

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

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

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
QGHLP
EHL YR
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.

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

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
A	0.29	0.03	0.03	0.03	0.02	0.03	0.04	0.08	0.01	0.04	0.06	0.04	0.02	0.02	0.03	0.09	0.05	0.01	0.02	0.07
R	0.04	0.34	0.04	0.03	0.01	0.05	0.05	0.03	0.02	0.02	0.05	0.12	0.02	0.02	0.02	0.04	0.03	0.01	0.02	0.03
N	0.04	0.04	0.32	0.08	0.01	0.03	0.05	0.07	0.03	0.02	0.03	0.05	0.01	0.02	0.02	0.07	0.05	0.00	0.02	0.03
D	0.04	0.03	0.07	0.40	0.01	0.03	0.09	0.05	0.02	0.02	0.03	0.04	0.01	0.01	0.02	0.05	0.04	0.00	0.01	0.02
C	0.07	0.02	0.02	0.02	0.48	0.01	0.02	0.03	0.01	0.04	0.07	0.02	0.02	0.02	0.02	0.04	0.04	0.00	0.01	0.06
Q	0.06	0.07	0.04	0.05	0.01	0.21	0.10	0.04	0.03	0.03	0.05	0.09	0.02	0.01	0.02	0.06	0.04	0.01	0.02	0.04
E	0.06	0.05	0.04	0.09	0.01	0.06	0.30	0.04	0.03	0.02	0.04	0.08	0.01	0.02	0.03	0.06	0.04	0.01	0.02	0.03
G	0.08	0.02	0.04	0.03	0.01	0.02	0.03	0.51	0.01	0.02	0.03	0.03	0.01	0.02	0.02	0.05	0.03	0.01	0.01	0.02
H	0.04	0.05	0.05	0.04	0.01	0.04	0.05	0.04	0.35	0.02	0.04	0.05	0.02	0.03	0.02	0.04	0.03	0.01	0.06	0.02
I	0.05	0.02	0.01	0.02	0.02	0.01	0.02	0.02	0.01	0.27	0.17	0.02	0.04	0.04	0.01	0.03	0.04	0.01	0.02	0.18
L	0.04	0.02	0.01	0.02	0.02	0.02	0.02	0.02	0.01	0.12	0.38	0.03	0.05	0.05	0.01	0.02	0.03	0.01	0.02	0.10
K	0.06	0.11	0.04	0.04	0.01	0.05	0.07	0.04	0.02	0.03	0.04	0.28	0.02	0.02	0.03	0.05	0.04	0.01	0.02	0.03
M	0.05	0.03	0.02	0.02	0.02	0.03	0.03	0.03	0.02	0.10	0.20	0.04	0.16	0.05	0.02	0.04	0.04	0.01	0.02	0.09
F	0.03	0.02	0.02	0.02	0.01	0.01	0.02	0.03	0.02	0.06	0.11	0.02	0.03	0.39	0.01	0.03	0.03	0.02	0.09	0.06
P	0.06	0.03	0.02	0.03	0.01	0.02	0.04	0.04	0.01	0.03	0.04	0.04	0.01	0.01	0.49	0.04	0.04	0.00	0.01	0.03
S	0.11	0.04	0.05	0.05	0.02	0.03	0.05	0.07	0.02	0.03	0.04	0.05	0.02	0.02	0.03	0.22	0.08	0.01	0.02	0.04
T	0.07	0.04	0.04	0.04	0.02	0.03	0.04	0.04	0.01	0.05	0.07	0.05	0.02	0.02	0.03	0.09	0.25	0.01	0.02	0.07
W	0.03	0.02	0.02	0.02	0.01	0.02	0.02	0.03	0.02	0.03	0.05	0.02	0.02	0.06	0.01	0.02	0.02	0.49	0.07	0.03
Y	0.04	0.03	0.02	0.02	0.01	0.02	0.03	0.02	0.05	0.04	0.07	0.03	0.02	0.13	0.02	0.03	0.03	0.03	0.32	0.05
V	0.07	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.01	0.16	0.13	0.03	0.03	0.04	0.02	0.03	0.05	0.01	0.02	0.27

Background frequencies

A 0.07400
R 0.05200
N 0.04500
D 0.05400
C 0.02500
Q 0.03400
E 0.05400
G 0.07400
H 0.02600
I 0.06800
L 0.09900
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