

BLAST:
Basic Local Alignment Search Tool

BLAST: Basic Local Alignment Search Tool

BLAST extracts “words” from the query sequence to search for similar sequences in the database

Scoring Matrix (*i.e.* BLOSUM)

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
A	2																			
R	-2	6																		
N	0	0	2																	
D	0	-1	2	4																
C	-2	-4	-4	-5	12															
Q	0	1	1	2	-5	4														
E	0	-1	1	3	-5	2	4													
G	1	-3	0	1	-3	-1	0	5												
H	-1	2	2	1	-3	3	1	-2	6											
I	-1	-2	-2	-2	-2	-2	-2	-3	-2	5										
L	-2	-3	-3	-4	-6	-2	-3	-4	-2	2	6									
K	-1	3	1	0	-5	1	0	-2	0	-2	-3	5								
M	-1	0	-2	-3	-5	-1	-2	-3	-2	2	4	0	6							
F	-4	-4	-4	-6	-4	-5	-5	-5	-2	1	2	-5	0	9						
P	1	0	-1	-1	-3	0	-1	-1	0	-2	-3	-1	-2	-5	6					
S	1	0	1	0	0	-1	0	1	-1	-1	-3	0	-2	-3	1	2				
T	1	-1	0	0	-2	-1	0	0	-1	0	-2	0	-1	-3	0	1	3			
W	-6	2	-4	-7	-8	-5	-7	-7	-3	-5	-2	-3	-4	0	-6	-2	-5	17		
Y	-3	-4	-2	-4	0	-4	-4	-5	0	-1	-1	-4	-2	7	-5	-3	-3	0	10	
V	0	-2	-2	-2	-2	-2	-2	-1	-2	4	2	-2	2	-1	-1	-1	0	-6	-2	4

Query word ($W = 3$)

⊥

Query: GSDFWQETRASFGCSLAALLNKCKT**PQG**QRLVNQWIKQPLMDKNRIEERLNLVEAFGCATSWPI

Neighborhood
words

PQG 18
PEG 15
PRG 14
PKG 14
PNG 13
PDG 13
PHG 13
PMG 13
PSG 13
PQA 12
PQN 12
...

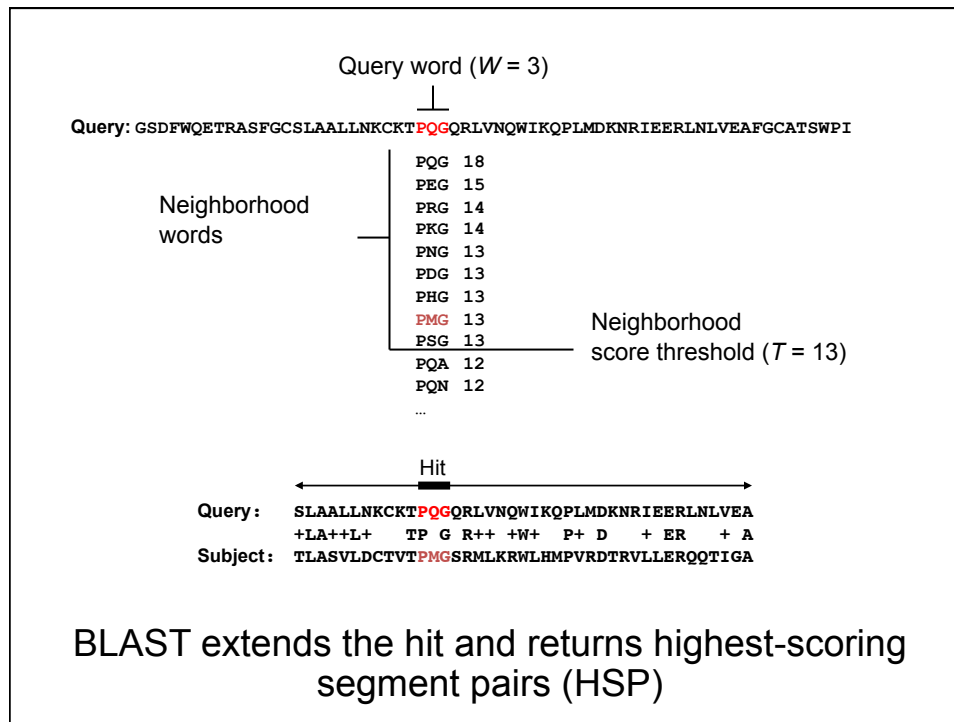
Neighborhood
score threshold ($T = 13$)

