Current Protocols in Protein Science
Unit 2.5

Sequence Similarity Searching

Gang Hu[1] and Lukasz Kurgan[2*]

[1]School of Mathematical Sciences and LPMC, Nankai University, Tianjin, China.

[2]Department of Computer Science, Virginia Commonwealth University, Richmond, USA.

* Corresponding author
Email: lkurgan@vcu.edu
Phone: 804-827-3986

ABSTRACT

Sequence similarity searching has become an important part of the daily routine of molecular biologists, bioinformaticians and biophysicists. With the rapidly growing sequence databanks, this computational approach is commonly applied to determine functions and structures of unannotated sequences, to investigate relationships between sequences, and to construct phylogenetic trees. We introduce arguably the most popular BLAST-based family of the sequence similarity search tools. We explain basic concepts related to the sequence alignment and demonstrate how to search the current databanks using website versions of BLASTP, PSI-BLAST and BLASTN. We also describe the standalone BLAST+ tool. Moreover, this unit discusses the inputs, parameter settings and outputs of these tools. Lastly, we cover recent advances in the sequence similarity searching, focusing on the fast MMseqs2 method.

Keywords: sequence similarity searching; alignment; BLAST; BLASTP, BLASTN, PSSM; MMseqs2.

INTRODUCTION TO SEQUENCE SIMILARITY SEARCHING AND BLAST PROGRAMS

Identification of similar sequences in large sequence databanks is typically the first step in analysis of new protein, DNA and RNA sequences. Structural and functional annotations of the similar sequences can be used to infer functions and structural features of the new sequences. Many databanks, such as NCBI's GenBank (Benson et al., 2018), EMBL's Nucleotide Sequence Database (Kulikova et al., 2007), Universal Protein Resource (UniProt) (Boutet et al., 2016; The UniProt, 2017), and Protein Data Bank (PDB) (Berman et al., 2000; Rose et al., 2017), provide a rich source of millions of sequences and annotations. More specifically, GenBank provides access to a comprehensive and non-redundant set of reference sequences, the Reference Sequence Database (RefSeq) (O'Leary et al., 2016). As of January 2018, RefSeq includes over 102 million protein sequences and over 21 million transcripts. UniProt is the main database of protein sequences and functional annotations. As of January 2018, it includes 556 thousand manually reviewed proteins and over 107 million computationally analyzed proteins. PDB provides access to 137,000 experimentally determined protein structures. Moreover, NCBI provides the nr database, a comprehensive collection of non-redundant protein sequences from RefSeq, UniProt, PDB and other protein databases, to facilitate searching for similar sequences.

Many algorithms that can be used to search for similar sequences were developed over the last three decades. They include FASTA (Pearson and Lipman, 1988), ClustalW (Larkin et al., 2007; Thompson et al., 1994), HMMER (Johnson et al., 2010), MMseqs2 (Steinegger and Soding, 2017), and Basic Local Alignment Search Tool (BLAST) (Altschul et al., 1990; Camacho et al., 2009). BLAST is arguably the most popular tool. The article that has introduced this algorithm was cited over 70 thousand times (source: Google Scholar on Feb. 22, 2018). BLAST performs and scores sequence alignment to find similar nucleotide or protein sequences for a given query sequence. The search is executed against a large database (typically the nr database) and the algorithm also quantifies statistical significance of the scored output matches. The family of the BLAST programs offers a variety of options that address searching of nucleotide and protein chains, see Table 1. One of the key factors that contribute to the popularity of BLAST is its omnipresent availability. BLAST algorithms are available in two main flavors: 1) WEB BLAST (Johnson et al., 2008), a web-based application that is accessible using any of the major web browsers for which the computations are performed on the webserver side; and 2) BLAST+ (Camacho et al., 2009), a standalone software that can be installed and run on a local computer. The latter option is particularly attractive for users who would like to include BLAST into a larger computational pipeline. The WEB BLAST can be accessed through the NCBI website at https://blast.ncbi.nlm.nih.gov/Blast.cgi. BLAST+ can be downloaded from ftp://ftp.ncbi.nlm.nih.gov/blast/executables/blast+/LATEST/. The programs listed in Table 1 can be found in both WEB BLAST and BLAST+. WEB BLAST also provides access to a variety of search databases including the nr, RefSeq, SWISS-PROT, PDB, human genome, mouse genome etc.

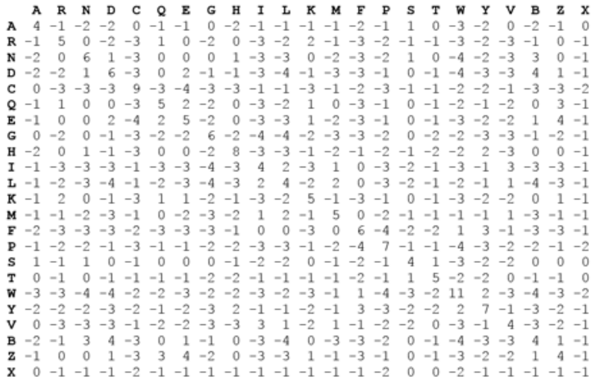
Next, we describe the WEB BLAST and BLAST+ for proteins, followed by the BLAST programs for the nucleotide chains. We conclude this Unit with a discussion of recently developed alternatives to BLAST.

Table 1. Summary of the BLAST programs.

Programs	Query sequence	Target database	Description
BLASTP suite	Protein	Protein	To search protein sequences against a protein database. Includes BLASTP, Quick BLASTP, PSI-BLAST, PHI-BLAST and DELTA-BLAST.
TBLASTN	Protein	Translated Nucleotide	To identify nucleotide sequences encoding proteins similar to the query protein.
BLASTN suite	Nucleotide	Nucleotide	To search nucleotide sequences against a nucleotide database. Includes MEGABLAST, discontiguous MEGABLAST and BLASTN.
BLASTX	Translated Nucleotide	Protein	To identify potential protein products encoded by a nucleotide query.
TBLASTX	Translated Nucleotide	Translated Nucleotide	To identify nucleotide sequences similar to the query transcript based on their coding potential.

BLASTP SUITE - BLAST AGAINST PROTEIN DATABASES

The sequence similarity search performed by BLAST has roots in early sequence alignment algorithms (Needleman and Wunsch, 1970; Smith and Waterman, 1981). The main difference is that BLAST performs a heuristic search that is characterized by a much faster convergence to a solution. The search maximizes a score that quantifies similarity between sequences, given a particular scoring matrix and gap penalty. The widely used scoring matrices for the protein sequences include Point Accepted Mutation (PAM) matrices (Schwarz and Dayhoff, 1979) and BLOcks SUbstitution Matrices (BLOSUM) (Altschul, 1991). These matrices quantify similarities between all pairs of amino acids, such that pairs of the same or similar amino acids have high scores while pairs of dissimilar amino acids are associated with low scores. Figure 1 gives one of the most popular BLOSUM62 scoring matrix. This symmetrical matrix has 20 rows and 20 columns. The number in the i^{th} row and j^{th} column is the similarity score for the pair of i^{th} and j^{th} amino acids. The gap penalty is for gaps (openings) that are often inserted into one of the aligned sequences to maximize the similarity score. The gap penalty is larger for a new gap when compared to an extension of an already existing gap. We show examples of alignments with gaps and discuss how to setup the gap penalty values in the "BLASTN Suite" and "Recent Advances" sections. BLAST can be parametrized to use different scoring matrices and gap penalties, resulting in different alignments for the same sequences.



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

Figure 1. The BLOSUM62 scoring matrix.

The input/query sequence and the sequences in the various databases are stored in a common format to expedite the search. The most popular sequence format is FASTA. This format was first introduced in the FASTA program for sequence alignment (Lipman and Pearson, 1985) and was adopted as a standard for representing nucleotide and protein sequences. Virtually all sequences databases support this format. FASTA is a text based format for both protein and nucleotide sequences. It uses multiple lines to store a single sequence. The first line starts with ">" symbol which is followed by comments that typically name and provide information about the sequence; multiple comment lines can be used. The subsequent lines give the one-letter encoded protein sequence. An example FASTA formatted sequence for the human arrestin protein follows:

```
>sp|P36575|ARRC_HUMAN Arrestin-C OS=Homo sapiens
MSKVFKKTSSNGKLSIYLGKRDFVDHVDTVEPIDGVVLVDPEYLKCRKLFVMLTCAFRYG
RDDLEVIGLTFRKDLYVQTLQVVPAESSSPQGPLTVLQERLLHKLGDNAYPFTLQMVTNL
PCSVTLQPGPEDAGKPCGIDFEVKSFCAENPEETVSKRDYVRLVVRKVQFAPPEAGPGPS
AQTIRRFLLSAQPLQLQAWMDREVHYHGEPISVNVSINNCTNKVIKKIKISVDQITDVVL
YSLDKYTKTVFIQEFTETVAANSSFSQSFAVTPILAASCQKRGLALDGKLKHEDTNLASS
TIIRPGMDKELLGILVSYKVRVNLMVSCGGILGDLTASDVGVELPLVLIHPKPSHEAASS
EDIVIEEFTRKGEEESQKAVEAEGDEGS
```

BLASTP suite searches protein sequences against a protein database. It includes five programs: BLASTP, Quick BLASTP, PSI-BLAST (Position-Specific Iterative BLAST), PHI-BLAST (Pattern-Hit Initiated BLAST) and DELTA-BLAST (Domain Enhanced Lookup Time Accelerated BLAST). BLASTP is a generic purpose protein sequence alignment program that compares a query protein sequence to sequences in a specific protein database. Quick BLASTP is an accelerated (faster) version of BLASTP that is specifically coupled with the nr database. PSI-BLAST (Altschul et al., 1997) finds more distant relatives (a wider range of similar sequences) by performing the BLAST search iteratively. It also produces the position specific scoring matrix (PSSM). We explain further details and how to use PSI-BLAST in the next section. PHI-BLAST (Zhang et al., 1998) performs search that is limited to alignments that match a specific pattern in the query sequence. Finally, DELTA-BLAST (Boratyn et al., 2012)s uses pre-computed PSSMs derived from the Conserved Domain Database (Marchler-Bauer et al., 2011) to provide stronger detection of remote homologs.

