Escuela Profesional de
Ciencia de la Computación

ICC Fase 1

Bioinformatics

Sequence alignment - Dinamic programming

MSc. Vicente Machaca Arceda

Universidad Nacional de San Agustín de Arequipa

June 26, 2020

Table of Contents

1 Introduction

- Objectives

2 Sequence alignment

- Definition
- Example
- Maths

Table of Contents

1 Introduction
- Objectives

2 Sequence alignment
- Definition
- Example
- Maths

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 Needleman–Wunsch algorithm.

Table of Contents

1 Introduction
- Objectives

2 Sequence alignment
- Definition
- Example
- Maths

Sequence alignment

Definition

How can we determine the similarity between two sequences?

Sequence alignment

Definition

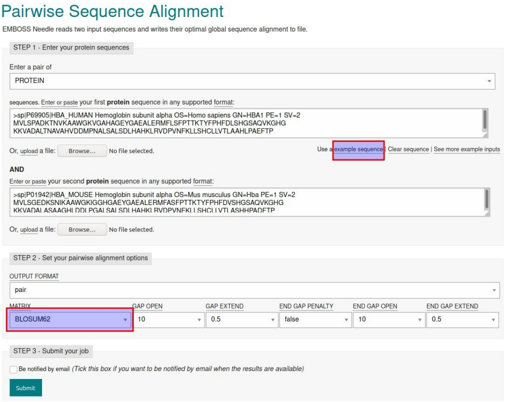
Sequence Alignment in Biology

The purpose of a sequence alignment is to line up all residues in the inputted sequence(s) for maximal level of similarity, in the sense of their functional or evolutionary relationship.

Sequence alignment

Pairwise sequence alignment

Visit EMBOSS, use the sample sequences and evaluate with BLOSUM62 matrix (no more that 62% of similarity).



Pairwise Sequence Alignment

EMBOSS Needle reads two input sequences and writes their optimal global sequence alignment to file.

STEP 1 - Enter your protein sequences

Enter a pair of

PROTEIN

sequences. Enter or paste your first **protein** sequence in any supported format:

```
>sp|P69905|HBA_HUMAN Hemoglobin subunit alpha OS=Homo sapiens GN=HBA1 PE=1 SV=2
MVLSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHFDLSHGSAQVKGHG
KKVADALTNAVAHVDDMPNALSALSDLHAHKLRVDPVNFKLLSHCLLVTLAAHLPAEFTP
```

Or, upload a file: Browse... No file selected.

Use a example sequence | Clear sequence | See more example inputs

AND

Enter or paste your second **protein** sequence in any supported format:

```
>sp|P01942|HBA_MOUSE Hemoglobin subunit alpha OS=Mus musculus GN=Hba PE=1 SV=2
MVLSGEDKSNIKAAWGKIGGHGAEYGAEALERMFASFPTTKTYFPHFDVSHGSAQVKGHG
KKVADALASAAGHLDDLPGALSALSDLHAHKLRVDPVNFKLLSHCLLVTLASHHPADFTP
```

Or, upload a file: Browse... No file selected.

STEP 2 - Set your pairwise alignment options

OUTPUT FORMAT

pair

MATRIX	GAP OPEN	GAP EXTEND	END GAP PENALTY	END GAP OPEN	END GAP EXTEND
BLOSUM62	10	0.5	false	10	0.5

STEP 3 - Submit your job

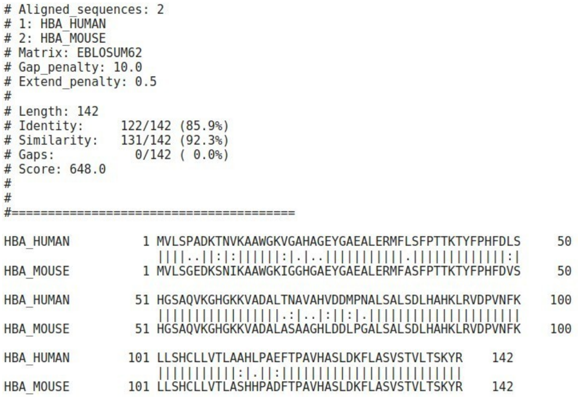
☐ Be notified by email *(Tick this box if you want to be notified by email when the results are available)*

Submit

Figure: Tool for pairwise sequence alignment.