Hit

Query: SLAALLNKCKT**PQG**QRLVNQWIKQPLMDKNRIEERLNLVEA

Subject: TLASVLDCTVT**PMG**SRMLKRWLHMPVRDTRVLLERQQTIGA

Identifies all word hits



E-values

- $E=10$ means 10 similar matches are expected by chance
- Lower values mean more stringent results
- Default is typically 10
- $E < 0.05$ is significant, but higher stringency is usually required
- Short regions of high identity are less significant than long regions of moderate similarity
- E depends on the database size

BLAST Programs

	Query	Database
blastn	Nucleotide	Nucleotide
blastp	Protein	Protein
blastx	Nucleotide	Protein
tblastn	Protein	Translated
tblastx	Nucleotide	Translated

NCBI BLAST

The screenshot shows the NCBI BLAST website interface. At the top, there are navigation tabs: Home, Recent Results, Saved Strategies, and Help. A search bar is present with the text "BLAST finds regions of similarity between biological sequences. mscr...". Below the search bar, there is a link to "New Aligning Multiple Protein Sequences? Try the COBAL Multiple Alignment Tool. (Go)".

The main content area is divided into several sections:

- BLAST Assembled RefSeq Genomes:** A section where users can choose a species genome to search. It lists various species including Human, Mouse, Rat, Arabidopsis thaliana, Oryza sativa, Bos taurus, Drosophila melanogaster, Gallus gallus, Pan troglodytes, Microbes, and Apis mellifera.
- Basic BLAST:** A section where users can choose a BLAST program to run. It lists:
 - nucleotide_blast:** Search a nucleotide database using a nucleotide query. Algorithms: blastn, megablast, discontinuous megablast.
 - protein_blast:** Search protein database using a protein query. Algorithms: blastp, psi-blast, phi-blast.
 - blastx:** Search protein database using a translated nucleotide query.
 - tblastn:** Search translated nucleotide database using a protein query.
 - tblastx:** Search translated nucleotide database using a translated nucleotide query.
- Specialized BLAST:** A section where users can choose a type of specialized search. It lists:
 - Make specific primers with **Primer-BLAST**
 - Search **trace archives**
 - Find **conserved domains** in your sequence (cds)
 - Find sequences with similar **conserved domain architecture** (cdart)
 - Search sequences that have **gene expression profiles** (GEO)
 - Search **immunoglobulins** (IgBLAST)
 - Search for **SNPs** (SNP)
 - Screen sequence for **vector contamination** (vecscreen)
 - Align** two (or more) sequences using BLAST (bl2seq)
 - Search **protein or nucleotide targets** in PubChem BioAssay
 - Search **SRA transcript and genomic libraries**
 - Constraint Based Protein Multiple Alignment Tool**
 - Needleman-Wunsch Global Sequence Alignment Tool**

At the bottom of the page, there are links for Copyright, Disclaimer, Privacy, Accessibility, Contact, and Send Feedback. The footer also includes the NCBI logo and the text "NCBI | NLM | NIH | Databases".

BLAST: Basic Local Alignment Search Tool

Basic BLAST

Choose a BLAST program to run.

nucleotide blast	Search a nucleotide database using a nucleotide query <i>Algorithms: blastn, megablast, discontinuous megablast</i>
protein blast	Search protein database using a protein query <i>Algorithms: blastp, psi-blast, phi-blast</i>
blastx	Search protein database using a translated nucleotide query
tblastn	Search translated nucleotide database using a protein query
tblastx	Search translated nucleotide database using a translated nucleotide query

BLAST: Basic Local Alignment Search Tool

The screenshot shows the BLAST web interface with several key sections and annotations:

- Enter Query Sequence:** A text area containing a DNA sequence: AAACCTCAAAAGCTCTAGAGAGAAGAGAGAGAGAGATCGAAGGTAAGAGAAAGATGT TAGAGATCGCA. Below it is a "Choose File" button for uploading a file.
- Choose Search Set:** A section for selecting the database. The "Database" dropdown is set to "Nucleotide collection (nr/nt)". A yellow callout bubble points to this dropdown with the text "Choose a database".
- Program Selection:** A section for selecting the BLAST program. The "Optimize for" radio buttons include "Highly similar sequences (megablast)", "More dissimilar sequences (discontiguous megablast)", and "Somewhat similar sequences (blastn)".
- Algorithm parameters:** A section for setting search parameters. The "Expect threshold" is set to 1e-100. A yellow callout bubble points to this field with the text "Choose an E value (likelihood that a similar match occurs by chance in the database used)".

