Biological Sequence Analysis

Sept 9
Pairwise Alignment

Topics for today

- Dotplot homework
- Pairwise alignment
- Linear & affine gaps
- Theoretical background for scoring systems
- BLOSUM matrices

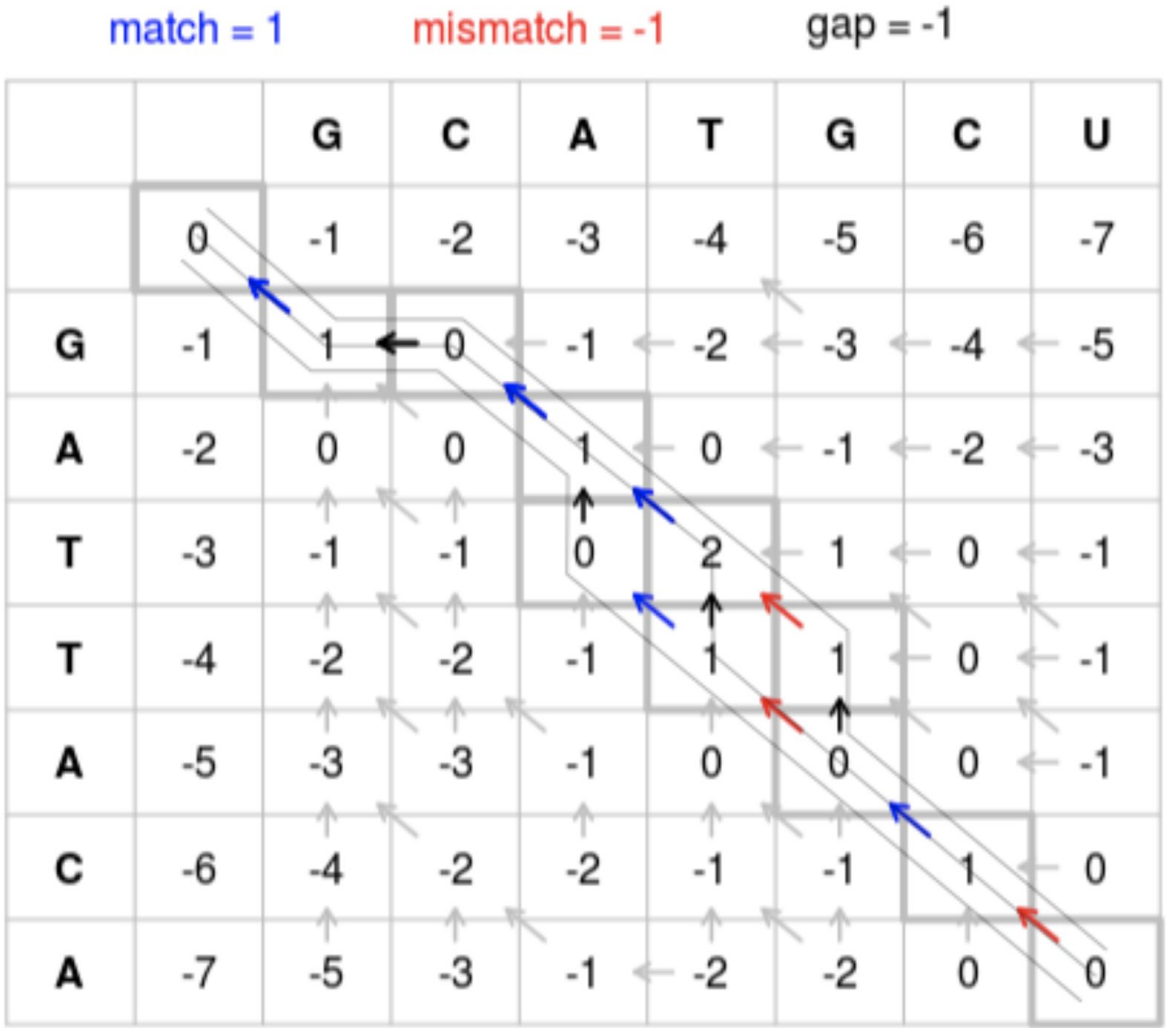
Thanks for the take-home messages!

