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

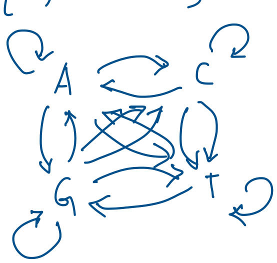
STEP 1 Repeat identification.
↳ transposons, LINEs, SINEs, viral seqs.
↳ Repeat Masker

STEP 2 Evidence alignment
eg. 1 RNA seq. aligned to assembly
eg. 2. ab initio gene prediction
(eg. motifs → AUG (start), TATAA boxes, etc).
stop codons,

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'

$P(x_i = t \mid x_{i-1} = s)$ ← 1

$a_{st} = \underbrace{\quad , \ldots , \quad , \quad}$

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

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

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

Markov property

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

observed

in your HMM:

$\boxed{\text{state sequence path} = \pi}$

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

TRANSITION PROBABILITIES

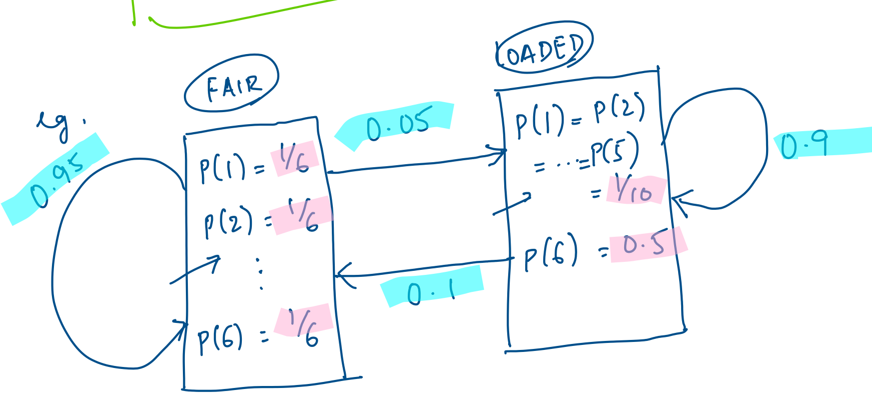
'SYMBOLS' → (LABELS) → hidden states

$e_k(b) = P(\underbrace{x_i = b} \mid \pi_i = k)$

EMISSION PROBABILITIES.

$\boxed{x = \text{unobserved label path}}$

eg.



VITERBI ALGORITHM

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

highest probability $P(x, \pi)$

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

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

then

$$v_l(i+1) = e_l(x_{i+1}) \max_k \left(v_k(i)\, a_{kl}\right)$$

$e_l(x_{i+1})$ ↓ Prob. of observing l in ith state.

a_{kl} ↓ 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

$$\begin{array}{c|cc} & H & T \\ \hline H & 0.5 & 0.5 \\ T & 0.5 & 0.5 \end{array}$$

LOADED COIN

$$\begin{array}{c|cc} & H & T \\ \hline H & 0.7 & 0.3 \\ T & 0.7 & 0.3 \end{array}$$

EMISSION PROBS

TRANSITION PROB.

$$\begin{array}{c|cc} & F & L \\ \hline F & 0.95 & 0.05 \\ L & 0.1 & 0.90 \end{array}$$

FORWARD ALGORITHM

Full prob. of a sequence x

$$P(x) = \sum_{x} P(x, \pi)$$

$$f_k(i) = P(x_1, x_2, \dots, x_i, \pi_i = k)$$

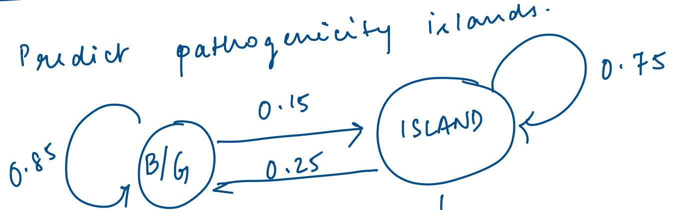
Prob. of the observed sequence up to and including x_i ; such that $\pi_i = k$

$$f_l(i+1) = \underbrace{e_l(x_i+1)}\sum_k f_k(i)\, a_{kl}$$

$$G(i|x) = \sum_k \overbrace{P(\pi_i = k | x)}\, g(k)$$

$g(k)$ = PRIOR on states

Predict pathogenicity islands.



A, T, C, G
P = 0.25

A : 0.15
T : 0.13
G : 0.3
C : 0.42

TRANSITION

	B	I
B	0.85	0.15
I	0.25	0.75

EMISSION.

	A	T	C	G
B	0.25	0.25	0.25	0.25
I	0.15	0.13	0.3	0.42