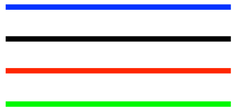
Alignment-based search (BLAST)

Mark Stenglein, MIP 280A4

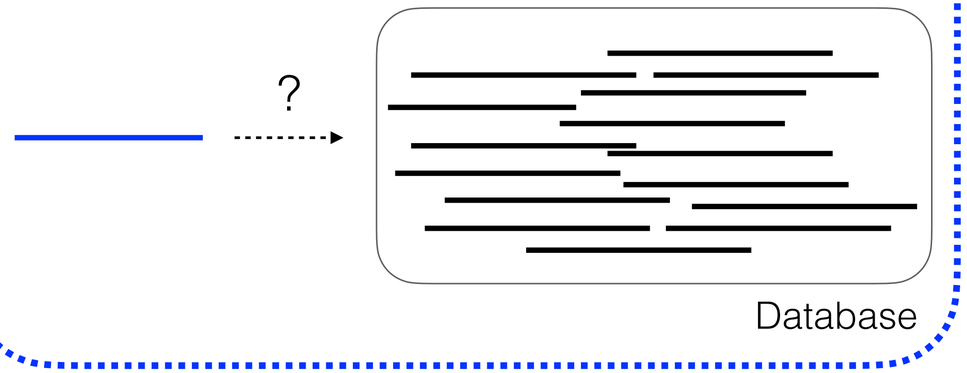
Today we will learn about alignment-based search (really: BLAST)

