CIIC 4025 Analysis and Design of Algorithms

WILFREDO LUGO, PHD

Biological Sequence Alignments

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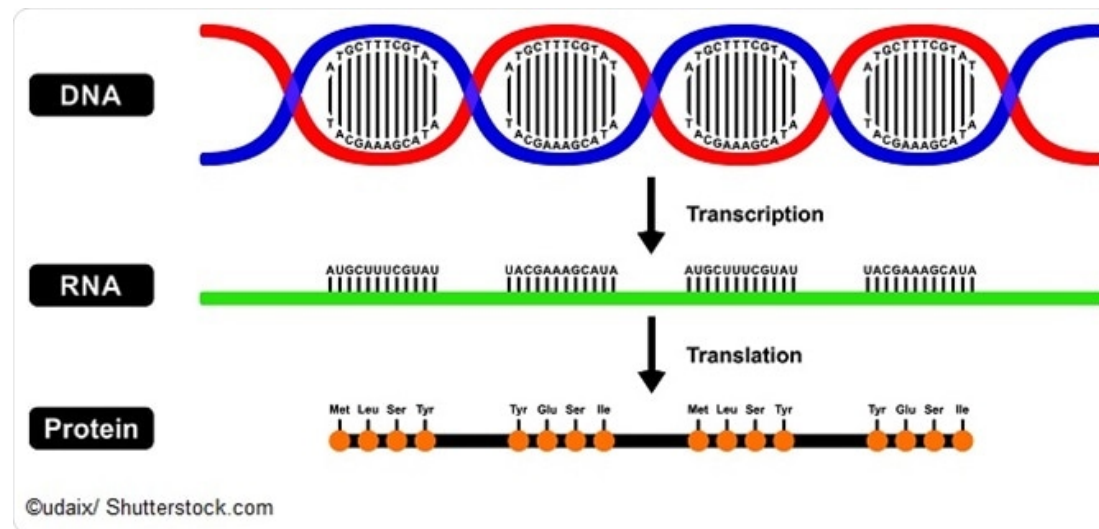   o Amino Acid Sequences

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

Biological Sequence Alignments

o Why do we need to align the sequences?
  o Get more information about
    o Functional Relationships
    o Structural Relationships
    o Evolutionary Relationships

o Types of Alignments
  o Global Alignments
    o Tries to align an entire sequence
    o Align all letters from query and target
    o Suitable for closely related sequences
  o Local Alignments
    o Align regions having highest similarities
    o Align substrings of target with substring of queries
    o Suitable for more divergent sequences.

input string

```
HEAGAWGHEEAHGEGAE
PAWHEAEHE
```

Global alignment

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

Local alignment

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

Biological Sequence Alignments

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

o 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

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

$$\begin{matrix} F_{0,j} = d * j \\ F_{i,0} = d * i \end{matrix}$$ 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

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

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

Scoring Matrix

Gap Penalty

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

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

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

○ 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

B_j

4	0
-3	5

A_i

Matched
(from larger to smaller)

B_j

3	0
-3	2

A_i

Mismatched
(from smaller to larger)

B_j

A_i

Gap

Choose the largest value