Blosum matrices

Morten Nielsen
DTU

Outline

- ## Alignment scoring matrices
  - What is a BLOSUM50 matrix and how is it different from a BLOSUM80 matrix?

- ## What are Blosum matrices good for?
  - Sequence alignment
  - Infer properties from one protein to another

BLOSUM = BLOck SUbstitution Matrices

- Focus on conserved domains, MSA's (multiple sequence alignment) are ungapped blocks.

- Compute pairwise amino acid alignment counts
  - Count amino acid replacement frequencies directly from columns in blocks
  - Sample bias:
    - Cluster sequences that are x% similar.
    - Do not count amino acid pairs within a cluster.
    - Do count amino acid pairs across clusters, treating clusters as an "average sequence".
    - Normalize by the number of sequences in the cluster.

- BLOSUM x matrices
  - Sequences that are x% similar were clustered during the construction of the matrix.

Log-odds scores

- BLOSUM is a log-likelihood matrix:
- Likelihood of observing j given you have i is
  - $P(j|i) = P_{ij}/P_i$
- The prior likelihood of observing j is
  - Q_j
- The log-likelihood score is
  - $S_{ij} = 2\log_2(P(j|i)/\log(Q_j)) = 2\log_2(P_{ij}/(Q_iQ_j))$

So what does this mean? An example

$N_{AA} = 14$

$N_{AD} = 5$

$N_{AV} = 5$

$N_{DA} = 5$

$N_{DD} = 8$

$N_{DV} = 2$

$N_{VA} = 5$

$N_{VD} = 2$

$N_{VV} = 2$

$P_{AA} = 14/48$

$P_{AD} = 5/48$

$P_{AV} = 5/48$

$P_{DA} = 5/48$

$P_{DD} = 8/48$

$P_{DV} = 2/48$

$P_{VA} = 5/48$

$P_{VD} = 2/48$

$P_{VV} = 2/48$

1: VVAD

2: AAAD

3: DVAD   MSA

4: DAAA

$Q_A = 8/16$

$Q_D = 5/16$

$Q_V = 3/16$

So what does this mean?

$P_{AA} = 0.29$

$P_{AD} = 0.10$

$P_{AV} = 0.10$

$P_{DA} = 0.10$

$P_{DD} = 0.17$

$P_{DV} = 0.04$

$P_{VA} = 0.10$

$P_{VD} = 0.04$

$P_{VV} = 0.04$

$Q_A Q_A = 0.25$

$Q_A Q_D = 0.16$

$Q_A Q_V = 0.09$

$Q_D Q_A = 0.16$

$Q_D Q_D = 0.10$

$Q_D Q_V = 0.06$

$Q_V Q_A = 0.09$

$Q_V Q_D = 0.06$

$Q_V Q_V = 0.03$

1: VVAD
2: AAAD
3: DVAD
4: DAAA

MSA

$Q_A = 0.50$

$Q_D = 0.31$

$Q_V = 0.19$

So what does this mean?

P_{AA} = 0.29	Q_AQ_A = 0.25	S_{AA} = 0.44
P_{AD} = 0.10	Q_AQ_D = 0.16	S_{AD} =-1.17
P_{AV} = 0.10	Q_AQ_V = 0.09	S_{AV} = 0.30
P_{DA} = 0.10	Q_DQ_A = 0.16	S_{DA} =-1.17
P_{DD} = 0.17	Q_DQ_D = 0.10	S_{DD} = 1.54
P_{DV} = 0.04	Q_DQ_V = 0.06	S_{DV} =-0.98
P_{VA} = 0.10	Q_VQ_A = 0.09	S_{VA} = 0.30
P_{VD} = 0.04	Q_VQ_D = 0.06	S_{VD} =-0.98
P_{VV} = 0.04	Q_VQ_V = 0.03	S_{VV} = 0.49

- BLOSUM is a log-likelihood matrix:

$$S_{ij} = 2\log_2(P_{ij}/(Q_iQ_j))$$

The Scoring matrix

	A	D	V
A	0.44	-1.17	0.30
D	-1.17	1.54	-0.98
V	0.30	-0.98	0.49

1: VVAD
2: AAAD
3: DVAD   MSA
4: DAAA

And what does the BLOSUMXX mean?

