Lecture 2: Sequence Alignment

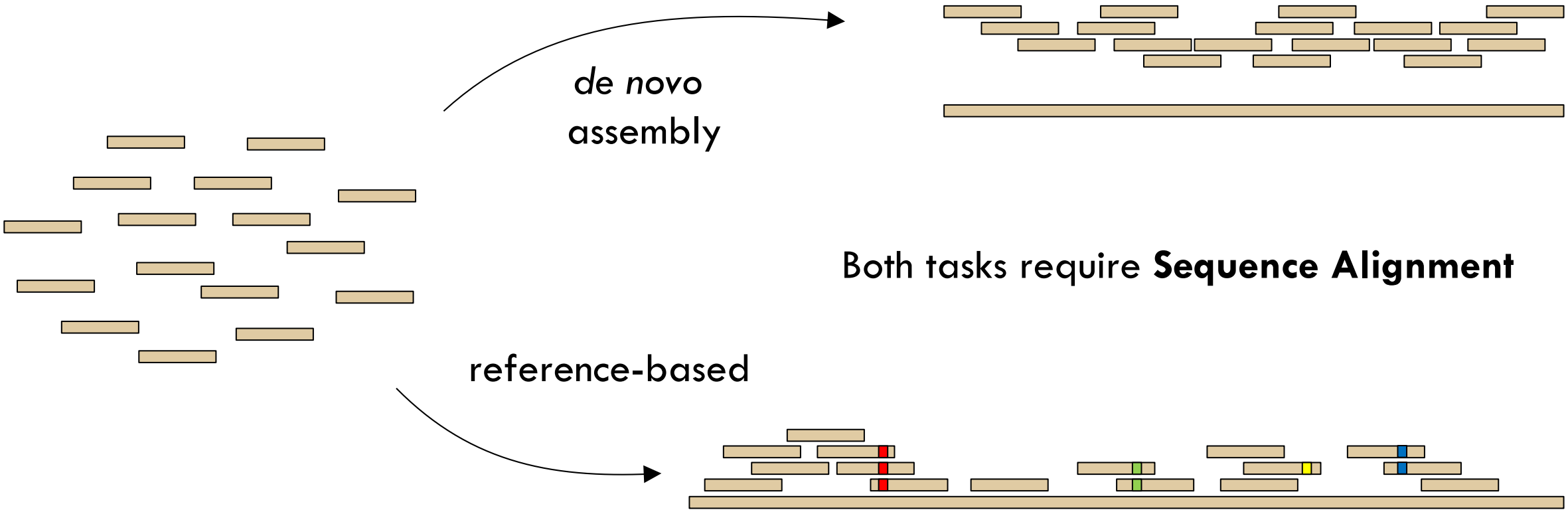
ECE 365 - Data Science and Genomics

Announcements

- Lab 1 (Introduction to Genomics) released
  - due March 11 at 11:59pm
  - Submit via gradescope

From last lecture: processing shotgun sequencing data

1. If a genome from that species has never been sequenced before



Both tasks require **Sequence Alignment**

2. If a genome from that species has already been sequenced

(slides adapted from Stanford CS262 lectures)

Sequence Alignment

x = **AGGCTATCACCTGACCTCCAGGCCGATGCCC**
y = **TAGCTAGCACGACCGCGGTCGATTTGCCCGAC**

"Definition": Given two sequences x and y, place gaps ('-') in them so that the resulting sequences "line up well"

```
-AGGCTATCACCTGACCTCCAGGCCGA--TGCCC---
TAG-CTAGCAC--GACCGC--GGTCGATTTGCCCGAC
```

(slides adapted from Stanford CS262 lectures)

What is a good alignment?

x = AGGCTAGTT
y = AGCGAAGTTT

```
AGGCTAGTT-
AGCGAAGTTT
```

6 matches, 3 mismatches, 1 gap

```
AGGCTA-GTT-
AG-CGAAGTTT
```

7 matches, 1 mismatch, 3 gaps

```
AGGC-TA-GTT-
AG-CG-AAGTTT
```

7 matches, 0 mismatch, 5 gaps

(slides adapted from Stanford CS262 lectures)

Scoring function

- We will score points for matches, and penalize mismatches and gaps:

mismatch:

```
CTATCAC
CTAGCAC
```

gap:

```
CTATCAC
CTA-CAC
```

- Scoring function:

match:	$+m$
mismatch:	$-s$
gap:	$-d$

$$\boxed{\text{Score} = (\#\text{ matches}) \times m - (\#\text{ mismatches}) \times s - (\#\text{ gaps}) \times d}$$

- Let's look at some examples!