- Short is good, focus on the most important part
- Please put questions in the discussion forum and not in the take-home message (simply because I might miss them)

template

		G	C	A	T	G	C	U
G								
A								
T								
T								
A								
C								
A								

backtracking



BLOSUM62

	Ala	Arg	Asn	Asp	Cys	Gln	Glu	Gly	His	Ile	Leu	Lys	Met	Phe	Pro	Ser	Thr	Trp	Tyr	Val
Ala	4																			
Arg	−1	5																		
Asn	−2	0	6																	
Asp	−2	−2	1	6																
Cys	0	−3	−3	−3	9															
Gln	−1	1	0	0	−3	5														
Glu	−1	0	0	2	−4	2	5													
Gly	0	−2	0	−1	−3	−2	−2	6												
His	−2	0	1	−1	−3	0	0	−2	8											
Ile	−1	−3	−3	−3	−1	−3	−3	−4	−3	4										
Leu	−1	−2	−3	−4	−1	−2	−3	−4	−3	2	4									
Lys	−1	2	0	−1	−3	1	1	−2	−1	−3	−2	5								
Met	−1	−1	−2	−3	−1	0	−2	−3	−2	1	2	−1	5							
Phe	−2	−3	−3	−3	−2	−3	−3	−3	−1	0	0	−3	0	6						
Pro	−1	−2	−2	−1	−3	−1	−1	−2	−2	−3	−3	−1	−2	−4	7					
Ser	1	−1	1	0	−1	0	0	0	−1	−2	−2	0	−1	−2	−1	4				
Thr	0	−1	0	−1	−1	−1	−1	−2	−2	−1	−1	−1	−1	−2	−1	1	5			
Trp	−3	−3	−4	−4	−2	−2	−3	−2	−2	−3	−2	−3	−1	1	−4	−3	−2	11		
Tyr	−2	−2	−2	−3	−2	−1	−2	−3	2	−1	−1	−2	−1	3	−3	−2	−2	2	7	
Val	0	−3	−3	−3	−1	−2	−2	−3	−3	3	1	−2	1	−1	−2	−2	0	−3	−1	4

```
# Entries for the BLOSUM80 matrix at a scale of ln(2)/2.0.
   A  R  N  D  C  Q  E  G  H  I  L  K  M  F  P  S  T  W  Y  V  B  J  Z  X  *
A  5 -2 -2 -2 -1 -1 -1  0 -2 -2 -2 -1 -1 -3 -1  1  0 -3 -2  0 -2 -2 -1 -1 -6
R -2  6 -1 -2 -4  1 -1 -3  0 -3 -3  2 -2 -4 -2 -1 -1 -4 -3 -3 -1 -3  0 -1 -6
N -2 -1  6  1 -3  0 -1 -1  0 -4 -4  0 -3 -4 -3  0  0 -4 -3 -4  5 -4  0 -1 -6
D -2 -2  1  6 -4 -1  1 -2 -2 -4 -5 -1 -4 -4 -2 -1 -1 -6 -4 -4  5 -5  1 -1 -6
