## PSSM's with pseudocounts

**Local multiple alignment**

*k* sequences, *w* positions, no gaps

| | | | | | |
|---|---|---|---|---|---|
| **W** | **E** | **I** | **R** | **D** |
| **W** | **E** | **I** | **R** | **D** |
| **W** | **E** | **I** | **R** | **E** |
| **W** | **E** | **I** | **Q** | **H** |

*k* rows, w columns

**Amino acid counts**

*c[i,j] :*
The number of copies of amino acid *i* 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

Pretend |Σ| = 7 . DON'T do that at home. Really |Σ| = 20 for proteins, |Σ| = 4 for DNA

**Frequency matrix**

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

*q[i.j]* is the frequency of amino acid *i* in column *j,* corrected with a pseudocount, *b* . We will 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** |
| **SUM** | 1.0 | 1.0 | 1.0 | 1.0 | 1.0 |

**Propensity matrix**

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

**Note**: this is a likelihood ratio

| | | | | | |
|---|---|---|---|---|---|
| **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**

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

This is a log likelihood ratio

| | | | | | |
|---|---|---|---|---|---|
| **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** |
|---|---|---|---|---|---|---|---|
| **9.8** | 5.0 | 0.8 | 2.7 | 0.6 | 0.8 | | |
| **5.6** | | 0.8 | 2.7 | 0.6 | 0.8 | 0.8 | |
| **16.8** | | | 5.0 | 2.9 | 3.1 | 2.8 | 3.0 |