# Introduction to Profile HMMs

Tandy Warnow

# 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)

# 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

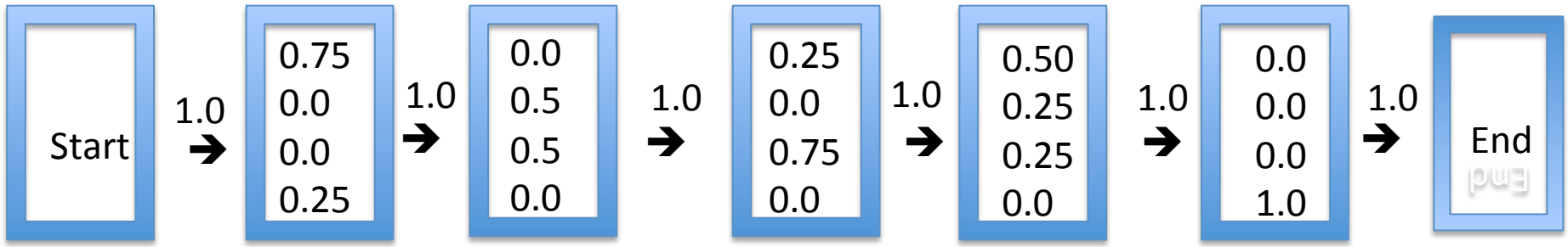
| | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| A | 0.75 | 0.0 | 0.25 | 0.50 | 0.0 |
| C | 0.00 | 0.5 | 0.00 | 0.25 | 0.0 |
| T | 0.00 | 0.5 | 0.75 | 0.25 | 0.0 |
| G | 0.25 | 0.0 | 0.00 | 0.00 | 1.0 |

# Dealing with zero-probability emissions

- Note that our profiles had some letters that had zero probability.
- Most profile HMMs don't allow zero-probability emissions for letters in the alphabet.
- One way of dealing with this is the "add-one" rule (see Mona Singh's lecture), but there are others.
- When you don't modify your empirical distribution at all to create the emission probabilities, this is called an "unadjusted profile".

# Using a profile

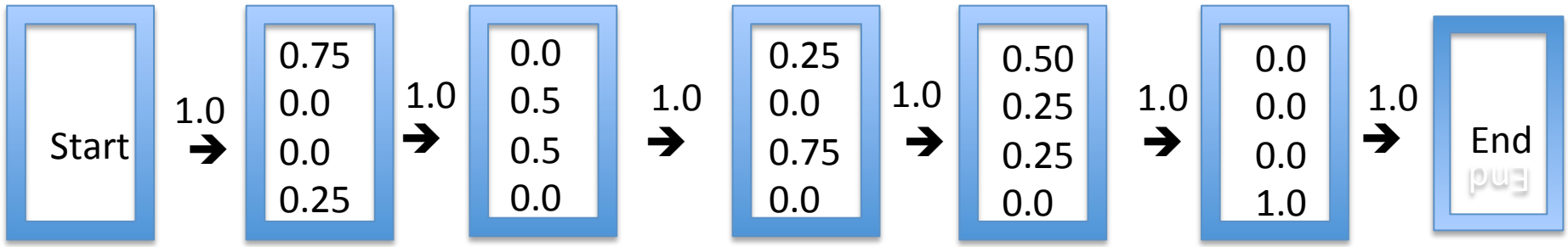
| | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| A | 0.75 | 0.0 | 0.25 | 0.50 | 0.0 |
| C | 0.00 | 0.5 | 0.00 | 0.25 | 0.0 |
| T | 0.00 | 0.5 | 0.75 | 0.25 | 0.0 |
| G | 0.25 | 0.0 | 0.00 | 0.00 | 1.0 |




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

# What is the probability of generating s = ATATG?

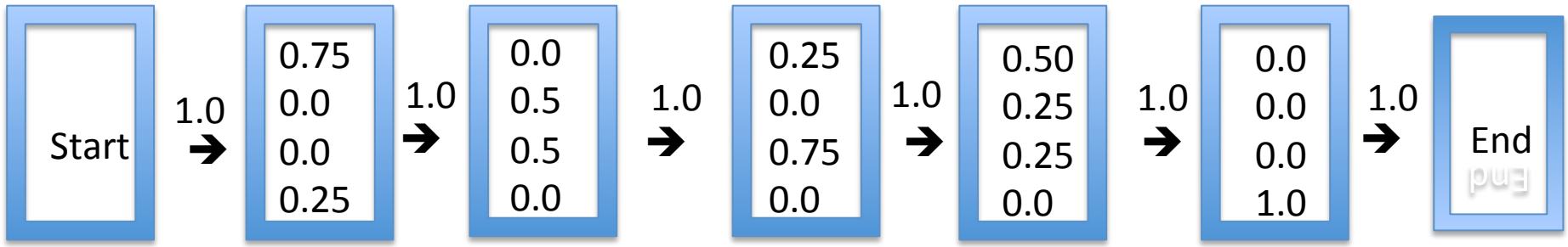
| | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| A | 0.75 | 0.0 | 0.25 | 0.50 | 0.0 |
| C | 0.00 | 0.5 | 0.00 | 0.25 | 0.0 |
| T | 0.00 | 0.5 | 0.75 | 0.25 | 0.0 |
| G | 0.25 | 0.0 | 0.00 | 0.00 | 1.0 |




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

# What is the probability of generating s = AAAAA?

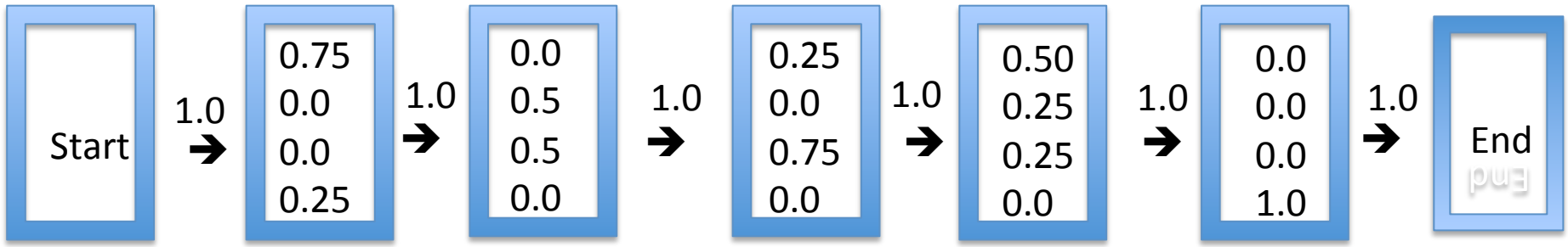
| | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| A | 0.75 | 0.0 | 0.25 | 0.50 | 0.0 |
| C | 0.00 | 0.5 | 0.00 | 0.25 | 0.0 |
| T | 0.00 | 0.5 | 0.75 | 0.25 | 0.0 |
| G | 0.25 | 0.0 | 0.00 | 0.00 | 1.0 |




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

# What are the most probable sequences for this model?

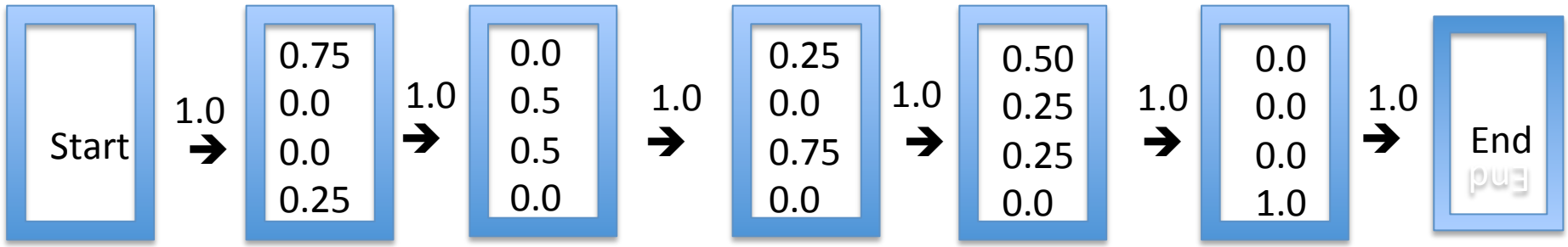
| | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| A | 0.75 | 0.0 | 0.25 | 0.50 | 0.0 |
| C | 0.00 | 0.5 | 0.00 | 0.25 | 0.0 |
| T | 0.00 | 0.5 | 0.75 | 0.25 | 0.0 |
| G | 0.25 | 0.0 | 0.00 | 0.00 | 1.0 |




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

# How would you compute the probability of generating a sequence s?

| | 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.

