Sequence alignment & search | Paper discussion

Day 09 | F, Jan 28

CompBio2022

BLAST

Your first paper discussion is on an influential algorithm in bioinformatics for sequence comparison called Basic local alignment search tool, or BLAST.

The instructions are on the class website at this link: https://github.com/krishnanlab/teaching/blob/master/2022-spring_compbio/Discussion-notes/Sequence-alignment-and-search_Discussion-notes.md

Group 1

Aidan, Ethan, Kaitlyn, Lydia, Tyler

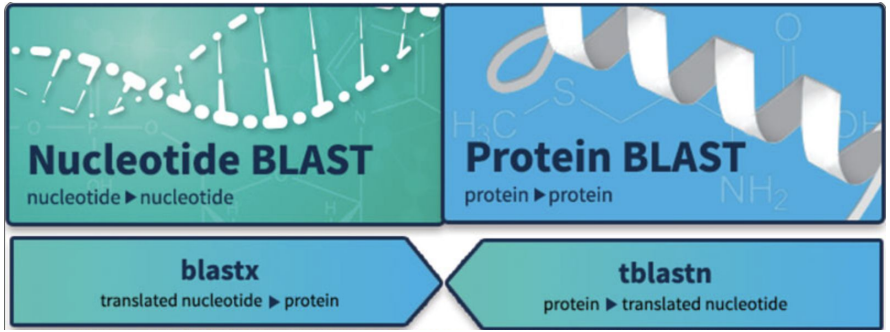
How does BLAST work?

How Does BLAST Work?

- Uses local alignment techniques to find sequences in a database that are similar to a query

  How?

- Breaks larger problem of sequence alignment into smaller subproblems; finds and scores similar pairs
- Accuracy/speed trade off: Takes liberties with Smith-Waterman algorithm to get 50x faster.



https://microbenotes.com/fasta-and-blast/

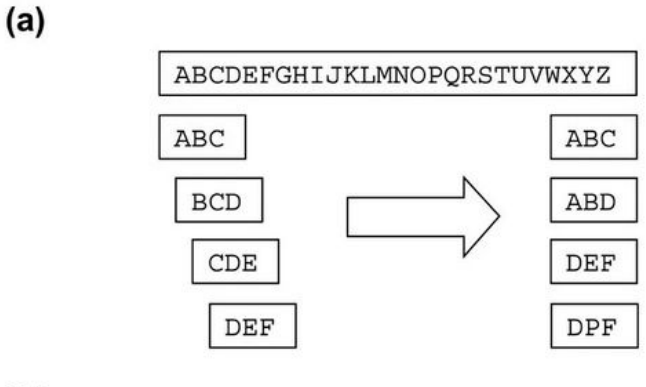
GCATACGGCATA
C - - - T - GGCATA

>

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

Step 1: Filter and Break into Subproblems

- Remove low complexity regions
- The query sequence is broken down into “words”
- For each word, use substitution matrix to find list of words above threshold.
- High scores above a threshold are kept