C -1 -4 -3 -4  9 -4 -5 -4 -4 -2 -2 -4 -2 -3 -4 -2 -1 -3 -3 -1 -4 -2 -4 -1 -6
Q -1  1  0 -1 -4  6  2 -2  1 -3 -3  1  0 -4 -2  0 -1 -3 -2 -3  0 -3  4 -1 -6
E -1 -1 -1  1 -5  2  6 -3  0 -4 -4  1 -2 -4 -2  0 -1 -4 -3 -3  1 -4  5 -1 -6
G  0 -3 -1 -2 -4 -2 -3  6 -3 -5 -4 -2 -4 -4 -3 -1 -2 -4 -4 -4 -1 -5 -3 -1 -6
H -2  0  0 -2 -4  1  0 -3  8 -4 -3 -1 -2 -2 -3 -1 -2 -3  2 -4 -1 -4  0 -1 -6
I -2 -3 -4 -4 -2 -3 -4 -5 -4  5  1 -3  1 -1 -4 -3 -1 -3 -2  3 -4  3 -4 -1 -6
L -2 -3 -4 -5 -2 -3 -4 -4 -3  1  4 -3  2  0 -3 -3 -2 -2 -2  1 -4  3 -3 -1 -6
K -1  2  0 -1 -4  1  1 -2 -1 -3 -3  5 -2 -4 -1 -1 -1 -4 -3 -3 -1 -3  1 -1 -6
M -1 -2 -3 -4 -2  0 -2 -4 -2  1  2 -2  6  0 -3 -2 -1 -2 -2  1 -3  2 -1 -1 -6
F -3 -4 -4 -4 -3 -4 -4 -4 -2 -1  0 -4  0  6 -4 -3 -2  0  3 -1 -4  0 -4 -1 -6
P -1 -2 -3 -2 -4 -2 -2 -3 -3 -4 -3 -1 -3 -4  8 -1 -2 -5 -4 -3 -2 -4 -2 -1 -6
S  1 -1  0 -1 -2  0  0 -1 -1 -3 -3 -1 -2 -3 -1  5  1 -4 -2 -2  0 -3  0 -1 -6
T  0 -1  0 -1 -1 -1 -1 -2 -2 -1 -2 -1 -1 -2 -2  1  5 -4 -2  0 -1 -1 -1 -1 -6
W -3 -4 -4 -6 -3 -3 -4 -4 -3 -3 -2 -4 -2  0 -5 -4 -4 11  2 -3 -5 -3 -3 -1 -6
Y -2 -3 -3 -4 -3 -2 -3 -4  2 -2 -2 -3 -2  3 -4 -2 -2  2  7 -2 -3 -2 -3 -1 -6
V  0 -3 -4 -4 -1 -3 -3 -4 -4  3  1 -3  1 -1 -3 -2  0 -3 -2  4 -4  2 -3 -1 -6
B -2 -1  5  5 -4  0  1 -1 -1 -4 -4 -1 -3 -4 -2  0 -1 -5 -3 -4  5 -4  0 -1 -6
J -2 -3 -4 -5 -2 -3 -4 -5 -4  3  3 -3  2  0 -4 -3 -1 -3 -2  2 -4  3 -3 -1 -6
Z -1  0  0  1 -4  4  5 -3  0 -4 -3  1 -1 -4 -2  0 -1 -3 -3 -3  0 -3  5 -1 -6
X -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -6
* -6 -6 -6 -6 -6 -6 -6 -6 -6 -6 -6 -6 -6 -6 -6 -6 -6 -6 -6 -6 -6 -6 -6 -6  1
```

