# Notes on using HMM forProtein Family Classification

- Pfam
- large collection of multiple sequence alignments and hidden Markov models

(http://www.sanger.ac.uk/Software/Pfam/index.shtml)

- covers many common protein domains and families
- Based on a statistical model (HMM) for a group of related protein sequences (e.g protein family)

# Profile Hidden Markov Models

- Basic tool in sequence analysis
- Look more complicated than they really are
- Used to model a family of sequences
- Can be built from a multiple sequence alignment
- Algorithms using profile HMMs are based on dynamic programming (much like Needleman-Wunsch)
- Given a Profile HMM computed for a multiple sequence alignment, you can use it to
  - Recognize related sequences
  - Add related sequences into the multiple sequence alignment

# Using profile HMMs to align sequences

- Given an MSA for a set **S** of representative sequences for a gene and set **X** of additional sequences (query sequences)
  - You build a profile HMM for the MSA on S.
  - For each **s in X** you find for the gene, you find the path through the profile HMM that is most likely to generate **s**.
  - The path specifies how to add the sequence into the MSA (only the match states count).
  - Transitivity gives you the final MSA after you add in all the other sequences.

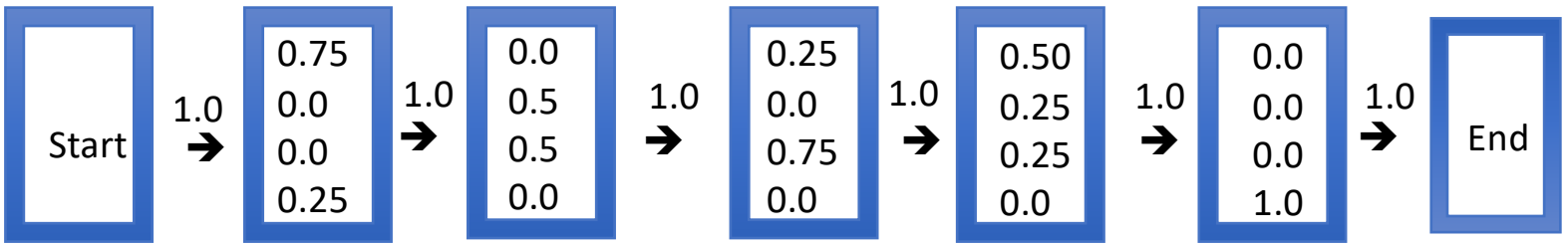
# Profile

- Given a gap-less multiple sequence alignment, we can build a profile describing what we see

- S1 = A C T A G
- S2 = A C A A G
- S3 = A T T T G
- S4 = G T T C G

| | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| A | 0.75 | 0.0 | 0.25 | 0.50 | 0.0 |
| C | 0.00 | 0.5 | 0.00 | 0.25 | 0.0 |
| T | 0.00 | 0.5 | 0.75 | 0.25 | 0.0 |
| G | 0.25 | 0.0 | 0.00 | 0.00 | 1.0 |

The profile yields a probability distribution of sequences – here, all of the same length.



# Adding in insertions

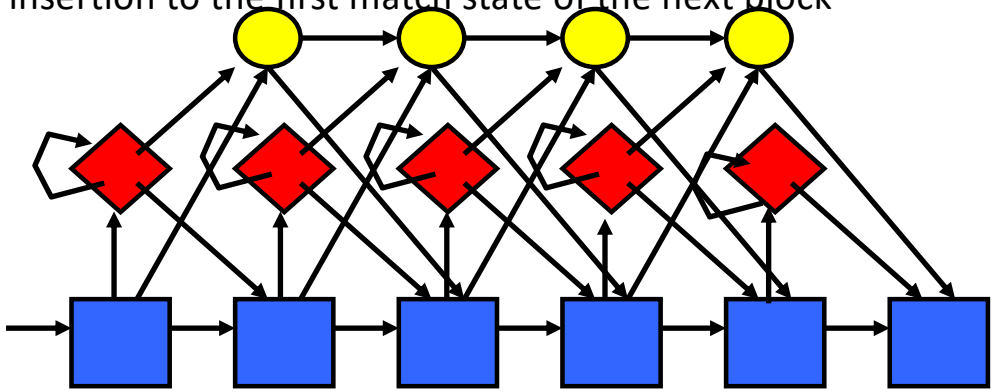
- The profile shown in the previous slide only had *match* states (indicated by rectangles). It doesn't allow any insertions or deletions.
- To model indels, we just have to add additional states to the graphical model.
  - Insertion states: Diamonds (have non-zero emission probabilities)
  - Deletion states: Circles (nothing emitted)

# Adding Insertions/Deletions to HMMs

- For a profile HMM, insertions and deletions are treated separately

- For insertions a new set of insert states is inserted, denoted by diamonds

- Transitions are needed from:
  - last match state in a block to the insertion
  - insertion to itself (to allow for multiple length insertions)
  - insertion to the first match state of the next block



```
HAKVPRG
HAR--DH
HDAV-MG
HYR--PD
FAD--RG
HFY--RG
HAK-PVL
HRKG-YG
HEKGGRG
HKP--RN
```

# Other uses of profile HMMs

- Given two profile HMMs (H1 and H2), and a sequence s, you can determine which one is more likely to generate s (again, using dynamic programming).
- Note that computing the probability that a profile HMM generates a sequence requires calculating the probability for *every* path and adding up the probabilities. This can be calculated in polynomial time, using dynamic programming.