Project 1

Alignment algorithm

Input

- Two sequences (from *fasta files), can be of different length, example:
  - GGVTTF
  - MEAIAKY
- Gap penalty (ex. g=-4)
- Substitution matrix (BLOSUM)

Steps

1. Fill scoring matrix
2. Backtracking to identify all possible alignments

Output

- All alignments

Input example

- 2 sequences (lengths m & n)
  - GGVTTF (m=6)
  - MGGETFA (n=7)
- Gap penalty g=-4

Score matrix S (compare p. 58)

- Create S (dimension = (m+1)x(n+1))
  - Rownames = 1st sequence
  - Colnames = 2nd sequence

		M	G	G	E	T	F	A
G								
G								
V								
T								
T								
F								

Input example

- 2 sequences (lengths m & n)
  - GGVTTF (m=6)
  - MGGETFA (n=7)
- Gap penalty g=-4

Score matrix S (compare p. 58)

- Create S (dimension = (m+1)x(n+1))
  - Rownames = 1st sequence
  - Colnames = 2nd sequence
- Fill first row/column with multiples of g

		M	G	G	E	T	F	A
	0	-4	-8	-12	-16	-20	-24	-28
G	-4							
G	-8							
V	-12							
T	-16							
T	-20							
F	-24							

Input example

- 2 sequences (lengths m & n)
  - GGVTTF (m=6)
  - MGGETFA (n=7)
- Gap penalty g=-4

Score matrix S (compare p. 58)

- Create S (dimension = (m+1)x(n+1))
  - Rownames = 1st sequence
  - Colnames = 2nd sequence
- Fill first row/column with multiples of g
- Fill each following line using the following formula

$$max\{S(i-1,j)+g, S(i,j-1)+g, S(i-1,j-1)+t(i,j)\}$$

		M	G	G	E	T	F	A
	0	-4	-8	-12	-16	-20	-24	-28
G	-4							
G	-8							
V	-12							
T	-16							
T	-20							
F	-24							

Input example

- 2 sequences (lengths m & n)
  - GGVTTF (m=6)
  - MGGETFA (n=7)
- Gap penalty g=-4

Score matrix S (compare p. 58)

- Create S (dimension = (m+1)x(n+1))
  - Rownames = 1st sequence
  - Colnames = 2nd sequence
- Fill first row/column with multiples of g
- Fill each following line using the following formula

$$max\{S(i-1,j)+g, S(i,j-1)+g, S(i-1,j-1)+t(i,j)\}$$

max{-4 + g, -4 +g, 0 +t('G','M')}

		M	G	G	E	T	F	A
	0	-4	-8	-12	-16	-20	-24	-28
G	-4							
G	-8							
V	-12							
T	-16							
T	-20							
F	-24							

Input example

- 2 sequences (lengths m & n)
  - GGVTTF (m=6)
  - MGGETFA (n=7)
- Gap penalty g=-4

Score matrix S (compare p. 58)

- Create S (dimension = (m+1)x(n+1))
  - Rownames = 1st sequence
  - Colnames = 2nd sequence
- Fill first row/column with multiples of g
- Fill each following line using the following formula

$$max\{S(i-1,j)+g, S(i,j-1)+g, S(i-1,j-1)+t(i,j)\}$$

max{-4 + g, -4 +g, 0 +t('G','M')} = max(-8, -8, -3) = -3

BLOSUM 62

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

		M	G	G	E	T	F	A
	0	-4	-8	-12	-16	-20	-24	-28
G	-4							
G	-8							
V	-12							
T	-16							
T	-20							
F	-24							

Input example

- 2 sequences (lengths m & n)
  - GGVTTF (m=6)
  - MGGETFA (n=7)
- Gap penalty g=-4

Score matrix S (compare p. 58)