Finding best global alignment

- We use an algorithm based on *dynamic programming*

match: $+1$
mismatch: -1
gap: -1

	x=	G	C	A	T	T	C
y	0						
G							
A							
T							
T							
A							
C							

Finding best global alignment

- We use an algorithm based on *dynamic programming*

		G	C	A	T	T	C
	0						
G							
A							
T							
T							
A							
C							

score of best alignment between GCATT and GA

match: $+1$
mismatch: -1
gap: -1

Finding best global alignment

- We use an algorithm based on *dynamic programming*

match: $+1$
mismatch: -1
gap: -1

		G	C	A	T	T	C
	0						
G							
A							
T							
T							
A							
C							

score of best alignment between GCA and GATTA

Finding best global alignment

- We use an algorithm based on *dynamic programming*

match: $+1$
mismatch: -1
gap: -1

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

$$H_{ij} = \max \begin{cases} H_{i-1,j-1} + m, \\ \quad (\text{if } x_i = y_j) \\ H_{i-1,j-1} - s, \\ \quad (\text{if } x_i \neq y_j) \\ H_{i-1,j} - d \\ H_{i,j-1} - d \end{cases}$$

Finding best global alignment

- We use an algorithm based on *dynamic programming*

G C A T T – C
G – A T T A C

5 - 0 - 2 = 3

match: $+1$
mismatch: -1
gap: -1

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

$$H_{ij} = \max \begin{cases} H_{i-1,j-1} + m, \\ \quad (\text{if } x_i = y_j) \\ H_{i-1,j-1} - s, \\ \quad (\text{if } x_i \neq y_j) \\ H_{i-1,j} - d \\ H_{i,j-1} - d \end{cases}$$

Finding best global alignment

- We use an algorithm based on *dynamic programming*

match: $+1$
mismatch: -1
gap: -1

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

$$H_{ij} = \max \begin{cases} H_{i-1,j-1} + m, \\ \quad (\text{if } x_i = y_j) \\ H_{i-1,j-1} - s, \\ \quad (\text{if } x_i \neq y_j) \\ H_{i-1,j} - d \\ H_{i,j-1} - d \end{cases}$$

```
GCATT-C
G-ATTAC
```

Finding best global alignment

- We use an algorithm based on *dynamic programming*

match: $+1$
mismatch: -1
gap: -1

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

Needleman-Wunsch algorithm

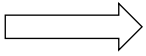
$$H_{ij} = \max \begin{cases} H_{i-1,j-1} + m, \\ \quad (\text{if } x_i = y_j) \\ H_{i-1,j-1} - s, \\ \quad (\text{if } x_i \neq y_j) \\ H_{i-1,j} - d \\ H_{i,j-1} - d \end{cases}$$

```
GCATT-C
G-ATTAC
```

Global vs Local Alignment

- Global alignment: Align x and y fully

x = AGGCTAGTT
y = AGCGAAGTTT

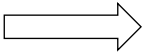


AGGC-TA-GTT-
AG-CG-AAGTTT

Global vs Local Alignment

- Global alignment: Align x and y fully

```
x = AGGCTAGTT
y = AGCGAAGTTT
```



```
AGGC-TA-GTT-
AG-CG-AAGTTT
```

- Local alignment: Align a substring of x to a substring of y

```
x = ACCTACTCGCAATATGCTAGCAGCAACGTA
y = GTCGGGATCGCAATCTGCAGTCCGCTTACC

-ACCTACTCGCAATATGCTAG-CAGCAACGTA
 |||||||||.|||  ||  |.|||
GTCGGGATCGCAATCTGC-AGTCCGCTTACC-
```

Finding best local alignment

- We can adapt the previous algorithm to find local alignments

Finding best local alignment

- We can adapt the previous algorithm to find local alignments

match: $+3$
mismatch: -3
gap: -2

		T	G	T	T	A	C	G	G
	0	0	0	0	0	0	0	0	0
G	0	0	3						
G	0								
T	0								
T	0								
G	0								
A	0								
C	0								
T	0								
A	0								

Smith-Waterman algorithm

$$H_{ij} = \max \begin{cases} H_{i-1,j-1} + m, \\ \quad (\text{if } x_i = y_j) \\ H_{i-1,j-1} - s, \\ \quad (\text{if } x_i \neq y_j) \\ H_{i-1,j} - d \\ H_{i,j-1} - d \\ 0 \end{cases}$$

