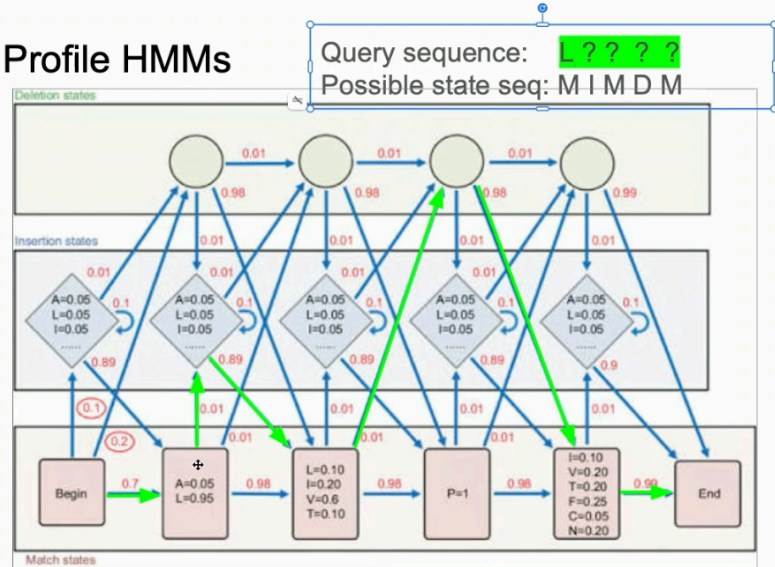


# 10/15/2026

- Multiple seq alignment
  - one comparison provides limited info
- naive MSA approach is not scalable - DP table gets too big
  - $O((2n)^k)$
- So we use heuristics
- Progressive alignment - uses UPGMA or neighbor joining to build tree
  - sequential pairwise
  - CLUSTAL
- AlphaFold2
  - deep learning model that predicts protein's 3d conformation from primary structure
  - Why is MSA useful in protein structure prediction?
    - MSA module finds conserved regions of amino acids between similar proteins of other species → if we know their 3d structure, this protein's 3d structure is probably similar
  - How do conserved regions inform protein struct
    - co-evolution: if 2 AAs sit close together in 3d protein, mutation in one often forces a matching mutation in the other to maintain structural stability
  - MSA in alphafold 2 works through HMM
    - profile HMM is trained to output sequences from protein family
    - Given arbitrary sequence, what is the chance that this model generated it?
      - given new sequence and past sequences from MSA, what is the most likely series of states that generated it

## Deeper look at Profile HMMs

- Think of a profile HMM as a sequence generating model
- We “train” this model by feeding it sequences from an MSA
- Given a new sequence, what is the most likely series of states that generated it?



- Alphafold2 queries a gene database across species to generate a reference MSA comprised of sequences similar to the input sequence
- Reference MSA is used to train profile HMM
  - profile HMM is collection of states and the possible transitions between them happening at certain probabilities
- Alignment can be thought of as sequence of transitions through the profile HMM, which contains states matching (mis)matches, insertions, and deletions
- Probabilities are calculated from relative frequencies