We need Markov's help

1. Old Markov's new role
2. A toy HMM
3. What's hidden?
4. Finding the best state path

1. Old Markov's new role

- Often, problems in biological sequence analysis are just a matter of putting the right label on each residue, such as, in gene finding and sequence alignment;
- Hidden Markov models (HMM) are a formal foundation for making probabilistic models of linear sequence "labeling" problems;
- HMMs provide a conceptual toolkit that allows building a model of almost any complexity; the search and alignment applications constitute probably the best-known use of HMMs for biological sequence analysis;
- HMMs are the heart of a diverse range of programs, and the Legos of computational sequence analysis.

2. A toy HMM: 5' splice site recognition

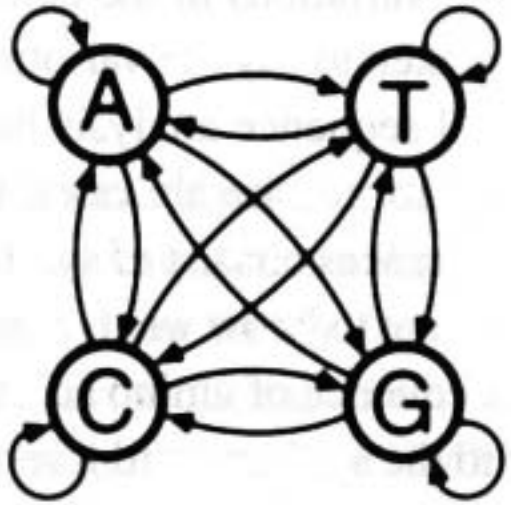
- Markov chain

For any probabilistic model of sequences we can write the probability of the sequence as

$$P(x) = P(x_L, x_{L-1}, \ldots, x_1)$$
$$= P(x_L|x_{L-1},\ldots,x_1)P(x_{L-1}|x_{L-2},\ldots,x_1)\ldots P(x_1)$$

the probability of each symbol (state) x_i depends only on the value of the preceding symbol x_{i-1}, i.e. $P(x_i|x_{i-1},\ldots,x_1) = P(x_i|x_{i-1})$

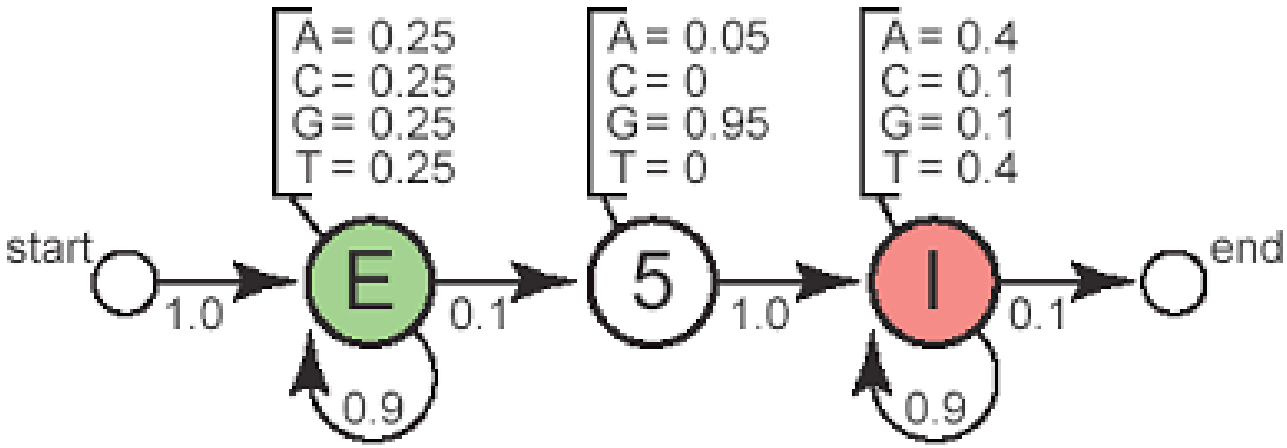
$$P(x) = P(x_L|x_{L-1})P(x_{L-1}|x_{L-2})\ldots P(x_2|x_1)P(x_1)$$



