In the Name of God, the Merciful, the Compassionate

Introduction to Bioinformatics
07 - Profiles and Hidden Markov Models

Instructor: Hossein Zeinali

Amirkabir University of Technology

Statistical Modeling

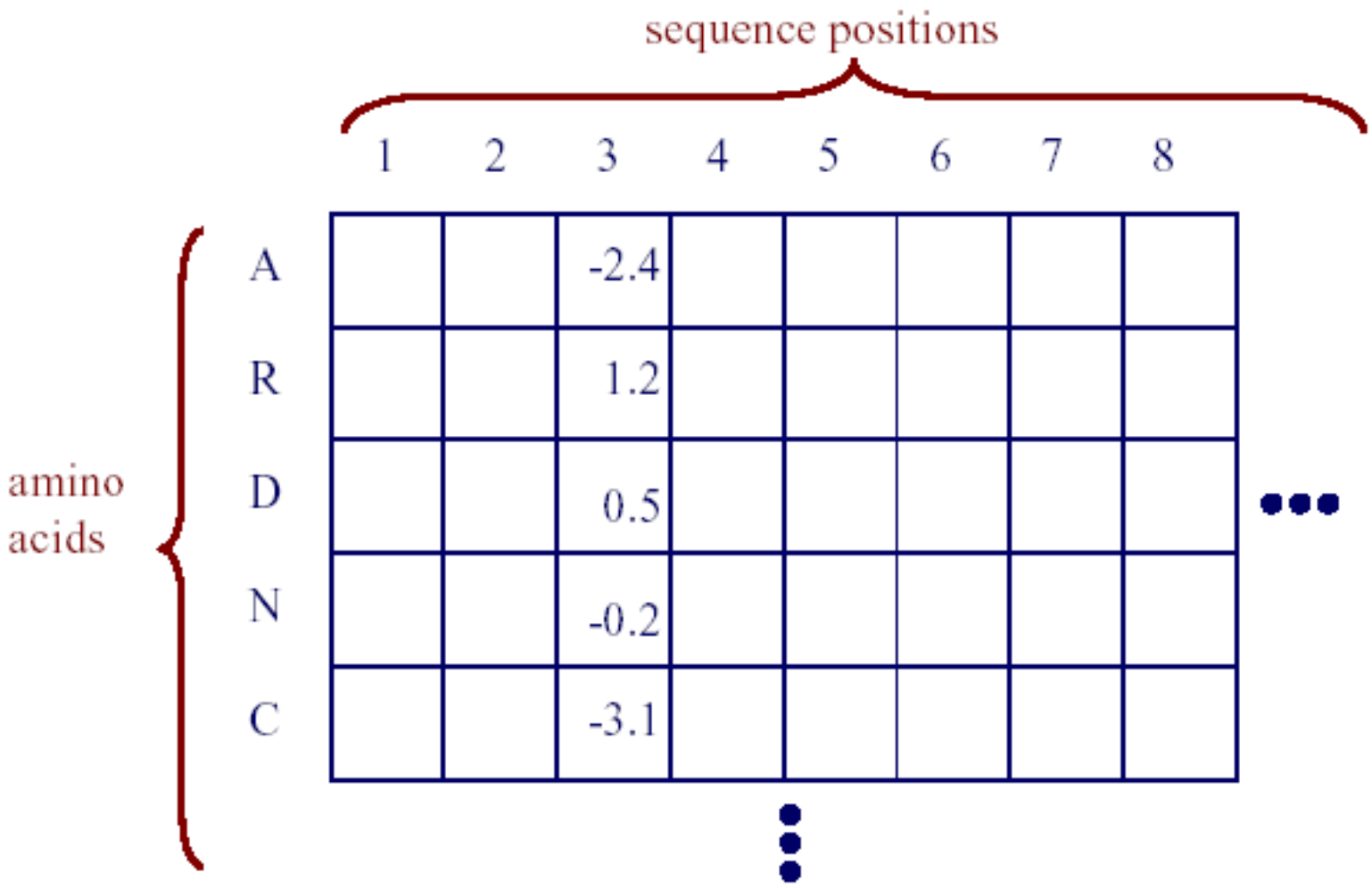
- Construction of statistical models using multiple sequence alignments (MSA):
  - Position-Specific Scoring Matrix (PSSM)
  - Profiles as a more general PSSM model
  - Hidden Markov Model (HMM)
- Statistical models reflect the frequency information of amino acid or nucleotide residues in an MSA.
  - They can be treated as a consensus for a given sequence family.
  - They can be used as a single sequence for database searching and alignment.
  - They can capture the observed frequencies and also predict frequencies of unobserved characters.

Position-Specific Scoring Matrix (PSSM)

- A PSSM is defined as a *table* that contains *probability information* of amino acids or nucleotides at each position of an *ungapped multiple sequence alignment*.
- In such a table:
  - The rows represent residue positions of a particular MSA
  - The columns represent the names of residues (or vice versa)
  - The values in the table represent log-odds scores of the residues calculated from the MSA.
  - A positive score represents identical residue or similar residue match; a negative score represents a non-conserved residue match.

Position-Specific Scoring Matrix (PSSM)

- A PSS … *nation* of am … *d* *multip*
- In suc
  - The
  - The
  - The … lculated fron
  - A p … ch; a neg



PSSM Entries = Log-Odds Scores

1. **Estimate probability of observing each residue in a particular position** *(probability of A given M, where M is PSSM model)*
2. **Divide by background probability of observing each residue** *(probability of A given B, where B is background model.)*
   - In a simple case, random chance can be considered as the background probability.
   - It can be estimated using sequences in the corresponding MSA.
   - In the best case, can be estimated using training data.
3. **Take log** so that can add (rather than multiply) scores

Observed frequency of residue "A"

Foreground model (i.e., the PSSM)

$$\log_2\left(\frac{\Pr(A|M)}{\Pr(A|B)}\right)$$

Background model

PSSM Example of Nucleotide Sequences

Position	1 2 3 4 5 6
Sequence 1	ATGTCG
Sequence 2	AAGACT
Sequence 3	TACTCA
Sequence 4	CGGAGG
Sequence 5	AACCTG

We will see how should deal with not seen residues.

1 Convert multiple alignment to a raw frequency table

Pos.	1	2	3	4	5	6	Overall freq.
A	0.6	0.6	—	0.4	—	0.2	0.30
T	0.2	0.2	—	0.4	0.2	0.2	0.20
G	—	0.2	0.6	—	0.2	0.6	0.27
C	0.2	—	0.4	0.2	0.6	—	0.23

2 Normalize the values by dividing them by overall freq.

Pos.	1	2	3	4	5	6	Overall freq.
A	2.0	2.0	—	1.33	—	0.67	0.30
T	1.0	1.0	—	2.0	1.0	1.0	0.20
G	—	0.74	2.22	—	0.74	2.22	0.27
C	0.87	—	1.74	0.87	2.61	—	0.23

3 Convert the values to log to base of 2

Pos.	1	2	3	4	5	6
A	1.0	1.0	—	0.41	—	-0.58
T	0.0	0.0	—	1.0	0.0	0.0
G	—	-0.43	1.15	—	-0.43	1.15
C	-0.2	—	0.8	-0.2	1.38	—

Inference from PSSN

- How well does the new sequence AACTCG fit into the matrix?
- First, we should find nucleotides at respective positions of the matrix.

Pos.	1	2	3	4	5	6
A	1.0	1.0	—	0.41	—	-0.58
T	0.0	0.0	—	1.0	0.0	0.0
G	—	-0.43	1.15	—	-0.43	1.15
C	-0.2	—	0.8	-0.2	1.38	—

- Then calculate the sum of log odds scores:

$$1.0 + 1.0 + 0.8 + 1.0 + 1.38 + 1.15 = 6.33$$

- The value can be interpreted as the probability of the sequence fitting the matrix as $2^{6.33}$, or 80 times more likely than by random chance.

Second Example

(A)

	12345
S1	**GCTCC**
S2	**AATCG**
S3	**TACGC**
S4	**GTGTT**
S5	**GTAAA**
S6	**CGTCC**

	1	2	3	4	5	Overall
A	.17	.33	.17	.17	.17	6/30 = .20
C	.17	.17	.17	.50	.50	9/30 = .30
G	.50	.17	.17	.17	.17	7/30 = .23
T	.17	.33	.50	.17	.17	8/30 = .27

Normalize by dividing by overall frequencies

	1	2	3	4	5	Overall
A	.85	1.65	.85	.85	.85	6/30 = .20
C	.57	.57	.57	1.67	1.67	9/30 = .30
G	2.17	.74	.74	.74	.74	7/30 = .23
T	.63	1.22	1.85	.63	.63	8/30 = .27

Convert to log to base of 2

	1	2	3	4	5
A	-0.23	0.72	-0.23	-0.23	-0.23
C	-0.81	-0.81	-0.81	0.74	0.74
G	1.11	-0.43	-0.43	-0.43	-0.43
T	-0.66	0.29	0.89	-0.66	-0.66

(B)

Match **GATCA** to PSSM

Find nucleotides at corresponding positions

	1	2	3	4	5
A	-0.23	0.72	-0.23	-0.23	-0.23
C	-0.81	-0.81	-0.81	0.74	0.74
G	1.11	-0.43	-0.43	-0.43	-0.43
T	-0.66	0.29	0.89	-0.66	-0.66

Sum corresponding log odds matrix scores

Score = 1.11 + 0.72 + 0.89 + 0.74 - 0.23 = **3.23**

PSSM Example of Amino Acid Sequences

NTEGEWI
NITRGEW
NIAGECC

Amino acid frequencies at every position of the alignment:

How we should deal with zeros?

