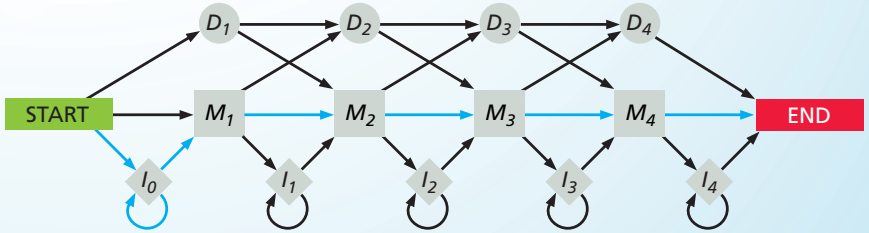
Profile Hidden Markov Models

Hamim Zafar

BSE633



IIT Kanpur
Biological Sciences & Bioengineering

Extending PSSM to Profile HMM

- Although a PSSM captures some conservation information, it is an **inadequate representation** of all the information in a multiple alignment of a protein family
  - Provide **ungapped scores**
- **How to account for gaps?**
  - Allow gaps at each position in the alignment, using the same gap score $\gamma(g)$ at each position
- Alignment gives us explicit indications of where gaps are more and less likely.
  - **position sensitive gap scores**
- **Solution:** build a hidden Markov model (HMM), with a repetitive structure of states, but different probabilities in each position.
  - full probabilistic model for sequences in the sequence family.
  - **Profile HMM**

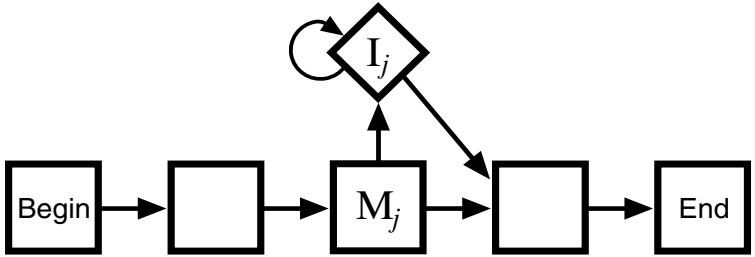
Basic Structure of Profile HMM: Match State

- PSSM can be viewed as a trivial HMM with a series of identical states called **match** states, separated by transitions of probability 1.
- Alignment is trivial because there is no choice of transitions.

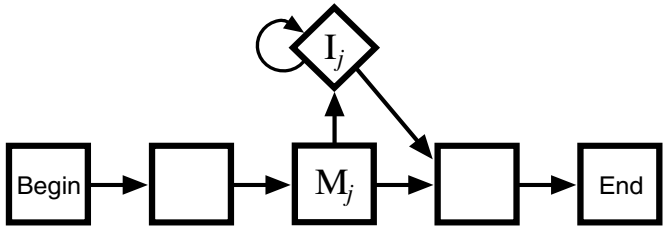
- Emission probabilities for the match states denoted by $e_{M_j}(a)$.
- Match state **matches (i.e., aligns) residues at a specific position** (column) of an alignment.
- linked to the matching state for the next alignment position by a transition that specifies the direction of the connection.
- Residues emitted by match states are related to the HMM, and so the emission probabilities depend on the multiple alignment. They are most closely related to the PSSMs, as they vary with position in the alignment.

Basic Structure of Profile HMM: Insertion State

- To handle **insertions**, i.e. portions of **query sequence** that do not match anything in the model, we introduce a set of new states I_j
- I_j will be used to match insertions after the residue matching the j^{th} column of the multiple alignment.
- I_j have emission distribution $e_{I_j}(a)$, but these are normally set to the background distribution q_a
  - If there is an insertion in the query sequence relative to the HMM, the residues involved are, by definition, not related to residue information in the HMM.



