

CSE182-L9

Protein domain analysis via HMMs
Gene finding

QUIZ!

- Question:
- Your 'friend' likes to gamble.
- She tosses a coin: HEADS, she gives you a dollar. TAILS, you give her a dollar.
- Usually, she uses a fair coin, but 'once in a while', she uses a loaded coin.
- Can you say what fraction of the times she loads the coin?

Representation 2: Profiles

- Profiles versus regular expressions
 - Regular expressions are intolerant to an occasional mis-match.
 - The Union operation ($I+V+L$) does not quantify the relative importance of I, V, L . It could be that V occurs in 80% of the family members.
 - Profiles capture some of these ideas.

Profiles

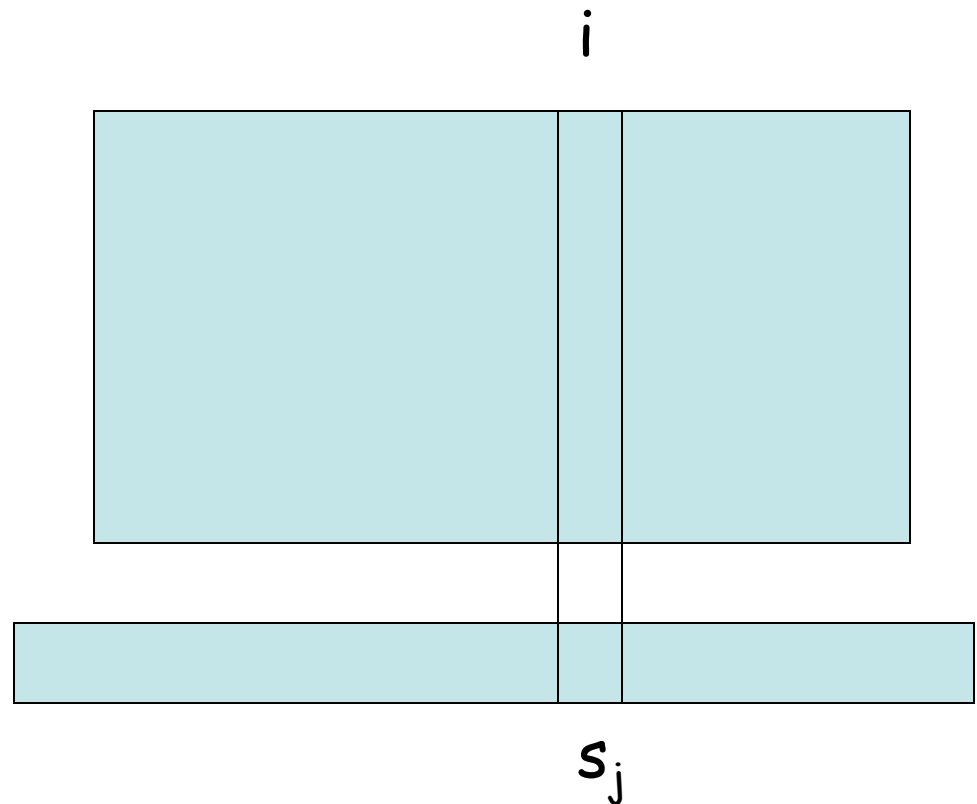
- Start with an alignment of strings of length m , over an alphabet A ,
- Build an $|A| \times m$ matrix $F=(f_{ki})$
- Each entry f_{ki} represents the frequency of symbol k in position i

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

A				
C				
D				
E				
F	0.71	0.14		
G				
H				
I				
K				
L				
M				
N				
P		0.28		
Q				
R				
S				
T				
V				
W				
Y	0.14			

Scoring matrices

- Given a sequence s , does it belong to the family described by a profile?
- We align the sequence to the profile, and score it
- Let $S(i,j)$ be the score of aligning position i of the profile to residue s_j
- The score of an alignment is the sum of column scores.