- Cluster sequence Blocks at XX% identity
- To statistics only across clusters

A)
AV
AP
AL
VL

B)
AV
GL
GL
GV

$\Big\}$ min XX % identify
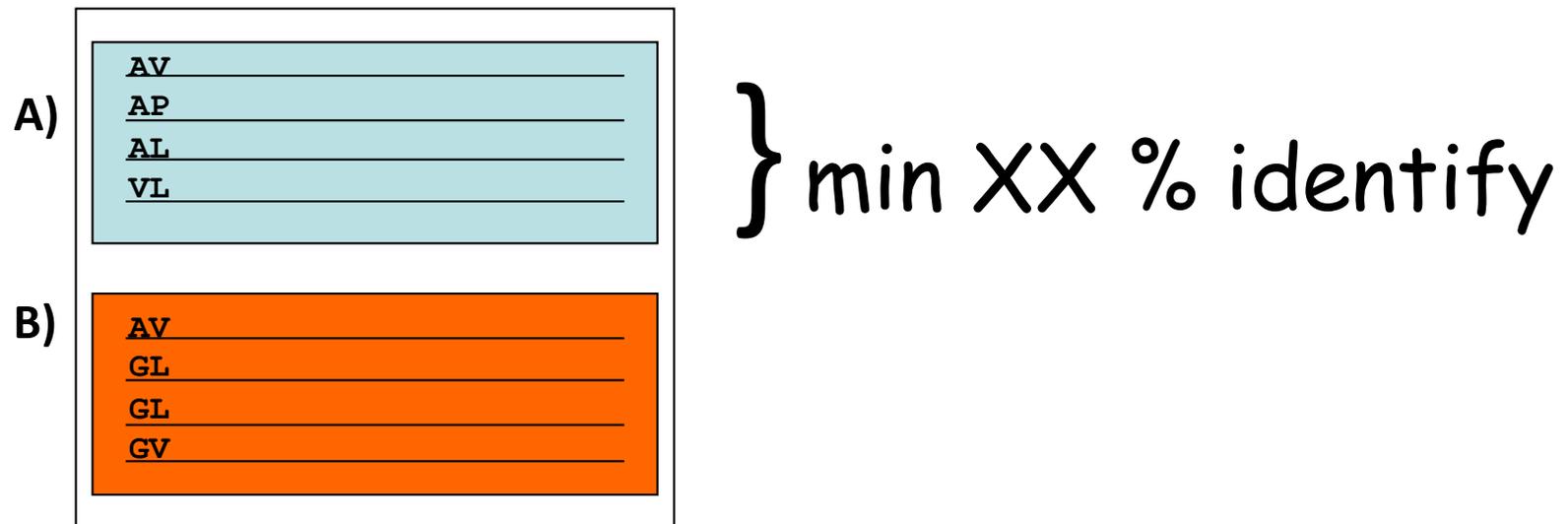
- Normalize statistics according to cluster size

And what does the BLOSUMXX mean?

A)

AV
AP
AL
VL

B)

AV
GL
GL
GV
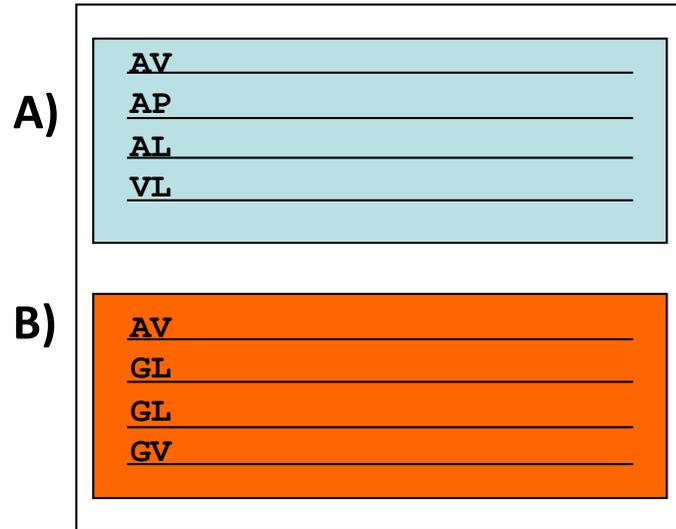
$$N_{AA} = \frac{3}{4} \cdot \frac{1}{4} = \frac{3}{16}$$

$$N_{AG} = \frac{3}{4} \cdot \frac{3}{4} = \frac{9}{16}$$

$$N_{VA} = \frac{1}{4} \cdot \frac{1}{4} = \frac{1}{16}$$

$$N_{VG} = \frac{1}{4} \cdot \frac{3}{4} = \frac{3}{16}$$

And what does the BLOSUMXX mean?

