

Say, you have the following 6 sequences

EDRYK
EHYLK
QGHLP
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 $\beta=5$ and no sequence weighting.

Say you want to calculate the values for A. Then

$$f(A) = 0.0$$

$$g(A) = f(E) \cdot q(A|E) + f(Q) \cdot q(A|Q) = 5/6 \cdot 0.06 + 1/6 \cdot 0.06 = 0.06$$

$$p(A) = (5 \cdot 0.0 + 5 \cdot 0.06) / 10 = 0.03$$

$$w(A) = 2 \cdot \log(0.03 / 0.074) / \log(2) = -2.61$$

since the other 18 amino acids have frequency values of 0.

	f_a	g_a	p_a	w_a
A	0	0.06	0.03	-2.61
R	0	0.053	0.027	-1.93
N	0	0.04	0.02	-2.33
D	0	0.083	0.042	-0.75
C	0	0.01	0.005	-4.64
Q	0.167	0.085	0.126	3.78
E	0.833	0.267	0.550	6.70
G	0	0.04	0.02	-3.78
H	0	0.03	0.015	-1.59
I	0	0.022	0.011	-5.30
L	0	0.042	0.021	-4.50
K	0	0.082	0.041	-1.01
M	0	0.012	0.006	-4.19
F	0	0.018	0.009	-4.72
P	0	0.028	0.014	-2.92
S	0	0.06	0.03	-1.85
T	0	0.04	0.02	-2.70
W	0	0.01	0.005	-2.76
Y	0	0.02	0.01	-3.36
V	0	0.032	0.016	-4.41