W06 - Ancestry Inference from RNA-seq reads using a hidden Markov model (HMM)

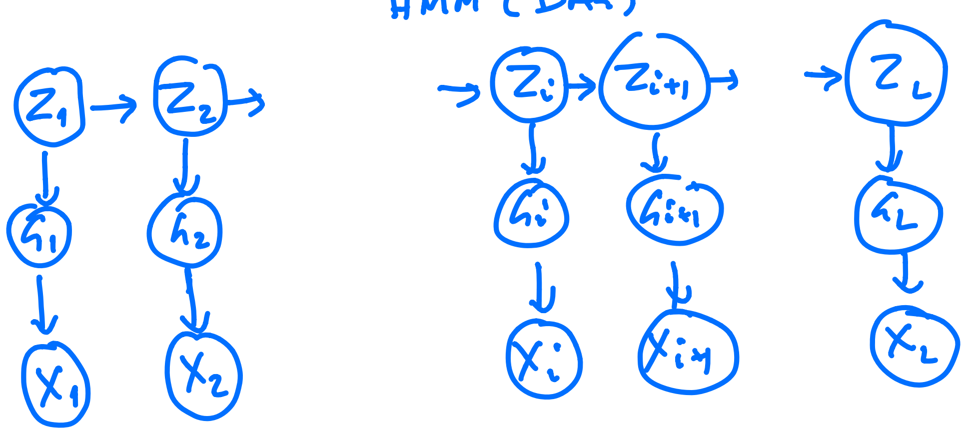
backcrossed flies $\quad \frac{A}{A} \times \frac{A}{B} \qquad \frac{B}{B} \times \frac{A}{B}$

2 breakpoints $\qquad\qquad$ 2 breaks

1 break $\qquad\qquad$ 1 break.

$Z = \left\{ \frac{A}{A}\ \frac{B}{B}\ \frac{A}{B} \right\}$ ancestry, z_i for each position i

HMM (DAG)



h_i = the 2 nts at the genome in that position. $\left(\frac{a}{a}, \frac{a}{c}, \frac{a}{g}, \frac{g}{t}, \frac{c}{c}, \frac{c}{g}, \ldots\right)$

$x_i = (n_a^i\ n_c^i\ n_g^i\ n_t^i)$ the reads at position i

$$P(x_1 q_1 z_1 \ldots x_L q_L z_L) = P(x_L q_L z_L \mid x_{L-1} q_{L-1} z_{L-1})$$

from DAG

$$\vdots$$

$$P(x_2 q_2 z_2 \mid x_1 q_1 z_1)$$

$$P(x_1 q_1 z_1)$$

$$= P(x_1 q_1 z_1) \cdot \prod_{i=2}^{L} P(x_i q_i z_i \mid x_{i-1} q_{i-1} z_{i-1})$$

From DAG:

$$P(x_i q_i z_i \mid x_{i-1} q_{i-1} z_{i-1}) = P(z_i \mid z_{i-1})\, P(q_i \mid z_i)\, P(x_i \mid q_i)$$

$\{x_i\}$ data

$\{z_i\}$ variables we want to know about

$\{q_i\}$ hidden variables, we need them to go from $z_i \to x_i$
but their actual values are not interesting to us

$$P(x_1 z_1 \ldots x_L z_L) = \sum_{q_1} \cdots \sum_{q_L} P(x_1 q_1 z_1 \ldots x_L q_L z_L)$$

$$= P(z_1) \sum_{q_1} P(x_1 \mid q_1) \cdot P(q_1 \mid z_1) \prod_{i=2}^{L} P(z_i \mid z_{i-1}) \cdot \underbrace{\sum_{q_i} P(x_i \mid q_i) P(q_i \mid z_i)}_{P(x_i \mid q_i)}$$

$$\boxed{P(X_1 Z_1 \ldots X_L Z_L) = P(Z_1)\,P(X_1|Z_1)\prod_{i=2}^{L} P(Z_i|Z_{i-1})\,P(X_i|Z_i)}$$

The parameters, where are they?

