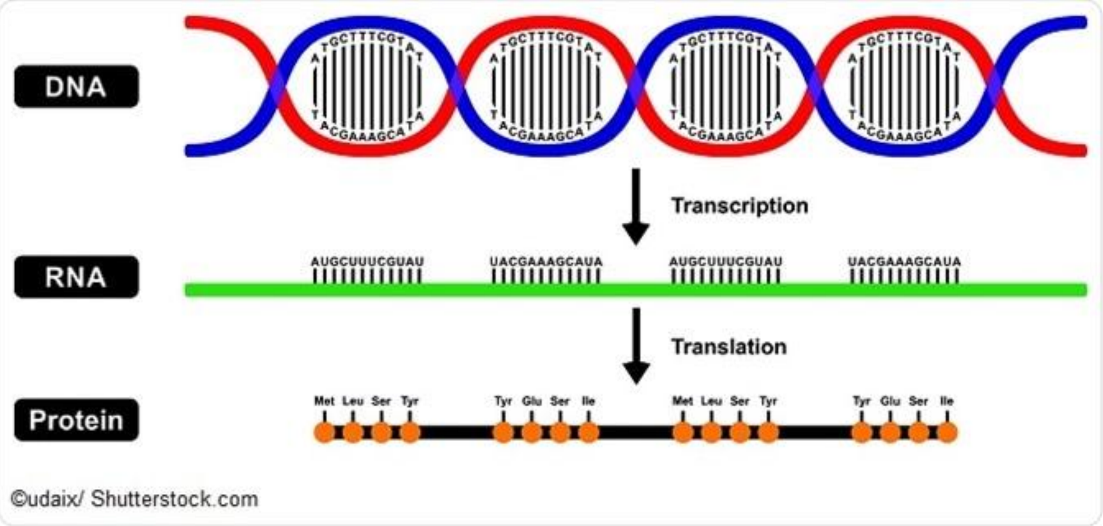
CIIC 4025 Analysis and Design of Algorithms

WILFREDO LUGO, PHD

Biological Sequence Alignments

- What are biological Sequences?
  - DNA/RNA
    - Nucleotide Base Sequences (e.g. Human Genome)
  - Proteins
    - Amino Acid Sequences

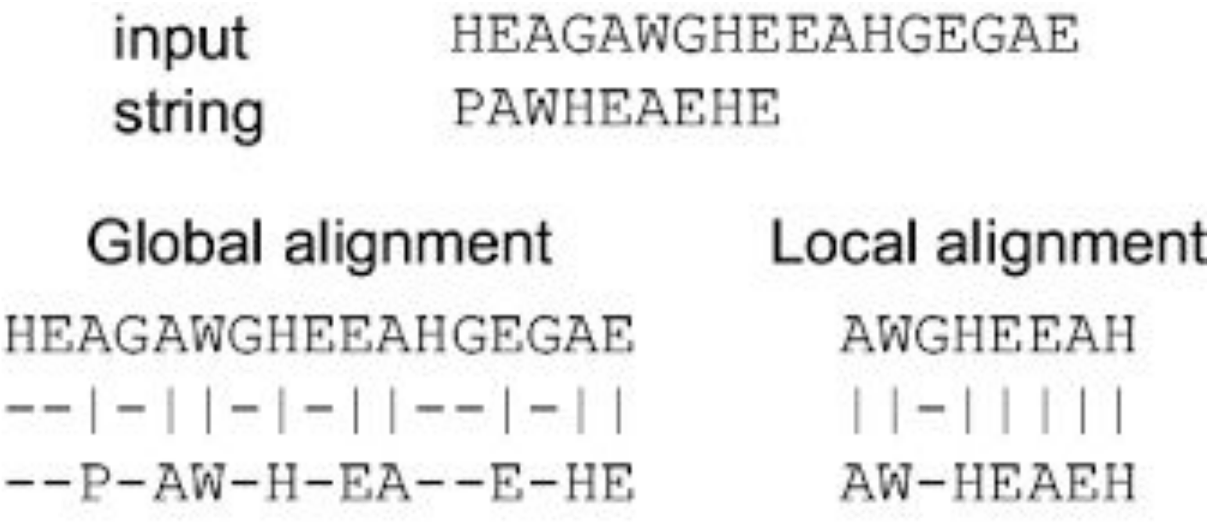


©udaix/ Shutterstock.com

Image Source: https://www.news-medical.net/life-sciences/Amino-Acids-and-Protein-Sequences.aspx

