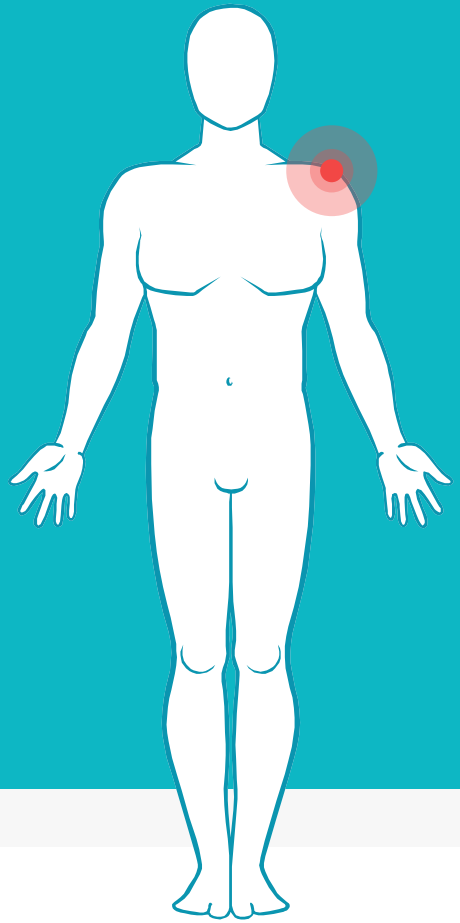
Sequence Alignment

Lecture – 4

Nafis Neehal, Lecturer, Department of CSE, DIU



CONTENTS

1. Sequence Alignment

Why and how align sequences

Sequence Alignment

A way of arranging the sequences of DNA, RNA, or protein to identify regions of similarity that may be a consequence of functional, structural, or evolutionary relationships between the sequences

```
CTGTCG-CTGCACG
-TGC-CG-TG----
```

2. Sequence Alignment Methods

Pairwise and Multiple

Pairwise Sequence Alignment

- A pair of sequences as input
- Align them in such a way that, for that particular alignment the assumed region of similarity produces higher score than all the other alignments
- Methods
  - Global Alignment (Needleman-Wunsch)
  - Local Alignment (Smith-Waterman)

```
CTGTCGCTGCACG--
--------TGC-CGTG
```

Multiple Sequence Alignment

Human specific
Primate specific
Ancient variant

```
Human     ATGAACGCATGC
Chimp.    ATGCACGCATGC
Gorilla   ATGCATGCATGC
Mouse     ATGCATGCATGC
Ancestor  ATGCATGCACGC
Horse     ATGCATGCACGC
```

- Three or more than three sequences as input
- Align all the sequences altogether in such a manner that the alignment produces highest score

3. Pairwise Sequence Alignment

Global and Local methods

Global Alignment (Needleman-Wunsch)

3 Major Steps

- Create 2D Matrix
- Trace back
- Final Alignment

Create 2D Matrix

- Row x Col 2D matrix draw (Row , Col size of seq1 and seq2 respectively)
- Place 2 seqs as Row and Column Header
- Cell (0,0) = 0
- Cell (0,1) to Cell (0,Column) and Cell (1,0) to Cell (Row,0) value = delete gap value from previous cell value
- For other cell values, follow equation in (1)

Trace back

- Start from Cell (Row, Col)
- Go back up to Cell (0,0)

Final Alignment

- Start from Cell (Row, Col)
- If ↖ then, place character in both seq
- If ← or ↑ then character in start seq & gap in end seq

Global Alignment (Needleman-Wunsch) - Example

Input
- seq1 = AAAC
- seq2 = AGC

Scoring Scheme

δ(x, x) = 1 (Match)
δ(x,-) = -2 (Gap)
δ(x, y) = -1 (Mis match)

