# 11 Exercises – Construction of scoring matrix

1. **Scoring schemes for protein alignments**

   Calculate the score of the alignment by using different scoring schemes.

```
Seq1 R-HIC
Seq2 RDDCC
```

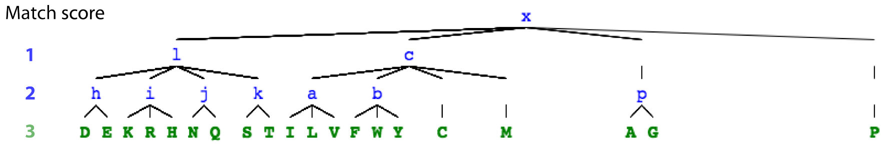
(a) Use the identity with a simple scoring scheme as match: 1, mismatch: 0, and gap penalty: 0.

(b) Use the genetic code.

| *First* | *Second position* | | | | *Third* |
|---|---|---|---|---|---|
| *position* | T | C | A | G | *position* |
| T | F | S | Y | C | T |
| | F | S | Y | C | C |
| | L | S | *Stop* | *Stop* | A |
| | L | S | *Stop* | W | G |
| C | L | P | H | R | T |
| | L | P | H | R | C |
| | L | P | Q | R | A |
| | L | P | Q | R | G |
| A | I | T | N | S | T |
| | I | T | N | S | C |
| | I | T | K | R | A |
| | M | T | K | R | G |
| G | V | A | D | G | T |
| | V | A | D | G | C |
| | V | A | E | G | A |
| | V | A | E | G | G |

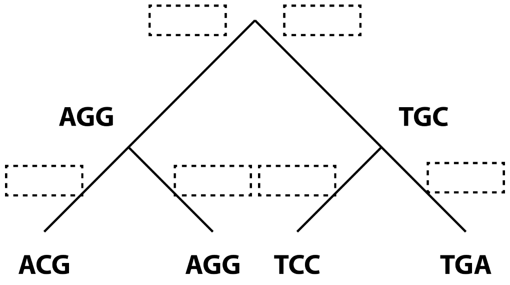
| | | |
|---|---|---|
| A | Ala | Alanine |
| C | Cys | Cysteine |
| D | Asp | Aspartic acid |
| E | Glu | Glutamic acid |
| F | Phe | Phenylalanine |
| G | Gly | Glycine |
| H | His | Histidine |
| I | Ile | Isoleucine |
| K | Lys | Lysine |
| L | Leu | Leucine |
| M | Met | Methionine |
| N | Asn | Asparagine |
| P | Pro | Proline |
| Q | Gln | Glutamine |
| R | Arg | Arginine |
| S | Ser | Serine |
| T | Thr | Threonine |
| V | Val | Valine |
| W | Trp | Tryptophan |
| Y | Tyr | Tyrosine |

(c) Use the AACH.



2. **Probabilities of accepted mutations**

   Use a phylogenetic tree below to calculate the probabilities of accepted mutations. The tree contains sequences of four OTUs.



   (a) Estimate the mutations and fill them in the boxes next to the edges.

   (b) Count the occurrences of mutations and fill them in the matrix. Note that a mutation A → B is equivalent with a mutation B → A.

|   | A | C | G | T |
|---|---|---|---|---|
| A |   |   |   |   |
| C |   |   |   |   |
| G |   |   |   |   |
| T |   |   |   |   |

   (c) Use the following definitions and calculate $f_{CG}$, $f_C$ and $f$.

$f_{ab}$ : The number of mutations from $a$ to $b$ or from $b$ to $a$<br>
$f_a$ : Total number of mutations in which $a$ takes part<br>
$f$ : Twice the total number of mutations

$f_{CG}$ :

$f_C$ :

$f$ :

   (d) Use the following definition and calculate $p_C$.

$p_a$ : The relative occurrence of $a$ in the observed sequences

$p_C$ :

3. **Relative mutability of PAM**

Relative mutability is calculated from frequencies of estimated mutilation and background probabilities.

$$m_a : \frac{1}{100p_a} \times \frac{f_a}{f}$$

$f_a$ : Total number of point mutations in which $a$ takes part
$f$ : Twice the total number of point mutations

$p_a$ : Relative occurrence of $a$ in the observed sequences

Assume that the frequencies are pre-calculated as follows.

- Frequencies of estimated mutations

  $f_A : 2, \quad f_G : 3, \quad f_C : 3, \quad f_T : 2$
  $f : 10$

- Background probabilities

  $p_A : 3/10, \quad p_G : 2/10, \quad p_C : 4/10, \quad p_T : 1/10$

(a) Calculate the probabilities of point mutations by $\frac{f_a}{f}$.

$\frac{f_A}{f}$ : $\qquad$ $\frac{f_G}{f}$ : $\qquad$ $\frac{f_C}{f}$ : $\qquad$ $\frac{f_T}{f}$ :

(b) Calculate $100p_a$.

$100p_A$ : $\qquad$ $100p_G$ : $\qquad$ $100p_C$ : $\qquad$ $100p_T$ :

(c) Calculate the relative mutability $m_a$.

$m_A$ : $\qquad$ $m_G$ : $\qquad$ $m_C$ : $\qquad$ $m_T$ :

4. **Mutation probabilities of PAM**

Mutation probabilities are calculated from relative mutability.

$$m_{ab} : m_a \times \frac{f_{ab}}{f_a}, \qquad m_{aa} : 1 - m_a$$

$$\begin{aligned} f_{ab} &: \text{Total number of point mutations in which } a \text{ takes part} \\ f_a &: \text{Twice the total number of point mutations} \\ m_a &: \text{Relative mutability of } a \end{aligned}$$

Assume that the frequencies are pre-calculated as follows.

- Frequencies of estimated mutations

  $f_{AC} : 8, \quad f_A : 32$

- Relative mutability

  $m_A : 0.004$

(a) Calculate $M_{AC}$.

(b) Calculate $M_{AA}$.

5. **Odds ratios of PAM**

Odds ratios are calculated from mutation probabilities and background probabilities.

$$O_{ab} = \frac{M_{ab}}{p_b} = m_a \times \frac{f_{ab}}{f_a} \times \frac{1}{f_b} = \frac{1}{100} \times \frac{f_{ab}}{f} \times \frac{1}{p_a p_b}$$

Assume that the frequencies are pre-calculated as follows.

$$f_{AC} : 16, \quad f : 400, \quad p_A : 0.2, \quad p_c : 0.4$$

(a) Calculate $O_{AC}$.

(b) Calculate $O_{CA}$.

6. **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 | C | G | T |
|---|---|---|---|---|
| A | | | | |
| C | | | | |
| G | | | | |
| T | | | | |

(b) Calculate $T$.

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

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

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

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