- A Markov chain for DNA
- state and transition probability

Sequence: C T T C A T G T G A A A G C A G A C G T A A G T C A

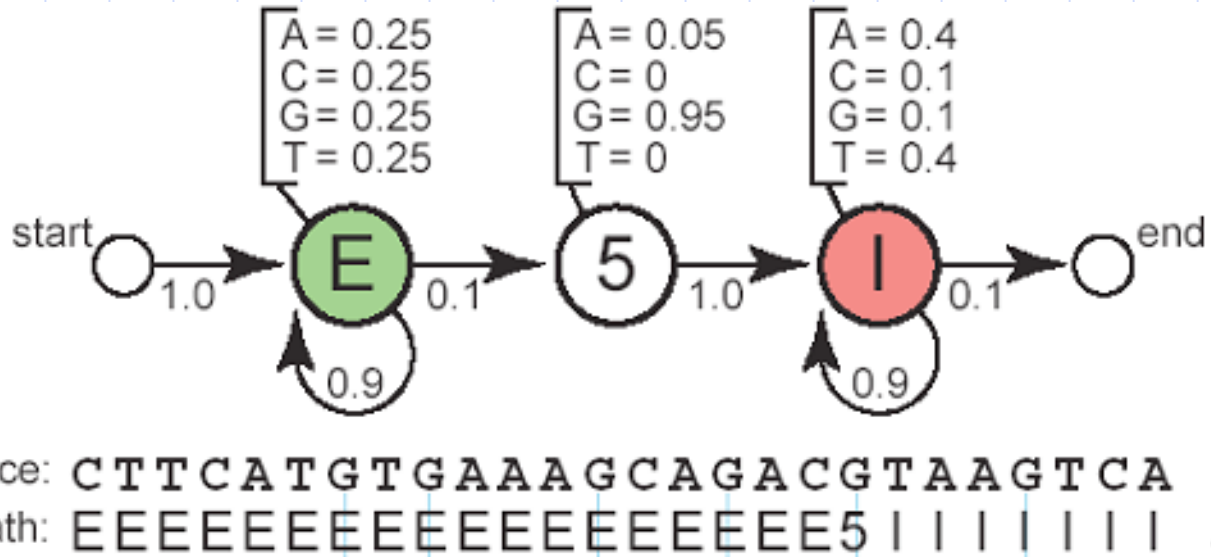
- HMM: emission probability



Sequence: C T T C A T G T G A A A G C A G A C G T A A G T C A

3. What's hidden?

- It's useful to imagine an HMM generate a sequence: an underlying state path (the labels) and an observed sequence (the DNA);
- The state path is a Markov chain: what state we go to next depends only on what state we are in;
- The state path also is a hidden Markov chain: if we're only given the observed sequence, this underlying path that we'd like to infer, is hidden.



- The essential difference between Markov chain and HMM: there is not a one-to-one correspondence between the states and the symbols for a HMM. It is no longer possible to tell what state the model was in when x_i was generated just by looking at x_i. In a Markov chain you always know exactly in which state a given observation belongs.

