

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

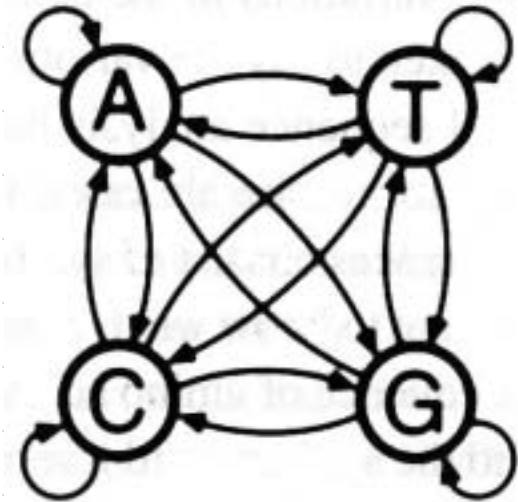
For any probabilistic model of sequences we can write the probability of the sequence as

$$\begin{aligned} P(x) &= P(x_L, x_{L-1}, \dots, x_1) \\ &= P(x_L | x_{L-1}, \dots, x_1) P(x_{L-1} | x_{L-2}, \dots, x_1) \dots P(x_1) \end{aligned}$$

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

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

◆ A Markov chain for DNA

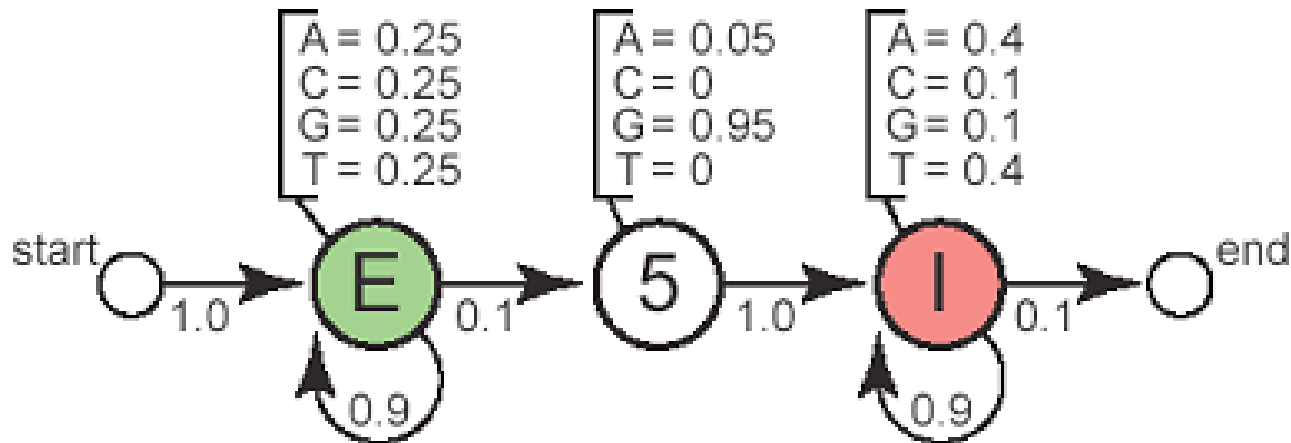


◆ state and transition probability



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

