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[a,j] :
The number of copies of amino acid *a* in column *j*. 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
***k*+|Σ|**	4	4	4	4	4

\|Σ\| rows, *w* columns

For this example, pretend \|Σ\| = 7 . Really, \|Σ\| = 20 for proteins,
\|Σ\| = 4 for DNA

Frequency matrix

$$q[a,j] = \frac{c[a,j]}{p[a]}$$

q[a.j] is the frequency of amino acid *a* in column *j*.

D	0.00	0.00	0.00	0.00	0.50
E	0.00	1.00	0.00	0.00	0.25
H	0.00	0.00	0.00	0.00	0.25
I	0.00	0.00	1.00	0.00	0.00
Q	0.00	0.00	0.00	0.25	0.00
R	0.00	0.00	0.00	0.75	0.00
W	1.00	0.00	0.00	0.00	0.00

Likelihood ratio:

$$P[a,j] = \frac{q[a,j]}{p[a]}$$

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

Background frequency, p_i

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

Log odds scoring matrix

$$S[a,j] = \log_2 \frac{q[a,j]}{p[a]}$$

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**
pseudocounts	**D**	**D**	**D**	**D**	**D**
	E	**E**	**E**	**E**	**E**
	H	**H**	**H**	**H**	**H**
	I	**I**	**I**	**I**	**I**
	Q	**Q**	**Q**	**Q**	**Q**
	R	**R**	**R**	**R**	**R**
	W	**W**	**W**	**W**	**W**

Amino acid counts

c[a,j] :
The number of copies of amino acid *a* in column *j*. 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
***k*+|Σ|**	11	11	11	11	11

\|Σ\| rows, *w* columns

For this example, pretend \|Σ\| = 7 . Really, \|Σ\| = 20 for proteins, \|Σ\| = 4 for DNA

Frequency matrix

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

q[a.j] is the frequency of amino acid *a* in column *j*, corrected with a pseudocount, *b*. In this example, we assume that *b=1*.

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

Likelihood ratio:

$$P[a,j] = \frac{q[a,j]}{p[a]}$$

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, p_i

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

Log odds scoring matrix

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