Biological Sequence Alignments

- Why do we need to align the sequences?
  - Get more information about
    - Functional Relationships
    - Structural Relationships
    - Evolutionary Relationships
- Types of Alignments
  - Global Alignments
    - Tries to align an entire sequence
    - Align all letters from query and target
    - Suitable for closely related sequences
  - Local Alignments
    - Align regions having highest similarities
    - Align substrings of target with substring of queries
    - Suitable for more divergent sequences.



```
input     HEAGAWGHEEAHGEGAE
string    PAWHEAEHE

Global alignment            Local alignment

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

Biological Sequence Alignments

- Dynamic Programing
  - Needleman-Wunsch – Most common Global Alignment Algorithm
  - Smith-Waterman – Most common Local Alignment Algorithm
- Needleman-Wunsch

$$F_{i,j} = max\begin{cases} F_{i-i,j-1} + S(A_i, B_j) \\ F_{i,j-1} + d \\ F_{i-1,j} + d \end{cases}$$

Scoring Matrix

Gap Penalty

$$\left.\begin{array}{l} F_{0,j} = d * j \\ F_{i,0} = d * \mathrm{i} \end{array}\right\}$$

Initialization Step

Biological Sequence Alignments

Needleman-Wunsch

- $S_1 = ATGCT$
- $S_2 = AGCT$
- Scoring Matrix
  - $S(A_i, B_j) = 1$, when $A_i = B_j$
  - $S(A_i, B_j) = -1$, when $A_i \neq B_j$
  - $d = -2$

$$F_{i,j} = max\begin{cases} F_{i-i,j-1} + S(A_i, B_j) \\ F_{i,j-1} + d \\ F_{i-1,j} + d \end{cases}$$

Scoring Matrix

Gap Penalty

$$\left.\begin{matrix} F_{0,j} = d * j \\ F_{i,0} = d * \text{i} \end{matrix}\right\}$$

Initialization Step

		A	T	G	C	T
	0	-2	-4	-6	-8	-10
A	-2					
G	-4					
C	-6					
T	-8					

$n = |S_2| + 1$

$m = |S_1| + 1$

Biological Sequence Alignments

$$F_{i,j} = max\begin{cases} F_{i-i,j-1} + S(A_i, B_j) \xrightarrow{} \text{Scoring Matrix} \\ F_{i,j-1} + d \xrightarrow{} \text{Gap Penalty} \\ F_{i-1,j} + d \xrightarrow{} \text{Gap Penalty} \end{cases}$$

$$\begin{rcases} F_{0,j} = d * j \\ F_{i,0} = d * \text{i} \end{rcases} \text{Initialization Step}$$

		A	T	G	C	T
	0	-2	-4	-6	-8	-10
A	-2	1	-1			
G	-4					
C	-6					
T	-8					

$$= max\begin{cases} 0 + 1 \\ -2 + -2 \\ -2 + -2 \end{cases}$$

$$= max\begin{cases} -2 + -1 \\ 1 + -2 \\ -4 + -2 \end{cases}$$

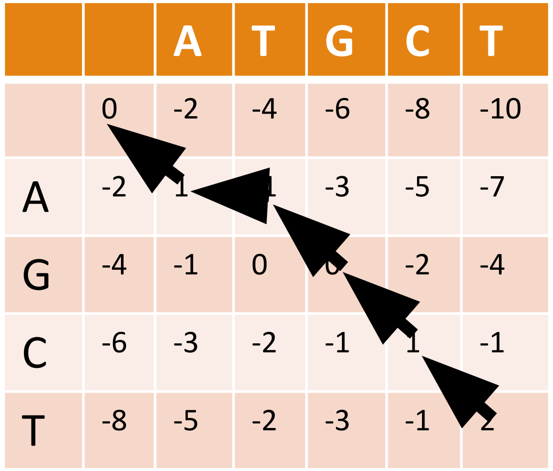
- Scoring Matrix
  - $S(A_i, B_j) = 1$, when $A_i = B_j$
  - $S(A_i, B_j) = -1$, when $A_i \neq B_j$
  - $d = -2$

Biological Sequence Alignments

- Final Matrix

		A	T	G	C	T
	0	-2	-4	-6	-8	-10
A	-2	1	-1	-3	-5	-7
G	-4	-1	0	0	-2	-4
C	-6	-3	-2	-1	1	-1
T	-8	-5	-2	-3	-1	2

- Backtracking



		A	T	G	C	T
	0	-2	-4	-6	-8	-10
A	-2	1	-1	-3	-5	-7
G	-4	-1	0	0	-2	-4
C	-6	-3	-2	-1	1	-1
T	-8	-5	-2	-3	-1	2

ATGCT
A-GCT