Scoring Profiles

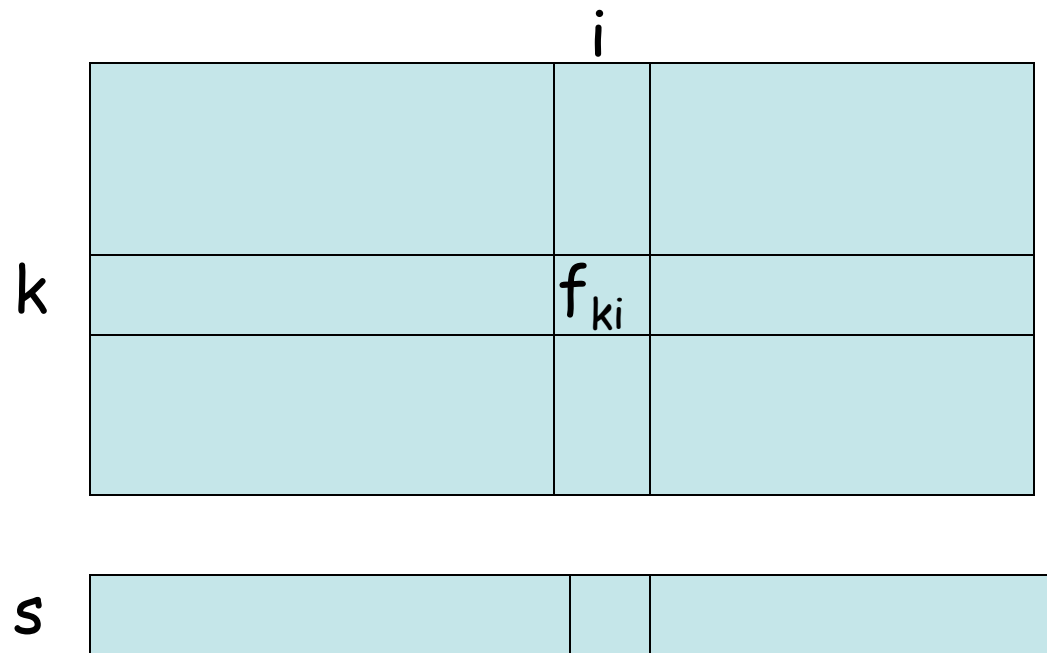
$$S(i, j) = \sum_k f_{ki} M[r_k, s_j]$$

Scoring Matrix

```

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

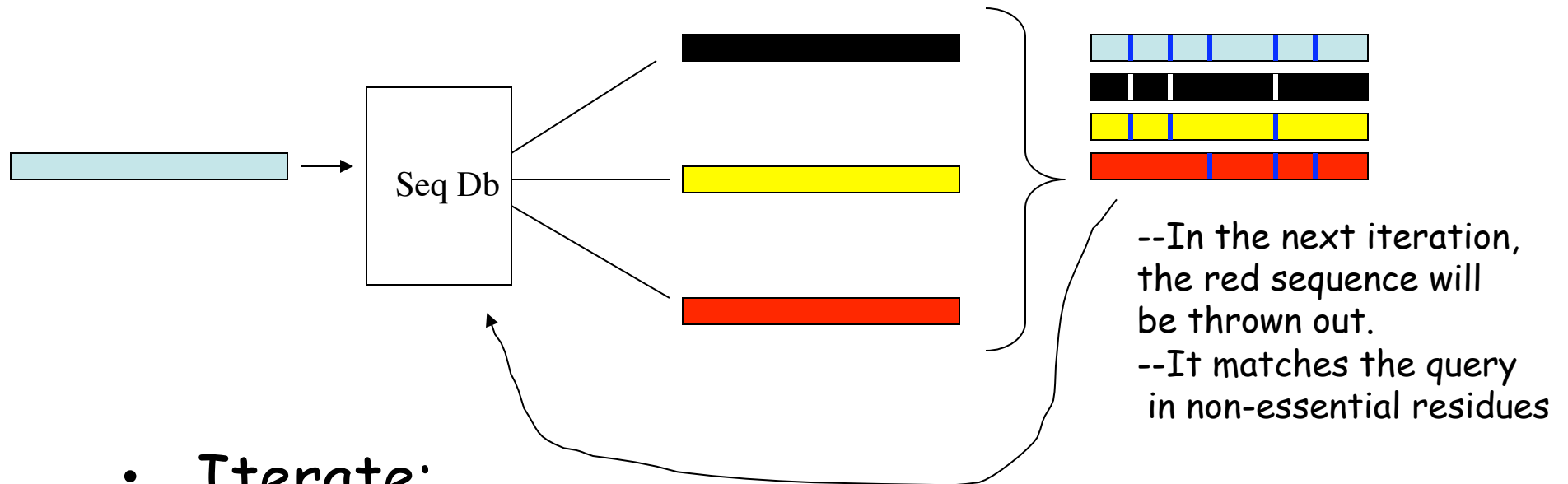
A	-18	-10	-1	-8	8	-3	3	-10	-2	-8
C	-22	-33	-18	-18	-22	-26	22	-24	-19	-7
D	-35	0	-32	-33	-7	6	-17	-34	-31	0
E	-27	15	-25	-26	-9	23	-9	-24	-23	-1
F	60	-30	12	14	-26	-29	-15	4	12	-29
G	-30	-20	-28	-32	28	-14	-23	-33	-27	-5
H	-13	-12	-25	-25	-16	14	-22	-22	-23	-10
I	3	-27	21	25	-29	-23	-8	33	19	-23
K	-26	25	-25	-27	-6	4	-15	-27	-26	0
L	14	-28	19	27	-27	-20	-9	33	26	-21
M	3	-15	10	14	-17	-10	-9	25	12	-11
N	-22	-6	-24	-27	1	8	-15	-24	-24	-4
P	-30	24	-26	-28	-14	-10	-22	-24	-26	-18
Q	-32	5	-25	-26	-9	24	-16	-17	-23	7
R	-18	9	-22	-22	-10	0	-18	-23	-22	-4
S	-22	-8	-16	-21	11	2	-1	-24	-19	-4
T	-10	-10	-6	-7	-5	-8	2	-10	-7	-11
V	0	-25	22	25	-19	-26	6	19	16	-16
W	9	-25	-18	-19	-25	-27	-34	-20	-17	-28
Y	34	-18	-1	1	-23	-12	-19	0	0	-18



Domain analysis via profiles

- Given a database of profiles of known domains/families, we can query our sequence against each of them, and choose the high scoring ones to functionally characterize our sequences.