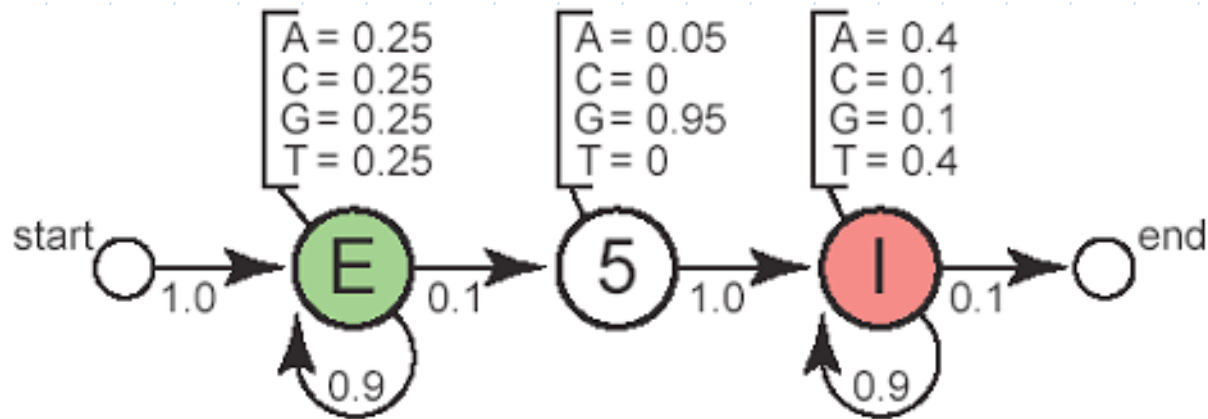
◆ HMM: emission probability



Sequence: **CTTCATGTGAAAGCAGACGTAAGTCA**

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.



Sequence: **CTTCATGTGAAAGCAGACGTAAGTCA**
 State path: **EEEEEEEEEEEEEEEEEEEE5IIIIII**

- ◆ 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

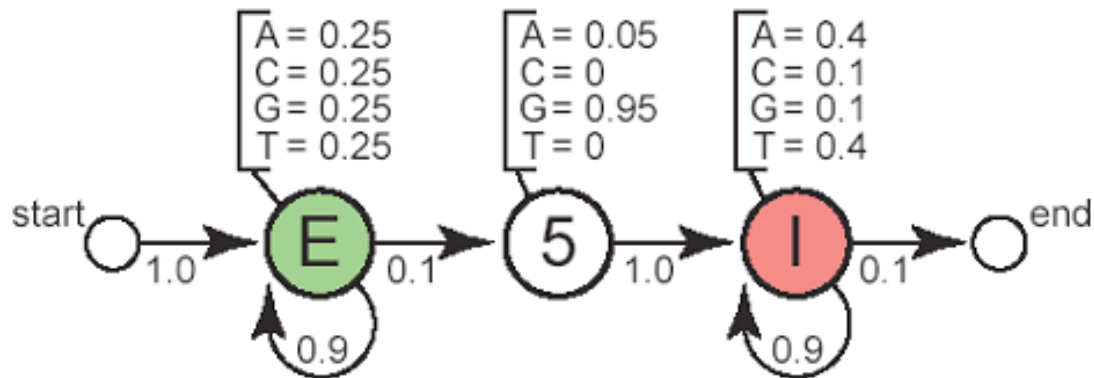
◆ $P(S, \pi | HMM, \theta)$

(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.

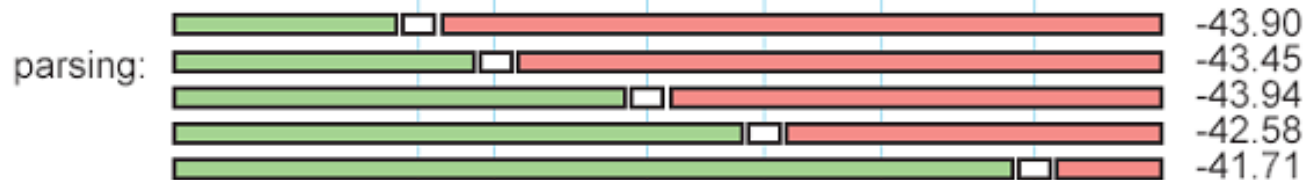
Total 14 possible paths that have non-zero probability for the example of 26 nucleotide sequence

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.



