11 Exercise solutions – 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.

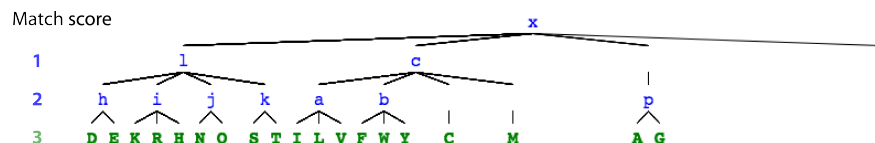   **Solution:** 2

   (b) Use the genetic code.

First position	Second position				Third position
	T	C	A	G	
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

   Solution: 9

   ```
           CGU --- CAU AUU UGU
           CGU GAU GAU UGU UGU
   ```
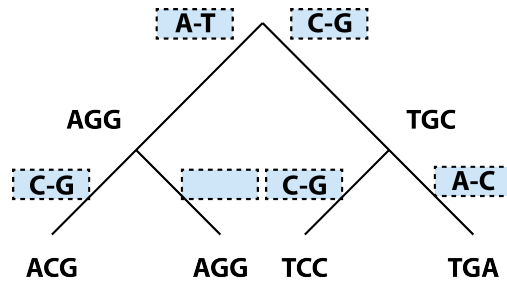
   (c) Use the AACH.

   Solution: $3 + 0 + 1 + 1 + 3 = 8$

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		1		1
C	1		3	
G		3		
T	1			

(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

f_a : The total number of mutations in which a takes part

f : Twice the total number of mutations

f_{CG} : 3

f_C : $1 + 3 = 4$

f : $2\,(1 + 1 + 3) = 10$

(d) Use the following definition and calculate p_C.

p_a : The relative occurrence of a in the observed sequences

p_C : $3/12$

3. **Relative mutability of PAM**

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

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

f_a : The total number of point mutations in which a takes part

f : Twice the total number of point mutations

p_a : The 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 $\dfrac{f_a}{f}$.

   $\dfrac{f_A}{f} :$ 2/10 $\dfrac{f_G}{f} :$ 3/10 $\dfrac{f_C}{f} :$ 3/10 $\dfrac{f_T}{f} :$ 2/10

   (b) Calculate $100 p_a$.

   $100 p_A :$ 30 $100 p_G :$ 20 $100 p_C :$ 40 $100 p_T :$ 10

   (c) Calculate the relative mutability m_a.

   $m_A :$ 2/300 $m_G :$ 3/200 $m_C :$ 3/400 $m_T :$ 2/100