BLOSUM 62 scoring matrix

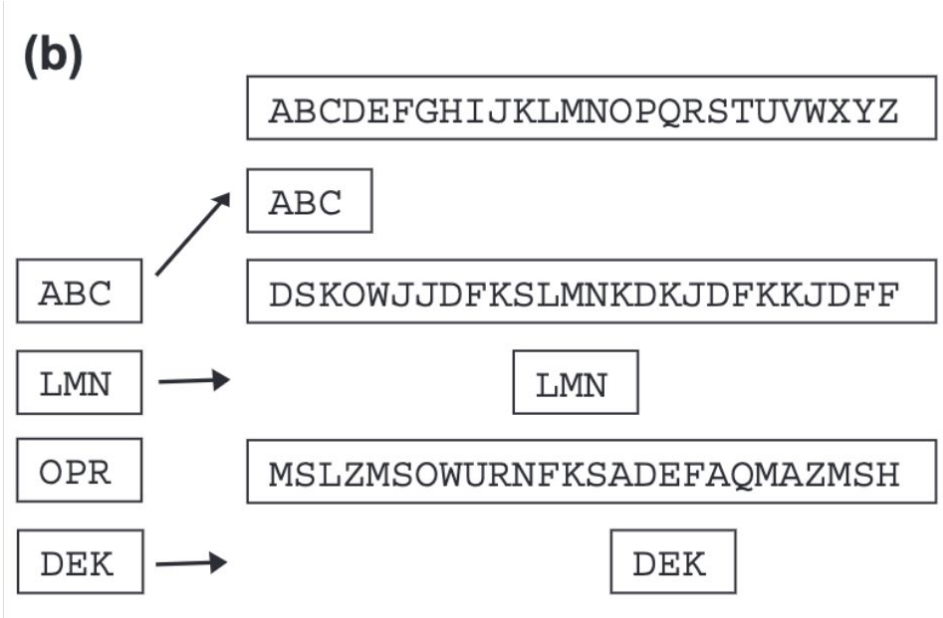
(positive values are shaded)

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
A	4																			
R	-1	5																		
N	-2	0	6																	
D	-2	-2	1	6																
C	0	-3	-3	-3	9															
Q	-1	1	0	0	-3	5														
E	-1	0	0	2	-4	2	5													
G	0	-2	0	-1	-3	-2	-2	6												
H	-2	0	1	-1	-3	0	0	-2	8											
I	-1	-3	-3	-3	-1	-3	-3	-4	-3	4										
L	-1	-2	-3	-4	-1	-2	-3	-4	-3	2	4									
K	-1	2	0	-1	-3	1	1	-2	-1	-3	-2	5								
M	-1	-1	-2	-3	-1	0	-2	-3	-2	1	2	-1	5							
F	-2	-3	-3	-3	-2	-3	-3	-3	-1	0	0	-3	0	6						
P	-1	-2	-2	-1	-3	-1	-1	-2	-2	-3	-3	-1	-2	-4	7					
S	1	-1	1	0	-1	0	0	0	-1	-2	-2	0	-1	-2	-1	4				
T	0	-1	0	-1	-1	-1	-1	-2	-2	-1	-1	-1	-1	-2	-1	1	5			
W	-3	-3	-4	-4	-2	-2	-3	-2	-2	-3	-2	-3	-1	1	-4	-3	-2	11		
Y	-2	-2	-2	-3	-2	-1	-2	-3	2	-1	-1	-2	-1	3	-3	-2	-2	2	7	
V	0	-3	-3	-3	-1	-2	-2	-3	-3	3	1	-2	1	-1	-2	-2	0	-3	-1	4

The values for amino acid substitutions were obtained from Henikoff S & Henikoff JG (1992) Amino acid substitutions matrices from protein blocks. *Proc. Natl. Acad. Sci.* **89**: 10915-10919.

Pertsemlidis, A., Fondon, J.W. Having a BLAST with bioinformatics (and avoiding BLASTphemy). *Genome Biol* **2,** reviews2002.1 (2001). https://doi.org/10.1186/gb-2001-2-10-reviews2002