Sequence: **CTTCATGTGAAAGCAGACGTAAGTCA**

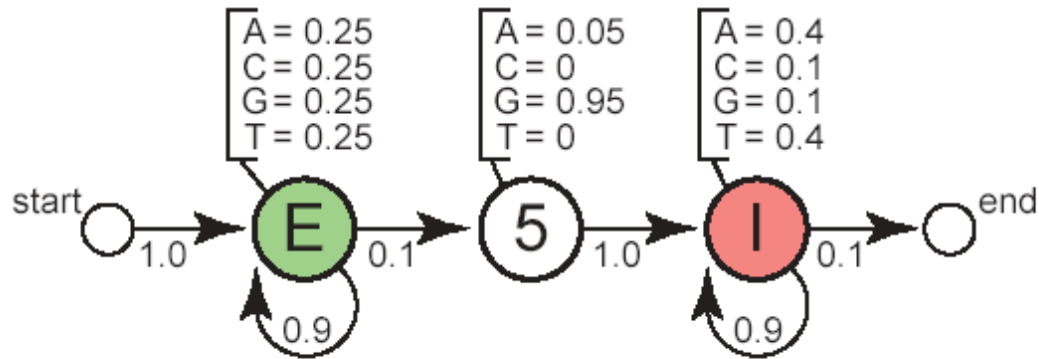
State path: **EEEEEEEEEEEEEEEEEEEE5 | | | | | | | |** $\log P$



◆ posterior decoding

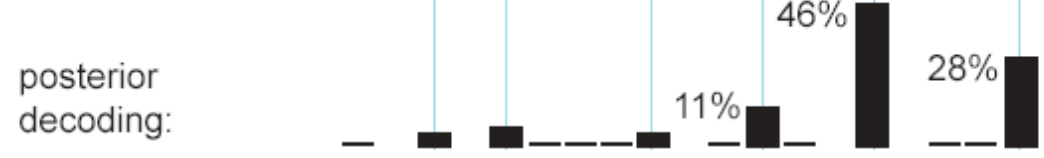
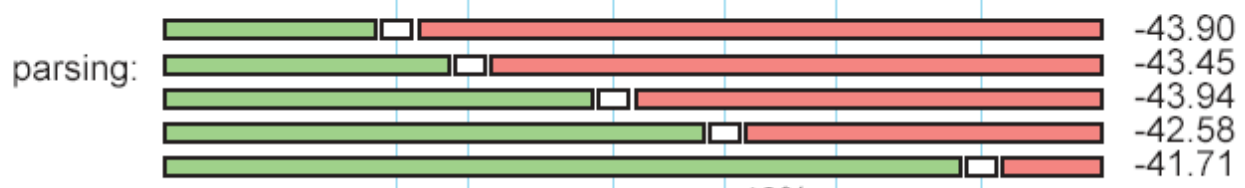
We want to know what the most probable state is for an observation x_i . More generally, we may want the probability that observation x_i came from state k given the observed sequence, i.e. $P(\pi_i = k | x)$

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



Sequence: **CTTCATGTGAAAGCAGACGTAAGTCA**

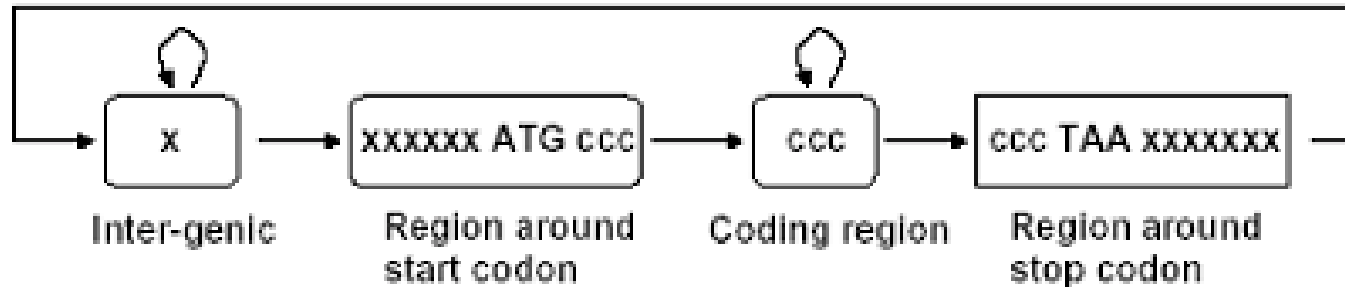
State path: **EEEEEEEEEEEEEEEEEEEE5IIIIIIII** $\log P$



