Week 14 Annotation

Monday, 18 April 2022 8:40 AM

1. Computation phase
   - ↳ proteins are identified
   - ↳ ab initio or evidence based

2. annotation phase
   - ↳ synthesize annotation

STEP 1 Repeat identification.

↳ transposons, LINEs, SINEs, viral seqs.

↳ RepeatMasker

STEP 2 Evidence alignment

eg.1 RNA seq. aligned to assembly

eg.2. ab initio gene prediction

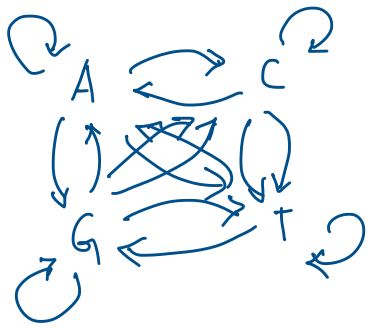
(eg. motifs → AUG (start), stop codons, TATAA boxes, etc).

STEP 3: UniProtKB, SwissProt, PDB.

(linking genes to proteins)

STEP 4: visualization (genome browsers)

HMM's $\{A, C, G, T\}$



TRANSITION PROB $= a_{st}$

= Prob. of transitioning from state 's' to state 't'

$$a_{st} = p(x_i = t \mid x_{i-1} = s) \quad \leftarrow (1)$$

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

$$= p(x_L \mid x_{L-1}, x_{L-2}, \ldots, x_1) \times p(x_{L-1} \mid x_{L-2}, \ldots, x_1) \times \ldots \times p(x_2 \mid x_1) \times p(x_1)$$

$$= p(x_L \mid x_{L-1}) \times p(x_{L-1} \mid x_{L-2}) \times p(x_{L-2} \mid x_{L-3}) \times \ldots \times p(x_2 \mid x_1) \times p(x_1)$$

Markov property

$$\therefore p(x) = p(x_1) \prod_{i=2}^{L} a_{x_{i-1} x_i} \quad \leftarrow (2)$$

in your HMM:

observed

state sequence path $= \pi$

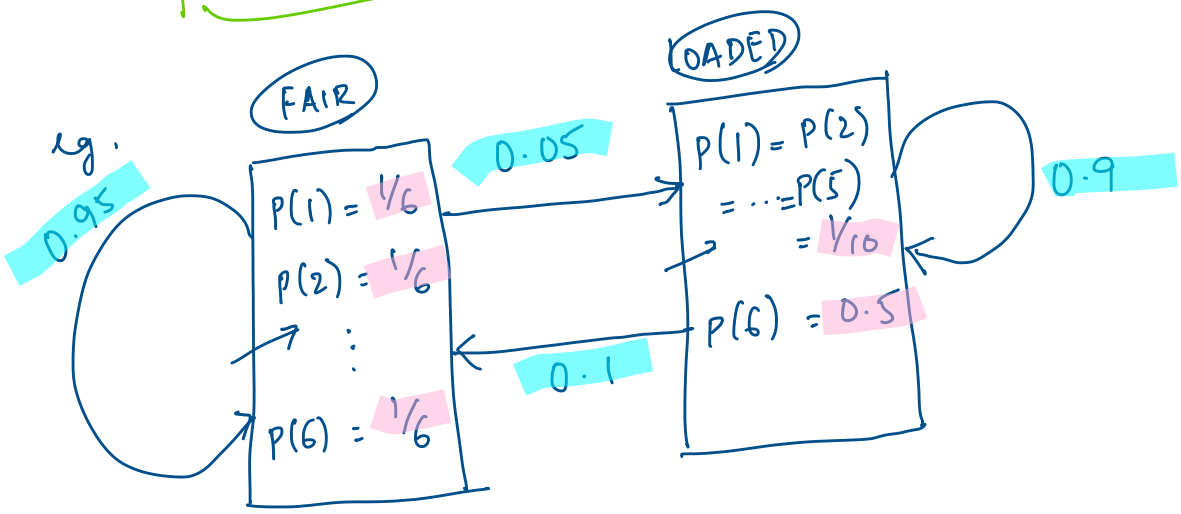
$$a_{kl} = p(\pi_i = l \mid \pi_{i-1} = k)$$

TRANSITION PROBABILITIES

'SYMBOLS' → (LABELS) → hidden states

$e_k(b) = P(x_i = b \mid \pi_i = k)$ ← EMISSION PROBABILITIES.

x = unobserved label path



VITERBI ALGORITHM

to pick best 'next' step, compute the highest probability seq.

$\pi^* = \arg\max_{\pi} P(x, \pi)$ ← state.

$v_k(i)$ = prob. of the most probable path ending in state k with observed symbol i

then $v_l(i+1) = e_l(x_{i+1}) \max (v_k(i)\, a_{kl})$ ↓ transition

l

Prob. of observing l in i^{th} state.

Prob. of transition of $k \to l$

eg. 2 1 2 3 6 2 4 .....

1 2 3 4 5 6 7

$$V_L(6,5) = e_L(6) \max\left(V_L(3,4), V_F(3,4)\right)$$

Fair coin or a biased coin

FAIR COIN

	H	T
H	0.5	0.5
T	0.5	0.5

LOADED COIN

	H	T
H	0.7	0.3
T	0.7	0.3

EMISSION PROBS

TRANSITION PROB.

	F	L
F	0.95	0.05
L	0.1	0.90