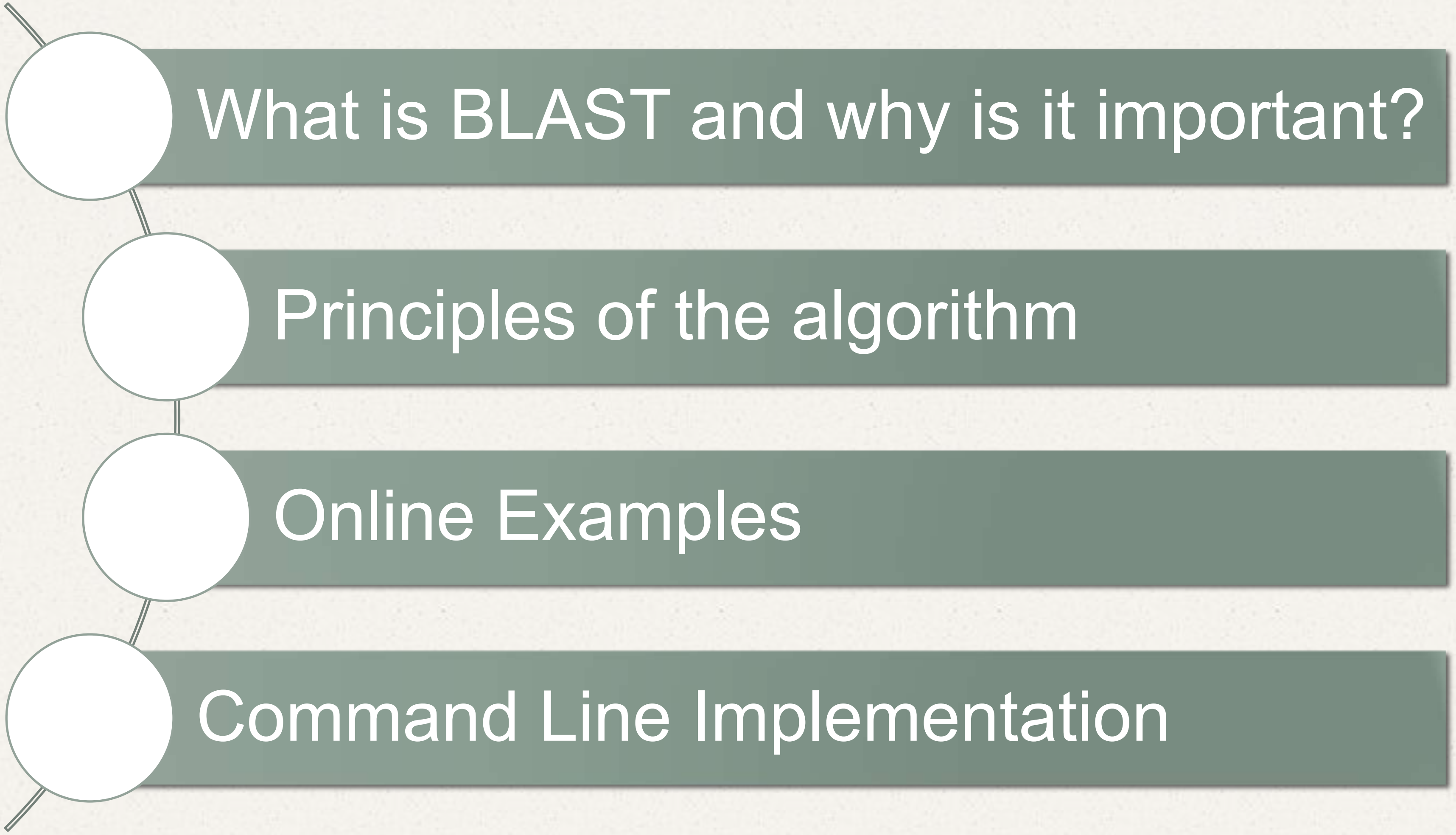
BLAST

Basic Local Alignment Search Tool

So useful – it is now a verb in the literature

Goals for Today:



What is BLAST and why is it important?

