HMM in sequence alignment

Sweta Mahajan

Under the supervision of Dr. Anirvan Chakraborty
Indian Institute of Science Education and Research,Kolkata

December 19, 2021

Outline of the talk

- Introduction to Pairwise Alignment
- Global Pairwise Sequence Alignment: Needleman Wunsch Algorithm(**GPSA**)
- Local Pairwise Sequence Alignment: Smith Waterman Algorithm(**LPSA**)
- Pairwise Alignment using HMM
- Multiple Sequence Alignment and Profile HMM

Introduction to Pairwise Alignment

What is Alignment?

Protein sequences(string of arrangement of letters from the set of 20 amino acids)

```
HBA_HUMAN  GSAQVKGHGKKVADALTNAVAHV---D--DMPNALSALSDLHAHKL
           ++ ++++H+ KV    + +A  ++           +L+ L+++H+ K
LGB2_LUPLU NNPELQAHAGKVFKLVYEAAIQLQVTGVVVTDATLKNLGSVHVSKG
```

'+' - conservative/similar amino acids who contribute positive score

' ' - non-conservative change which contribute negative score

Introduction to Pairwise Alignment

Why alignment?

Deletions, insertions and substitutions. Two similar functioning sequences have diverged from each other.

One protein sequence - function is known.

Align the query sequence with the given one.

Assign score to an aligned pair of residues, to gaps and add to find total.

Score is more -similarity is not by chance.

Gather information about one Biological Sequence from another.

Introduction to Pairwise Alignment

Example alignment

```
HBA_HUMAN   GSAQVKGHGKKVADALTNAVAHV---D--DMPNALSALSDLHAHKL
            ++ ++++H+ KV    + +A   ++           +L+ L+++H+ K
LGB2_LUPLU NNPELQAHAGKVFKLVYEAAIQLQVTGVVVTDATLKNLGSVHVSKG
```

Figure: This particular alignment is meaningful as the two sequences are evolutionarily related, have the same three dimensional structure and same function in oxygen binding.

```
HBA_HUMAN   GSAQVKGHGKKVADALTNAVAHVDDMPNALSALSD----LHAHKL
            GS+ + G +    +D L   ++ H+ D+   A +AL D     ++AH+
F11G11.2    GSGYLVGDSLTFVDLL--VAQHTADLLAANAALLDEFPQFKAHQE
```

Figure: However this is an illegitimate alignment. The sequences have different three dimensional structures and functions.

Introduction to Pairwise Alignment

Good Scoring Model

Takes into account the evolutionary history of the biological sequence. (expert molecular biologist may be able to give score by hand)

Natural selection screens the mutations and hence we see some patterns.

Assign score to an aligned pair of residues, to gaps and add to find total. Logarithm of the relative likelihood of the sequences being related, compared to being unrelated.

Identities and conservative substitutions are more likely to be in alignment - constitute positive score terms.
Non-conservative changes are less likely to be aligned- contribute negative score terms.

Introduction to Pairwise Alignment

Substitution Matrix

Ungapped Alignment

$x = x_1, x_2, \cdots, x_n \quad y = y_1, y_2, \cdots, y_n$

$$\text{relative likelihood} = \frac{\text{sequences are related}}{\text{sequences are random}}$$

Under the random model, we have

$$P(x, y|R) = \prod_i q_{x_i} \prod_j q_{y_j}$$

q_{x_i}=Background probabilities
Under the Match model,

$$P(x, y|M) = \prod_i p_{x_i y_i}$$

Introduction to Pairwise Alignment

Substitution Matrix

Under the random model, we have

$$P(x, y|R) = \prod_i q_{x_i} \prod_j q_{y_j}$$

Under the Match model,

$$P(x, y|M) = \prod_i p_{x_i y_i}$$

$$\text{odds ratio} = \frac{P(x, y|M)}{P(x, y|R)} = \frac{\prod_i p_{x_i y_i}}{\prod q_{x_i} q_{y_i}}$$

$$\text{log odds ratio} = S = \sum_i s(x_i, y_i)$$

where $s(x_i, y_i) = log\left(\frac{p_{x_i y_i}}{q_{x_i} q_{y_i}}\right)$

Introduction to Pairwise Alignment

Substitution Matrix

BLOSUM50 and PAM are two substitution matrices which are derived in this way.

