Fast Alignment Heuristics

COMP462/561: Computational Biology Methods

*Based on Course Notes by Dr. Mathieu Blanchette

1

Smith-Waterman: Local Alignment (1981)

Problem:

Given two sequences, *S* and *T*, of lengths *m* and *n*, find the substring *s* of *S* and the substring *t* of *T* such that the alignment score of *s* against *t* is maximized

Algorithm: Dynamic programming algo (very similar to NW)

1. Initialization matrix
2. Fill matrix with appropriate alignment scores
3. Trace back from highest scoring cell(s) to find best alignment(s)

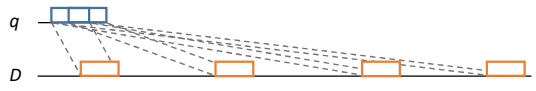
2

Motivation

Problem:

Given a query sequence, *q*, of length *m* (small, ~1000 nucleotides) and a large database (target), *D*, of size *n* (billions of nucleotides)

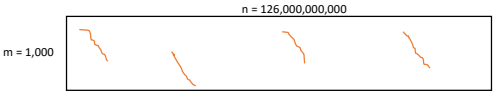
Find all local alignments of *q* within *D* that have a score above threshold, *T*



3

Smith-Waterman?

- SW is **too slow**…would take O(mn + m*hits)
- How?
  - Trace back all entries of a dynamic programming matrix with a score>*T*



- ***Too slow, too much memory!***

4

SW Initialization

A = CGTGAATTCAT

B = GCTTAC

$H(i,0) = 0, 0 \leq i \leq m$

$H(0,j) = 0, 0 \leq j \leq n$

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

5

SW Matrix Filling

Similar to **Needleman-Wunsch (NW)**, fill in the matrix such that:

$$H(i,j) = max^1 \begin{cases} H(i-1,j-1) + s(a_i, b_j) & Match/Mismatch \\ max_{k \geq 1}\{H(i-k,j) + g\} & Deletion \\ max_{l \geq 1}\{H(i,j-l) + g\} & Insertion \end{cases}, 1 \leq i \leq m, 1 \leq j \leq n$$

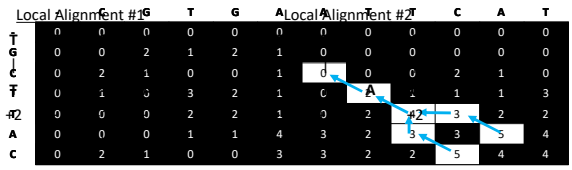
	-	C	G	T	G	A	A	T	T	C	A	T
-	0	0	0	0	0	0	0	0	0	0	0	0
G	0	0	2	1	2	1	0	0	0	0	0	0
C	0	2	1	0	1	1	0	0	0	2	1	0
T	0	0	0	3	2	1	0	0	1	1	2	1
T	0	0	0	2	2	1	0	0	2	1	1	3
A	0	0	0	1	1	4	3	2	1	3	2	2
C	0	2	1	0	0	3	3	2	2	5	4	4

With a match score of +2 and a mismatch & indel score equal to -1.

6

SW Trace Back

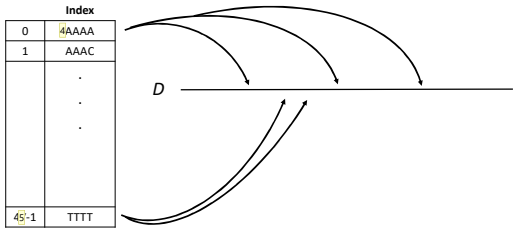
- With NW, we trace back from the bottom right-most cell of the matrix
  - Slightly different with SW. **How?**



7

7

Preprocess Database to Build Indices



10

10

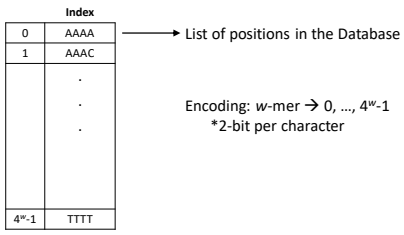
Basic Local Alignment Search Tool Idea

- Give up on (guaranteed) optimality
  - *Heuristic* approach
  - Search only for local-alignments with **high-scoring gapless alignments (HSPs)**
- Pre-process the database, *D*, so that queries can be answered in constant time with respect to *n*
- **BLAST** was published in 1990 by Altschul, Lipman, Miller,
  - cited by more than 10^5 papers

8

8

Indexing the Database



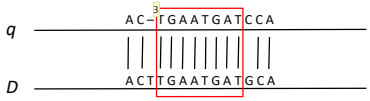
Encoding: *w*-mer → 0, …, 4^w-1
*2-bit per character