Pairwise alignment

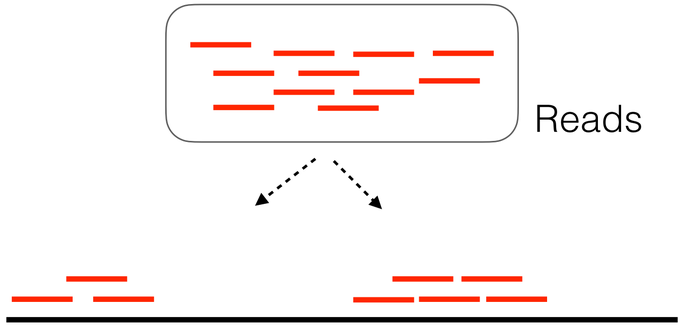
Multiple sequence alignment



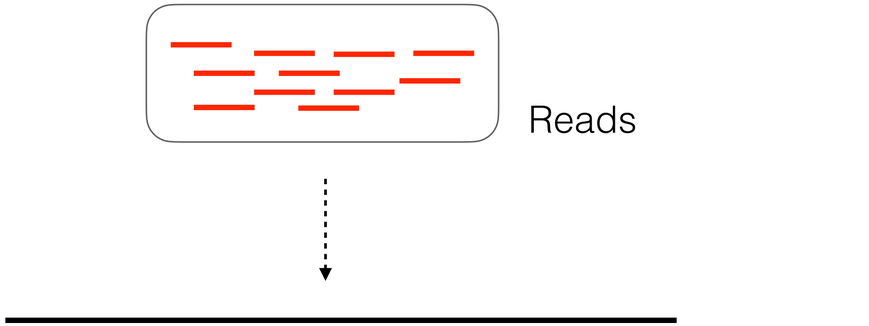
Alignment-based database searches



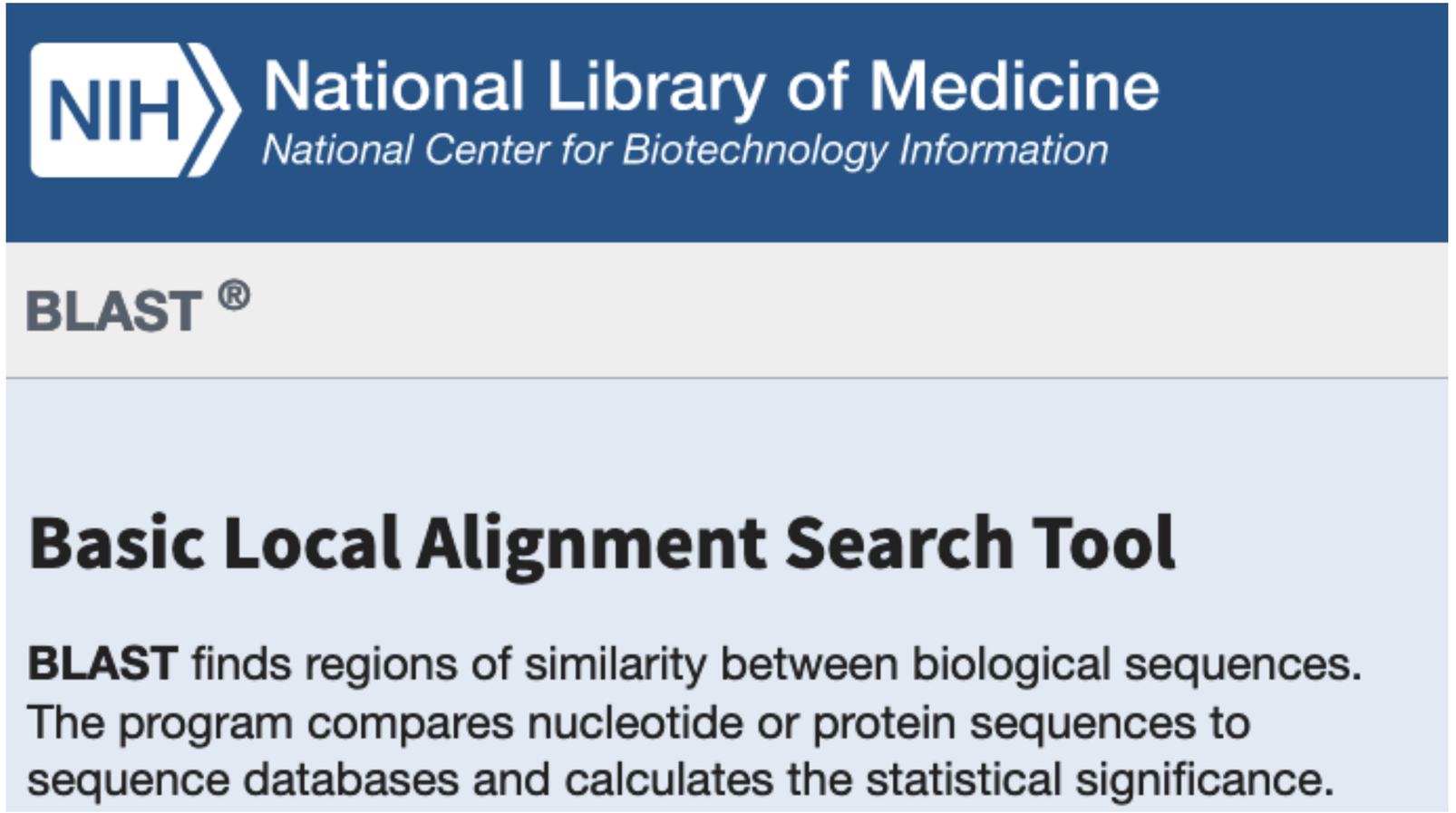
Alignment to reference (mapping)



Assembly



https://blast.ncbi.nlm.nih.gov/

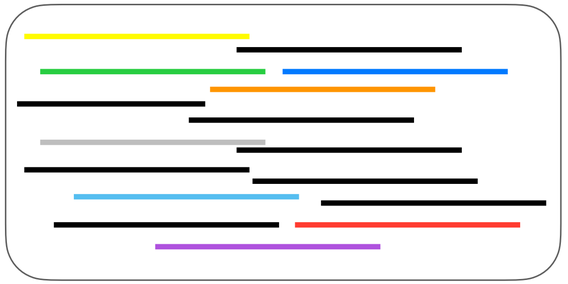


Anatomy of a BLAST search

One or more sequences (**queries**)

