

Sequence alignment theory and applications

BLAST algorithm

Learning Objectives

- Understand the principles of the BLAST algorithm
- Understand the different BLAST programs, parameters and their applications
- Be able to adjust the sensitivity and specificity of BLAST searches by adjusting parameters
- Understand and evaluate BLAST results

Learning Outcomes

- Select the correct BLAST algorithm and database for the appropriate biological question being asked, i.e. select between blastp, blastn, blastx etc. based on the question being addressed
- Understand the meaning of the various output metrics/results
- Comment on results based on the various output metrics

Why do a BLAST search?

- You can get important clues about the function of an as yet uncharacterised sequence.
- Identify homologous species for unknown species.
- Locate domains in protein sequence.
- Establish phylogeny.
- Mapping DNA from unknown location.

Principles of BLAST

- Basic Local Alignment Search Tool (Altschul et al., 1990) finds regions of local similarity between sequences.
- Compares nucleotide or protein sequences to sequence databases and calculates the statistical significance of matches.
- Heuristic method for local alignment.
- Based on the assumption that good alignments contain short lengths of exact matches.

<https://blast.ncbi.nlm.nih.gov/Blast.cgi>

NIH U.S. National Library of Medicine NCBI National Center for Biotechnology Information Sign in to NCBI

BLAST® Home Recent Results Saved Strategies Help

Basic Local Alignment Search Tool

BLAST finds regions of similarity between biological sequences. The program compares nucleotide or protein sequences to sequence databases and calculates the statistical significance.

Learn more

NEWS

Search Betacoronavirus Database

We have created a new BLAST database focused on the SARS-CoV-2 (Severe acute respiratory syndrome coronavirus 2) Sequences. For further detail please visit [NCBI GenBank](#).

Mon, 03 Feb 2020 10:00:00 EST [More BLAST news...](#)

Web BLAST

Nucleotide BLAST nucleotide ▶ nucleotide

blastx translated nucleotide ▶ protein

tblastn protein ▶ translated nucleotide

Protein BLAST protein ▶ protein

BLAST Genomes

Enter organism common name, scientific name, or tax id

Human Mouse Rat Microbes Search

Specialized Searches

Specialized searches

SmartBLAST



Find proteins highly similar to your query

Primer-BLAST



Design primers specific to your PCR template

Global Align



Compare two sequences across their entire span (Needleman-Wunsch)

CD-search



Find conserved domains in your sequence

IgBLAST



Search immunoglobulins and T cell receptor sequences

VecScreen



Search sequences for vector contamination

CDART



Find sequences with similar conserved domain architecture

Multiple Alignment



Align sequences using domain and protein constraints

MOLE-BLAST



Establish taxonomy for uncultured or environmental sequences

3 Parts of BLAST

1. Set-up
2. Find local alignments between the query sequence and a sequence in a database.
3. Produce p-values and a rank ordering of the local alignments according to the p-values.

BLAST - Part 1

1. Set-up

- Read in query, search parameters and database
- Check for low complexity or other repeats
(optional)

3 Parts of BLAST

1. Set-up
2. Find local alignments between the query sequence and a sequence in a database.
3. Produce p-values and a rank ordering of the local alignments according to the p-values.

BLAST - Part 2

2. Find local alignments

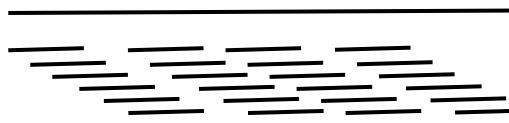
- Break query sequence into “words”
 - Typically, 3 for protein sequence and 11 for nucleotide sequence
- Matches between the “words” in the “query” and “words” in the database sequences are found.
- These can be exact matches (for nt-nt search) or matches that satisfy some positive-valued threshold score (for prot-prot search) as determined using a scoring matrix.

Sequences to “words”

MRRGRLLEIALGFTVLLASYTSHGA

MRR
RRG
RGR
GRL
RLL
LLE
LEI
EIA
IAL
...

