- **Q1: Could we use HMM for pairwise alignment?**
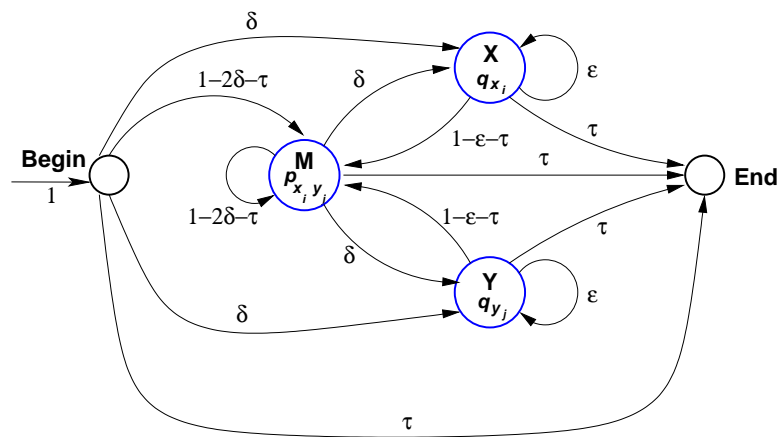- **Q2: Could we do more [with "pair" HMM]?**

or:

How should we read *BSA, Durbin et al., 1998, Ch. 4?*

Part 1 *(mainly an application of already known issues)*

0. Durbin et al. proposes (see Fig 4.2) a pair HMM for global alignment of two sequences with affine gaps.

- Show how two different alignments (among those below) are recognized/generated by the (non-deterministic FSA corresponding to) this HMM.

```
HEAGAWGHEE        HEAGAWGHEE        HEAGAWGHEE        HEAGAWGHEE
-P-A-WHEAE        -P--AWHEAE        P---AWHEAE        -P-A-WHEAE


HEAGAWGHE-E       HEAGAWGHE-E       HEAGAWGHE-E       HEAGAWGHEE
-PA--W-HEAE       --P-AW-HEAE       -P--AW-HEAE       --PA-WHEAE
```

- Q: How does the pair HMM recognizes the best pairwise alignment (among all those alignments that it recognizes)?
  A:

1. Design the Viterbi algorithm for the given pair HMM.
   [i.e, adapt the general HMM Viterbi algorithm (see Ch. 3) to the particular (pair) HMM given above.]

- Q: Could we elaborate a logg-odds version of the Viterbi algorithm for pair HMM? [Attention: not log version!] If so, outline the basic idea.
  A:

2. Design a non-deterministic FSA to recognize pairs of unrelated sequences, including the pair of empty sequences.

3. Elaborate the logg-odds version of the Viterbi algorithm for the pair HMM.

4. Find out the mathematical correspondence between the parameters of pair HMM and the FSA for global alignment with affine gaps which inspired it.
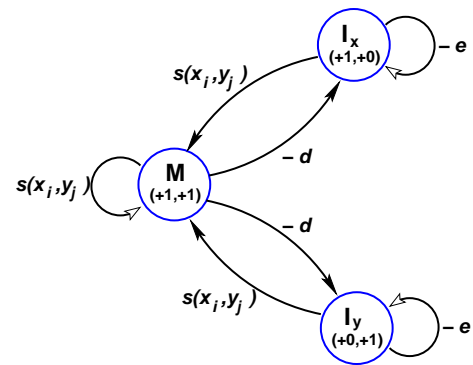
5. Design the **forward algorithm** for the pair HMM.

6. Design the **backward algorithm** for the pair HMM.

Part 2 *(introducing/"discovering" new issues)*

7. Q: How can we compute the **full probability that x and y are related** (independently of the alignment path)?
   Hint: $P(x, y) = \sum_{alignments \; \pi} P(x, y, \pi)$
   A:

• Q: How do you get $P(x, y, \pi^\star)$, where $\pi^\star$ is the best alignment path?
   A:

• Q: What does $P(\pi^\star \mid x, y)$ represent?
   A: The posterior probability of the Viterbi/best alignment path, given that the two sequences are related.

   Q: How do you compute it?
   A:

• Q: True or False: $P(x, y, \pi^\star) \leq P(x, y)$?
   A:

8. Can we adapt the back-tracing procedure of the Viterbi algorithm for pair HMM so to **sample** (i.e, randomly generate) **suboptimal alignments**?

9. Q: How can we compute the probability that two positions x_i and y_j are aligned, regardless of the alignment path?
   [Notation: $P(x_i \Diamond y_j \mid x, y)$]
   A:

   Q: The same question for $P(x, y, x_i \Diamond -)$, and $P(x, y, - \Diamond y_j)$.
   A:

10. The **overall accuracy of an alignment** π is defined as $\mathcal{A}(\pi) = \sum_{(i,j) \; \in \pi} P(x_i \Diamond y_j)$.
    Q: Could we design a dynamic programming **algorithm** that finds the alignment/path having the best overall accuracy for a given pair of sequences?

For your laboratory portfolio

1. Implement the algorithm for finding suboptimal alignments of two sequences based on probabilistic sampling.

2. a. Given a set/file of pairwise alignments, estimate (using MLE) the p_{ab} and q_x probabilities.

   b. Using the EM algorithm for pair HMM and some fresh alignments, compute some good values for the parameters ϵ, δ, and τ. (You will arbitrarily set initial values for these parameters.)

3. Implement the alignment algorithm based on overall accuracy.