Blosum matrices

What are they?

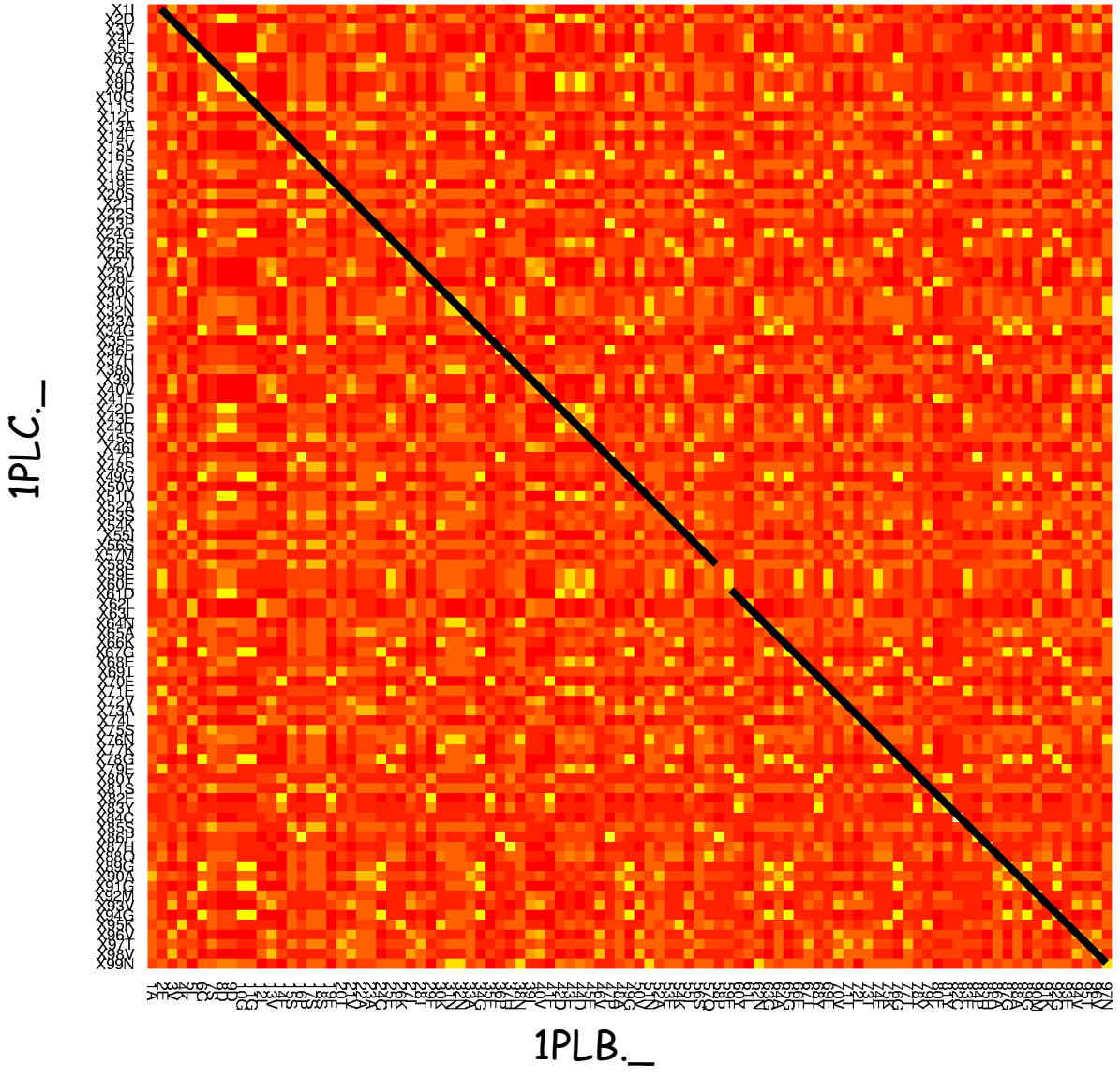
Morten Nielsen
BioSys, DTU

Outline

- Alignment scoring matrices
  - How are Blosum matrices constructed?
  - What is a BLOSUM50 matrix and how is it different from a BLOSUM80 matrix?
  - What is the difference between a Blosum scoring matrix and the Blosum frequency substitution matrix?

Sequence Alignment

CENTERFO
RBIOLOGI
CALSEQU
ENCEANA
LYSIS CBS



Where is the active site?