Principles of the algorithm

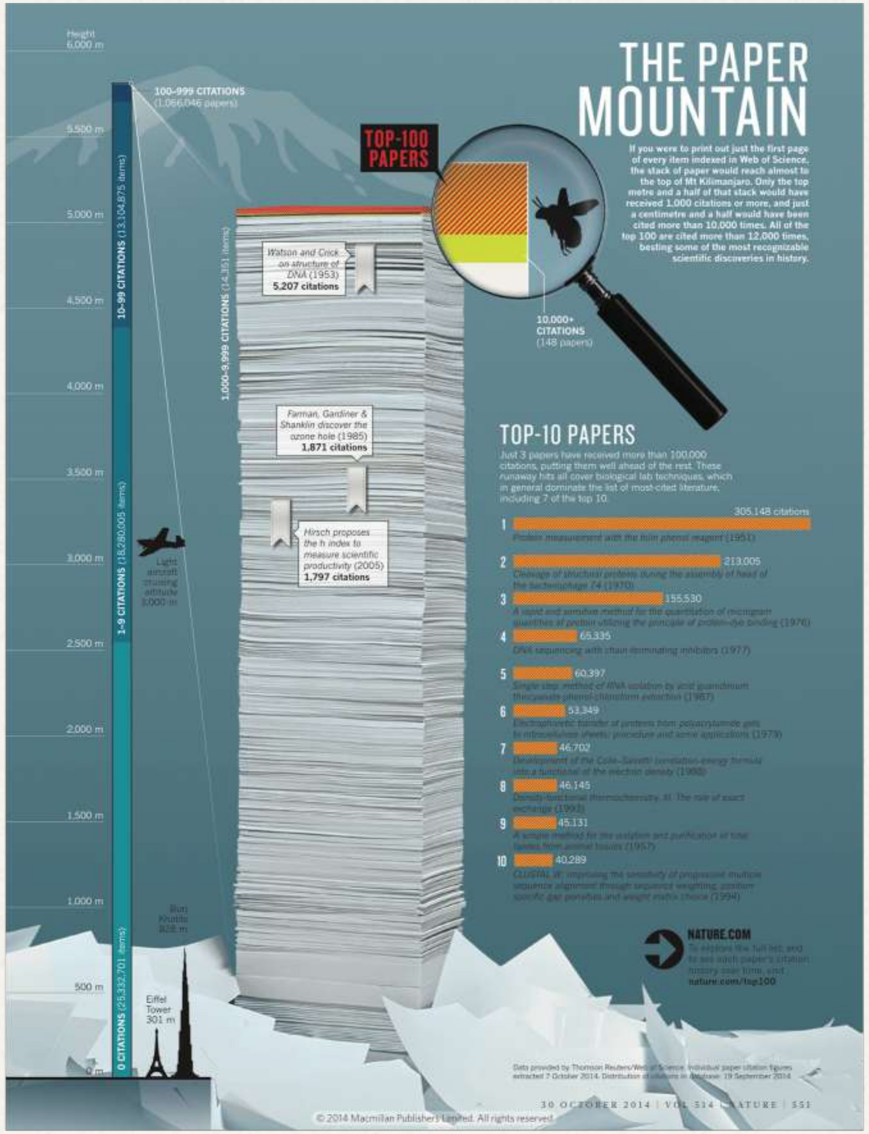
Online Examples

Command Line Implementation

A Lot of Blasting

- Where is BLAST on this list?
  - Altschul et al. 1990
    - #12 – 38,380 citations
      - 53,672 (Web of Science 6/1/2019)
  - Altschul et al. 1997
    - #14 – 36,410 citations
      - 48,001 (Web of Science 6/1/2019)
  - Combined: 4th!

Van Noorden et al. 2014, *Nature*



THE PAPER MOUNTAIN

If you were to print out just the first page of every item indexed in Web of Science, the stack of paper would reach almost to the top of Mt Kilimanjaro. Only the top metre and a half of that stack would have received 1,000 citations or more, and just a centimetre and a half would have been cited more than 10,000 times. All of the top 100 are cited more than 12,000 times, besting some of the most recognizable scientific discoveries in history.

TOP-10 PAPERS

Just 3 papers have received more than 100,000 citations, putting them well ahead of the rest. These runaway hits all cover biological lab techniques, which in general dominate the list of most-cited literature, including 7 of the top 10.

NATURE.COM To explore the full list and to see each paper's citation history over time, visit nature.com/top100

Data provided by Thomson Reuters/Web of Science. Individual paper citation figures extracted 7 October 2014. Distribution of citations in database: 19 September 2014.