Genscan: HMM for gene finding

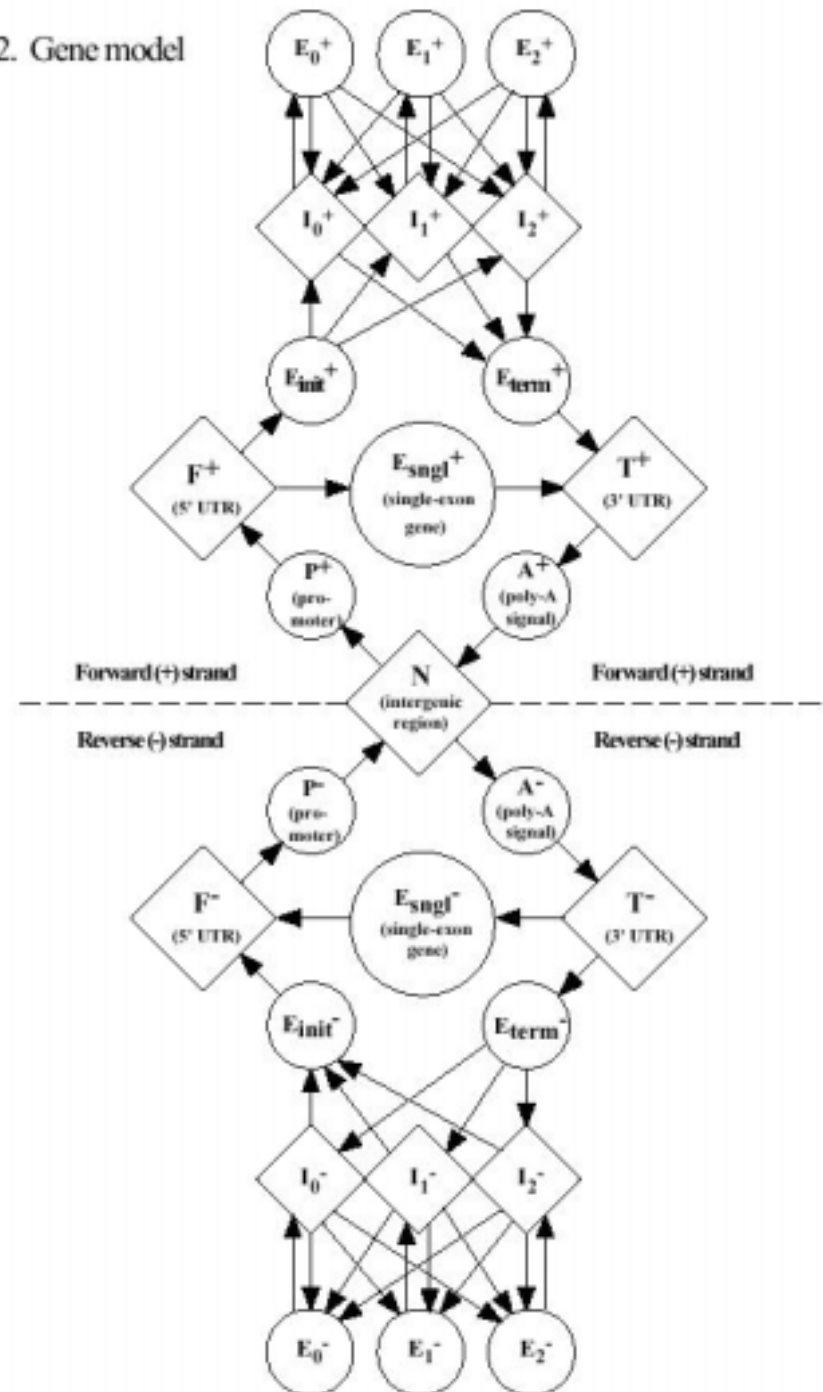
◆ More realistic models

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.