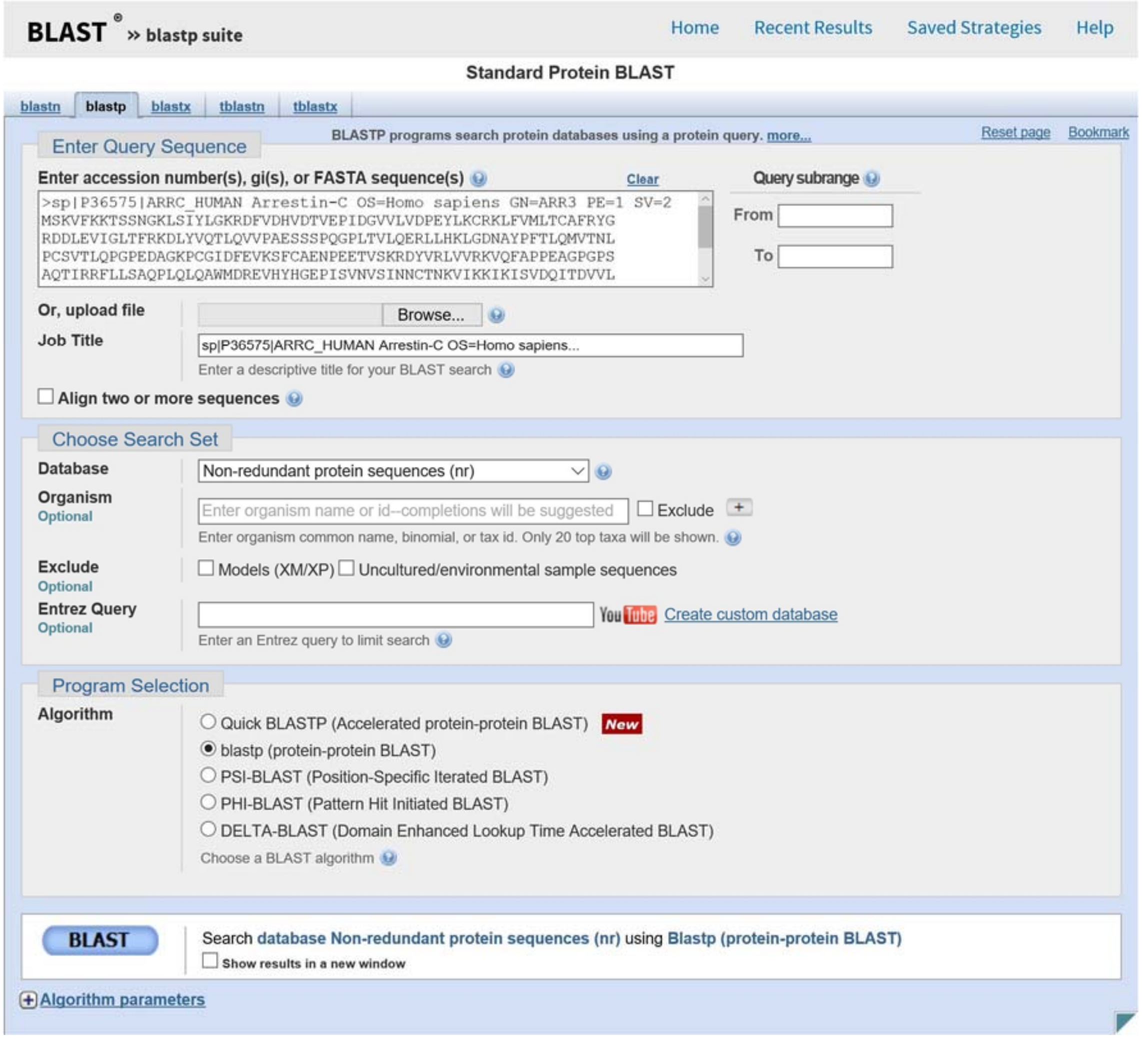
The web interface of the BLASTP suite, which is shown in Figure 2, includes three parts:

(1) "Enter Query Sequence" box facilitates inputting of the query sequence(s). The sequence(s) can be provided either directly using the text box or they can be uploaded from a file. The sequences can be provided either directly in the FASTA format or the query protein can be identified with either accession number or NCBI gi number. The accession number and NCBI gi number can be found in the NCBI databases. User can also opt to align part of the query sequence that is defined with the query subrange. Finally, BLASTP can be used to align two query proteins with each other using the "Align two or more sequences" checkbox.

(2) "Choose Search Set" box allows for selection of the target databases and organisms. Eight databases are currently available including nr, RefSeq, PDB and SWISS-PROT (Bairoch and Apweiler, 2000). List of organisms is optional and it should be used to limit the results to this selection of species.

(3) "Program Selection" box offers choice of the five BLASTP suite programs where BLASP is selected by default.

Figure 2 gives a query that concerns human arrestin protein which is set up to be searched against the nr database using the default BLASTP program.



Standard Protein BLAST

BLASTP programs search protein databases using a protein query. more...

Enter Query Sequence

Enter accession number(s), gi(s), or FASTA sequence(s)

```
>sp|P36575|ARRC_HUMAN Arrestin-C OS=Homo sapiens GN=ARR3 PE=1 SV=2
MSKVFKKTSSNGKLSIYLGKRDFVDHVDTVEPIDGVVLVDPEYLKCRKLFVMLTCAFRYG
RDDLEVIGLTFRKDLYVQTLQVVPAESSSPQGPLTVLQERLLHKLGDNAYPFTLQMVTNL
PCSVTLQPGPEDAGKPCGIDFEVKSFCAENPEETVSKRDYVRLVVRKVQFAPPEAGPGPS
AQTIRRFLLSAQPLQLQAWMDREVHYHGEPISVNVSINNCTNKVIKKIKISVDQITDVVL
```

Query subrange

From

To

Or, upload file

Job Title: sp|P36575|ARRC_HUMAN Arrestin-C OS=Homo sapiens...

Enter a descriptive title for your BLAST search

☐ Align two or more sequences

Choose Search Set

Database: Non-redundant protein sequences (nr)

Organism Optional: Enter organism name or id--completions will be suggested ☐ Exclude

Enter organism common name, binomial, or tax id. Only 20 top taxa will be shown.

Exclude Optional: ☐ Models (XM/XP) ☐ Uncultured/environmental sample sequences

Entrez Query Optional:

Enter an Entrez query to limit search

Program Selection

Algorithm

○ 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)

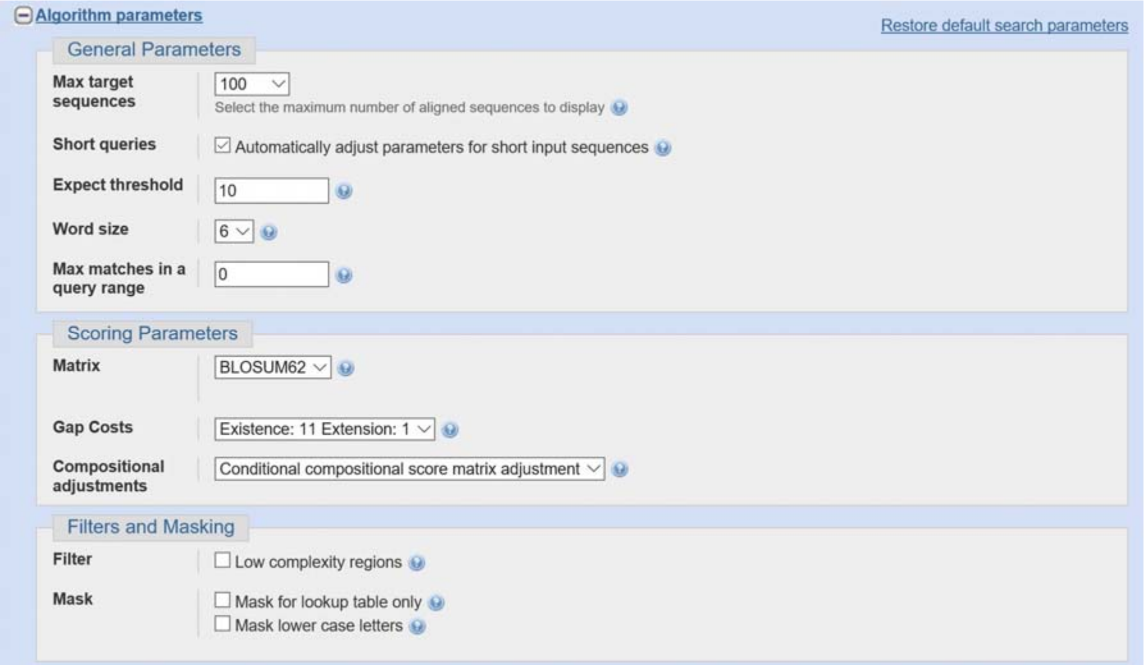
Choose a BLAST algorithm

Search database Non-redundant protein sequences (nr) using Blastp (protein-protein BLAST)

☐ Show results in a new window

Figure 2. Web interface for the BLASTP suite.