Sequence alignment

Pairwise sequence alignment



```
# Aligned_sequences: 2
# 1: HBA_HUMAN
# 2: HBA_MOUSE
# Matrix: EBLOSUM62
# Gap_penalty: 10.0
# Extend_penalty: 0.5
#
# Length: 142
# Identity:     122/142 (85.9%)
# Similarity:   131/142 (92.3%)
# Gaps:           0/142 ( 0.0%)
# Score: 648.0
#
#
#=======================================

HBA_HUMAN          1 MVLSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHFDLS     50
                     ||||..||:|:||||||:|.|..|||||||||||.|||||||||||||:|
HBA_MOUSE          1 MVLSGEDKSNIKAAWGKIGGHGAEYGAEALERMFASFPTTKTYFPHFDVS     50

HBA_HUMAN         51 HGSAQVKGHGKKVADALTNAVAHVDDMPNALSALSDLHAHKLRVDPVNFK    100
                     |||||||||||||||||.:|..|:||:|.|||||||||||||||||||||
HBA_MOUSE         51 HGSAQVKGHGKKVADALASAAGHLDDLPGALSALSDLHAHKLRVDPVNFK    100

HBA_HUMAN        101 LLSHCLLVTLAAHLPAEFTPAVHASLDKFLASVSTVLTSKYR    142
                     |||||||||||:|.||:|||||||||||||||||||||||||
HBA_MOUSE        101 LLSHCLLVTLASHHPADFTPAVHASLDKFLASVSTVLTSKYR    142
```

Figure: Alignment: "|" stands for equality, ":" for similarity and "." for non-similarity.

Pairwise Sequence Alignment

in Maths

Input data:

- Two sequences S_1 and S_2

Parameters:

- A scoring function f for **substitutions** and **gaps**.

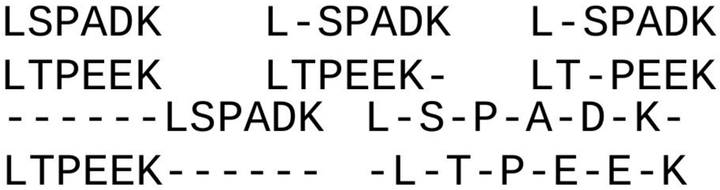
Output:

- The optimal alignment of S_1 and S_2, which has the maximal score.

Pairwise Sequence Alignment

in Maths

There are too many possible solution for sequence alignment.



```
LSPADK          L-SPADK        L-SPADK
LTPEEK          LTPEEK-        LT-PEEK
------LSPADK    L-S-P-A-D-K-
LTPEEK------    -L-T-P-E-E-K
```

$$\binom{2n}{n} = \frac{(2n)!}{(n!)^2}$$

Pairwise Sequence Alignment

in Maths

if $n = 300$

$$\binom{2n}{n} = \frac{(2n)!}{(n!)^2} = \frac{2 * 300}{(300!)^2} \approx 7 * 10^{88}$$

The visible universe is estimated to contain $10^{78} \sim 10^{80}$ atoms.

Pairwise Sequence Alignment

Dynamic Programming

Dynamic Programming

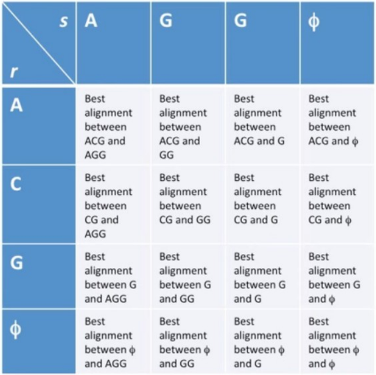
Dynamic Programming solves problems by combining the solutions to sub-problems.