Amino Acid	1	2	3	4	5	6	7
N	1	0	0	0	0	0	0
T	0	0.33	0.33	0	0	0	0
E	0	0	0.33	0	0.66	0.33	0
G	0	0	0	0.66	0.33	0	0
W	0	0	0	0	0	0.33	0.33
I	0	0.66	0	0	0	0	0.33
H	0	0	0	0	0	0	0
R	0	0	0	0.33	0	0	0
A	0	0	0.33	0	0	0	0
C	0	0	0	0	0	0.33	0.33
...	...	...	...	...	...	...	...

PSSM Example (Cont.)

- In order to model every possible sequence:
  - Amino acids that *do not appear* at a specific position of a multiple alignment must also be considered. Log(0) = negative infinity which mean impossible sequence!
- Pseudo-counts procedure:
  - Assign minimal scores to residues that do not appear at a certain position

$$Score(x) = \frac{Frequency + Pseudocount}{N + B \times Pseudocount}$$

- Where:
  - *Frequency* is the frequency of residue *i* in column *j* (the count of occurrences).
  - *Pseudocount* is a number higher or equal to 1.
  - *N* is the number of sequences in the MSA.
  - *B* is the total number of allocated pseudocounts in each position (i.e. 20 for all amino acids)

PSSM Example (Cont.)

- In our example, N = 3 and let's use pseudocount = 1:
  - Score(N) at position 1 = 3/3 = **1**.
  - Score(I) at position 1 = 0/3 = **0**.
- Readjust by pseudocount:
  - Score(N) at position 1 -> (**3**+1) / (3+20) = 4/23 = **0.174**.
  - Score(I) at position 1 -> (**0**+1) / (3+20) = 1/23 = **0.044**.

The PSSM is obtained by taking the logarithm of (the values obtained above divided by the background frequency of the residues).

- To simplify for this example we'll assume that every amino acid appears equally in protein sequences, i.e. f_i = 0.05 for every i):
  - PSSM Score(N) at position 1 = log(**0.174** / 0.05) = **0.541**.
  - PSSM Score(I) at position 1 = log(**0.044** / 0.05) = **-0.061**.

PSSM Example (Cont.)

Amino acid	1	2	3	4	5	6	7
N	0.541	-0.061	-0.061	-0.061	-0.061	-0.061	-0.061
T	-0.061	0.240	0.240	-0.061	-0.061	-0.061	-0.061
E	-0.061	-0.061	0.240	-0.061	0.416	0.240	-0.061
G	-0.061	-0.061	-0.061	0.416	0.240	-0.061	-0.061
W	-0.061	-0.061	-0.061	-0.061	-0.061	0.240	0.240
I	-0.061	0.416	-0.061	-0.061	-0.061	-0.061	0.240
H	-0.061	-0.061	-0.061	-0.061	-0.061	-0.061	-0.061
R	-0.061	-0.061	-0.061	0.240	-0.061	-0.061	-0.061
A	-0.061	-0.061	0.240	-0.061	-0.061	-0.061	-0.061
C	-0.061	-0.061	-0.061	-0.061	-0.061	0.240	0.240
...	...	...	...	...	...	...	...

Profiles

- A profile is a PSSM with penalty information regarding insertions and deletions for a sequence family.
  - Can be considered as a generalized form of MSSM
  - *profile* is often used interchangeably with PSSM

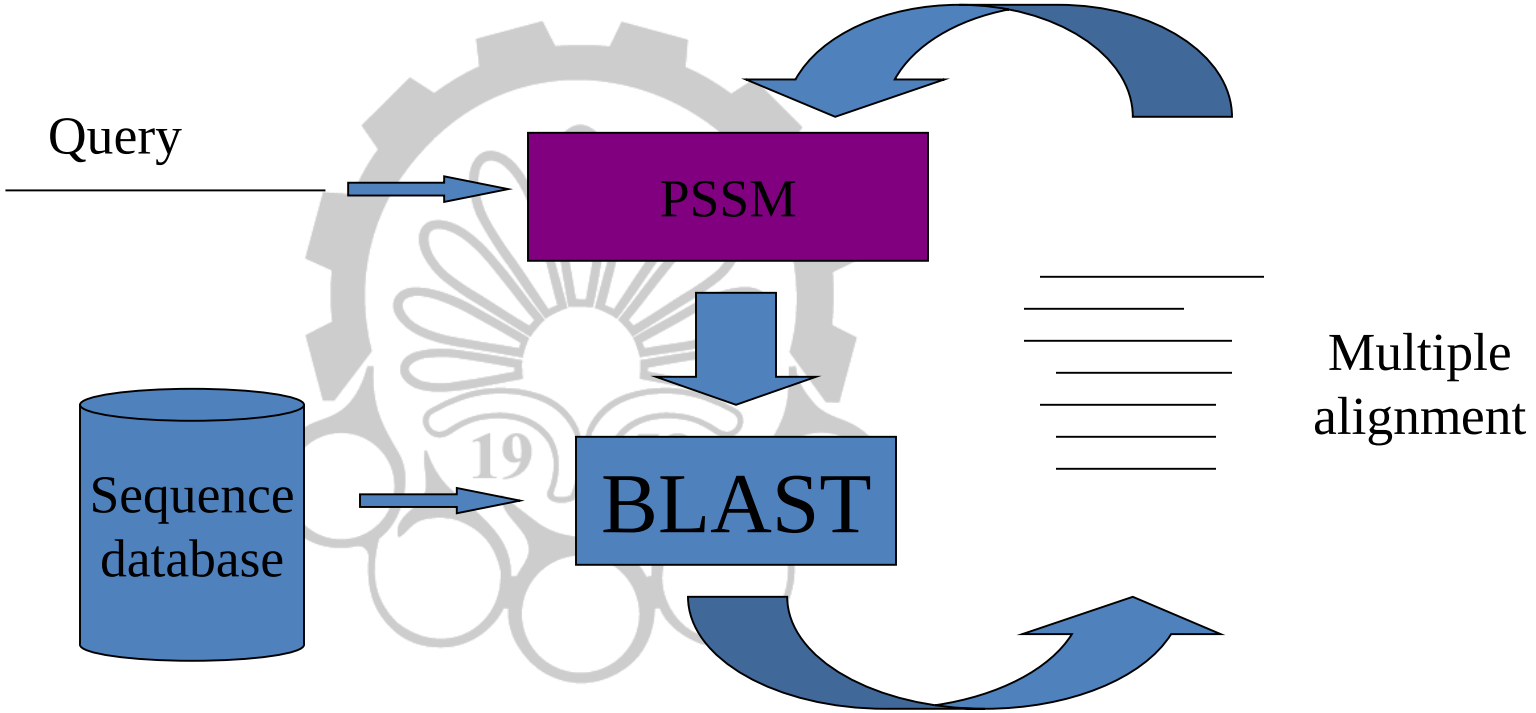
```
     12345
S1   KLM-K
S2   KLKLK
S3   KMML-
S4   ML-LM
```

	1	2	3	4	5	Overall
K	.75		.25		.50	6/20
L		.75		.75		6/20
M	.25	.25	.50		.25	5/20
-			.25	.25	.25	3/20

PSI-BLAST

- ***P*osition *S*pecific *I*terated BLAST**
- **Intuition**: substitution matrices should be specific to a particular position
- **Basic idea:**
  - Use BLAST with high stringency to get a set of closely related sequences
  - Align those sequences to create a new substitution matrix for each position
  - Then use that matrix (iteratively) to find additional sequences