- What if the sequence matches some other sequences weakly (using BLAST), but does not match any known profile?

Psi-BLAST idea



- Iterate:
 - Find homologs using Blast on query
 - Discard very similar homologs
 - Align, make a profile, search with profile.
 - Why is this more sensitive?

Representation 3: HMMs

- Building good profiles relies upon good alignments.
 - Difficult if there are gaps in the alignment.
 - Psi-BLAST/BLOCKS etc. work with gapless alignments.
- An HMM representation of Profiles helps put the alignment construction/membership query in a uniform framework.
- Also allows for position specific gap scoring.

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

The generative model

- Think of each column in the alignment as generating symbols according to a distribution.
- For each column, build a node that outputs an a.a. with the appropriate probability

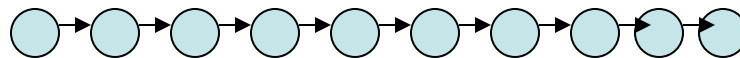
	F	
	F	
	Y	
	F	
	F	
	L	
	F	
A		
C		
D		
E		
F	0.71	
G		
H		
I		
K		
L		
M		
N		
P		
Q		
R		
S		
T		
V		
W		
Y	0.14	



Pr[F]=0.71
Pr[Y]=0.14

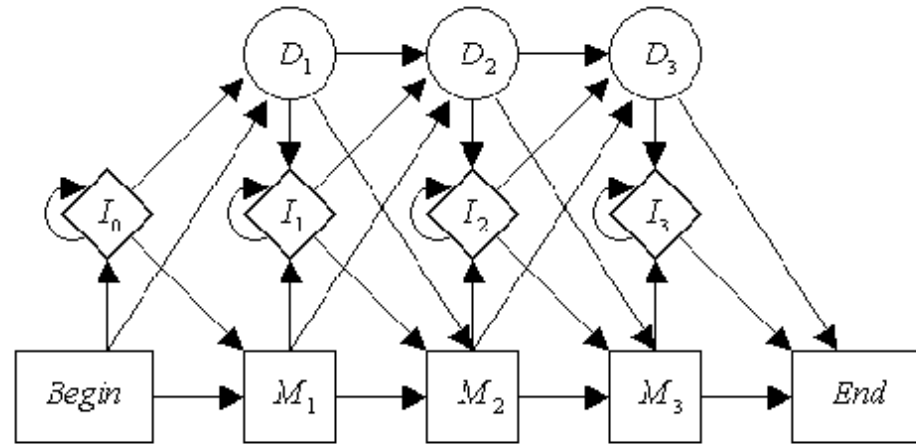
A simple Profile HMM

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



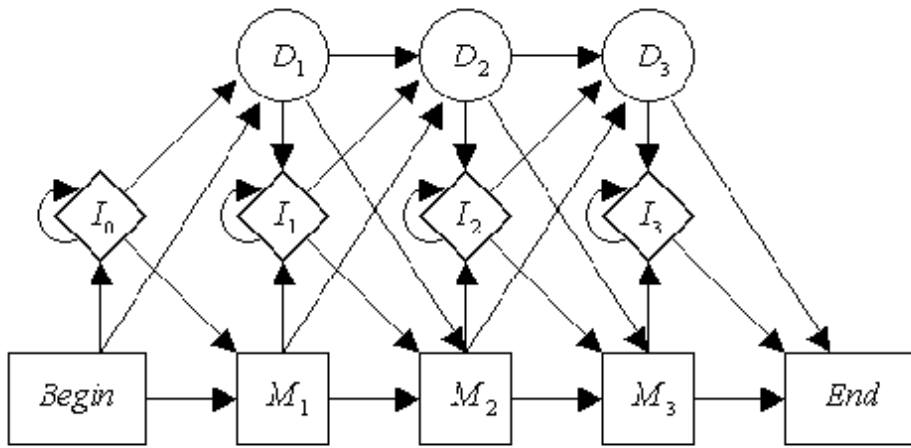