BLAST output

The screenshot shows the NCBI BLAST interface. At the top, it displays the query ID (k13391), description (Nucleic acid), and length (988). A graphic summary shows the distribution of 13 Blast Hits on the query sequence. Below this is a table of sequences producing significant alignments. A yellow callout bubble points to the 'Max score' column, containing the text: "Score: a measure of the similarity between the query and subject sequences".

Accession	Description	Max score	Total score	Query coverage	E-value	Max ident	Links
NM_110155.3	Arabidopsis thaliana PPI8 (NAMED PLASMA MEMBRANE INTRINSIC PROTEIN 18); water channel	1808	1808	99%	0.0	99%	View Map
NM_001084528.1	Arabidopsis thaliana PPI8 (NAMED PLASMA MEMBRANE INTRINSIC PROTEIN 18); water channel	1220	1790	99%	0.0	99%	View Map
NM_002825184.1	Arabidopsis thaliana PPI8 (NAMED PLASMA MEMBRANE INTRINSIC PROTEIN 18); water channel	1613	1615	99%	0.0	98%	View Map
NM_001084527.1	Arabidopsis thaliana PPI8 (NAMED PLASMA MEMBRANE INTRINSIC PROTEIN 18); water channel	1461	1469	82%	0.0	99%	View Map
NM_002727586.1	Arabidopsis thaliana PPI8 (NAMED PLASMA MEMBRANE INTRINSIC PROTEIN 18); water channel	1020	1020	87%	0.0	88%	View Map
NM_116060.3	Arabidopsis thaliana PPI4 (PLASMA MEMBRANE INTRINSIC PROTEIN 1A); water channel (PPI4)	1000	1000	86%	0.0	87%	View Map
NM_001084524.1	Arabidopsis thaliana PPI4 (PLASMA MEMBRANE INTRINSIC PROTEIN 1A); water channel (PPI4)	1000	1000	88%	0.0	87%	View Map
NM_002825183.1	Arabidopsis thaliana PPI4 (PLASMA MEMBRANE INTRINSIC PROTEIN 1A); water channel (PPI4)	723	723	87%	0.0	82%	View Map
NM_1001045.4	Arabidopsis thaliana PPI4 (PLASMA MEMBRANE INTRINSIC PROTEIN 1A); water channel (PPI4)	723	723	87%	0.0	82%	View Map
NM_002825182.1	Arabidopsis thaliana PPI4 (PLASMA MEMBRANE INTRINSIC PROTEIN 1A); water channel (PPI4)	612	612	24-172	79%	82%	View Map
NM_001084522.4	Arabidopsis thaliana PPI4 (PLASMA MEMBRANE INTRINSIC PROTEIN 1A); water channel (PPI4)	602	602	12-100	81%	81%	View Map
NM_002328231.1	Rattus norvegicus Aquaporin PIP3, putative, mRNA	501	501				View Map
NM_001302093.1	Populus trichocarpa Aquaporin, PIP family, PIP subfamily, mRNA	523	523				View Map

Different BLAST servers

The image shows two different BLAST web interfaces. On the left is the Cassava Genome Database interface, which includes a search bar, a dropdown menu for 'Program' (set to 'blastn') and 'Database' (set to 'Cassava ESTs'), and various options for filtering and displaying results. On the right is the Phytosome v5.0 search page, which features a 'Run BLAST' button, a 'QUERY' section for entering the sequence, a 'TARGET' section for selecting a genome, and a 'PARAMETERS' section for configuring search options like 'BLAST Program', 'Output Format', and 'Expect threshold'.