# Notes

- Each profile generates strings of exactly the same length. This is not very useful for dealing with biological data.

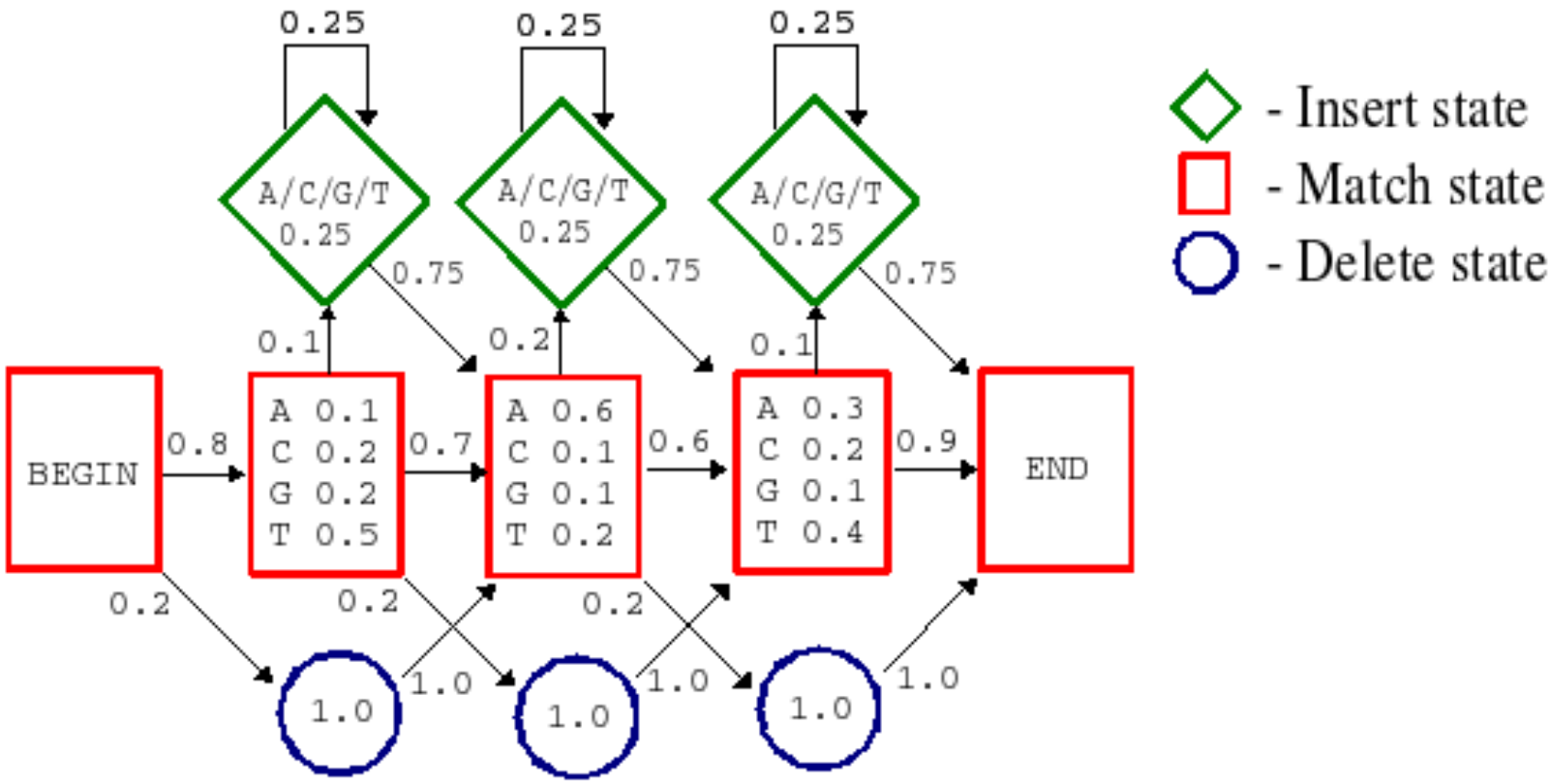
- If you are given a sequence that is generated by the profile, then you know exactly what path was used to generate the sequence. That is, none of the states are hidden.

# Insertions and Deletions (Indels)

- Insertions: events that increase the sequence length, such as:
  - AAAAA -> AATCGAAA
  - AAAAA -> AAAAATCGATTA
- Deletions: events that decrease the sequence length, such as:
  - ATCGA -> AGA
  - AAAAA -> AA

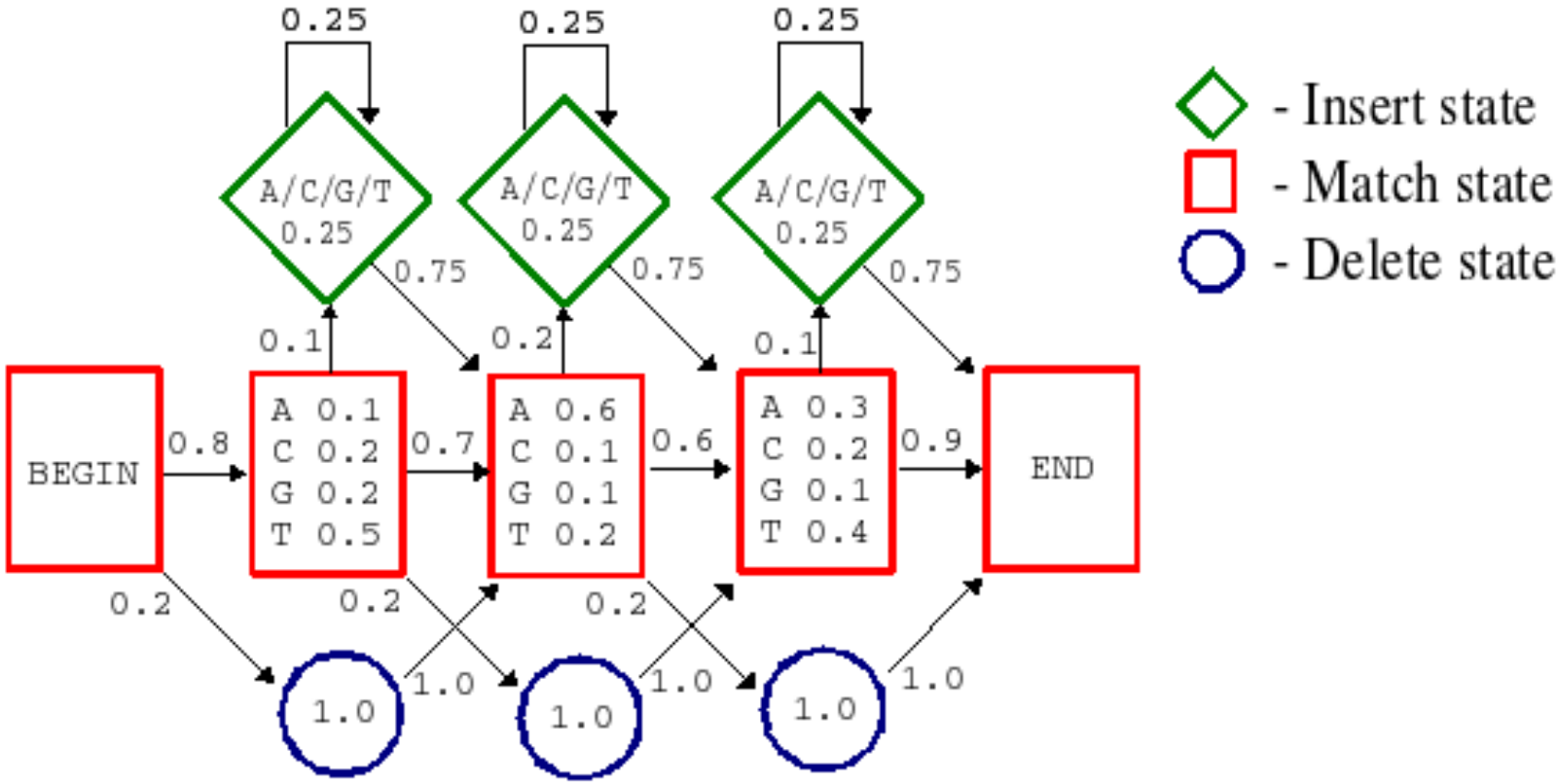
# Allowing for sequence length heterogeneity