4. Finding the best state path

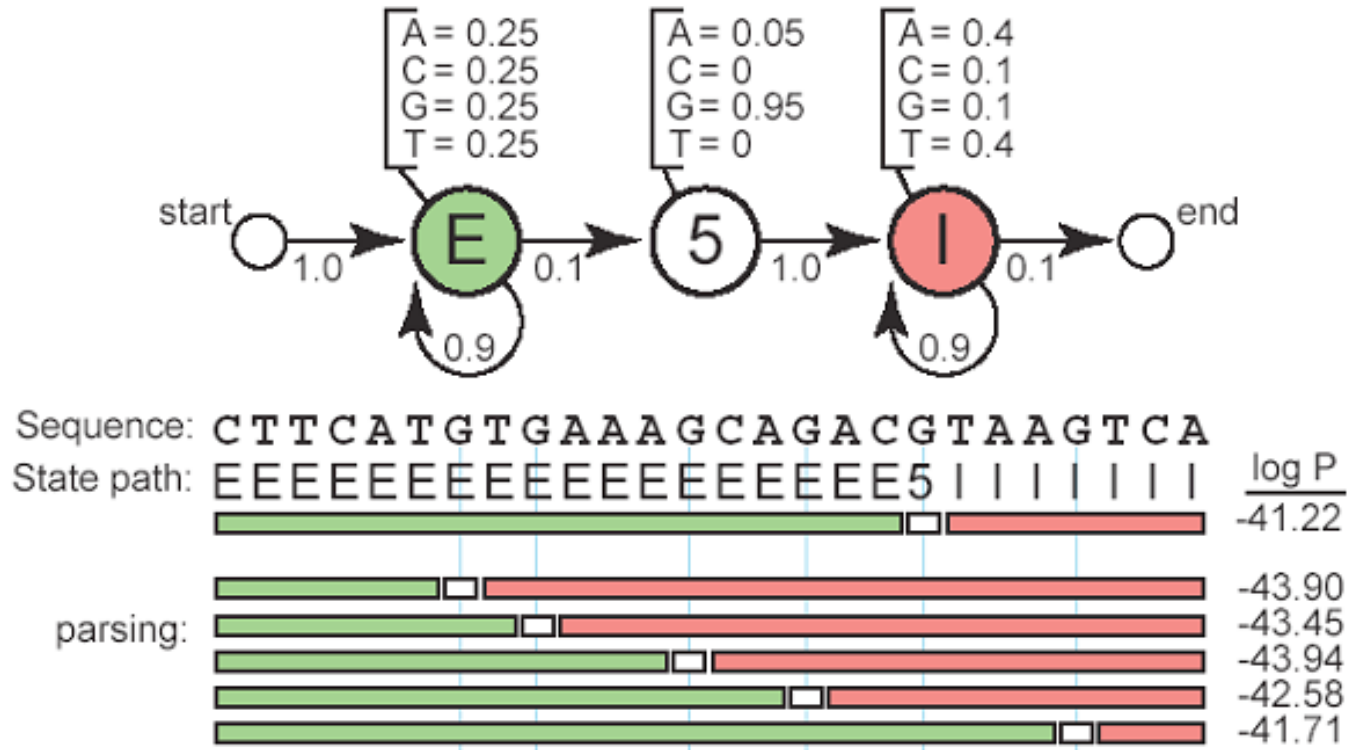
- $P(S, \pi \mid HMM, \theta)$

(An HMM with parameters θ generates a state path π and an observed sequence S)

is the product of all the emission and transition probabilities were used.

Total 14 possible paths that have non-zero probability for the example of 26 nucleotide sequence

For most real problems, there are so many possible state sequences that could not afford to enumerate them: Viterbi algorithm (a dynamic programming algorithm) is guaranteed to find the most probable state path given a sequence and an HMM.