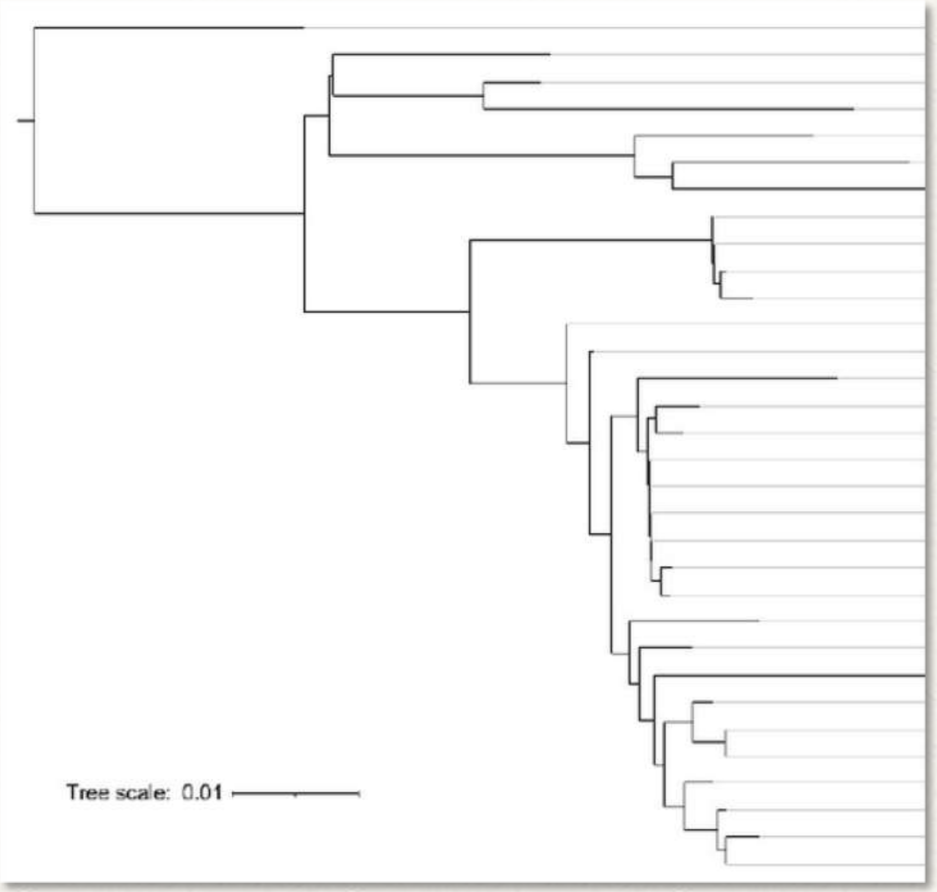
30 OCTOBER 2014 | VOL 514 | NATURE | 551

© 2014 Macmillan Publishers Limited. All rights reserved.

Height 6,000 m
5,500 m
5,000 m
4,500 m
4,000 m
3,500 m
3,000 m
2,500 m
2,000 m
1,500 m
1,000 m
500 m
0 m
100–999 CITATIONS (1,066,046 papers)
10–99 CITATIONS (13,104,875 items)
1,000–9,999 CITATIONS (14,351 items)
1–9 CITATIONS (18,280,005 items)
0 CITATIONS (25,332,701 items)
TOP-100 PAPERS
10,000+ CITATIONS (148 papers)
Watson and Crick on structure of DNA (1953) 5,207 citations
Farman, Gardiner & Shanklin discover the ozone hole (1985) 1,871 citations
Hirsch proposes the h index to measure scientific productivity (2005) 1,797 citations
Light aircraft cruising altitude 3,000 m
Burj Khalifa 828 m
Eiffel Tower 301 m
1 305,148 citations
Protein measurement with the folin phenol reagent (1951)
2 213,005
Cleavage of structural proteins during the assembly of head of the bacteriophage T4 (1970)
3 155,530
A rapid and sensitive method for the quantitation of microgram quantities of protein utilizing the principle of protein–dye binding (1976)
4 65,335
DNA sequencing with chain-terminating inhibitors (1977)
5 60,397
Single-step method of RNA isolation by acid guanidinium thiocyanate-phenol-chloroform extraction (1987)
6 53,349
Electrophoretic transfer of proteins from polyacrylamide gels to nitrocellulose sheets: procedure and some applications (1979)
7 46,702
Development of the Colle–Salvetti correlation-energy formula into a functional of the electron density (1988)
8 46,145
Density-functional thermochemistry. III. The role of exact exchange (1993)
9 45,131
A simple method for the isolation and purification of total lipides from animal tissues (1957)
10 40,289
CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice (1994)

Goals

- Search a query against a database
  - Identify species
  - Locate domains
  - Assess function
  - Establish phylogeny
  - Mapping

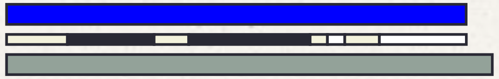


What is BLAST?

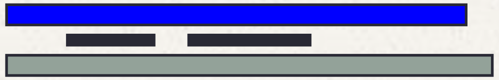
- Sequence searching algorithm
- Finds the best local alignments
- Calculates statistical significance
- Similarity suggests homology
- Less sensitive than Smith-Waterman, but FASTER!

