# From Smith-Waterman to *BLAST*

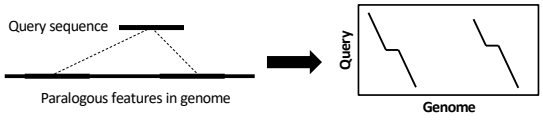
Jeremy Buhler (*in absentia*)

Wilson Leung 12/25/2022

1

## Key limitations of the Smith-Waterman **local** alignment algorithm

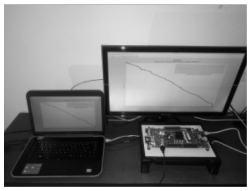
- **Quadratic** in time and space complexity
- Report only **one optimal alignment**
  - Usually want all interesting alignments
  - Example: map a mRNA against a genome




2

## Smith-Waterman implementations

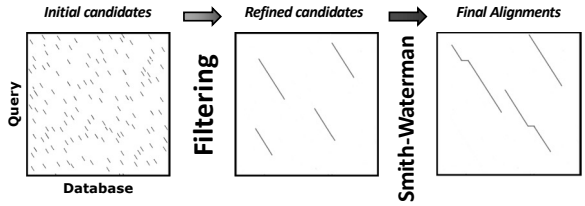
- Bill Pearson's ssearch
  - https://fasta.bioch.virginia.edu/fasta_www2/fasta_list2.shtml
- *Water* (EBI / EMBOSS)
  - https://www.ebi.ac.uk/Tools/psa/emboss_water/
- Hardware solutions:
  - Graphical Processing Units (GPUs)
  - Field-Programmable Gate Arrays (FPGAs)



Oliveira FF, Dias LA, Fernandes MAC. Proposal of Smith-Waterman algorithm on FPGA to accelerate the forward and backtracking steps. PLoS One. 2022 Jun 30;17(6):e0254736.

3

## BLAST alignment strategy: generate and filter




- Goal: minimize the need for calculating Smith-Waterman alignments

4

## Challenges with the BLAST alignment strategy

1. Identify candidate patterns
2. Find the best alignment "near" a candidate

5

## Identify candidate patterns

- High-scoring alignment between two sequences will contain some **consecutive matches**
- Treat **k-mer** (word) matches as candidates

**(k = 4)**

…atacatcactacgatcc-a…
…agacatg--tgcaatccca…

6

## Locate k-mer matches (k=3)

Query:

| 1 | 2 | 3 | 4 | 5 | 6 | 7 |
|---|---|---|---|---|---|---|
| a | c | g | t | a | c | g |

**3-mers in query**

| 3-mer | Position(s) |
|---|---|
| acg | 1, 5 |
| cgt | 2 |
| gta | 3 |
| tac | 4 |

**3-mer matches in database**

| 1 | 2 | 3 | 4 | 5 | 6 | 7 |
|---|---|---|---|---|---|---|
| a | c | g | c | g | t | a |

| k-mer match | Database position(s) | Query position(s) |
|---|---|---|
| acg | 1 | 1, 5 |
| cgt | 4 | 2 |
| gta | 5 | 3 |

7

## Use a hash table to more efficiently store k-mers

- A table of **$4^k$ entries** is required to store all possible k-mers of a DNA query sequence
- BLAST uses a **hash table** to store k-mers
  - Space requirement proportional to the query size
- Reduces the time required to the **sum** of the lengths of the two sequences

8

## Other "Build a Table" abstractions

- Search multiple queries against a database
  - BLAT: index the database
- More space-efficient index structures
  - Suffix array
  - Burrows-Wheeler transform
  - FM-index
- Used by second-generation sequence aligners (e.g., BWA, Bowtie)

Li H and Homer N. A survey of sequence alignment algorithms for next-generation sequencing. Briefings in Bioinformatics. 2010 Sep;11(5):473-83.

9

## k-mer size affects the sensitivity and specificity of the search

- How "good" are the candidate matches?
- Trade off between **sensitivity** (true positives) and **specificity** (true negatives)
  - k = 1 (high sensitivity)
  - k = entire sequence (high specificity)

10

## Quantifying specificity

- Given DNA sequences S and T
  - i.i.d. random with equal base frequencies

Probability of 1 bp match: $\frac{1}{4}$

Probability of k-mer match: $\left(\frac{1}{4}\right)^k$

Expected number of k-mer matches: $|S| \cdot |T| \left(\frac{1}{4}\right)^k$

Search 1kb pattern against a 1Gb database:

$\log_4 10^{(3+9)} \approx 20\,bp$

11

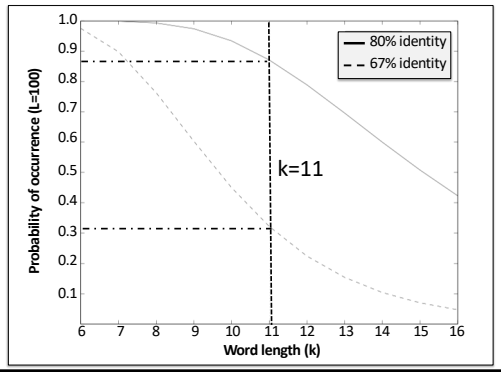
## Quantifying sensitivity

- Require **at least one k-mer match** to detect an alignment between S and T
- Sequences with lower percent identity have fewer k-mer matches
- How large a value of *k* is likely to detect most alignments?