For each query: find the most closely related subect(s)

One or more sequences in the database (**subjects**)



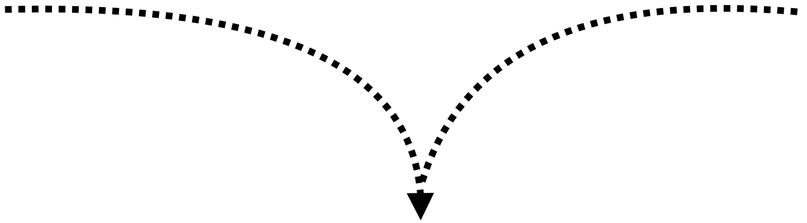
Database

Report best “hits” for each query

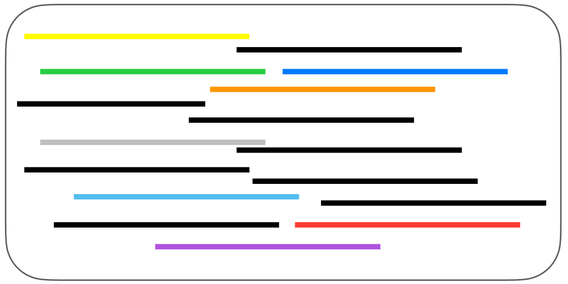
Anatomy of a BLAST search

queries can be nucleotide or protein sequences

For each query: find the most closely related subect(s)