PSI-BLAST

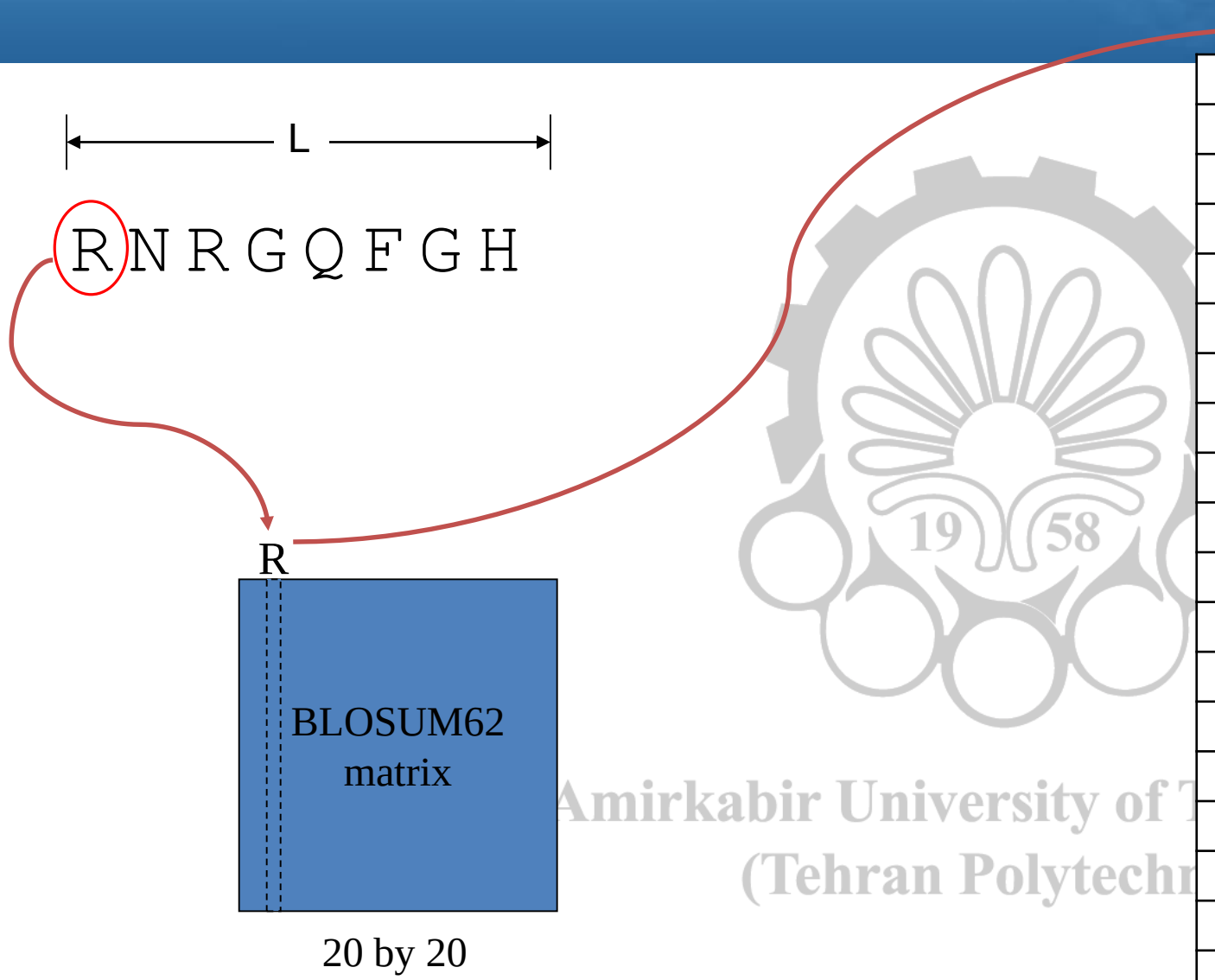


PSI-BLAST pseudocode

```
Convert query to PSSM
do {
   BLAST database with PSSM
   Stop if no new homologs are found
   Add new homologs to PSSM
}
Print current set of homologs
```

This step requires a user-defined threshold

Creating a PSSM from 1 sequence

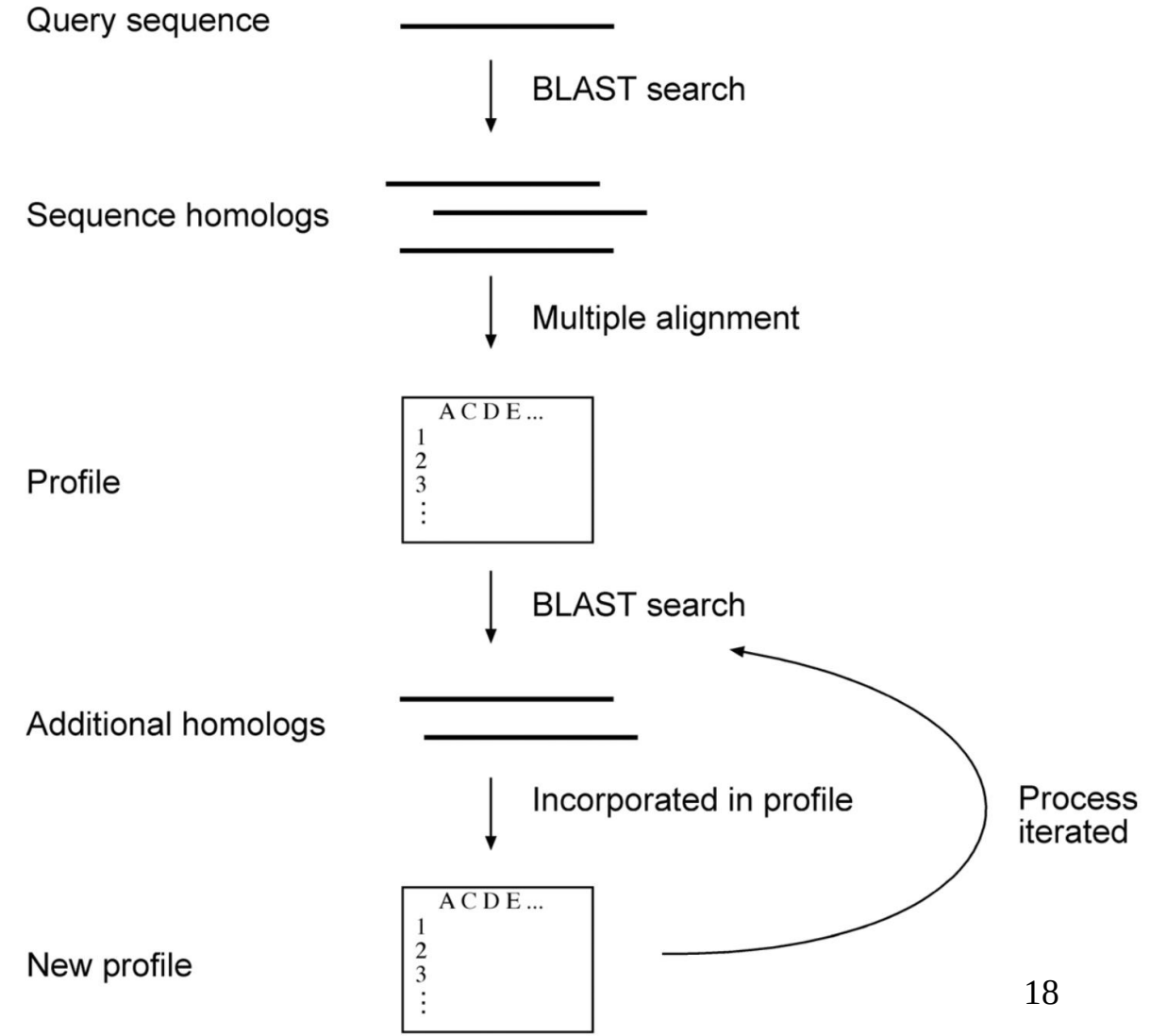


	R
A	-1
R	5
N	0
D	-2
C	-3
Q	1
E	0
G	-2
H	0
I	-3
L	-2
K	2
M	-1
F	-3
P	-2
S	-1
T	-1
W	-3
Y	-2
V	-3

A	-1	-2	-1	0	-1	-2	0	-2
R	5	0	5	-2	1	-3	-2	0
N	0	6	0	0	0	-3	0	1
D	-2	1	-2	-1	0	-3	-1	-1
C	-3	-3	-3	-3	-3	-2	-3	-3
Q	1	0	1	-2	5	-3	-2	0
E	0	0	0	-2	2	-3	-2	0
G	-2	0	-2	6	-2	-3	6	-2
H	0	1	0	-2	0	-1	-2	8
I	-3	-3	-3	-4	-3	0	-4	-3
L	-2	-3	-2	-4	-2	0	-4	-3
K	2	0	2	-2	1	-3	-2	-1
M	-1	-2	-1	-3	0	0	-3	-2
F	-3	-3	-3	-3	-3	6	-3	-1
P	-2	-2	-2	-2	-1	-4	-2	-2
S	-1	1	-1	0	0	-2	0	-1
T	-1	0	-1	-2	-1	-2	-2	-2
W	-3	-4	-3	-2	-2	1	-2	-2
Y	-2	-2	-2	-3	-1	3	-3	2
V	-3	-3	-3	-3	-2	-1	-3	-3

Another Schematic Diagram of PSI-BLAST

- The program employs a weighting scheme in the profile construction in each iteration to increase sensitivity.
- It uses pseudocounts to provide extra weight to unobserved residues to make the profile more inclusive.
- It can detect weak but biologically significant similarities between sequences.
- It is associated with low selectivity caused by the false-positives generated in the automated profile construction process. This problem is known as *profile drift*.



Why (not) PSI-BLAST

- Weights sequence according to observed diversity *specific* to family of interest