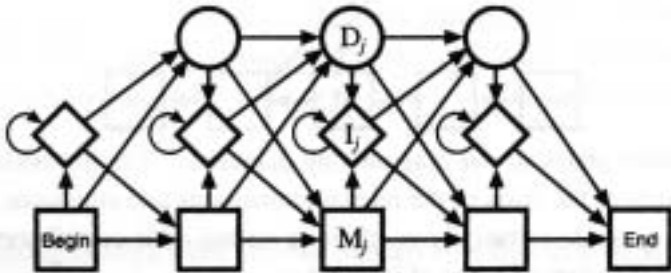
Fig. 2. Gene model



Profile HMMs for sequence families

◆ “profile” (Gribskov et al, 1987): non-probabilistic profile

◆ profile HMM:

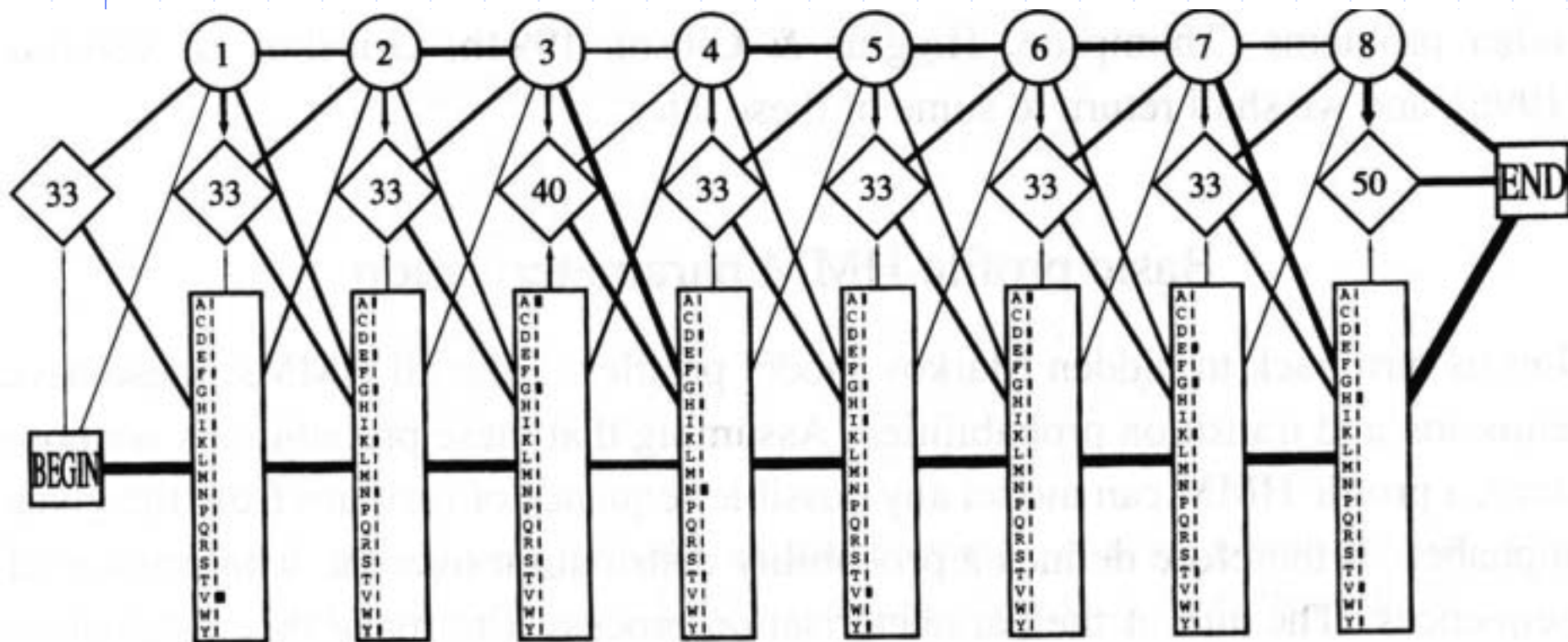


General information about the entry	
Entry name	INTRN_C_TER
Accession number	P550618
Entry type	MATRIX
Date	MAY-2002 (CREATED), MAY-2002 (DATA UPDATE), MAY-2002 (INFO UPDATE)
PROSITE documentation	PS00361
Name and characterization of the entry	
Description	Intein C-terminal splicing motif profile
/GENERAL SPEC: ALPHABET=ACDEFGHIKLRNPQRSTVWY; LENGTH=41;	
/DISJOINT: DEFINITION=PROTECT; M1=0; M2=0;	
/NORMALIZATION: MODE=1; FUNCTION=LINEAR; S1=0.8533; S2=0.0226959; TEST=MScore;	
/CUT OFF: LEVEL=1; SCORE=150; H SCORE=7; MEE=1; TEST=1;	
/CUT OFF: LEVEL=1; SCORE=149; H SCORE=6; MEE=1; TEST=7;	
/DEFAULT: M0=8; D=-24; I=-20; B0=60; E1=60; E2=60; E3=60; M1=10; M2=10; IM=10; DM=10;	
A B C D E F G H I K L N P Q R S T V W Y Z	
/I:	M0=0; E1=0; E2=10; M=10;
/M:	SY='Y': M=-15,-11,-23,-8,-8, 3,-22,-8,-14,-1,-17,-10,-20,-2,-8, 8,-12,-8,-13,-8,-12,-8;
/M:	SY='V': M=-1,-23,-13,-27,-28,-8,-24,-27,-20,-18, 7, 4,-24,-24,-24,-17,-8, 1,-24,-27,-8,-28;
/M:	SY='T': M=-13,-13,-24,-18,-17, 29,-28, 8,-2,-11,-1,-1,-24,-27,-11,-8,-18,-10,-8, 27,-28,-17;