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
A	**5**	−2	−1	−2	−1	−1	−1	0	−2	−1	−2	−1	−1	−3	−1	1	0	−3	−2	0
R	−2	**7**	−1	−2	−4	1	0	−3	0	−4	−3	3	−2	−3	−3	−1	−1	−3	−1	−3
N	−1	−1	**7**	2	−2	0	0	0	1	−3	−4	0	−2	−4	−2	1	0	−4	−2	−3
D	−2	−2	2	**8**	−4	0	2	−1	−1	−4	−4	−1	−4	−5	−1	0	−1	−5	−3	−4
C	−1	−4	−2	−4	**13**	−3	−3	−3	−3	−2	−2	−3	−2	−2	−4	−1	−1	−5	−3	−1
Q	−1	1	0	0	−3	**7**	2	−2	1	−3	−2	2	0	−4	−1	0	−1	−1	−1	−3
E	−1	0	0	2	−3	2	**6**	−3	0	−4	−3	1	−2	−3	−1	−1	−1	−3	−2	−3
G	0	−3	0	−1	−3	−2	−3	**8**	−2	−4	−4	−2	−3	−4	−2	0	−2	−3	−3	−4
H	−2	0	1	−1	−3	1	0	−2	**10**	−4	−3	0	−1	−1	−2	−1	−2	−3	2	−4
I	−1	−4	−3	−4	−2	−3	−4	−4	−4	**5**	2	−3	2	0	−3	−3	−1	−3	−1	4
L	−2	−3	−4	−4	−2	−2	−3	−4	−3	2	**5**	−3	3	1	−4	−3	−1	−2	−1	1
K	−1	3	0	−1	−3	2	1	−2	0	−3	−3	**6**	−2	−4	−1	0	−1	−3	−2	−3
M	−1	−2	−2	−4	−2	0	−2	−3	−1	2	3	−2	**7**	0	−3	−2	−1	−1	0	1
F	−3	−3	−4	−5	−2	−4	−3	−4	−1	0	1	−4	0	**8**	−4	−3	−2	1	4	−1
P	−1	−3	−2	−1	−4	−1	−1	−2	−2	−3	−4	−1	−3	−4	**10**	−1	−1	−4	−3	−3
S	1	−1	1	0	−1	0	−1	0	−1	−3	−3	0	−2	−3	−1	**5**	2	−4	−2	−2
T	0	−1	0	−1	−1	−1	−1	−2	−2	−1	−1	−1	−1	−2	−1	2	**5**	−3	−2	0
W	−3	−3	−4	−5	−5	−1	−3	−3	−3	−3	−2	−3	−1	1	−4	−4	−3	**15**	2	−3
Y	−2	−1	−2	−3	−3	−1	−2	−3	2	−1	−1	−2	0	4	−3	−2	−2	2	**8**	−1
V	0	−3	−3	−4	−1	−3	−3	−4	−4	4	1	−3	1	−1	−3	−2	0	−3	−1	**5**

Figure: **BLOSUM50 substitution matrix. The entries are scaled and rounded for computational ease.**

Introduction to Pairwise Alignment

Gap Penalty

We are expected to penalise the gaps. So, the cost corresponding to 'g' number of gaps can be linear,

$$\gamma(g) = -gd$$

or can be affine

$$\gamma(g) = -d - (g-1)e$$

d=gap open penalty
e=gap extension penalty.

GPSA: Needleman Wunsch Algorithm

- Takes gaps into account
- gives optimal global alignment
- uses previous optimal solutions to subsequences and builds recursively on that using dynamic programming
- For the time being, we will use linear gap penalty

GPSA: Needleman Wunsch Algorithm

$x = x_1, x_2, \cdots, x_n \qquad y = y_1, y_2, \cdots, y_m$

We construct a matrix F in which $F(i, j)$ is the score of the best alignment of the sequence x and y up to the i^{th} and j^{th} position respectively.

I G K x_i
L G — y_j

We calculate $F(i, j)$ recursively from previous entries of the matrix and reach till the end $F(n, m)$, which by definition is the best score of the alignment.