$P(Z_i|Z_{i-1})$ " 3 probability distributions

$P\left(\begin{smallmatrix}A\\A\end{smallmatrix}\middle|\begin{smallmatrix}A\\A\end{smallmatrix}\right) = p$

$P\left(\begin{smallmatrix}B\\B\end{smallmatrix}\middle|\begin{smallmatrix}A\\A\end{smallmatrix}\right) = 0$

$P\left(\begin{smallmatrix}A\\B\end{smallmatrix}\middle|\begin{smallmatrix}A\\A\end{smallmatrix}\right) = 1-p$

$P\left(\begin{smallmatrix}A\\A\end{smallmatrix}\middle|\begin{smallmatrix}B\\B\end{smallmatrix}\right) = 0$

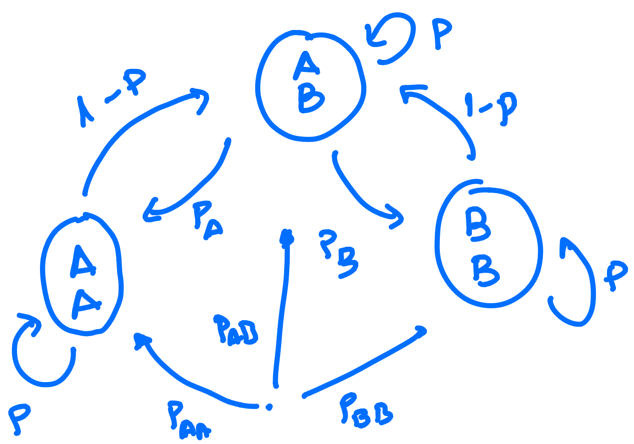
$P\left(\begin{smallmatrix}B\\B\end{smallmatrix}\middle|\begin{smallmatrix}B\\B\end{smallmatrix}\right) = p$

$P\left(\begin{smallmatrix}A\\B\end{smallmatrix}\middle|\begin{smallmatrix}B\\B\end{smallmatrix}\right) = 1-p$

$P\left(\begin{smallmatrix}A\\A\end{smallmatrix}\middle|\begin{smallmatrix}A\\B\end{smallmatrix}\right) = p_A$

$P\left(\begin{smallmatrix}B\\B\end{smallmatrix}\middle|\begin{smallmatrix}A\\B\end{smallmatrix}\right) = p_B$

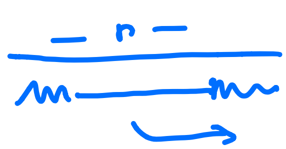
$P\left(\begin{smallmatrix}A\\B\end{smallmatrix}\middle|\begin{smallmatrix}A\\B\end{smallmatrix}\right) = p$



$p + p_A + p_B = 1$

State diagram $\neq$ DAG

p?



length between breaks = geometric distribution

$p^n(1-p) = P(n) \qquad \langle n \rangle = \frac{p}{1-p}$

the "emission" parameters

$$Z_i \to G_i \to X_i$$

$$P(G_i|Z_i) \quad \begin{cases} P(G_i|AA) = \begin{cases} 1 & \text{if } G_i = g_i^A g_i^A \\ 0 & - \end{cases} \\ P(G_i|BB) = \begin{cases} 1 & \text{if } G_i = g_i^B g_i^B \\ 0 & - \end{cases} \\ P(G_i|AB) = \begin{cases} 1 & \text{if } G_i = g_i^A g_i^B \\ 0 & - \end{cases} \end{cases}$$

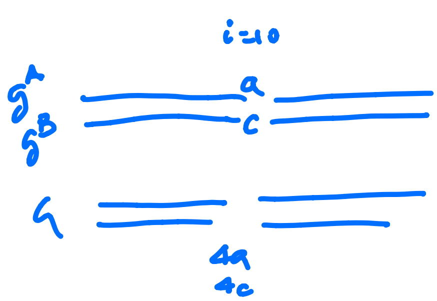
$P(X_i|G_i)$

- no parameters
- assumes no errors in the genomes
- Andolfatto has an ϵ error parameter

$$P(X_i|G_i) = \begin{cases} p_1^{n_a} q_1^{n_c+n_g+n_t} & \left.\begin{matrix} G_i = aa \\ \vdots \end{matrix}\right] \text{4 cases} \\ p_2^{n_a+n_c} q_2^{n_g+n_t} & \left.\begin{matrix} G_i = ac \\ \vdots \end{matrix}\right\} \text{6 cases} \end{cases}$$

$$\left.\begin{matrix} p_1 = 1-\epsilon \\ q_1 = \epsilon/3 \end{matrix}\right\} \quad \begin{matrix} p_2 = \frac{1-\epsilon}{2} \\ q_2 = \epsilon/2 \end{matrix}$$