/M:	SY='D': M=-12, 27,-1, 32, 8,-29,-32,-8,-28,-8,-23,-23, 24,-17,-8,-12, 4,-8,-22,-8,-21,-14;
/M:	SY='L': M=-4,-25,-24,-27,-22, 3,-30,-24, 22,-25, 26, 13,-24,-25,-21,-21,-18,-4, 20,-23,-3,-23;
/M:	SY='I': M=-1, 8,-18,-1, 4,-10,-30,-18,-14,-7,-16,-19, 3,-12,-3,-8, 11, 12,-9,-29,-11, 8;
/M:	SY='V': M= 3,-18,-3,-21,-21,-8,-21,-25, 18,-18, 1, 0,-27,-24,-21,-19,-4, 8, 23,-35,-13,-22;
/I:	I=-5; M0=17;
/M:	SY='E': M=-10, 8,-25, 7, 15,-17,-19,-7,-21, 2,-19,-13,-8, 11, 8,-4,-8,-9,-21,-21,-8, 8; D=-5;
/I:	I=-5; M0=17;
/M:	SY='H': M=-2, 19,-17, 7, 0,-18, 8,-1,-18,-2,-18,-13, 19,-11,-3,-3, 2,-3,-16,-22,-13,-2; D=-3;
/I:	I=-5; M0=17; IM=17; IM=17; DM=17;
/M:	SY='N': M=12, 8,-24, 1, 8,-22,-8, 42,-28,-8,-18,-8, 8,-18, 7,-4,-1,-13,-24,-27, 8, 4; D=-3;
/I:	I=-3; DM=17;
/M:	SY='R': M=-5, 14,-19, 2,-8, 50,-8,-2,-12,-4,-18,-11, 27,-28,-8,-3, 7, 0,-27,-30,-12,-8;
/M:	SY='T': M=-14,-24,-24,-30,-24, 60,-24,-8, 8,-22, 8, 2,-20,-29,-28,-17,-20,-10, 0, 8, 27,-29;
/I:	I=-8; M0=12;
/M:	SY='V': M=-4,-21,-18,-20,-21, 7,-30,-29, 24,-24, 18, 12,-22,-24,-22,-20,-18,-4, 23,-21,-1,-20; D=-4;
/I:	I=-6; M0=12; IM=12; DM=12;
/M:	SY='A': M=15,-14,-14,-20,-15,-7,-14,-19, 1,-15,-2,-8,-22,-27,-13,-17, 4, 8, 8,-23,-11,-14;
/I:	I=-5; M0=12;

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