**Profile HMM (Haussler et al., 1993)**

**Motivation**

- Biological sequences are typically grouped into families with a certain functionality
- A relevant task is that of detecting whether a target sequence belongs to a certain family
- This could be done aligning the sequence to each of the sequences from the family
- However, pairwise alignments alone can miss cases of distantly related sequences
- A better way to detect such relationship would be:

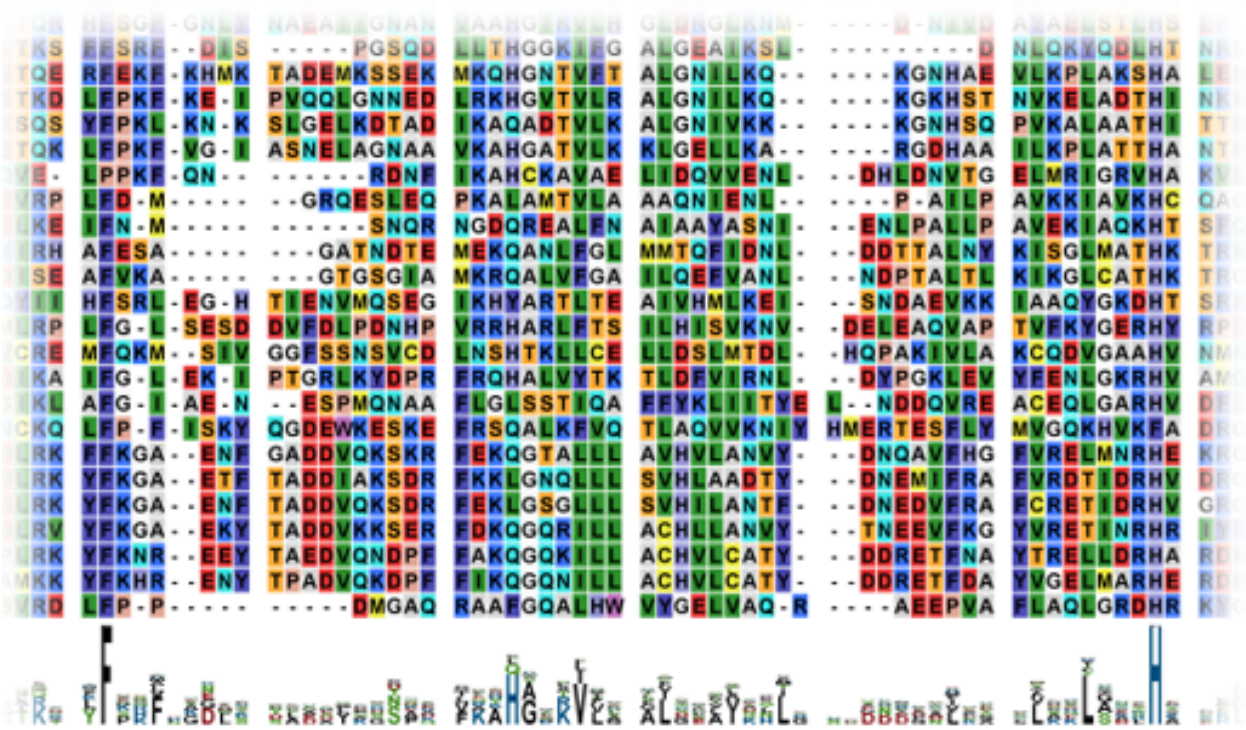
    1. building a model of the family
    2. testing whether the target sequence is compatible with the model

**Profile HMM**

**Multiple alignments**

- A multiple alignment consists of the simultaneous alignment of a set of sequences
- All sequences from a certain family could be aligned to form a multiple alignment representing the family
- The family model should be a compact probabilistic representation of such multiple alignment

**Multiple alignment: example for the globin family**

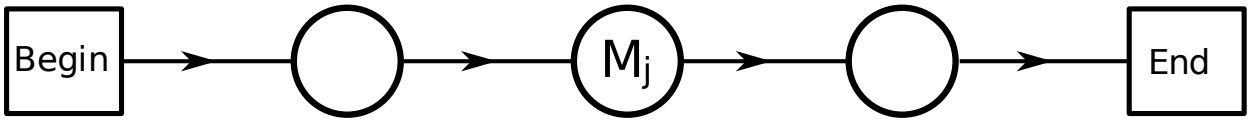


**Profile HMM**

**Dealing with ungapped regions**

- Large portions of multiple alignments for a protein family consist of ungapped sequences of residues
- Each position in such regions has a certain amino-acid *profile*, representing the frequencies with which each amino-acid occurs in the column of the alignment
- By normalizing such profiles, it is possible to derive a probability of observing a certain residue *in that position*.