subjects can be nucleotide or protein sequences

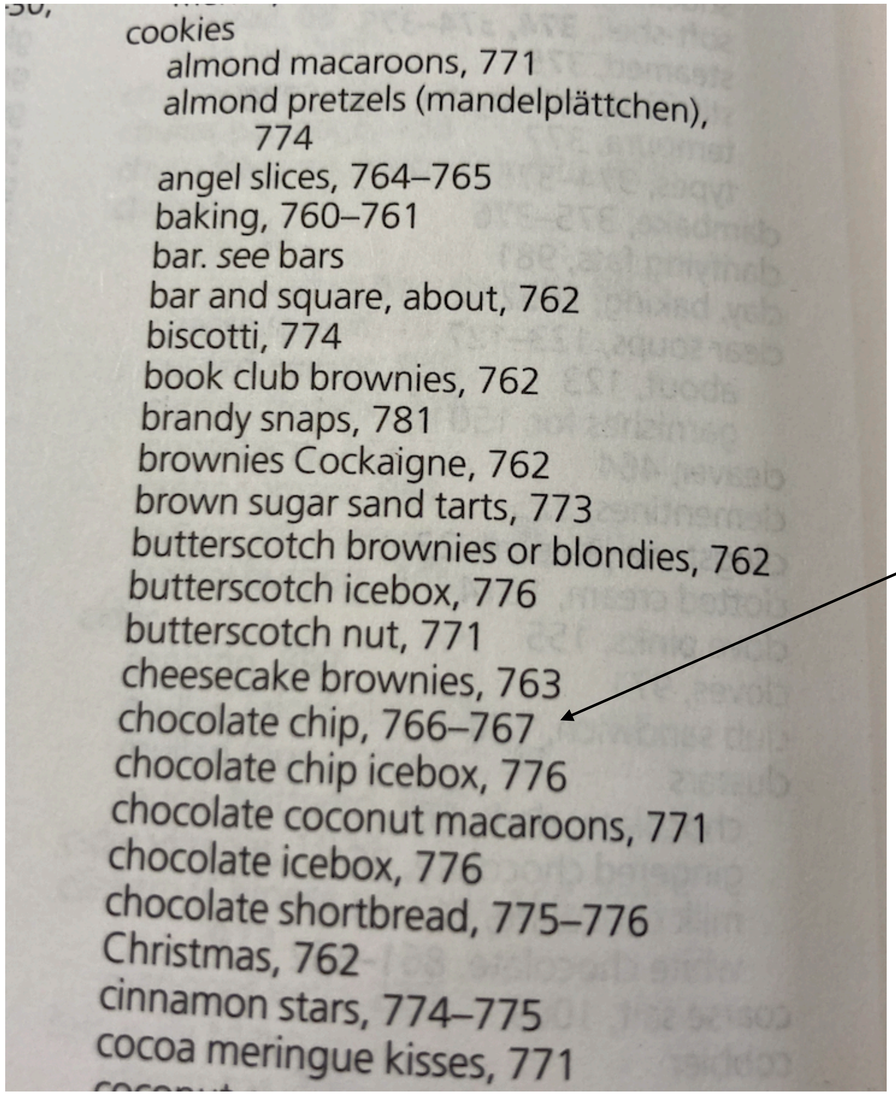


Database

Report best “hits” for each query

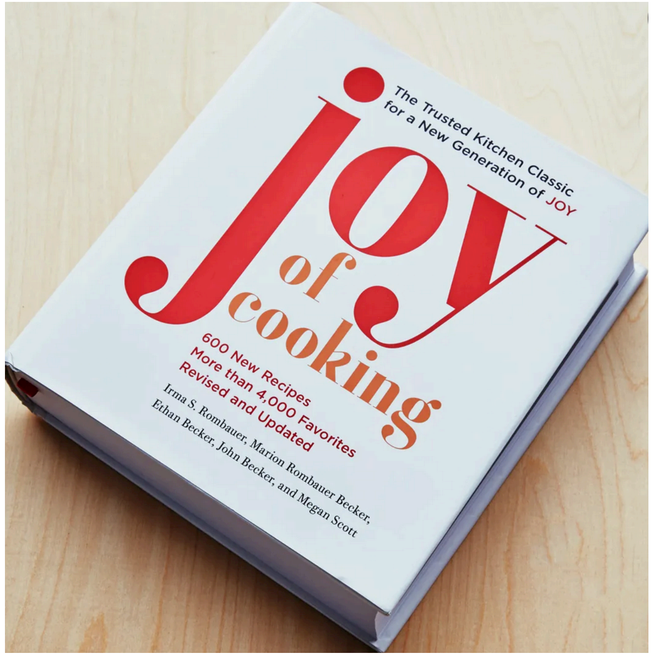
You can use preexisting databases from NCBI or you can make your own database

BLAST uses pre-built database indexes to speed up search



Indexes help you find things faster

chocolate chip cookies on page 766



Pre-indexing is a common strategy to speed-up alignment
We'll see it again when we learn about mapping

A simplified version of the BLAST workflow

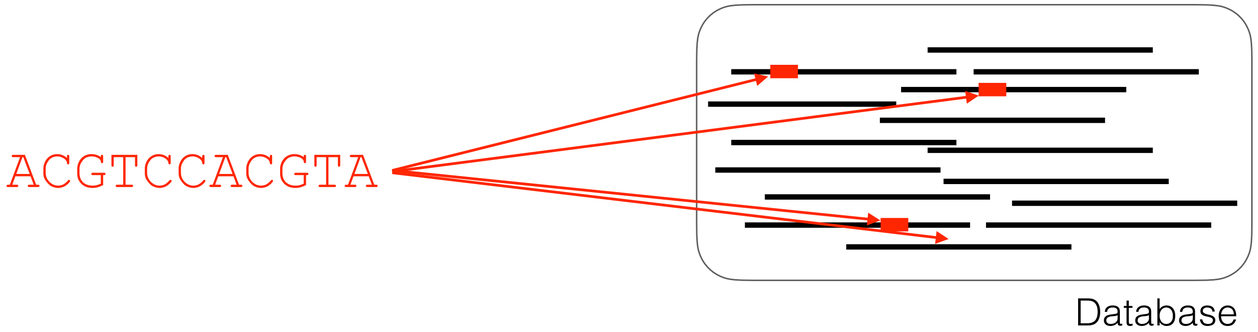
Break query sequence into “words” of a particular length

```
>a query sequence
ACGTCCACGTACGA

ACGTCCACGTA
 CGTCCACGTAC
  GTCCACGTACG
   TCCACGTACGA
```