The BLAST parameters can be adjusted using a link at the bottom of the web interface page. This is optional and in most cases BLAST is used with pre-selected default parameters, which are shown in Figure 3. Using the parameters menu users can restrict the maximal number of aligned (similar) target sequences (default value is 100) and can set the maximal E-value. The E-value quantifies statistical significance of the similarity where lower value corresponds to more significant similarity; the default value is 10, which means that about 10 of the similar sequences are expected to be found by chance (Karlin and Altschul, 1990). Users can also influence key parameters that govern the sequence similarity search. BLAST uses a filter-based technique to speed up the search for similar sequences. Instead of aligning the sequences directly it finds word matches between the query and database sequences (filter step) and then it extends each word match to a gapped alignment (alignment step). Use of smaller word sizes makes the search more sensitive. The default word size for proteins is set to 6. The "maximal matches in a query range" parameter limits the number of matches and has the default value of 0, which means that there is no limit. The default scoring matrix is BLOSUM62 (Figure 1) and the default setup of the gap penalties is 11 for opening the gap and 1 for the extension of the gap. Different scoring matrices are designed to detect similarities among sequences that diverge by differing degrees (Altschul, 1991; Altschul, 1993). BLOSUM62 is preferred for the detection of weak similarities (Henikoff and Henikoff, 1992). Other matrices are more suitable for specific scenarios. For instance, BLOSUM45 is designed to detect long and weak alignments (Choudhuri, 2014). The "Filters and Masking" section in Figure 3 allows for masking low complexity regions in the protein sequences when the top option is checked. This means that the low complexity regions will not be used in the alignment. If the second option "Mask for lookup table only" is checked then the low complexity regions will be masked only for constructing the lookup table or words list that are used in the filter step while the alignment step is performed without masking. The third option masks lower case letters in the query FASTA sequence; this allows user to choose which parts of the sequence should not be aligned.



General Parameters

Max target sequences: 100
Select the maximum number of aligned sequences to display

Short queries: ☑ Automatically adjust parameters for short input sequences

Expect threshold: 10

Word size: 6

Max matches in a query range: 0

Scoring Parameters

Matrix: BLOSUM62

Gap Costs: Existence: 11 Extension: 1

Compositional adjustments: Conditional compositional score matrix adjustment

Filters and Masking

Filter: □ Low complexity regions

Mask: □ Mask for lookup table only
□ Mask lower case letters

Figure 3. Parameters of the BLASTP suite.

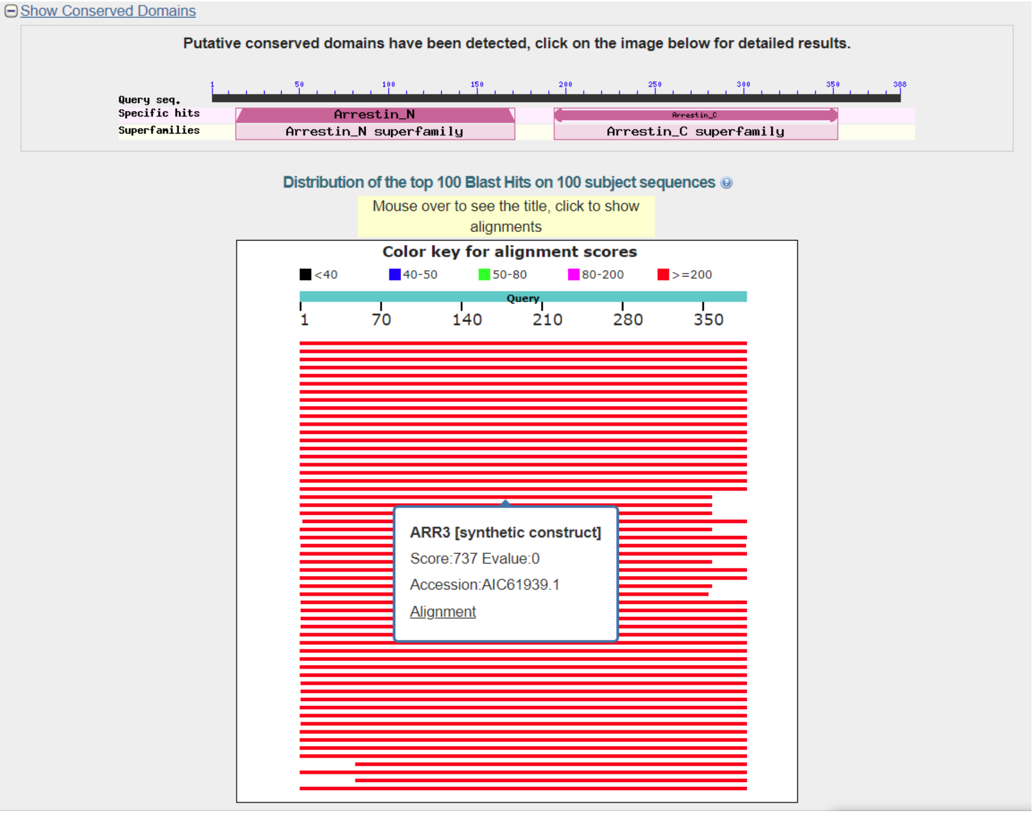
We run the query shown in Figures 2 and 3 and the results are summarized in Figures 4, 5 and 6. The webpage that gives the search results is divided into four parts:

(1) General information that is located at the top includes query identifier and description, target database name and description as well as a detailed summary of key search parameters. These parameters include word size, E-value, gap penalties, scoring matrix, number of sequences scanned, date and time of the request, and about a dozen of other statistics.

(2) "Graphic Summary" box provides a simplified, graphical representation of the alignment of the query sequence with the top scoring similar sequences (Figure 4). The summary includes the alignment score and E-value for each aligned sequence. Higher alignment score implies higher degree of similarity between the query sequence and a given hit sequence. The E-value quantifies the number of hits that are expected to happen by chance. The multiple alignment for several top hits for our query protein is given in Figure 5. The conserved regions are shown in red. The multiple alignment can be downloaded using a wide variety of formats including FASTA, clustal, phylip, nexus and ASN.1. The "Graphic Summary" box also provides putative conserved domains that can be used to classify the query protein (see the top part of Figure 4). Users can find further details about these conserved domains by clicking on the image. This opens a new webpage and gives a list of domain hits with their sequences and Pfam or smart accession numbers. We find three hits for the arrestin example, each with 2 domains (Figure 6). The first domain is found in the 14 to 171 sequence segment in the query sequence chain and it corresponds to Pfam id 00339 (Arrestin N-terminal domain). The second domain stretches between positions 193 to 353 and concerns Pfam id 02752 (Arrestin C-terminal domain) or alternatively smart id 01017.

(3) "Descriptions" box provides additional details for each aligned sequence. These details include it's accession number, alignment score, E-value, query coverage and identity with the query protein that is expressed in percentage points. The query coverage is the fraction of the query length that is included in the aligned segments. The aligned sequences are arranged by their E-values in the ascending order.

(4) "Alignments" box provides side-by-side alignments between the query sequence and the top hits.



⊟ Show Conserved Domains

Putative conserved domains have been detected, click on the image below for detailed results.

Distribution of the top 100 Blast Hits on 100 subject sequences

Mouse over to see the title, click to show alignments

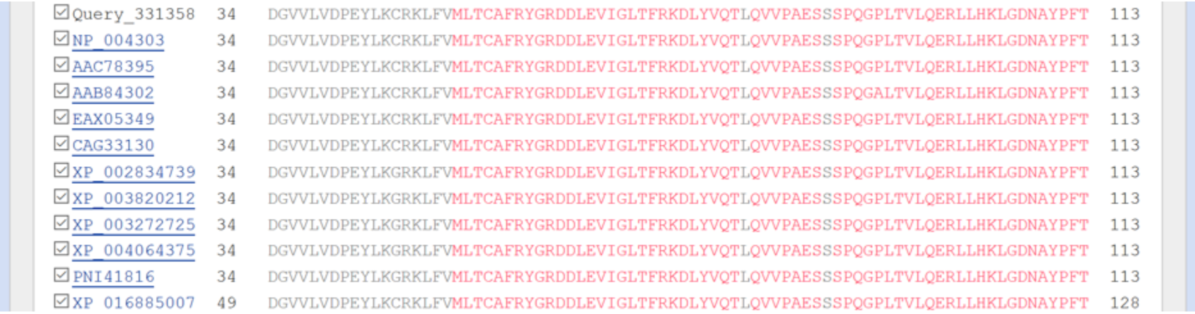
ARR3 [synthetic construct]