- Connect nodes for each column into a chain. This chain generates random sequences.
- What is the probability of generating FKVVGQVILD?
- In this representation
 - $\text{Prob}[\text{New sequence } S \text{ belongs to a family}] = \text{Prob}[\text{HMM generates sequence } S]$
- What is the difference with Profiles?

Profile HMMs can handle gaps



- The match states are the same as on the previous page.
- Insertion and deletion states help introduce gaps.
 - When in an insert state, generate any amino-acid
 - When in delete, generate a -
 - A sequence may be generated using different paths.

Example



A L - L
A I V L
A I - L

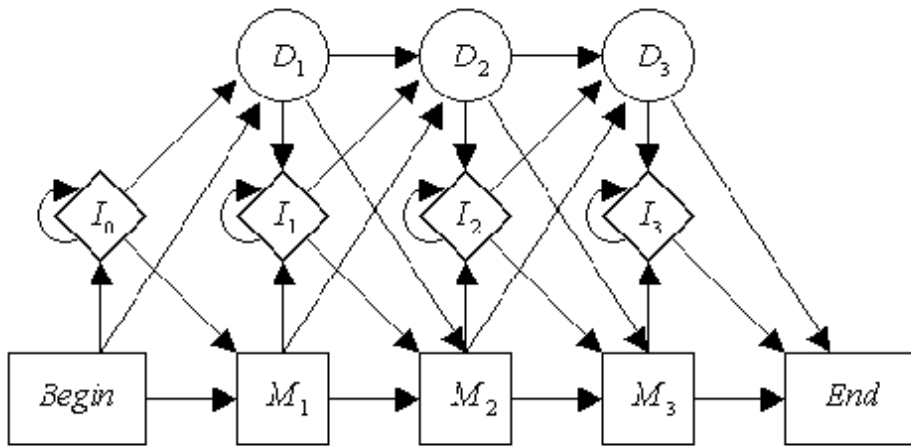
- Probability [ALIL] is part of the family?
- Note that multiple paths can generate this sequence.

- 1 Go to M1, and generate A
- 2 Go to I1 and generate L
- 3 Go to M2 and generate I
- 4 Go to M3 and generate L

OR

- 1 Go to M1, and generate A
- 2 Go to M2 and generate L
- 3 Go to I2 and generate I
- 4 Go to M3 and generate L

Example



A L - L
A I V L
A I - L

- Probability [ALIL] is part of the family?
- Note that multiple paths can generate this sequence.
 - $M_1 I_1 M_2 M_3$
 - $M_1 M_2 I_2 M_3$