Basic Structure of Profile HMM: Insertion State



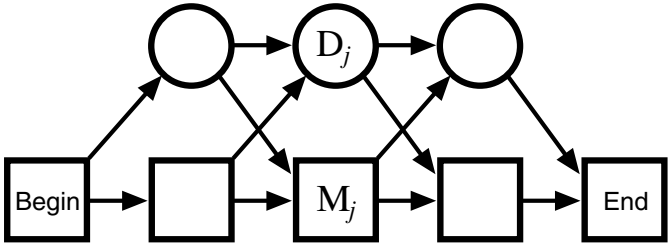
- Three transitions involving I_j
  - from M_j to I_j
  - a loop transition from I_j to itself, to accommodate multi-residue insertions
  - a transition back from I_j to M_{j+1}
- The log-odds cost of an insert is the sum of the costs of the relevant transitions and emissions
- Assuming $e_{I_j}(a) = q_a$, there is no log-odds contribution from the emission, and the score of a gap of length k is

$$\log a_{\mathrm{M}_j\mathrm{I}_j} + \log a_{\mathrm{I}_j\mathrm{M}_{j+1}} + (k-1)\log a_{\mathrm{I}_j\mathrm{I}_j}.$$

Affine gap Scoring model

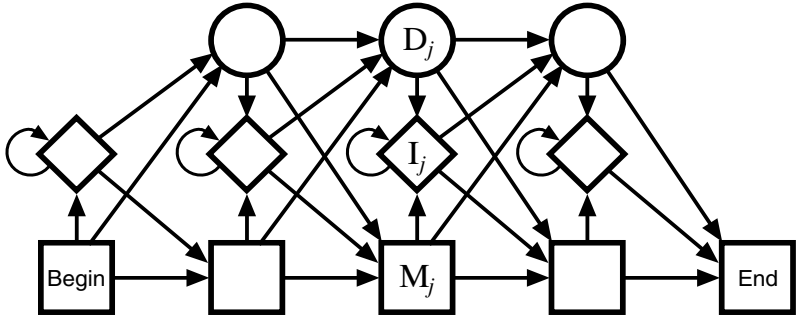
Basic Structure of Profile HMM: Deletion State

- To model **Deletions**, i.e. segments of the multiple alignment that are not matched by any residue in the query sequence, introduce **silent deletion states** D_j
- Because the **silent states do not emit any residues**, it is possible to use a sequence of them to get from any match state to any later one, between two residues in the sequence.



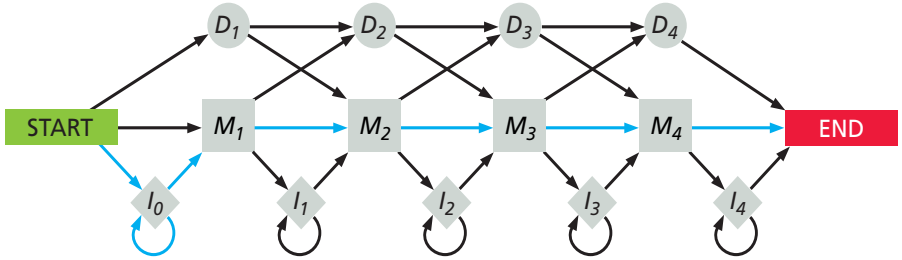
- The **cost of a deletion** is the **sum** of the costs of an $M \rightarrow D$ transition followed by a number of $D \rightarrow D$ transitions, then a $D \rightarrow M$ transition.
- It is possible that the $D \rightarrow D$ transitions will have different probabilities, and hence contribute differently to the score