Score:737 Evalue:0

Accession:AIC61939.1

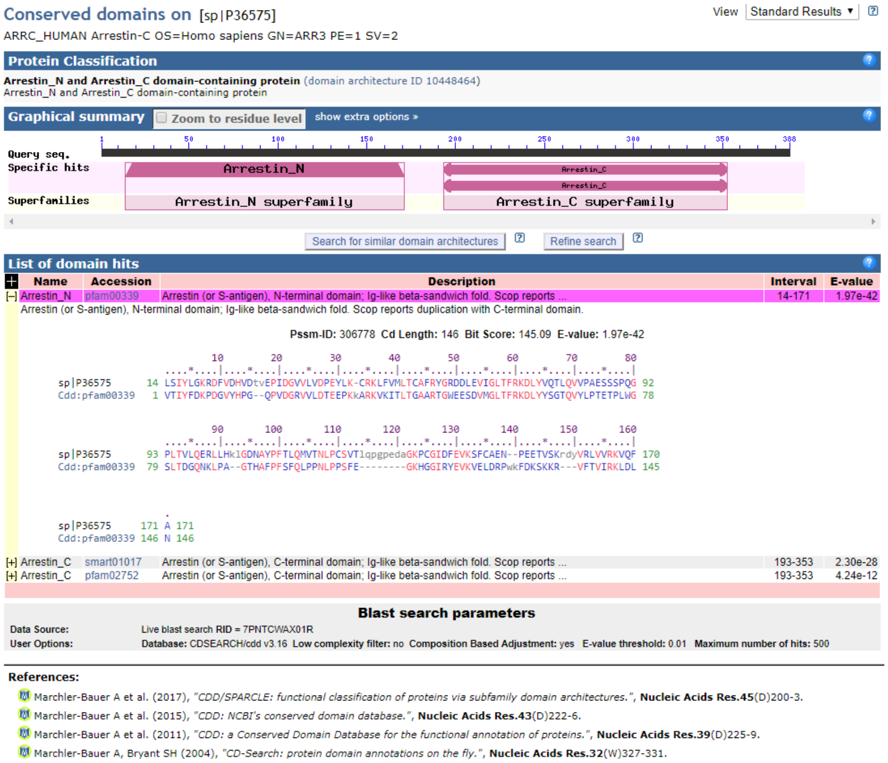
Alignment

Figure 4. Graphic summary box generated by the BLASTP suite for human arrestin protein.



```
☑Query_331358   34    DGVVLVDPEYLKCRKLFVMLTCAFRYGRDDLEVIGLTFRKDLYVQTLQVVPAESSSPQGPLTVLQERLLHKLGDNAYPFT   113
☑NP_004303      34    DGVVLVDPEYLKCRKLFVMLTCAFRYGRDDLEVIGLTFRKDLYVQTLQVVPAESSSPQGPLTVLQERLLHKLGDNAYPFT   113
☑AAC78395       34    DGVVLVDPEYLKCRKLFVMLTCAFRYGRDDLEVIGLTFRKDLYVQTLQVVPAESSSPQGPLTVLQERLLHKLGDNAYPFT   113
☑AAB84302       34    DGVVLVDPEYLKCRKLFVMLTCAFRYGRDDLEVIGLTFRKDLYVQTLQVVPAESSSPQGALTVLQERLLHKLGDNAYPFT   113
☑EAX05349       34    DGVVLVDPEYLKCRKLFVMLTCAFRYGRDDLEVIGLTFRKDLYVQTLQVVPAESSSPQGPLTVLQERLLHKLGDNAYPFT   113
☑CAG33130       34    DGVVLVDPEYLKCRKLFVMLTCAFRYGRDDLEVIGLTFRKDLYVQTLQVVPAESSSPQGPLTVLQERLLHKLGDNAYPFT   113
☑XP_002834739   34    DGVVLVDPEYLKGRKLFVMLTCAFRYGRDDLEVIGLTFRKDLYVQTLQVVPAESSSPQGPLTVLQERLLHKLGDNAYPFT   113
☑XP_003820212   34    DGVVLVDPEYLKGRKLFVMLTCAFRYGRDDLEVIGLTFRKDLYVQTLQVVPAESSSPQGPLTVLQERLLHKLGDNAYPFT   113
☑XP_003272725   34    DGVVLVDPEYLKGRKLFVMLTCAFRYGRDDLEVIGLTFRKDLYVQTLQVVPAESSSPQGPLTVLQERLLHKLGDNAYPFT   113
☑XP_004064375   34    DGVVLVDPEYLKGRKLFVMLTCAFRYGRDDLEVIGLTFRKDLYVQTLQVVPAESSSPQGPLTVLQERLLHKLGDNAYPFT   113
☑PNI41816       34    DGVVLVDPEYLKGRKLFVMLTCAFRYGRDDLEVIGLTFRKDLYVQTLQVVPAESSSPQGPLTVLQERLLHKLGDNAYPFT   113
☑XP_016885007   49    DGVVLVDPEYLKCRKLFVMLTCAFRYGRDDLEVIGLTFRKDLYVQTLQVVPAESSSPQGPLTVLQERLLHKLGDNAYPFT   128
```

Figure 5. Multiple alignment generated by the BLASTP suite for the human arrestin proteins. The first column lists the accession identifiers of several top hit sequences. The second and last column are positions of sequences. The third column is the alignment where conserved regions are shown in red.



Conserved domains on [sp|P36575]

ARRC_HUMAN Arrestin-C OS=Homo sapiens GN=ARR3 PE=1 SV=2

Protein Classification

Arrestin_N and Arrestin_C domain-containing protein (domain architecture ID 10448464)
Arrestin_N and Arrestin_C domain-containing protein

List of domain hits

	Name	Accession	Description	Interval	E-value
[–]	Arrestin_N	pfam00339	Arrestin (or S-antigen), N-terminal domain; Ig-like beta-sandwich fold. Scop reports ...	14-171	1.97e-42

Arrestin (or S-antigen), N-terminal domain; Ig-like beta-sandwich fold. Scop reports duplication with C-terminal domain.

Pssm-ID: 306778 **Cd Length:** 146 **Bit Score:** 145.09 **E-value:** 1.97e-42

```
                          10        20        30        40        50        60        70        80
                  ....*....|....*....|....*....|....*....|....*....|....*....|....*....|....*....|
sp|P36575      14 LSIYLGKRDFVDHVDtvEPIDGVVLVDPEYLK-CRKLFVMLTCAFRYGRDDLEVIGLTFRKDLYVQTLQVVPAESSSPQG 92
Cdd:pfam00339   1 VTIYFDKPDGVYHPG--QPVDGRVVLDTEEPKkARKVKITLTGAARTGWEESDVMGLTFRKDLYYSGTQVYLPTETPLWG 78

                          90       100       110       120       130       140       150       160
                  ....*....|....*....|....*....|....*....|....*....|....*....|....*....|....*....|
sp|P36575      93 PLTVLQERLLHklGDNAYPFTLQMVTNLPCSVTlqpgpedaGKPCGIDFEVKSFCAEN--PEETVSKrdyVRLVVRKVQF 170
Cdd:pfam00339  79 SLTDGQNKLPA--GTHAFPFSFQLPPNLPPSFE--------GKHGGIRYEVKVELDRPwkFDKSKKR---VFTVIRKLDL 145


                  .
sp|P36575     171 A 171
Cdd:pfam00339 146 N 146
```

	Name	Accession	Description	Interval	E-value
[+]	Arrestin_C	smart01017	Arrestin (or S-antigen), C-terminal domain; Ig-like beta-sandwich fold. Scop reports ...	193-353	2.30e-28
[+]	Arrestin_C	pfam02752	Arrestin (or S-antigen), C-terminal domain; Ig-like beta-sandwich fold. Scop reports ...	193-353	4.24e-12

Blast search parameters

Data Source: Live blast search RID = 7PNTCWAX01R
User Options: **Database:** CDSEARCH/cdd v3.16 **Low complexity filter:** no **Composition Based Adjustment:** yes **E-value threshold:** 0.01 **Maximum number of hits:** 500

References:

- Marchler-Bauer A et al. (2017), *"CDD/SPARCLE: functional classification of proteins via subfamily domain architectures."*, **Nucleic Acids Res.45**(D)200-3.
- Marchler-Bauer A et al. (2015), *"CDD: NCBI's conserved domain database."*, **Nucleic Acids Res.43**(D)222-6.
- Marchler-Bauer A et al. (2011), *"CDD: a Conserved Domain Database for the functional annotation of proteins."*, **Nucleic Acids Res.39**(D)225-9.
- Marchler-Bauer A, Bryant SH (2004), *"CD-Search: protein domain annotations on the fly."*, **Nucleic Acids Res.32**(W)327-331.

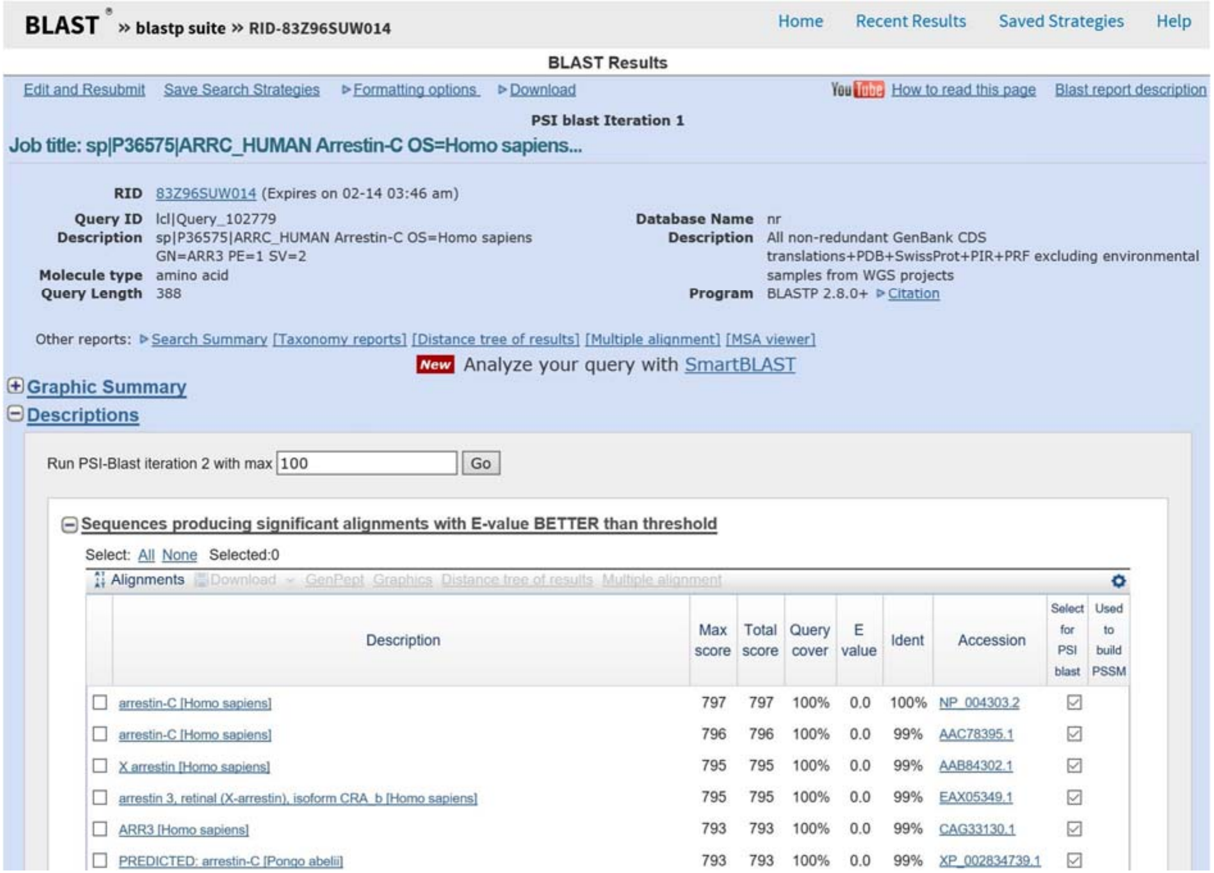
Figure 6. Conserved domains found by BLASTP suite for the human arrestin protein. These results include accession numbers of the domains, the alignment between the query sequence and domain sequence, and the corresponding alignment scores and E-values.