**Profile HMM**




**Probabilistic model for ungapped regions**

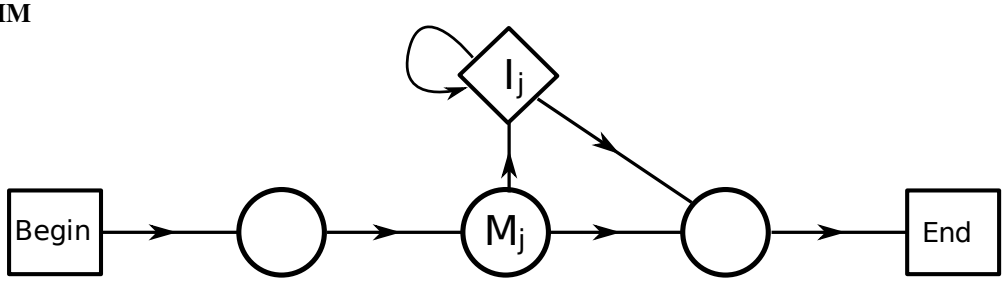
- Each position in the region can be modelled with a match state with position specific emission probabilities
- The whole region can be modelled as a sequence of match states, with transitions only between successive states
- Beginning and end of the region can be modelled with special non-emitting begin and end states

**Profile HMM**

**Dealing with gaps**

- Gaps in the alignment tend to occur at certain positions (i.e. gaps align columnwise)
- Gaps can be dealt with by modelling the two type of corresponding modifications:
  - insertions of a sequence of residues
  - deletions of a sequence of residues

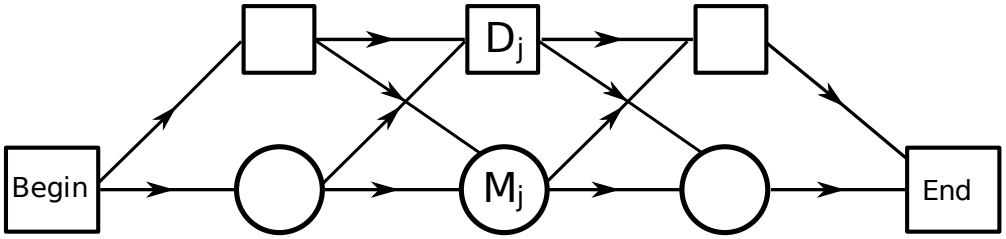
**Profile HMM**




**Probabilistic model with insertions**

- An insertion should be modelled with a specific insertion state $I$ (represented as a diamond)
- As insertions in different positions have different probabilities, transition probabilities should be position specific
- An insertion state should also have a self transition to account for insertions of sequences of residues
- Emission probabilities could instead be set for all insertion states equal to the background probability $q_a$ of observing a certain amino-acid $a$ in an arbitrary sequence.