CENTERFO
RBIOLOGI
CALSEQU
ENCEANA
LYSIS CBS

Sequence alignment. Infer function (and functional residues) from one protein to another

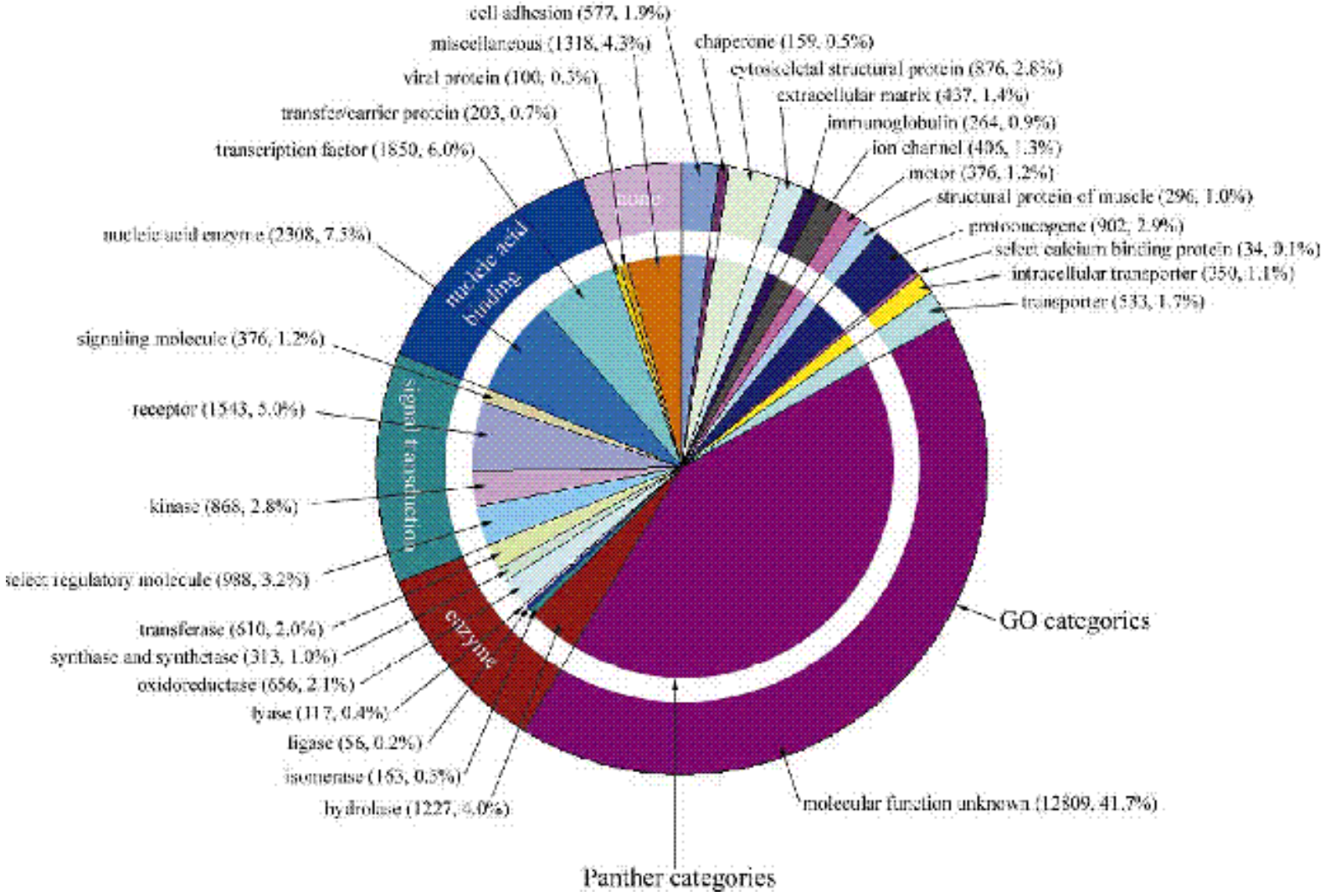
```
1K7C.A TVYLAGDSTMAKNGGGSGTNGWGEYLASYLSATVVNDAVAGRSARSYTREGRFENIADVVTAGDYVIVEFGHNDGGSLSTDN
              S                                  G                                  N
1WAB._ EVVFIGDSLVQLMHQCE---IWRELFS---PLHALNFGIGGDSTQHVLW--RLENGELEHIRPKIVVVWVGTNNHG------

1K7C.A GRTDCSGTGAEVCYSVYDGVNETILTFPAYLENAAKLFTAK--GAKVILSSQTPNNPWETGTFVNSPTRFVEYAEL-AAEVA
1WAB._ ----------------------HTAEQVTGGIKAIVQLVNERQPQARVVVLGLLPRGQ-HPNPLREKNRRVNELVRAALAGHP

1K7C.A GVEYVDHWSYVDSIYETLGNATVNSYFPIDHTHTSPAGAEVVAEAFLKAVVCTGTSL
                                        H
1WAB._ RAHFLDADPG---FVHSDG--TISHHDMYDYLHLSRLGYTPVCRALHSLLLRL---L
```