12

## Word length versus probability of occurrence

Target length (L) = 100




13

## Adjust k-mer size based on the level of sequence similarity

- BLAT (k=15)
  - Find highly similar sequences
- *blastn* (k=11)
  - Find most medium to high similarity alignments
  - Most candidates are false positives
- RepeatMasker (k=8)
  - Find highly diverged repeat copies

14

## Use more sensitive parameters to identify the initial transcribed exon

- Program Selection:
  - From *megablast* to ***blastn***
- Word Size:
  - From 11 to **7**
- Match/Mismatch Scores:
  - From +2/-3 to **+1/-1**
- Gap Costs:
  - Existence: from 5 to **2**
  - Extension: from 2 to **1**

15

## Word match for protein sequences

- Use shorter k-mer:
  - *blastp* (k=3)
- Allow approximate matches using similarity:
  - Keep all word matches with score ≥ T (**neighborhood**)
- Reduce number of spurious candidates:
  - Require two word matches along the same diagonal (**two-hit algorithm**)

16

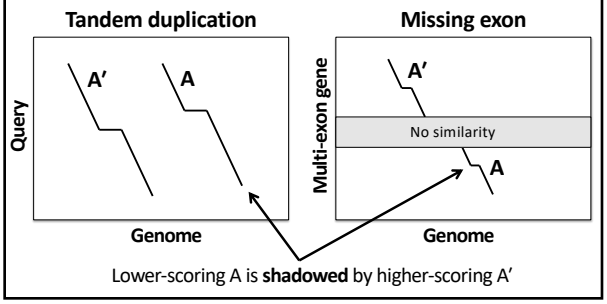
## Use dynamic programming (DP) to filter candidates

- Search the region surrounding each candidate
- Define the **size** and **shape** of the search region
- Report **multiple** high-scoring alignments
  - Align a multi-exon mRNA against a genome
  - Report alignments to all exons

17

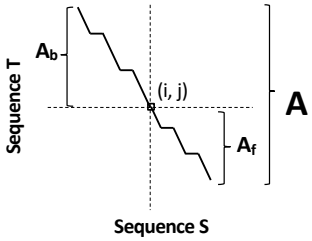
## The "shadowing problem"

- A "good" alignment might be omitted because of a better alignment within the search region




18

## Solution: **pin** the alignment




- Candidate match is centered on S[i], T[j]
- Compute optimal alignments that **pass through (i, j)**
  - Half-anchor alignments

$A_f$ = Best alignment that starts from (i, j)

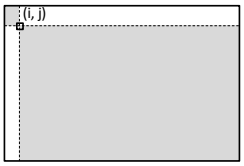
$A_b$ = Best alignment that ends at (i, j)

**A** = Best alignment (combine $A_f$ and $A_b$)

19

## Define the size of the two search regions

- One option: bound the search regions by the **ends of the two sequences**
  - Best case: half of the entire DP matrix
  - Worst case: cost as much as not filtering

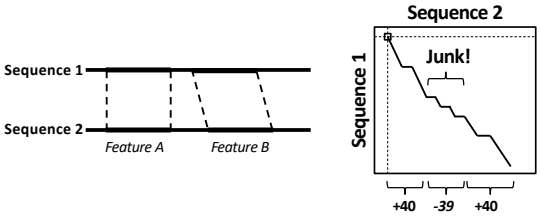



*DP fill region ≥ half of matrix*

20

## The "chaining problem"

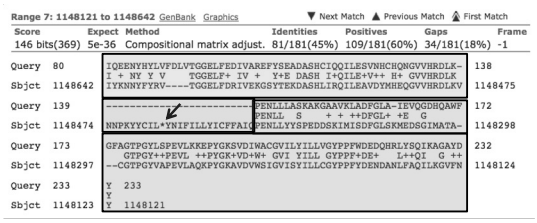
- Opposite problem to shadowing
  - Connect multiple features into a single alignment




21

## BLAST often chains multiple alignment blocks into a single alignment

*tblastn* of CaMKII-PA (query) against the *D. mojavensis* genome (subject)




Mills LJ and Pearson WR. Adjusting scoring matrices to correct overextended alignments. Bioinformatics. 2013 Dec 1;29(23):3007-13.