- The profile shown in the previous slides 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)



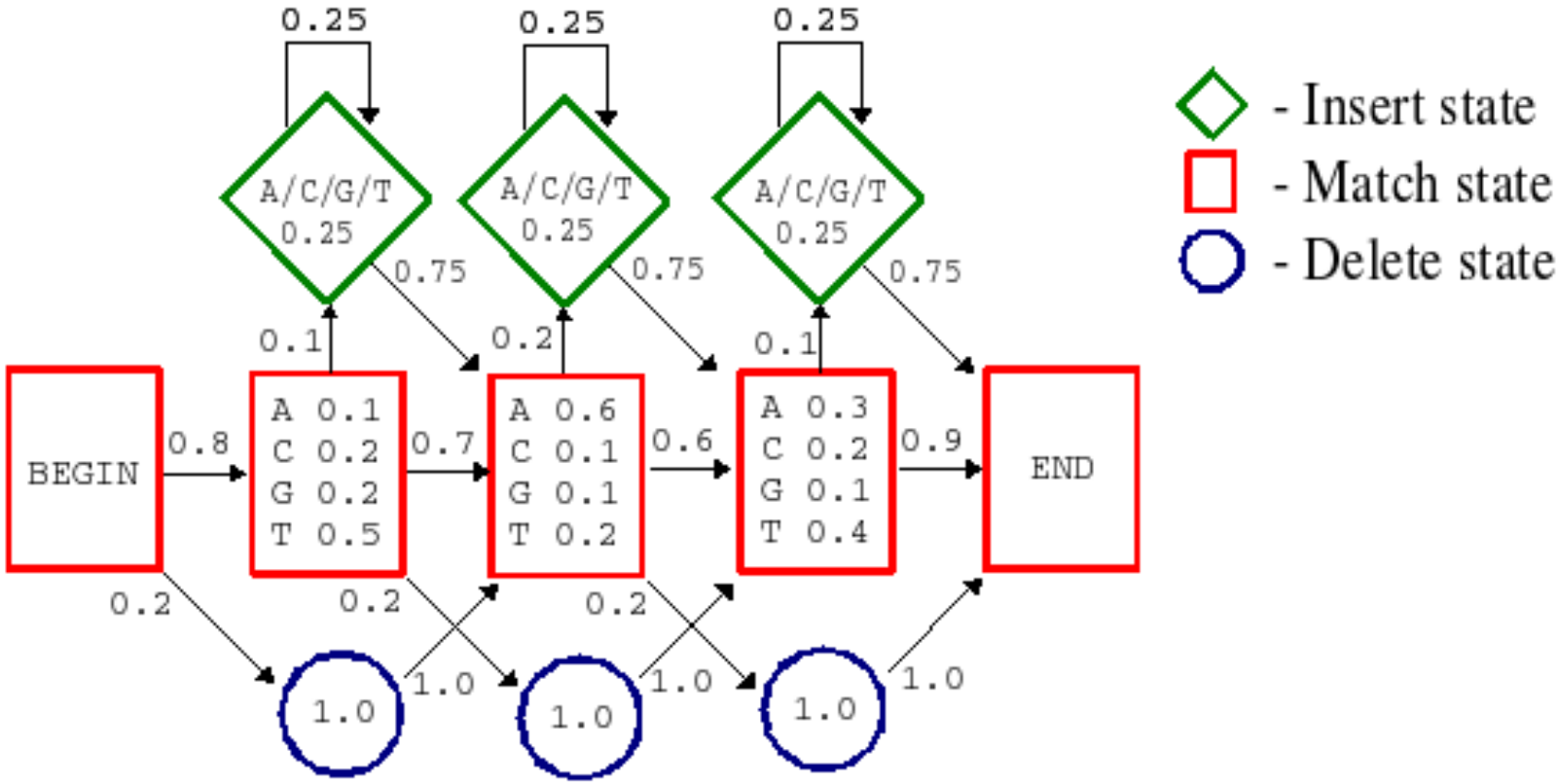

From http://www.cbs.dtu.dk/~kj/bioinfo_assign2.html

# How many paths can generate s = AAA?



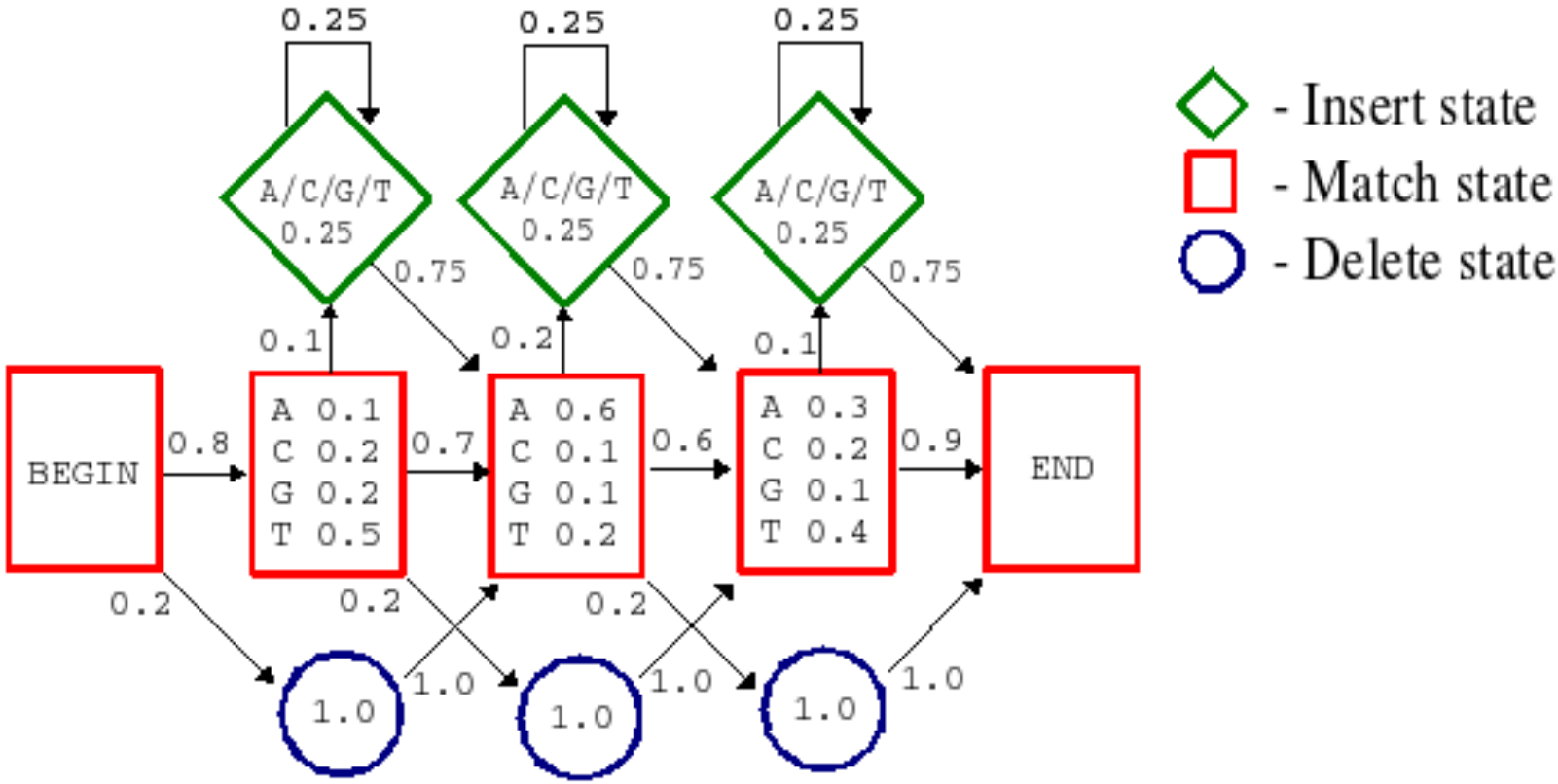

From http://www.cbs.dtu.dk/~kj/bioinfo_assign2.html

# How many paths can generate s = AA?



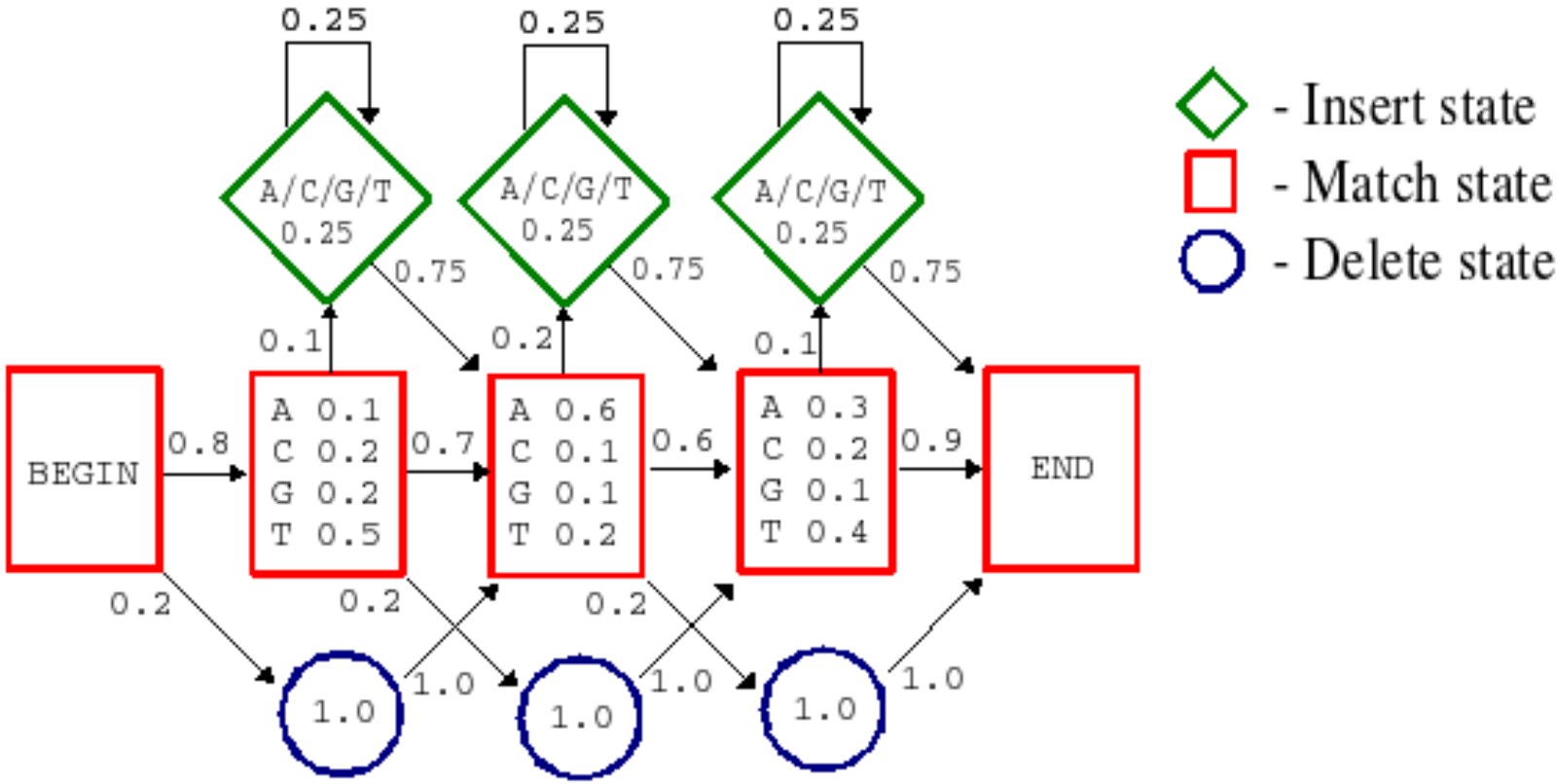

From http://www.cbs.dtu.dk/~kj/bioinfo_assign2.html

# How many paths can generate the empty string?



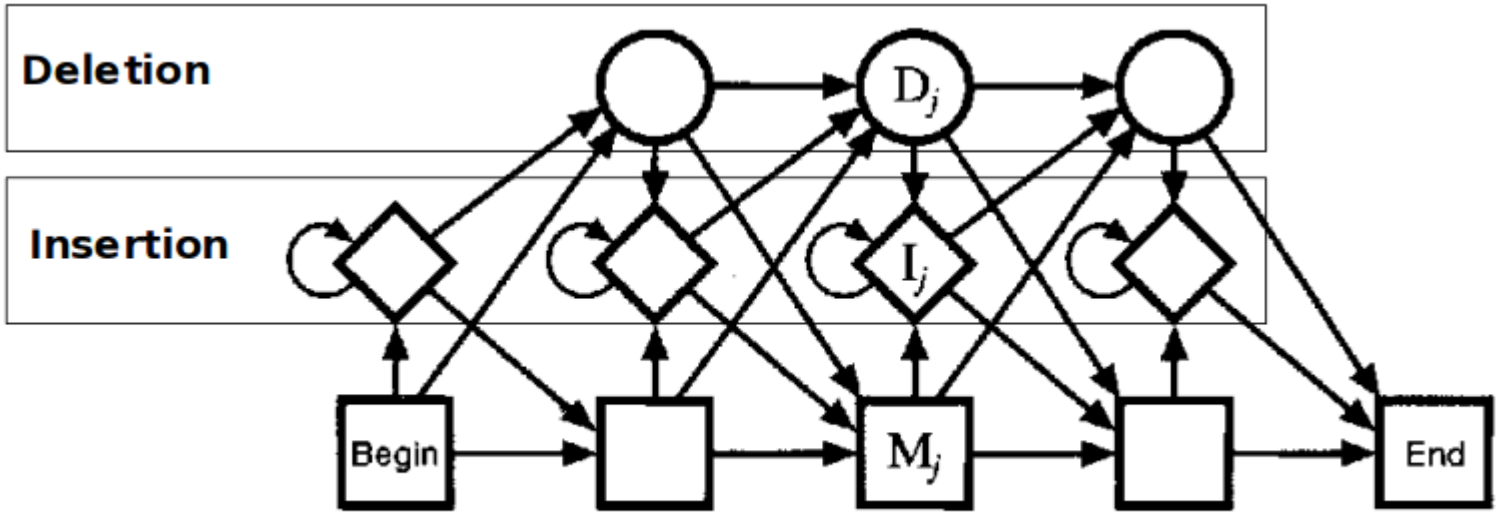

From http://www.cbs.dtu.dk/~kj/bioinfo_assign2.html

# Can you find a string for which only one path can generate it?



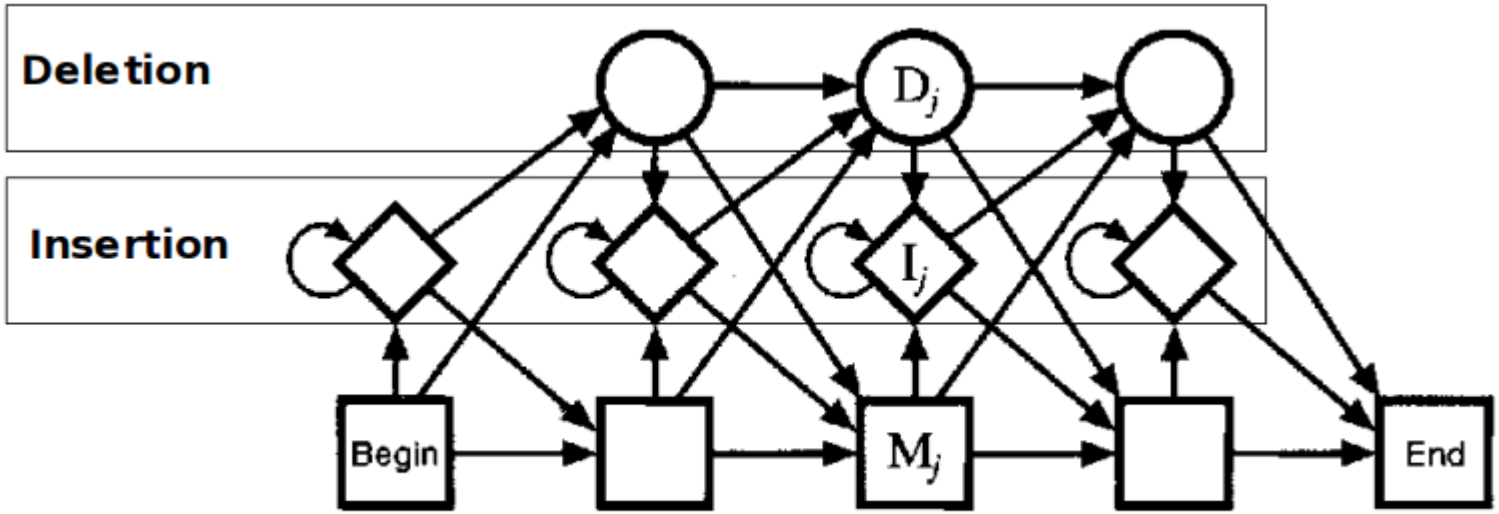

From http://www.cbs.dtu.dk/~kj/bioinfo_assign2.html