- Break the problem into smaller sub-problems.
- Solve these sub-problems optimally recursively.
- Use these optimal solutions to construct an optimal solution for the original problem.

vmachacaa@unsa.edu.pe

Dynamic Programming

Example



s / r	A	G	G	ϕ
A	Best alignment between ACG and AGG	Best alignment between ACG and GG	Best alignment between ACG and G	Best alignment between ACG and ϕ
C	Best alignment between CG and AGG	Best alignment between CG and GG	Best alignment between CG and G	Best alignment between CG and ϕ
G	Best alignment between G and AGG	Best alignment between G and GG	Best alignment between G and G	Best alignment between G and ϕ
ϕ	Best alignment between ϕ and AGG	Best alignment between ϕ and GG	Best alignment between ϕ and G	Best alignment between ϕ and ϕ

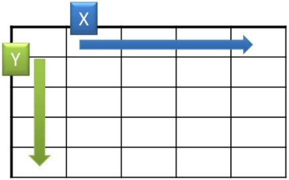
Figure: How we could divide the problem into sub-problems.

Dynamic Programming

Example

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

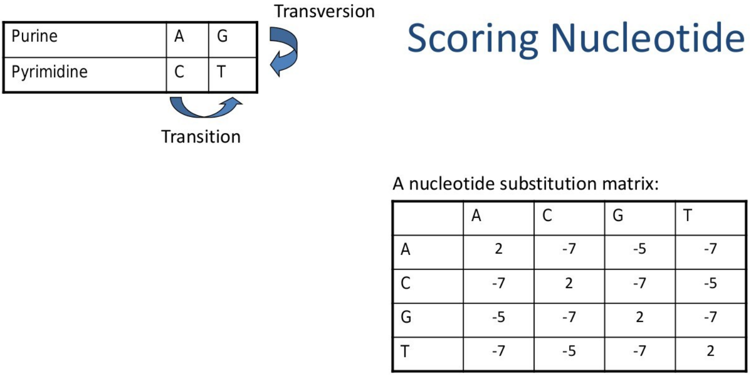
$$F(i,j) = \max \begin{cases} F(i-1,j-1) + s(x_i, y_j) & \textbf{x}_\textbf{i} \textbf{ aligned to } \textbf{y}_\textbf{j} \\ F(i-1,j) + d & \textbf{x}_\textbf{i} \textbf{ aligned to } \textbf{\textit{a gap}} \\ F(i,j-1) + d & \textbf{y}_\textbf{j} \textbf{ aligned to } \textbf{\textit{a gap}} \end{cases}$$



$$\begin{tikzcd}
F(i-1,j-1) \arrow[drr] & & & F(i,j-1) \arrow[d] \\
& & s(x_i,y_j) \arrow[dr] & d \arrow[d] \\
F(i-1,j) \arrow[r] & d \arrow[rr] & & F(i,j)
\end{tikzcd}$$

Dynamic Programming

Example



Scoring Nucleotide

Purine	A	G
Pyrimidine	C	T

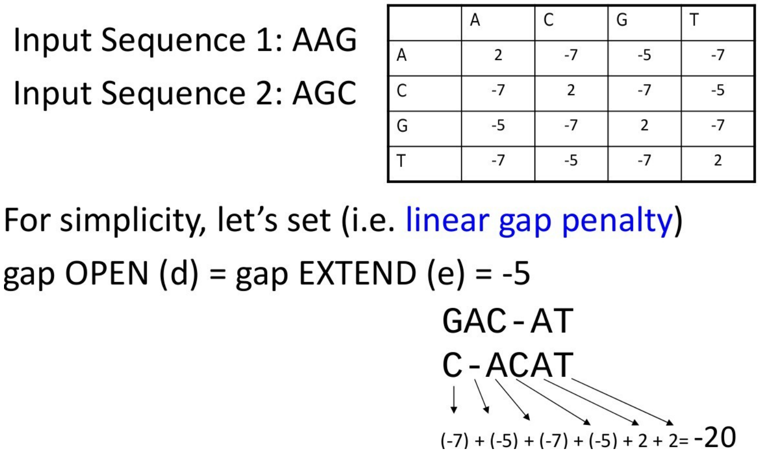
A nucleotide substitution matrix:

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

Figure: Example of $s(x_i, y_i)$.