Step 2: Find the Matches

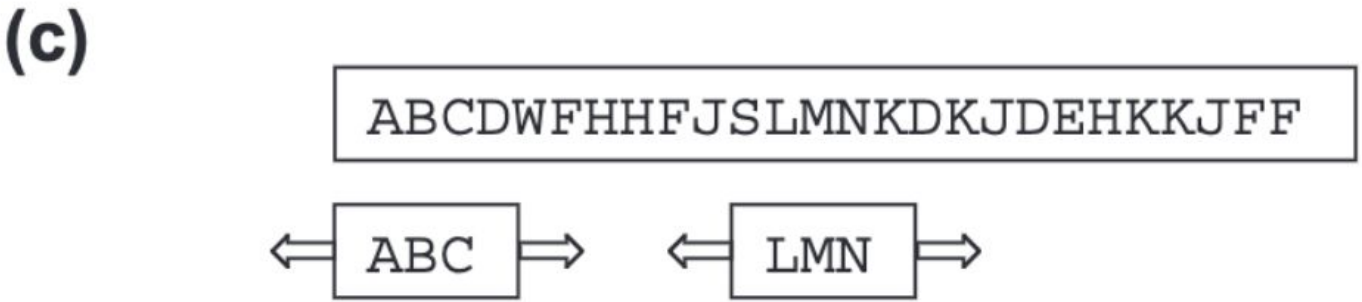


- BLAST has generated HSP word list; searches through target sequences to find matches
- Uses any match to find a possible sequence match from database
- Align these high scoring words to target sequences

Pertsemlidis, A., Fondon, J.W. Having a BLAST with bioinformatics (and avoiding BLASTphemy). *Genome Biol* **2,** reviews2002.1 (2001). https://doi.org/10.1186/gb-2001-2-10-reviews2002

Step 3: Match the Matches!

- Last step: Extend alignments one base at a time until score falls below threshold
- Result are MSPs (maximal scoring pairs)
- Behind the scenes, probability of this alignment is calculated
- BLAST outputs these MSPs as results



Pertsemlidis, A., Fondon, J.W. Having a BLAST with bioinformatics (and avoiding BLASTphemy). *Genome Biol* **2,** reviews2002.1 (2001). https://doi.org/10.1186/gb-2001-2-10-reviews2002

Group 2

Alder, Carly, Gary, Jerry, and Josie

What does the output from the BLAST search look like?

What are the rules to consider when using sequence similarity to infer homology?

Blast Output - Part A and B

(a) **BLASTX** 2.1.2 [Nov-13-2000]

Reference:
Altschul, Stephen F., Thomas L. Madden, Alejandro A. Schäffer,
Jinghui Zhang, Zheng Zhang, Webb Miller, and David J. Lipman (1997),
"Gapped BLAST and PSI-BLAST: a new generation of protein database search programs", Nucleic
Acids Res. 25:3389-3402.

Query = ACAT1_3_2 (101 letters)

Database: /server/blast-db/nr 691,164 sequences;
217,777,941 total letters

Searching.....done

(b)

Sequences producing significant alignments:	Score (bits)	E Value
ref|NP_003092.1| sterol O-acyltransferase (acyl-Coenzyme A:...	70	5e-12
prf||2201440A acyl-CoA/cholesterol acyltransferase [Homo sa...	70	5e-12
gb|AAC62931.1| (AF053337) acyl-CoA:cholesterol acyltransfer...	69	6e-12
gb|AAC62930.1| (AF053336) acyl-CoA:cholesterol acyltransfer...	69	6e-12
pir||I47040 sterol O-acyltransferase (EC 2.3.1.26) - rabbit...	66	9e-11

...

Blast Output - Part C and D

(c)

```
>pir||I47040 sterol O-acyltransferase (EC 2.3.1.26) - rabbit (fragment)
 gb|AAB06959.1| (U65393) acyl-CoA:cholesterol acyltransferase [Oryctolagus
             cuniculus]
           Length = 305

 Score = 65.6 bits (157), Expect = 9e-11
 Identities = 29/32 (90%), Positives = 32/32 (99%)
 Frame = +3

Query: 3   GSHFDDFVTNLIEKSATLDNGGCALTTFSVLE 98
           GSHFDDFVTNLIEKSA+LDNGGCALTTFS+L+
Sbjct: 8   GSHFDDFVTNLIEKSASLDNGGCALTTFSILK 39
```

(d)

```
Database: /server/blast-db/nr
   Posted date:  May 22, 2001  4:03 PM
 Number of letters in database: 217,777,941
 Number of sequences in database:  691,164

Lambda        K        H
   0.318    0.135    0.401

Gapped
Lambda        K        H
   0.270   0.0470    0.230

Matrix: BLOSUM62
Gap Penalties: Existence: 11, Extension: 1
effective length of query: 21
effective length of database: 209,483,973
effective search space: 4399163433
effective search space used: 4399163433
frameshift window, decay const: 50,  0.1
...
```