Break query sequence into “words”



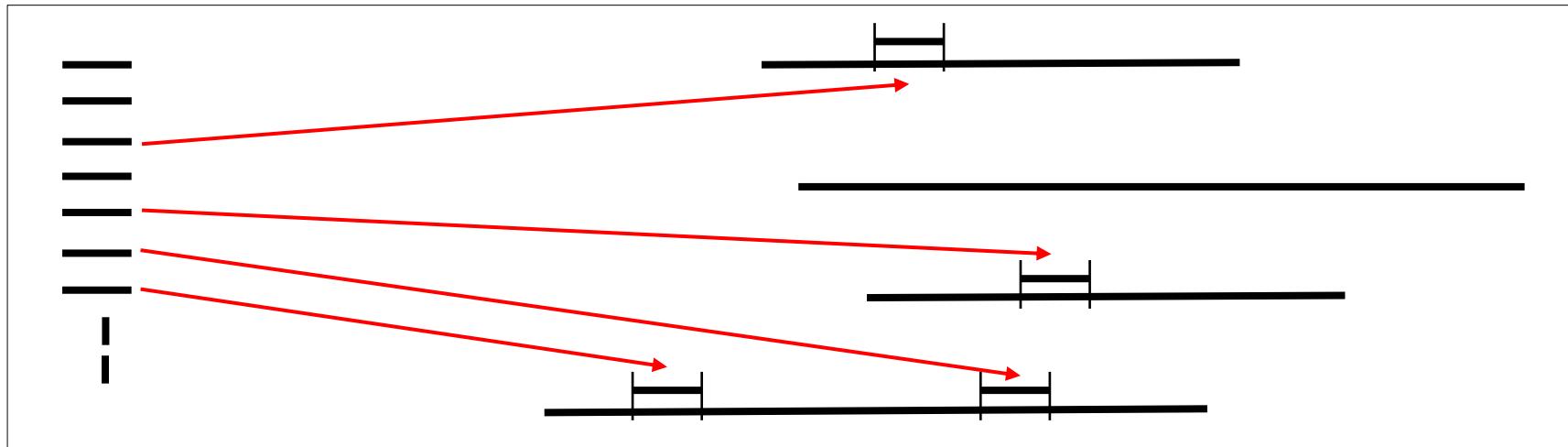
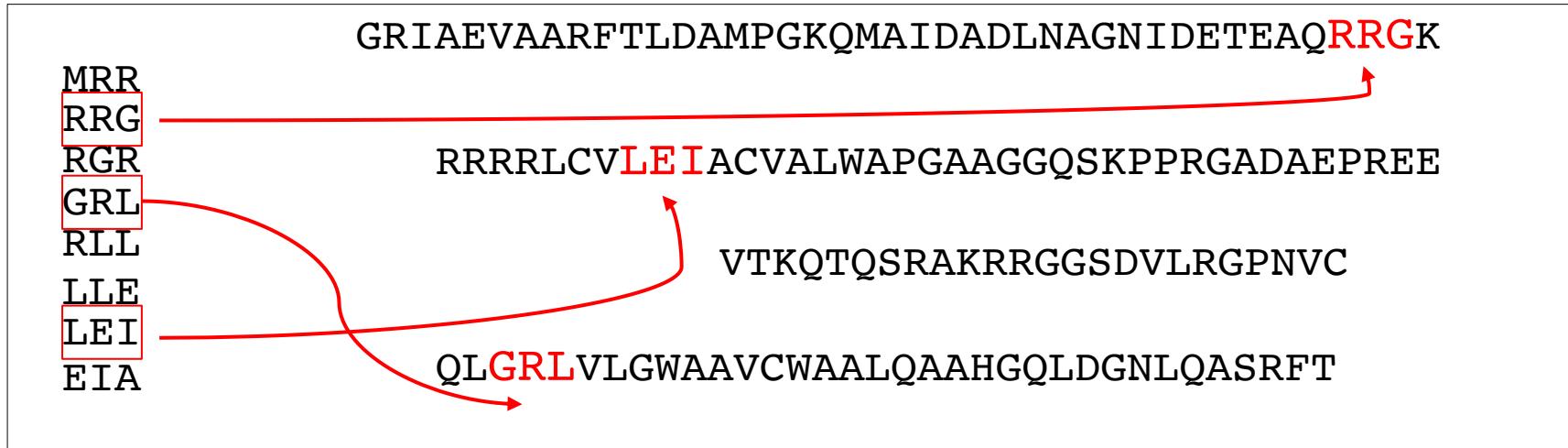
Break database sequences into “words”