- Create S (dimension = (m+1)x(n+1))
  - Rownames = 1st sequence
  - Colnames = 2nd sequence
- Fill first row/column with multiples of g
- Fill each following line using the following formula

$$max\{S(i-1,j)+g, S(i,j-1)+g, S(i-1,j-1)+t(i,j)\}$$

max{-4 + g, -4 +g, 0 +t('G','M')} = max(-8, -8, -3) = -3

BLOSUM 62

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

		M	G	G	E	T	F	A
	0	-4	-8	-12	-16	-20	-24	-28
G	-4	-3						
G	-8							
V	-12							
T	-16							
T	-20							
F	-24							

Input example

- 2 sequences (lengths m & n)
  - GGVTTF (m=6)
  - MGGETFA (n=7)
- Gap penalty g=-4

Score matrix S (compare p. 58)

- Create S (dimension = (m+1)x(n+1))
  - Rownames = 1st sequence
  - Colnames = 2nd sequence
- Fill first row/column with multiples of g
- Fill each following line using the following formula

$$max\{S(i-1,j)+g, S(i,j-1)+g, S(i-1,j-1)+t(i,j)\}$$

max{-8 + g, -3 +g, -4 +t('G','G')} = max(-12, -7, 2)=2

BLOSUM 62

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

		M	G	G	E	T	F	A
	0	-4	-8	-12	-16	-20	-24	-28
G	-4	-3	2					
G	-8							
V	-12							
T	-16							
T	-20							
F	-24							

Input example

- 2 sequences (lengths m & n)
  - GGVTTF (m=6)
  - MGGETFA (n=7)
- Gap penalty g=-4

Score matrix S (compare p. 58)

- Create S (dimension = (m+1)x(n+1))
  - Rownames = 1st sequence
  - Colnames = 2nd sequence
- Fill first row/column with multiples of g
- Fill each following line using the following formula

$$max\{S(i-1,j)+g, S(i,j-1)+g, S(i-1,j-1)+t(i,j)\}$$

BLOSUM 62

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

		M	G	G	E	T	F	A
	0	-4	-8	-12	-16	-20	-24	-28
G	-4	-3	2	-2	-6	-10	-14	-18
G	-8	-7	3	8	4	0	-4	-8
V	-12	-7	-1	4	6	4	0	-4
T	-16	-11	-5	0	3	11	7	3
T	-20	-15	-9	-4	-1	8	9	7
F	-24	-19	-13	-8	-5	4	14	10

Steps

1. Fill scoring matrix
2. Backtracking to identify all possible alignments

Algorithm

1. Start with the element in the last row, last column

		M	G	G	E	T	F	A
	0	-4	-8	-12	-16	-20	-24	-28
G	-4	-3	2	-2	-6	-10	-14	-18
G	-8	-7	3	8	4	0	-4	-8
V	-12	-7	-1	4	6	4	0	-4
T	-16	-11	-5	0	3	11	7	3
T	-20	-15	-9	-4	-1	8	9	7
F	-24	-19	-13	-8	-5	4	14	10

Steps

1. Fill scoring matrix
2. Backtracking to identify all possible alignments

Algorithm

1. Start with the element in the last row, last column
2. Identify the previous step that resulted in this value:
   - 14+g?
   - 7+g?
   - 9+t('F','A')?

		M	G	G	E	T	F	A
	0	-4	-8	-12	-16	-20	-24	-28
G	-4	-3	2	-2	-6	-10	-14	-18
G	-8	-7	3	8	4	0	-4	-8
V	-12	-7	-1	4	6	4	0	-4
T	-16	-11	-5	0	3	11	7	3
T	-20	-15	-9	-4	-1	8	9	7
F	-24	-19	-13	-8	-5	4	14	10

Steps

1. Fill scoring matrix
2. Backtracking to identify all possible alignments

Algorithm

1. Start with the element in the last row, last column
2. Identify the previous step that resulted in this value:
   - 14+g?
   - 7+g?
   - 9+t('F','A')?