GB.c

Rules for Sequence Similarity

- **Compare protein sequences if the query sequences encode proteins**
- **Pay close attention to the statistics**
- **Avoid reporting raw BLAST scores**
- **Know the difference between sensitivity and selectivity**
  - Sensitivity: the ability of a method to recognize distantly related sequence
  - Selectivity: lowering the scores for unrelated sequences
- **Remember that sequence data include experimental artifacts**

Group 3

Annalise,Maria, Mitchell, Mitch, and Sneha

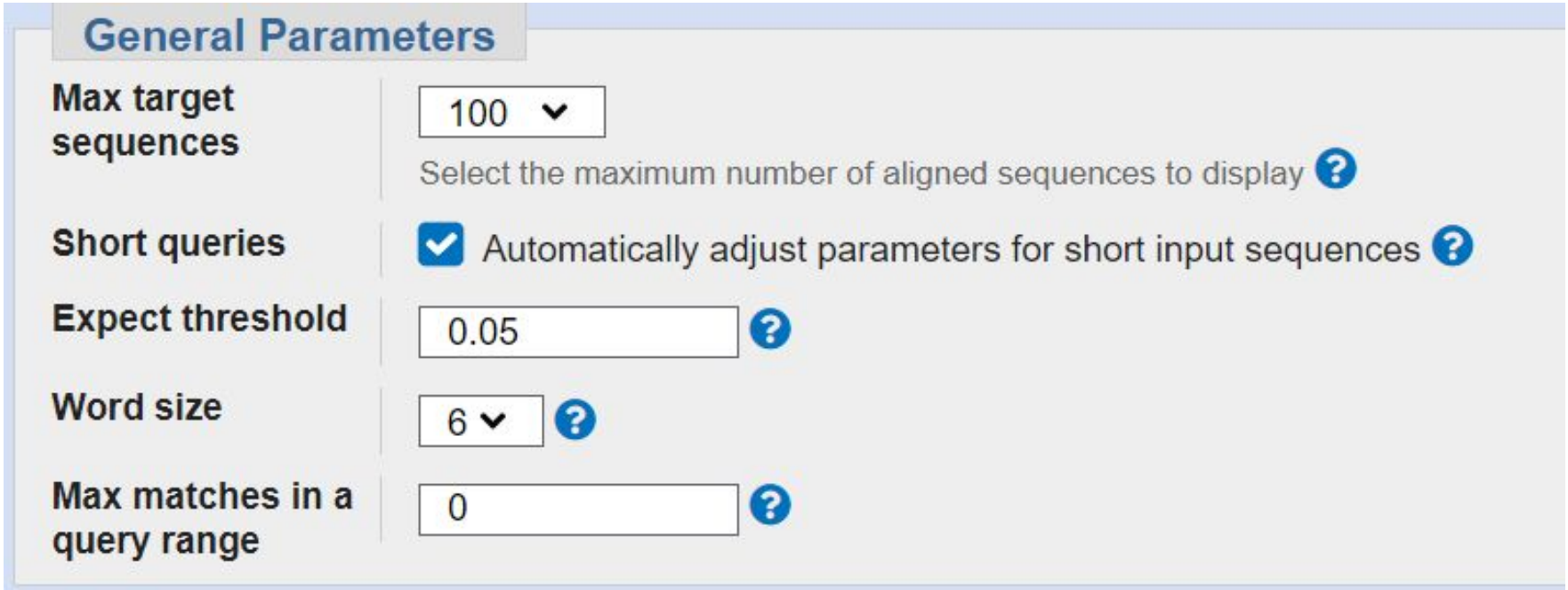
In the BLAST webserver, what do these parameters do?