- In order to compute the probabilities, we must assign probabilities of transition between states

Profile HMMs

- Directed Automaton \mathcal{M} with nodes and edges.
 - Nodes emit symbols according to 'emission probabilities'
 - Transition from node to node is guided by 'transition probabilities'
- Joint probability of seeing a sequence S , and path P
 - $\Pr[S, P | \mathcal{M}] = \Pr[S | P, \mathcal{M}] \Pr[P | \mathcal{M}]$
 - $\Pr[ALIL \text{ AND } M_1 I_1 M_2 M_3 | \mathcal{M}]$
 $= \Pr[ALIL | M_1 I_1 M_2 M_3, \mathcal{M}] \Pr[M_1 I_1 M_2 M_3 | \mathcal{M}]$
- $\Pr[ALIL | \mathcal{M}] = ?$

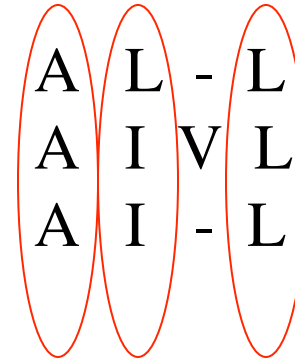
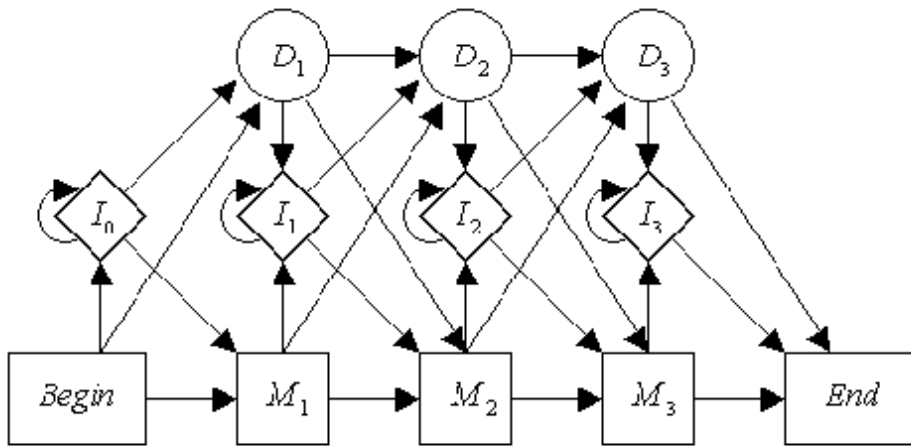
Formally

- The emitted sequence is $S=S_1S_2\dots S_m$
- The path traversed is $P_1P_2P_3\dots$
- $e_j(s)$ = emission probability of symbol s in state P_j
- Transition probability $T[j,k]$: Probability of transitioning from state j to state k .
- $\Pr(P,S|\mathcal{M}) = e_{P_1}(S_1) T[P_1,P_2] e_{P_2}(S_2) \dots\dots$
- What is $\Pr(S|\mathcal{M})$?

Two solutions

- An unknown (hidden) path is traversed to produce (emit) the sequence S .
- The probability that M emits S can be either
