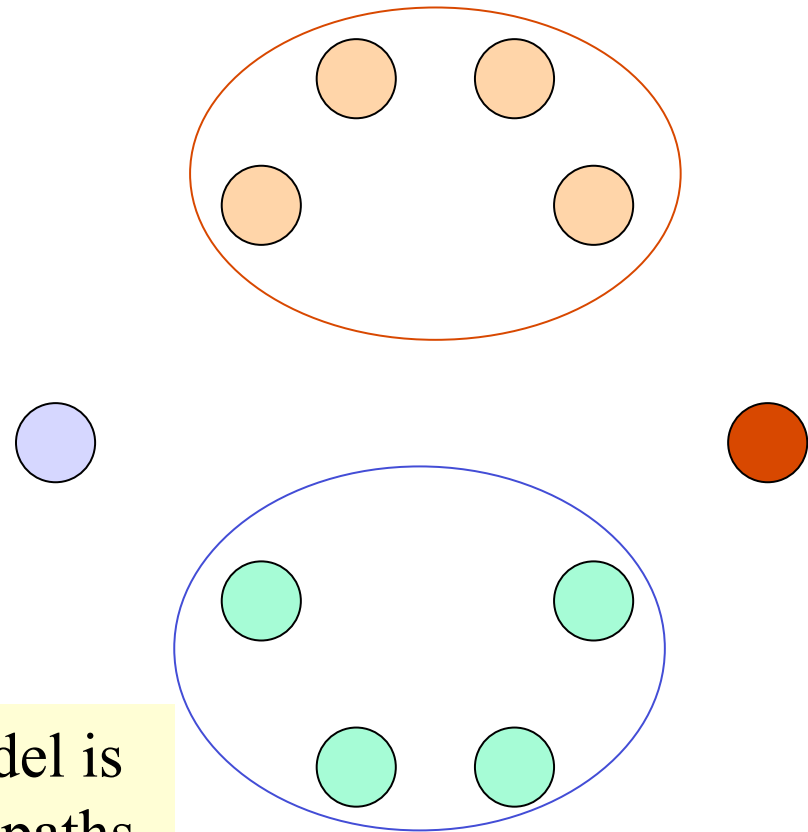


Hidden Markov Model (HMM)

- 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 $P[i,j]$.

Profile Method

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

Location in Seq.	Sequence							Protein Name
	1	2	3	4	5	6	7	
14	G	V	S	A	S	A	V	Ka RbtR
32	G	V	S	E	M	T	I	Ec DeoR
33	G	V	S	P	G	T	I	Ec RpoD
76	G	A	G	I	A	T	I	Ec TrpR
178	G	C	S	R	E	T	V	Ec CAP
205	C	L	S	P	S	R	L	Ec AraC
210	C	L	S	P	S	R	L	St AraC
13	G	V	N	K	E	T	I	Br MerR

FREQUENCY TABLE

	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
1	0	2	0	0	0	6	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2	1	1	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	4	0	0
3	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	6	0	0	0	0
4	1	0	0	1	0	0	0	1	1	0	0	0	3	0	1	0	0	0	0	0
5	1	0	0	2	0	1	0	0	0	0	1	0	0	0	0	3	0	0	0	0
6	1	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	5	0	0	0
7	0	0	0	0	0	0	0	4	0	2	0	0	0	0	0	0	0	2	0	0

7

Profile Method

FREQUENCY TABLE

	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
1	0	2	0	0	0	6	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2	1	1	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	4	0	0
3	0	0	0	0	0	1	0	0	0	0	1	0	0	0	6	0	0	0	0	0
4	1	0	0	1	0	0	0	1	1	0	0	0	3	0	1	0	0	0	0	0
5	1	0	0	2	0	1	0	0	0	0	1	0	0	0	0	3	0	0	0	0
6	1	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	5	0	0	0
7	0	0	0	0	0	0	0	4	0	2	0	0	0	0	0	0	0	2	0	0

WEIGHT MATRIX

	A	C	E	G	I	K	L	M	N	P	R	S
1	0	108	0	101	0	0	0	0	0	0	0	0
2	21	78	0	0	0	0	44	0	0	0	0	0
3	0	0	0	23	0	0	0	0	46	0	0	102
4	21	0	32	0	38	32	0	0	0	86	39	0
5	21	0	62	23	0	0	0	74	0	0	0	72
6	21	0	0	0	0	0	0	0	0	0	69	0
7	0	0	0	0	98	0	44	0	0	0	0	0

$$Weight[i, AA] = \log \left(\frac{Freq[i, AA]}{p[AA] \cdot N} \right) \cdot 100$$