– General Parameters

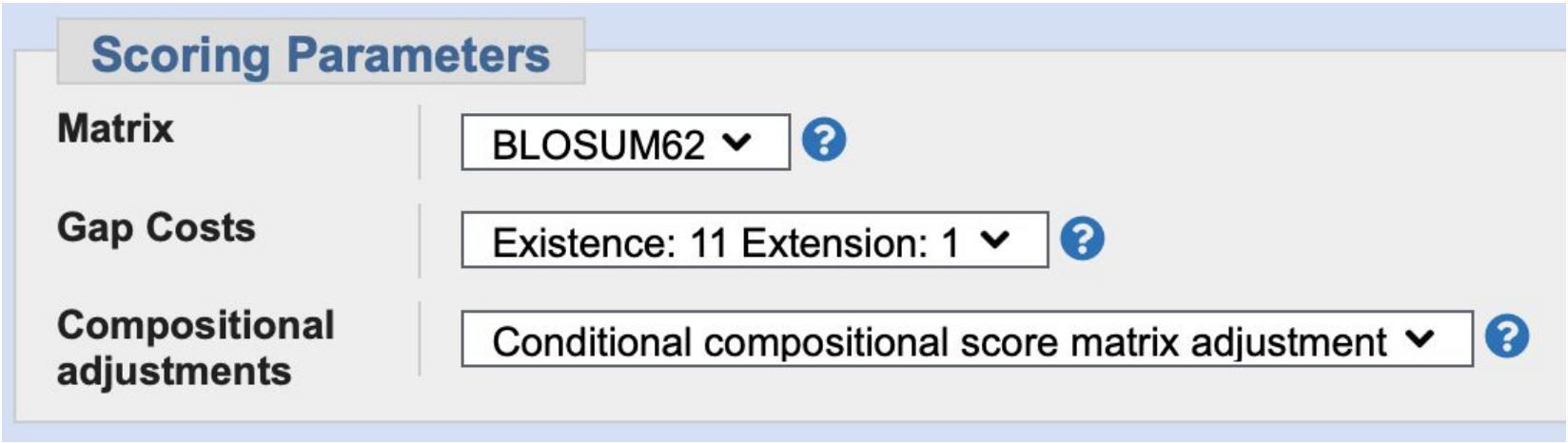
– Scoring Parameters

– Filtering & Masking

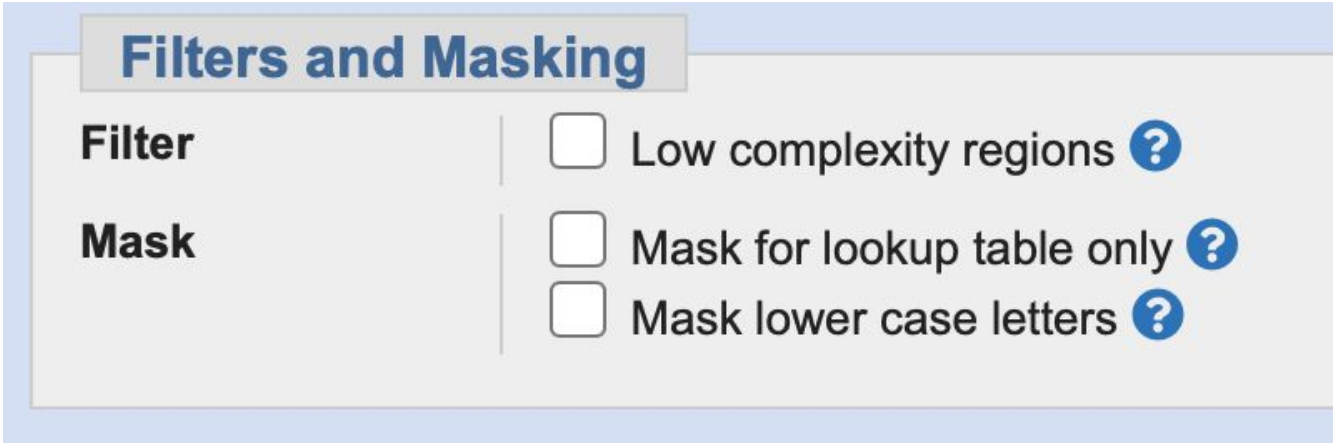
General Parameters



Scoring Parameters



Filtering & Masking



Group 4

Aaron, Isabel, Joanna, Josh, Mehrsa, and Tyus

What is *one* example of a contribution of BLAST in terms of:

– Approach

– Algorithmic techniques

– Computational ideas

Contribution of BLAST in:

Approach:

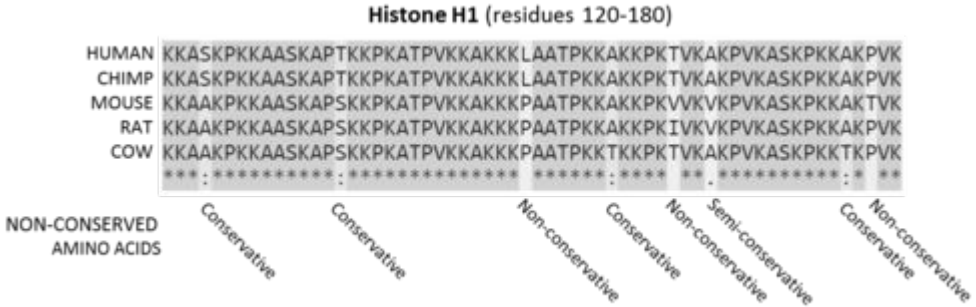
- Homology
- Ancestral sequences
- Identify pairs of similar segments within the sequence (nucleotide or amino acid) whose similarity is greater than a given threshold score

Algorithm technique(s):

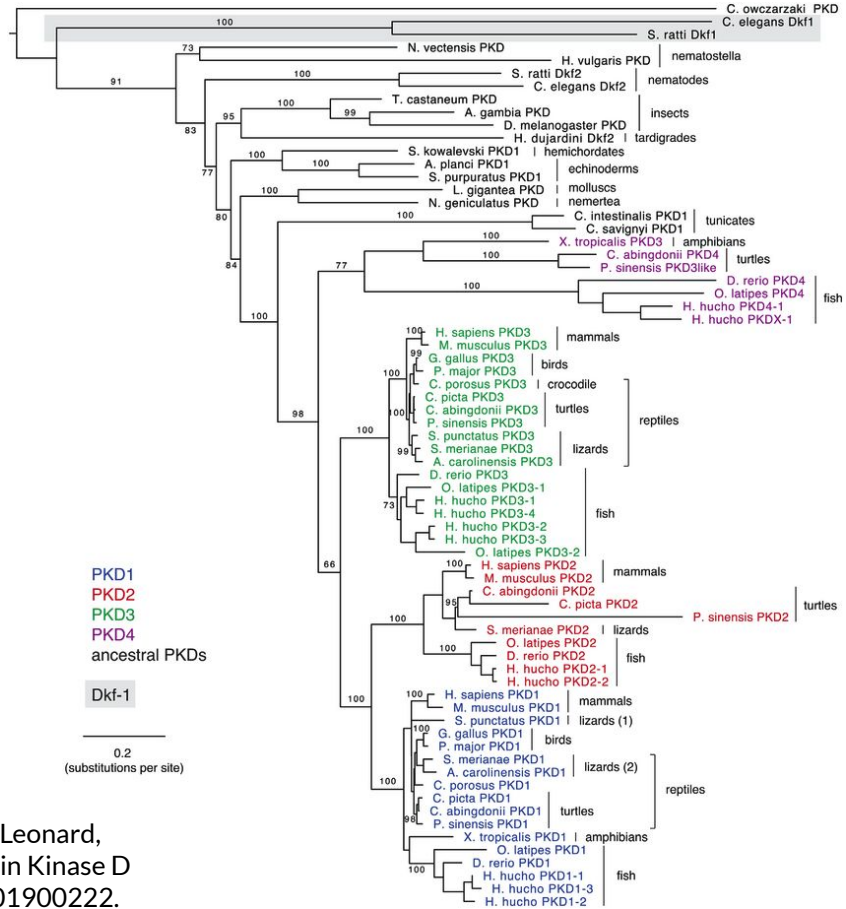
- Iteration
- Scoring matrix (default: BLOSUM62)
- Smith-Waterman alignment