 - The sum over the joint probabilities over all paths.
 - $\Pr(S|M) = \sum_p \Pr(S,P|M)$
 - OR, it is the probability of the most likely path
 - $\Pr(S|M) = \max_p \Pr(S,P|M)$
- Both are appropriate ways to model, and have similar algorithms to solve them.

Viterbi Algorithm for HMM

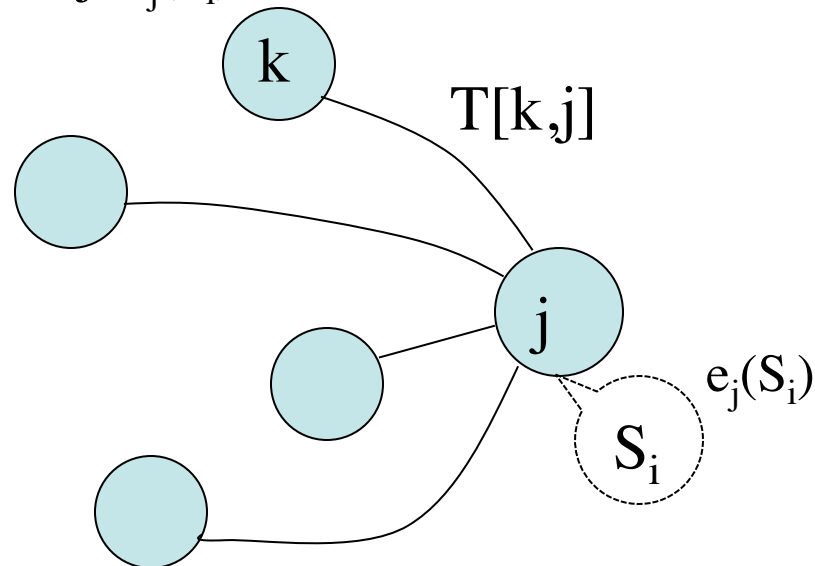


- Let $P_{\max}(i,j|M)$ be the probability of the most likely solution that emits $S_1 \dots S_i$, and ends in state j (is it sufficient to compute this?)
- $P_{\max}(i,j|M) = \max_k P_{\max}(i-1,k) T[k,j] e_j(S_i)$ (Viterbi)
- $P_{\text{sum}}(i,j|M) = \sum_k (P_{\text{sum}}(i-1,k) T[k,j]) e_j(S_i)$

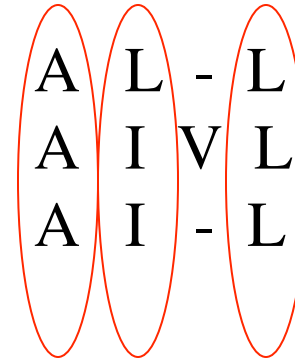
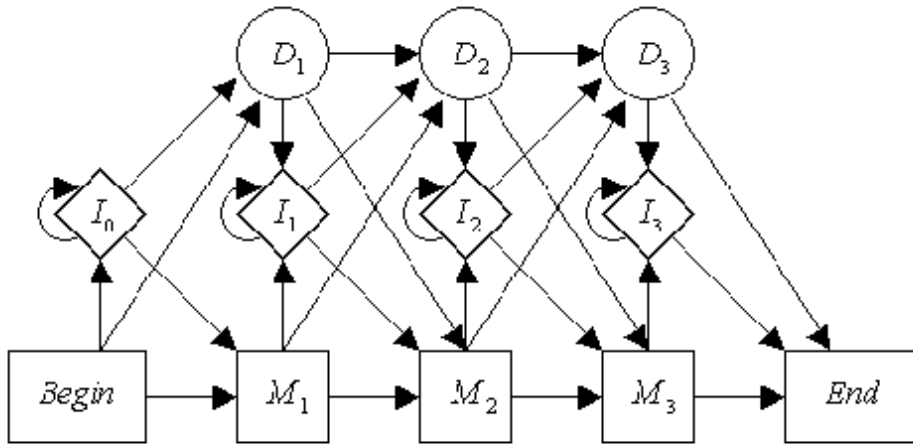
Viterbi illustration

- Let $P_{\max}(i,j|M)$ be the probability of the most likely solution that emits $S_1 \dots S_i$, and ends in state j (is it sufficient to compute this?)

$$P_{\max}(i,j|M) = \max_k P_{\max}(i-1,k) T[k,j] e_j(S_i)$$



Profile HMM membership



Path: A L I L
 M_1 M_2 I_2 M_3

- We can use the Viterbi/Sum algorithm to compute the probability that the sequence belongs to the family.
- Backtracking can be used to get the path, which allows us to give an alignment

