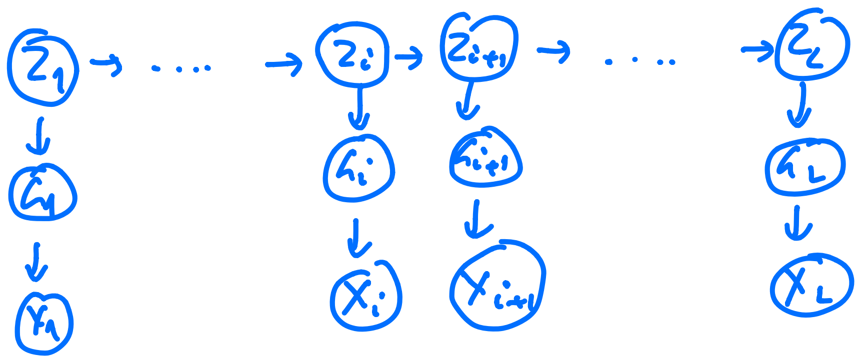
W07 - Hidden Markov Models - Inference
Expectation - Maximization

Inference by dynamic programming



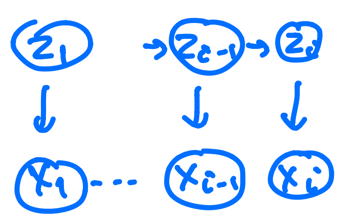
Data = $[X_1 \ldots X_L]$ reads $X_i = \{n_a^i, n_c^i, n_g^i, n_t^i\}$

Inference $[Z_1 \ldots Z_L]$ ancestry $Z_i \in \{AA, BB, AB\}$

Ignored (integrated) $[G_1 \ldots G_L]$ genome $G_i \in \{aa, ac, ag, \ldots\}$

Forward algorithm

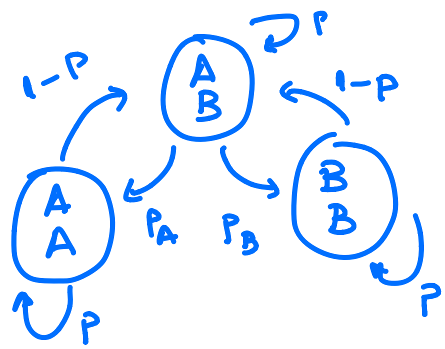
$$f_z(i) = P(X_1 \ldots X_{i-1}\, X_i\, Z_i = z)$$



$$f_z(i) = \sum_{Z_{i-1}=AA,AB,BB} f_{Z_{i-1}}(i-1) \cdot P(z \mid Z_{i-1})\, P(X_i \mid z)$$

Forward

$$\boxed{f_z(i) = \sum_{z_{i-1}=\substack{ABA\\ABB}} f_{z_{i-1}}(i-1)\, P(z|z_{i-1})\, P(x_i|z)}$$



$$\begin{aligned} f_z(i) = {} & f_{AA}(i-1)\, P(z|AA)\, P(x_i|z) \\ & + f_{BB}(i-1)\, P(z|BB)\, P(x_i|z) \\ & + f_{AB}(i-1)\, P(z|AB)\, P(x_i|z) \end{aligned}$$

$z = AA$

$$\begin{aligned} f_{AA}(i) = {} & f_{AA}(i-1)\, p \cdot P(x_i|AA) \\ & + f_{AB}(i-1)\, p_A\, P(x_i|AA) \end{aligned}$$

$z = BB$

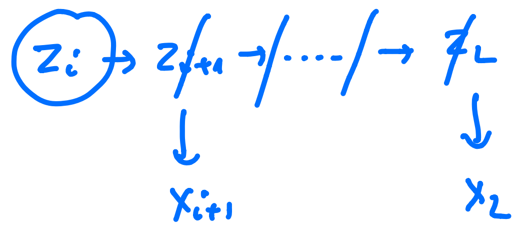
$$\begin{aligned} f_{BB}(i) = {} & f_{BB}(i-1)\, p\, P(x_i|BB) \\ & + f_{AB}(i-1)\, p_B\, P(x_i|BB) \end{aligned}$$

$z = AB$

$$\begin{aligned} f_{AB}(i) = {} & f_{AB}(i-1)\, p\, P(x_i|AB) \\ & + f_{AA}(i-1)\,(1-p)\, P(x_i|AB) \\ & + f_{BB}(i-1)\,(1-p)\, P(x_i|AB) \end{aligned}$$

