BLAST Paper Presentation
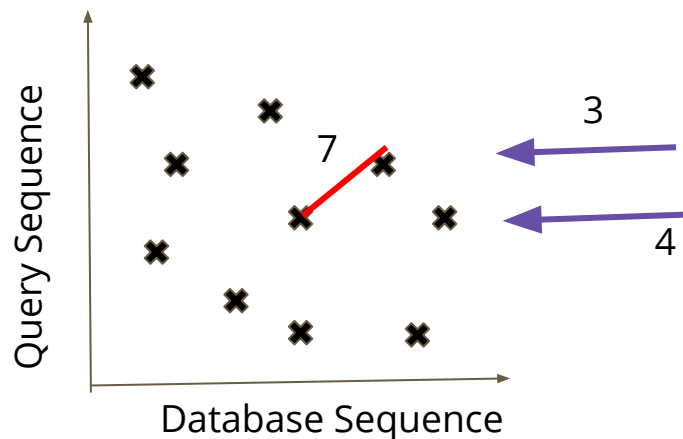
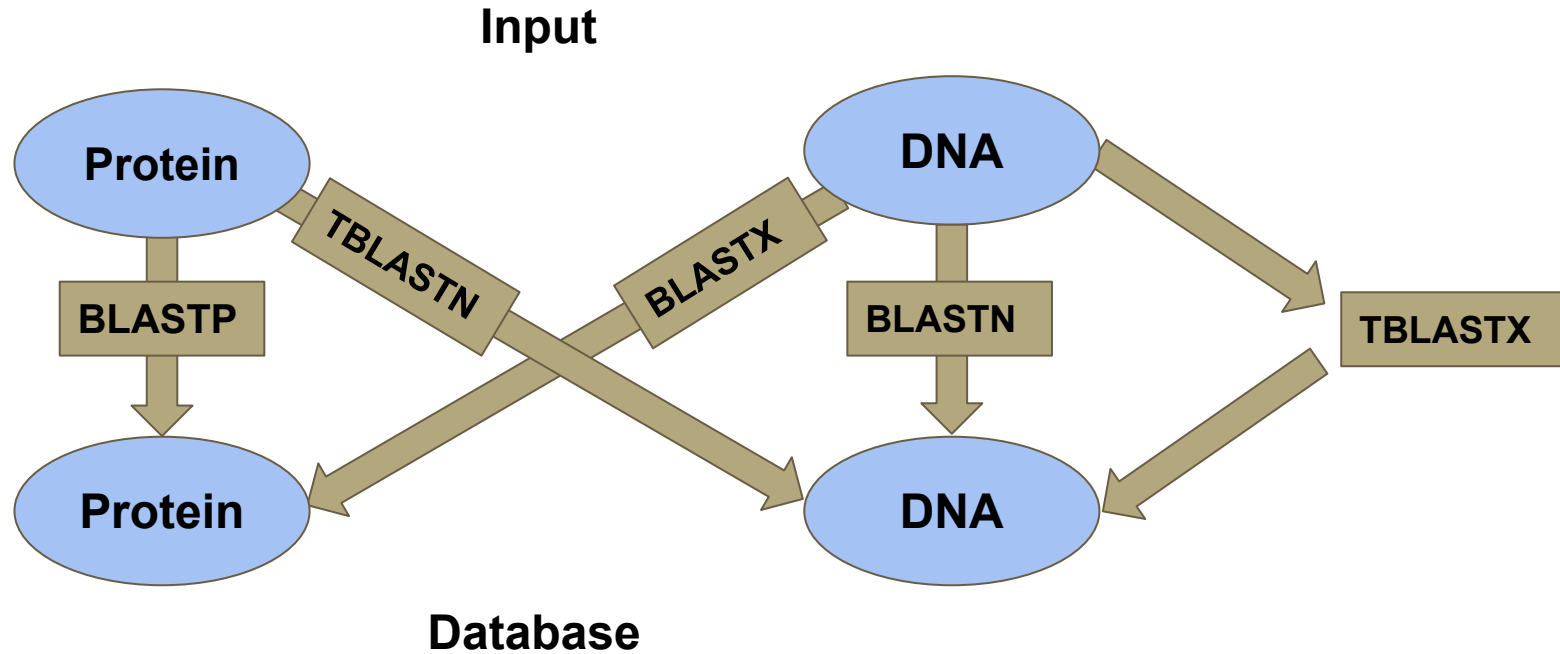Ali Cramer, Larissa Ford, Katherine Perry

Basic Local Alignment Search Tool

Works similarly to Smith-Waterman, but works much faster. Trade off for some accuracy.