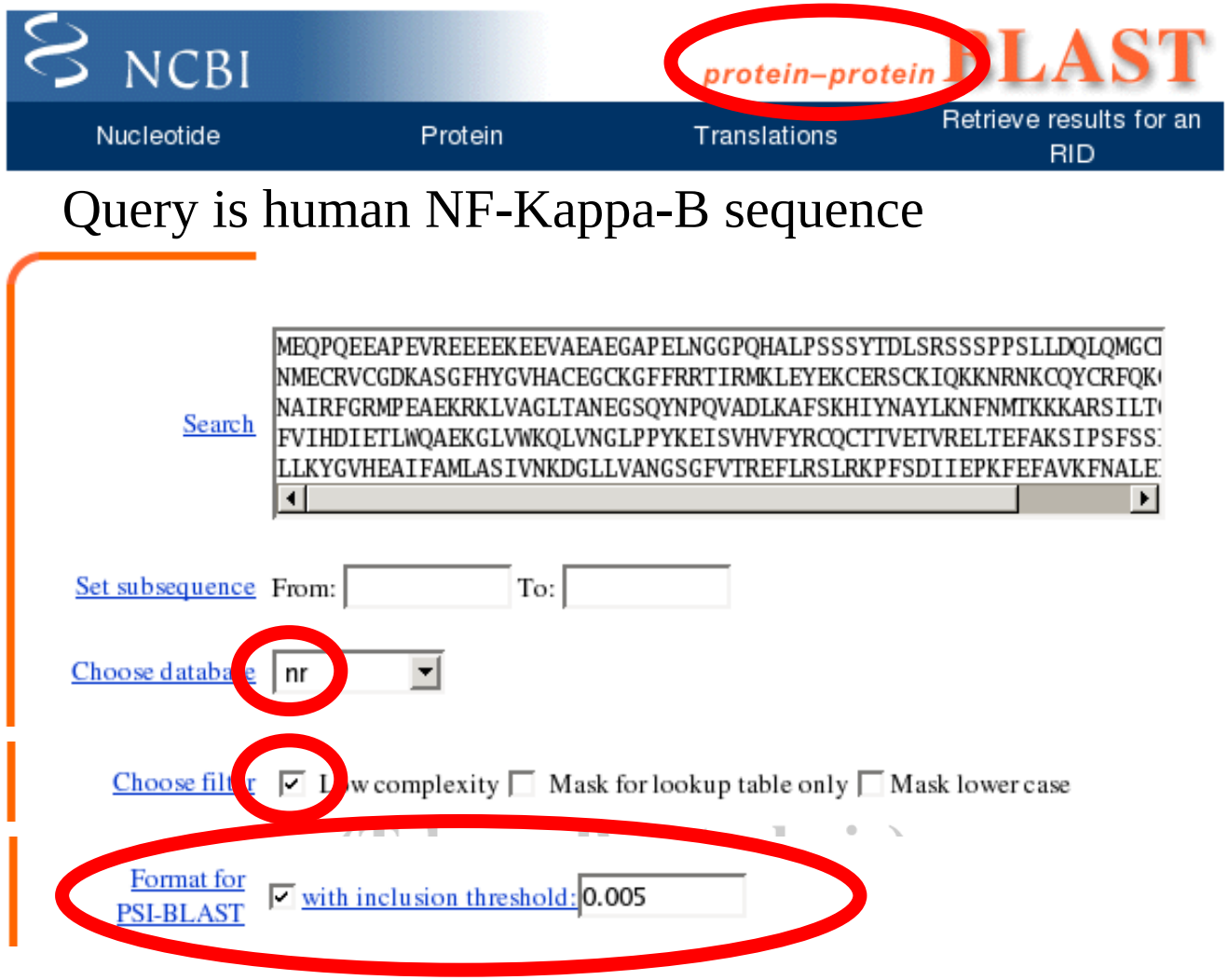
- **Advantage:** If sequences used to PSSMs are all homologous, *sensitivity* at a given *specificity* improves significantly

- **Disadvantage:** However, if any non-homologous sequences are included in PSSMs, they are “corrupted.” Then they "pull in" addition non-homologous sequences, and become worse than generic.

How to use PSI BLAST

- Set initial thresholds high
- Inspect each iteration's result for suspicious sequences
- Do several iterations (~5), or until no new sequences are found
- Even if only looking for a small set of sequences, make initial search very broad
  - First, use NR (large, inclusive database) with up to 5 iterations to set PSSM
  - Then use *that* PSSM to search in restricted domain

PSI-BLAST example



First Iteration

Distribution of 2373 Blast Hits on the Query Sequence

Legend:

NEW - means that the alignment score was below the threshold on the previous iteration

- means that the alignment was checked on the previous iteration

Run PSI-Blast iteration 2

Hit list size 1000

Sequences with E-value BETTER than threshold

			Score	E	
NEW	☑	gi|21759000|sp|Q12955|ANK3_HUMAN Ankyrin-3 (ANK-3) (Ankyrin G)	86.3	4e-16	G
NEW	☑	gi|21542418|sp|P19838|NFKB1_HUMAN Nuclear factor NF-kappa-B p...	1835	0.0	G
		...			
NEW	☑	gi|28558069|sp|Q99LW0|ANRX_MOUSE Ankyrin repeat domain protein 1	45.4	9e-04	G
NEW	☐	gi|13626132|sp|Q9QZH2|BARD1_RAT BRCA1-associated RING domain pro	45.1	0.001	G

Run PSI-Blast iteration 2

Sequences with E-value WORSE than threshold

☑	gi|20531989|sp|Q8WWX0|ASB5_HUMAN Ankyrin repeat and SOCS box pro	45.1	0.001	G
☐	gi|2493567|sp|Q60773|CDN7_MOUSE Cyclin-dependent kinase 4 inhibi	45.1	0.001	G

Second iteration

Distribution of 4921 Blast Hits on the Query Sequence

	gi|72077024|ref|XP_789126.1|	PREDICTED: similar to Ankyrin-2 ...	315	2e-84	G
NEW	gi|72022177|ref|XP_789744.1|	PREDICTED: similar to ankyrin 3,...	314	3e-84	G
	gi|72020988|ref|XP_792296.1|	PREDICTED: similar to Ankyrin-1 ...	314	5e-84	G
	gi|71981411|ref|NP_001021268.1|	UNCoordinated family member (unc	312	2e-83	G
NEW	gi|72165808|ref|XP_794269.1|	PREDICTED: similar to ankyrin 1,...	312	2e-83	G

1: XP_789744. Reports PREDICTED: simila...[gi:72022177]

```
LOCUS       XP_789744                488 aa            linear   INV 09-AUG-2005
DEFINITION  PREDICTED: similar to ankyrin 3, epithelial isoform b, partial
            [Strongylocentrotus purpuratus].
ACCESSION   XP_789744
VERSION     XP_789744.1  GI:72022177
DBSOURCE    REFSEQ: accession XM_784651.1
KEYWORDS    .
SOURCE      Strongylocentrotus purpuratus
  ORGANISM  Strongylocentrotus purpuratus
```

Format

Show Graphical Overview Linkout Sequence Retrieval NCBI gi PSSM in Text format

PSSM:2

425A683131415926535949FA6E0200707C7E547DF8001000017FE03FEFFFA0D0007F80003060969BE000000000000001579E
A5F5929E3B9D24A513E8D59B1112B6C84D6508A152910AA52AAB5912A9B312953595140EDB5A0AF7BD803C8001D074007

Representing a Profile as a Logo

- The score parameters of a PSSM are useful for obtaining alignments, but do not easily show the residue preferences or conservation at particular positions.
- This residue information is of interest because it is suggestive of the key functional sites of the protein family.
- A suitable graphical representation would make the identification of the se key residues easier.
- One solution to this problem uses information theory, and produces diagrams that are called logos.

Representing a Profile as a Logo (Cont.)

- In any PSSM (without pseudocount) column $\underline{u}$, residue type $\underline{a}$ will occur with a frequency $f_{u,a}$.
- The entropy in that position is defined by:

$$H_u = -\sum_a f_{u,a} \log_2 f_{u,a}$$

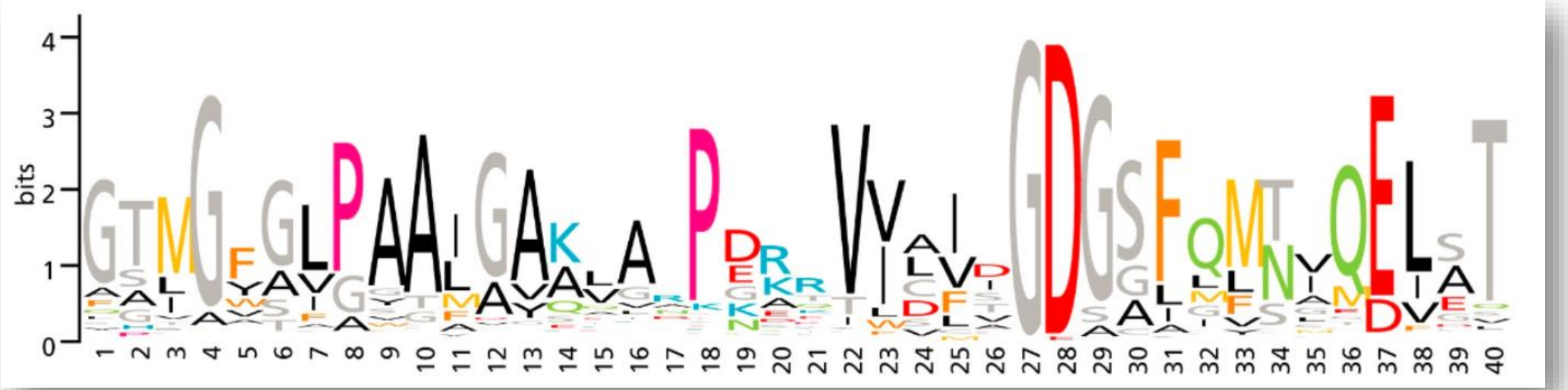
- The maximum value of H_u occurs if all residues are present with equal frequency, in which case H_u takes the value $\log_2 20$ for proteins.

Representing a Profile as a Logo (Cont.)