Init: $f_z(1) = P_r(z)\, P(x_1|z)$

Backward Algorithm



$$b_z(i) = P(x_{i+1} \dots x_L \mid z_i = z)$$

$$b_z(i) = \sum_{z_{i+1}} \dots \sum_{z_L} P(x_{i+1} z_{i+1} \dots x_L z_L \mid z_i = z)$$

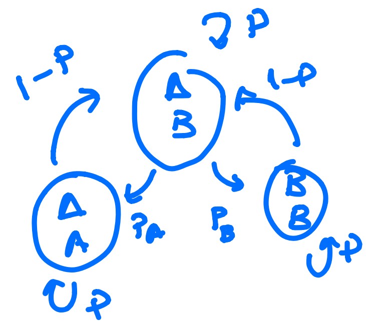
$$= \sum_{z_{i+1}} \sum_{z_{i+2}} \dots \sum_{z_L} P(x_{i+1} z_{i+1} \mid z_i = z) \cdot P(x_{i+2} z_{i+2} \dots x_L z_L \mid z_{i+1})$$

$$= \sum_{z_{i+1}} P(x_{i+1} z_{i+1} \mid z_i = z)\, b_{z_{i+1}}(i+1)$$

$$= \sum_{z_{i+1}} P(x_{i+1} \mid z_{i+1})\, P(z_{i+1} \mid z) \cdot b_{z_{i+1}}(i+1)$$

$$\boxed{b_z(i) = \sum_{z_{i+1}} b_{z_{i+1}}(i+1) \cdot P(z_{i+1} \mid z)\, P(x_{i+1} \mid z_{i+1})}$$

$$\begin{array}{ccccccccc} Z_i & \to & Z_{i+1} & \to & Z_{i+2} & \to & \cdots & \to & Z_L \\ & & \downarrow & & \downarrow & & & & \downarrow \\ & & X_{i+1} & & X_{i+2} & & & & X_L \end{array}$$



$$b_Z(i) = b_{AA}(i+1)\, P(AA|Z)\, P(X_{i+1}|AA)$$
$$+\, b_{BB}(i+1)\, P(BB|Z)\, P(X_{i+1}|BB)$$
$$+\, b_{AB}(i+1)\, P(AB|Z)\, P(X_{i+1}|AB)$$

$\underline{Z = AA}$

$$b_{AA}(i) = b_{AA}(i+1)\, p\, P(X_{i+1}|AA) + b_{AB}(i+1)\,(1-p)\, P(X_{i+1}|AB)$$

$\underline{Z = BB}$

$$b_{BB}(i) = b_{BB}(i+1)\, p\, P(X_{i+1}|BB) + b_{AB}(i+1)\,(1-p)\, P(X_{i+1}|AB)$$

$\underline{Z = AB}$

$$b_{AB}(i) = b_{AB}(i+1)\, p\, P(X_{i+1}|AB) + b_{AA}(i+1)\, p_A\, P(X_{i+1}|AA) + b_{BB}(i+1)\, p_B\, P(X_{i+1}|BB)$$

<u>Initialization</u>

$$b_{AA}(L) = b_{AB}(L) = b_{BB}(L) = 1$$

$$P(X_1 \ldots X_L | Z_L = AA) = 1$$

$P(x_1 \dots x_L)$ = Likelihood

$$P(x_1 \dots x_L) = \sum_{z_i} P(\overbrace{x_1 \dots x_{i-1}\, x_i\, z_i}^{B}\, \overbrace{x_{i+1} \dots x_L}^{A})$$

$$= \sum_{z_i} P(x_{i+1} \dots x_L \mid z_i)\; P(x_1 \dots x_{i-1}\, x_i\, z_i)$$

$$= \sum_{z_i} b_{z_i}(i) \cdot f_{z_i}(i)$$

$$\boxed{P(x_1 \dots x_L) = f_{AA}(i)\, b_{AA}(i) + f_{BB}(i)\, b_{BB}(i) + f_{AB}(i)\, b_{AB}(i) \quad \forall i}$$

Good test for forward/backward corectness

Decoding - Finally the Inference we are up to!