$$p_1 + 3q_1 = 1 \qquad p_2 + p_2 + q_2 + q_2 = 1$$



$$P(G_{10}|AA) = 1 \quad \text{if } G_{10} = aa$$

$$P(G_{10}|BB) = 1 \quad \text{if } G_{10} = cc$$

$$P(G_{10}|AB) = 1 \quad \text{if } G_{10} = ac$$

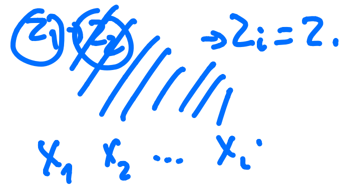
$$P(X_{10} \mid AA) = \sum_{G_{10}} P(X_{10} \mid G_{10}) P(G_{10} \mid AA)$$

$$= P(X_{10} \mid aa) = p_1^{4+4} q_1^{0} = p_1^{8}$$

$$P(X_{10} \mid AB) = P(X_{10} \mid ac) = p_2^{8} q_2^{0} = p_2^{8}$$

$$P(X_{10} \mid BB) = P(X_{10} \mid cc) = p_1^{4} q_1^{4}$$

The forward algorith



$$f_z(i) = P(x_1 x_2 \ldots x_i\, z_i = z)$$

$$= \sum_{z_1} \ldots \sum_{z_{i-1}} P(x_1 z_1, x_2 z_2, \ldots x_i\, z_i = z)$$

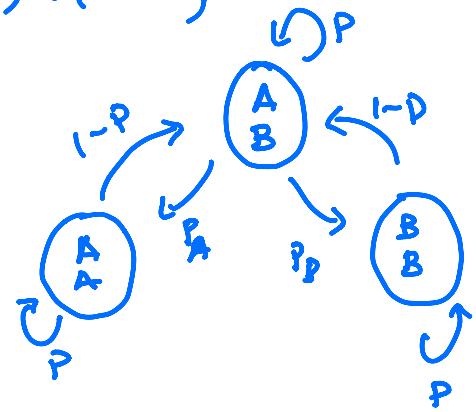
$$= \sum_{z_1} \sum_{z_{i-1}} P(x_1 z_1) \cdot P(x_2 z_2 | x_1 z_1) \ldots P(x_{i-1} z_{i-1} | x_{i-2} z_{i-2}) \; P(x_i\, z_i = z | x_{i-1} z_{i-1})$$

$$= \sum_{z_{i-1}} \underbrace{\sum_{z_1} \ldots \sum_{z_{i-2}} P(x_1 z_1) P(x_2 z_2 | x_1 z_1) \ldots P(x_{i-1} z_{i-1} | x_{i-2} z_{i-2})}_{f_{z_{i-1}}(i-1)} \cdot P(x_i\, z_i = z | x_{i-1} z_{i-1})$$

$$= \sum_{z_{i-1}} f_{z_{i-1}}(i-1) \cdot P(x_i\, z_i = z | x_{i-1} z_{i-1})$$

$$= \sum_{z_{i-1}} f_{z_{i-1}}(i-1) \cdot P(z_i = z | z_{i-1}) \, P(x_i | z)$$

$$f_z(i) = \sum_{z_{i-1}} f_{z_{i-1}}(i-1) \; P(z|z_{i-1}) \; P(x_i|z)$$



$z = AA \qquad z_{i-1} = AA, BB, AB$

$$f_{AA}(i) = + f_{AA}(i-1) \; P(AA|AA) \; P(x_i|AA)$$

$$+ f_{BB}(i-1) \; P(AA|BB) \; P(x_i|AA)$$

$$+ f_{AB}(i-1) \; P(AA|AB) \; P(x_i|AA)$$

$$= f_{AA}(i-1) \cdot p \cdot P(x_i|AA)$$

$$+ 0$$

$$+ f_{AB}(i-1) \cdot p_A \; P(x_i|AA)$$

$$f_{BB}(i) = + f_{AA}(i-1) \; P(BB|AA) \; P(x_i|AA)$$

$$+ f_{BB}(i-1) \; P(BB|BB) \; P(x_i|BB)$$

$$+ f_{AB}(i-1) \; P(BB|AB) \; P(x_i|BB)$$

$$= f_{BB}(i-1) \cdot p \; P(x_i|BB) + f_{AB}(i-1) \; p_B \; P(x_i|BB)$$

$$f_{AB}(i) = f_{AA}(i-1)\, P(AB|AA)\; P(x_i|AB)$$
$$+ f_{BB}(i-1)\, P(AB|BB)\; P(x_i|AB)$$
$$+ f_{AB}(i-1)\, P(AB|AB)\; P(x_i|AB)$$