GPSA: Needleman Wunsch Algorithm

To calculate $F(i, j)$, we can have three cases as follows:

- x_i is aligned with y_j in which case the score $F(i, j)$ becomes $F(i-1, j-1) + s(x_i, y_j)$
- x_i is aligned with a gap in which case the score $F(i, j)$ becomes $F(i-1, j) - d$
- y_j is aligned with a gap in which case the score $F(i, j)$ becomes $F(i, j-1) - d$

I G A x_i	A I G A x_i	G A x_i — —
L G V y_j	G V y_j — —	S L G V y_j

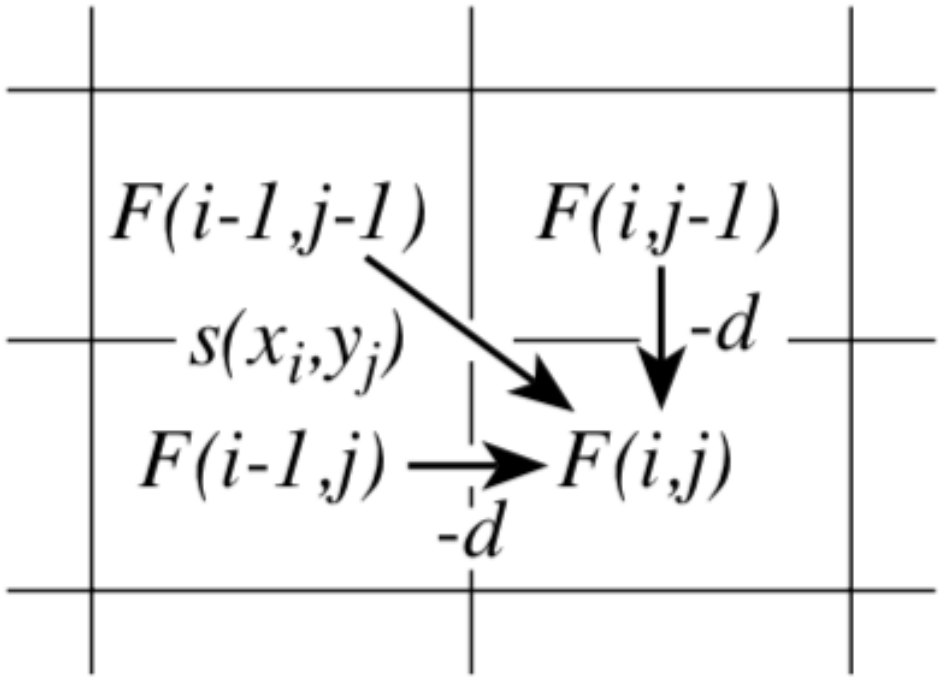
Figure: **x_i aligned to y_j, x_i aligned to gap, y_j aligned to gap**

GPSA: Needleman Wunsch Algorithm

$$F(0,0) = 0$$

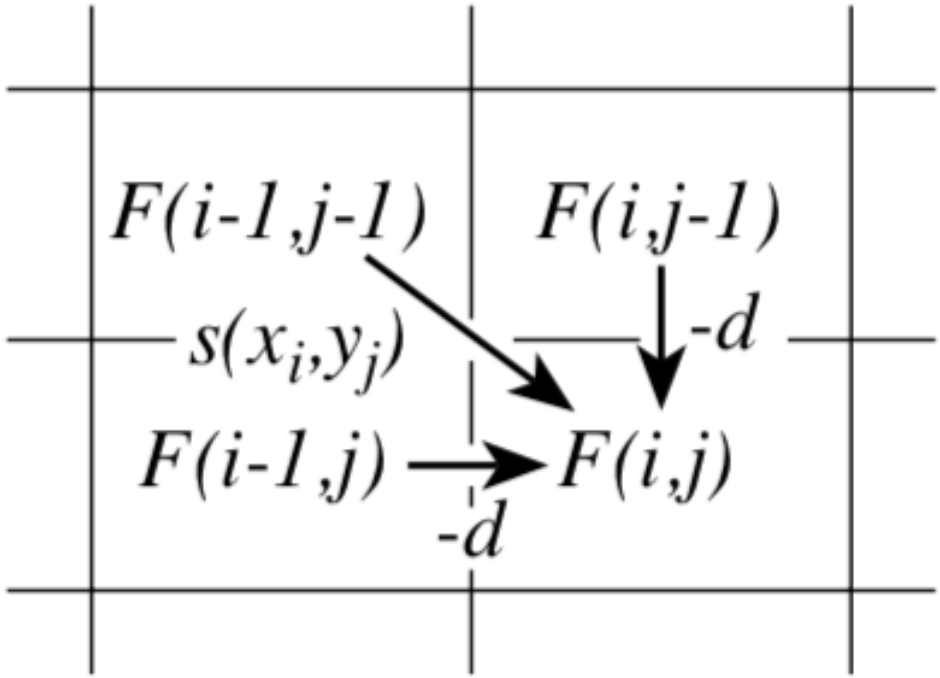
$$F(i,j) = max \begin{cases} F(i-1,j-1) + s(x_i, y_j) \\ F(i-1,j) - d \\ F(i,j-1) - d \end{cases}$$

The above procedure can be summarised in the picture below:



GPSA: Needleman Wunsch Algorithm

$F(i,0) = -id \qquad F(0,j) = -jd$



This algorithm works because the final score essentially is a addition of maximum terms from start to end.