Find location of matching “words” in the database sequences

Query “words”

Database sequences

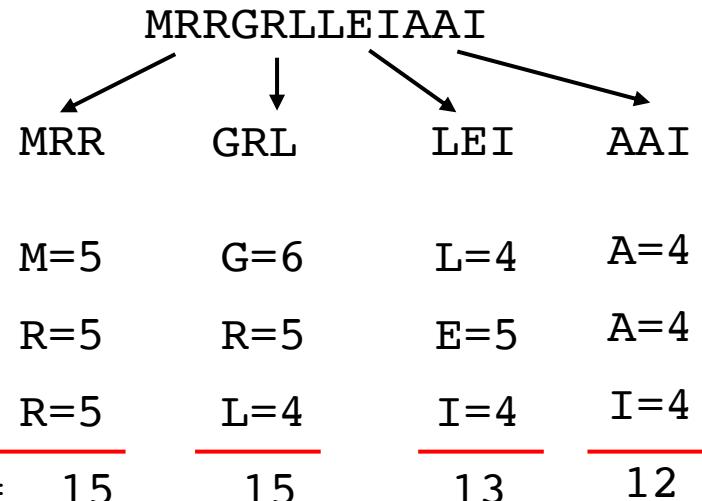


BLAST - Part 2

2. Find local alignments, continued (1)

- Each “word” match is scored
 - The score is computed by assigning a value to each aligned pair of letters and then summing these values over the length of the alignment.
 - For protein sequence alignments, scores for every possible amino acid letter pair are given in a “substitution matrix” where likely substitutions have positive values and unlikely substitutions have negative values. Default matrix used by BLASTP is BLOSUM62.
 - For nucleotide alignments, BLAST uses a reward of +2 for aligned pairs of identical letters and a penalty of –3 for each non-identical aligned pair.

BLAST - Part 2

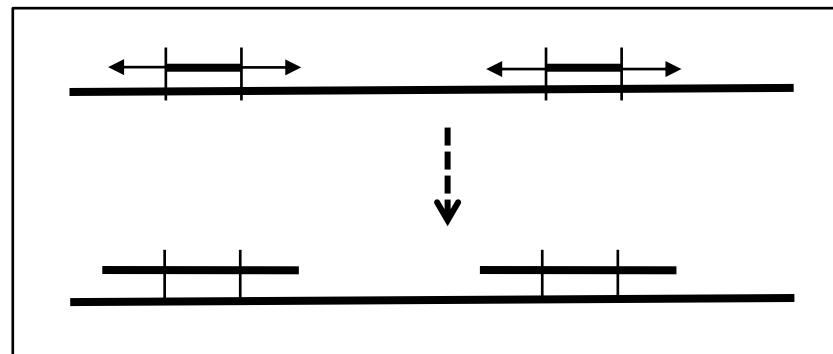
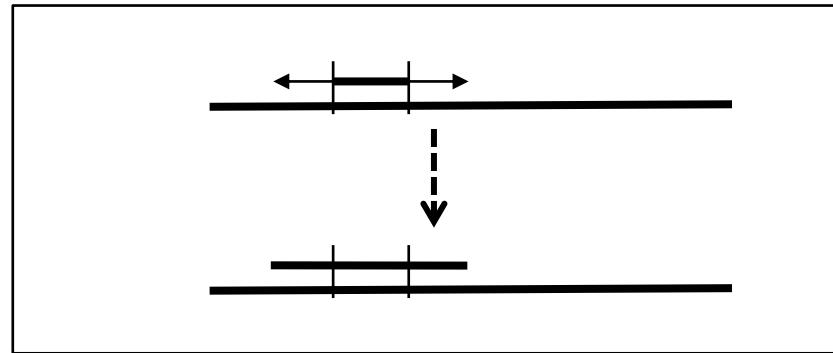


- Words that do not score above a given threshold (T) are removed from the list.
 - E.g. When T is 13, AAI is removed.
 - Matches with a score above a given threshold are used to “seed” an ungapped alignment.