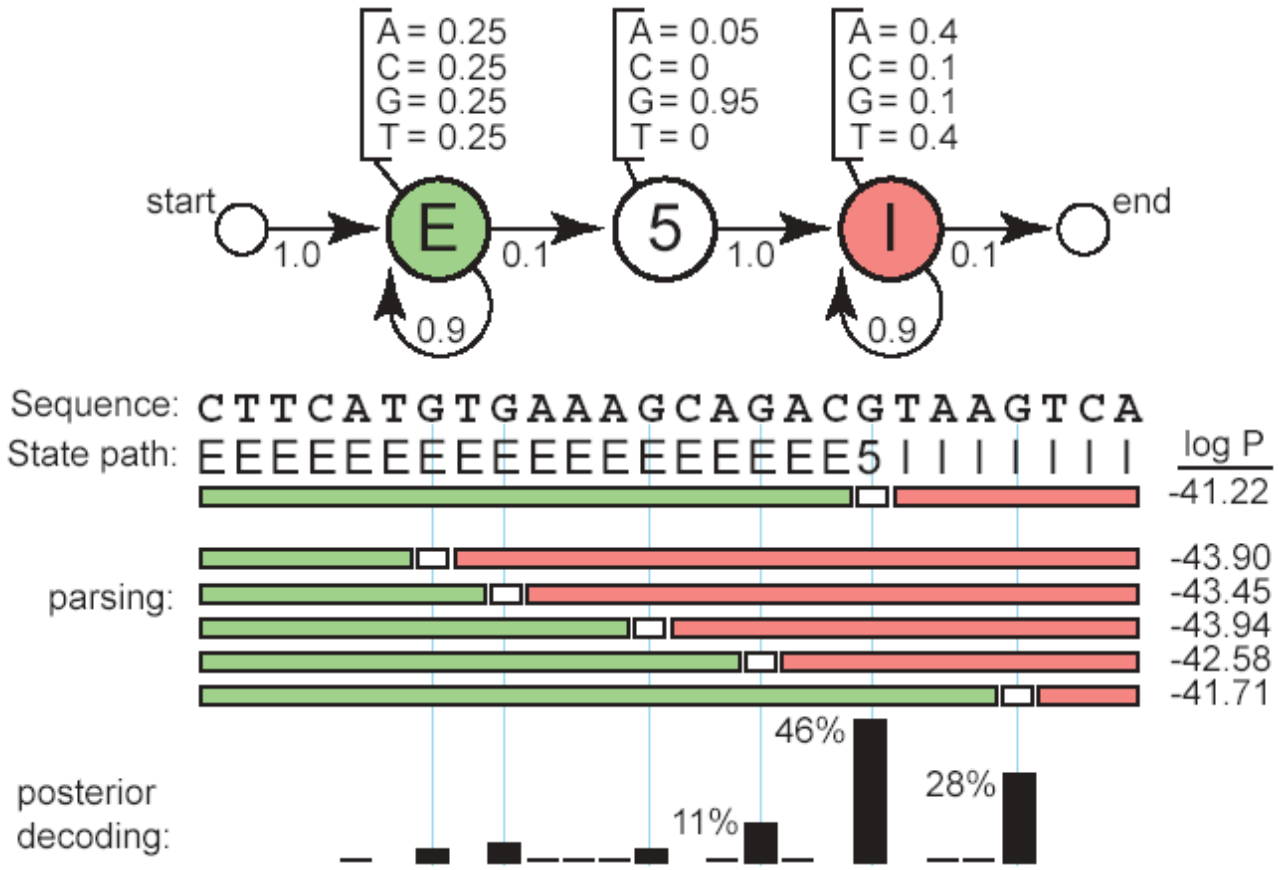

A = 0.25
C = 0.25
G = 0.25
T = 0.25
A = 0.05
C = 0
G = 0.95
T = 0
A = 0.4
C = 0.1
G = 0.1
T = 0.4
start
E
5
I
end
1.0
0.1
1.0
0.1
0.9
0.9
Sequence: CTTCATGTGAAAGCAGACGTAAGTCA
State path: EEEEEEEEEEEEEEEEEEE5IIIIIIII
log P
-41.22
parsing:
-43.90
-43.45
-43.94
-42.58
-41.71

posterior decoding

We want to know what the most probable state is for an observation x_i. More generally, we may want the probability that observation x_i came from state k given the observed sequence, i.e. $P(\pi_i = k \mid x)$

$$P(\pi_i = k \mid x) = \frac{f_k(i)b_k(i)}{P(x)} = \frac{P(x_1,\ldots,x_i,\pi_i = k)P(x_{i+1}\ldots x_L \mid \pi_i = k)}{\sum_{\pi} P(x,\pi)}$$

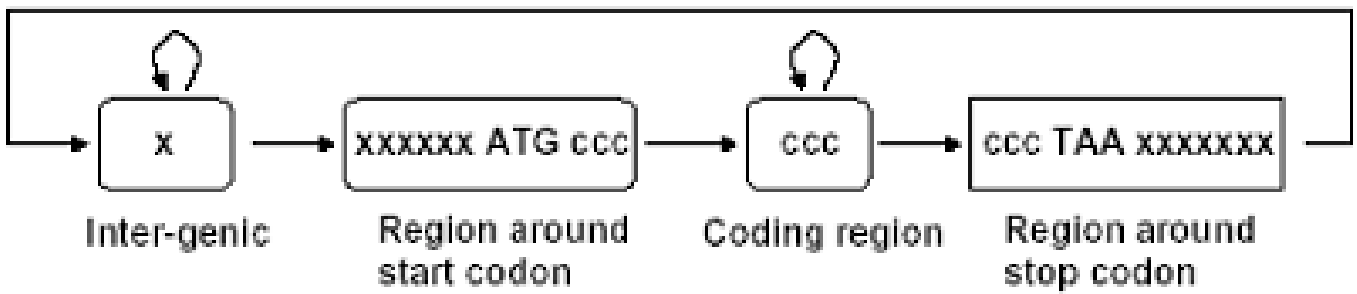
(R.S.A.G.'s book, pp57-61)