Complete Structure of Profile HMM

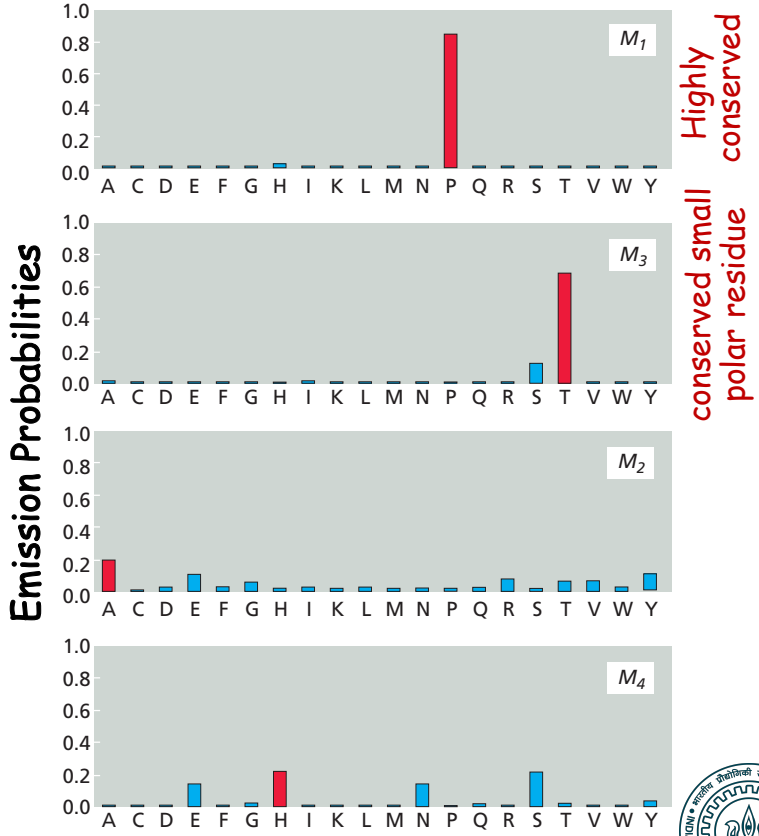


- Transitions between insert and delete states are added but usually very improbable. Leaving them out has negligible effect on scoring a match, but can create problems when building the model.
- Transitions between insert states I_j and I_{j+1} are not allowed because they would imply the absence of a match state between them, and hence delete state D_{j+1} should be involved
- The path taken through the model, resulting in the emission of a sequence, can be interpreted as assigning each of the residues to be either aligned to a particular HMM position (corresponding to the particular match state) or else inserted at a particular position (corresponding to the particular insert state)

Profile HMM: an Example



- Transition to M_1 when query sequence amino-terminal residue matches the first position
- If the query sequence has extra residues at its amino terminus, these will be emitted by the I_0 state
- If first-position residue is missing from the query the transition used will be from start to D_1
- Any extra residues in the query sequence after the profile will be emitted by the last insert state.
- The existence of insert states allows the profile to occur anywhere within a larger sequence.

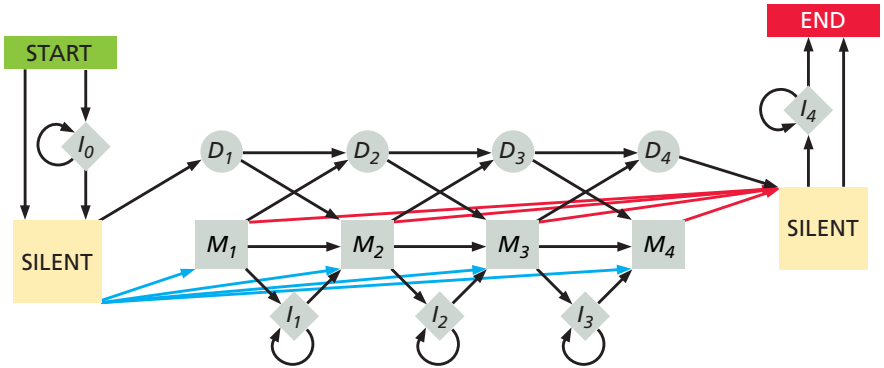


Alignment of Unrelated Sequence to Profile HMM

- Profile HMM represents a profile against which a query sequence is to be aligned.
- We must allow for the query sequence to be wholly unrelated to the profile.
- **Unrelated sequence** would be emitted by paths that visited **many delete and insert** states.
  - The sequence emitted by the match states can be regarded as aligned to the profile HMM, and any residues emitted by insert states are unaligned by definition.
- An **unrelated sequence** will align to a profile HMM with a **very low probability or log-odds score.**
- Unrelated sequences will usually be identified on these quantitative scores rather than a qualitative assessment of the alignment itself.

Profile HMM for Local Profile Alignment

- If the model is parameterized for sequences with the full-length profile, alignment of only a part of the profile will involve a path through many delete states, which will occur with very low probability



- Two new silent states directly connected to every match state of the model
- The blue and red transitions connect the silent states to all match states, allowing only a part of the profile to be included in the path.

- I_0 state models all the sequence prior to the profile, and the I_4 state all the sequence after it.
- Models flanking sequences of the query.

Profile HMM Parameterization using Aligned Sequences

Goal: Estimate the emission and transition probabilities and length of a profile HMM from a multiple sequence alignment

- A profile HMM can model any possible sequence of residues from the given alphabet
- It defines a probability distribution over the whole space of sequences.
- The aim of the parameterization process it to make this distribution peak around members of the family.
- The choice of length of the model corresponds more precisely to a decision on which multiple alignment columns to assign to match states, and which to assign to insert states.

