Global alignement Needleman-wunsch

INFOF-434 Macromolecular sequences analysis and biological databases

Problem

Input

- Two sequences of variable length, for example:
  - GGVTTF
  - MEAIAKY
- Gap penalty, for example g = -4
- Substitution matrix, for example BLOSUM 62

Steps

1. Compute scores for scoring matrix
2. Backtrack to identify all possible alignments

Output

k alignments (NW)

Example

GGVTTF (m=6)
MGGETFA (n=7)
Gap = -4

1. Create a matrix S of dimension (m+1)x(n+1) with the first sequence as rows and the second sequences as columns

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

Example

GGVTTF (m=6)
MGGETFA (n=7)
g = -4

1. Create a matrix S of dimension (m+1)x(n+1) with the first sequence as rows and the second sequences as columns
2. Fill the 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							

Example

GGVTTF (m=6)
MGGETFA (n=7)
g = -4

1. Create a matrix S of dimension (m+1)x(n+1) with the first sequence as rows and the second sequences as columns
2. Fill the first row/column with multiples of g
3. Fill other cells according to:

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

		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							

Example

GGVTTF (m=6)
MGGETFA (n=7)
g = -4

1. Create a matrix S of dimension (m+1)x(n+1) with the first sequence as rows and the second sequences as columns
2. Fill the first row/column with multiples of g
3. Fill other cells according to:

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

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							

Example

GGVTTF (m=6)
MGGETFA (n=7)
g = -4

1. Create a matrix S of dimension (m+1)x(n+1) with the first sequence as rows and the second sequences as columns
2. Fill the first row/column with multiples of g
3. Fill other cells according to:

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

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							

Example

GGVTTF (m=6)
MGGETFA (n=7)
g = -4

1. Create a matrix S of dimension (m+1)x(n+1) with the first sequence as rows and the second sequences as columns
2. Fill the first row/column with multiples of g
3. Fill other cells according to:

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

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							

Example

GGVTTF (m=6)
MGGETFA (n=7)
g = -4

1. Create a matrix S of dimension (m+1)x(n+1) with the first sequence as rows and the second sequences as columns
2. Fill the first row/column with multiples of g
3. Fill other cells according to:

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

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

Example : global

1. Begin with the element of 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

Example : global

1. Begin with the element of last row, last column
2. Identify the previous step resulting 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

Example : global

1. Begin with the element of last row, last column
2. Identify the previous step resulting 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

Example : global

1. Begin with the element of last row, last column
2. Identify the previous step resulting in this value:
   - 14 + g ?
   - 7 + g ?
   - 9 + t('F','A') ?
3. Repeat step 2 until you reach (0,0)

		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

Example : global

1. Begin with the element of last row, last column
2. Identify the previous step resulting in this value:
   - 14 + g ?
   - 7 + g ?
   - 9 + t('F','A') ?
3. Repeat step 2 until you reach (0,0)
4. Find all possible alignments

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

Z-score

- To know if a score is significative or not, you will compare it to a distribution of scores obtained with random sequences (for example the sequences you are working with, but scrambled)

- Mean and SD of distribution : $\mu_{al} = \frac{1}{N_{al}} \sum_{i=1}^{N_{al}} S_{al}(i)$ $\sigma_{al} = \sqrt{\frac{1}{N_{al}} \sum_{i=1}^{N_{al}} \left(S_{al}(i) - \mu_{al}\right)^2}$

- You compare the real score with the distribution according to the formula :

$$Z = \frac{S_{rée1} - \mu_{al}}{\sigma_{al}}$$