Global vs Local Alignment

- Global alignment: entire sequences



- Local alignment: segments of sequences



- Local alignment often the most relevant
  - Depends on biological assumptions

Blast Types

Name	Query	Database
blastn	nucleotide	nucleotide
blastp	protein	protein
blastx	nucleotide	protein
tblastx	nucleotide	nucleotide
tblastn	protein	nucleotide
PSI-blast	protein	protein

Blast Databases: Protein

Name	Host	Description
nr	NCBI	Non-redundant, general
Refseq_protein	NCBI	Annotated and curated protein collection
SwissProt	SIB	Manually curated and reviewed proteins form UniProt
Trembl	EBI	Automatically annotated, non-reviewed proteins
PDB	Rutgers/UCSD/UCSC	Proteins with 3D structural information

Blast Databases: Nucleotide

Name	Host	Description
nt	NCBI	Non-redundant, general
Refseq_RNA	NCBI	Annotated and curated RNA sequence collection
Refseq_Genomics	NCBI	Sequenced and curated genomes
EST	NCBI	Expressed sequence tags
UNIVEC	NCBI	Vector contaminant database
WGS	NCBI	Draft, whole genome shotgun sequence assemblies
SRA	NCBI	Raw NGS datasets
Many more databases, e.g. barcoding, viral, tRNA, etc, custom-built databases		

How it works: Making words

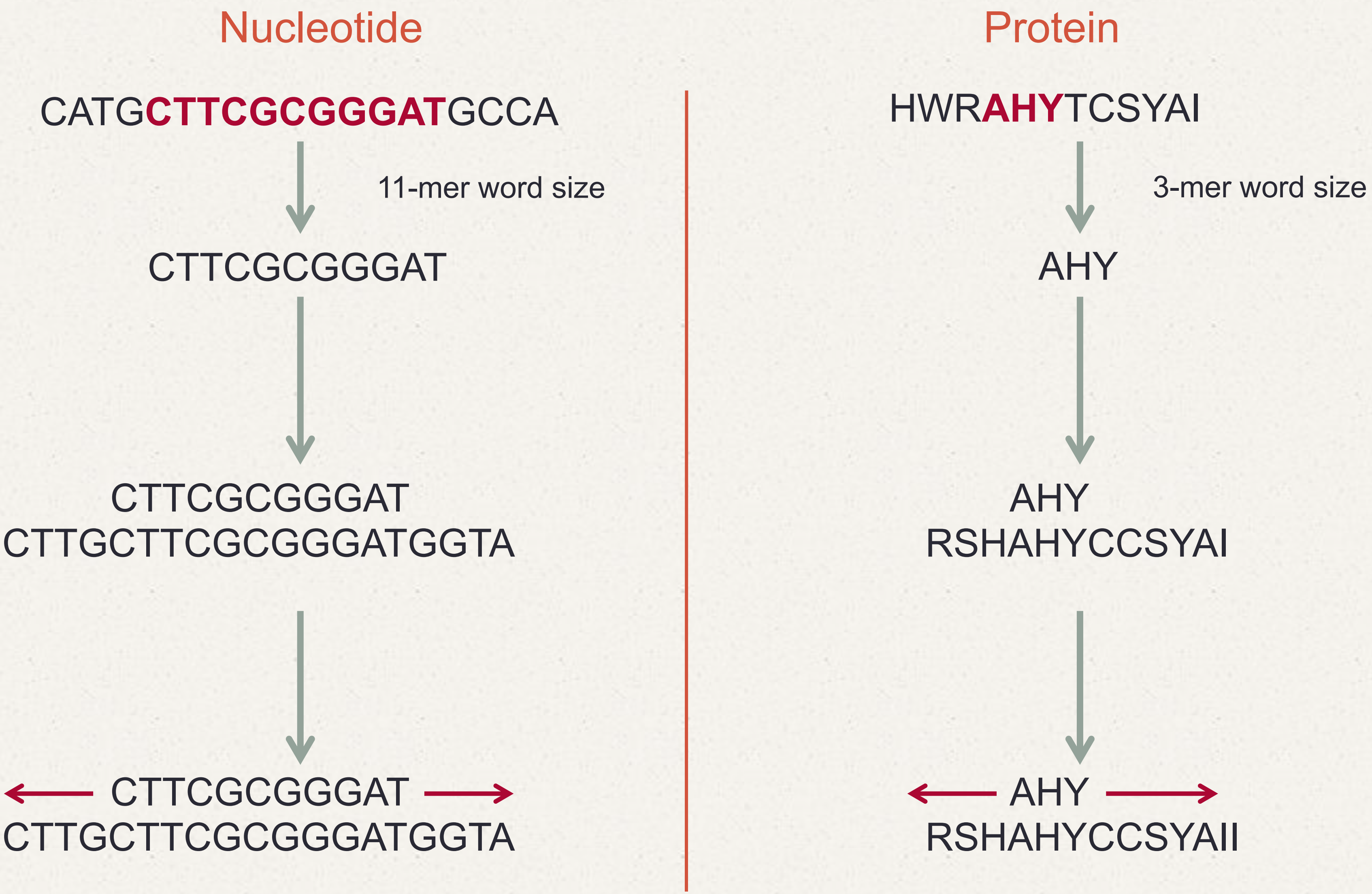
Nucleotide

- 11-letter words (seeds)
  - ACTACGTGCTATGC
    - ACTACGTGCTA
    - CTACGTGCTAT
    - TACGTGCTATG
    - ACGTGCTATGC