8

Profile Method

WEIGHT MATRIX

	A	C	E	G	I	K	L	M	N	P	R	S
1	0	108	0	101	0	0	0	0	0	0	0	0
2	21	78	0	0	0	0	44	0	0	0	0	0
3	0	0	0	23	0	0	0	0	46	0	0	102
4	21	0	32	0	38	32	0	0	0	86	39	0
5	21	0	62	23	0	0	0	74	0	0	0	72
6	21	0	0	0	0	0	0	0	0	0	69	0
7	0	0	0	0	98	0	44	0	0	0	0	0

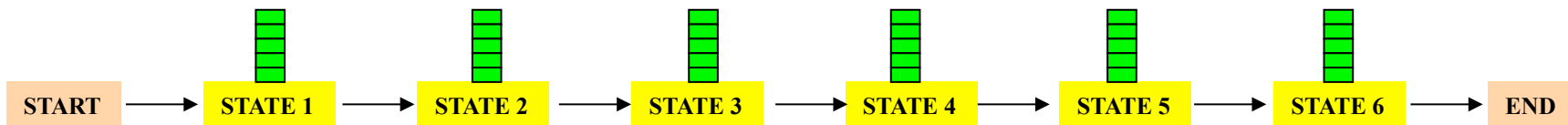
Given the following protein sequence:

```
M T E D L F G D L Q D D T I L A H L D N
P A E D T S R F P A L L A E L N D L L R
G E L S R L G V D P A H S L E I V V A I
C K H L G G G Q V Y I P R G Q A L D S L
I R D L R I W N D F N G R N V S E L T T
R Y G V T F N T V Y K A I R R M R R L K
```

Profile HMMs

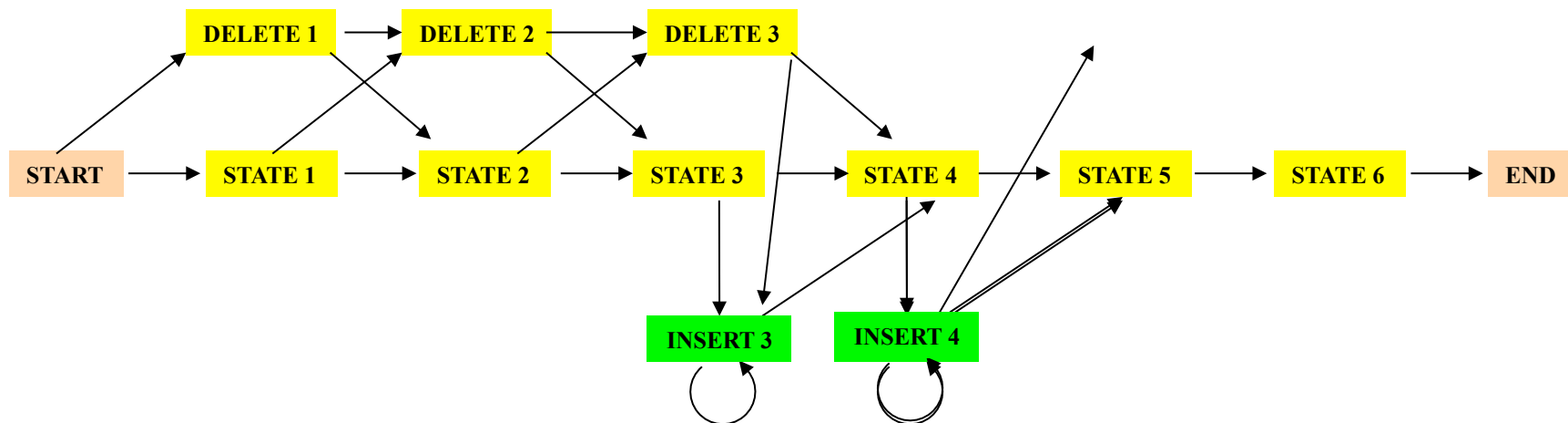
PROFILE METHOD, [M. Gribskov et al., '90]