Now, we trace back to find the optimal alignment.

LPSA: Smith Waterman Algorithm

Given pair of highly diverged sequence, the similarity often is found locally. The global alignment fails to say they are similar. Less overall similarities, but share common motifs. We tweak the global alignment algorithm a bit to obtain the local alignment as follows:

$$F(0,0) = 0$$

$$F(i,j) = max \begin{cases} 0 \\ F(i-1,j-1) + s(x_i, y_j) \\ F(i-1,j) - d \\ F(i,j-1) - d \end{cases}$$

- score becomes negative, new alignment starts.
- option zero, a consequence of it is that the first row and the first column in this matrix will now be filed with 0's.

Affine Gap Penalty

We are going to use the affine gap penalty system now.
M- Match state(the residues need not be identical)
X- Insert at sequence X
Y- Insert at sequence Y

$M(i,j)$- Score of best alignment between $x_1, x_2, \cdots, x_i$ & $y_1, y_2, \cdots, y_j$ given x_i is aligned with y_j

$X(i,j)$- Score of best alignment between $x_1, x_2, \cdots, x_i$ & $y_1, y_2, \cdots, y_j$ given x_i is aligned with a gap

$Y(i,j)$- Score of best alignment between $x_1, x_2, \cdots, x_i$ & $y_1, y_2, \cdots, y_j$ given y_j is aligned with a gap

I G A x_i	A I G A x_i	G A x_i — —
L G V y_j	G V y_j — —	S L G V y_j

Affine Gap Penalty

$M(i,j)$- Score of best alignment between $x_{1,\dots,i}$ & $y_{1,\dots,j}$ given x_i is aligned with y_j

$$M(i,j) = max \begin{cases} M(i-1,j-1) + s(x_i, y_j) \\ X(i-1,j-1) + s(x_i, y_j) \\ Y(i-1,j-1) + s(x_i, y_j) \end{cases}$$

I	G	x_{i-1}	x_i		I	G	x_{i-1}	x_i		I	G	—	x_i
L	G	y_{j-1}	y_j		L	G	–	y_j		L	G	y_{j-1}	y_j

Affine Gap Penalty

$X(i,j)$- Score of best alignment between $x_{1,\dots,i}$ & $y_{1,\dots,j}$ given x_i is aligned with a gap

$$X(i,j) = max \begin{cases} M(i-1,j) - d \\ X(i-1,j) - e \end{cases}$$

I	G	x_{i-1}	x_i		I	G	x_{i-1}	x_i
L	G	y_j	—		L	G	—	—

Affine Gap Penalty

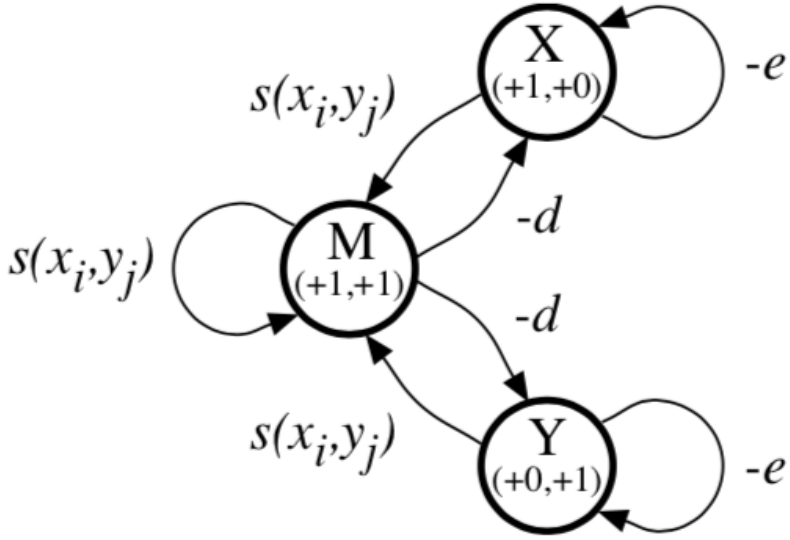
$Y(i,j)$- Score of best alignment between $x_{1,\dots,i}$ & $y_{1,\dots,j}$ given y_j is aligned with a gap

