Position Specific Scoring Matrices

PSSM's, profiles, weight matrices, templates…

Assume pattern has already been discovered.

Input: local MSA, *k x n* matrix

*A[i,j]: j*th symbol in *i*th sequence

Output: Scoring matrix, |Σ| x n

S[i,j]: score of symbol *i* at position *j*

Position Specific Scoring Matrices

PSSM's, profiles, weight matrices, templates…

Example:

```
WEIRD
WEIRD
WEIRE
WEIQH
```

See spreadsheets…

Position Specific Scoring Matrices

PSSM's, profiles, weight matrices, templates…

Given *A[w,k] (k* sequences*, w* positions),

the frequency of amino acid *i* at position *j* is

$$F[i, j] = \frac{n_{ij}}{k}$$

where n_{ij} is the number of instances of aa *i* at site *j*

The propensity of amino acid *i* at position *j is*

$$P[i, j] = \frac{F[i, j]}{f(i)}$$

where *f(i)* is the background frequency of *i.*

From this, we obtain, a position specific scoring matrix

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

Scoring a potential new instance of the pattern:

Given a sequence *t,* a window of length *w* starting at position *L* is scored as follows*:*

$$S[t,L] = \sum_{j=0}^{w-L} S[t[j+L], j]$$

Sequence *t:*

w

ATGTTGCACGGTGGCATTTCTGCGGCATTGGCGGAAACTA TGACATTCGTGACGAGCAAGATAAACTATGCTGCGTTTCACGTTTAACTCTTT

L

A PSSM can be considered is a log odds scoring matrix

Note that the score of a window of length *w* at position *L* in *t*, is a log likelihood ratio of the form

$$S[t,L] = \log_2 \frac{P[data \mid H_a]}{P[data \mid H_0]}$$

where the *data* is the subsequence at *L*, H_a is the alternate hypothesis that *t* contains the pattern and H_0 is the null hypothesis (no pattern)

$$\begin{aligned} S[t,L] &= \sum_{j=0}^{w-L} S[t[j+L], j] \\ &= \sum_{j=0}^{w-L} \log_2 P[t[j+L], j] \\ &= \sum_{j=0}^{w-L} \log_2 \frac{F[t[j+L], j]}{f(t[i])} \\ &= \log_2 \frac{\prod_{j=0}^{w-L} F[t[j+L], j]}{\prod_{j=0}^{w-L} f(t[j+L])} \\ &= \log_2 \frac{P[data \mid H_a]}{P[data \mid H_0]} \end{aligned}$$

Amino acid background frequencies

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

Pseudocounts

Example:

```
AAAAA
CCCCC
DDDDD
. . .
YYYYY
WEIRD
WEIRD
WEIRE
WEIQH
```

$$F[i, j] = \frac{n_i + b}{k + |\Sigma| b}$$

The pseudocount, ***b,*** avoids the problem of zero entries in the frequency matrix (and negative infinity in the log odds scoring matrix.)

Frequently, ***b = 1***, is chosen.

Also, see Durbin, 5.6