A = 0.25
C = 0.25
G = 0.25
T = 0.25
A = 0.05
C = 0
G = 0.95
T = 0
A = 0.4
C = 0.1
G = 0.1
T = 0.4
start
E
5
I
end
1.0
0.1
1.0
0.1
0.9
0.9
Sequence: C T T C A T G T G A A A G C A G A C G T A A G T C A
State path: E E E E E E E E E E E E E E E E E E E 5 I I I I I I I
log P
-41.22
parsing:
-43.90
-43.45
-43.94
-42.58
-41.71
posterior decoding:
46%
28%
11%

Genscan: HMM for gene finding

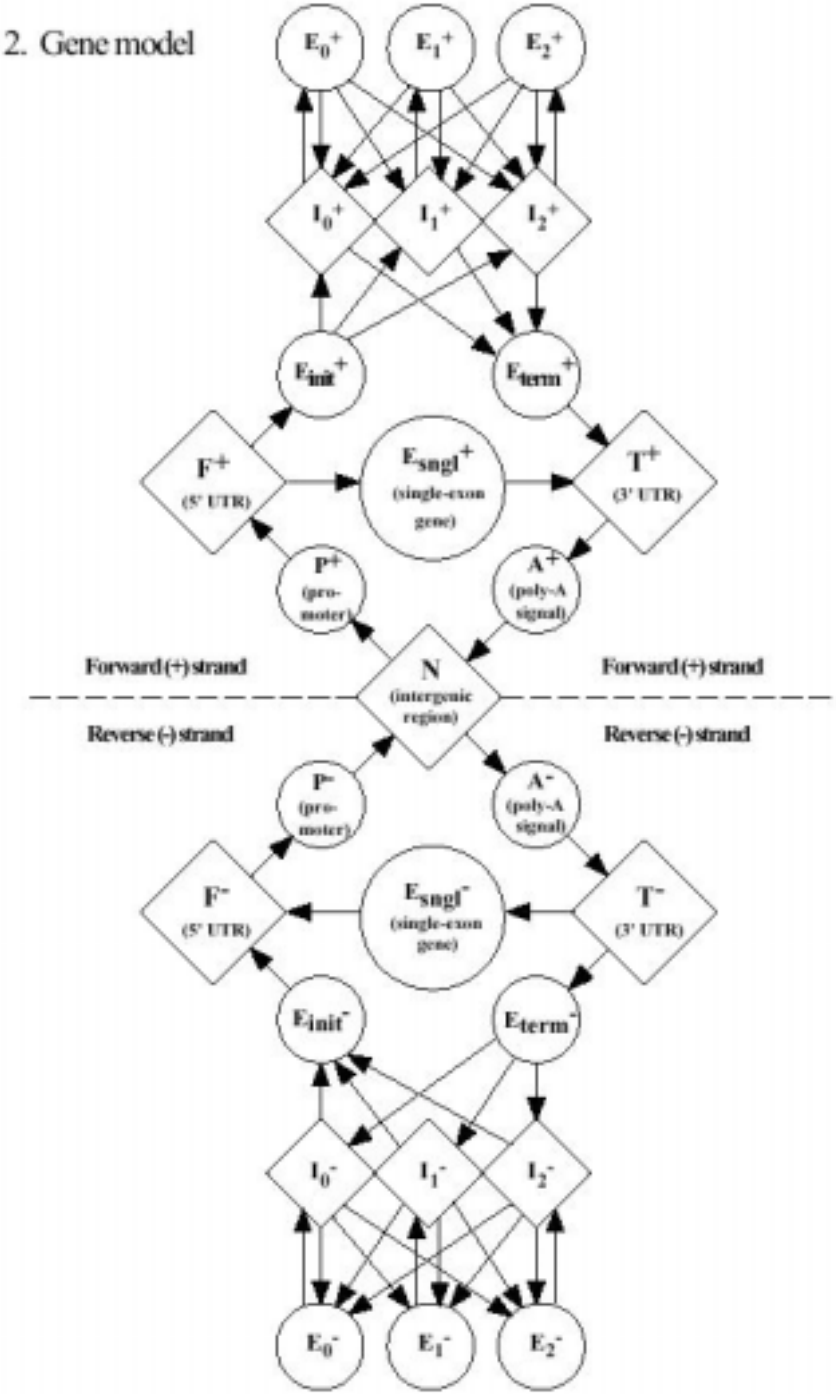
- **More realistic models**