Approximate local alignments

BLOSUM (BLOcks SUbstitution Matrix)[1]

Substitution matrix used for protein sequence alignment

Scores based on frequency of substitutions in multiple alignments

BLOSUM62=default for protein BLAST; 62=62% shared identity

- higher/lower percentages allow for search within more distant/closer evolutionary relatives

- Allows us to derive T value from common scoring system

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

[1]

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.

Steps to Implement BLAST [4]

1. Filter sequence for useless segments for alignment
2. Make a list of words for query

1	2	3	4	5	6	7
a	b	c	d	e	f	g
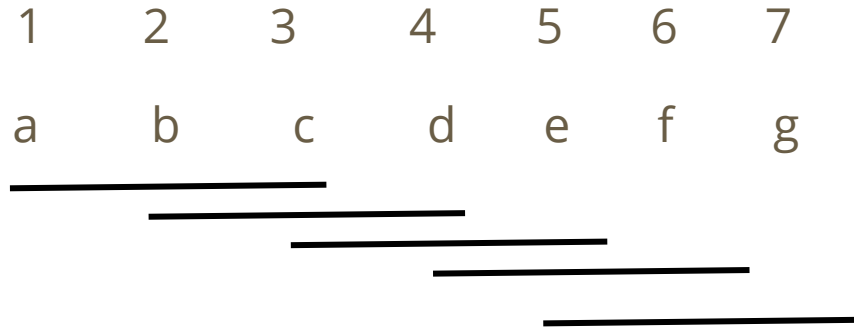
List of words: abc, bcd, cde, def, efg

Steps to Implement BLAST (continued)

3. Use BLOSUM62 to evaluate potential matches; calculate log-odds score

our words: a b c   b c d   c d e

database: a b a   b c d   a a e

   T = 9   10   12   6

4. Eliminate words that scored below the selected neighborhood word score threshold (T)

Steps to Implement BLAST (continued)

5. Organize words that made the cut into an organized search tree

our word: a b c

database: a a c   a c c   ⟨ a b c ⟩

6. Scan database sequences for exact matches to each word set

Steps to Implement BLAST (continued)[2]