PSI-BLAST

Compared to BLASTP, PSI-BLAST aims to find distant relatives for the query protein, i.e., proteins that are similar to proteins that share similarity with the query sequence. To do that, PSI-BLAST performs the search iteratively. In the first iteration, a list of similar sequences is created using BLASTP. The resulting multiple alignment is used to compute PSSM which is used to substitute the original scoring matrix from the first iteration. This produces a larger and more diverse group of similar proteins. Next, the alignment is run again and the PSSM is recomputed based on the subsequent multiple alignment. The search performed based on PSSM is more sensitive than the search based on the default BLOSUM or PAM matrices since PSSM is based on the multiple alignment that is specific to the query protein. This two-step process can be repeated multiple times.

PSSM is an N by 20 matrix (N is the length of the query sequence) which is calculated from the frequencies of residues at specific positions in the multiple alignment. The scores at the position j in the query sequence (j^{th} row of PSSM) and the amino acid i (i^{th} column of PSSM) are defined as $\log(f_{ij}/q_i)$ where f_{ij} is the frequency of amino acid i at position j in the multiple alignment and q_i is the expected relative frequency of amino acid i. Positive scores mean that a given amino acid is more likely than expected to appear at the specific position. Besides being used for alignment, PSSM is also used to calculate evolutionary conservation of the query sequence (Valdar, 2002) and to predict protein secondary structure (Jones, 1999; Meng and Kurgan, 2016; Wang et al., 2016), solvent accessibility (Heffernan et al., 2015; Magnan and Baldi, 2014), functional sites (Yan et al., 2016; Zhang and Kurgan, 2017; Zhang et al., 2017), and intrinsic disorder (Disfani et al., 2012; Meng et al., 2017; Mizianty et al., 2010).

PSI-BLAST includes three extra parameters on the top of the parameters that are available for BLAST. They appear below the "Filters and Masking" box (Figure 3). User can provide his/her own PSSM, specify the PSI-BLAST threshold that determines statistical significance of hits that are included in the computation of the PSSM for the next iteration, and define value of the pseudocount parameter. The default value of the threshold is set to 0.005 and we recommended this value as a reasonably good choice. The hits with the E-values that are lower than the threshold are included in calculations for the next iteration. The pseudocount value is used to initialize the frequencies that otherwise would be set to zero. The default value of pseudocount it based on the minimum length description principle (Altschul et al., 2009).



PSI blast Iteration 1

Job title: sp|P36575|ARRC_HUMAN Arrestin-C OS=Homo sapiens...

RID 83Z96SUW014 (Expires on 02-14 03:46 am)

Query ID lcl|Query_102779
Description sp|P36575|ARRC_HUMAN Arrestin-C OS=Homo sapiens GN=ARR3 PE=1 SV=2
Molecule type amino acid
Query Length 388

Database Name nr
Description All non-redundant GenBank CDS translations+PDB+SwissProt+PIR+PRF excluding environmental samples from WGS projects
Program BLASTP 2.8.0+ ▷Citation

Other reports: ▷Search Summary [Taxonomy reports] [Distance tree of results] [Multiple alignment] [MSA viewer]
New Analyze your query with SmartBLAST
⊕Graphic Summary
⊖Descriptions

Run PSI-Blast iteration 2 with max 100 Go

⊖Sequences producing significant alignments with E-value BETTER than threshold

Select: All None Selected:0

Alignments Download GenPept Graphics Distance tree of results Multiple alignment

	Description	Max score	Total score	Query cover	E value	Ident	Accession	Select for PSI blast	Used to build PSSM
□	arrestin-C [Homo sapiens]	797	797	100%	0.0	100%	NP_004303.2	☑	
□	arrestin-C [Homo sapiens]	796	796	100%	0.0	99%	AAC78395.1	☑	
□	X arrestin [Homo sapiens]	795	795	100%	0.0	99%	AAB84302.1	☑	
□	arrestin 3, retinal (X-arrestin), isoform CRA_b [Homo sapiens]	795	795	100%	0.0	99%	EAX05349.1	☑	
□	ARR3 [Homo sapiens]	793	793	100%	0.0	99%	CAG33130.1	☑	
□	PREDICTED: arrestin-C [Pongo abelii]	793	793	100%	0.0	99%	XP_002834739.1	☑	

Figure 7. Output produced by the first iteration of PSI-BLAST for the human arrestin protein. The output is similar to the output from BLASTP except for the Descriptions box where the user can select sequences to generate PSSM for next iteration.

Figure 7 shows the results of the first iteration of PSI-BLAST for the sample query protein, the human arrestin, with the PSI-BLAST threshold = 0.001 and the pseudocount set to the default value. Similar to BLASTP, the corresponding webpage includes the general information (summary of search parameters) and the Graphic Summary, Descriptions and Alignments boxes. One difference from the output of BLASTP is that the Descriptions box allows for selection of the aligned sequences that will be used to generate PSSM for the next iteration. The aligned sequences that score below the threshold from the previous iteration are marked in yellow. Clicking the “Go” button (middle of Figure 7) results in the computation of the next iteration (iteration 2) of PSI-BLAST. The search typically converges (no new sequences below threshold are found) after several iterations. The search for the human arrestin protein converges after just four iterations. The results for every iteration include the multiple alignment and PSSM; except for the first iteration when PSSM is not yet available. They can be downloaded using a link at the top of the webpage. The PSSM is formatted based on the ASN.1 format from the BLASTP suite that can be read by AsnTool from the NCBI’s toolkit. A plain text PSSM is generated by a standalone BLAST+ program. Correspondingly, we give an example plain-text PSSM in next section that describes BLAST+.

BLAST+ - STANDALONE BLAST

BLAST+ must be run on a local computer. It can be installed on the Windows, Linux and MacOS platforms. The installation package and sources can be downloaded from ftp://ftp.ncbi.nlm.nih.gov/blast/executables/blast+/LATEST/. BLAST+ includes BLASTP, BLASTN, BLASTX, TBLASTN, TBLASTX, PSI-BLAST, DELTA-BLAST, and reverse PSI-BLAST (RPS-BLAST) programs. These programs are the same as the programs available via WEB BLAST, except for RPS-BLAST that searches a query sequence against a database of PSSMs. PHI-BLAST from the WEB BLASTP suite is integrated into PSI-BLAST in BLAST+ and can be accessed with the -phi_pattern parameter. Similarly, MEGABLAST and discontiguous MEGABLAST from the WEB BLASTN suite are integrated into BLASTN in BLAST+ and can be used with the -task parameter.

One advantage of BLAST+ compared to the web-based versions is that users have the choice to use their own databases for the search or to utilize one of dozens of already preprocessed databases that can be downloaded from ftp://ftp.ncbi.nlm.nih.gov/blast/db/. The former choice requires users to provide the sequences in a specific format that is compatible with BLAST. The makeblastdb program that is part of BLAST+ supports this task. First, the sequences must be prepared in the FASTA format. The makeblastdb takes multiple FASTA-formatted files as input and it outputs a properly formatted BLAST database. Here is an example call to run this program:

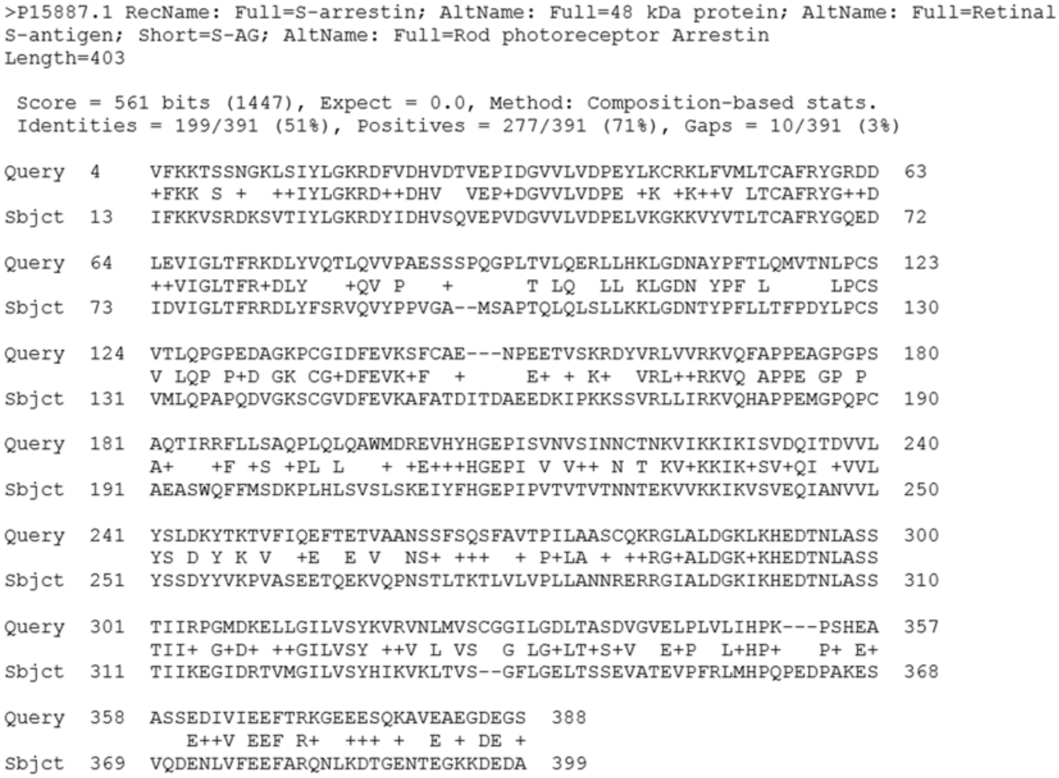
```
makeblastdb –in ‘swiss1.fa swiss2.fa’ –dbtype prot –out swiss
```

This above command creates a BLAST database named “swiss” using swiss1.fa and swiss2.fa FASTA files. The database type (dbtype) need be specified as prot (for proteins) or nucl (for nucleotides).