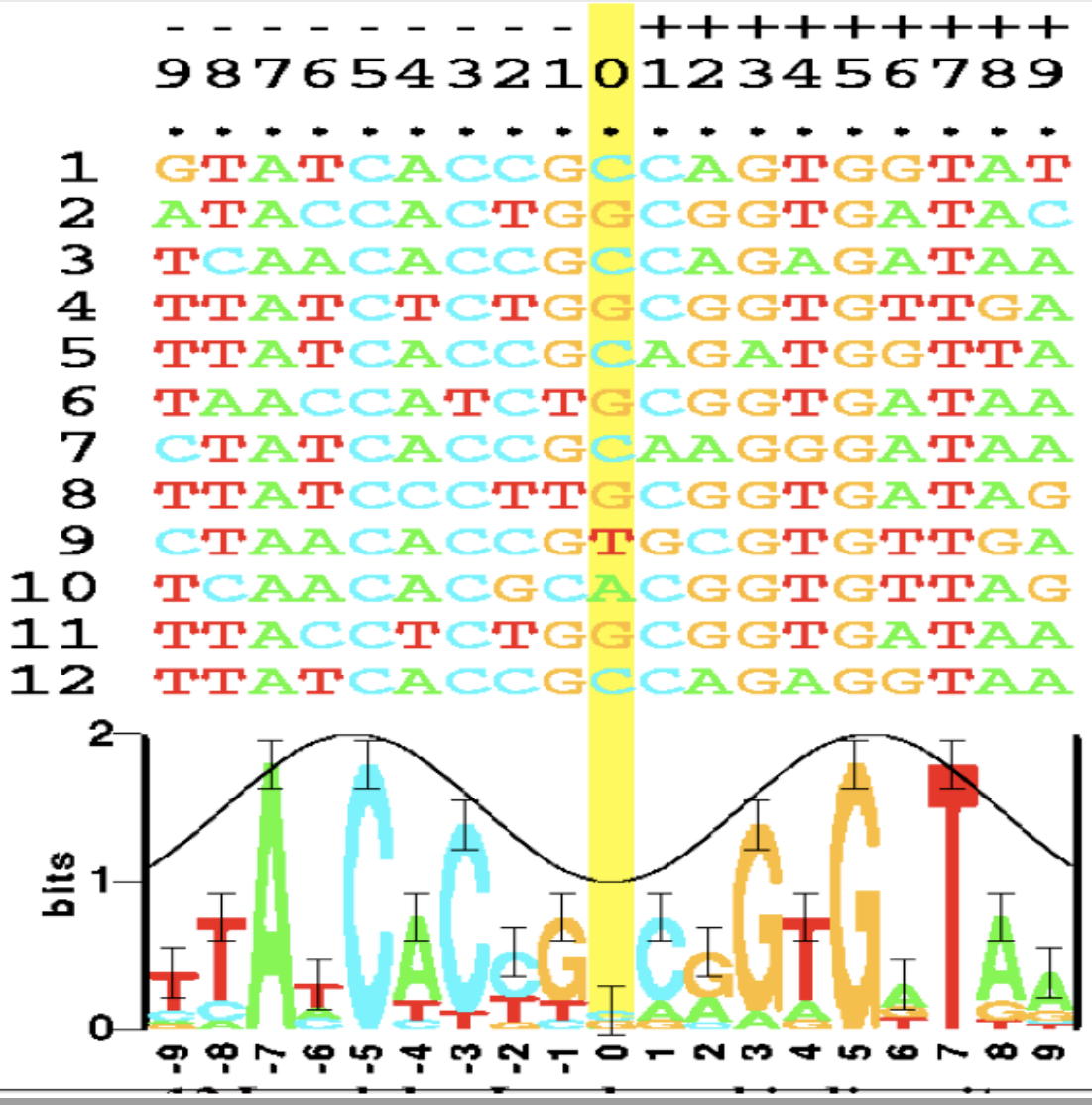
- The information present in the pattern at position $\underline{u}$ is given by:

$$I_u = \log_2 20 - H_u$$

- If the contribution of a residue is defined as $f_{u,a} I_u$, then a logo can be produced where at every position the residues are represented by their one-letter code, with each letter having a height proportional to its contribution.



Representing a Profile as a Logo: Example



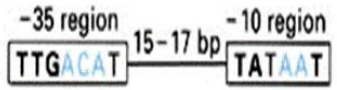
Discover Conserved Patterns by MSA

- Is there a *conserved* ***cis-acting*** regulatory sequence?
- *Rationale:* if sequences are homologous (derived from a common ancestor), they may be structurally/functionally equivalent.

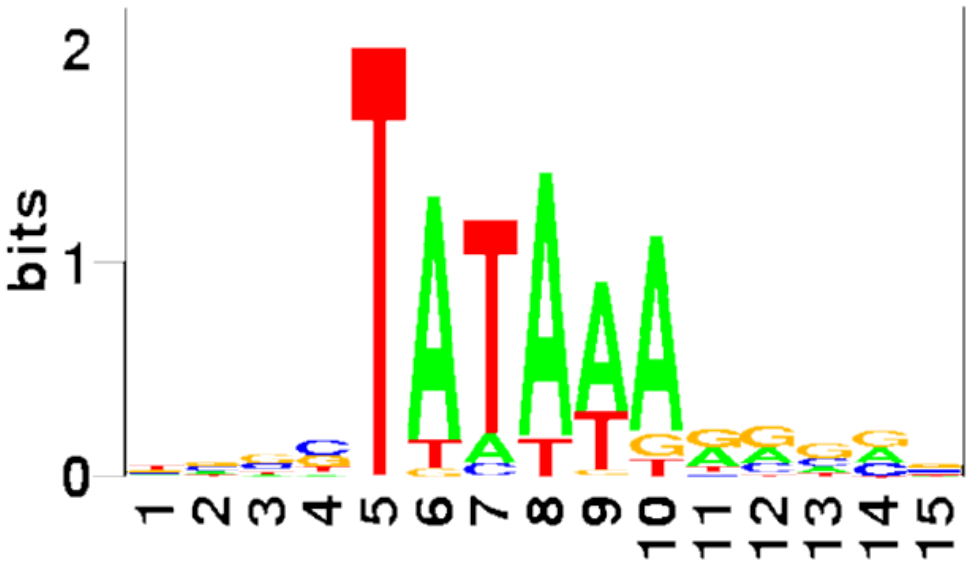
(a) Strong *E. coli* promoters

```
tyr tRNA     AACACTTTACAGCGGCG··CGTCATTTGATATGATGC·GCCCCGCTTCCCGA
rrn D1       AATACTTGTGCAAAAAA··TTGGGATCCCTATAATGCGCCTCCGTTGAGACG
rrn X1       TCCGCTTGTCTTCCTGA··GCCGACTCCCTATAATGCGCCTCCATCGACACG
rrn (DXE)2   CAGGGTTGACTCTGAAA··GAGGAAAGCGTAATATAC·GCCACCTCGCGACA
rrn E1       TTCTATTGCGGCCTGCG··GAGAACTCCCTATAATGCGCCTCCATCGACACG
rrn A1       TCCTCTTGTCAGGCCGG··AATAACTCCCTATAATGCGCCACCACTGACACG
rrn A2       AATGCTTGACTCTGTAG··CGGGAAGGCGTATTATGC·ACACCCCGCGCCGC
λ PR         GCGTGTTGACTATTTTA·CCTCTGGCGGTGATAATGG··TTGCATGTACTAA
λ PL         CGGTGTTGACATAAATA·CCACTGGCGGTGATACTGA··GCACATCAGCAGG
T7 A3        AACGGTTGACAACATGA·AGTAAACACGGTACGATGT·ACCACATGAAACGA
T7 A1        GAGTATTGACTTAAAGT·CTAACCTATAGGATACTTA·CAGCCATCGAGAGG
T7 A2        AGGTATTGACAACATGAAGTAACATGCAGTAAGATAC·AAATCGCTAGGTAA
fd VIII      CTCCGTTGTACTTTGTT··TCGCGCTTGGTATAATCG·CTGGGGGTCAAAGA
                 -35                          -10          +1
```

(b) Consensus sequences of promoters



TATA box:
transcriptional promoter element



A Full Example

TACGAT
TATAAT
TATAAT
GATACT
TATGAT
TATGTT
TATAGT

Consensus sequence: **TATAAT**
Regular expression: **[TG]A[TC][GA]XT**

Pseudocount = 1
Background Pro.=0.25

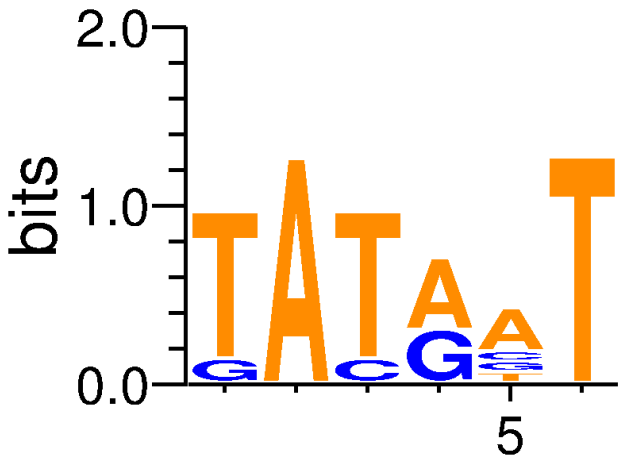
	1	2	3	4	5	6
A	0	7	0	4	4	0
C	0	0	1	0	1	0
G	1	0	0	3	1	0
T	6	0	6	0	1	7

Apply Pseudocount and Normalize by 7+4=11

	1	2	3	4	5	6
A	.09	.73	.09	.45	.45	.09
C	.09	.09	.18	.09	.18	.09
G	.18	.09	.09	.36	.18	.09
T	.64	.09	.64	.09	.18	.73

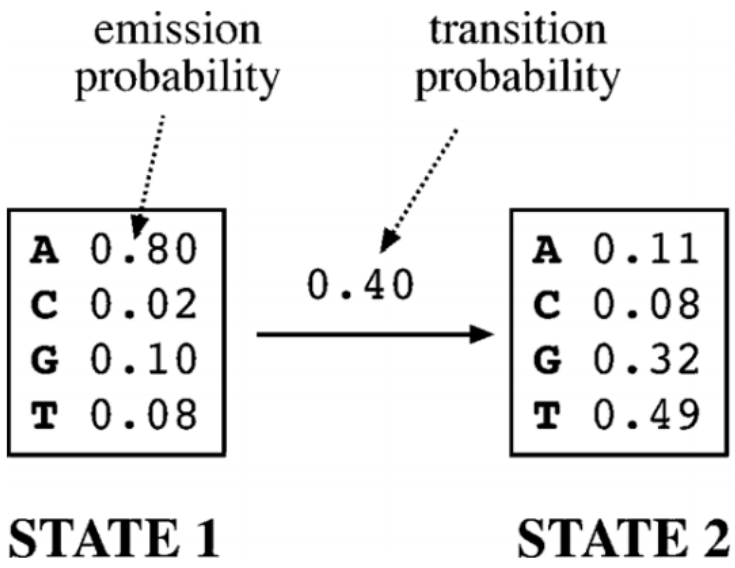
Normalize by Background and take log2

	1	2	3	4	5	6
A	-1.46	1.54	-1.46	0.86	0.86	-1.46
C	-1.46	-1.46	-0.46	-1.46	-0.46	-1.46
G	-0.46	-1.46	-1.46	0.54	-0.46	-1.46
T	1.35	-1.46	1.35	-1.46	-0.46	1.54