$$= f_{AA}(i-1) \cdot (1-p)\; P(x_i|AB)$$
$$+ f_{BB}(i-1)\,(1-p)\; P(x_i|AB)$$
$$+ f_{AB}(i-1)\;\; p\;\; P(x_i|AB$$

$$f_{AA}(1) = P_r(AA) \cdot P(x_1|AA)$$
$$f_{BB}(1) = P_r(BB) \cdot P(x_1|BB)$$
$$f_{AB}(1) = P_r(AB) \cdot P(x_i|AB)$$

$$P(x_1 \ldots x_L) = \sum_{z_i} P(\overbrace{x_1 \ldots x_i}^{B}\; z_i = z\; \overbrace{x_{i+1} \ldots x_L}^{A})$$

$$= \sum_{z_i} P(x_{i+1} \ldots x_L \mid x_1 \ldots x_i z_i = z)$$
$$P(x_1 \ldots x_i z_i)$$

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

good check of your algorithm. $= f_{AA}(i)\, g_{AA}(i) + f_{BB}(i)\, g_{BB}(i)$
$+ f_{AB}(i)\, g_{AB}(i)$

Backward algorithm

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

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

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

$$= \sum_{z_{i+1}} P(x_{i+1}, z_{i+1} \mid z_i = z) \cdot \underbrace{\sum_{z_{i+2}} \dots \sum_{z_L} P(x_{i+2} z_{i+2} \mid x_{i+1} z_{i+1}) \cdots}_{b_{z_{i+1}}(i+1)}$$

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

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

$$b_z(i) = \sum_{z_{i+1}} P(z_{i+1} | z) P(x_{i+1} | z_{i+1}) b_{z_{i+1}}(i+1)$$

$$b_{AA}(i) = b_{AA}(i+1) P(AA | AA) P(x_{i+1} | AA)$$
$$+ b_{BB}(i+1) P(BB | AA) P(x_{i+1} | BB)$$
$$+ b_{AB}(i+1) P(AB | AA) P(x_{i+1} | AB)$$

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

$$b_{BB}(i) = + b_{AA}(i+1) P(AA | BB) P(x_{i+1} | AA)$$
$$+ b_{BB}(i+1) P(BB | BB) P(x_{i+1} | BB)$$
$$+ b_{AB}(i+1) P(AB | BB) P(x_{i+1} | AB)$$

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

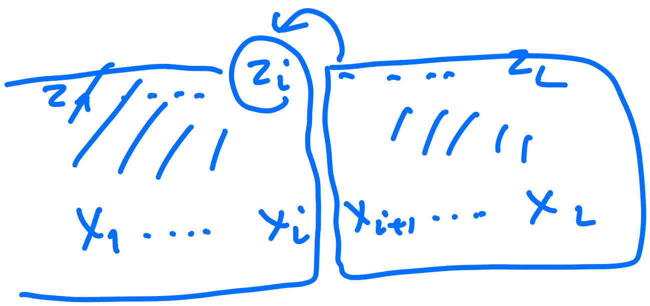
$$b_{AB}(i) = b_{AA}(i+1)\ P(AA|AB)\ P(x_{i+1}|AA)$$
$$+\ b_{BB}(i+1)\ P(BB|AB)\ P(x_{i+1}|BB)$$
$$+\ b_{AB}(i+1)\ P(AB|AB)\ P(x_{i+1}|AB)$$

$$= P_A\ b_{AA}(i+1)\ P(x_{i+1}|AA)$$
$$P_B\ b_{BB}(i+1)\ P(x_{i+1}|BB)$$
$$P\ b_{AB}(i+1)\ P(x_{i+1}|AB)$$

Initialization

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

Posterior decoding



$$P(z_i \mid x_1 \dots x_L) = \frac{P(x_1 \dots x_i z_i \dots x_L)}{P(x_1 \dots x_L)}$$

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

$$= \frac{f_{z_i}(i)\, b_{z_i}(i)}{P(x_1 \dots x_L)}$$

$$P(z_i = AA \mid x_1 \dots x_L) = \frac{f_{AA}(i)\, b_{AA}(i)}{f_{AA}(i)\, b_{AA}(i) + f_{BB}(i)\, b_{BB}(i) + f_{AB}(i)\, b_{AB}(i)}$$

$$P(z_i = BB \mid x_1 \dots x_L) = \frac{f_{BB}(i)\, b_{BB}(i)}{\text{same}}$$

$$P(z_i = AB \mid x_1 \dots x_L) = \frac{f_{AB}(i)\, b_{AB}(i)}{\text{same}}$$