For example: gene model



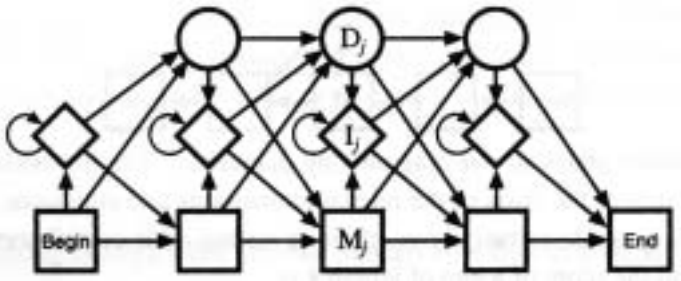
- A Semi-Markov Model: there are no transitions from the state to the state itself;
- Tracks “phase” of exon or intron (0 coincides with codon boundary, or 1 or 2);
- long-range correlation: 5-order Markov model.

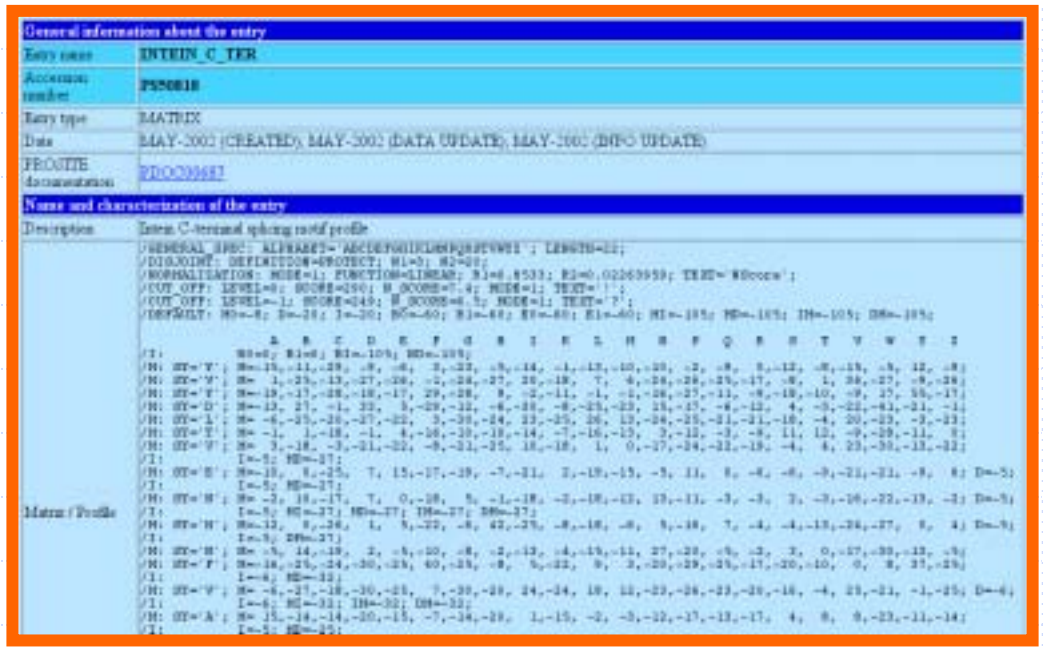
Fig. 2. Gene model



Profile HMMs for sequence families

- "profile"(Gribskov et al, 1987): non-probabilistic profile
- profile HMM:

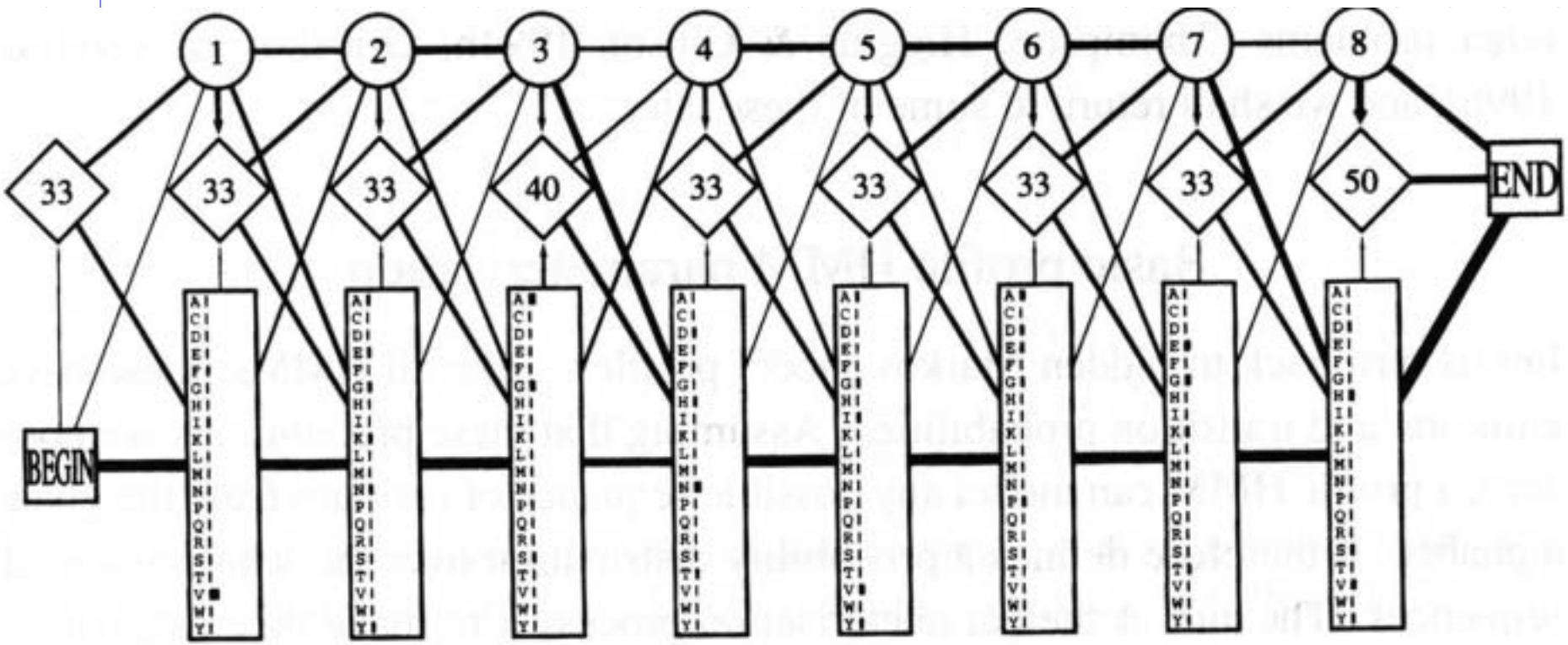




An example

- seven global protein sequences

```
HBA_HUMAN    ...VGA--HAGEY...
HBB_HUMAN    ...V----NVDEV...
MYG_PHYCA    ...VEA--DVAGH...
GLB3_CHITP   ...VKG------D...
GLB5_PETMA   ...VYS--TYETS...
LGB2_LUPLU   ...FNA--NIPKH...
GLB1_GLYDI   ...IAGADNGAGV...
                  ***   *****
```



Exercise

- Give a pair HMM for pairwise alignment