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

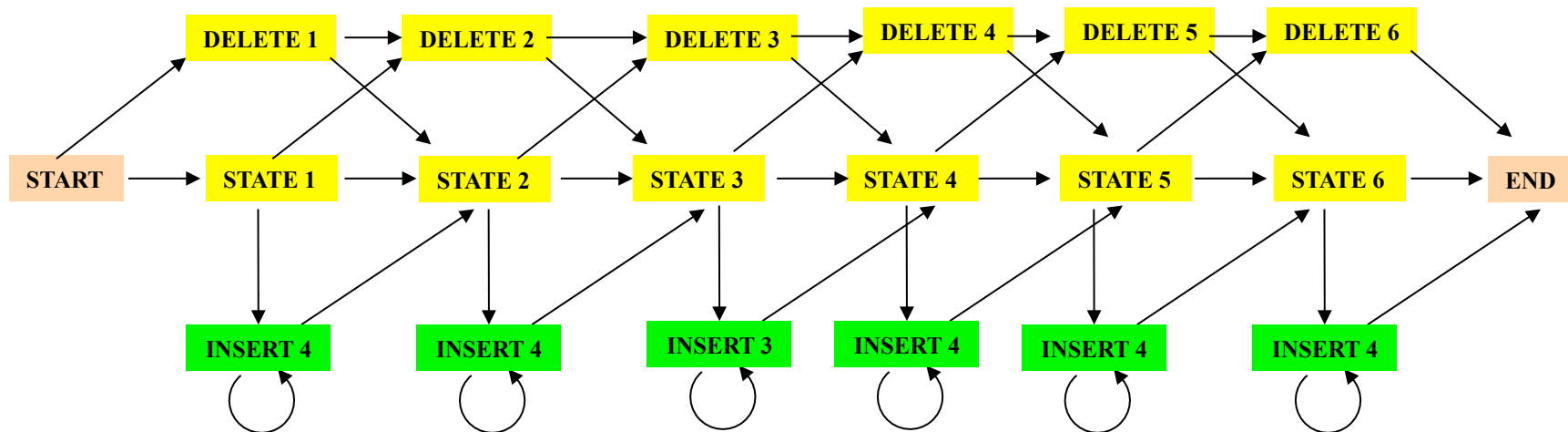


Profile HMMs with InDels

- Insertions
- Deletions
- Insertions & Deletions



Profile HMMs with InDels

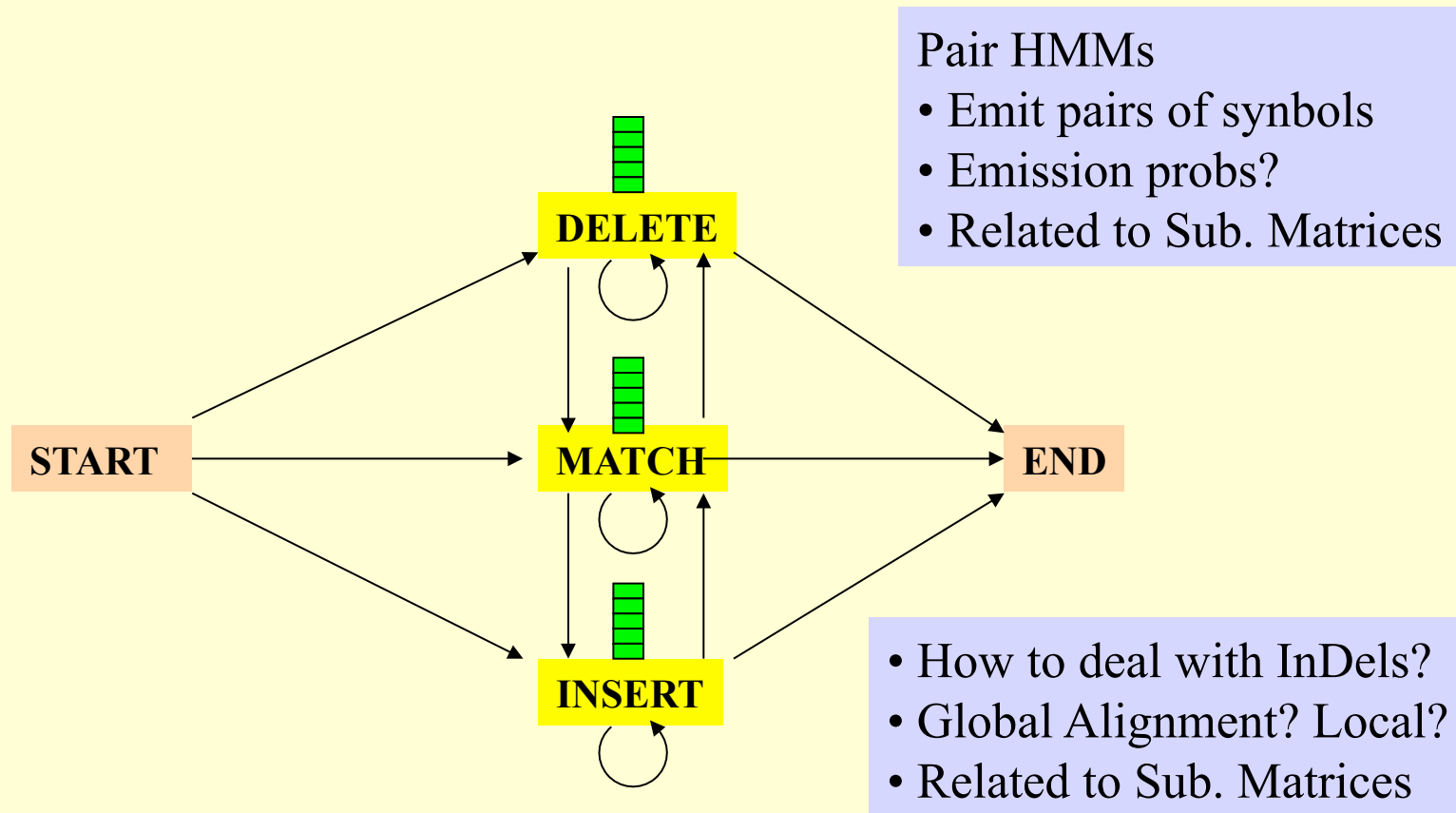


