

# INF281 Exercise 09 solutions

## 1. BLOSUM

BLOSUM uses several thousand blocks to calculate the probabilities of accepted mutation. Use the following definitions and Block1 & Block2 to solve the problems.

$f_{ab}$  : Frequencies of an observed pair  $a$  and  $b$ .

$T$  : Total number of pairs from all blocks.

The number of pairs can be calculated as  $1/2wm(m-1)$ .

$$p_a : p_a = f_{aa} + \sum_{e \neq a} f_{ae}/2$$

$$e_{aa} : p_a p_a$$

$$e_{ab} : p_a p_b + p_b p_a = 2p_a p_b$$

|        |        |
|--------|--------|
| Block1 | Block2 |
| CAGC   | GGA    |
| GTAC   | GTA    |
| CAGC   |        |

(a) Count the occurrences of all pairs.

|   | A | G | C | T |
|---|---|---|---|---|
| A | 2 | 2 | 0 | 2 |
| G | 2 | 2 | 2 | 1 |
| C | 0 | 2 | 4 | 0 |
| T | 2 | 1 | 0 | 0 |

(b) Calculate  $T$ .

**Solution:**  $(1/2 \times 4 \times 3 \times 2) + (1/2 \times 3 \times 2 \times 1) = 12 + 3 = 15$

(c) Calculate  $f_{AA}$  and  $f_{AG}$ .

**Solution:**  $f_{AA} : 2/15, \quad f_{AG} : 2/15$

(d) Calculate  $p_A$  and  $p_G$ .

**Solution:**  $p_A : 8/30, \quad p_G : 9/30$

(e) Calculate  $e_{AA}$  and  $e_{AG}$ .

**Solution:**  $e_{AA} : 64/900, \quad e_{AG} : 144/900$

(f) Calculate  $f_{AA}/e_{AA}$  and  $f_{AG}/e_{AG}$ .

**Solution:**  $f_{AA}/e_{AA} : 1.875$ ,  $f_{AG}/e_{AG} : 0.833$

## 2. PPM (Position probability matrix)

PWM (position weight matrix) is a popular method to find sequenced patterns. It can be generated from PPM (position probability matrix) and PFM (position frequency matrix).

Seq1 CAA

Seq2 CAG

Seq3 GAC

Seq4 ATT

(a) Create a PFM from Seq1, Seq2, Seq3, and Seq4.

|   | 1 | 2 | 3 |
|---|---|---|---|
| A | 1 | 3 | 1 |
| G | 1 | 0 | 1 |
| C | 2 | 0 | 1 |
| T | 0 | 1 | 1 |

(b) Create a PPM from Seq1, Seq2, Seq3, and Seq4.

|   | 1    | 2    | 3    |
|---|------|------|------|
| A | 0.25 | 0.75 | 0.25 |
| G | 0.25 | 0    | 0.25 |
| C | 0.5  | 0    | 0.25 |
| T | 0    | 0.25 | 0.25 |

## 3. Sequence profile

A sequence profile is similar to PWM, but it uses a scoring scheme. Use the following definitions to calculate the profile values.

$$Prof_{ra} : \frac{1}{m_r} \sum_{b \in M} R_{ba} F_{rb}$$

$F_{rb}$  : The number of occurrences of  $b$  at position  $r$

$R_{ba}$  : Pairwise score between  $b$  and  $a$

$m_r$  : The number of residues without gaps at position  $r$

Scoring matrix:

|   | A  | G  | C  | T  |
|---|----|----|----|----|
| A | 2  | 1  | -3 | -2 |
| G | 1  | 3  | -2 | -1 |
| C | -3 | -2 | 4  | 1  |
| T | -2 | -1 | 1  | 2  |

## MSA

Seq1 GT  
Seq2 -G  
Seq3 CA

(a) Calculate the profile values of position 1.

$$A1: (1/2) \times ((2 \times 0) + (1 \times 1) + (-3 \times 1) + (-2 \times 0)) = -1$$

$$G1: (1/2) \times ((1 \times 0) + (3 \times 1) + (-2 \times 1) + (-1 \times 0)) = 1/2$$

$$C1: (1/2) \times ((-3 \times 0) + (-2 \times 1) + (4 \times 1) + (1 \times 0)) = 1$$

$$T1: (1/2) \times ((-2 \times 0) + (-1 \times 1) + (1 \times 1) + (2 \times 0)) = 0$$

(b) Calculate the profile values of position 2.

$$A2: (1/3) \times ((2 \times 1) + (1 \times 1) + (-3 \times 0) + (-2 \times 1)) = 1/3$$

$$G2: (1/3) \times ((1 \times 1) + (3 \times 1) + (-2 \times 0) + (-1 \times 1)) = 1$$

$$C2: (1/3) \times ((-3 \times 1) + (-2 \times 1) + (4 \times 0) + (1 \times 1)) = -4/3$$

$$T2: (1/3) \times ((-2 \times 1) + (-1 \times 1) + (1 \times 0) + (2 \times 1)) = -1/3$$

(c) Make a profile matrix.

|   | 1   | 2    |
|---|-----|------|
| A | -1  | 1/3  |
| G | 1/2 | 1    |
| C | 1   | -4/3 |
| T | 0   | -1/3 |

## 4. Profile search

A sequence profile can take gap penalties into account. Calculate the score of the alignment between the DNA profile below and a DNA segment.

A DNA profile of length 4

|   | A  | G  | C  | T  | Gap |
|---|----|----|----|----|-----|
| 1 | 5  | -5 | -2 | -1 | 10  |
| 2 | -2 | 3  | 4  | -7 | 10  |
| 3 | 1  | 2  | 1  | -1 | 5   |
| 4 | -3 | 3  | -2 | 7  | 10  |

- P1, P2, P3, P4: profile blocks at positions 1 - 4
- Gap penalty (for segments): 4

(a) Profile search on segment D1

|          |    |    |    |    |
|----------|----|----|----|----|
| Profile: | P1 | P2 | P3 | P4 |
| D1:      | A  | C  | G  | T  |

**Solution:**  $5 + 4 + 2 + 7 = 18$

(b) Profile search on segment D2

|          |    |    |   |    |    |
|----------|----|----|---|----|----|
| Profile: | P1 | P2 | - | P3 | P4 |
| D2:      | A  | C  | C | G  | T  |

**Solution:**  $5 + 4 + (-5) + 2 + 7 = 13$

(c) Profile search on segment D3

|          |    |    |    |    |
|----------|----|----|----|----|
| Profile: | P1 | P2 | P3 | P4 |
| D3:      | A  | -  | G  | T  |

**Solution:**  $5 + (-4) + 2 + 7 = 10$

(d) Profile search on segment D4

|          |    |    |   |    |    |
|----------|----|----|---|----|----|
| Profile: | P1 | P2 | - | P3 | P4 |
| D4:      | -  | A  | C | G  | T  |

**Solution:**  $(-4) + (-2) + (-5) + 2 + 7 = -2$