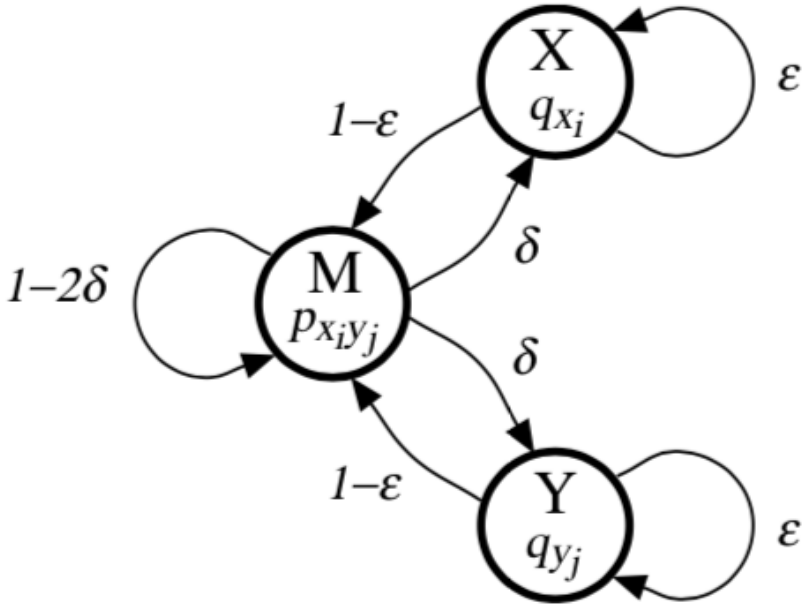
$$Y(i,j) = max \begin{cases} M(i,j-1) - d \\ X(i,j-1) - e \end{cases}$$

I G	x_i	—
L G	y_{j-1}	y_j

I G	—	—
L G	y_{j-1}	y_j

Pairwise Alignment using HMM





Pairwise Alignment using HMM

To find the optimal alignment, we use the viterbi algorithm. We can find out probability of similarity using the Forward algorithm.

$$P(x, y) = \sum_{\text{all alignment } \pi} P(x, y, \pi)$$

Profile HMM

- most often functional biological sequences come in a family having similar kind of function in different organisms.
- Sequences of a particular family are usually diverged from each other in their primary sequence due to duplication in the genome during cell division or by speciation which give rise to sequences with similar functions in related organisms.
- So, by identifying the family where the query sequence belongs gives us a hint about its function.
- The family shares common domain(conserved mutations) which is our focus.

Profile HMM

```
Helix                 AAAAAAAAAAAAAAAA   BBBBBBBBBBBBBBBBCCCCCCCCCCCC
HBA_HUMAN  ---------VLSPADKTNVKAAWGKVGA--HAGEYGAEALERMFLSFPTTKTYFPHF
HBB_HUMAN  --------VHLTPEEKSAVTALWGKV----NVDEVGGEALGRLLVVYPWTQRFFESF
MYG_PHYCA  ---------VLSEGEWQLVLHVWAKVEA--DVAGHGQDILIRLFKSHPETLEKFDRF
GLB3_CHITP ----------LSADQISTVQASFDKVKG------DPVGILYAVFKADPSIMAKFTQF
GLB5_PETMA PIVDTGSVAPLSAAEKTKIRSAWAPVYS--TYETSGVDILVKFFTSTPAAQEFFPKF
LGB2_LUPLU --------GALTESQAALVKSSWEEFNA--NIPKHTHRFFILVLEIAPAAKDLFS-F
GLB1_GLYDI ---------GLSAAQRQVIAATWKDIAGADNGAGVGKDCLIKFLSAHPQMAAVFG-F
Consensus            Ls....  v a W kv .  .    g . L.. f . P .   F  F

Helix          DDDDDDDEEEEEEEEEEEEEEEEEEE              FFFFFFFFFFFF
HBA_HUMAN  -DLS-----HGSAQVKGHGKKVADALTNAVAHV---D--DMPNALSALSDLHAHKL-
HBB_HUMAN  GDLSTPDAVMGNPKVKAHGKKVLGAFSDGLAHL---D--NLKGTFATLSELHCDKL-
MYG_PHYCA  KHLKTEAEMKASEDLKKHGVTVLTALGAILKK----K-GHHEAELKPLAQSHATKH-
GLB3_CHITP AG-KDLESIKGTAPFETHANRIVGFFSKIIGEL--P---NIEADVNTFVASHKPRG-
GLB5_PETMA KGLTTADQLKKSADVRWHAERIINAVNDAVASM--DDTEKMSMKLRDLSGKHAKSF-
LGB2_LUPLU LK-GTSEVPQNNPELQAHAGKVFKLVYEAAIQLQVTGVVVTDATLKNLGSVHVSKG-
GLB1_GLYDI SG----AS---DPGVAALGAKVLAQIGVAVSHL--GDEGKMVAQMKAVGVRHKGYGN
Consensus   .  t    .. . v..Hg kv. a   a...l   d   . a l. l   H  .

Helix       FFGGGGGGGGGGGGGGGGG     HHHHHHHHHHHHHHHHHHHHHHHHH
HBA_HUMAN  -RVDPVNFKLLSHCLLVTLAAHLPAEFTPAVHASLDKFLASVSTVLTSKYR------
HBB_HUMAN  -HVDPENFRLLGNVLVCVLAHHFGKEFTPPVQAAYQKVVAGVANALAHKYH------
MYG_PHYCA  -KIPIKYLEFISEAIIHVLHSRHPGDFGADAQGAMNKALELFRKDIAAKYKELGYQG
GLB3_CHITP --VTHDQLNNFRAGFVSYMKAHT--DFA-GAEAAWGATLDTFFGMIFSKM-------
GLB5_PETMA -QVDPQYFKVLAAVIADTVAAG---------DAGFEKLMSMICILLRSAY-------
LGB2_LUPLU --VADAHFPVVKEAILKTIKEVVGAKWSEELNSAWTIAYDELAIVIKKEMNDAA---
GLB1_GLYDI KHIKAQYFEPLGASLLSAMEHRIGGKMNAAAKDAWAAAYADISGALISGLQS-----
Consensus    v.   f  l . ..  ....    f   . aa. k. .     l sky
```