Missing transitions from **DELETE j** to **INSERT j** and
from **INSERT j** to **DELETE $j+1$** .

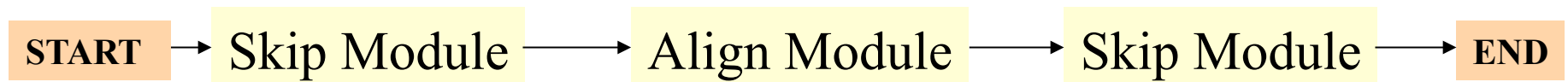
How to model Pairwise Sequence Alignment

LEAPVE

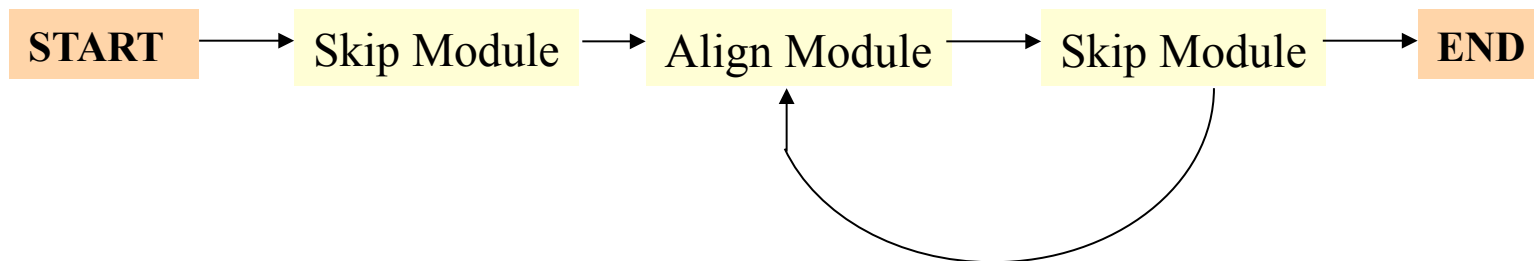
LAPVIE



How to model Pairwise Local Alignments?