Summary

- HMMs allow us to model position specific gap penalties, and allow for automated training to get a good alignment.
- Patterns/Profiles/HMMs allow us to represent families and focus on key residues
- Each has its advantages and disadvantages, and needs special algorithms to query efficiently.

Protein Domain databases

- A number of databases capture proteins (domains) using various representations
- Each domain is also associated with structure/function information, parsed from the literature.
- Each database has specific query mechanisms that allow us to compare our sequences against them, and assign function



3D



HMM



Biology

- In our discussion of BLAST, we alternated between looking at DNA, and protein sequences, treating them as strings.
 - DNA, RNA, and proteins are the 3 important molecules
- What is the relation between the three?

Genetics: What is a gene?

Helen Pearson¹

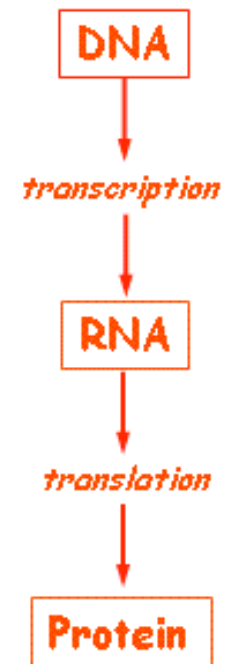
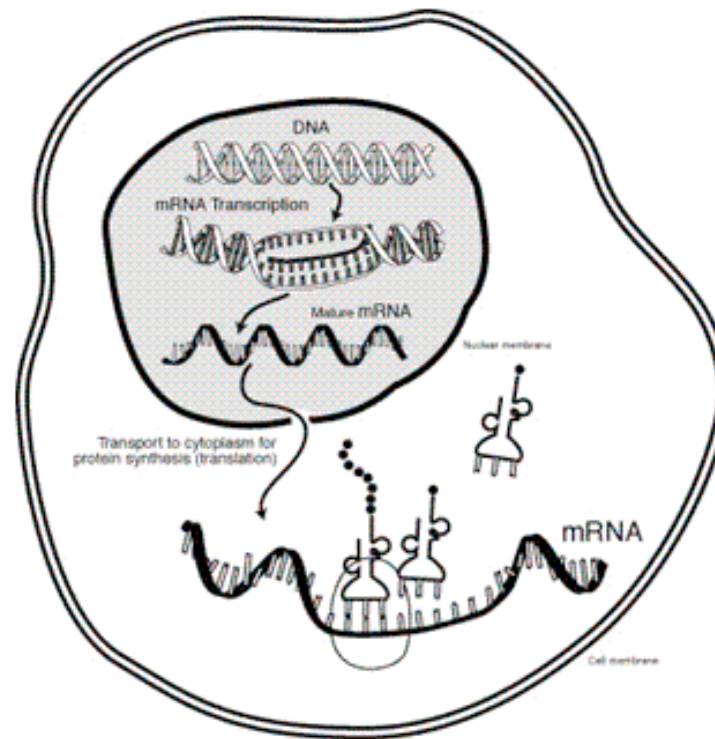
1. Helen Pearson is a reporter working for Nature in New York.

The idea of genes as beads on a DNA string is fast fading. Protein-coding sequences have no clear beginning or end and RNA is a key part of the information package, reports Helen Pearson.

