

PSSM's without pseudocounts

Local multiple alignment

k sequences
 w sites
 ungapped

W	E	I	R	D
W	E	I	R	D
W	E	I	R	E
W	E	I	Q	H

Amino acid counts

$c[x, i]$:

Number of copies of amino acid x in column i . For brevity, only residues that occur at least once are shown.

D	0	0	0	0	2
E	0	4	0	0	1
H	0	0	0	0	1
I	0	0	4	0	0
Q	0	0	0	1	0
R	0	0	0	3	0
W	4	0	0	0	0

$|\Sigma|$ rows, w columns

For this example, pretend $|\Sigma| = 7$.

Really, $|\Sigma| = 20$ for proteins, $|\Sigma| = 4$ for DNA

Frequency matrix

$$q[x, i] = \frac{c[x, i]}{k}$$

$q[x, i]$:

Frequency of amino acid x in column i .

D	0.0	0.0	0.0	0.0	0.5
E	0.0	1.0	0.0	0.0	0.3
H	0.0	0.0	0.0	0.0	0.3
I	0.0	0.0	1.0	0.0	0.0
Q	0.0	0.0	0.0	0.3	0.0
R	0.0	0.0	0.0	0.8	0.0
W	1.0	0.0	0.0	0.0	0.0

Likelihood ratio

$$P[x, i] = \frac{q[x, i]}{p[x]}$$

$P[x, i]$:

Probability the amino acid x in column i , given the background of x .

D	0.0	0.0	0.0	0.0	9.6
E	0.0	16.1	0.0	0.0	4.0
H	0.0	0.0	0.0	0.0	10.9
I	0.0	0.0	18.9	0.0	0.0
Q	0.0	0.0	0.0	6.1	0.0
R	0.0	0.0	0.0	14.7	0.0
W	71.4	0.0	0.0	0.0	0.0

Background frequency

D	0.052
E	0.062
H	0.023
I	0.053
Q	0.041
R	0.051
W	0.014

PSSM: Log-odds scoring matrix

$$S[x, i] = \log_2 P[x, i]$$

$S[x, i]$:

Log-odds score of observing amino acid x in column i .

D					3.3
E		4.0			2.0
H					3.4
I			4.2		
Q				2.6	
R				3.9	
W	6.2				

PSSM's with pseudocounts

Local multiple alignment

k sequences
 w sites
 ungapped

W	E	I	R	D
W	E	I	R	D
W	E	I	R	E
W	E	I	Q	H

Amino acid counts

$c[x, i] + b$:

Number of copies of amino acid x in column i . For brevity, only residues that occur at least once are shown.

D	1	1	1	1	3
E	1	5	1	1	2
H	1	1	1	1	2
I	1	1	5	1	1
Q	1	1	1	2	1
R	1	1	1	4	1
W	5	1	1	1	1

$|\Sigma|$ rows, w columns
 $b = 1$

For this example, pretend $|\Sigma| = 7$.
 Really, $|\Sigma| = 20$.

Frequency matrix

$$q[x, i] = \frac{c[x, i] + b}{k + |\Sigma|b}$$

$q[x, i]$:

Frequency of amino acid x in column i , corrected with pseudocount b .

D	0.09	0.09	0.09	0.09	0.27
E	0.09	0.45	0.09	0.09	0.18
H	0.09	0.09	0.09	0.09	0.18
I	0.09	0.09	0.45	0.09	0.09
Q	0.09	0.09	0.09	0.18	0.09
R	0.09	0.09	0.09	0.36	0.09
W	0.45	0.09	0.09	0.09	0.09

$k + |\Sigma|b = 11$

Likelihood ratio

$$P[x, i] = \frac{q[x, i]}{p[x]}$$

$P[x, i]$:

Probability the amino acid x in column i , given the background of x .

D	1.7	1.7	1.7	1.7	5.2
E	1.5	7.3	1.5	1.5	2.9
H	4.0	4.0	4.0	4.0	7.9
I	1.7	1.7	8.6	1.7	1.7
Q	2.2	2.2	2.2	4.4	2.2
R	1.8	1.8	1.8	7.1	1.8
W	32.5	6.5	6.5	6.5	6.5

Background frequency

D	0.052
E	0.062
H	0.023
I	0.053
Q	0.041
R	0.051
W	0.014

PSSM: Log-odds scoring matrix

$$S[x, i] = \log_2 P[x, i]$$

$S[x, i]$:

Log-odds score of observing amino acid x in column i .

D	0.8	0.8	0.8	0.8	2.4
E	0.6	2.9	0.6	0.6	1.6
H	2.0	2.0	2.0	2.0	3.0
I	0.8	0.8	3.1	0.8	0.8
Q	1.1	1.1	1.1	2.1	1.1
R	0.8	0.8	0.8	2.8	0.8
W	5.0	2.7	2.7	2.7	2.7

Scoring a new sequence

W	I	W	E	I	R	H	Total Score
5.0	0.8	2.7	0.6	0.8			9.8
	0.8	2.7	0.6	0.8	0.8		5.6
		5.0	2.9	3.1	2.8	3.0	16.8