Final Alignment

-AGC

AAAC

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

$$V_{i,j} = \max \begin{cases} V_{i-1,j} + \delta(s_i, -) \\ V_{i,j-1} + \delta(-, t_j) \\ V_{i-1,j-1} + \delta(s_i, t_j) \end{cases}$$

Eq. 1: Cell Value

Local Alignment (Smith-Waterman)

3 Major Steps

- Create 2D Matrix
- Trace back
- Final Alignment

Create 2D Matrix

- Row x Col 2D matrix draw (Row , Col size of seq1 and seq2 respectively)
- Place 2 seqs as Row and Column Header
- First Row, First Column all value = 0
- For other cell values, follow equation in (2)

Trace back

- Start from each Cell which has the maximum value in the entire matrix
- Go back up to the Cell where first time 0 occurs

Final Alignment

- Start from each Cell with max value
- If ↖ then, place character in both seq
- If ← or ↑ then character in start seq & gap in end seq

Local Alignment (Smith-Waterman) - Example

Input
- seq1 = AAAC
- seq2 = AAG

Scoring Scheme

δ(x, x) = 1 (Match)
δ(x,-) = -2 (Gap)
δ(x, y) = -1 (Mis match)

$$A[i,j] = \max \begin{cases} A[i,j-1] + \text{gap} \\ A[i-1,j] + \text{gap} \\ A[i-1,j-1] + \text{match}(i,j) \\ 0 \end{cases}$$