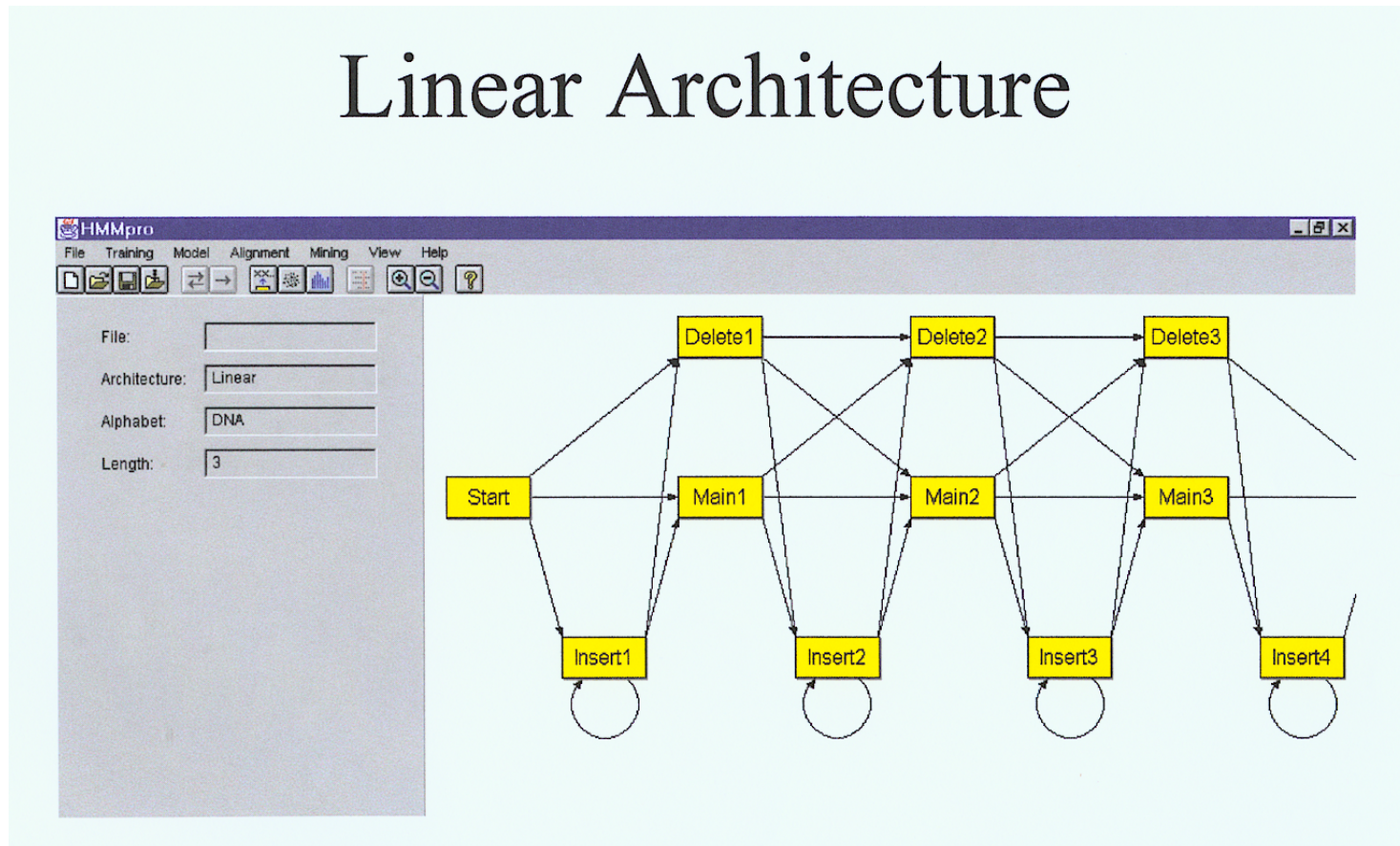


How to model Pairwise Local Alignments with gaps?