gaps tend to align with each other leaving ungapped regions in between

Profile HMM

Position Specific Score Matrix

First we try to model the ungapped region and then will deal with insertions and deletions.
The Position Specific Score Matrix(PSSM) gives the distribution of the residues at each position of a conserved motif(ungapped region).

$$P(x|M) = \prod_{i=1}^{L} e_i(x_i)$$

$$S = \sum_{i=1}^{L} log \frac{e_i(x_i)}{q_{x_i}}$$

Notice the similarity between substitution matrix and this.
A PSSM can also be used for match in a longer sequence x of length N by finding the score for each starting point k from 1 to $N - L - 1$, L being the length of the PSSM.

Profile HMM

We develop a probabilistic model called "profile HMM" to model insertions and deletions.
Backbone= columns that represent conserved motif of the family

Chain of repetitive match states corresponding to the backbone of the MSA, but with different emission probabilities.

PSSM can be modelled by this. Alignment is trivial.

Profile HMM

We take an example to help illustrate how to build profile HMMs. Suppose we have a motif **"WEIRD"** and an MSA as follows.

WEIRD

WEIRE

WEIQH

WECIRD

WECLIRD

WEID

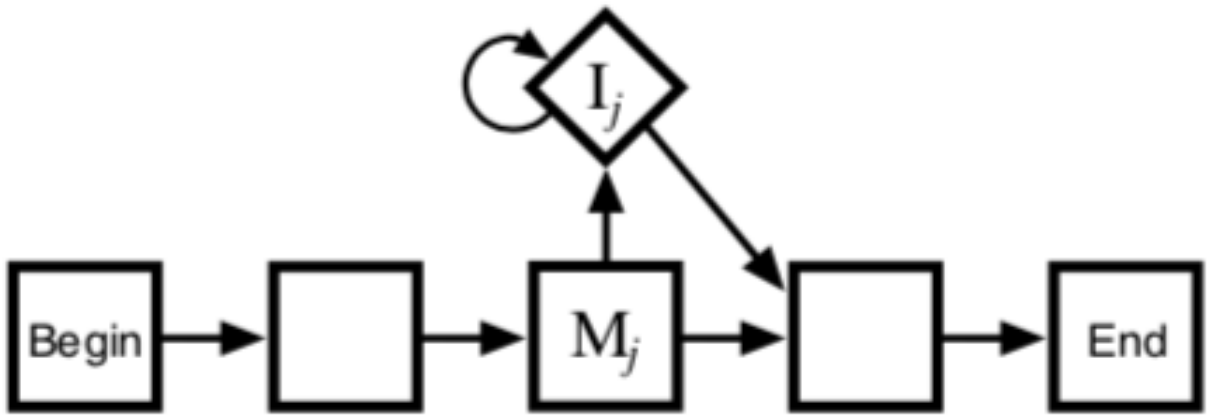
WED

Profile HMM

Insertions

When the query sequence contains a region that is not present in the model, that is an insertion in the query sequence.