Eq. 2: Cell Value

Final Alignment

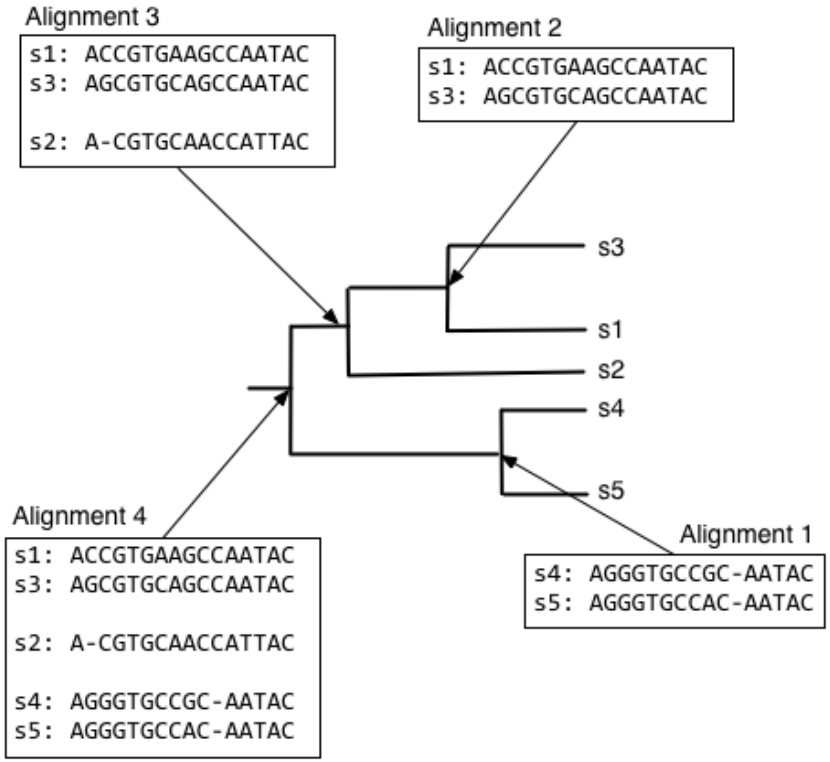
```
-AAG
AAAC
```

		A	A	G
	0	0	0	0
A	0	1	1	0
A	0	1	2	0
A	0	1	2	1
C	0	0	0	1

4. Multiple Sequence Alignment

Progressive, Iterative

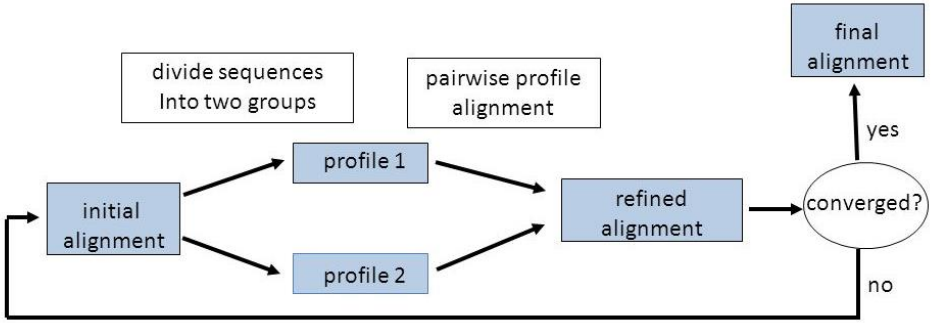
Progressive Method



- Two major steps – Guide Tree build up and Multiple Pairwise Alignment
- Steps
  - Take each pair, align
  - Generate consensus of that alignment
  - Align new sequence with the consensus of the previous one
  - Go back, Until all sequences are finished
- Example
  - Clustal ω
  - MAFFT
  - KALIGN
  - T-COFFEE

Iterative Method

- Works similarly to progressive methods
- Repeatedly realign the initial sequences as well as add new sequences to the growing MSA
- Example
  - DIALIGN
  - MUSCLE
  - POA



MSA Challanges

- Computationally Expensive
- Difficult to score. Multiple comparison necessary in each column of the MSA for a cumulative score
- Placement of gaps and scoring of substitution is more difficult
- Difficulty increases with diversity
- Relatively easy for a set of closely related sequences. Identifying the correct ancestry relationships for a set of distantly related sequences is more challenging
- Even difficult if some members are more alike compared to others

95%

Of Human DNA is identical to Chimpanzees

2 gm DNA

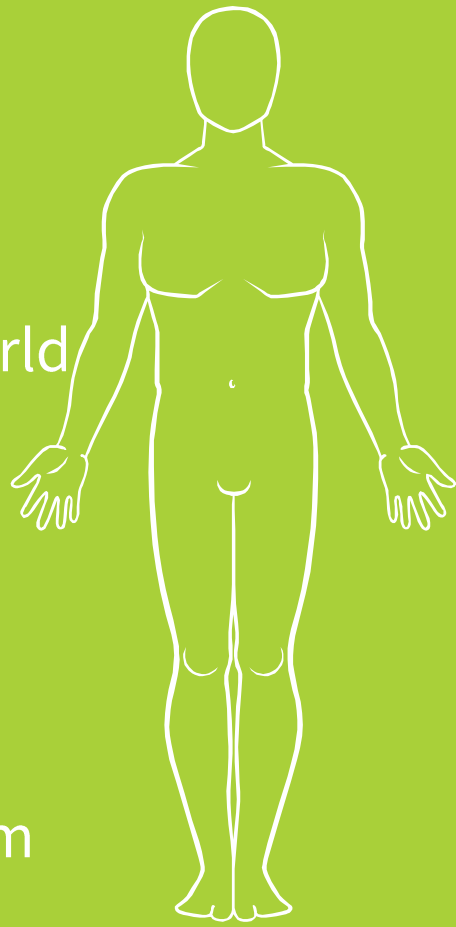
Can contain digital information of whole world

510 DNA Codes

Lost throughout human evolution

1.8 Meter

Long DNA is squeezed into a space of 0.09 µm



TO BE CONTINUED

Shocked?

Youtube Links

- Global Alignment Part 1 - https://www.youtube.com/watch?v=vqxc2EfPWdk
- Global Alignment Part 2 - https://www.youtube.com/watch?v=zwA-6_1bLgE
- Local Alignment - https://www.youtube.com/watch?v=latoW0sJ35Q