Dynamic Programming

Example



Input Sequence 1: AAG

Input Sequence 2: AGC

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

For simplicity, let’s set (i.e. linear gap penalty)

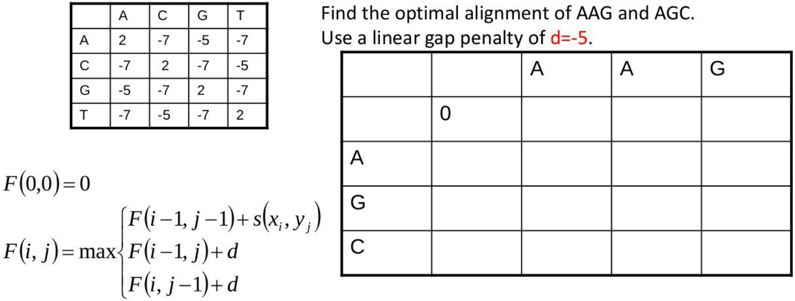
gap OPEN (d) = gap EXTEND (e) = -5

GAC-AT
C-ACAT

(-7) + (-5) + (-7) + (-5) + 2 + 2= -20

Dynamic Programming

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 alignment of AAG and AGC.
Use a linear gap penalty of d=-5.

		A	A	G
	0			
A				
G				
C				

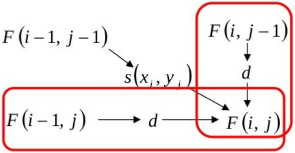
$$F(0,0) = 0$$

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

Dynamic Programming

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

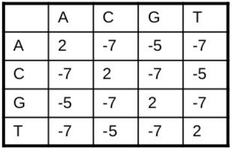


Find the optimal alignment of AAG and AGC.
Use a linear gap penalty of d=-5.

		A	A	G
	0 →	-5 →	-10 →	-15
A	↓ -5			
G	↓ -10			
C	↓ -15			

Dynamic Programming

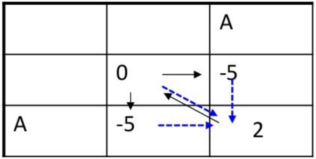
Example



	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 alignment of AAG and AGC.
Use a linear gap penalty of d=-5.

$$\begin{tikzcd}
F(i-1,\, j-1) \arrow[dr] & & F(i,\, j-1) \arrow[d] \\
& s(x_i, y_j) \arrow[dr] & d \arrow[d] \\
F(i-1,\, j) \arrow[r] & d \arrow[r] & F(i,\, j)
\end{tikzcd}$$



		A
	0	-5
A	-5	2

$$-5 + (-5) = -10$$

$$\boxed{0 + 2 = 2}$$

$$-5 + (-5) = -10$$

vmachacaa@unsa.edu.pe

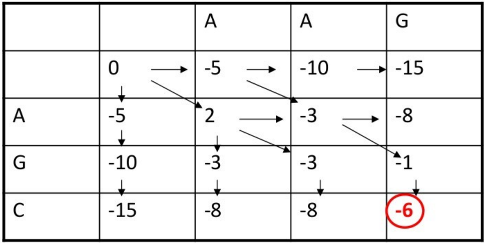
Dynamic Programming

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

$$\begin{tikzcd}
F(i-1,\ j-1) \arrow[dr] & & F(i,\ j-1) \arrow[d] \\
 & s(x_i, y_j) \arrow[dr] & d \arrow[d] \\
F(i-1,\ j) \arrow[r] & d \arrow[r] & F(i,\ j)
\end{tikzcd}$$

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



		A	A	G
	0	-5	-10	-15
A	-5	2	-3	-8
G	-10	-3	-3	-1
C	-15	-8	-8	-6

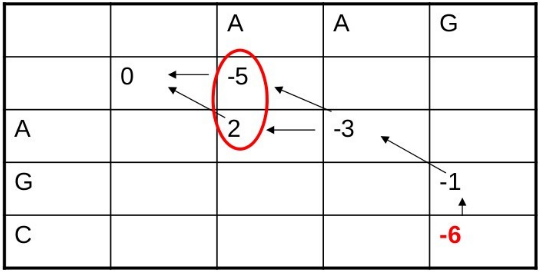
Dynamic Programming

Example

Trace back to the upper left. Each arrow introduces one symbol at the end of each aligned sequence.

```
A A G -
- A G C

A A G -
A - G C
```



		A	A	G
	0	-5		
A		2	-3	
G				-1
C				**-6**

Questions?

vmachacaa@unsa.edu.pe