Protein

- 3-letter words (seeds)
  - PQGDEF
    - PQG
    - QGD
    - GDE
    - DEF

How it works



Nucleotide

CATG**CTTCGCGGGAT**GCCA

11-mer word size

CTTCGCGGGAT

CTTCGCGGGAT
CTTGCTTCGCGGGATGGTA

CTTCGCGGGAT
CTTGCTTCGCGGGATGGTA

Protein

HWR**AHY**TCSYAI

3-mer word size

AHY

AHY
RSHAHYCCSYAI

AHY
RSHAHYCCSYAII

Blast Scoring and E-values

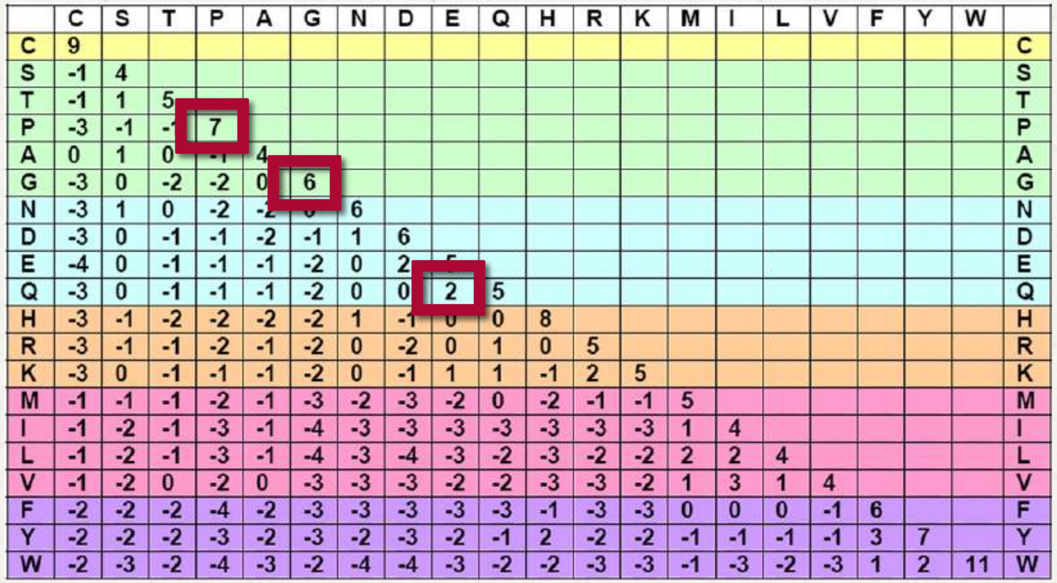
- Nucleotide sequences search for 11-letter matches
  - (4^11 = 4,194,304 combinations)
  - Match = +5, mismatch = -4
  - Only scores above a threshold (T) are kept

ACTACGTGCTA
ACTACGTGCTA
5+5+5+5+5+5+5+5+5+5+5 = 55

ACTACGTGCTA
ACAAGATGGTA
5+5-4+5-4-4+5+5-4+5+5 = 19

Blast Scoring and E-values

- Proteins use a BLOSUM62 scoring matrix
  - 20x20x20 = 8,000 possible 3-letter words
  - All possible amino acid pairs are given a score
  - All combinations above a threshold (T) are kept
    - Minimizes search space



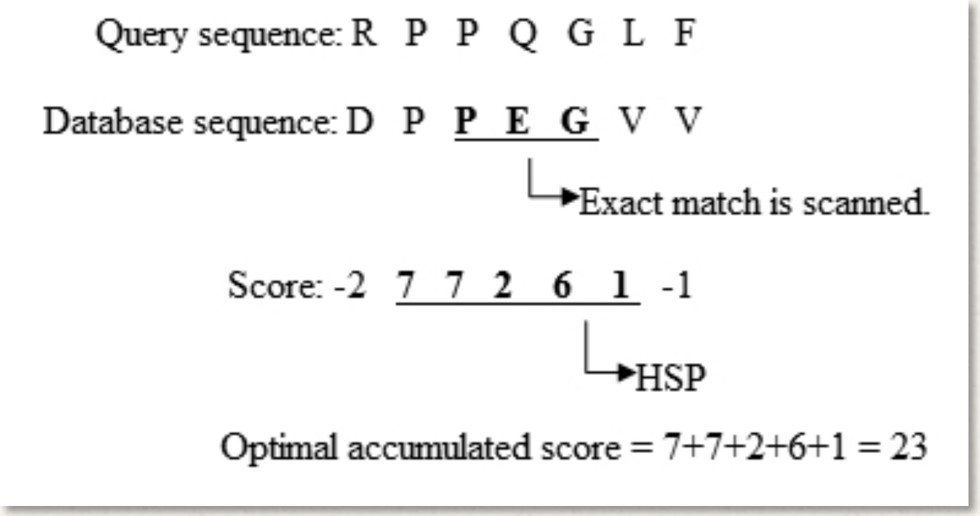
	C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W	
C	9																				C
S	-1	4																			S
T	-1	1	5																		T
P	-3	-1	-1	7																	P
A	0	1	0	-1	4																A
G	-3	0	-2	-2	0	6															G
N	-3	1	0	-2	-2	0	6														N
D	-3	0	-1	-1	-2	-1	1	6													D
E	-4	0	-1	-1	-1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