BLAST+ should be executed from command line. We give an example command to run PSI-BLAST. The query sequence is the same that we use in the WEB BLASTP example, the human arrestin protein, and we use the available SWISS-PROT database as the target database:

```
psiblast -query query.fa –db ./blastdb/swissprot -num_iterations 3 -out result.out -
out_ascii_pssm pssm.out -E-value 0.001
```

The input query.fa (“-query query.fa”) query file is in the FASTA format and includes only the arrestin. We specify the target database as swissprot (“-db ./blastdb/swissprot”). The number of PSI_BLAST iterations is set to 3 (“-num_iterations 3”) and the E-value threshold is specified as 0.001 (“-E-value 0.001”). The results will be saved into the result.out file (“-out_result.out”) and the corresponding PSSM will be stored in the “pssm.out” file (“-out_ascii_pssm pssm.out”). The command line also allows to specify values of other parameters, such as the maximal number of sequences used to generate PSSM, scoring matrix, gap penalty, word size, filters, etc. The default values are used if they are not specified. We recommend the documentation available at https://www.ncbi.nlm.nih.gov/books/NBK279690/ for further reading on how to setup and use BLAST+.



```
>P15887.1 RecName: Full=S-arrestin; AltName: Full=48 kDa protein; AltName: Full=Retinal
S-antigen; Short=S-AG; AltName: Full=Rod photoreceptor Arrestin
Length=403

 Score = 561 bits (1447), Expect = 0.0, Method: Composition-based stats.
 Identities = 199/391 (51%), Positives = 277/391 (71%), Gaps = 10/391 (3%)

Query  4    VFKKTSSNGKLSIYLGKRDFVDHVDTVEPIDGVVLVDPEYLKCRKLFVMLTCAFRYGRDD  63
            +FKK S +  ++IYLGKRD++DHV  VEP+DGVVLVDPE +K +K++V LTCAFRYG++D
Sbjct  13   IFKKVSRDKSVTIYLGKRDYIDHVSQVEPVDGVVLVDPELVKGKKVYVTLTCAFRYGQED  72

Query  64   LEVIGLTFRKDLYVQTLQVVPAESSSPQGPLTVLQERLLHKLGDNAYPFTLQMVTNLPCS  123
            ++VIGLTFR+DLY   +QV P   +      T LQ  LL KLGDN YPF L     LPCS
Sbjct  73   IDVIGLTFRRDLYFSRVQVYPPVGA--MSAPTQLQLSLLKKLGDNTYPFLLTFPDYLPCS  130

Query  124  VTLQPGPEDAGKPCGIDFEVKSFCAE---NPEETVSKRDYVRLVVRKVQFAPPEAGPGPS  180
            V LQP P+D GK CG+DFEVK+F  +     E+ + K+  VRL++RKVQ APPE GP P
Sbjct  131  VMLQPAPQDVGKSCGVDFEVKAFATDITDAEEDKIPKKSSVRLLIRKVQHAPPEMGPQPC  190

Query  181  AQTIRRFLLSAQPLQLQAWMDREVHYHGEPISVNVSINNCTNKVIKKIKISVDQITDVVL  240
            A+   +F +S +PL L   + +E+++HGEPI V V++ N T KV+KKIK+SV+QI +VVL
Sbjct  191  AEASWQFFMSDKPLHLSVSLSKEIYFHGEPIPVTVTVTNNTEKVVKKIKVSVEQIANVVL  250

Query  241  YSLDKYTKTVFIQEFTETVAANSSFSQSFAVTPILAASCQKRGLALDGKLKHEDTNLASS  300
            YS D Y K V  +E  E V  NS+ +++  + P+LA + ++RG+ALDGK+KHEDTNLASS
Sbjct  251  YSSDYYVKPVASEETQEKVQPNSTLTKTLVLVPLLANNRERRGIALDGKIKHEDTNLASS  310

Query  301  TIIRPGMDKELLGILVSYKVRVNLMVSCGGILGDLTASDVGVELPLVLIHPK---PSHEA  357
            TII+ G+D+ ++GILVSY ++V L VS  G LG+LT+S+V  E+P  L+HP+   P+ E+
Sbjct  311  TIIKEGIDRTVMGILVSYHIKVKLTVS--GFLGELTSSEVATEVPFRLMHPQPEDPAKES  368

Query  358  ASSEDIVIEEFTRKGEEESQKAVEAEGDEGS  388
               E++V EEF R+  +++ +  E + DE +
Sbjct  369  VQDENLVFEEFARQNLKDTGENTEGKKDEDA  399
```

Figure 8. Results for an example similar sequence generated by the standalone PSI-BLAST for the query human arrestin protein.

The above example PSI-BLAST command generates two files. The first, result.out, lists 129 sequences that are sufficiently similar to our query, i.e., E-value < 0.001. The file provides detailed information including the accession numbers, multitude of scores that quantify quality of the alignment, and the alignment itself. Sample results for one of the hits are shown in Figure 8. The UniProt identifier of the aligned sequence is P15887.1. The alignment score, which is calculated based on PSSM, is 561 and this sequence shares 51% identical residues with the arrestin chain. The alignment, which is shown below these statistics, organizes the sequences into blocks. The first line in each block is the query sequence and the third line is the aligned/hit sequence. The positions for both sequences are given at the beginning and end of the lines. The sequence in the second line in each block (between the chains of the query and hit) is the match sequence. Letters in the match sequence denote matching residues while "+" means that the residues do not match and that the corresponding score is positive. Moreover, blank space denotes that the score is negative. The second file provides PSSM that is shown in Figure 9. PSSM is formatted as *N* by 20 matrix where *N* is the length of the query sequence. We show the first 13 rows that correspond to the first 13 residues in the arrestin sequence. Each row is one position in the query sequence. The log odds scores in a given row are calculated based on the frequency of the 20 amino acids at this position in the multiple alignment generated in the last iteration of PSI-BLAST.

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

Figure 9. PSSM generated by the standalone PSI-BLAST based on the query for the human arrestin protein. We shows values for the first 13 residues in the arresting sequences. The first two columns are the sequence position and one code amino acid type at this position. The following 20 columns are the scores for every amino acid type at a given position.

BLASTN SUITE - BLAST AGAINST NUCLEOTIDE DATABASES

BLASTN suite searches nucleotide sequence query against a nucleotide database. It includes three programs: MEGABLAST, discontiguous MEGABLAST and BLASTN. MEGABLAST was designed for searching highly similar sequences, such as in case of the intra-species comparisons (Morgulis et al., 2008). Discontiguous MEGABLAST should be used to search more dissimilar sequences, which are typical for cross-species comparisons. Finally, BLASTN is suitable for short queries and cross-species comparisons.

The webpage of the BLASTN suite is shown in Figure 10. Similar to the BLASTP suite, it has three parts:

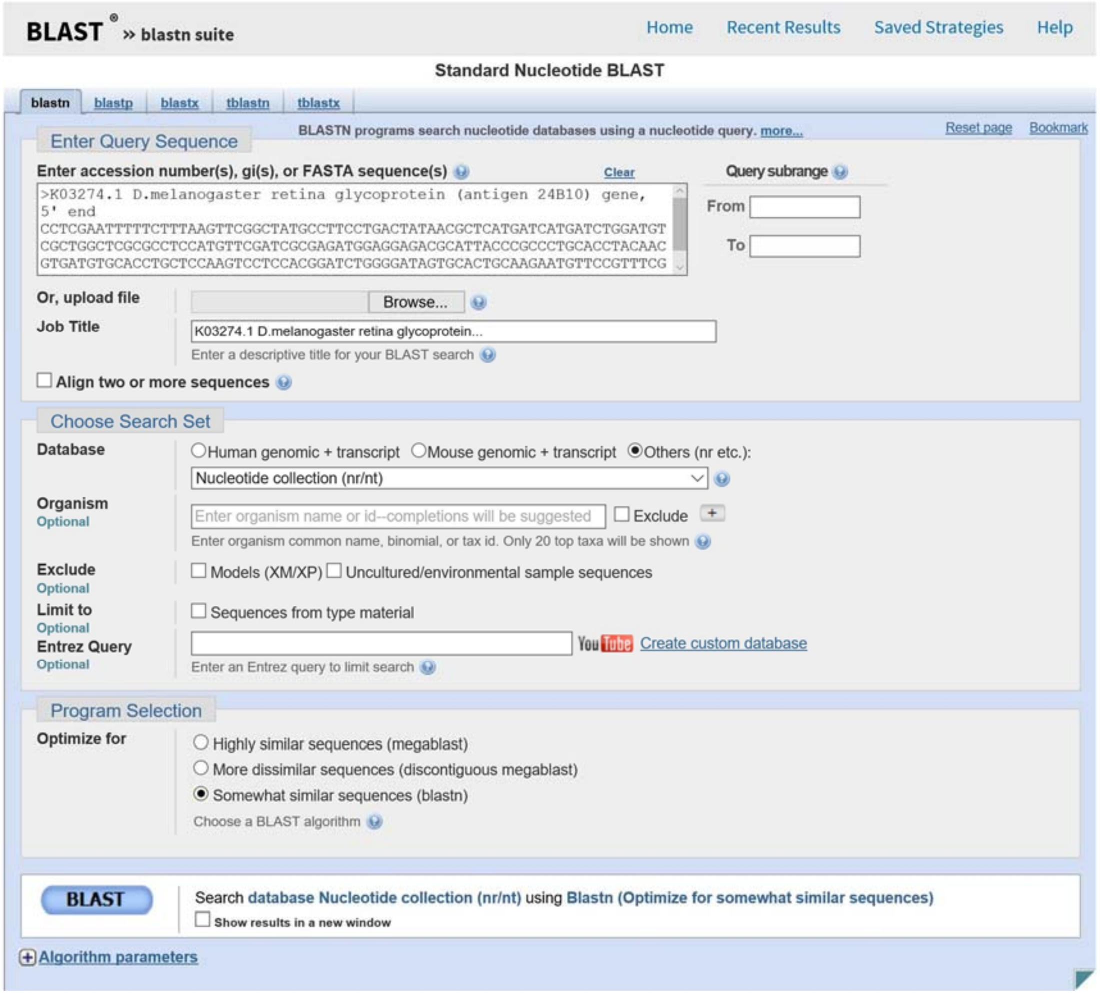
(1) "Enter Query Sequence" box should be used to provide the query sequences in one of the two ways: directly enter query sequence into the text box or upload it in a file. Besides the FASTA-formatted sequences, users can provide database accession number or NCBI gi number. Also, a segment of the query sequence that can be specified in "Query subrange" box may be used for the search. Like for BLASTP, BLASTN suite also allows