BLAST - Part 2

Extend Hits

- The “words” are extended in both directions into gap-free extensions for as long as the cumulative alignment score increases or stays the same.
- Gap-free extensions with alignment score above a given threshold are used to seed gapped extensions
- Gapped extensions (HSPs) scoring above an empirically determined cut-off score, S , are retained.



BLAST - Part 2

- HSP is the fundamental unit of BLAST algorithm output.
- An HSP consists of two sequence fragments of arbitrary but equal length whose alignment is locally maximal and for which the alignment score meets or exceeds a cut-off score, S.

```
Query 1 MRRGRLLEIALGFTVLLASYTSHGA 25
          M RGRL+ +A+G    +L+ +T    G
Sbjct 1 MCRGRLVRLAVGLVAVLSLWTEPGG 25
```

High-Scoring Segment Pair (HSP)

- 2 sequences
- Scoring system
- Cut-off score

3 Parts of BLAST

1. Set-up
2. Find local alignments between the query sequence and a sequence in a database.
3. Produce p-values and a rank ordering of the local alignments according to the p-values.

BLAST - Part 3

3. Determine the statistical significance of each HSP score (Raw score -> P-value -> E-value).
 - What is the probability (P-value) of the HSP alignment occurring by chance?
 - Are the two sequences descended from the same common ancestor?
 - P-value is the probability of finding exactly a HSPs with score $\geq S$

BLAST - Part 3

Problems:

- Longer sequences are more likely to find higher scoring pairs.
- Longer databases are more likely to result in higher scoring pairs.

Solution:

- Convert Probability values (P-values) to Expectation values (E-values)

BLAST – Part 3

- The “Expect Value” (E-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.

$$E = \frac{\text{Length of Database}}{\text{Length of Sequence}} \times \text{Probability}$$

- E-value is dependent on the length of the sequence and the length of the database
As the database grows the E-value will change.
- E-value $< 1e-179$ is written as 0.0

Bit Score

- Bit score (S') is a normalised raw score that is independent of the search space.
 - Search space (N) = Query length (n) \times Sum of length of sequences in the database (m)
- It measures sequence similarity independent of query sequence length and database size and is normalized based on the raw pairwise alignment score.
 - It is linearly related to the raw score.
 - The higher the bit score, the more significant the match is.
- An alignment may have different E-values if searched against different databases but will have the same bit score.

Take a moment to reflect

1. How are the scores and E-value related? Will a high score give a high or low E-value?
2. Would a longer sequence give a higher or lower E-value?
3. How would an increase in database size affect the E-value?
4. What effect will adjusting the gap penalty have on the alignment?

BLAST Output (1)

BLAST® » blastp suite » results for RID-E29U13HT016

Home Recent Results Saved Strategies Help

[Edit Search](#) Save Search Search Summary ▾

Job Title sequence

RID E29U13HT016 Search expires on 06-12 02:32 am [Download All](#) ▾

Program BLASTP ⓘ Citation ▾

Database nr See details ▾

Query ID lcl|Query_34004

Description sequence

Molecule type amino acid

Query Length 59

Other reports Distance tree of results Multiple alignment MSA viewer ⓘ

Filter Results

Organism only top 20 will appear exclude

Type common name, binomial, taxid or group name

+ Add organism

Percent Identity E value Query Coverage

to to to

[Filter](#) [Reset](#)

Descriptions Graphic Summary Alignments Taxonomy

BLAST Output (2)