# A more general topology for a profile HMM




From http://codecereal.blogspot.com/2011/07/protein-profile-with-hmm.html

# What has changed between this model and the previous one?




From http://codecereal.blogspot.com/2011/07/protein-profile-with-hmm.html

# HMMER

- http://hmmer.janelia.org
- One of the most popular collection of tools to perform analyses based on profile HMMs.
- **HMMER web server: interactive sequence similarity searching, NAR 2011,** http://nar.oxfordjournals.org/content/39/suppl_2/W29
- See PFAM, http://pfam.xfam.org, for how profile HMMs are used to represent groups of functionally and structurally related proteins.

# Building Profile HMMs

- Profile HMMs can be built from a given multiple sequence alignment – this is not too difficult.
- Profile HMMs can also be built from unaligned sequences. This is a bit complicated, and often uses the Baum-Welch algorithm.

# Using Profile HMMs

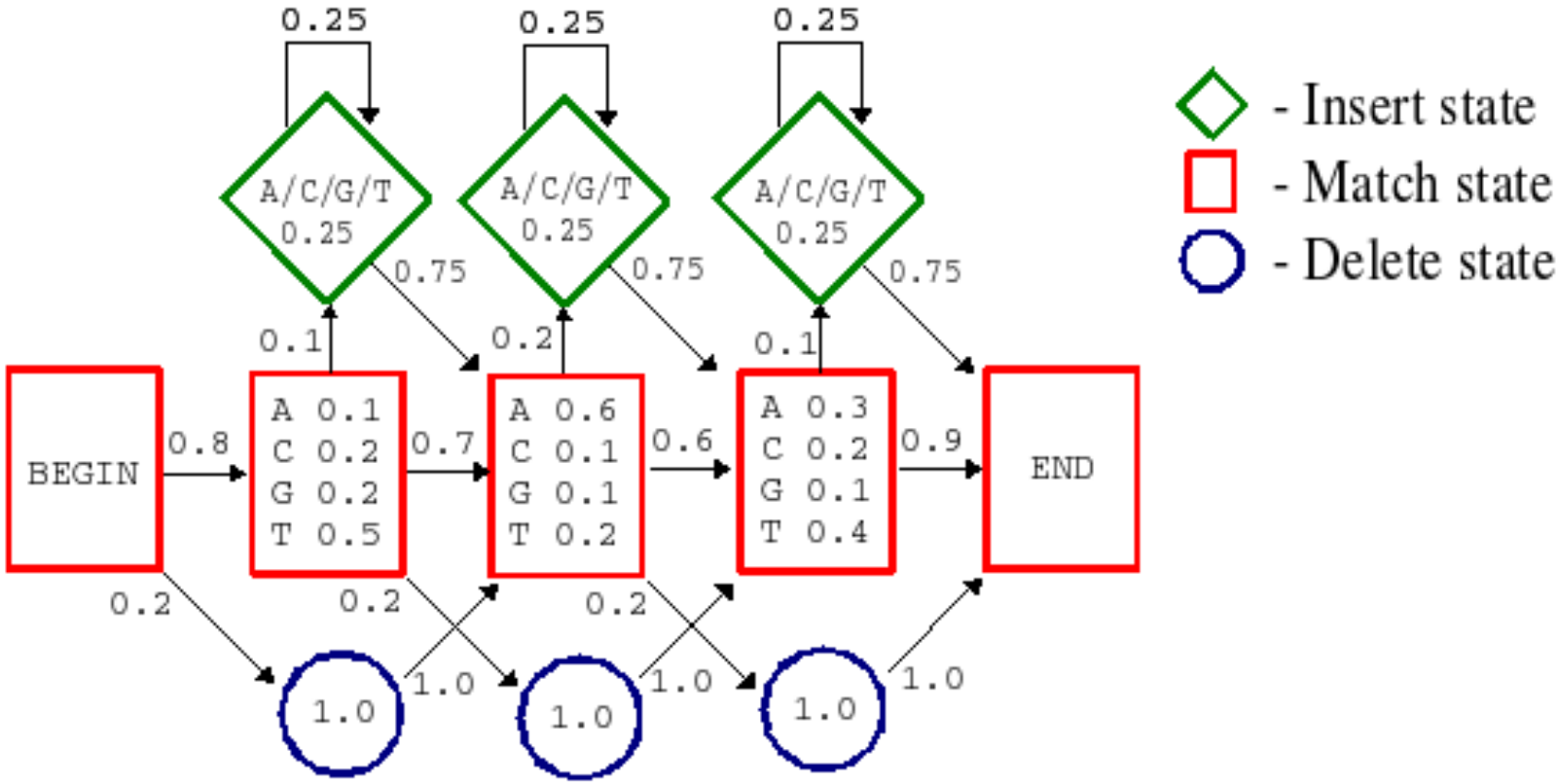
- Given a Profile HMM computed for a multiple sequence alignment, you can use it to
  - Classify new sequences into families (e.g., protein families and superfamilies)
  - Infer function of new sequences
  - Add related sequences into the multiple sequence alignment
  - Compute multiple sequence alignments for groups of related sequences

# Using profile HMMs for Protein Family Classification

- Given two profile HMMs (H1 and H2), and a sequence s, you can determine which one is more likely to generate s using dynamic programming.
- Computing Pr(s|H) can be done using dynamic programming (the "forward algorithm").