P Q G
P E G
7+2+6 = 15

P Q G
E Q R
-1+5+-2 = 2

Extending Matches

- Match = HSP (High-scoring Sequence Pair)
  - Match is found and extended as long as score stays above a threshold value
    - After finished extending, the HSP is kept if above the cutoff score (S)



Query sequence: R P P Q G L F

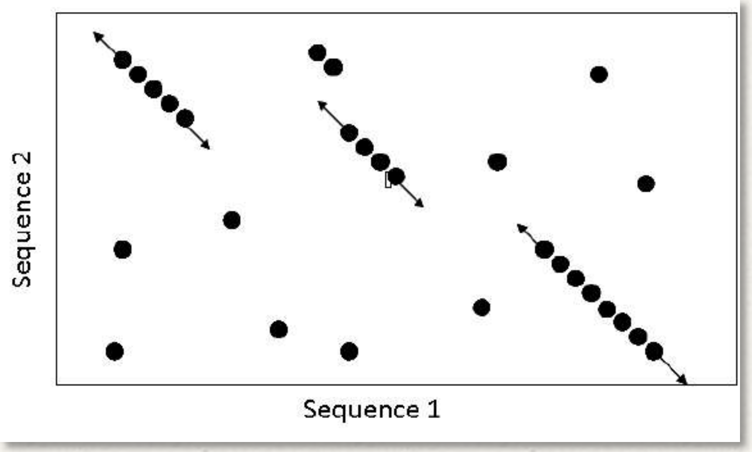
Database sequence: D P **P E G** V V

Exact match is scanned.