11

11

Gapless Alignments

- If *q* has a good alignment, *X*, somewhere in *D*



- Then *X* is likely to contain a HSP

9

9

Scanning for Hits in *D*

- Given a query, *q*

For each *w*-mer in q — O(|q|)
  For each matches c: q in *D* — O(w)
    Investigate match further…

How many hits do we expect for a *w*-mer of size 11?

$$\frac{3 \times 10^9}{4^{11}} = 1000$$

12

12

Slide 9

2 Assumption:
Lê Nhật Hưng, 23-Sep-19

1 "High scoring pair"
Lê Nhật Hưng, 23-Sep-19

3 Portion in q 100% identical to portion in D
Lê Nhật Hưng, 23-Sep-19

Slide 10

4 Size of portion expect to have perfect match is 4.

Usually, it's 11
Lê Nhật Hưng, 23-Sep-19

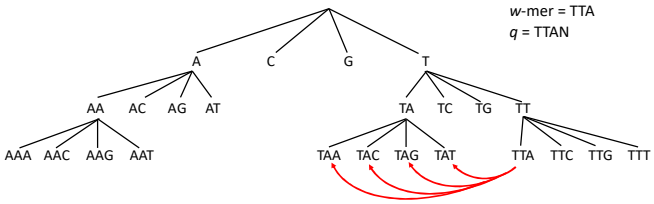
5 w=4 here
Lê Nhật Hưng, 23-Sep-19

Slide 12

6 TYPO: must be w
Lê Nhật Hưng, 23-Sep-19

Improving Scan Times

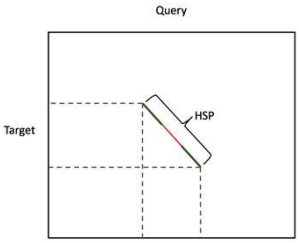
- Encode database indices in a **trie**



13

13

BLAST HSP



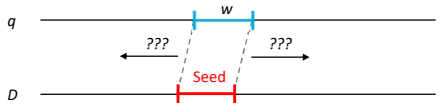
How can we improve the hit further?

16

16

Extending hits to find High Scoring Pair (HSP)

Goal: Quickly determine if a match is promising or not



14

14

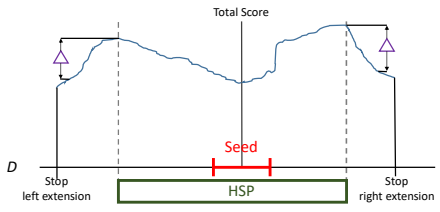
Statistics of Local Alignments

- Even if *D* was completely random, we would expect to observe some pretty high scoring HSPs
  - How do we know when we should get excited?
- **E-value** (score(HSP))
  - The expected number of local alignments with a score greater or equal to HSP's that would be found in a random *D*

17

17

Ungapped Extension Phase



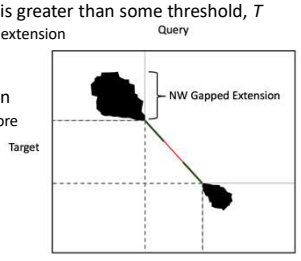
Time? Linear in size of extension

15

15

Gapped Extension

- If the HSP's alignment score is greater than some threshold, *T*
  - Do a more expensive gapped extension
    - Using a variant NW
- Perform NW in each direction
  - Consider only entries with score greater than "best so far"



18

18

Choosing the size of *w*

Small (<=11)
- High probability of finding exact *w*-mer in HSP
- Lots of false positive seeds
- High sensitivity
- Slow

Large (>12)
- Miss many HSPs
- Few false positives
- Low sensitivity
- Fast

19

19

Karlin-Altschul (1990)

$$E(S) = Kmn$$

- $E(S) = K\ m\ n\ e$
- S is the score of the ungapped HSP alignment
- K and depend on the scoring scheme and background probabilities
  - scales scores scheme
- A low E-value (10^{-1} - 10^{-100}) is a good match
  - Low chance of observing HSP given random chance alone

20

20

Variants

For proteins: inexact matches are considered
- Based on a **point accepted matrix (PAM)**

Query	Target	BLAST variant
DNA	DNA	blastn
Protein	Protein	blastp
Protein	DNA	tblastn
DNA	protein	blastx

21

21

Optimizations

- **Dealing with repeats in q or D**
- **Two-Hit method**
  - Lower *T* to allow more hits, but only extend if two hits fall on the same diagonal
    - Within a window of fixed length
    - *Increases* hits and *lowers* extensions
- **Gapped seeds**

22

22

Upcoming Topics

- Wednesday – **multiple sequence alignment (MSA)**
  - Dr. Blanchette will return!
- End of the semester – **Burrows-Wheeler Transform (BWT)**
  - https://en.wikipedia.org/wiki/Burrows%E2%80%93Wheeler_transform
  - In pattern matching: https://www.youtube.com/watch?v=z5EDLODQPtg

23

23