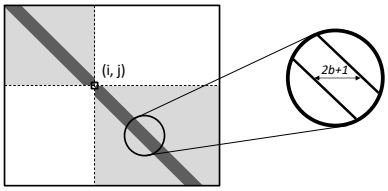
22

## Ignore alignments that are "not promising"

- Ignore alignments with very large gaps
  - Usually have poor score
  - Can identify second feature from its own candidates
- Limit search region to the diagonal surrounding the candidate
  - The **bandwidth** (b) parameter controls the width of the diagonal

23

## Use **banded alignments** to reduce the search space

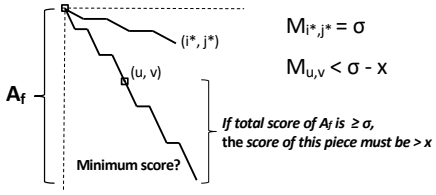



- Number of DP entries to compute is proportional to the **length of the shorter sequence** (times b)

24

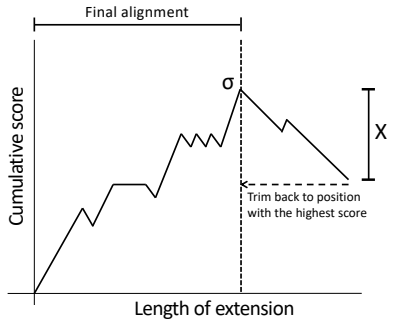
## Use **X-drop** to further reduce the search space

- Terminate the alignment if the score drops **below x** compared to the optimal score




25

## BLAST X-drop strategy




Korf, I., Yandell, M., and Bedell, J. (2003). *BLAST*. O'Reilly Media, Inc.

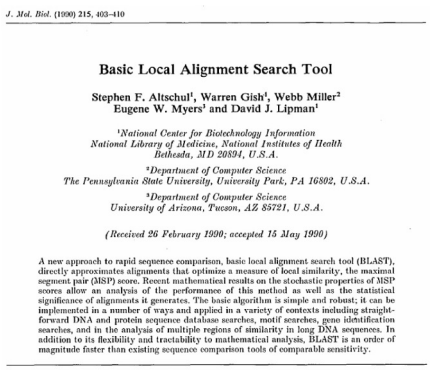
26

## Summary

- BLAST uses a **generate and filter** strategy
  - Generate candidate matches
  - Filter using dynamic programming (DP)
- Mitigates problems with shadowing and chaining
- Minimizes the amount of time spent on DP
  - Banded alignment
  - X-drop

27

## Questions?



*J. Mol. Biol.* (1990) 215, 403–410

**Basic Local Alignment Search Tool**

Stephen F. Altschul[1], Warren Gish[1], Webb Miller[2]
Eugene W. Myers[3] and David J. Lipman[1]

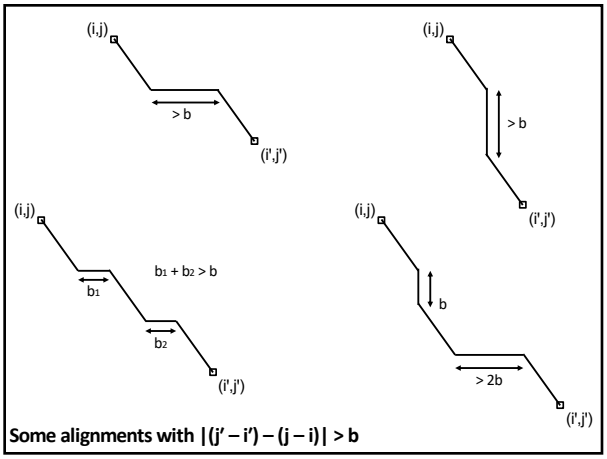
[1]*National Center for Biotechnology Information*
*National Library of Medicine, National Institutes of Health*
*Bethesda, MD 20894, U.S.A.*
[2]*Department of Computer Science*
*The Pennsylvania State University, University Park, PA 16802, U.S.A.*
[3]*Department of Computer Science*
*University of Arizona, Tucson, AZ 85721, U.S.A.*

*(Received 26 February 1990; accepted 15 May 1990)*

A new approach to rapid sequence comparison, basic local alignment search tool (BLAST), directly approximates alignments that optimize a measure of local similarity, the maximal segment pair (MSP) score. Recent mathematical results on the stochastic properties of MSP scores allow an analysis of the performance of this method as well as the statistical significance of alignments it generates. The basic algorithm is simple and robust; it can be implemented in a number of ways and applied in a variety of contexts including straightforward DNA and protein sequence database searches, motif searches, gene identification searches, and in the analysis of multiple regions of similarity in long DNA sequences. In addition to its flexibility and tractability to mathematical analysis, BLAST is an order of magnitude faster than existing sequence comparison tools of comparable sensitivity.

28

29




**Some alignments with $|(j' - i') - (j - i)| > b$**

30