- ## High Blosum values mean <u>high similarity between clusters</u>
  - Conserved substitution allowed
- ## Low Blosum values mean <u>low similarity between clusters</u>
  - Less conserved substitutions allowed

BLOSUM80

```
    A   R   N   D   C   Q   E   G   H   I   L   K   M   F   P   S   T   W   Y   V
A   7  -3  -3  -3  -1  -2  -2   0  -3  -3  -3  -1  -2  -4  -1   2   0  -5  -4  -1
R  -3   9  -1  -3  -6   1  -1  -4   0  -5  -4   3  -3  -5  -3  -2  -2  -5  -4  -4
N  -3  -1   9   2  -5   0  -1  -1   1  -6  -6   0  -4  -6  -4   1   0  -7  -4  -5
D  -3  -3   2  10  -7  -1   2  -3  -2  -7  -7  -2  -6  -6  -3  -1  -2  -8  -6  -6
C  -1  -6  -5  -7  13  -5  -7  -6  -7  -2  -3  -6  -3  -4  -6  -2  -2  -5  -5  -2
Q  -2   1   0  -1  -5   9   3  -4   1  -5  -4   2  -1  -5  -3  -1  -1  -4  -3  -4
E  -2  -1  -1   2  -7   3   8  -4   0  -6  -6   1  -4  -6  -2  -1  -2  -6  -5  -4
G   0  -4  -1  -3  -6  -4  -4   9  -4  -7  -7  -3  -5  -6  -5  -1  -3  -6  -6  -6
H  -3   0   1  -2  -7   1   0  -4  12  -6  -5  -1  -4  -2  -4  -2  -3  -4   3  -5
I  -3  -5  -6  -7  -2  -5  -6  -7  -6   7   2  -5   2  -1  -5  -4  -2  -5  -3   4
L  -3  -4  -6  -7  -3  -4  -6  -7  -5   2   6  -4   3   0  -5  -4  -3  -4  -2   1
K  -1   3   0  -2  -6   2   1  -3  -1  -5  -4   8  -3  -5  -2  -1  -1  -6  -4  -4
M  -2  -3  -4  -6  -3  -1  -4  -5  -4   2   3  -3   9   0  -4  -3  -1  -3  -3   1
F  -4  -5  -6  -6  -4  -5  -6  -6  -2  -1   0  -5   0  10  -6  -4  -4   0   4  -2
P  -1  -3  -4  -3  -6  -3  -2  -5  -4  -5  -5  -2  -4  -6  12  -2  -3  -7  -6  -4
S   2  -2   1  -1  -2  -1  -1  -1  -2  -4  -4  -1  -3  -4  -2   7   2  -6  -3  -3
T   0  -2   0  -2  -2  -1  -2  -3  -3  -2  -3  -1  -1  -4  -3   2   8  -5  -3   0
W  -5  -5  -7  -8  -5  -4  -6  -6  -4  -5  -4  -6  -3   0  -7  -6  -5  16   3  -5
Y  -4  -4  -4  -6  -5  -3  -5  -6   3  -3  -2  -4  -3   4  -6  -3  -3   3  11  -3
V  -1  -4  -5  -6  -2  -4  -4  -6  -5   4   1  -4   1  -2  -4  -3   0  -5  -3   7
```

$\langle S_{ii} \rangle = 9.4$

$\langle S_{ij} \rangle = -2.9$

BLOSUM30

```
      A   R   N   D   C   Q   E   G   H   I   L   K   M   F   P   S   T   W   Y   V
A     4  -1   0   0  -3   1   0   0  -2   0  -1   0   1  -2  -1   1   1  -5  -4   1
R    -1   8  -2  -1  -2   3  -1  -2  -1  -3  -2   1   0  -1  -1  -1  -3   0   0  -1
N     0  -2   8   1  -1  -1  -1   0  -1   0  -2   0   0  -1  -3   0   1  -7  -4  -2
D     0  -1   1   9  -3  -1   1  -1  -2  -4  -1   0  -3  -5  -1   0  -1  -4  -1  -2
C    -3  -2  -1  -3  17  -2   1  -4  -5  -2   0  -3  -2  -3  -3  -2  -2  -2  -6  -2
Q     1   3  -1  -1  -2   8   2  -2   0  -2  -2   0  -1  -3   0  -1   0  -1  -1  -3
E     0  -1  -1   1   1   2   6  -2   0  -3  -1   2  -1  -4   1   0  -2  -1  -2  -3
G     0  -2   0  -1  -4  -2  -2   8  -3  -1  -2  -1  -2  -3  -1   0  -2   1  -3  -3
H    -2  -1  -1  -2  -5   0   0  -3  14  -2  -1  -2   2  -3   1  -1  -2  -5   0  -3
I     0  -3   0  -4  -2  -2  -3  -1  -2   6   2  -2   1   0  -3  -1   0  -3  -1   4
L    -1  -2  -2  -1   0  -2  -1  -2  -1   2   4  -2   2   2  -3  -2   0  -2   3   1
K     0   1   0   0  -3   0   2  -1  -2  -2  -2   4   2  -1   1   0  -1  -2  -1  -2
M     1   0   0  -3  -2  -1  -1  -2   2   1   2   2   6  -2  -4  -2   0  -3  -1   0
F    -2  -1  -1  -5  -3  -3  -4  -3  -3   0   2  -1  -2  10  -4  -1  -2   1   3   1
P    -1  -1  -3  -1  -3   0   1  -1   1  -3  -3   1  -4  -4  11  -1   0  -3  -2  -4
S     1  -1   0   0  -2  -1   0   0  -1  -1  -2   0  -2  -1  -1   4   2  -3  -2  -1
T     1  -3   1  -1  -2   0  -2  -2  -2   0   0  -1   0  -2   0   2   5  -5  -1   1
W    -5   0  -7  -4  -2  -1  -1   1  -5  -3  -2  -2  -3   1  -3  -3  -5  20   5  -3
Y    -4   0  -4  -1  -6  -1  -2  -3   0  -1   3  -1  -1   3  -2  -2  -1   5   9   1
V     1  -1  -2  -2  -2  -3  -3  -3  -3   4   1  -2   0   1  -4  -1   1  -3   1   5
```