		M	G	G	E	T	F	A
	0	-4	-8	-12	-16	-20	-24	-28
G	-4	-3	2	-2	-6	-10	-14	-18
G	-8	-7	3	8	4	0	-4	-8
V	-12	-7	-1	4	6	4	0	-4
T	-16	-11	-5	0	3	11	7	3
T	-20	-15	-9	-4	-1	8	9	7
F	-24	-19	-13	-8	-5	4	14	10

Steps

1. Fill scoring matrix
2. Backtracking to identify all possible alignments

Algorithm

1. Start with the element in the last row, last column
2. Identify the previous step that resulted in this value:
   - 14+g?
   - 7+g?
   - 9+t('F','A')
3. Repeat Step 2

		M	G	G	E	T	F	A
	0	-4	-8	-12	-16	-20	-24	-28
G	-4	-3	2	-2	-6	-10	-14	-18
G	-8	-7	3	8	4	0	-4	-8
V	-12	-7	-1	4	6	4	0	-4
T	-16	-11	-5	0	3	11	7	3
T	-20	-15	-9	-4	-1	8	9	7
F	-24	-19	-13	-8	-5	4	14	10

Steps

1. Fill scoring matrix
2. Backtracking to identify all possible alignments

Algorithm

1. Start with the element in the last row, last column
2. Identify the previous step that resulted in this value:
   - 14+g?
   - 7+g?
   - 9+t('F','A')
3. Repeat Step 2
4. Determine all possible alignments

		M	G	G	E	T	F	A
	0	-4	-8	-12	-16	-20	-24	-28
G	-4	-3	2	-2	-6	-10	-14	-18
G	-8	-7	3	8	4	0	-4	-8
V	-12	-7	-1	4	6	4	0	-4
T	-16	-11	-5	0	3	11	7	3
T	-20	-15	-9	-4	-1	8	9	7
F	-24	-19	-13	-8	-5	4	14	10

Steps

1. Fill scoring matrix
2. Backtracking to identify all possible alignments

Algorithm

1. Start with the element in the last row, last column
2. Identify the previous step that resulted in this value:
   - 14+g?
   - 7+g?
   - 9+t('F','A')
3. Repeat Step 2
4. Determine all possible alignments (here it's only 1)

```
M G G – E T F A
|   |   |   |   |   |   |   |
-   G G V T T F -
```

		M	G	G	E	T	F	A
	0	-4	-8	-12	-16	-20	-24	-28
G	-4	-3	2	-2	-6	-10	-14	-18
G	-8	-7	3	8	4	0	-4	-8
V	-12	-7	-1	4	6	4	0	-4
T	-16	-11	-5	0	3	11	7	3
T	-20	-15	-9	-4	-1	8	9	7
F	-24	-19	-13	-8	-5	4	14	10

Affine gap penalty

- Different costs for initial gap I and extended gap E:
  AB-BBD vs AB----BBD
- Consequences:
  - Initialization needs to take this into account
  - Need two additional matrices to keep the gap information, one for each sequence

$$max\{S(i-1,j)+g, S(i,j-1)+g, S(i-1,j-1)+t(i,j)\}$$

We cannot know whether
a gap was introduced before
from the values S(i-1,j) and S(I,j-1)

		M	G	G	E	T	F	A
G								
G								
V								
T								
T								
F								

Affine gap penalty

- Different costs for initial gap I and extended gap E:
  AB-BBD vs AB----BBD
- Consequences:
  - Initialization needs to take this into account
  - Need two additional matrices to keep the gap information, one for each sequence

- **First sequence**:
  - Previous value was a gap -> use V(i-1,j) and E
  - Previous value was not a gap -> use S(i-1,j) and I

    $V(i,j)=\max\{S(i-1,j)-I\ ,\ V(i-1,j)-E\}$

- **Second sequence:**
  - Previous value was a gap -> use W(i,j-1) and E
  - Previous value was not a gap -> use S(i,j-1) and I

    $W(i,j)=\max\{S(i,j-1)-I\ ,\ W(i,j-1)-E\}$

- **Scoring matrix S**:

  $S(i,j)=\max\{S(i-1,j-1)+t(x(i),y(j)),\ V(i,j),\ W(i,j)\}$

Example: I = 4 and E = 1

		M	G	G	E	T	F	A
	0	-4	-5	-6	-7	-8	-9	-10
G	-4							
G	-5							
V	-6							
T	-7							
T	-8							
F	-9							

Global vs local alignment

- **Global**
  - Negative values are possible in the scoring matrix
  - Start alignment from the **last value** in the matrix

- **Local**
  - Negative values are replaced by 0 (p.65)
  - Start alignment from the **maximal value** in the scoring matrix