Escuela Profesional de
Ciencia de la Computación

Bioinformatics

Local Alignment

MSc. Vicente Machaca Arceda

Universidad Nacional de San Agustín de Arequipa

June 19, 2020

Table of Contents

vmachacaa@unsa.edu.pe

Table of Contents

vmachacaa@unsa.edu.pe

Introduction

Objectives

- Understand the importance of sequence alignment in Bioinformatics.

vmachacaa@unsa.edu.pe

Introduction

Objectives

- Understand the importance of sequence alignment in Bioinformatics.
- Understand and implement the Smith-Waterman algorithm.

vmachacaa@unsa.edu.pe

Table of Contents

vmachacaa@unsa.edu.pe

Global vs Local Alignment

Input sequences:
S_1 : ATGCGT
S_2 : ACGGCGT

Global Alignment **(3 of optimal score)**:

```
A_TGCGT    AT_GCGT    ATG_CGT
ACGGCGT    ACGGCGT    ACGGCGT
```

Local Alignment **(4 of optimal score)**:

```
3 GCGT 6
4 GCGT 7
```

vmachacaa@unsa.edu.pe

Global vs Local Alignment

input	HEAGAWGHEEAHGEGAE
string	PAWHEAEHE

Global alignment

```
HEAGAWGHEEAHGEGAE
--|-||-|-||--|-||
--P-AW-H-EA--E-HE
```

Local alignment

```
AWGHEEAH
||-|||||
AW-HEAEH
```

vmachacaa@unsa.edu.pe

Global vs Local Alignment

Global and local alignment are solve with dynamic programming.

Global alignment

Porposed in: **A general method applicable to the search for similarities in the amino acid sequence of two proteins** by Needleman in 1970 [1].

Local alignment

Porposed in: **Identification of common molecular subsequences** by Smith in 1981 [2].

vmachacaa@unsa.edu.pe

Local alignment

Definition

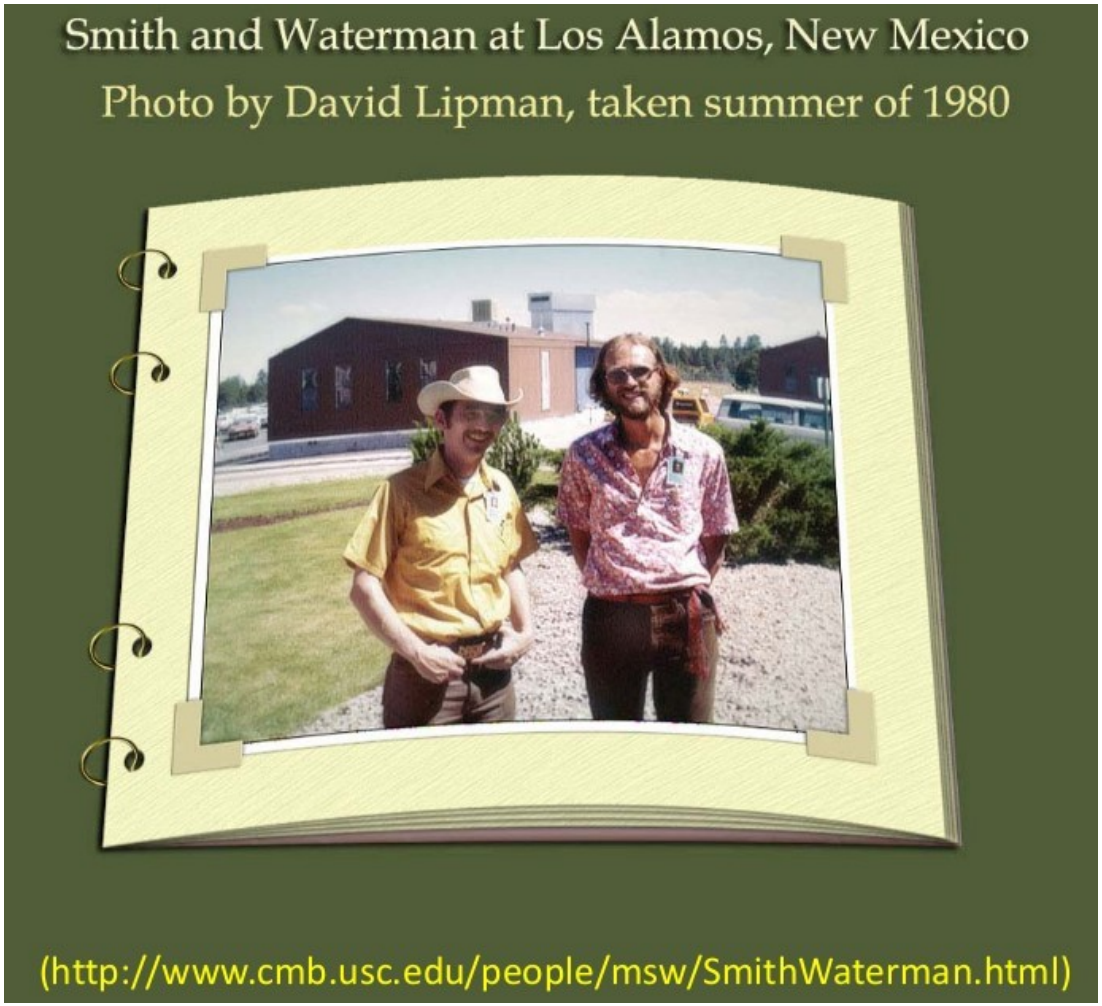
No negative scores are used (use zero instead). A similar tracing-back procedure is used in dynamic programming. However, the alignment path may begin and end internally along the main diagonal. It starts with the highest scoring position and proceeds diagonally up to the left until reaching a cell with a zero [3].