Rick Young, a geneticist at the Whitehead Institute in Cambridge, Massachusetts, says that when he first started teaching as a young professor two decades ago, it took him about two hours to teach fresh-faced undergraduates what a gene was and the nuts and bolts of how it worked. Today, he and his colleagues need three months of lectures to convey the concept of the gene, and that's not because the students are any less bright. "It takes a whole semester to teach this stuff to talented graduates," Young says. "It used to be we could give a one-off definition and now it's much more complicated."

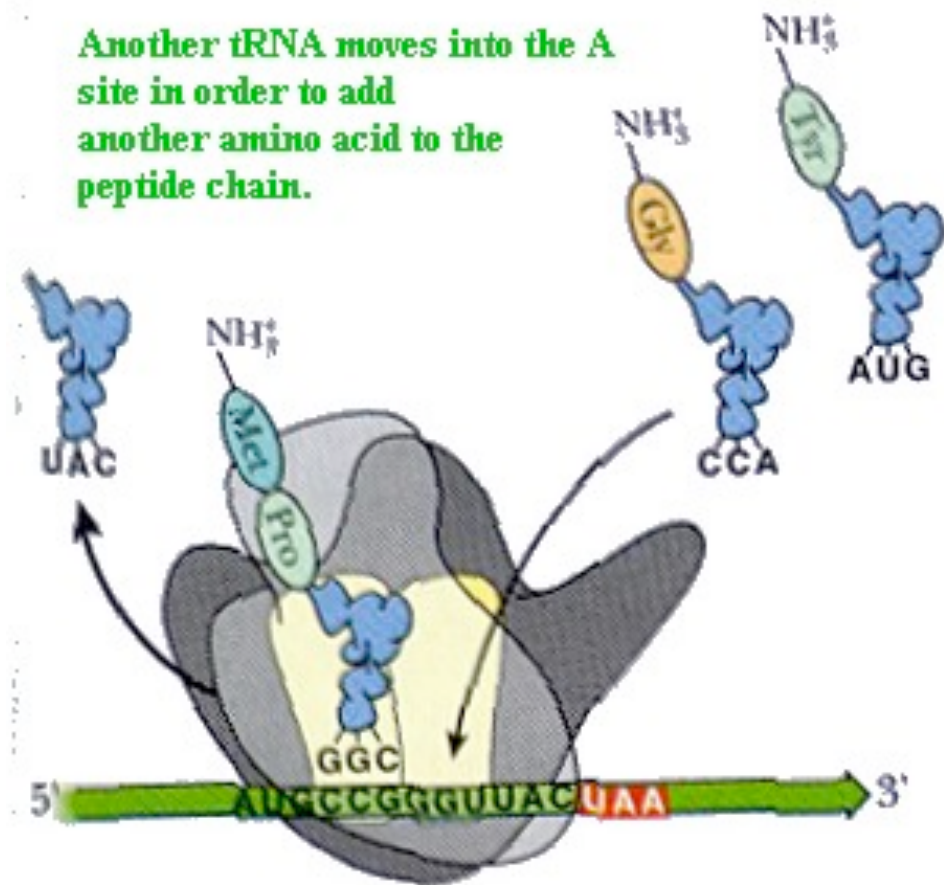
Transcription and translation

- We define a gene as a location on the genome that codes for proteins.
- The genic information is used to manufacture proteins through transcription, and translation.
- There is a unique mapping from triplets to amino-acids



Translation

- The ribosomal machinery reads mRNA.
- Each triplet is translated into a unique amino-acid until the STOP codon is encountered.
- There is also a special signal where translation starts, usually at the ATG (M) codon.



End of L9