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

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

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

	М	G	G	Ε	Т	F	Α
G							
G							
V							
Т							
Т							
F							

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

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

		М	G	G	E	Т	F	Α
	0	-4	-8	-12	-16	-20	-24	-28
G	-4							
G	-8							
V	-12							
Т	-16							
Т	-20							
F	-24							

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

- 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

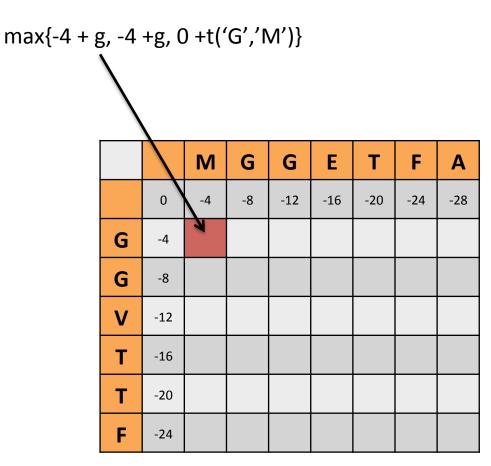
$$\max\{S(i-1,j)+g,S(i,j-1)+g,S(i-1,j-1)+t(i,j)\}$$

		М	G	G	Ε	Т	F	Α
	0	-4	-8	-12	-16	-20	-24	-28
O	-4							
G	-8							
٧	-12							
Т	-16							
Т	-20							
F	-24							

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

- 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

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

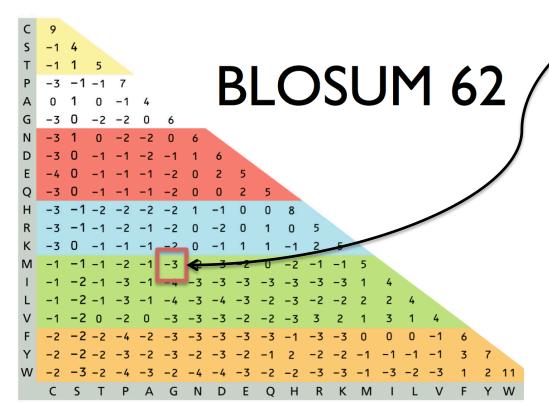


- 2 sequences (lengths m & n)
 - GGVTTF (m=6)
 - MGGETFA (n=7)
- Gap penalty g=-4

- 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

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

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



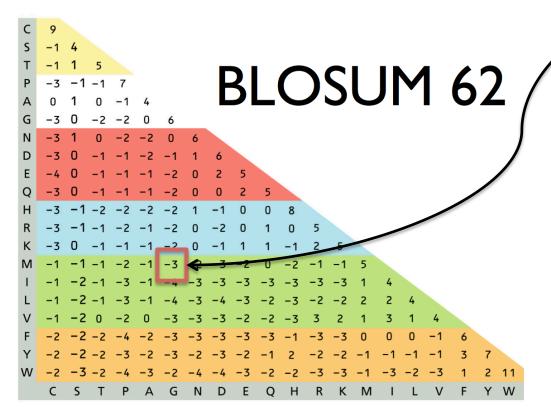
		Μ	G	G	Е	۲	F	Α
	0	-4	-8	-12	-16	-20	-24	-28
G	-4							
G	-8							
V	-12							
Т	-16							
Т	-20							
F	-24							

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

- 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

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

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



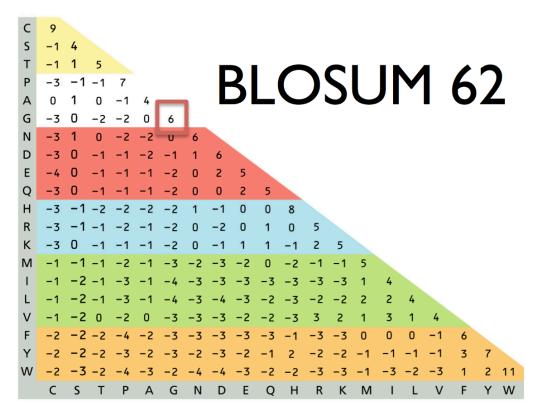
		М	G	G	E	Т	F	Α
	0	-4	-8	-12	-16	-20	-24	-28
G	-4	-3						
G	-8							
V	-12							
Т	-16							
Т	-20							
F	-24							

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

- 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

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

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



		Μ	G	G	Ε	Т	F	Α
	0	-4	-8	-12	-16	-20	-24	-28
G	-4	-3	2					
G	-8							
V	-12							
Т	-16							
Т	-20							
F	-24							

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

- 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
```

		М	G	G	Ε	Т	F	Α
	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
٧	-12	-7	-1	4	6	4	0	-4
Т	-16	-11	-5	0	3	11	7	3
Т	-20	-15	-9	-4	-1	8	9	7
F	-24	-19	-13	-8	-5	4	14	10

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

Algorithm

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

		М	G	G	E	Т	F	Α
	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
Т	-16	-11	-5	0	3	11	7	3
Т	-20	-15	-9	-4	-1	8	9	7
F	-24	-19	-13	-8	-5	4	14	10

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

- 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')?

		М	G	G	E	Т	F	Α
	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
Т	-16	-11	-5	0	3	11	7	3
Т	-20	-15	-9	-4	-1	8	9	7
F	-24	-19	-13	-8	-5	4	14	10

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

- 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	Е	T	H	Α
	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
Т	-16	-11	-5	0	3	11	7	3
Т	-20	-15	-9	-4	-1	8	9	7
F	-24	-19	-13	-8	-5	4	14	10

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

- 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	Ε	Т	F	Α
	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
Т	-16	-11	-5	0	3	11	7	3
Т	-20	-15	-9	-4	-1	8	9	7
F	-24	-19	-13	-8	-5	4	14	10

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

- 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

		Μ	G	G	Е	T	F	Α
	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
Т	-16	-11	-5	0	3	11	7	3
Т	-20	-15	-9	-4	-1	8	9	7
F	-24	-19	-13	-8	-5	4	14	10

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

- 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	Т	F.	Α
-	G	G	٧	Т	Т	F	_

		M	G	G	Ε	Т	F	Α
	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
Т	-16	-11	-5	0	3	11	7	3
Т	-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:

- 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	Т	F	Α
G							
G							
V							
Т							
Т					_	-	
F							

Affine gap penalty

Different costs for initial gap I and extended gap E:

- 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	Е	۲	F	A
	0	-4	-5	-6	-7	-8	-9	-10
G	-4							
G	-5							
V	-6							
Т	-7							
Т	-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