vmachacaa@unsa.edu.pe

Local alignment

Waterman

[Interview with Waterman]



vmachacaa@unsa.edu.pe

Local alignment

Definition

$$F(0,0) = 0$$

$$F(i, j) = \max \begin{cases} F(i-1, j-1) + s(x_i, y_j) \\ F(i-1, j) + d \\ F(i, j-1) + d \end{cases}$$

Global alignment

$$F(0,0) = 0$$

$$F(i, j) = \max \begin{cases} F(i-1, j-1) + s(x_i, y_j) \\ F(i-1, j) + d \\ F(i, j-1) + d \\ 0 \end{cases}$$

Local alignment

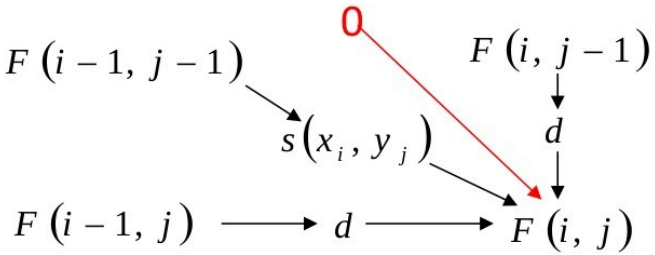
vmachacaa@unsa.edu.pe

Local alignment

Definition

$$F(0,0) = 0$$

$$F(i,j) = \max \begin{cases} F(i-1,\ j-1) + s(x_i,\ y_j) \\ F(i-1,\ j) + d \\ F(i,\ j-1) + d \\ 0 \end{cases}$$

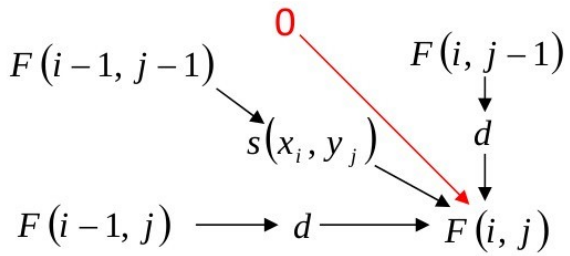


vmachacaa@unsa.edu.pe

Local alignment

Example

	A	C	G	T
A	2	-7	-5	-7
C	-7	2	-7	-5
G	-5	-7	2	-7
T	-7	-5	-7	2



Find the optimal local alignment of AAG and AGC.
Use a linear gap penalty of d = -5.

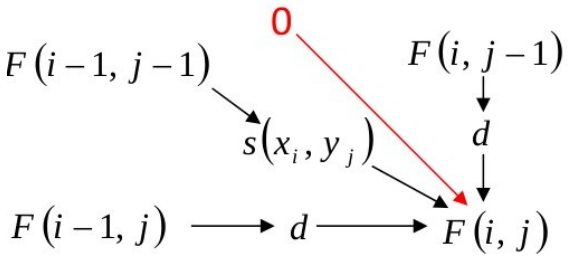
		A	A	G
A				
G				
C				

vmachacaa@unsa.edu.pe

Local alignment

Example

	A	C	G	T
A	2	-7	-5	-7
C	-7	2	-7	-5
G	-5	-7	2	-7
T	-7	-5	-7	2



Find the optimal local alignment of AAG and AGC.
Use a linear gap penalty of d = -5.

		A	A	G
	0	0	0	0
A	0	2	2	0
G	0	0	0	4
C	0	0	0	0

vmachacaa@unsa.edu.pe

Local alignment

Example

Trace back begins at the highest score in the matrix and continues until you reach 0.

A G
A G

		A	A	G
	0	0	0	0
A	0	2	2	0
G	0	0	0	4
C	0	0	0	0

vmachacaa@unsa.edu.pe

Local alignment

Example

And also the secondary best alignment.

A
A

		A	A	G
	0	0	0	0
A	0	2	2	0
G	0	0	0	4
C	0	0	0	0

vmachacaa@unsa.edu.pe

Local alignment

Global vs. Local

$$F(0,0) = 0$$

$$F(i, j) = \max \begin{cases} F(i-1, j-1) + s(x_i, y_j) \\ F(i-1, j) + d \\ F(i, j-1) + d \end{cases}$$

A	A	G	-	A	A	G	-
-	A	G	C	A	-	G	C

$$F(0,0) = 0$$

$$F(i, j) = \max \begin{cases} F(i-1, j-1) + s(x_i, y_j) \\ F(i-1, j) + d \\ F(i, j-1) + d \\ 0 \end{cases}$$

A	G
A	G

A
A

vmachacaa@unsa.edu.pe

Semi Global Alignment

Definition

In semi global similarity we seek a global alignment where we do not penalize for gaps at one or another end of the string.

The three scores are in general in the following relationship:
Global score $\leq$ **semi-global score** $\leq$ **local score**

vmachacaa@unsa.edu.pe

Exercises

Local align the following sequences:

- S_1 : ACCGTGA
- S_2 : GTGAATA

Use this scores: match = +1, mismatch = -1, gap = -1

Questions?

References I

S. B. Needleman and C. D. Wunsch, "A general method applicable to the search for similarities in the amino acid sequence of two proteins," *Journal of molecular biology*, vol. 48, no. 3, pp. 443–453, 1970.

T. F. Smith, M. S. Waterman *et al.*, "Identification of common molecular subsequences," *Journal of molecular biology*, vol. 147, no. 1, pp. 195–197, 1981.

J. Xiong, *Essential bioinformatics*. Cambridge University Press, 2006.