Estimation of pseudocounts

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 sequences in the multiple alignment (minus 1), β is the weight on prior (or weight on pseudocounts), 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 (read as “the substitution frequency of a given b”).

Once you have estimated the frequency p_a, the weight matrix values at a given position are calculated using the relation

$$w_a = 2 \times \log_2 \frac{p_a}{q_a}$$

where p_a is the estimated frequency of amino acid a, and q_a is the background frequency of amino acid a (see last page). Remember that you can always calculate the base 2 logarithm using this relation:

$$\log_2 x = \frac{\log x}{\log 2}$$

where “log” is any logarithm function, e.g. base 10 or ln (the natural logarithm).

This calculation is repeated at every position in the motif, so the full equation becomes:

$$w_{ia} = 2 \times \log_2 \frac{p_{ia}}{q_a}$$

where p_a is the frequency of amino acid a at position i in the motif.

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

Example

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 position 1** 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				
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

Table. The Blosum frequency substitution matrix. Each row gives the probabilities for substituting an amino acid to each of the 20 conventional amino acids. That is, the first row gives the probabilities P(aa|A) etc..

Examples: P(A|A) = 0.29, P(E|A) = 0.04, P(A|E) = 0.06, P(D|E) = 0.09, P(D|N) = 0.08

```
# Background frequencies
A 0.074
R 0.052
N 0.045
D 0.054
C 0.025
Q 0.034
E 0.054
G 0.074
H 0.026
I 0.068
L 0.099
K 0.058
M 0.025
F 0.047
P 0.039
S 0.057
T 0.051
W 0.013
Y 0.032
V 0.073
```