$$P(z_i \mid x_1 .. x_L) = P(z_i \mid \text{Data}, \text{HMM } p)$$

3 cases

$$P(z_i = AA \mid x_1 .. x_L)$$
$$P(z_i = BB \mid x_1 ... x_L)$$
$$P(z_i = AB \mid x_1 .. x_L)$$
$$\sum = 1$$

$$P(\overset{A}{z_i} \mid \overset{B}{x_1 ... x_L}) = \frac{P(x_1 ... x_i z_i ... x_L)}{P(x_1 ... x_L)}$$

$$= \frac{P(\overline{x_1 ... x_i z_i}^{B} \; \overline{x_{i+1} ... x_L}^{A})}{P(x_1 ... x_L)}$$

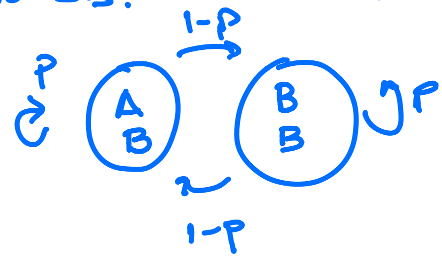
$$= \frac{P(x_{i+1} ... x_L \mid z_i) \cdot P(x_1 ... x_i z_i)}{P(x_1 ... x_L)}$$

$$\boxed{P(z_i \mid x_1 .. x_L) = \frac{b_{z_i}(i) \cdot f_{z_i}(i)}{\sum_z b_z(i) f_z(i)}}$$

→ class code

W07 - How much does the value of p matter?

Consider only backcrosses to BB. AA not possible



$$P(z_i | z_{i-1}) = \begin{cases} p & z_i = z_{i-1} \\ 1-p & z_i \neq z_{i-1} \end{cases}$$

$$P(\overbrace{\begin{smallmatrix}A\\B\end{smallmatrix} \cdots \begin{smallmatrix}A\\B\end{smallmatrix}}^{n} \begin{smallmatrix}B\\B\end{smallmatrix}) = p^n (1-p)$$

$$P(\begin{smallmatrix}B\\B\end{smallmatrix} \cdots \begin{smallmatrix}B\\B\end{smallmatrix} \begin{smallmatrix}A\\B\end{smallmatrix}) = p^n (1-p)$$

$$\langle n \rangle = \frac{p}{1-p}$$

Drosophila $\langle n \rangle = L/2$

chr 2L len $= 23.01 \cdot 10^6$ $\quad p = 0.99999991$

chr 4 len $= 1.35 \cdot 10^6$ $\quad p = 0.99999852$?

does it matter?

→ class code

Maximum Likelihood estimates

To do ML estimates, we need "labelled data".

i.e. Data for which z_i are known.

$$D = \{ x_1^m z_1^m \dots x_L^m z_L^m \}_{m=1}^{M} \qquad M = \# \text{ of flies}$$

$$P(D \mid HMM) = \prod_{m=1}^{M} P(x_1^m z_1^m \dots x_L^m z_L^m \mid p)$$

$$= \prod_{m=1}^{M} P(z_1^m) P(x_1^m \mid z_1^m) \prod_{i=2}^{L} P(z_i^m \mid z_{i-1}^m) P(x_i^m \mid z_i^m)$$

$$C_i^m(AA \to AA) + C_i^m(AB \to AB) = C_i^m(\text{same})$$

$$C_i^m(AB \to BB) + C_i^m(BB \to AB) = C_i^m(\text{break})$$

$$P(D \mid p) = \prod_{m=1}^{M} P(z_i^m) \prod_{i=1}^{L} p^{C_i^m(\text{same})} (1-p)^{C_i^m(\text{break})} \cdot \prod_{i=1}^{L} P(x_i \mid z_i)$$

$$\log P(D \mid p) \propto \sum_{m=1}^{M} \sum_{i=1}^{L} C_m^i(s) \log p + \sum_m \sum_i C_m^i(b) \log(1-p)$$

$$\log P(D \mid p) \propto C(s) \log p + C(b) \log(1-p)$$

$$C(s) = \sum_m \sum_i \left[C_m^i(AB \rightarrow AB) + C_m^i(BB \rightarrow BB) \right]$$

$$C(b) = \sum_m \sum_i \left[C_m^i(AB \rightarrow BB) + C_m^i(BB \rightarrow AB) \right]$$

$$\frac{\delta \log P(D \mid p)}{\delta p} = \frac{C(s)}{p} - \frac{C(b)}{1-p} = 0$$

$$\frac{C(s)}{p_{ML}} = \frac{C(b)}{1-p^*}$$

$$\boxed{p^* = \frac{C(s)}{C(s) + C(b)}}$$