BLOSUM80: more closely related proteins

BLOSUM45: more distant seqs

```
# Entries for the BLOSUM45 matrix at a scale of ln(2)/3.0.
   A  R  N  D  C  Q  E  G  H  I  L  K  M  F  P  S  T  W  Y  V  B  J  Z  X  *
A  5 -2 -1 -2 -1 -1 -1  0 -2 -1 -1 -1 -1 -2 -1  1  0 -2 -2  0 -1 -1 -1 -1 -5
R -2  7  0 -1 -3  1  0 -2  0 -3 -2  3 -1 -2 -2 -1 -1 -2 -1 -2 -1 -3  1 -1 -5
N -1  0  6  2 -2  0  0  0  1 -2 -3  0 -2 -2 -2  1  0 -4 -2 -3  5 -3  0 -1 -5
D -2 -1  2  7 -3  0  2 -1  0 -4 -3  0 -3 -4 -1  0 -1 -4 -2 -3  6 -3  1 -1 -5
C -1 -3 -2 -3 12 -3 -3 -3 -3 -3 -2 -3 -2 -2 -4 -1 -1 -5 -3 -1 -2 -2 -3 -1 -5
Q -1  1  0  0 -3  6  2 -2  1 -2 -2  1  0 -4 -1  0 -1 -2 -1 -3  0 -2  4 -1 -5
E -1  0  0  2 -3  2  6 -2  0 -3 -2  1 -2 -3  0  0 -1 -3 -2 -3  1 -3  5 -1 -5
G  0 -2  0 -1 -3 -2 -2  7 -2 -4 -3 -2 -2 -3 -2  0 -2 -2 -3 -3 -1 -4 -2 -1 -5
H -2  0  1  0 -3  1  0 -2 10 -3 -2 -1  0 -2 -2 -1 -2 -3  2 -3  0 -2  0 -1 -5
I -1 -3 -2 -4 -3 -2 -3 -4 -3  5  2 -3  2  0 -2 -2 -1 -2  0  3 -3  4 -3 -1 -5
L -1 -2 -3 -3 -2 -2 -2 -3 -2  2  5 -3  2  1 -3 -3 -1 -2  0  1 -3  4 -2 -1 -5
K -1  3  0  0 -3  1  1 -2 -1 -3 -3  5 -1 -3 -1 -1 -1 -2 -1 -2  0 -3  1 -1 -5
M -1 -1 -2 -3 -2  0 -2 -2  0  2  2 -1  6  0 -2 -2 -1 -2  0  1 -2  2 -1 -1 -5
F -2 -2 -2 -4 -2 -4 -3 -3 -2  0  1 -3  0  8 -3 -2 -1  1  3  0 -3  1 -3 -1 -5
P -1 -2 -2 -1 -4 -1  0 -2 -2 -2 -3 -1 -2 -3  9 -1 -1 -3 -3 -3 -2 -3 -1 -1 -5
S  1 -1  1  0 -1  0  0  0 -1 -2 -3 -1 -2 -2 -1  4  2 -4 -2 -1  0 -2  0 -1 -5
T  0 -1  0 -1 -1 -1 -1 -2 -2 -1 -1 -1 -1 -1 -1  2  5 -3 -1  0  0 -1 -1 -1 -5
W -2 -2 -4 -4 -5 -2 -3 -2 -3 -2 -2 -2 -2  1 -3 -4 -3 15  3 -3 -4 -2 -2 -1 -5
Y -2 -1 -2 -2 -3 -1 -2 -3  2  0  0 -1  0  3 -3 -2 -1  3  8 -1 -2  0 -2 -1 -5
V  0 -2 -3 -3 -1 -3 -3 -3 -3  3  1 -2  1  0 -3 -1  0 -3 -1  5 -3  2 -3 -1 -5
B -1 -1  5  6 -2  0  1 -1  0 -3 -3  0 -2 -3 -2  0  0 -4 -2 -3  5 -3  1 -1 -5
J -1 -3 -3 -3 -2 -2 -3 -4 -2  4  4 -3  2  1 -3 -2 -1 -2  0  2 -3  4 -2 -1 -5
Z -1  1  0  1 -3  4  5 -2  0 -3 -2  1 -1 -3 -1  0 -1 -2 -2 -3  1 -2  5 -1 -5
X -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -5
* -5 -5 -5 -5 -5 -5 -5 -5 -5 -5 -5 -5 -5 -5 -5 -5 -5 -5 -5 -5 -5 -5 -5 -5  1
```

PAM vs BLOSUM

PAM	BLOSUM
Based on global alignments of closely related proteins.	Based on local alignments.
PAM1 is the matrix calculated from comparisons of sequences with no more than 1% divergence but corresponds to 99% sequence identity.	BLOSUM 62 is a matrix calculated from comparisons of sequences with a pairwise identity of no more than 62%.
Other PAM matrices are extrapolated from PAM1.	Based on observed alignments; they are not extrapolated from comparisons of closely related proteins.
Higher numbers in matrices naming scheme denote larger evolutionary distance.	Larger numbers in matrices naming scheme denote higher sequence similarity and therefore smaller evolutionary distance.[19]

Compare BLOSUM scores for different chemical groups

ILV - aliphatic

DERK - polar/charged

Do both within-group and between-group comparisons