Profile HMM Parameterization: Assigning States

- An alignment column that contains **no gaps** should be assigned to a **match state**
- One with a **majority of gaps** should be assigned to an **insert state**
- A threshold proportion of gaps is selected to determine the assignments.
  - columns that are more than half gap characters are modelled by inserts

```
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...
                  ***   *****
```

Starred columns will be treated as 'matches' in the profile HMM.

Profile HMM Parameterization: Assigning Probabilities

- After match and insert state assignments, the path that each sequence follows through the HMM can be deduced.
- The alignment data is translated into the frequencies of transitions between particular states along the path and the frequencies of emission of individual residues from particular states.
- Count up the number of times each transition or emission is used, and assign probabilities according to

$$a_{kl} = \frac{A_{kl}}{\sum_{l'} A_{kl'}} \quad \text{and} \quad e_k(a) = \frac{E_k(a)}{\sum_{a'} E_k(a')}$$

- k and l are indices over states, and a_{kl} and e_k are the transition and emission probabilities, and A_{kl} and E_k are the corresponding frequencies.
- Emission probabilities for insert states are based on the overall amino acid composition of the dataset (background probability)

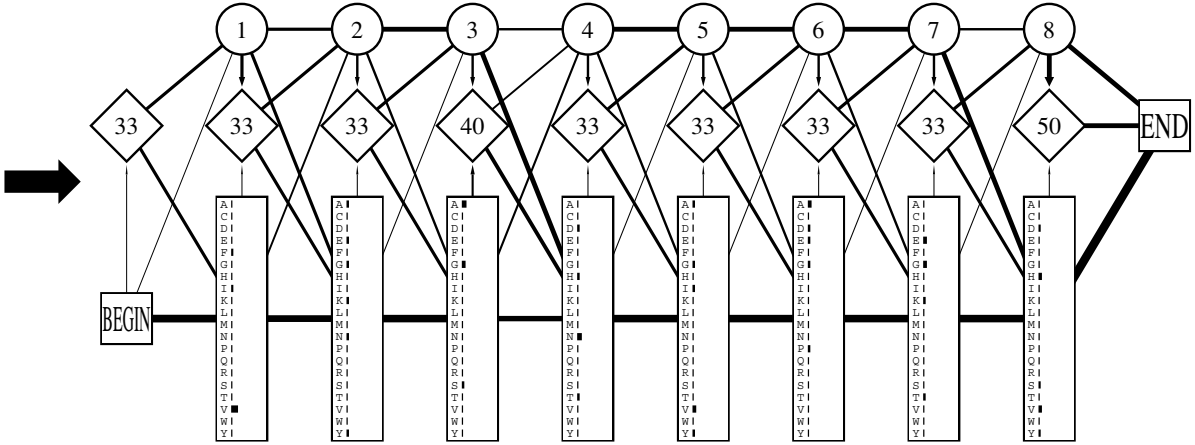
$$e_{I_j}(a) = q_a$$

Profile HMM Parameterization: Small Training Data

- Some transitions or emissions may not be seen in the training alignment containing small number of sequences
  - Poor estimates of the probabilities. Zero probabilities for transitions or emissions not observed, which would mean they would never be allowed in the future.
  - especially the case for emission parameters
- For protein sequence HMMs a minimum of 20 sequences is required simply to observe each possible match state emission, but many more are required to obtain realistic estimates of the emission probabilities.
- An HMM that has zero probability for emission of a particular residue from a given match state cannot align a sequence with that residue at that position.
- Even when a relatively large quantity of sequence data is available, parameterization problems occur for positions with very highly conserved residues.
- Solution: Pseudocounts
  - Remember Laplace's rule: to add 1 to each frequency.

Profile HMM Parameterization: An Example

```
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...
                ***  *****
```



Profile HMM derived using SAM package (Laplace's rule). Emission probabilities are shown as bars opposite the different amino acids for each match state, and transition probabilities are indicated by the thickness of the lines. The $I \rightarrow I$ transition probabilities times 100 are shown in the insert states.