aligning sequences provided by the users. To do so, users should check the “Align two or more sequences” checkbox and input their own sequences for the alignment.

(2) “Choose Search Set” box provides selection of the target database or organism. The search can be restricted to a subset of selected database using “Organism” “Exclude” “Limit to” and “Entrez Query” options.

(3) “Program Selection” box should be used to select one of the three program: MEGABLAST, discontiguous MEGABLAST and BLASTN.

The “Algorithm parameters” link at the bottom of the page opens additional menu where users can choose the scoring matrix and can set values of the gap penalties and other search parameters.



Standard Nucleotide BLAST

BLASTN programs search nucleotide databases using a nucleotide query. more...

Enter Query Sequence

Enter accession number(s), gi(s), or FASTA sequence(s)

```
>K03274.1 D.melanogaster retina glycoprotein (antigen 24B10) gene,
5' end
CCTCGAATTTTTCTTTAAGTTCGGCTATGCCTTCCTGACTATAACGCTCATGATCATGATCTGGATGT
CGCTGGCTCGCGCCTCCATGTTCGATCGCGAGATGGAGGAGACGCATTACCCGCCCTGCACCTACAAC
GTGATGTGCACCTGCTCCAAGTCCTCCACGGATCTGGGGATAGTGCACTGCAAGAATGTTCCGTTTCG
```

Query subrange
From
To

Or, upload file

Job Title: K03274.1 D.melanogaster retina glycoprotein...

Enter a descriptive title for your BLAST search

□ Align two or more sequences

Choose Search Set

Database: ○Human genomic + transcript ○Mouse genomic + transcript ●Others (nr etc.):

Nucleotide collection (nr/nt)

Organism Optional: Enter organism name or id--completions will be suggested □ Exclude

Enter organism common name, binomial, or tax id. Only 20 top taxa will be shown

Exclude Optional: □ Models (XM/XP) □ Uncultured/environmental sample sequences

Limit to Optional: □ Sequences from type material

Entrez Query Optional: You Tube Create custom database

Enter an Entrez query to limit search

Program Selection

Optimize for

○ Highly similar sequences (megablast)

○ More dissimilar sequences (discontiguous megablast)

● Somewhat similar sequences (blastn)

Choose a BLAST algorithm

Search database Nucleotide collection (nr/nt) using Blastn (Optimize for somewhat similar sequences)

□ Show results in a new window

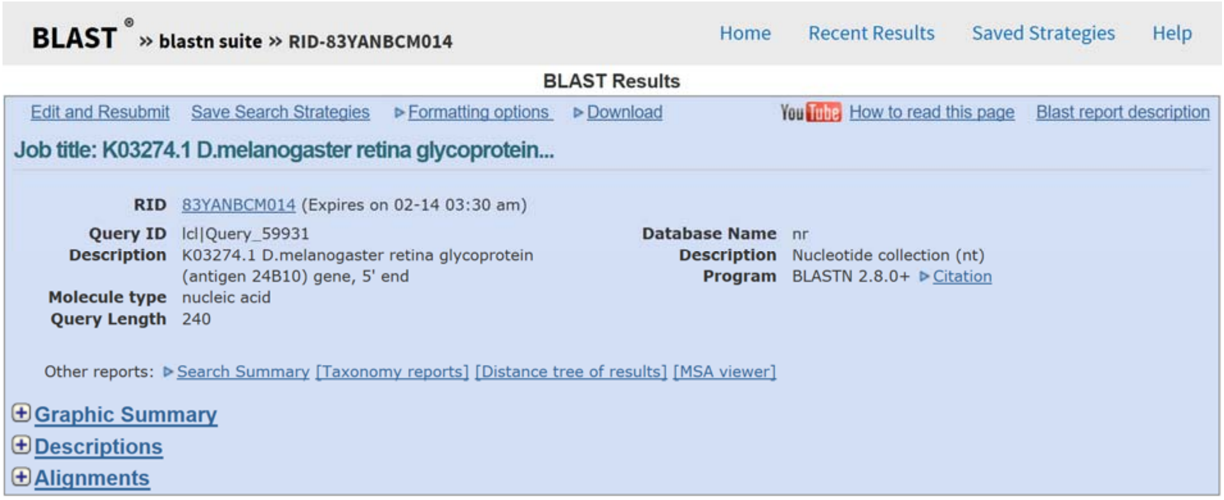
Figure 10. Web interface for the BLASTN suite.

We run an example query for the retina glycoprotein gene from *Drosophila Melanogaster* using the nr database and BLASTN. We use the default settings for the algorithm. The output of BLASTN is formatted the same way as the output of BLASTP. At the top of the page it gives general information (Figure 11) that includes the target database name, query sequence identifier and length, type and description of the query molecule as well as details of the BLASTN parameters. This is followed by:

(1) "Graphic Summary" box shown in Figure 12 that provides a graphical representation of the alignments of the 79 sequence hits. Each hit is shown as a colored bar where the color represents the alignment score. Detailed information for a specific hit will be shown when the corresponding bar is clicked. It includes description of the hit sequence, alignment score and E-value. The alignment scores quantify the degree of similarity while the E-values quantify number of hits one can expect to observe by chance.

(2) "Descriptions" box shown in Figure 13 provides detailed statistics and database accession identifiers for each hit. The hits are sorted by E-value in the ascending order. Besides the E-values and alignment scores, this box also gives the maximal score, query coverage and identity. The maximal score is the highest alignment score among the aligned segments between the query and hit sequences. The query coverage is the fraction of the query sequence length that is covered by the aligned segments. The identity gives the % of the aligned segments where the two sequences match. The best hit with the 100% query coverage and 100% identity in Figure 13 is the exact same sequences as our query sequence. The following hits are the sequences of the chromosome 3R from *Drosophila Melanogaster* which localize our query.

(3) "Alignments" box gives detailed alignments along the query sequence. Figure 14 shows an example alignment for one of the hits for the retina glycoprotein gene query. This alignment contains one segment in the 7 to 242 range of the query sequence. Vertical lines between query sequence (first line in each alignment block) and hit sequence (third/subject line in each alignment block) correspond to matched identical nucleotides. In this example there are 197 identities and one gap in the middle of the last block.

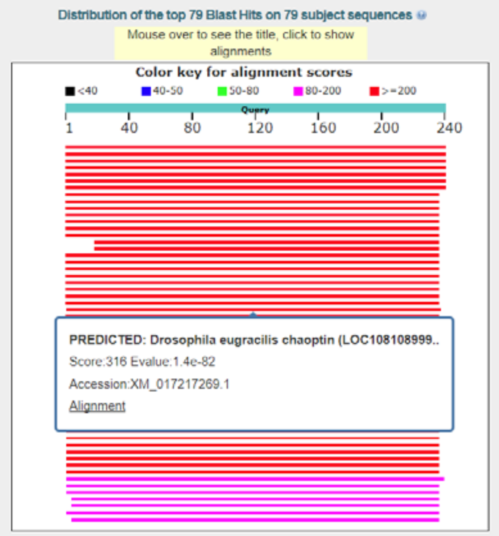


Job title: K03274.1 D.melanogaster retina glycoprotein...

RID	83YANBCM014 (Expires on 02-14 03:30 am)		
Query ID	lcl|Query_59931	**Database Name**	nr
Description	K03274.1 D.melanogaster retina glycoprotein (antigen 24B10) gene, 5' end	**Description**	Nucleotide collection (nt)
Molecule type	nucleic acid	**Program**	BLASTN 2.8.0+ ▷Citation
Query Length	240		

Other reports: ▷Search Summary [Taxonomy reports] [Distance tree of results] [MSA viewer]
⊕Graphic Summary
⊕Descriptions
⊕Alignments

Figure 11. Summary information that is shown at the top of the page with the results generated by the BLASTN suite for the retina glycoprotein gene.