Score: -2 $\underline{7\ \ 7\ \ 2\ \ 6\ \ 1}$ -1

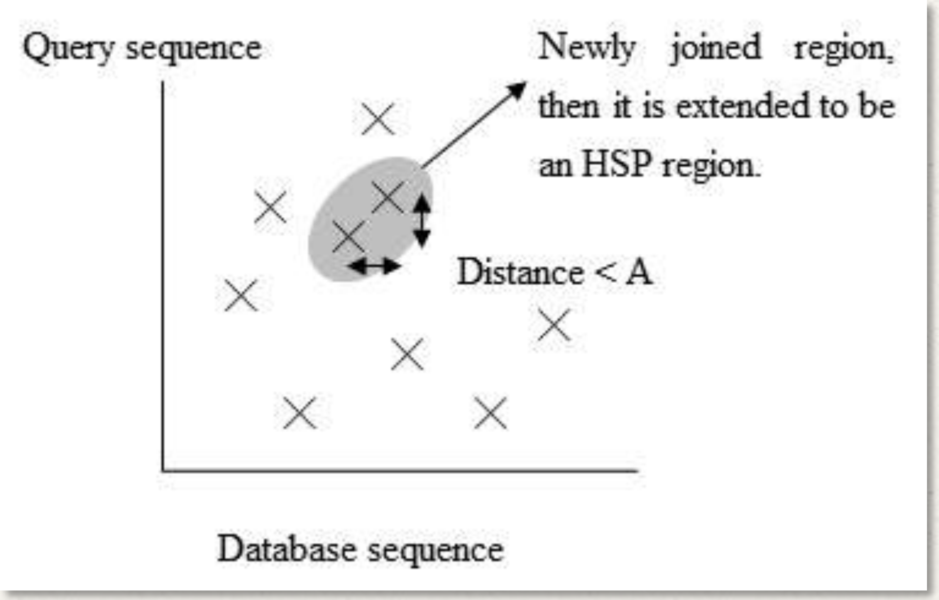
HSP

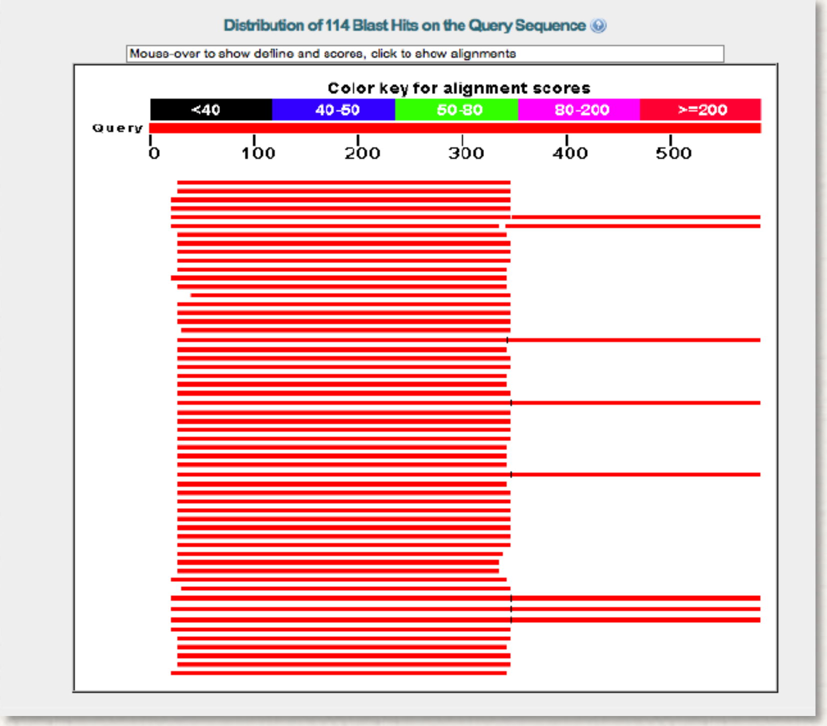
Optimal accumulated score = 7+7+2+6+1 = 23



Assembling HSPs

- HSPs, after extension, are assembled into a longer alignment


Query sequence
Newly joined region, then it is extended to be an HSP region.
Distance < A
Database sequence


Distribution of 114 Blast Hits on the Query Sequence

Mouse-over to show defline and scores, click to show alignments

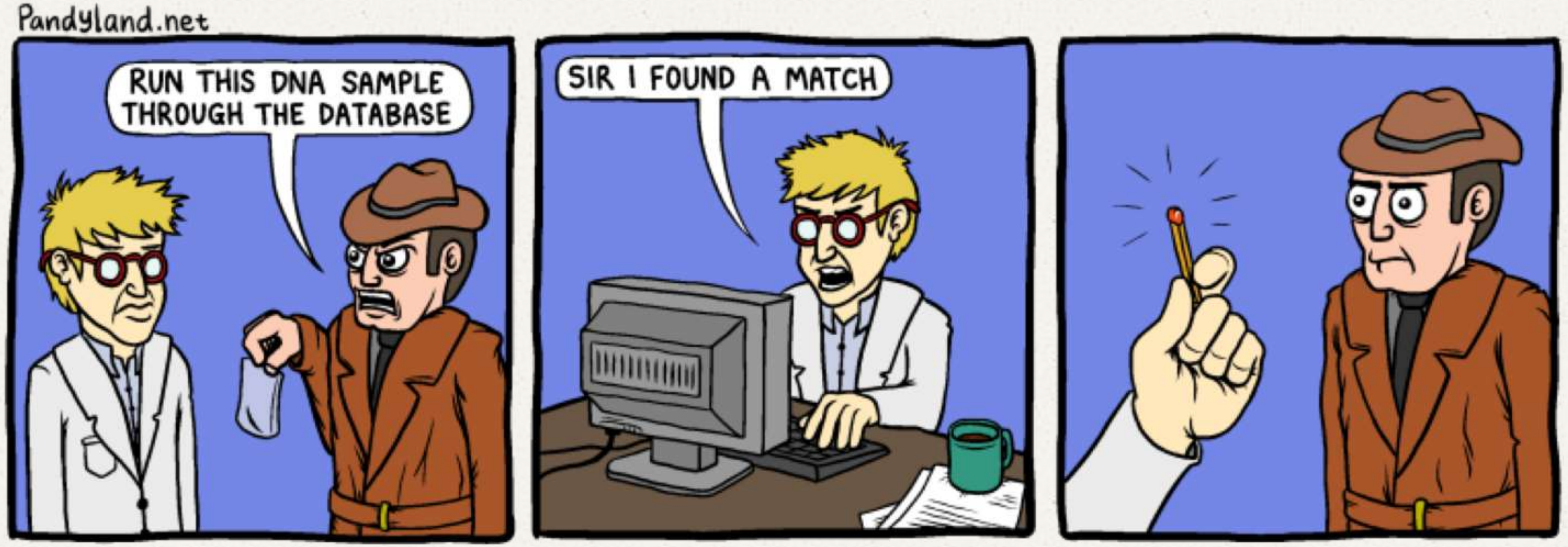
Output

- Max/Total Score: quality of the alignment
  - Higher the score the better the match
- Query Coverage: what proportion of the query the particular HSP covers
- E-value: probability that a match ≥ Max Score occurs by random chance (based on database size)
- Max Identity: For that HSP, the % of bases that match

Accession	Total Score	Query Coverage	E-value	Max Ident
X56286.1	579	54%	7e-162	99%
AF091629.1	573	54%	3e-160	99%
L48348.1	481	55%	2e-132	93%

Interpretation

- The matches you get are only acceptable matches, not necessarily the optimal match
- Your search is only as good as your database
  - If the optimal match is not in the database, you will not find it.
  - If you have sequences not in the database, SUBMIT THEM!