Hidden Markov Model in Bioinformatics

- A more efficient way of computing matching scores between a sequence and a sequence profile is through the use of HMMs.
  - It was originally developed for use in speech recognition.
- Each state composes of a number elements or symbols:
  - For nucleotide sequences, there are four possible symbols: A, T, G, and C.
  - For amino acid sequences, there twenty symbols.
- A partial HMM example:



- Probability of AG sequence:

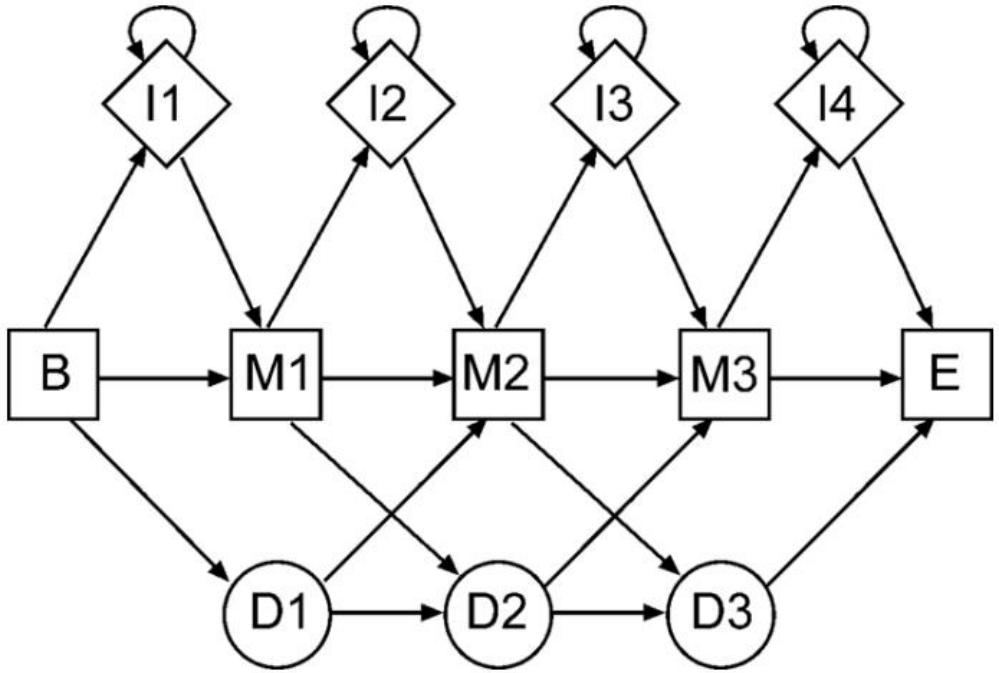
$$0.80 \times 0.40 \times 0.32 = 0.102$$

Hidden Markov Model in Bioinformatics

- To use an HMM to describe *gapped multiple sequence alignment*, a character in the alignment can be in one of three states:
  - Match, Insertion, and Deletion
- To represent the three states in an HMM, a special graphical representation has been used:
  - Transitions from state to state proceed from left to right
  - There are various paths through the model representing all possible combinations of matches, mismatches, and gaps to generate an alignment.

Hidden Markov Model in Bioinformatics

- Squares indicate match states (M), diamonds insert states (I), and circles delete states (D)
- The beginning and end of the match states are indicated by B and E, respectively.



- The circles on top of the insert state indicate self-looping, which allows insertions of any number of residues to fit into the model.

From Alignment to Profile

	1	2	3	4	5			6	7	8
	A	C	D	E	F	A	C	A	D	F
	A	F	D	A	–	–	–	C	C	F
Alignment	A	–	–	E	F	D	–	F	D	C
	A	C	A	E	F	–	–	A	–	C
	A	D	D	E	F	A	A	A	D	F

- First, remove columns if the fraction of gap symbols (“-”) exceeds θ, **the maximum fraction of insertions threshold**.

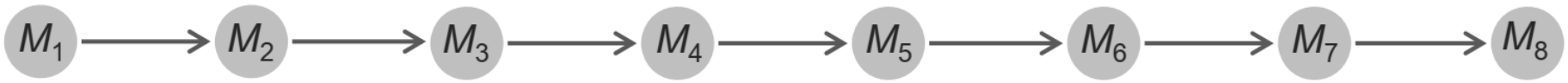
From Alignment to Profile (Cont.)

	1	2	3	4	5		6	7	8
Alignment	A	C	D	E	F	A C	A	D	F
	A	F	D	A	–	– –	C	C	F
	A	–	–	E	F	D –	F	D	C
	A	C	A	E	F	– –	A	–	C
	A	D	D	E	F	A A	A	D	F

*Alignment**	A	C	D	E	F	A	D	F
	A	F	D	A	–	C	C	F
	A	–	–	E	F	F	D	C
	A	C	A	E	F	A	–	C
	A	D	D	E	F	A	D	F

PROFILE(*Alignment**)	A	1	0	0	1/5	0	3/5	0	0
	C	0	2/4	0	0	0	1/5	1/4	2/5
	D	0	1/4	3/4	0	0	0	3/4	0
	E	0	0	0	4/5	0	0	0	0
	F	0	1/4	0	0	1	1/5	0	3/5

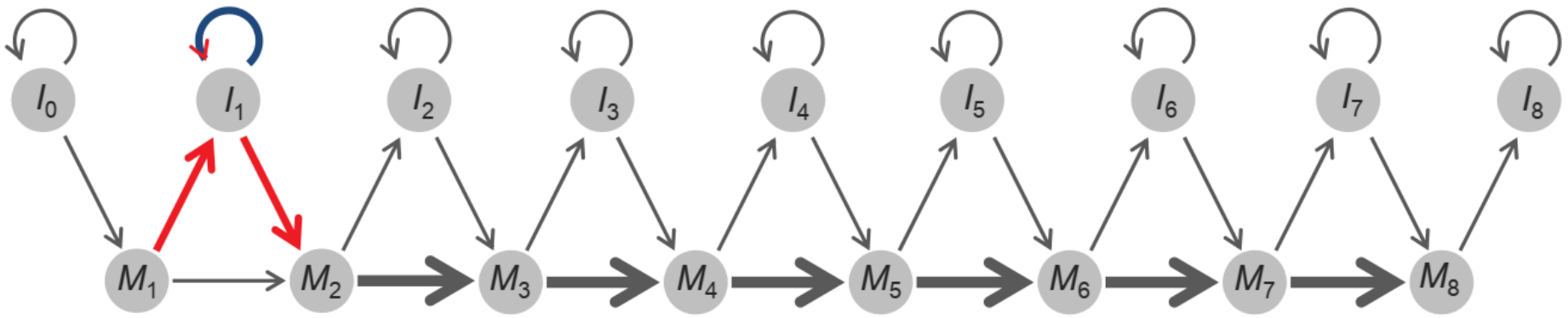
Toward a Profile HMM



A **F** D D A F F D F

How do we model insertions?