Computational idea(s):

- Searching only the previous list that has passed the filtering to be faster/searching things it has already processed
- Pre-processing to remove common sequences

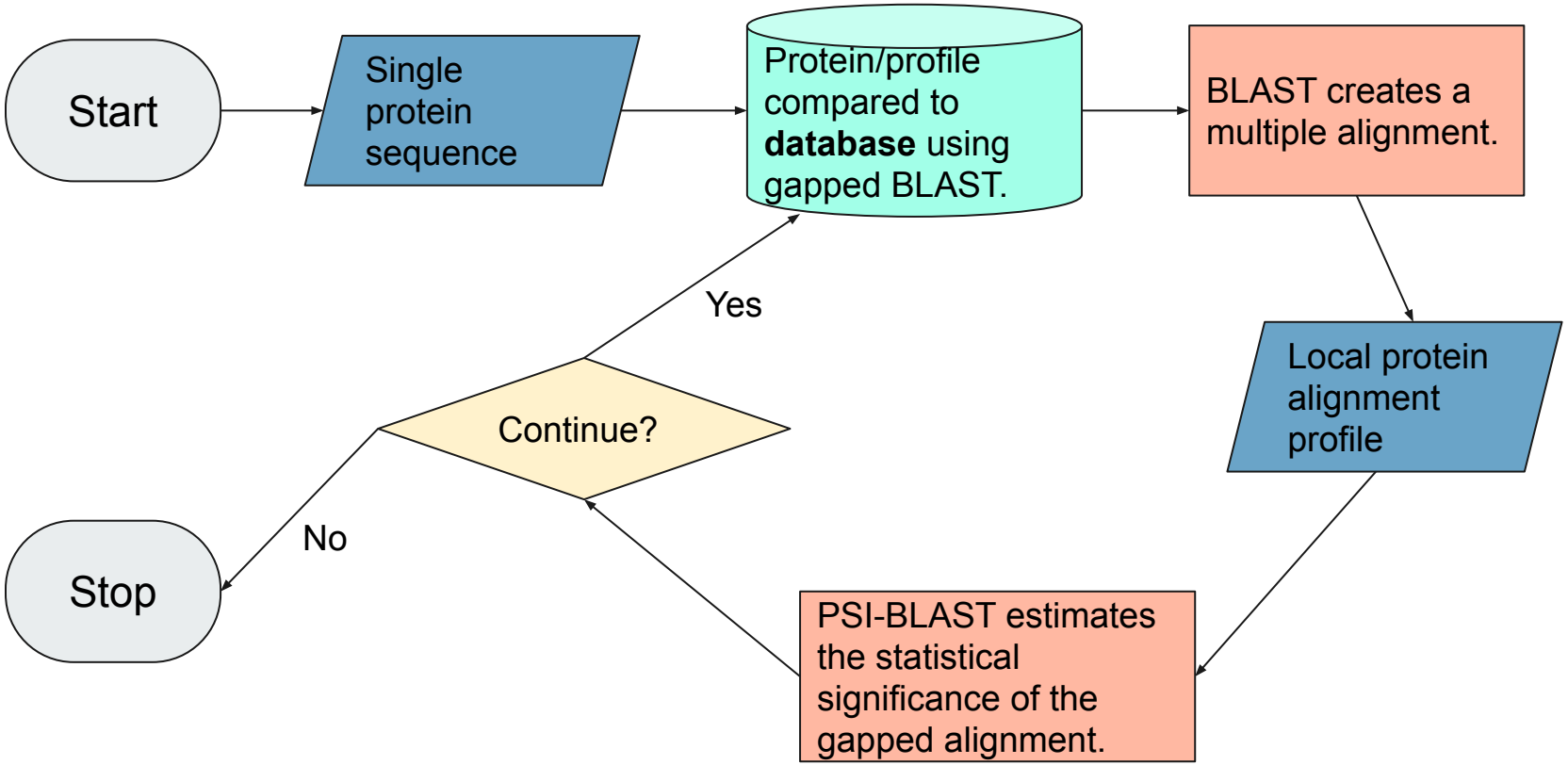


Example using BLAST - Protein Kinase D (PKD)



Reinhardt, Ronja & Truebestein, Linda & Schmidt, Heiko & Leonard, Thomas. (2020). It Takes Two to Tango: Activation of Protein Kinase D by Dimerization. BioEssays. 42. 1900222. 10.1002/bies.201900222.

Algorithmic Iteration using PSI-BLAST



Iterated profile searches with PSI-BLAST—a tool for discovery in protein databases
SF Altschul, EV Koonin - Trends in biochemical sciences, 1998