7. Match shorter regions and lower-scoring words by lowering T (BLAST2)

Isolated words

Word clusters

T = 12

T = 14

T = 16
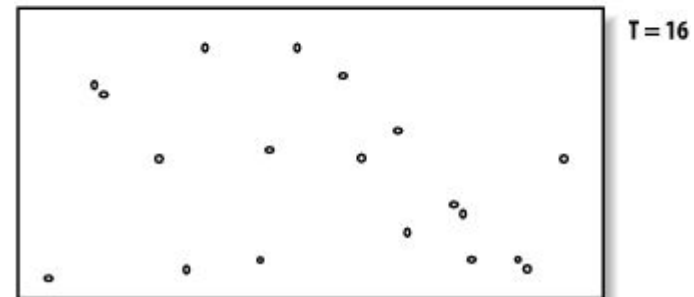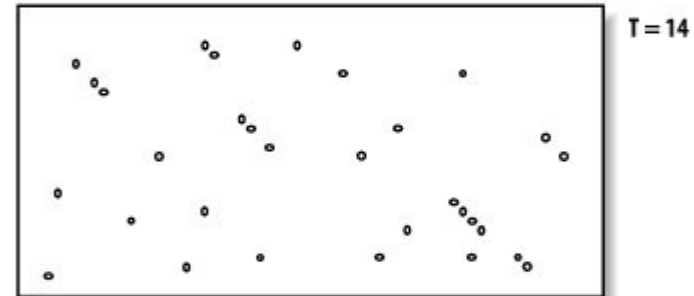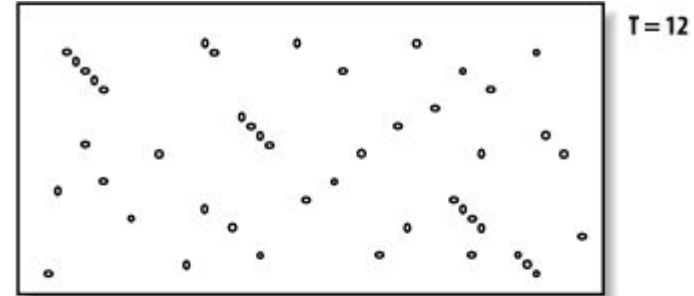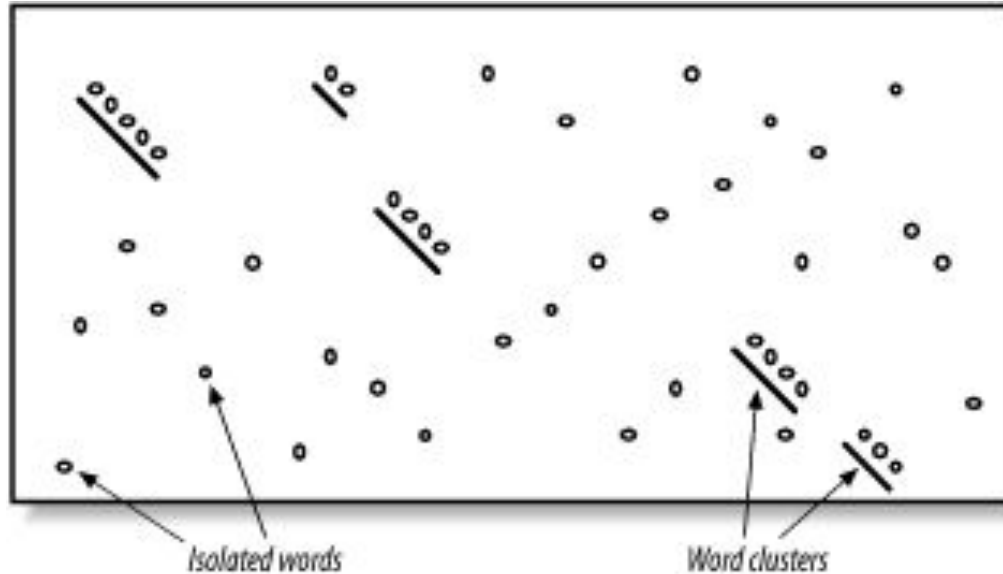
Steps to Implement BLAST (continued)

8. Extend matched segments in both directions until score falls

database: a c a ⟦a b c⟧ c b c

query: c b a ⟦a b c⟧ c b a

total score: 8 10 16   10   14 8 0

Steps to Implement BLAST (continued)