Descriptions

Graphic Summary

Alignments

Taxonomy

ⓘ hover to see the title ⚡ click to show alignments

Show Conserved Domains

Alignment Scores

█ < 40

█ 40 - 50

█ 50 - 80

█ 80 - 200

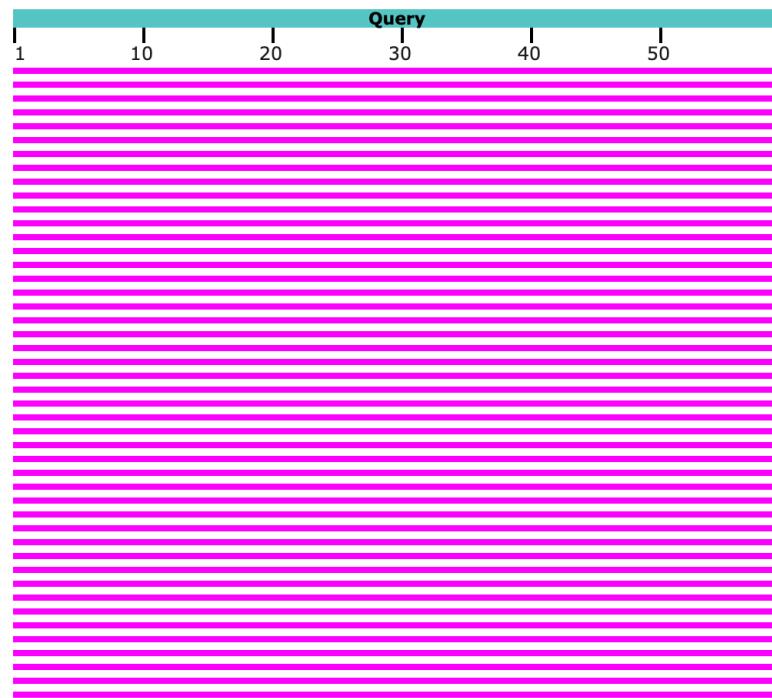
█ >= 200



100 sequences selected ⓘ

No putative conserved domains have been detected

Distribution of the top 100 Blast Hits on 100 subject sequences



BLAST Output (3)

Descriptions Graphic Summary Alignments Taxonomy

Sequences producing significant alignments Download ▾ Manage Columns ▾ Show 100 ▾ ?

select all 100 sequences selected GenPept Graphics Distance tree of results Multiple alignment

	Description	Max Score	Total Score	Query Cover	E value	Per. Ident	Accession
<input checked="" type="checkbox"/>	fibrillin_1_(Marfan syndrome)_isoform CRA_a [Homo sapiens]	126	126	100%	6e-32	100.00%	EAW77353.1
<input checked="" type="checkbox"/>	fibrillin_1 [Homo sapiens]	126	126	100%	6e-32	100.00%	BAD16739.1
<input checked="" type="checkbox"/>	fibrillin [Homo sapiens]	126	126	100%	6e-32	100.00%	AAB02036.1
<input checked="" type="checkbox"/>	fibrillin-1 preproprotein [Homo sapiens]	126	126	100%	6e-32	100.00%	NP_000129.3
<input checked="" type="checkbox"/>	epididymis secretory sperm binding protein [Homo sapiens]	126	126	100%	6e-32	100.00%	ADO22212.1
<input checked="" type="checkbox"/>	fibrillin_1 [Homo sapiens]	125	125	100%	8e-32	100.00%	BAD16738.1
<input checked="" type="checkbox"/>	fibrillin [Homo sapiens]	124	124	100%	1e-31	100.00%	CAA45118.1
<input checked="" type="checkbox"/>	unnamed protein product [Homo sapiens]	124	124	100%	2e-31	100.00%	BAG65498.1
<input checked="" type="checkbox"/>	fibrillin_1 [Homo sapiens]	114	114	100%	2e-30	100.00%	BAD16737.1
<input checked="" type="checkbox"/>	FBN1 isoform 5 [Pan troglodytes]	124	124	100%	2e-31	98.31%	PNI74996.1
<input checked="" type="checkbox"/>	fibrillin-1 [Pan troglodytes]	124	124	100%	2e-31	98.31%	XP_001149266.4
<input checked="" type="checkbox"/>	fibrillin-1 [Carlito syrichta]	123	123	100%	7e-31	98.31%	XP_008068093.1
<input checked="" type="checkbox"/>	FBN1 isoform 3 [Pan troglodytes]	118	118	100%	1e-30	98.31%	PNI74995.1
<input checked="" type="checkbox"/>	fibrillin_1 [Grampus griseus]	116	116	100%	7e-32	96.61%	SEQ26544.1
<input checked="" type="checkbox"/>	PREDICTED: LOW QUALITY PROTEIN: fibrillin-1 [Rousettus aegyptiacus]	122	122	100%	7e-31	96.61%	XP_016018715.1
<input checked="" type="checkbox"/>	fibrillin-1 isoform X2 [Delphinapterus leucas]	122	122	100%	9e-31	96.61%	XP_022442869.1
<input checked="" type="checkbox"/>	fibrillin-1 isoform X2 [Equus caballus]	122	122	100%	9e-31	96.61%	XP_023473665.1