$\langle S_{ii} \rangle = 8.3$

$\langle S_{ij} \rangle = -1.16$

Relationship between the two Blosum matrices

- What happens when $\alpha = 0$?
  - we only have one sequence, ILVKAIPHL

$$p_{1,A} = \frac{\alpha \cdot f_{1,A} + \beta \cdot g_{1,A}}{\alpha + \beta} = g_{1,A}$$

$$g_{1,A} = \sum_a f_a \cdot q(A \mid a) = q(A \mid I) = \frac{q_{IA}}{q_I}$$

$$W_{1,A} = \log(\frac{p_{1,A}}{q_{1,A}}) = \log(\frac{g_{1,A}}{q_{1,A}}) = \log(\frac{q_{IA}}{q_I \cdot q_A}) = Bl(A,I)$$

Relationship between the two Blosum matrices

```
        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 0.01 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  0.05 0.05 0.05 0.04 0.01 0.04 0.05 0.04 0.35 0.01 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 0.02 0.01 0.01 0.02 0.03 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
```

```
# Background
A 0.07400
R 0.05200
N 0.04500
D 0.05400
C 0.02500
Q 0.03400
E 0.05400
G 0.07400
H 0.02600
I 0.06800
L 0.09900
K 0.05800
M 0.02500
F 0.04700
P 0.03900
S 0.05700
T 0.05100
W 0.01300
Y 0.03200
V 0.07300
```

```
A    5
R   -2   7
N   -1  -1   7
D   -2  -2   2   8
C   -1  -4  -2  -4  13
Q   -1   1   0   0  -3   7
E   -1   0   0   2  -3   2   6
G    0  -3   0  -1  -3  -2  -3   8
H   -2   0   1  -1  -3   1   0  -2  10
I   -1  -4  -3  -4  -2  -3  -4  -4  -4   5
L   -2  -3  -4  -4  -2  -2  -3  -4  -3   2   5
K   -1   3   0  -1  -3   2   1  -2   0  -3  -3   6
M   -1  -2  -2  -4  -2   0  -2  -3  -1   2   3  -2   7
F   -3  -3  -4  -5  -2  -4  -3  -4  -1   0   1  -4   0   8
P   -1  -3  -2  -1  -4  -1  -1  -2  -2  -3  -4  -1  -3  -4  10
S    1  -1   1   0  -1   0  -1   0  -1  -3  -3   0  -2  -3  -1   5
T    0  -1   0  -1  -1  -1  -1  -2  -2  -1  -1  -1  -1  -2  -1   2   5
W   -3  -3  -4  -5  -5  -1  -3  -3  -3  -3  -2  -3  -1   1  -4  -4  -3  15
Y   -2  -1  -2  -3  -3  -1  -2  -3   2  -1  -1  -2   0   4  -3  -2  -2   2   8
V    0  -3  -3  -4  -1  -3  -3  -4  -4   4   1  -3   1  -1  -3  -2   0  -3  -1   5
     A    R    N    D    C    Q    E    G    H    I    L    K    M    F    P    S    T    W    Y    V
```

$$W(V,R) = 2 * \log_2(p(R|V)/q(R))$$
$$= 2 * \log_2(0.02/0.052)$$
$$= -2.76$$