9. Determine whether HSP score with extension is greater than cutoff score S

10. Determine statistical significance of HSP score to decide local alignment

11. Produce a single alignment (with gaps) of all initial HSPs
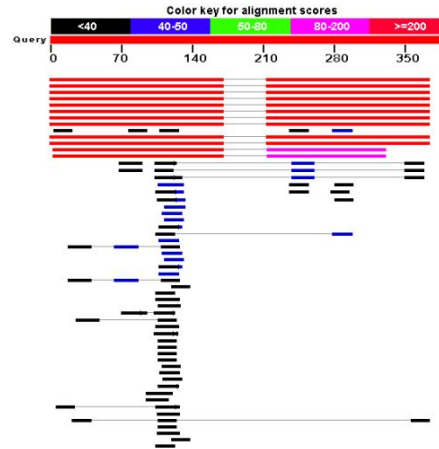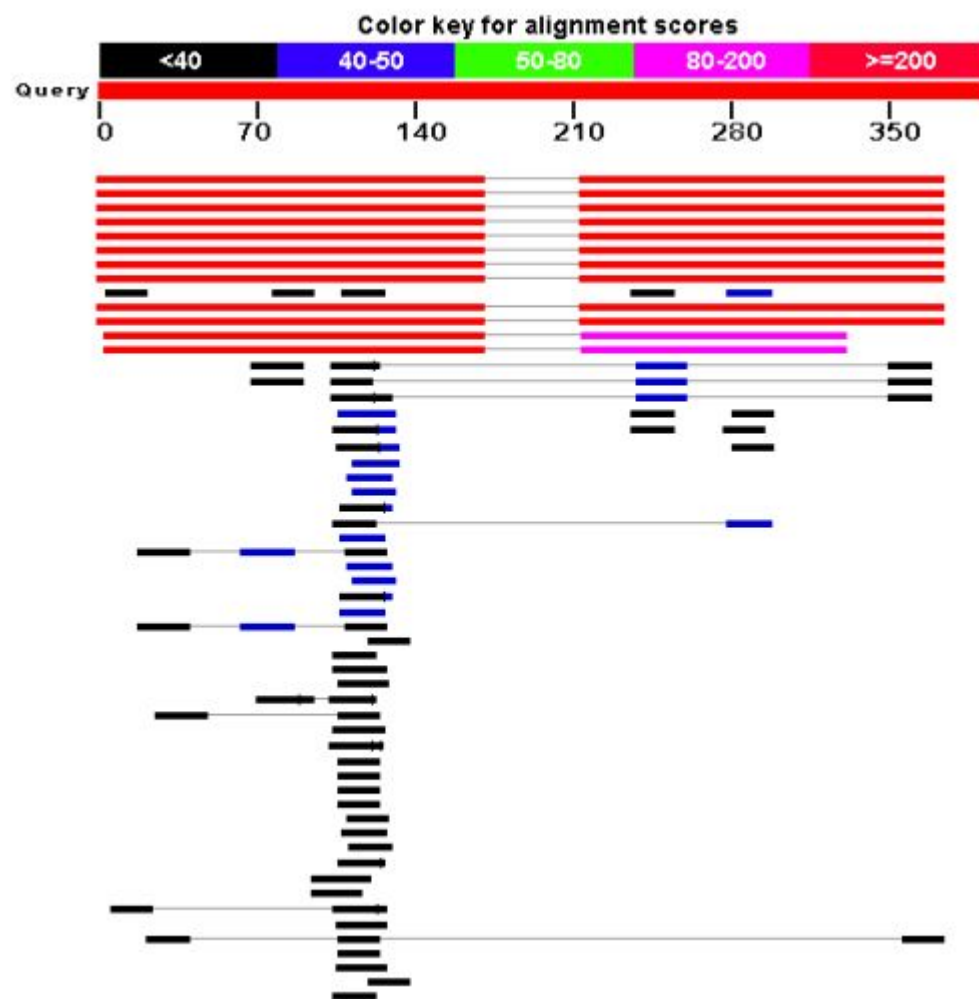
P and E

P	probability of observing a score S ≥ x	• This determines whether or not the region's score is statistically significant
E_{HSP}	expected frequency of observing a score S ≥ x for one region	• determine which regions will produce local alignments • must be calculated every time (constantly changing database)
E	expected frequency of observing a score S ≥ x for the final alignment	• determine whether the alignment meets the user threshold

Limitations [3]

Lower accuracy

May miss out on similarities that are gapped

May not find the best possible match

A C G T G A C
G T G T G G A

A T A A T C A T
T T C A G C T T

Glossary

BLAST: algorithm for comparing biological sequences

BLASTN: nucleotide-nucleotide BLAST, finds similar DNA sequences

BLASTX: compares the protein translation of a shorter sequence of nucleotides against a protein database

TBLASTX: find distant similarities between nucleotide sequences, by translating both the queried and database nucleotide sequences

BLASTP: protein-protein BLAST, finds similar protein sequences

TBLASTN: compares a protein sequence against the 6-frame translations of nucleotide sequences

BLOSUM: BLOcks SUbstitution Matrix

HSP: high-scoring segment pair, larger stretch of matched sequence

Sources

1. https://capricorn.bc.edu//bi204/wp-content/uploads/2015/08/Chapter-9-2015.pdf

2. http://etutorials.org/Misc/blast/Part+III+Practice/Chapter+5.+BLAST/5.2+The+BLAST+Algorithm/

3. https://www.gqlifesciences.com/3-problems-with-using-blast-for-sequence-alignments-in-ip-searching/#:~:text=and%20NCBI%20BLAST%20makes%20it,attempts%20to%20align%20two%20sequences%E2%80%A6

4. http://cshprotocols.cshlp.org/content/2007/7/pdb.ip41.full

5. http://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/650/Explanation_BLAST_output.html