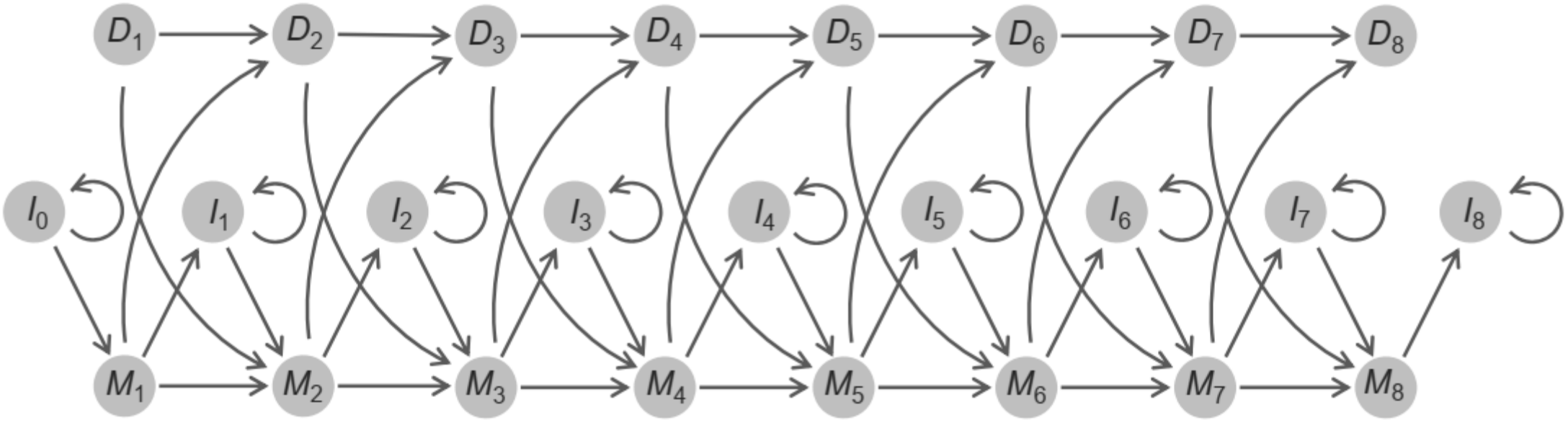
Toward a Profile HMM: Insertions



A **F** D D A F F D F

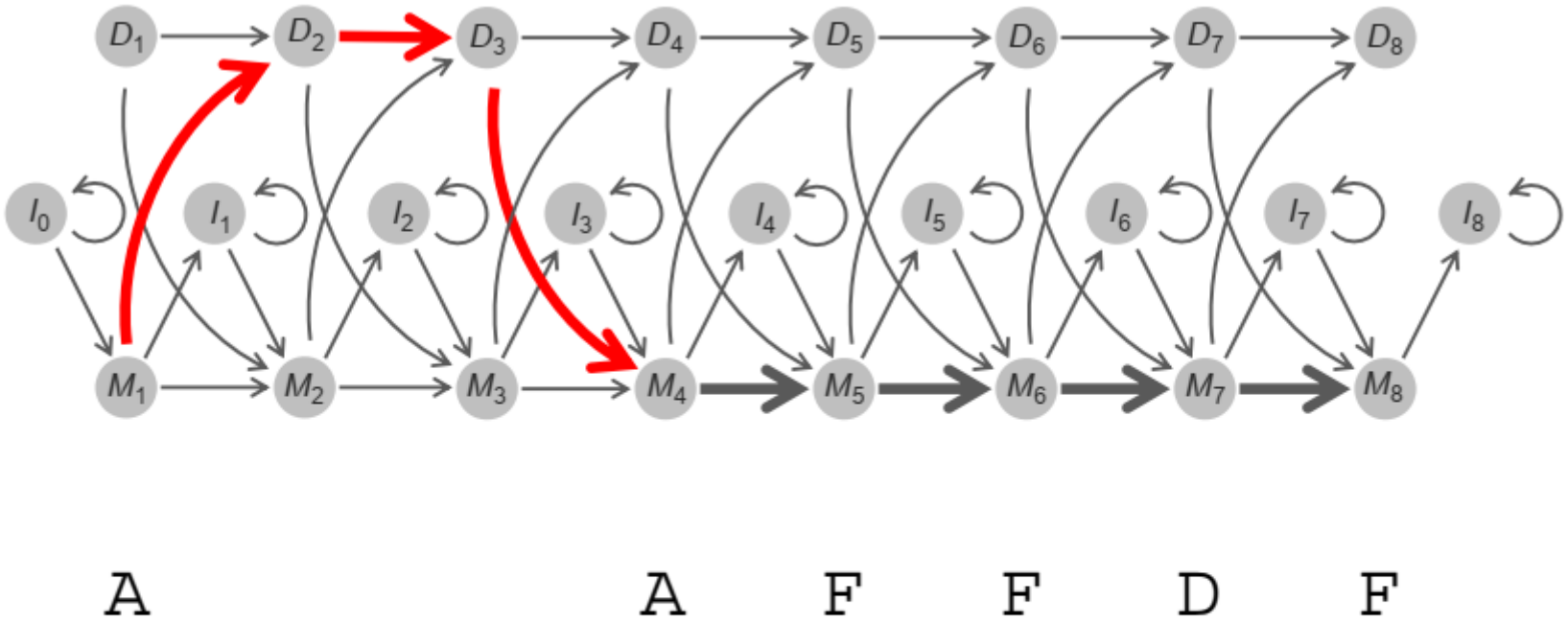
How do we model deletions?

Toward a Profile HMM: Deletions



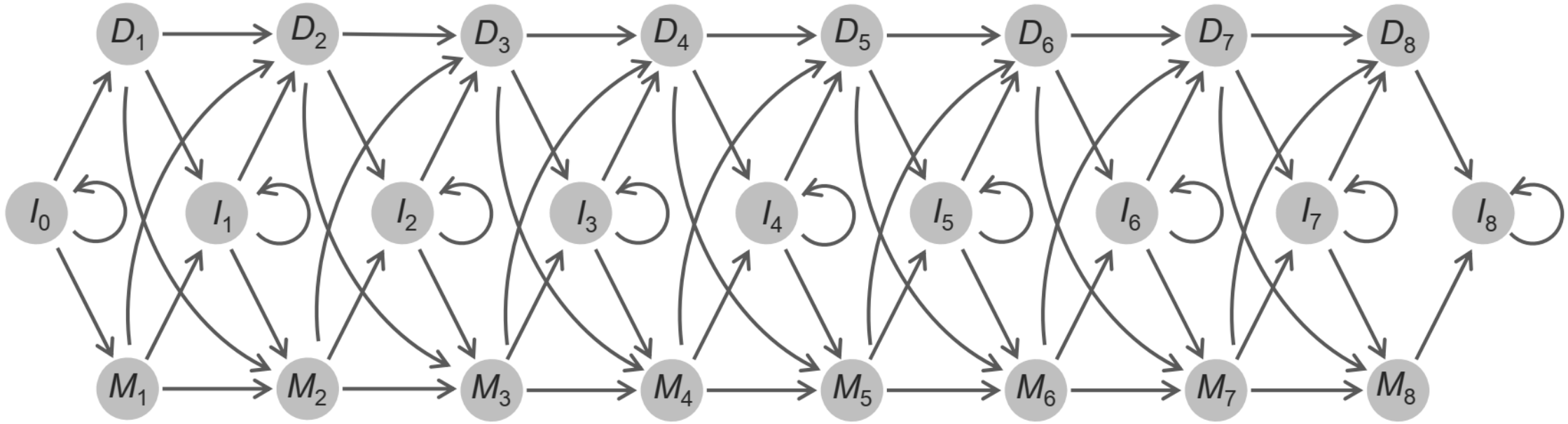
A A F F D F

Adding "Deletion States"

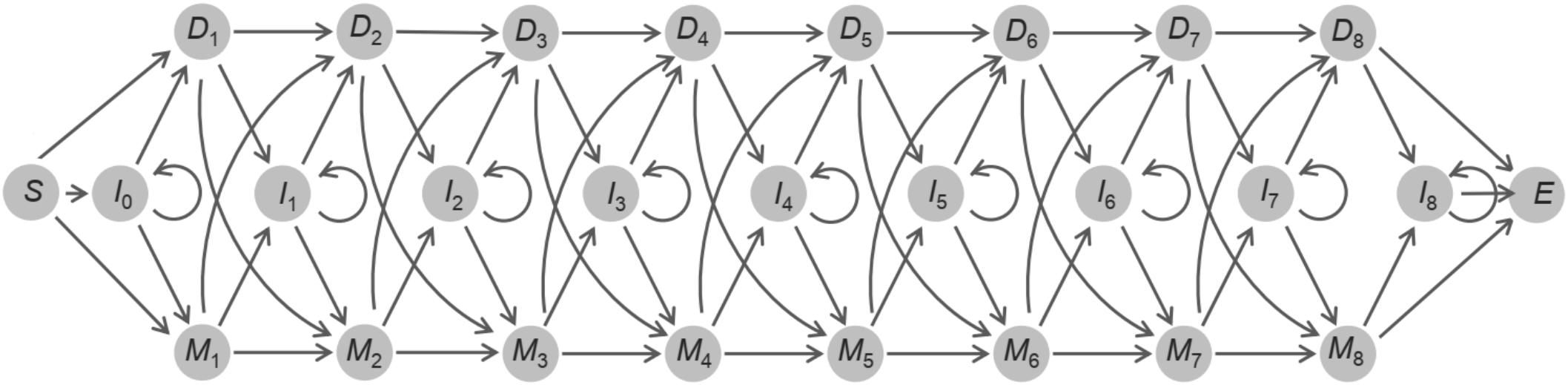


Are any edges still missing in this HMM diagram?

Adding Edges Between Deletion/Insertion States



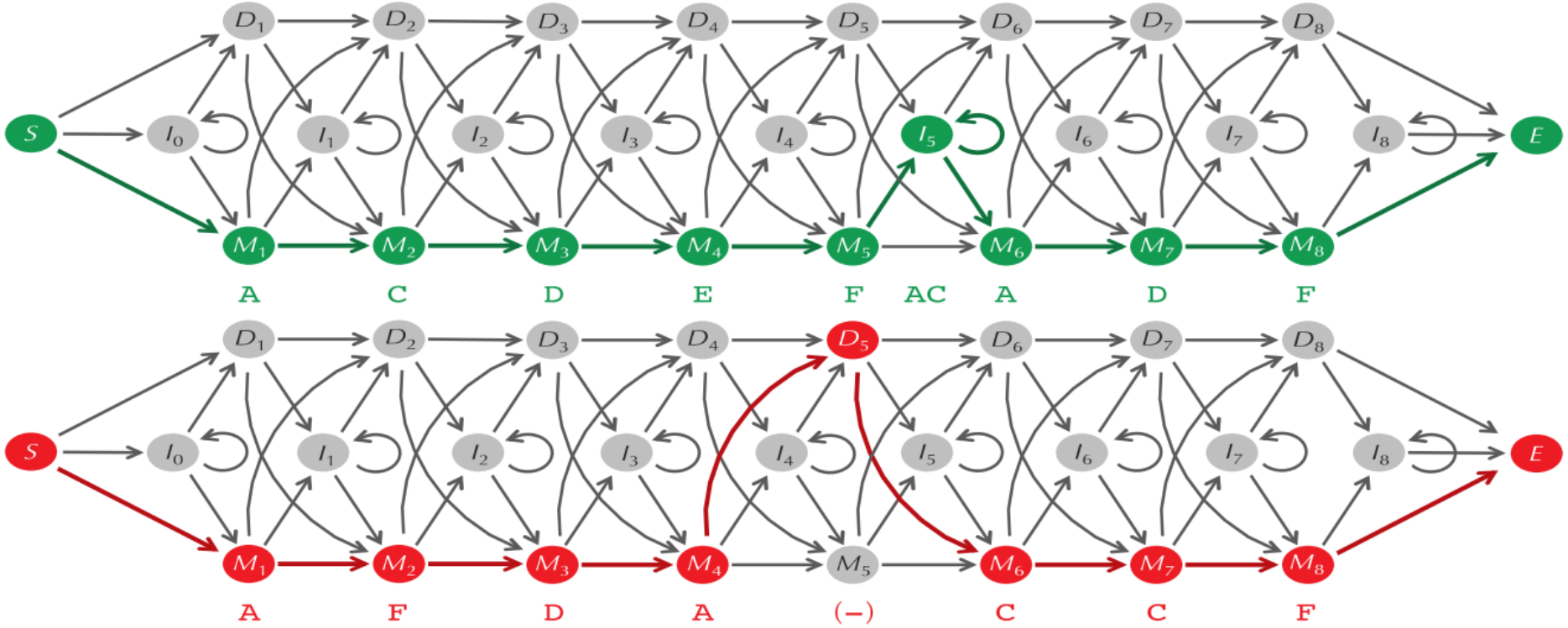
The Profile HMM is Ready to Use!