- $e_{M_1}(V) = \frac{6}{27}$,
- $e_{M_1}(I) = e_{M_1}(F) = 2/27$
- $e_{M_1}(a) = 1/27$ for all residue types a other than V, I, F.
- $a_{M_1M_2} = \frac{7}{10}, a_{M_1D_2} = \frac{2}{10}$
- $a_{M_1I_1}$ =1/10
- Following column 1 there are six transitions from match to match, one transition to a delete state, in HBB_HUMAN, and no insertions

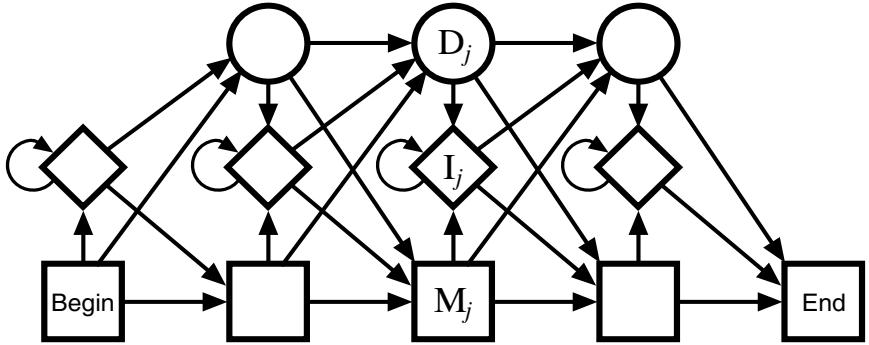
Scoring a Sequence Against a Profile HMM

- Profile HMMs are used to **detect potential membership in a family** by obtaining significant matches of a sequence to the profile HMM.
- Given a parameterized profile HMM, any given path through the model will emit a sequence with an associated probability. This path probability will be the product of all the transition and emission probabilities along the path.
- In addition, the path defines how the emitted sequence is aligned to the model.
- Two choices of scoring a sequence $\boldsymbol{x}$
  - Viterbi equations to give the most probable alignment $\boldsymbol{\pi}^*$ of a sequence $\boldsymbol{x}$ together with its probability $\boldsymbol{P(x, \pi^* | M)}$
  - forward equations to calculate the full probability of $\boldsymbol{x}$ summed over all possible paths $\boldsymbol{P(x|M)}$.
- We want to consider the log-odds ratio of the resulting probability to the probability of $\boldsymbol{x}$ given a standard random model

$$P(x|R) = \prod_i q_{x_i}.$$

Most Probable Path using Viterbi Algorithm

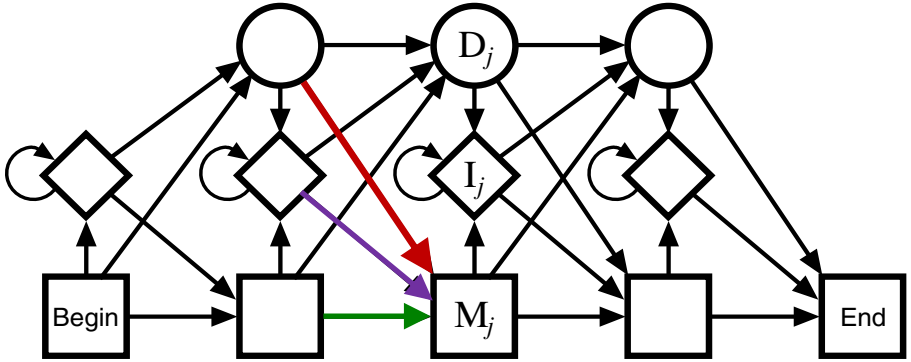
- Profile HMM with an arbitrary number of sets of match M, delete D, and insert I states.
- $\boldsymbol{V_j^M(i)}$: log-odds score of the best path matching subsequence $x_{1\ldots i}$ to the submodel up to state j, ending with x_i being emitted by state M_j



- $\boldsymbol{V_j^I(i)}$: log-odds score of the best path ending with x_i being emitted by state I_j
- $\boldsymbol{V_j^D(i)}$: log-odds score of the best path ending in state D_j

Most Probable Path using Viterbi Algorithm

- Recurrence Relation for $V_j^M(i)$
- $V_j^M(i)$: log-odds score of the best path matching subsequence $x_{1\ldots i}$ to the submodel up to state j, ending with x_i being emitted by state M_j