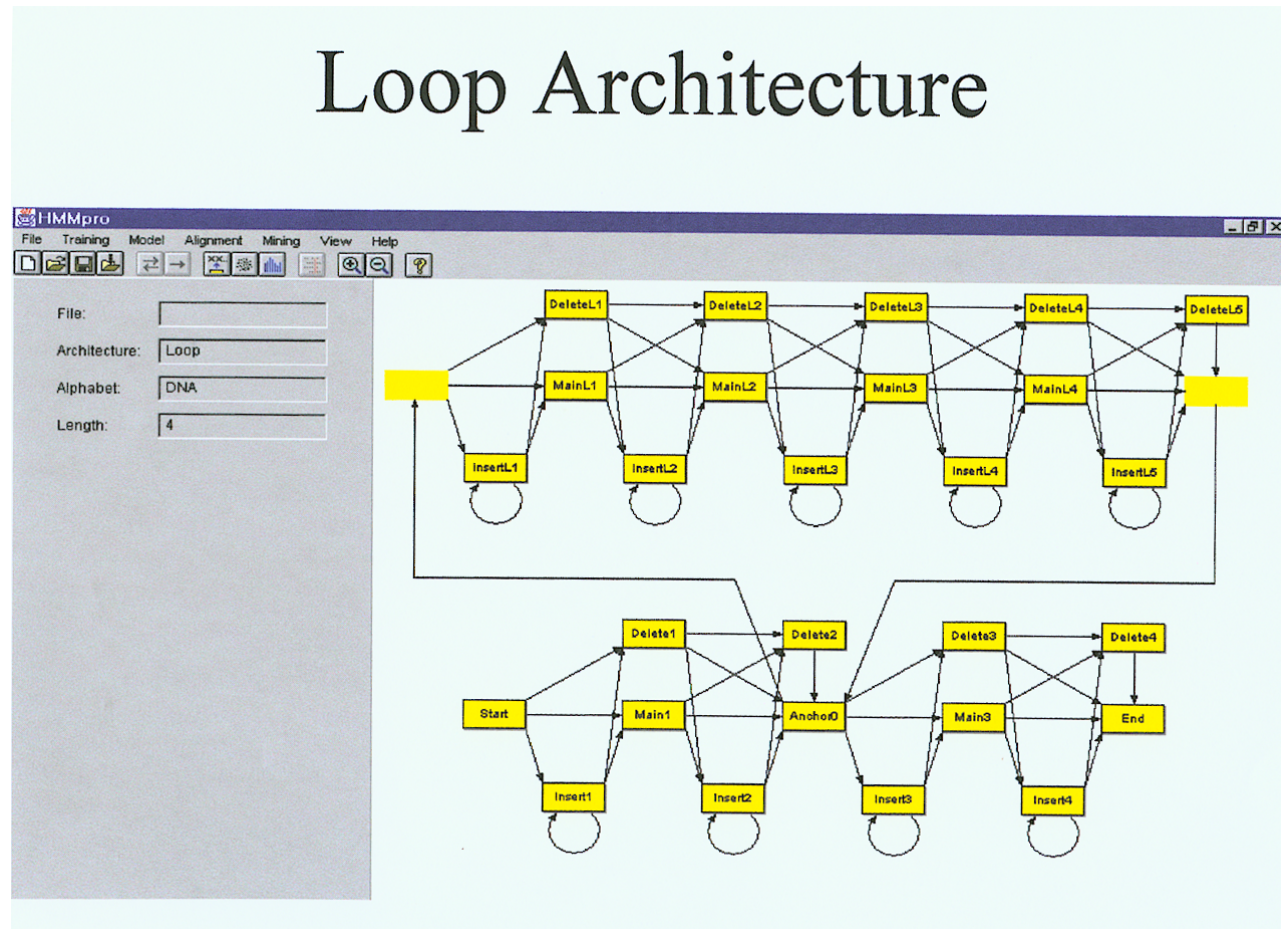
Standard HMM architectures

Linear Architecture



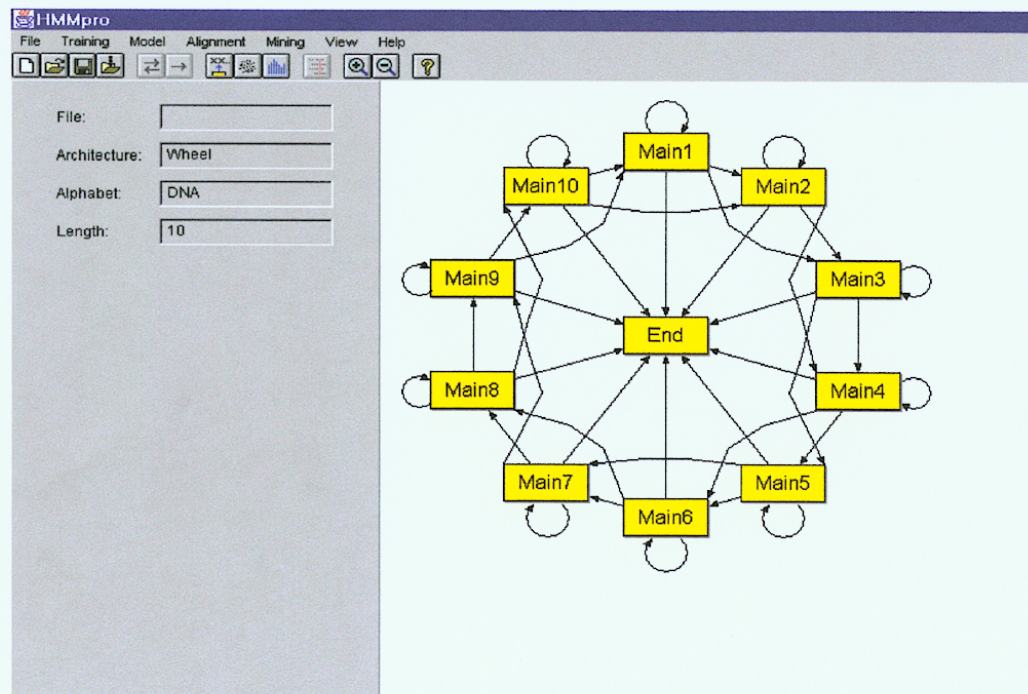
Standard HMM architectures

Loop Architecture



Standard HMM architectures

Wheel Architecture



Profile HMMs from Multiple Alignments

```
HBA_HUMAN   VGA--HAGEY
HBB_HUMAN   V----NVDEV
MYG_PHYCA   VEA--DVAGH
GLB3_CHITP  VKG-----D
GLB5_PETMA  VYS--TYETS
LGB2_LUPLU  FNA--NIPKH
GLB1_GLYDI  IAGADNGAGV
```

Construct Profile HMM from above multiple alignment.

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

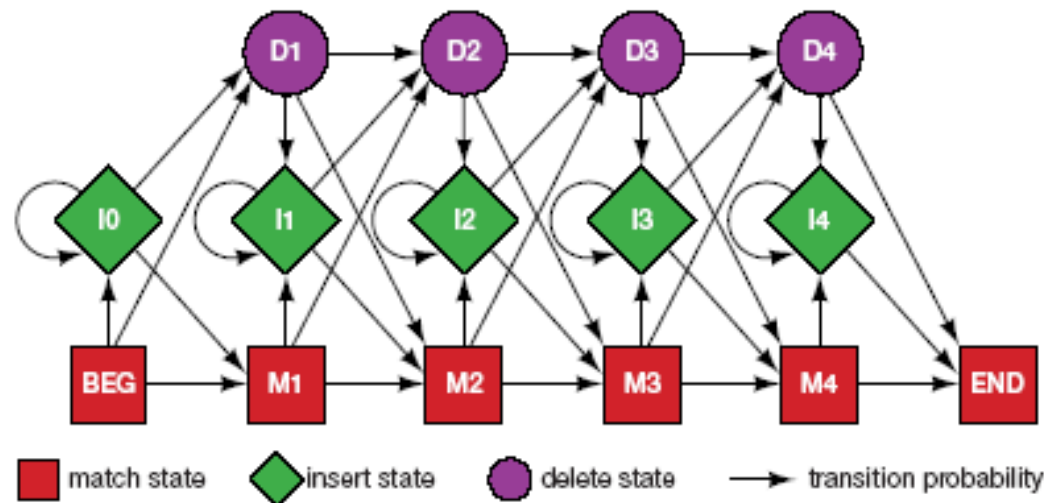


FIGURE 5.16. Relationship between the sequence alignment and the hidden Markov model of the alignment (Krogh et al. 1994). This particular form for the HMM was chosen to represent the sequence, structural, and functional variation expected in proteins. The model accommodates the identities, mismatches, 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* positions), 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 any one of many pathways from one type of sequence variation to another (states) along the state transition arrows and terminating in the ending state labeled END. Any sequence can be generated by the model and each pathway has a probability associated with it. Each square match state stores an amino acid distribution such that the probability of finding an amino acid depends on

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 Θ
- **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 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)

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.