AMRITA VISHWA VIDYAPEETHAM
DEEMED TO BE UNIVERSITY

22BIO211: Intelligence of Biological Systems - 2

INTRODUCTION TO SEQUENCE ALIGNMENT

Dr. Manjusha Nair M
Amrita School of Computing, Amritapuri

Email : manjushanair@am.amrita.edu
Contact No: 9447745519

Sequence alignment

- Compare two sequences
  - *The Hamming distance - counts mismatches in two strings*
    - rigidly assumes that we align the i-th symbol of one sequence against the i-th symbol of the other.
- Since biological sequences are subject to insertions and deletions
  - *it is often the case that the i-th symbol of one sequence corresponds to a symbol at a completely different position in the other sequence*
- The goal, is to find the most appropriate correspondence of symbols.

Sequence alignment

- For example, ATGCATGC and TGCATGCA have no matching positions, and so their Hamming distance is equal to 8:

ATGCATGC

TGCATGCA

- These strings have seven matching positions if we align them differently:

ATGCATGC-

- TGCATGCA

- Another alignment of two sequences

ATGC-TTA-

-TGCATTAA

A Good alignment - one that matches as many symbols as possible.

Sequence alignment

- We now define an **alignment of sequences** v and w as a two-row matrix
  - *such that the first row contains the symbols of v (in order),*
  - *the second row contains the symbols of w (in order),*
  - *and space symbols may be interspersed throughout both strings,*
    - as long as two space symbols are not aligned against each other.

```
A T - G T T A T A
A T C G T - C - C
```

Sequence alignment

- An alignment presents one possible scenario where v could have evolved into w.
  - *Columns containing the same letter in both rows are called **matches** and represent conserved nucleotides.*
  - *Columns containing different letters are called **mismatches** and represent single-nucleotide substitutions*
- Columns containing a space symbol are called **indels**
  - *A column containing a space symbol in the top row of the alignment is called an **insertion***
    - as it implies the insertion of a symbol when transforming v into w;

Sequence alignment

- Columns containing a space symbol are called **indels**
  - *A column containing a space symbol in the bottom row of the alignment is called a **deletion***
    - as it indicates the deletion of a symbol when transforming v into w.

```
A T - G T T A T A
A T C G T - C - C
```

    - The alignment above has four matches, two mismatches, one insertion, and two deletions.

Sequence alignment – Longest Common Subsequence

- Matches in an alignment of two strings define a **common subsequence** of the two strings
  - *or a sequence of symbols appearing in the same order (although not necessarily consecutively) in both strings.*

```
A T - G T T A T A
A T C G T - C - C
```

- **ATGT** is a common subsequence of the **above.**
- An alignment of two strings maximizing the number of matches corresponds to the **longest common subsequence** of these strings.
- Note that two strings may have more than one longest common subsequence.

Sequence alignment – Longest Common Subsequence

- **Longest Common Subsequence Problem:** *Find a longest common subsequence of two strings.*

- **Input:** Two strings.
- **Output:** A longest common subsequence of these strings.

We need to devise an algorithm for the longest common subsequence of two strings

Sequence alignment as a game

- We can define a two-player game for sequence alignment
- At each turn of the game, the player has two choices
  - *He can remove the first symbol from each sequence*
    - in which case he earns a point if the symbols match
  - *He can remove the first symbol from either of the two sequences*
    - in which case he earns no points but may set himself up to earn more in later moves.
- The goal is to maximize the number of points.

Sequence alignment as a game

- Example :
  - *Two sequences ATGTTATA and ATCGTCC.*
  - *At each step, we choose to remove either one or both symbols from the left of the two sequences.*
    - we add it to a growing alignment of ATGTTATA and ATCGTCC on the right.
  - *If we remove both symbols, then we align them in the "growing alignment".*
  - *If we remove only one symbol, then we align this symbol with a space symbol in the growing alignment.*

Sequence alignment as a game : Example

Growing alignment	Remaining symbols	Score	
	A T G T T A T A A T C G T C C		Both A' s are removed
A A	T G T T A T A T C G T C C	+1	Both T' s are removed
A T A T	G T T A T A C G T C C	+1	C is removed from the second
A T – A T C	G T T A T A G T C C		Both G' s are removed
A T – G A T C G	T T A T A T C C	+1	Both T' s are removed
A T – G T A T C G T	T A T A C C	+1	T is removed from the first
A T – G T T A T C G T –	A T A C C		A is removed from the first, C is removed from the second
A T – G T T A A T C G T – C	T A C		T is removed from the first
A T – G T T A T A T C G T – C –	A C		A is removed from the first, C is removed from the second
A T – G T T A T A A T C G T – C – C			

- Total Score = 4
- Many such alignments are possible

Summary

- Sequence Alignment
- Longest Common Subsequence Problem
- Sequence Alignment as a Game
  - *Example*