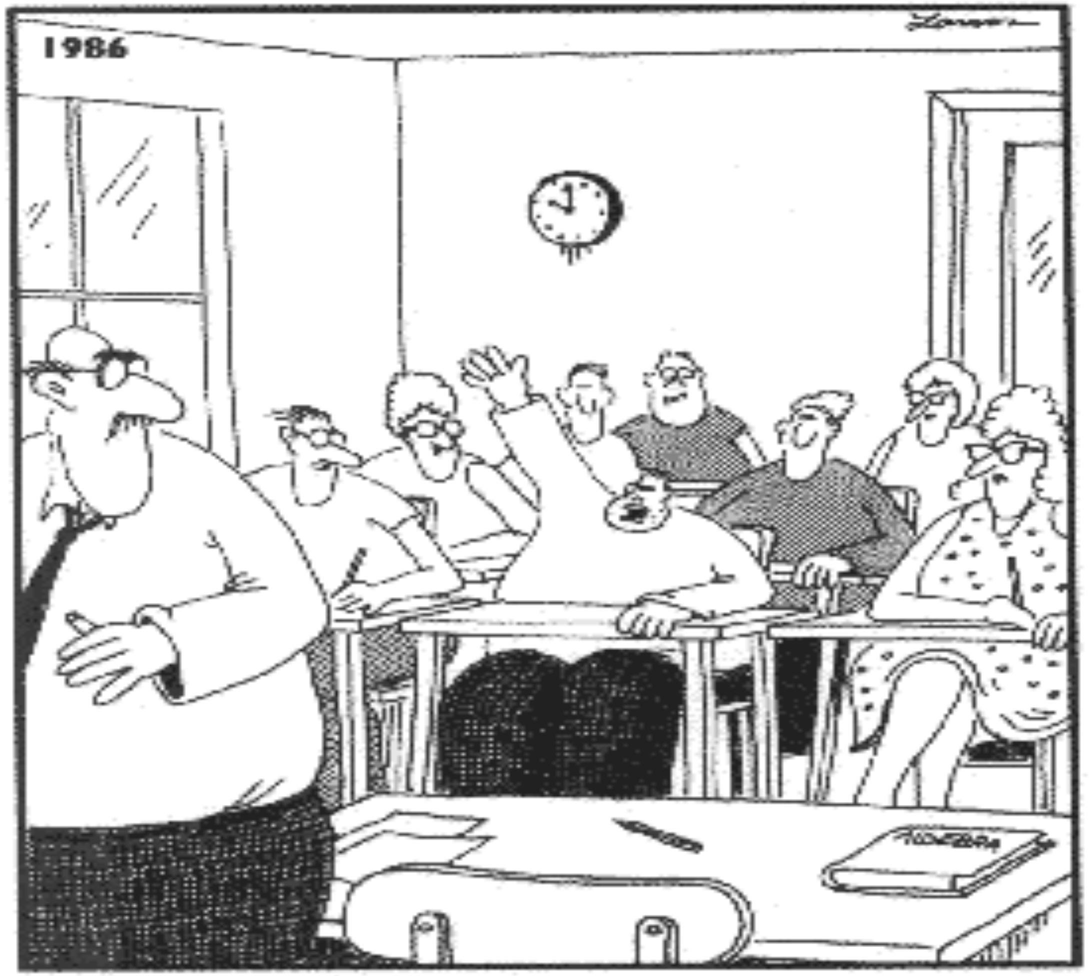
Take Home Points

- Blast is a powerful tool for database searching
- Very fast, but at the expense of sensitivity
- Flexible (types, databases)
- Interpret results carefully
- Help make it grow!

Several examples

- Example 1: SRA Blast (https://www.ncbi.nlm.nih.gov/sra)
  - Query: M55627.1
    - *Coccidiodes immitis* (Valley fever fungus) ssuRNA
  - Project: SRX633288
    - Puma 454 transcriptome reads

- Example 2: Blast an assembly (https://blast.ncbi.nlm.nih.gov/)
  - Query:
    - TruSeq Universal Adapter
    - AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCT
  - Database: nt
  - Organism: *Cyprinus carpio (taxid:7962)*



"Mr. Osborne, may I be excused? My brain is full."