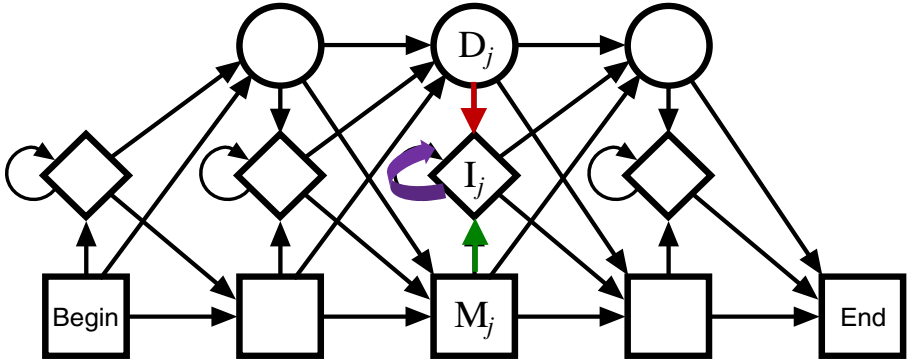
$$V_j^{\mathrm{M}}(i) = \log \frac{e_{\mathrm{M}_j}(x_i)}{q_{x_i}} + \max \begin{cases} V_{j-1}^{\mathrm{M}}(i-1) + \log a_{\mathrm{M}_{j-1}\mathrm{M}_j}, \\ V_{j-1}^{\mathrm{I}}(i-1) + \log a_{\mathrm{I}_{j-1}\mathrm{M}_j}, \\ V_{j-1}^{\mathrm{D}}(i-1) + \log a_{\mathrm{D}_{j-1}\mathrm{M}_j}; \end{cases}$$

Emission

Possible Transitions

Most Probable Path using Viterbi Algorithm

- Recurrence Relation for $V_j^I(i)$
- $V_j^I(i)$: log-odds score of the best path ending with x_i being emitted by state I_j



Probabilities cancel
No emission score

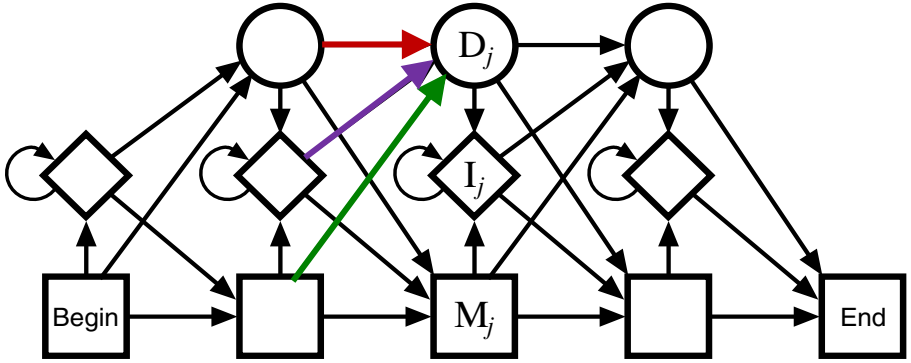
$$V_j^{\mathrm{I}}(i) = \log \frac{e_{\mathrm{I}_j}(x_i)}{q_{x_i}} + \max \begin{cases} V_j^{\mathrm{M}}(i-1) + \log a_{\mathrm{M}_j \mathrm{I}_j}, \\ V_j^{\mathrm{I}}(i-1) + \log a_{\mathrm{I}_j \mathrm{I}_j}, \\ V_j^{\mathrm{D}}(i-1) + \log a_{\mathrm{D}_j \mathrm{I}_j}; \end{cases}$$

Emission

Possible Transitions

Most Probable Path using Viterbi Algorithm

- Recurrence Relation for $V_j^D(i)$
- $V_j^D(i)$: log-odds score of the best path ending in state D_j



Silent State
No Emission

$$V_j^{\mathrm{D}}(i) = \max \begin{cases} V_{j-1}^{\mathrm{M}}(i) + \log a_{\mathrm{M}_{j-1}\mathrm{D}_j}, \\ V_{j-1}^{\mathrm{I}}(i) + \log a_{\mathrm{I}_{j-1}\mathrm{D}_j}, \\ V_{j-1}^{\mathrm{D}}(i) + \log a_{\mathrm{D}_{j-1}\mathrm{D}_j}. \end{cases}$$