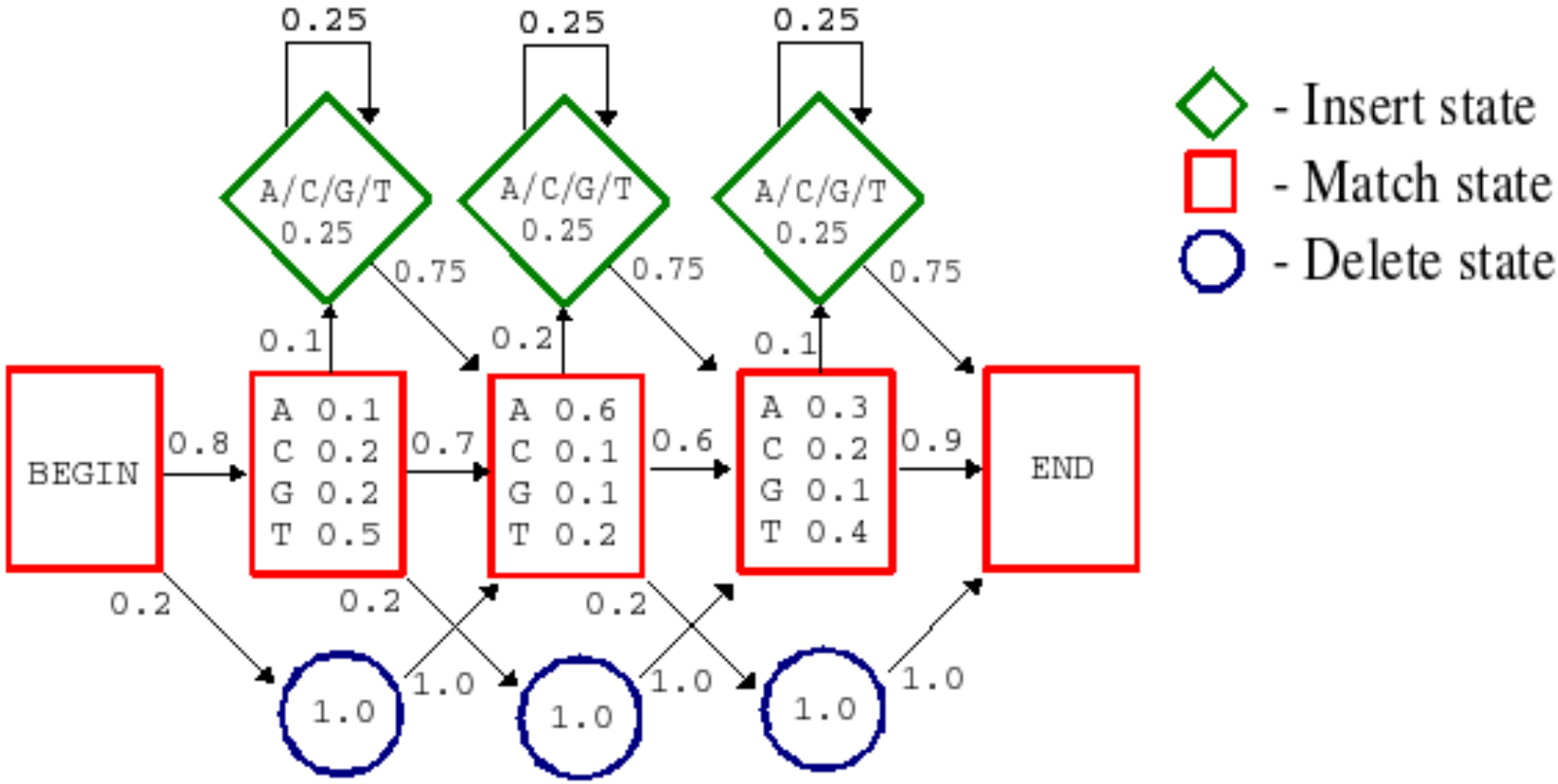
Given a set of protein families, compute profile HMMs for each, and then find the family whose profile HMM is most likely to generate s.