Query sequences= **WECIRD**
But the insertion could be anywhere. Hence, we need to add an insert state between any two consecutive match states.



Query sequence=**WECLIRD**
The emission probabilities for the insert states are the background probabilities.

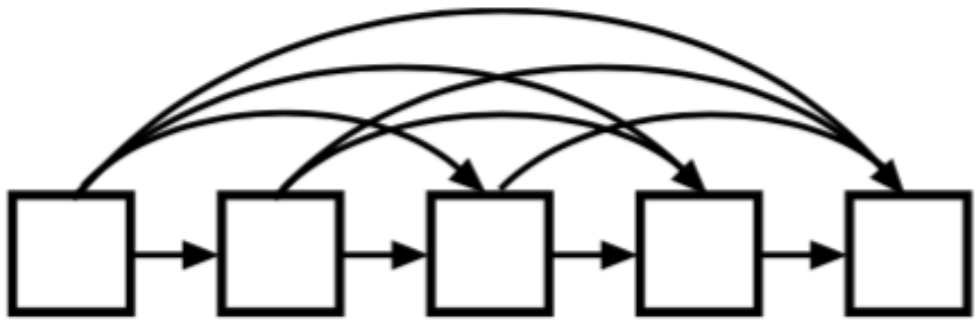
Profile HMM

Deletions

When there is a region in the model that is not present in the query sequence, there is a deletion in the query sequence.

Query sequence= **WEID**
Could be multiple deletions as in **WED**.

So, the deletions could be handled by adding jumps from any match state to any later non-neighbouring match state.

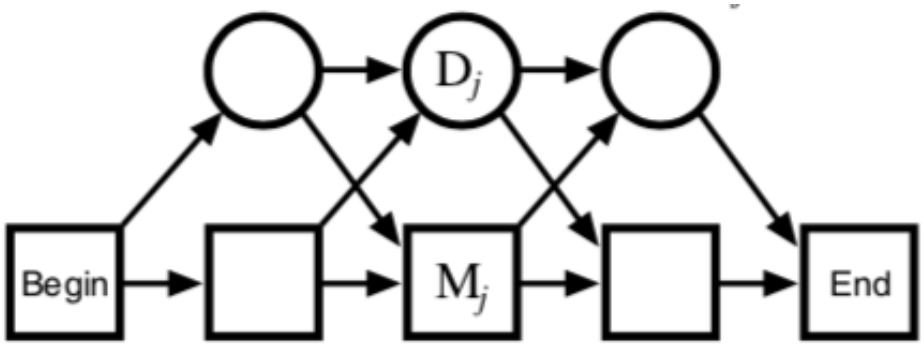


Problem? Give rise to a lot of unknown parameters estimate.