PSSM

POS	PROBE	CONSENSUS	PROFILE																				+/-	
			A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y		
1	E C V L	V	3	-2	3	4	0	4	-1	3	-1	4	4	1	1	1	-2	1	2	6	-6	-2	9	
2	L L S P	L	2	-2	-2	-1	3	0	-1	3	-1	6	5	-1	3	0	-1	3	1	4	1	-1	9	
3	V V V V	V	2	2	-2	-2	2	2	-3	11	-2	8	6	-2	1	-2	-2	0	2	15	-9	-1	9	
4	K E A T	A	6	-2	5	6	-5	4	1	0	5	-2	0	3	3	1	3	6	0	-6	-4	9		
5	A P L P	P	6	-1	0	1	-2	2	0	1	0	2	2	0	8	2	0	2	2	3	-5	-4	9	
6	G G G G	G	7	1	7	5	-6	15	-1	-3	0	-4	-3	4	3	2	-3	6	4	2	-11	-7	9	
7	S S Q E	D	4	-1	7	7	-6	7	2	-2	2	-3	-2	4	3	6	1	6	2	-1	-6	-5	9	
8	S S T P	S	4	4	2	2	-4	4	-1	0	2	-3	-2	2	7	0	1	10	6	0	-2	-4	9	
9	V L V A	V	5	0	-1	-1	3	1	-2	7	-2	7	6	-1	1	-1	-3	0	2	10	-5	-1	9	
10	K R R S	R	0	-1	1	1	-5	0	2	-2	8	-3	1	3	3	10	5	1	-2	7	-5	9		
11	M L I I	I	0	-2	-3	-2	7	-3	-3	11	-1	11	10	-2	-2	-1	-2	-2	1	9	-3	1	9	
12	S S T S	S	4	6	2	2	-3	5	-1	0	2	-3	-2	3	4	-1	1	12	6	0	0	-4	9	
13	C C C C	C	3	15	-5	-5	-1	2	-1	3	-5	-8	-6	-3	1	-6	-3	7	3	3	-13	10	9	
14	K S Q R	K	1	-2	3	3	-6	1	3	-2	7	-3	0	3	3	5	7	4	1	-2	2	-5	9	
15	A A G S	A	10	3	4	3	-5	8	-1	-1	1	-2	-1	3	4	1	-2	7	4	2	-6	-4	9	
16	T S D S	S	4	3	5	4	-5	6	0	0	2	-3	-2	4	3	1	1	9	6	0	-3	-4	9	
17	G C S Q	G	5	1	6	5	-6	9	1	-2	1	-3	-2	4	3	4	0	6	3	0	-6	-6	9	
18	Y F L S	F	-1	2	-4	-3	9	-3	0	4	-3	6	3	-1	-3	-3	-3	1	-1	2	7	7	9	
19	T T R L	T	1	-2	0	1	0	0	0	2	2	2	3	1	1	1	3	1	7	2	1	-2	9	
20	F F . L	F	-2	-3	-6	-4	10	-4	-1	6	-4	9	6	-3	-4	-4	-3	-2	-1	3	7	8	4	
21	S . . D	S	3	2	5	4	-4	5	0	-1	2	-3	-2	4	3	1	1	8	2	0	-1	-2	4	
22	S . . S	S	2	3	1	1	-2	3	-1	9	1	-2	-1	2	2	0	1	8	2	0	1	-2	4	
23	. . . G	G	2	0	2	1	-2	4	0	0	0	-1	-1	1	1	1	-1	2	1	1	-3	-2	4	
24	. . . D	D	1	-1	4	3	-2	2	1	0	1	-1	-1	2	1	2	0	1	1	0	-3	-1	4	
25	. . . G	G	2	0	2	1	-2	4	0	0	0	-1	-1	1	1	1	-1	2	1	1	-3	-2	4	
26	. . A G N	A	6	0	4	3	-4	6	1	-1	1	-2	-1	5	2	2	-1	3	3	1	-5	-3	4	
27	Y N Y T	Y	0	5	0	-1	5	-1	2	1	-1	0	-1	4	-3	-2	-2	0	3	0	3	6	4	
28	E D D Y	D	2	-2	9	8	-3	3	4	-1	1	-3	2	5	-1	4	-1	1	1	-1	-6	0	9	
29	L M A L	L	3	-5	-3	-1	6	-1	-2	6	-1	10	10	-2	0	0	-2	-1	0	6	-1	0	9	
30	Y N A W	N	4	1	3	2	0	2	3	-1	1	-1	-1	8	0	1	-1	2	1	-1	-1	2	9	
.
48	S G N S	S	4	3	5	3	-4	7	0	-2	2	-4	-3	6	3	1	0	10	3	0	-2	-4	9	
49	S S N Y	S	2	5	2	1	1	2	1	0	1	-2	-2	5	1	-1	0	8	1	-1	3	1	9	

Gribkov et al. 1987

BLAST against the SRA