# What is the probability of generating s = AAA?



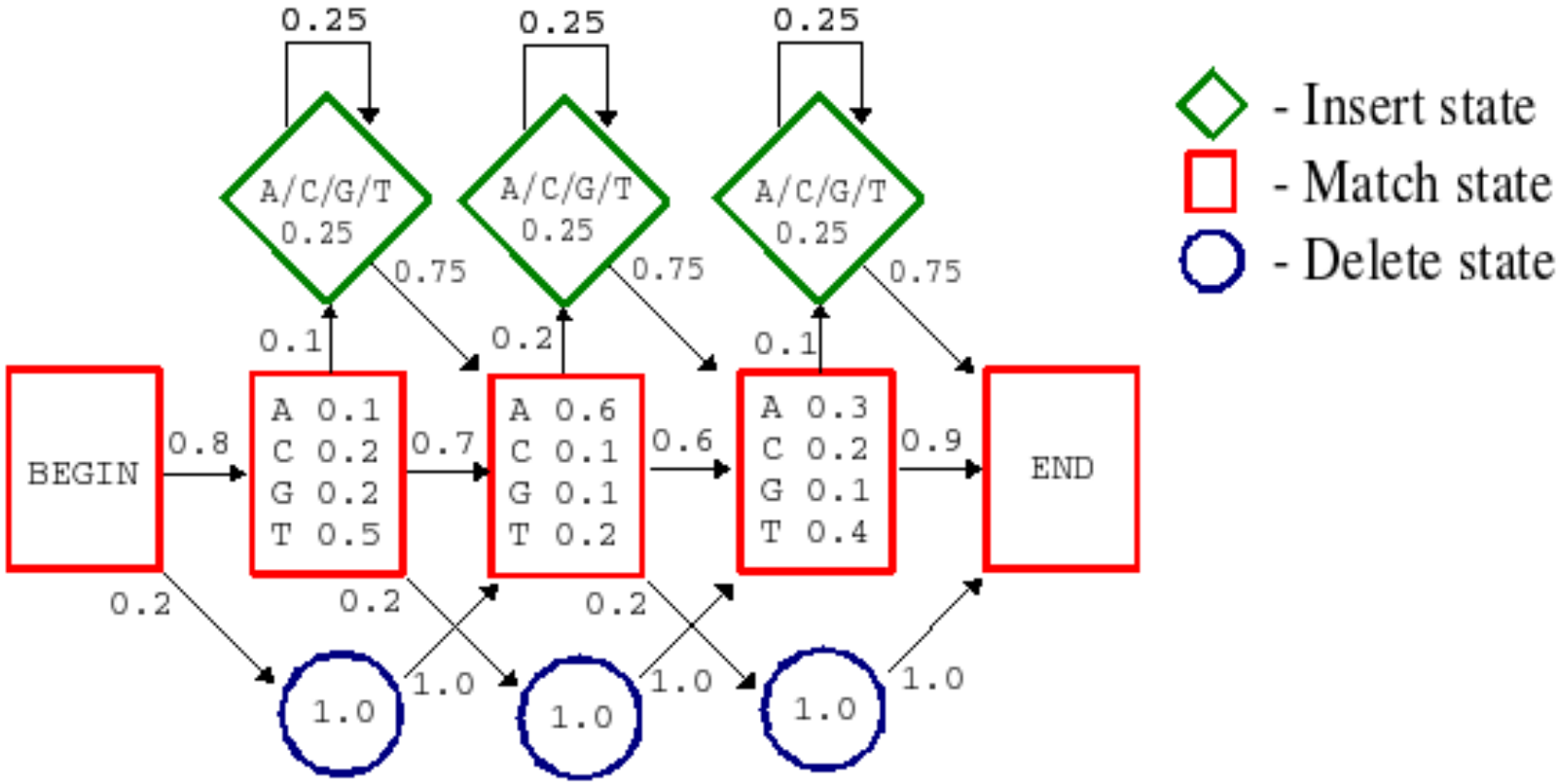

From http://www.cbs.dtu.dk/~kj/bioinfo_assign2.html

# What paths can generate s = AAA?



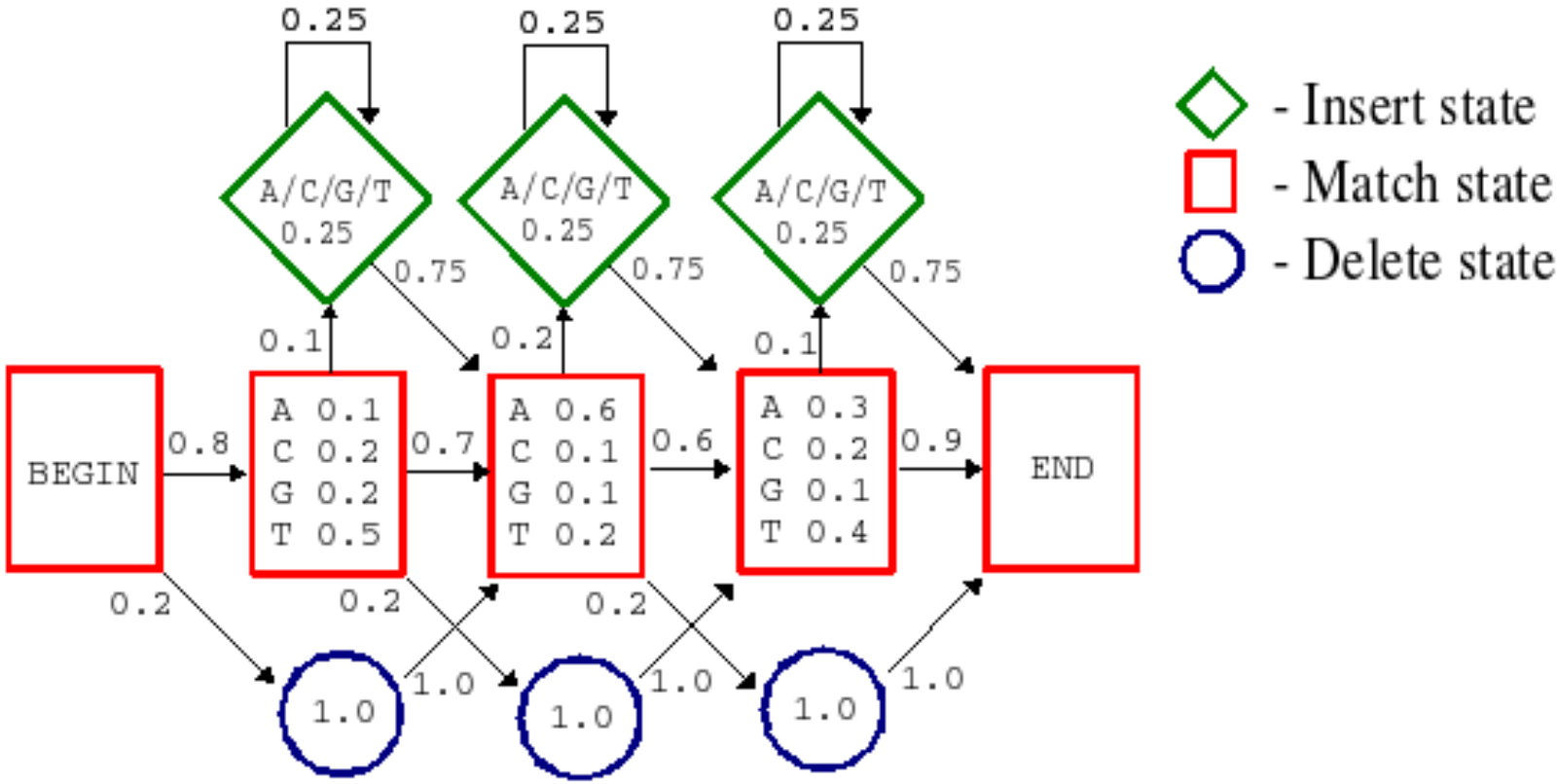

From http://www.cbs.dtu.dk/~kj/bioinfo_assign2.html

# For each path that can generate s = AAA, what is its probability?



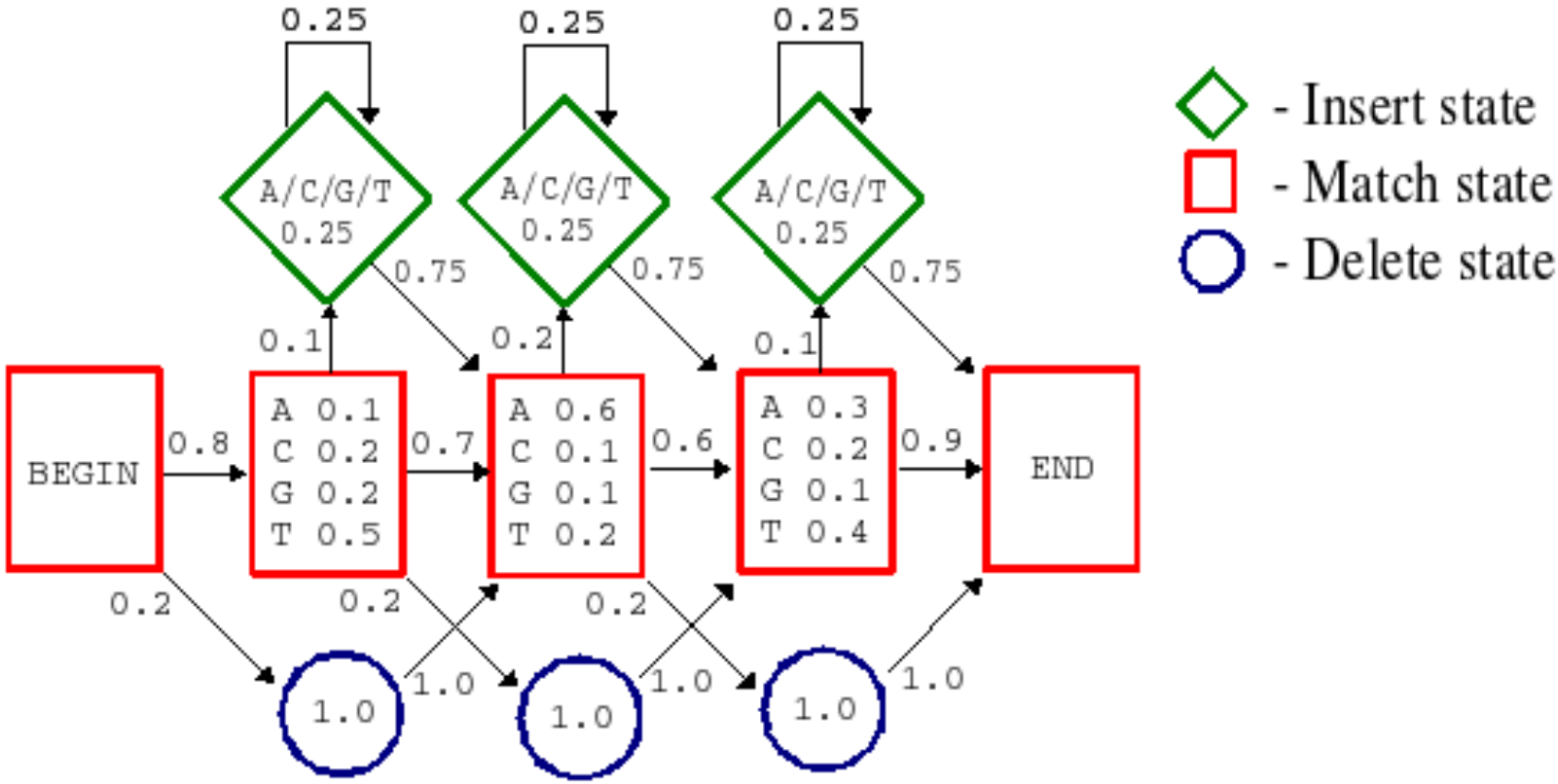

From http://www.cbs.dtu.dk/~kj/bioinfo_assign2.html

# Computing Pr(AAA|H) is not trivial




From http://www.cbs.dtu.dk/~kj/bioinfo_assign2.html

# What about Pr(A|H)?

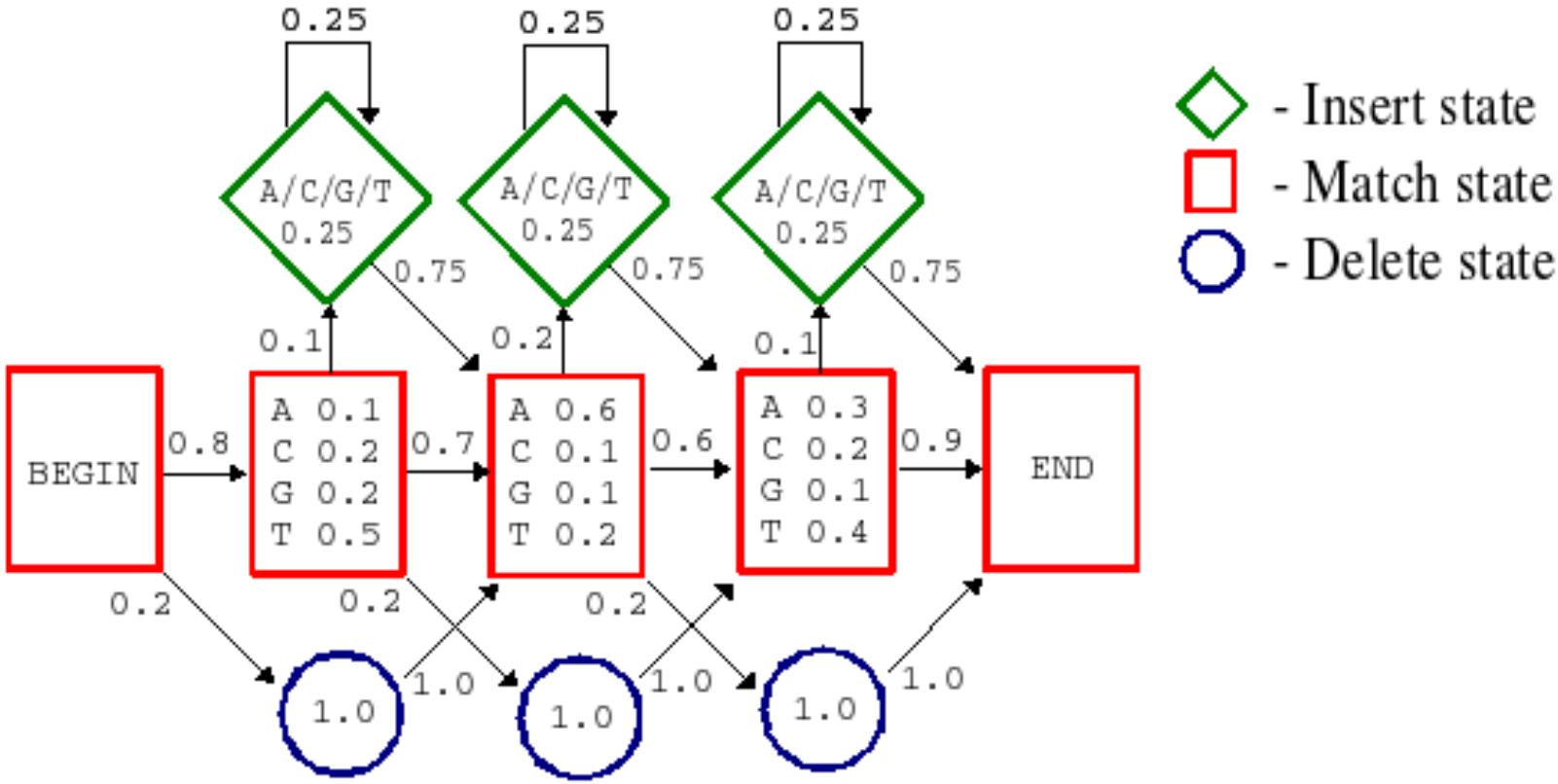



From http://www.cbs.dtu.dk/~kj/bioinfo_assign2.html

