Estimation of pseudo counts

The equation used to estimate frequencies in a weight matrix is

$$p_a = \frac{\alpha \cdot f_a + \beta \cdot g_a}{\alpha + \beta}$$

where α is the number of sequence in the multiple alignment (minus 1), β is the weight on prior (or weight on pseudo counts), f_a is the observed frequency for amino acid a and g_a is the pseudo frequency for amino acid a.

The pseudo frequency is estimated using the relation

$$g_a = \sum_b f_b \cdot q(a \mid b)$$

where f_b is the observed frequency for amino acid b, and q(a|b) is the Blosum substitution frequency for the amino acid a, conditional on the observation of amino acid b.

Once you have estimated the frequency p_a , the weight matrix values are calculated using the relation

$$W_a = 2 * \frac{\log(\frac{p_a}{q_a})}{\log 2}$$

where p_a is the frequencies of amino acid a at position i in the motif, and q_a is the background frequency of amino acid a (see last page).

The Blosum62 substitution matrix and a table of the 20 background frequencies are given on the last page.

Say, you have the following 6 sequences

EDRYK

EHYLK

OGHLP

EHLYR

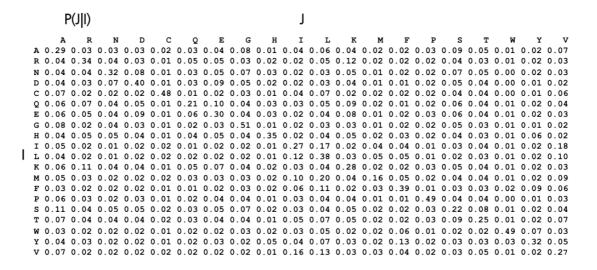
EHQEA

EHYLR

Estimate the observed frequencies (f_a), the pseudo frequencies (g_a), and the combined frequencies p_a at P1 for the 20 amino acids (fill out the table below). Use β =5 and no sequence weighting.

	f_a	g_a	p_a	W_a
A				
R				
N				
D				
C Q E				
Q				
Е				
G				
Н				
I				
L				
K				
M				
F				
P				
S T				
T				
W				
Y				
V				

Blosum substitution frequencies. Each row gives the frequency of q(J|I). That is the first row gives the frequencies of observing the 20 amino acids given you have observed the amino acid I.



Background frequencies (q_a)

```
A 0.07400
R 0.05200
N 0.04500
D 0.05400
C 0.02500
Q 0.03400
E 0.05400
G 0.07400
н 0.02600
I 0.06800
K 0.05800
M 0.02500
F 0.04700
P 0.03900
s 0.05700
T 0.05100
W 0.01300
Y 0.03200
```

V 0.07300