Homology modeling and the human genome

CENTERFO
RBIOLOGI
CALSEQU
ENCEANA
LYSIS **CBS**



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.
- BLOSUMXX matrices
  - Sequences that are xx% similar are clustered during the construction of the matrix.

Log-odds scores

- Log-odds scores are given by
  - Log(Observation/Expected)
- The log-odd score of matching amino acid j with amino acid i in an alignment is

$$\log(\frac{P_{ij}}{Q_i \cdot Q_j})$$

- where P_{ij} is the frequency of observing amino i aligned with j, and Q_i, Q_j are the frequencies of amino acids i and j in the data set.
- The log-odd score is (in half bit units)

$$S_{ij} = 2 \cdot \log_2(\frac{P_{ij}}{Q_i \cdot Q_j})$$

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
4: DAAA

MSA

Q_A	= 8/16
Q_D	= 5/16
Q_V	= 3/16

So what does this mean?

CENTERFO
RBIOLOGI
CALSEQU
ENCEANA
LYSIS CBS

$P_{AA} = 0.29$	$Q_A Q_A = 0.25$
$P_{AD} = 0.10$	$Q_A Q_D = 0.16$
$P_{AV} = 0.10$	$Q_A Q_V = 0.09$
$P_{DA} = 0.10$	$Q_D Q_A = 0.16$
$P_{DD} = 0.17$	$Q_D Q_D = 0.10$
$P_{DV} = 0.04$	$Q_D Q_V = 0.06$
$P_{VA} = 0.10$	$Q_V Q_A = 0.09$
$P_{VD} = 0.04$	$Q_V Q_D = 0.06$
$P_{VV} = 0.04$	$Q_V Q_V = 0.03$

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?

$P_{AA} = 0.29$	$Q_A Q_A = 0.25$	$S_{AA} = 0.44$
$P_{AD} = 0.10$	$Q_A Q_D = 0.16$	$S_{AD} = -1.17$
$P_{AV} = 0.10$	$Q_A Q_V = 0.09$	$S_{AV} = 0.30$
$P_{DA} = 0.10$	$Q_D Q_A = 0.16$	$S_{DA} = -1.17$
$P_{DD} = 0.17$	$Q_D Q_D = 0.10$	$S_{DD} = 1.54$
$P_{DV} = 0.04$	$Q_D Q_V = 0.06$	$S_{DV} = -0.98$
$P_{VA} = 0.10$	$Q_V Q_A = 0.09$	$S_{VA} = 0.30$
$P_{VD} = 0.04$	$Q_V Q_D = 0.06$	$S_{VD} = -0.98$
$P_{VV} = 0.04$	$Q_V Q_V = 0.03$	$S_{VV} = 0.49$

- BLOSUM is a log-likelihood matrix:

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

The Scoring matrix

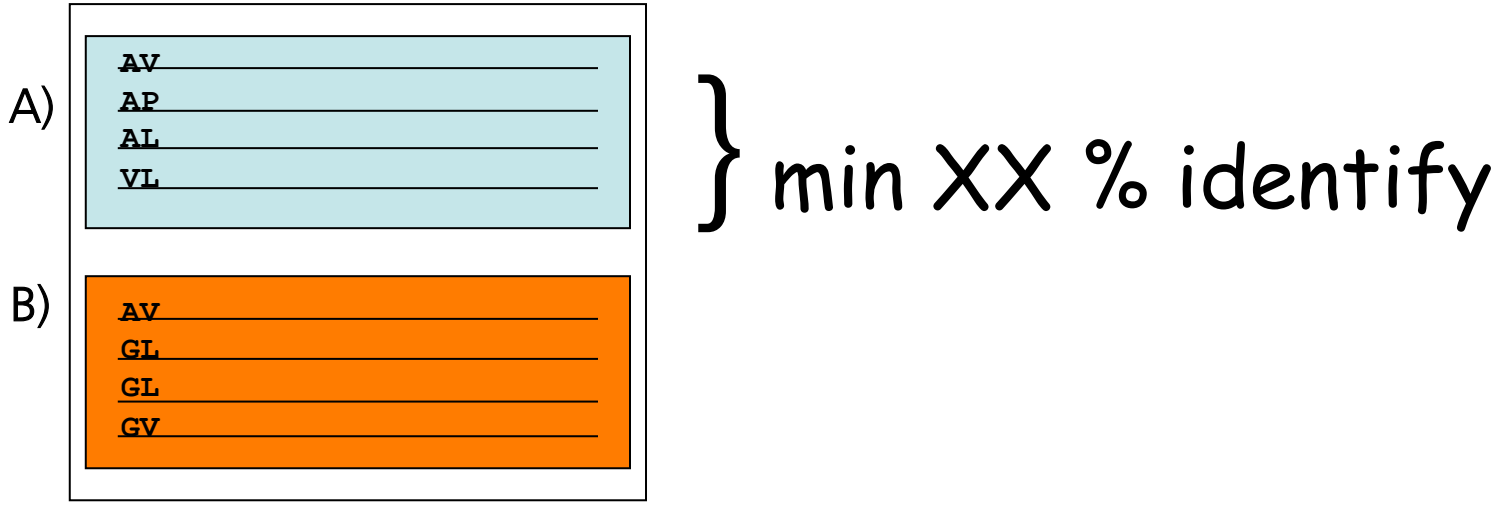
	A	D	V
A	0.44	-1.17	0.30
D	-1.17	1.54	-0.98
V	0.30	-0.98	0.49

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

MSA

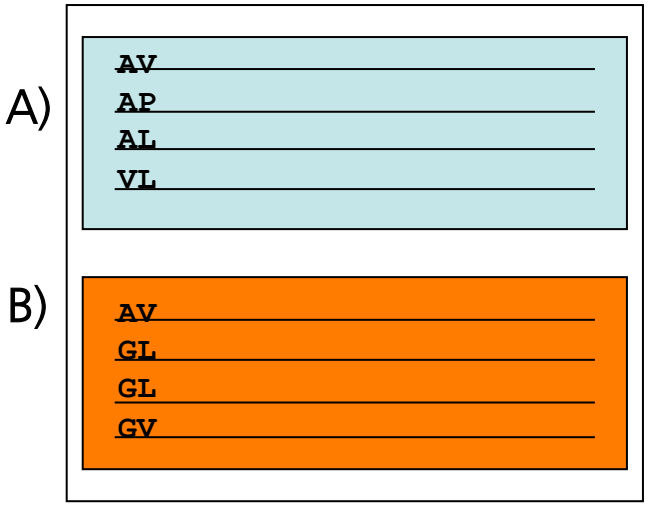
And what does the BLOSUMXX mean?

- Cluster sequence Blocks at XX% identity
- Do statistics only *<u>across</u>* clusters



- Normalize statistics according to cluster size

And what does the BLOSUMXX mean?



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

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

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

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

And what does the BLOSUMXX mean?

- High Blosum values mean *high similarity between* clusters
  - Conserved substitution dominate
- Low Blosum values mean *low similarity between* clusters
  - Less conserved substitutions dominate

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

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

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

The different Blosum matrices

- The BLOSUM alignment scoring matrix is a log-likelihood matrix

$$S_{ij} = 2 \cdot \log_2\left(\frac{P_{ij}}{Q_i \cdot Q_j}\right)$$

- The Blosum frequency substitution matrix, is a conditional probability matrix of matching amino acids j given you have amino acid i

$$P(j \mid i) = \frac{P_{ij}}{Q_i}$$

- and

$$S_{ij} = 2 \cdot \log_2\left(\frac{P_{ij}}{Q_i \cdot Q_j}\right) = 2 \cdot \log_2\left(\frac{P(j \mid i)}{Q_j}\right)$$

The way from frequencies to log-odds

CENTERFO
RBIOLOGI
CALSEQU
ENCEANA
LYSIS CBS

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

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

$$S_{AA} = 2 \cdot \log_2\left(\frac{P(A \mid A)}{Q_A}\right) = 2 \cdot \log_2\left(\frac{0.29}{0.074}\right) = 3.9$$

$$S_{AR} = 2 \cdot \log_2\left(\frac{P(R \mid A)}{Q_R}\right) = 2 \cdot \log_2\left(\frac{0.03}{0.052}\right) = -1.6$$