BLAST Output

[Descriptions](#)[Graphic Summary](#)[Alignments](#)[Taxonomy](#)Alignment view [Pairwise](#)[Restore defaults](#)[Download](#)

100 sequences selected

[Download](#) [GenPept](#) [Graphics](#)[Next](#) [Previous](#) [Descriptions](#)

fibrillin 1 (Marfan syndrome), isoform CRA_a [Homo sapiens]

Sequence ID: [EAW77353.1](#) Length: 2869 Number of Matches: 1Range 1: 1 to 59 [GenPept](#) [Graphics](#)[▼ Next Match](#) [▲ Previous Match](#)

Score	Expect	Method	Identities	Positives	Gaps
126 bits(316)	6e-32	Composition-based stats.	59/59(100%)	59/59(100%)	0/59(0%)

Query 1	MRRGRLLEIALGFTVLLASYTSHGADANLEAGNVKETRASRAKRGGGGHDALKGPNC	59
	MRRGRLLEIALGFTVLLASYTSHGADANLEAGNVKETRASRAKRGGGGHDALKGPNC	
Sbjct 1	MRRGRLLEIALGFTVLLASYTSHGADANLEAGNVKETRASRAKRGGGGHDALKGPNC	59

Related Information

[Gene](#) - associated gene details[Download](#) [GenPept](#) [Graphics](#)[Next](#) [Previous](#) [Descriptions](#)

fibrillin 1 [Homo sapiens]

Sequence ID: [BAD16739.1](#) Length: 2871 Number of Matches: 1Range 1: 1 to 59 [GenPept](#) [Graphics](#)[▼ Next Match](#) [▲ Previous Match](#)

Score	Expect	Method	Identities	Positives	Gaps
126 bits(316)	6e-32	Composition-based stats.	59/59(100%)	59/59(100%)	0/59(0%)

Query 1	MRRGRLLEIALGFTVLLASYTSHGADANLEAGNVKETRASRAKRGGGGHDALKGPNC	59
	MRRGRLLEIALGFTVLLASYTSHGADANLEAGNVKETRASRAKRGGGGHDALKGPNC	
Sbjct 1	MRRGRLLEIALGFTVLLASYTSHGADANLEAGNVKETRASRAKRGGGGHDALKGPNC	59

Related Information

[Gene](#) - associated gene details

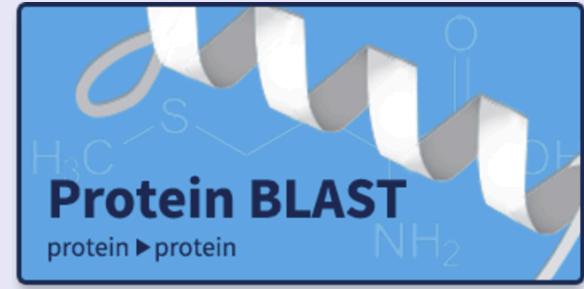
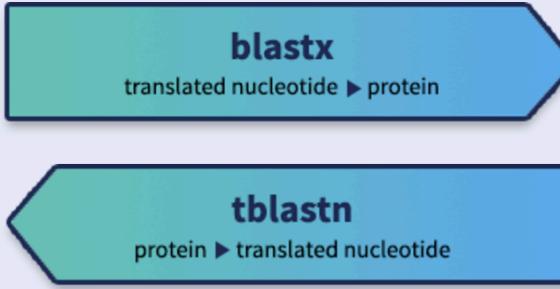
BLAST Output (nucleotide)