Finding best local alignment

- We can adapt the previous algorithm to find local alignments

		T	G	T	T	A	C	G	G
	0	0	0	0	0	0	0	0	0
G	0	0	3	1	0	0	0	3	3
G	0	0	3	1	0	0	0	3	6
T	0	3	1	6	4	2	0	1	4
T	0	3	1	4	9	7	5	3	2
G	0	1	6	4	7	6	4	8	6
A	0	0	4	3	5	10	8	6	5
C	0	0	2	1	3	8	13	11	9
T	0	3	1	5	4	6	11	10	8
A	0	1	0	3	2	7	9	8	7

match: +3
mismatch: -3
gap: -2

Smith-Waterman algorithm

$$H_{ij} = \max \begin{cases} H_{i-1,j-1} + m, \\ \quad (\text{if } x_i = y_j) \\ H_{i-1,j-1} - s, \\ \quad (\text{if } x_i \neq y_j) \\ H_{i-1,j} - d \\ H_{i,j-1} - d \\ 0 \end{cases}$$

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

Finding best local alignment

- We can adapt the previous algorithm to find local alignments

match: +3
mismatch: s= +3
gap: d= +2

		T	G	T	T	A	C	G	G
	0	0	0	0	0	0	0	0	0
G	0	0	3	1	0	0	0	3	3
G	0	0	3	1	0	0	0	3	6
T	0	3	1	6	4	2	0	1	4
T	0	3	1	4	9	7	5	3	2
G	0	1	6	4	7	6	4	8	6
A	0	0	4	3	5	10	8	6	5
C	0	0	2	1	3	8	13	11	9
T	0	3	1	5	4	6	11	10	8
A	0	1	0	3	2	7	9	8	7

Smith-Waterman algorithm

$$H_{ij} = \max \begin{cases} H_{i-1,j-1} + m, \\ \quad (\text{if } x_i = y_j) \\ H_{i-1,j-1} - s, \\ \quad (\text{if } x_i \neq y_j) \\ H_{i-1,j} - d \\ H_{i,j-1} - d \\ 0 \end{cases}$$

GTT-AC
GTTGAC

Alignment problem variations

Global alignment

GCATT-C
G-ATTAC

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

Local alignment

T **GTT-AC** GG
G **GTTGAC** TA

		T	G	T	T	A	C	G	G
	0	0	0	0	0	0	0	0	0
G	0	0	3	1	0	0	0	3	3
G	0	0	3	1	0	0	0	3	6
T	0	3	1	6	4	2	0	1	4
T	0	3	1	4	9	7	5	3	2
G	0	1	6	4	7	6	4	8	6
A	0	0	4	3	5	10	8	6	5
C	0	0	2	1	3	8	13	11	9
T	0	3	1	5	4	6	11	10	8
A	0	1	0	3	2	7	9	8	7

Alignment problem variations

Global alignment

GCATT-C
G-ATTAC

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

Local alignment

T **GTT-AC** GG
G **GTTGAC** TA

		T	G	T	T	A	C	G	G
	0	0	0	0	0	0	0	0	0
G	0	0	3	1	0	0	0	3	3
G	0	0	3	1	0	0	0	3	6
T	0	3	1	6	4	2	0	1	4
T	0	3	1	4	9	7	5	3	2
G	0	1	6	4	7	6	4	8	6
A	0	0	4	3	5	10	8	6	5
C	0	0	2	1	3	8	13	11	9
T	0	3	1	5	4	6	11	10	8
A	0	1	0	3	2	7	9	8	7

Alignment to reference

GCACT-C → short
TCT **G-ACTAC** GCGT

long

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

Alignment problem variations

Global alignment

GCATT-C
G-ATTAC

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

Local alignment

T **GTT-AC** GG
G **GTTGAC** TA

		T	G	T	T	A	C	G	G
	0	0	0	0	0	0	0	0	0
G	0	0	3	1	0	0	0	3	3
G	0	0	3	1	0	0	0	3	6
T	0	3	1	6	4	2	0	1	4
T	0	3	1	4	9	7	5	3	2
G	0	1	6	4	7	6	4	8	6
A	0	0	4	3	5	10	8	6	5
C	0	0	2	1	3	8	13	11	9
T	0	3	1	5	4	6	11	10	8
A	0	1	0	3	2	7	9	8	7

Alignment to reference

GCACT-C
TCT **G-ACTAC** GCGT

		T	C	T	G	A	C	T	A	C	G	C	G	T
G														
C														
A														
C														
T														
C														

Overlap alignment

GCACT-C GCGTTCA
TTCT **G-ACTAC**

		G	C	A	C	T	C	G	C	G	T	T	C	A
T														
T														
C														
T														
G														
A														
C														
T														
A														
C														