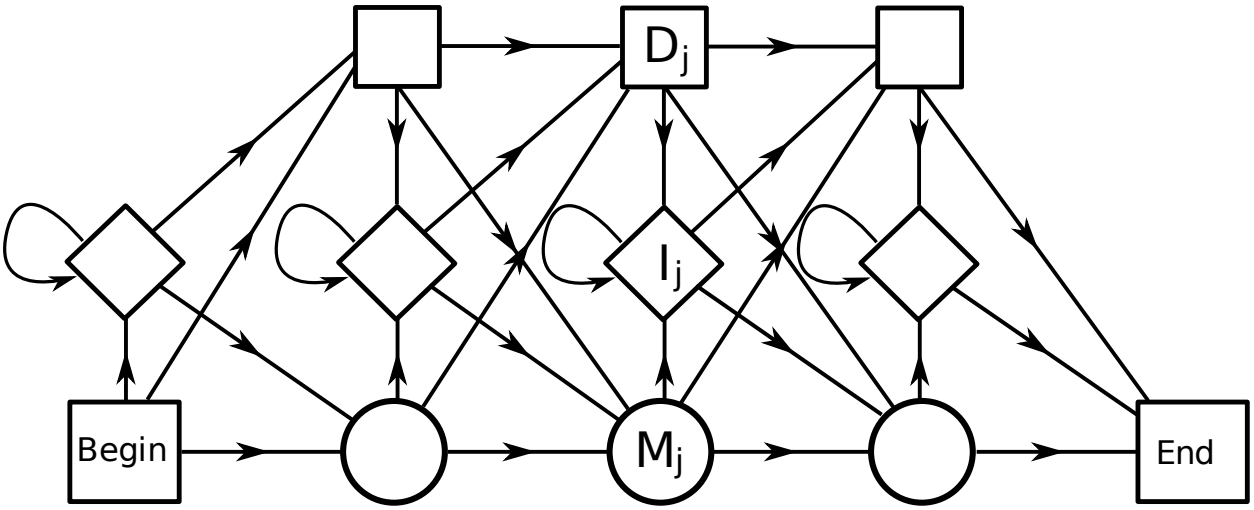
**Profile HMM**




**Probabilistic model with deletions**

- Deletions should be modelled as special *silent* states $D$ which do not emit symbols (represented as a square)
- Allowing self transitions as in insertions would complicate inference algorithms
- Sequences of deletions are instead modelled as sequences of deletion states
- This also allows to specify different transition probabilities between deletion states

**Profile HMM: full model**




*Note*

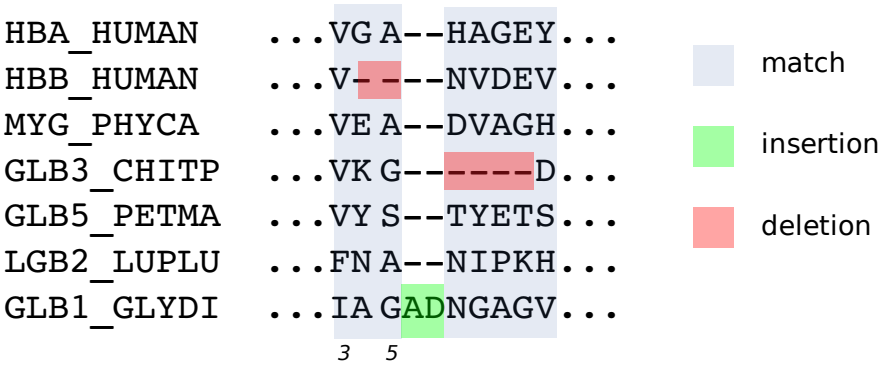
- We allow direct transitions between insertion and deletion states
- These situations are quite rare, but leaving such transitions out would give zero probability to these cases

**Profile HMM**
**Parameter estimation**

- We assume a multiple alignment profile for the family of interest is available (created with multiple alignment algorithms, possibly relying on 3D information)

- We need to estimate transition probabilities between states, and emission probabilities for match states (those for insertion states are set to background probabilities for arbitrary sequences)
- We first decide which positions in the alignment correspond to match states, and which to insertions or deletions:
  - A reasonable approach is that if half of the column elements in a position are gaps, the position is **not** a match state
- This allows to turn our alignment in a fully observed set of training examples: probabilities can be estimated from counts

**Profile HMM**




**Parameter estimation: examples**

- Non-zero emission probabilities for match state $M_3$:

$$e_{M_3}(V) = 5/7 \quad e_{M_3}(F) = 1/7 \quad e_{M_3}(I) = 1/7$$

- Non-zero transition probabilities from match state $M_3$:

$$a_{M_3M_4} = 6/7 \quad a_{M_3D_4} = 1/7$$

- Non-zero transition probabilities from match state $M_5$:

$$a_{M_5M_6} = 5/7 \quad a_{M_5I_5} = 1/7 \quad a_{M_5D_6} = 1/7$$

**Profile HMM**
**Parameter estimation: adding pseudocounts**

- All transitions and emissions never observed in the multiple alignment will be set to zero using only counts.
- This can be a problem if an unsufficient number of examples is available (i.e. always)
- A simple solution consists of adding a non-zero prior probability for any transition or emission, to be combined to the counts observed on data
- Such prior probability can be thought of coming from *pseudocounts* of hypothetical observations of emissions/transitions
- The simplest pseudocount (Laplace smoother) consists of adding a single hypothetical observation of any possible emission/transition