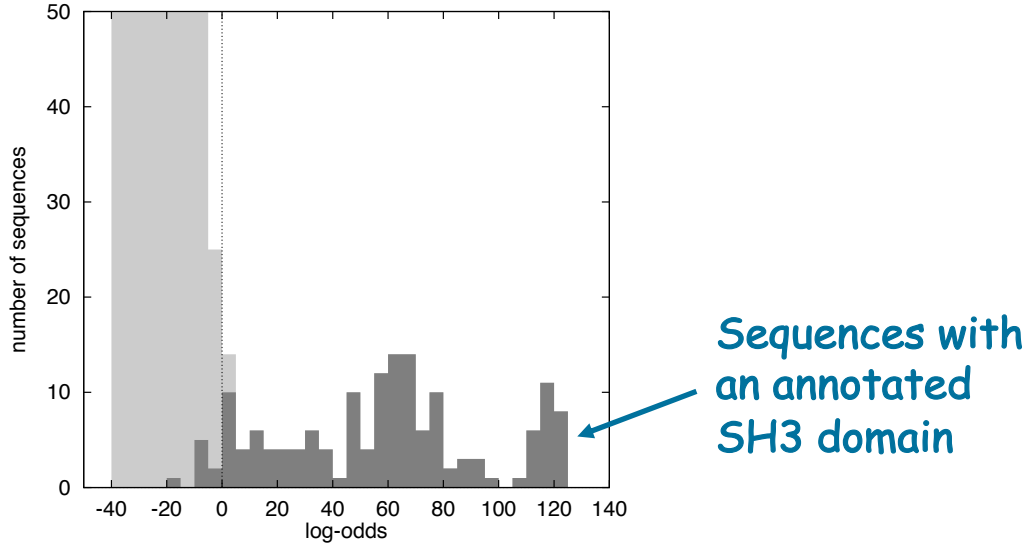
Possible Transitions

Most Probable Path using Viterbi Algorithm

- Initialization and Termination:
- We want to allow the alignment to start and end in a delete or insert state, in case the beginning or end of the sequence does not match the first or the last match state of the model
- Rename the Begin state as M_0 and set $V_0^M(0) = 0$
- We then allow transitions to I_0 and D_1 .
- At the end we can collect together possible paths ending in insert and delete states by renaming the End state to M_{L+1} and using the top relation without the emission term to calculate $V_{L+1}^M(n)$ as the final score.

Searching a Database using Profile HMM

- Once we find the most probable path of a sequence in the profile HMM using Viterbi Algorithm, we also get the probability of the sequence for that alignment
- The log-odds score found in this manner can be used to search databases for members of the same family.



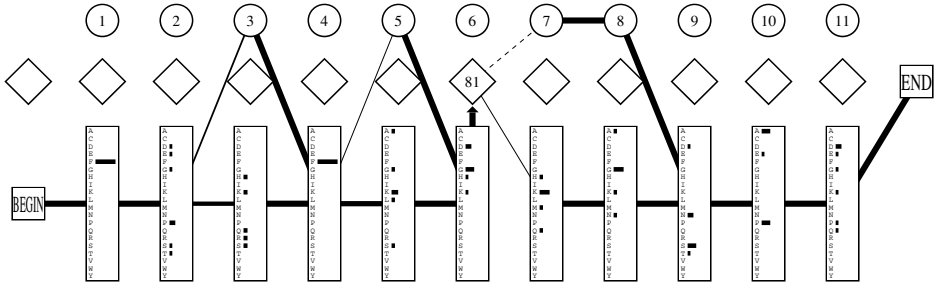
Distribution of log-odds scores from a search of Swissprot with a profile HMM of the SH3 domain

Multiple Alignment with a Known Profile HMM

- This problem arises in sequence analysis, when we have a multiple alignment and a model of a small representative set of sequences in a family, and we wish to use that model to align a large number of other family members together.
- Constructing a multiple alignment requires calculating a Viterbi alignment for each individual sequence.
- Residues aligned to the same profile HMM match state are aligned in columns.
- An important difference between profile HMM multiple alignments and traditional multiple alignments

