CS 576

SPARSH AGARWAL
9075905142

A2) (a) $P(X=x, Y=y, O=4)$

$= P(X=x, Y=y \mid 4)\, P(O=4)$

$= \left(\prod_{i=1}^{\min(3,4)} q_{x_i}\right)\left(\prod_{i=4} P_{x_i y_i - 3}\right)\left(\prod_{j=\max(2,1)}^{4} q_{y_j}\right) P(O=4)$

$= \left(\prod_{i=1}^{3} q_{x_i}\right)\left(P_{x_4 y_1}\right)\left(\prod_{j=2}^{4} q_{y_j}\right) P(O=4)$

$= (q^3)(s)(q^3)(1-\theta)^3\theta$

$= q^6 s (1-\theta)^3\theta$

$= \left(\frac{1}{4}\right)^6 s \left(\frac{2}{3}\right)^3 \frac{2}{3}$

$= s\left(\frac{1}{4}\right)^4\left(\frac{1}{3}\right)^4 = 0.0000\,10334$

(b) $P(X=x, Y=y) = \sum_{O=1}^{\infty} P(X=x, Y=y, O=o)$

$= \sum_{O=1}^{4} P(X=x, Y=y, O=o) + \sum_{O=5}^{\infty} P(X=x, Y=y, O=o)$

$\Rightarrow$ Dividing the answer in 2 parts

$\sum_{O=5}^{\infty} P(X=x, Y=y, O=o) = \left(\prod_{i=1}^{4} q_{x_i}\right)\left(\text{not valid } \sum_{i=5}\right)\left(\prod_{j=1}^{4} q_{y_j}\right)(1-\theta)^5\theta$

$+ \quad " \qquad " \qquad " \quad (1-\theta)^5\theta$

$+ \quad " \qquad " \qquad " \quad (1-\theta)^6\theta$

$= q^8\left[(1-\theta)^4\theta + (1-\theta)^5\theta + (1-\theta)^6\theta + \ldots\ldots (1-\theta)^6\theta\right]$

$= q^8 (1-\theta)^4 \theta \left[1 + (1-\theta) + (1-\theta)^2 + (1-\theta)^3 + \cdots\right]$

$= q^8 (1-\theta)^4 \theta \left[\frac{1-(1-\theta)^{\infty}}{1-(1-\theta)}\right]$

$= q^8 (1-\theta)^4$

⇒ First part, $\sum_{i=1}^{4} P(X=x, Y=y, \theta=\theta)$

$= P(X=x, Y=y \mid \theta=1) P(\theta=1) + P(X=x, Y=y \mid \theta=2) P(\theta=2) +$
$P(X=x, Y=y \mid \theta=3) P(\theta=3) + P(X=x, Y=y \mid \theta=4) P(\theta=4)$

$= (P_{x_1y_1} \times P_{x_2y_2} \times P_{x_3y_3} \times P_{x_4y_4})\theta + q^2 (P_{x_2y_1} \times P_{x_3y_2} + P_{x_4y_3})(1-\theta)\theta$
$+ q^4 (P_{x_3y_1} \times P_{x_4y_2})(1-\theta)^2\theta + q^6 (1-\theta)^3 \theta (P_{x_4y_1})$

$= (m \times m \times s \times s)\theta + q^2 (m \times s \times m)(1-\theta)\theta + q^4 (m \times m)(1-\theta)^2\theta$
$+ q^6 (1-\theta)^3 \theta (s)$

$= m^2 s^2 \theta + m^2 q^2 s (1-\theta)\theta + q^4 m^2 (1-\theta)^2\theta + q^6 s (1-\theta)^3\theta$

∴ Total answer is

$= m^2 s^2 \theta + m^2 q^2 s(1-\theta)\theta + q^4 m^2 (1-\theta)^2\theta + q^6 s (1-\theta)^3\theta + q^8 (1-\theta)^4$

$= \frac{1}{1382976} + \frac{1}{2370816} + \frac{1}{12192768} + \frac{1}{143536} + \frac{1}{331776}$

$= 0.000000723 + 0.000000422 + 0.000000082 + 0.000005167 + 0.000003014$

$= 0.000009408$

$P(\theta = 4 \mid X=x, Y=y)$

(c) $= P(\theta = 4, X=x, Y=y) \div P(X=x, Y=y)$

∴ taking answers from first two parts

$$= \frac{0.0000010334}{0.000009408}$$

(d) $P(\theta \leq 4 \mid X=x, Y=y)$

using previous parts

$$P(\theta \leq 4 \mid X=x, Y=y) = \frac{\sum_{\theta=1}^{4} P(\theta=\theta, X=x, Y=y)}{P(X=x, Y=y)}$$

$$= \frac{0.000009408 - 0.000003014}{0.000009408}$$

(e) considering that q^n is valid for $n=1$, there are 16 possibilities of $P(X=x, Y=y)$

$$\sum_{\theta=1}^{\infty} P(X=x, Y=y, \theta=\theta) = \sum_{\theta=1}^{1} P(X=x, Y=y, \theta=\theta) + \sum_{\theta=2}^{\infty} P(X=x, Y=y, \theta=\theta)$$

$$= \theta \left(\prod_{i=1}^{0} q_{x_i} \right) \left(\prod_{i=1}^{1} p_{x_i, y_i} \right) \left(\prod_{j=\max(2,1)}^{1} q_{y_j} \right)$$

$$+ (1-\theta)\theta + (1-\theta)^2\theta + (1-\theta)^3\theta \cdots$$

$$= \theta m + (1-\theta)\theta \left[\frac{1-(1-\theta)^{\infty}}{1-(1-\theta)} \right]$$

$$= \theta m + 1 - \theta$$

$\therefore$ the total probability $= 1 = 16\left(\theta_m + 1 - \theta\right)$

$\therefore 16\left(\theta_m + \frac{2}{3}\right) = 1$

$\theta\, m + \frac{2}{3} = \frac{1}{16}$

$\frac{m+2}{3} = \frac{1}{16}$

$m = \frac{3}{16} - 2$

$m = \frac{-29}{16}$

A3) (b) Seed methods like Blast are very fast & reliable in a statistical sence. (very good for large sequences) Whereas, DP methods like Needleman-Wunsch alignment are relatively slow & computational steps increase as the square or cube of the sequence lengths.

(a) In protein alignment, we look for ~~transpol~~ transpositions inversions, deletions, insertions & substitutions. For short proteins we consider only substitutions & insertion/deletions which are represented as ~~mismatches~~ match/mismatch & gaps respectively. But for long proteins we also consider transposition & inversion.

Some insertion & deletions may not significantly affect the structure of protein ∴ we need to look for similarities instead of just match/mismatch. We look for partial matches (i.e. some amino acid pairs are more substitutable than others) We represent this similarity in form of a score matrix, where we produce all possible scores for different alignments feasable. Then we select the alignment with max score.