Profile HMM

Silent States

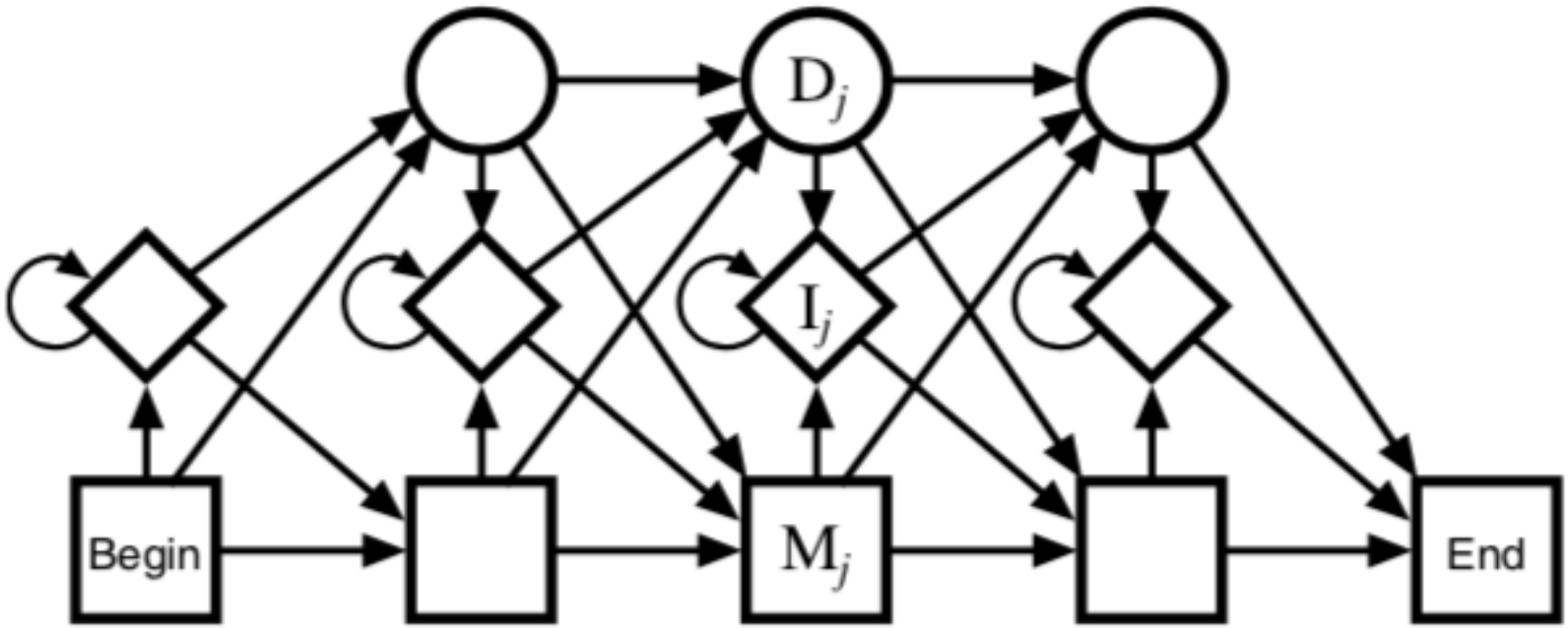
Solution? We use Silent states.



There is a trade-off

Not Possible: probability of transition from state **1** to state **4** is **low** but from state **1** to state **5** is **high** or a model where 1 to 4 is high but 1 to 5 is low.

Profile HMM



Profile HMM

Parameter Estimation

Let us review the whole process in steps:

- The MSA λ of a set of sample sequences from the protein family is provided.
- We choose the most conserved columns $1, 2, \cdots, L$ of the MSA λ as our backbone and define match states $M_1, M_2, \cdots, M_L$.
- Estimate probabilities a_{kl} and $e_k(a)$

$$a_{kl} = \frac{A_{kl} + 1}{\sum_q (A_{kq} + 1)} \qquad \text{and} \qquad e_k(a) = \frac{E_k(a) + 1}{\sum_b [E_k(b) + 1]}$$

where,
- A_{kl}= the count of transitions from k $\Rightarrow$ l in λ
- $E_k(a)$= the count of emissions of 'a' from state k in λ

The additional '1' added in the numerator and denominator is due to Laplace rule of pseudocounts.

Profile HMM

Example of Profile HMM

We are given with a multiple sequence alignment as follows:

```
VG--H
V---N
VE--D
IAADN
```

length of profile HMM = average of the length of the sequences in the MSA.
In this example the lengths are 3,2,3,5(before inserting gaps) whose average is 3.25.

Match states = M_1, M_2, M_3
Insertion states = I_0, I_1, I_2, I_3
Deletion states = D_1, D_2, D_3

Profile HMM

Example of Profile HMM

V	G	-	-	H
V	-	-	-	N
V	E	-	-	D
I	A	A	D	N
M_1	M_2	I_2	I_2	M_3

This gives us the following labeled sequences:

V G H
M_1 M_2 M_3

V – N
M_1 D_2 M_3

V E D
M_1 M_2 M_3

I A A D N
M_1 M_2 I_2 I_2 M_3

Profile HMM

Example of Profile HMM

V	G	H		
M_1	M_2	M_3		
V	–	N		
M_1	D_2	M_3		
V	E	D		
M_1	M_2	M_3		
I	A	A	D	N
M_1	M_2	I_2	I_2	M_3

$$e_{M_1}(V) = \frac{3+1}{(3+1)+(1+1)+(18)}$$

Profile HMM

Example of Profile HMM

V	G	H		
M_1	M_2	M_3		
V	–	N		
M_1	D_2	M_3		
V	E	D		
M_1	M_2	M_3		
I	A	A	D	N
M_1	M_2	I_2	I_2	M_3

$$a_{M_2 I_2} = \frac{1+1}{(2+1)+(1+1)+(0+1)}$$

Viterbi algorithm - optimal alignment, probability of that alignment gives the probability that the sequence belongs to that family.

Forward algorithm- find probability that the query sequence belongs to the protein family irrespective of any alignment.

Thank You!