# Hidden states

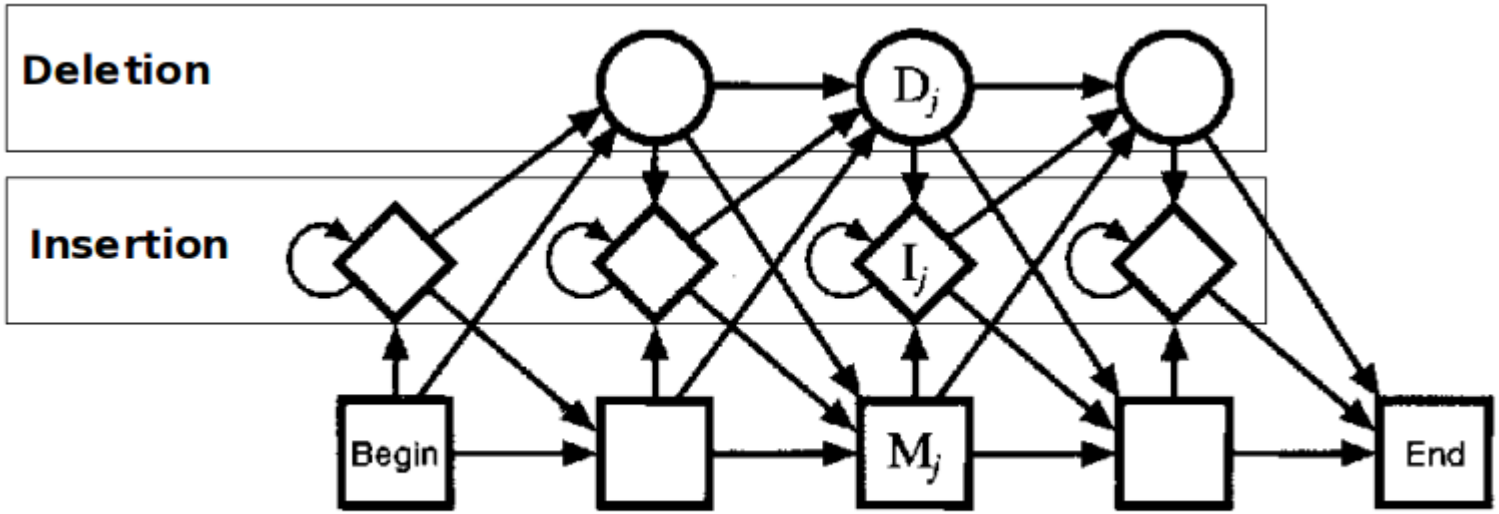
- For some sequences, there is no path through the profile HMM that can generate the sequence.
- For some other sequences, there is exactly one path.
- And then for some others, there is more than one path… and you can't tell which one generated the sequence! More generally, you can't tell which state generated each letter in the sequence – so the states are said to be "hidden".

# What path is most likely to generate sequence AA?




From http://www.cbs.dtu.dk/~kj/bioinfo_assign2.html

# Can we compute the maximum probability path for the general case?




From http://codecereal.blogspot.com/2011/07/protein-profile-with-hmm.html