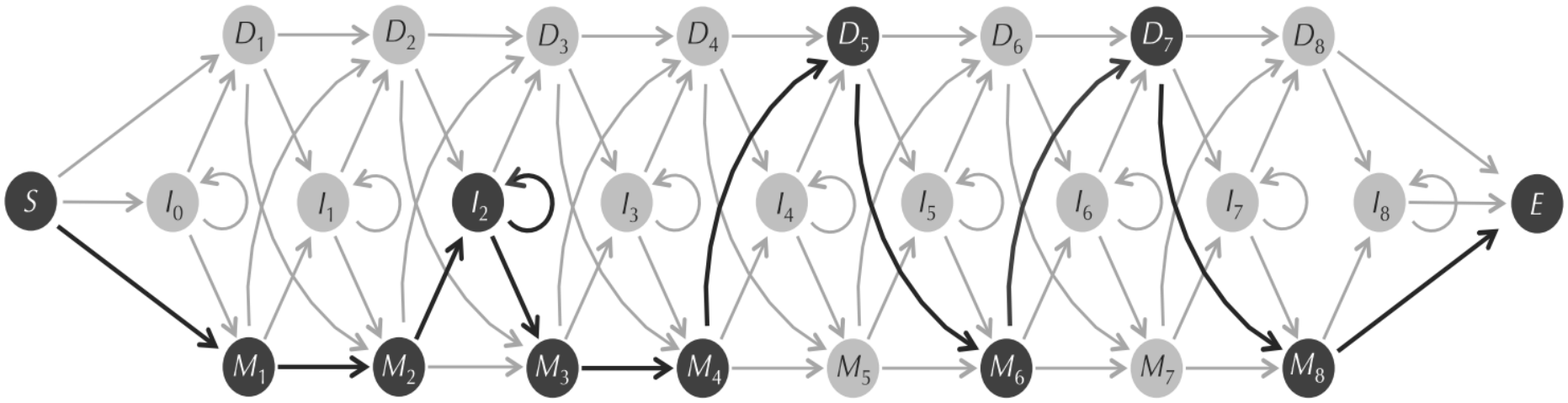
- **Profile HMM Problem**: *Construct a profile HMM from a multiple alignment.*
  - **Input:** A multiple alignment *Alignment* and a threshold θ (maximum fraction of insertions per column).
  - **Output:** Transition and emission matrices of the profile HMM *HMM*(*Alignment*,θ).

Hidden Paths Through Profile HMM

A	C	D	E	F	A	C	A	D	F
A	F	D	A	–	–	–	C	C	F
A	–	–	E	F	D	–	F	D	C
A	C	A	E	F	–	–	A	–	C
A	D	D	E	F	A	A	A	D	F



Aligning a Protein Against a Profile HMM

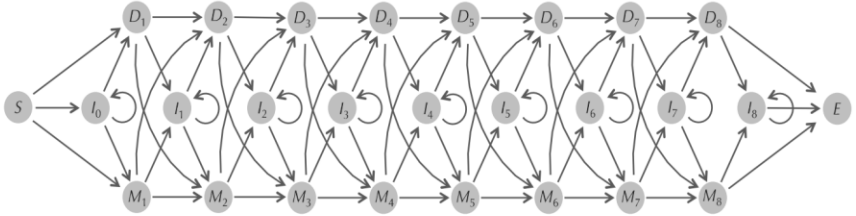


A C AF D E (–) A (–) F

Alignment

A	C	--	D	E	F	AC	A	D	F
A	F	--	D	A	–	--	C	C	F
A	–	--	–	E	F	D–	F	D	C
A	C	--	A	E	F	--	A	–	C
A	D	--	D	E	F	AA	A	D	F

Protein **ACAFDEAF**



Forbidden Transitions

	S	I_0	M_1	D_1	I_1	M_2	D_2	I_2	M_3	D_3	I_3	M_4	D_4	I_4	M_5	D_5	I_5	M_6	D_6	I_6	M_7	D_7	I_7	M_8	D_8	I_8	E
S			1																								
I_0																											
M_1						.8	.2																				
D_1																											
I_1																											
M_2									1																		
D_2																											
I_2																											
M_3												1															
D_3																											
I_3												1															
M_4															.8	.2											
D_4																											
I_4																											
M_5																	.25	.75									
D_5																	.33	.67									
I_5																		1									
M_6																					.8	.2					
D_6																											
I_6																											
M_7																								1			
D_7																											
I_7																								1			
M_8																											1
D_8																											
I_8																											
E																											

Gray cells: edges in the HMM diagram.

Clear cells: *forbidden* transitions.

Don't forget pseudocounts:
HMM(*Alignment*,θ,σ)

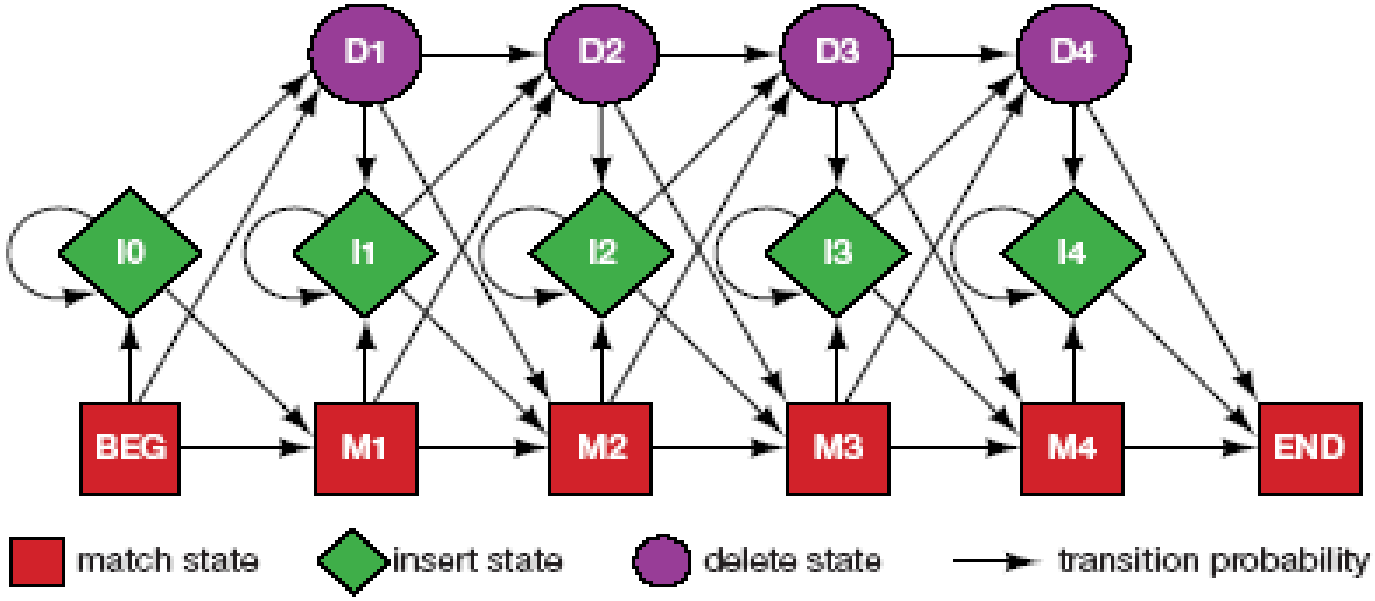
HMM Topology Example

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



Applications

- HMMs can be used for database searching to detect distant sequence homologs.
- HMMs are also used in protein family classification through motif and pattern identification.
- Advanced gene and promoter prediction also employ HMMs.
- HMMer (http://hmmer.wustl.edu/) is an HMM package for sequence analysis.
- The probability modeling in HMMs has more predictive power than profiles.

References

- Mostly used:
  - Essential bioinformatics, Chapter 6 (Profiles and Hidden Markov Models)
- Second reference:
  - Bioinformatics and functional genomics, Chapter 5 (Advanced Database Searching)

- IP notice: some slides were selected from Drena Dobbs' and Pevzner's slides.

Thanks for your attention