Computational Improvement in BLAST using BLAST+

- BLAST is an open source tool that allows researchers to create new modules
- Examples include iBLAST, BLAST+, megablast, and CPU- and GPU-accelerated BLAST
- BLAST+ improves BLASTX search speed for long queries by splitting the queries into smaller segments and processing them in parallel.

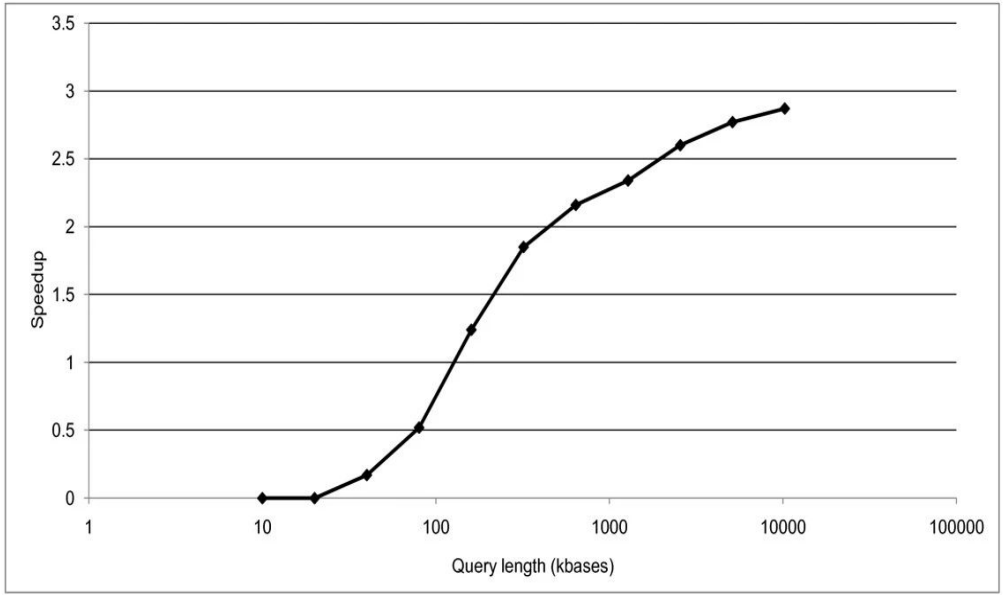


Figure 2. Speedup of BLASTX searches for differently sized queries with and without query splitting.

Camacho, C., Coulouris, G., Avagyan, V. *et al.* BLAST+: architecture and applications. *BMC Bioinformatics* 10, 421 (2009). https://doi.org/10.1186/1471-2105-10-421