Score 366 bits(198)		Expect 1e-94	Identities 274/311(88%)	Gaps 4/311(1%)	Strand Plus/Minus	
Query	48468796	CGGTGGCTCACGCCCTGTAATCCCAGCACTTGGGAGGCCGAGGCAGGCACGAGGT				48468855
Sbjct	11303	CGGTGGTTACGCCCTGTAATCCCAGCACTTGAGAGGCCGAGGCAGGCAGGATCACAAGGT				11244
Query	48468856	CAGGAGATCGAGACCATCCTGGCTAACACACAGTGAAACCTCATCTACTAAAAATACAAA				48468915
Sbjct	11243	CAGGAGATTGAGACTGTCCCTGGCTAACACACAGTGAAACCCCTGTCTACTAAAAATACAAA				11184
Query	48468916	AAATTAGCCAGGCGTGGTGGTGGTGCCTGTAGTCCCAGCTACTCGGGAGGCTGAGGCAG				48468975
Sbjct	11183	AAATTAGCTGGGCATGGTGGCACATGCCCTGTAGTCCCAGCTACTCGGGAGGCTGAGGCAG				11124
Query	48468976	GAGAATGGCATGAACCTGGGAGGCCGAGCTTGCAGTGAGCCAAGATGGTACCACTGTACT				48469035
Sbjct	11123	GAGAATGGCATGAACCTGGGAGACGGAGCTTGCAGTGAGCAGAGATTGCGCCACTGCACT				11064
Query	48469036	CCAGCCTGGCGACAGAGCGAGACTCCGTCTCaaaaaaaaaaa--aaa--aaaatatatata				48469092
Sbjct	11063	CCAGCCTGGGTGACAGAGCGAGACTCCATCTCAAAAAAACAAAGCAAAACAAA-ATAAATT				11005
Query	48469093	taaaatataat	48469103			
Sbjct	11004	TACAATAAAAT	10994			

Low Complexity Region

- “Low complexity region” is a region of sequence composed of few kinds of elements (low compositional complexity).
- They might give high scores that confuses the program to find the actual significant sequences in the database.
- Marked with X (protein sequences) or N (nucleotide sequences) and ignored in the BLAST program.
- SEG (proteins) or DUST (nucleotides)

BLAST Algorithms



- Highly similar sequences (megablast)
- More dissimilar sequences (discontiguous megablast)
- Somewhat similar sequences (blastn)

- Quick BLASTP (Accelerated protein-protein BLAST)
- blastp (protein-protein BLAST)
- PSI-BLAST (Position-Specific Iterated BLAST)
- PHI-BLAST (Pattern Hit Initiated BLAST)
- DELTA-BLAST (Domain Enhanced Lookup Time Accelerated BLAST)

Enter Query Sequence BLASTN programs search nucleotide subjects using a nucleotide query. [more...](#)

Enter accession number(s), gi(s), or FASTA sequence(s) [?](#) Clear Query subrange [?](#)

From _____
To _____

Or, upload file [Choose File](#) No file chosen [?](#)

Job Title _____
Enter a descriptive title for your BLAST search [?](#)

Align two or more sequences [?](#)

Enter Subject Sequence

Enter accession number(s), gi(s), or FASTA sequence(s) [?](#) Clear Subject subrange [?](#)

From _____
To _____

Or, upload file [Choose File](#) No file chosen [?](#)

Can align two or more sequences in all BLAST algorithms

PSI-BLAST

- Used to find distant relatives of a protein.
- Does an iterative search.
 - First search a regular BlastP
 - Generates a multiple alignment of the HSPs above an E-value threshold and calculates a profile or a Position Specific Scoring Matrix (PSSM)
 - PSSM captures the conservation pattern in the alignment and records it as a scoring matrix
 - In the next iteration this profile is used instead of the original substitution matrix to detect sequences that match the conservation pattern specified by the PSSM
 - After every iteration new sequences above the threshold are added to the the PSSM
- In this way, PSI-BLAST allows detection of distant sequence similarities.

Summary

- The algorithm underlying a BLAST search is complex.
- By understanding it you can adjust the sensitivity and specificity of the search.
- Higher the Bit score the lower the E-value
 - E-value depends on the size of the database and the length of the query.
 - A “significant” E-value of a result will depend on your sequence and goal of the search.
- How to choose amongst the different “flavours” of BLAST for proteins and nucleotides

Useful Resources

- [BLAST QuickStart](#)
- [BLAST Handbook](#)
- [PSI-BLAST](#)
- NCBI – Expect Values, part 1 – [video](#)
- NCBI – Expect Values, part 2 – [video](#)