words of length 11

The pre-built index tells you where in database sequences each word occurs



“Seed” alignments are extended, and those scoring sufficiently high are reported

A key statistic of a BLAST search result is the expect (E)-value

"The Expect Value is **the number of times that an alignment as good or better than that found by BLAST would be expected to occur by chance**, given the size of the database searched"

https://www.ncbi.nlm.nih.gov/books/NBK1734/

E-values are like p-values: Lower is better

E-value	Interpretation
10	Bad
1	Bad
0.1	Bad
0.01	So-so
$1x10^{-3}$	So-So
$1x10^{-10}$	Good
$1x10^{-100}$	Great
0	Best

Unlike N-W and S-W, BLAST is not guaranteed to find the highest scoring alignment
(But it almost always does and is good enough)

Heuristic (computer science)

From Wikipedia, the free encyclopedia

For other uses, see Heuristic (disambiguation).

In mathematical optimization and computer science, **heuristic** (from Greek εὑρίσκω "I find, discover") is a technique designed for solving a problem more quickly when classic methods are too slow or for finding an approximate solution when classic methods fail to find any exact solution. This is achieved by trading optimality, completeness, accuracy, or precision for speed. In a way, it can be considered a shortcut.

BLAST exercise: monkeypox virus DNA polymerase

Monkeypox

CDC > Poxvirus > Monkeypox

Monkeypox

About Monkeypox

Frequently Asked Questions

What CDC is Doing

2022 Outbreak Cases & Data

Signs & Symptoms

How It Spreads

Testing

Prevention

Vaccines

If You Are Sick

About Monkeypox

Updated July 22, 2022 Español Print

What is Monkeypox?

Monkeypox is a rare disease caused by infection with the monkeypox virus. Monkeypox virus is part of the same family of viruses as variola virus, the virus that causes smallpox. Monkeypox symptoms are similar to smallpox symptoms, but milder, and monkeypox is rarely fatal. Monkeypox is not related to chickenpox.

Monkeypox was discovered in 1958 when two outbreaks of a pox-like disease occurred in colonies of monkeys kept for research. Despite being named "monkeypox," the source of the disease remains unknown. However, African rodents and non-human primates (like monkeys) might harbor the virus and infect people.

The first human case of monkeypox was recorded in 1970. Prior to the 2022 outbreak, monkeypox had been reported in people in several central and western African countries. Previously, almost all monkeypox cases in people outside of Africa were linked to international travel to countries where the disease commonly occurs or through imported animals. These cases occurred on multiple continents.

Changing word size changes search sensitivity and speed

```
>a query sequence
ACGTCCACGTACGA
```

```
ACGTCCACGTA
 CGTCCACGTAC
  GTCCACGTACG
   TCCACGTACGA
```

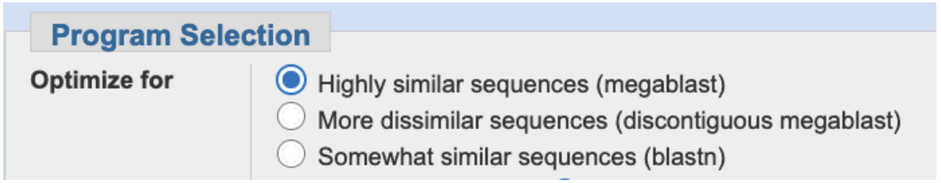
words of length 11

Longer words result in faster searches
But searches will only find more closely related matches.

```
ACGTCCA
 CGTCCAC
  GTCCACG
   TCCACGT
    etc…
```

words of length 7

Shorter words make search slower
But searches will find more dissimilar matches



BLAST variants

Program	Database	Query
BLASTN	Nucleotide	Nucleotide
BLASTP	Protein	Protein
BLASTX	Protein	Nucleotide translated into protein
TBLASTN	Nucleotide translated into protein	Protein
TBLASTX	Nucleotide translated into protein	Nucleotide translated into protein