Multiple Alignment with a Known Profile HMM

An Example



Profile HMM

Alignment

```
FPHF-DLS-----HGSAQ
FESFGDLSTPDAVMGNPK
FDRFKHLKTEAEMKASED
FTQFAG-KDLESIKGTAP
FPKFKGLTTADQLKKSAD
FS-FLK-GTSEVPQNNPE
FG-FSG----AS---DPG
```

Inserts

The same seven sequences were realigned to the model

Multiple Alignment with a Known Profile HMM

Original Alignment

```
FPHF-DLS-----HGSAQ
FESFGDLSTPDAVMGNPK
FDRFKHLKTEAEMKASED
FTQFAG-KDLESIKGTAP
FPKFKGLTTADQLKKSAD
FS-FLK-GTSEVPQNNPE
FG-FSG----AS--DPG
```

Most Probable Paths through the Model

Position	1	2	3	4	5	6	insert	7	8	9	10	11
	F	P	H	F	–	D	LS	H	G	S	A	Q
	F	E	S	F	G	D	LSTPDAV	M	G	N	P	K
	F	D	R	F	K	H	LKTEAEM	K	A	S	E	D
	F	T	Q	F	A	G	KDLESI	K	G	T	A	P
	F	P	K	F	K	G	LTTADQL	K	K	S	A	D
	F	S	–	F	L	K	GTSEVP	Q	N	N	P	E
	F	G	–	F	S	G	AS	–	–	D	P	G

After Realignment

```
FPHF-Dls.....HGSAQ
FESFGDlstpdavMGNPK
FDRFKHlkteaemKASED
FTQFAGkdlesi.KGTAP
FPKFKGlttadqlKKSAD
FS-FLKgtsevp.QNNPE
FG-FSGas.....--DPG
```

- Original alignment and the new alignment are the same
- A profile HMM does not attempt to align the residues assigned to insert states.
- The insert state residues usually represent parts of the sequences which are atypical, unconserved, and not meaningfully alignable.
  - biologically realistic view of multiple alignment
- In contrast, many other multiple alignment algorithms align the whole sequences, regardless of what parts of the sequence are meaningfully alignable or not.

HMMER: biosequence analysis using profile hidden Markov models

Get the latest version

v3.3

Download source

(archived older versions)

HMMER is used for searching sequence databases for sequence homologs, and for making sequence alignments. It implements methods using probabilistic models called profile hidden Markov models (profile HMMs).

HMMER is often used together with a profile database, such as Pfam or many of the databases that participate in Interpro. But HMMER can also work with query *sequences*, not just profiles, just like BLAST. For example, you can search a protein query sequence against a database with **phmmer**, or do an iterative search with **jackhmmer**.

HMMER is designed to detect remote homologs as sensitively as possible, relying on the strength of its underlying probability models. In the past, this strength came at significant computational expense, but as of the new HMMER3 project, HMMER is now essentially as fast as BLAST.

HMMER can be downloaded and installed as a command line tool on your own hardware, and now it is also more widely accessible to the scientific community via new search servers at the European Bioinformatics Institute.

PERFORM A SEARCH

An online interactive search service is available at the European Bioinformatics Institute. Go there to search against the latest Uniprot databases.

DOCUMENTATION

The HMMER User's Guide: [PDF].

NEWS

See the blog Cryptogenomicon for more information and discussion about HMMER3.

ebi.ac.uk/Tools/hmmer/

EMBL-EBI | Services | Research | Training | About us | EMBL-EBI Hinxton

HMMER

Biosequence analysis using profile hidden Markov Models

Home | Search | Results | Software | Help | About | Contact

Quick search

Paste in your sequence or use the example

Reference Proteomes | UniProtKB | SwissProt | Pfam

Submit | Reset

Alternative search options

The HMMER web server: fast and sensitive homology searches. This site has been designed to provide near **interactive searches** for most queries, coupled with **intuitive and interactive results** visualisations.

Quickstart tutorial | Online documentation

Reading Materials

- Chapter 5 and 6 of the book "Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids", by Durbin et al., Cambridge University Press.
- Chapter 6 of the book "Understanding Bioinformatics", by M. Zvelebil and J.O. Baum. Published by Garland Science, 2008.
- Krogh, Anders. "An Introduction to Hidden Markov Models for Biological Sequences" In Computational Methods in Molecular Biology, edited by S. L. Salzberg, D. B. Searls and S. Kasif, pages 45-63. Elsevier, 1998.