Distribution of the top 79 Blast Hits on 79 subject sequences

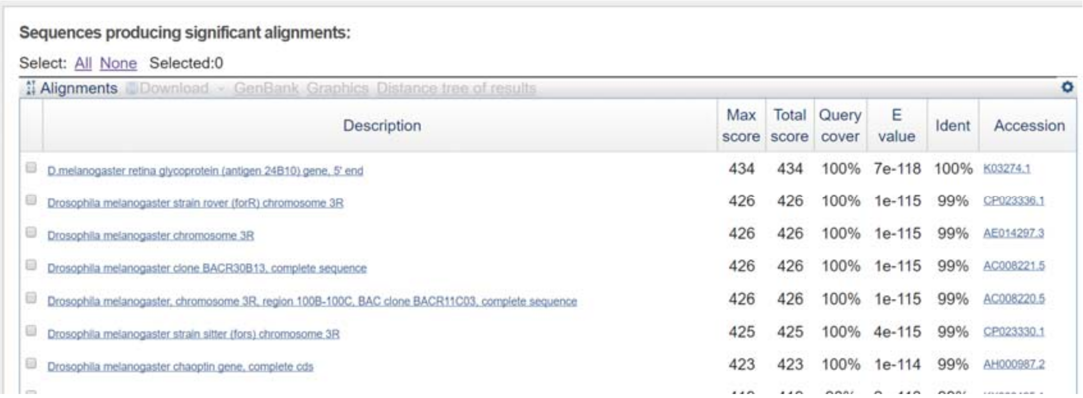
Mouse over to see the title, click to show alignments

Color key for alignment scores

■ <40 ■ 40-50 ■ 50-80 ■ 80-200 ■ >=200

PREDICTED: Drosophila eugracilis chaoptin (LOC108108999..
Score:316 Evalue:1.4e-82
Accession:XM_017217269.1
Alignment

Figure 12. Graphic summary box generated by the BLASTN suite for the retina glycoprotein gene.



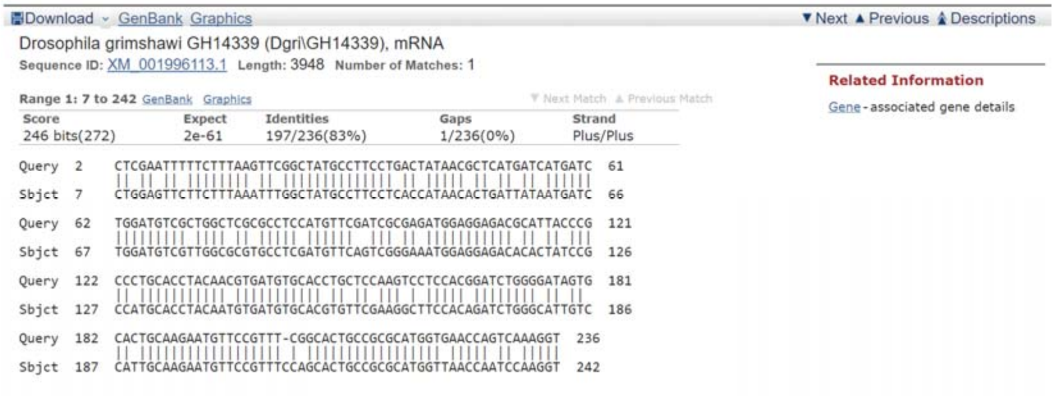
Sequences producing significant alignments:

Select: All None Selected:0

Alignments Download GenBank Graphics Distance tree of results

	Description	Max score	Total score	Query cover	E value	Ident	Accession
□	D.melanogaster retina glycoprotein (antigen 24B10) gene, 5' end	434	434	100%	7e-118	100%	K03274.1
□	Drosophila melanogaster strain rover (forR) chromosome 3R	426	426	100%	1e-115	99%	CP023336.1
□	Drosophila melanogaster chromosome 3R	426	426	100%	1e-115	99%	AE014297.3
□	Drosophila melanogaster clone BACR30B13, complete sequence	426	426	100%	1e-115	99%	AC008221.5
□	Drosophila melanogaster, chromosome 3R, region 100B-100C, BAC clone BACR11C03, complete sequence	426	426	100%	1e-115	99%	AC008220.5
□	Drosophila melanogaster strain sitter (fors) chromosome 3R	425	425	100%	4e-115	99%	CP023330.1
□	Drosophila melanogaster chaoptin gene, complete cds	423	423	100%	1e-114	99%	AH000987.2

Figure 13. Part of the description box output by the BLASTN suite for the retina glycoprotein gene.



Download GenBank Graphics
▼Next ▲Previous Descriptions

Drosophila grimshawi GH14339 (Dgri\GH14339), mRNA
Sequence ID: XM_001996113.1 Length: 3948 Number of Matches: 1

Related Information
Gene - associated gene details

Range 1: 7 to 242 GenBank Graphics

▼ Next Match ▲ Previous Match

```
Score           Expect   Identities       Gaps         Strand
246 bits(272)   2e-61    197/236(83%)     1/236(0%)    Plus/Plus

Query  2    CTCGAATTTTTCTTTAAGTTCGGCTATGCCTTCCTGACTATAACGCTCATGATCATGATC  61
            || || || |||||||| || |||||||||||||| || ||||| || || || ||||||
Sbjct  7    CTGGAGTTCTTCTTTAAATTTGGCTATGCCTTCCTCACCATAACACTGATTATAATGATC  66

Query  62   TGGATGTCGCTGGCTCGCGCCTCCATGTTCGATCGCGAGATGGAGGAGACGCATTACCCG  121
            ||||||||| |||| || ||||| ||||||  ||| || ||||||||||| || || |||
Sbjct  67   TGGATGTCGTTGGCGCGTGCCTCGATGTTCAGTCGGGAAATGGAGGAGACACACTATCCG  126

Query  122  CCCTGCACCTACAACGTGATGTGCACCTGCTCCAAGTCCTCCACGGATCTGGGGATAGTG  181
            || ||||||||||| ||||||||||| || || ||| | ||||| |||||||| || || 
Sbjct  127  CCATGCACCTACAATGTGATGTGCACGTGTTCGAAGGCTTCCACAGATCTGGGCATTGTC  186

Query  182  CACTGCAAGAATGTTCCGTTT-CGGCACTGCCGCGCATGGTGAACCAGTCAAAGGT  236
            || |||||||||||||||||| | ||||||||||||||||| ||||| || |||||
Sbjct  187  CATTGCAAGAATGTTCCGTTTCCAGCACTGCCGCGCATGGTTAACCAATCCAAGGT  242
```

Figure 14. An example alignment produced by the BLASTN suite for the retina glycoprotein gene.

RECENT ADVANCES

Recent developments in high throughput sequencing technologies resulted in the generation of massive datasets and a rapid growth of the sequence databases. For instance, RefSeq has grown from one million records in 2003, to 10 million in 2009 and 150 million in 2018 (https://www.ncbi.nlm.nih.gov/refseq/statistics/). Corresponding, searching for similar sequences becomes more time-consuming. Several tools that are faster than the popular BLAST were developed in recent years to address this issue. They include UBLAST (Edgar, 2010), RAPSearch2 (Zhao et al., 2012), Diamond (Buchfink et al., 2015), MMSeq (Hauser et al., 2016), and MMseqs2 (Steinegger and Soding, 2017). Here we focus on the most recent MMseqs2 algorithm.

MMseqs2 (Many-against-Many sequence searching) is specifically designed to search and cluster huge protein sequence datasets. According to published empirical results MMseqs2 is 10,000 times faster than BLAST, and at 100 times speedup it achieves the same sensitivity as BLAST (Steinegger and Soding, 2017). MMSeq2 software can be freely downloaded from https://github.com/soedinglab/MMseqs2. Users must prepare their database before searching with MMSeq2. Here, we use SWISS-PROT as the target database, like we did with BLAST. We downloaded the SWISS-PROT sequences and generated the corresponding sequence database for MMseqs2 as follows:

```
mmseqs createdb swissprot.fa swissprot
```

The swissprot.fa file includes the FASTA-formatted SWISS-PROT sequences. The above command generates the MMseqs2 database named swissprot. Unlike for BLAST+, the query sequences must be also converted into the MMseqs2's database format. The corresponding command follows:

```
mmseqs createdb arrestin.fa arrestin
```

The assrestin.fa file includes the sequence of the human arrestin proteins that we also used with BLAST. MMseqs2 can also accept multiple sequences as the input.

MMseqs2 includes five workflows:

(1) mmseqs search for searching query sequences against the target database
(2) mmseqs cluster for clustering sequences by similarity
(3) mmseqs linclust which is a less sensitive but faster program for clustering sequences by similarity
(4) mmseqs clusterupdate for incremental updating of an existing clustering
(5) mmseqs taxonomy for assignment of the query chain to a taxonomy by computing the lowest common ancestor

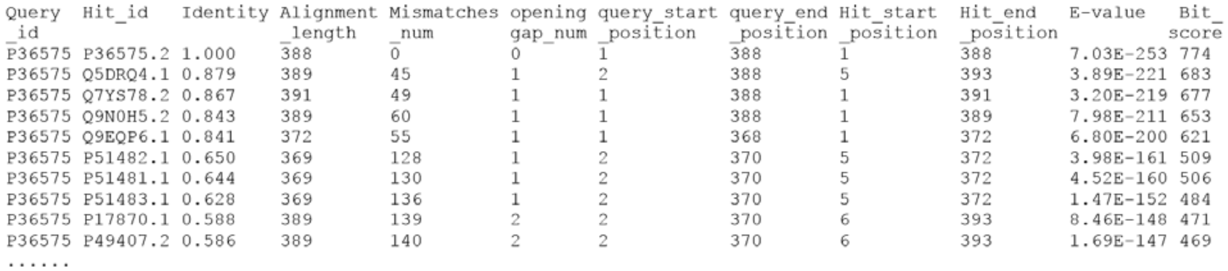
We focus our example on the mmseqs search that is equivalent to the BLASTP search:

```
mmseqs search arrestin swissprot result.out tmp –e 0.0005 -a
```

The human arrestin is our query, SWISS-PROT is the target database, and the search results are saved in the result.out file. The "tmp" is directory where temporary files are saved and "-e" parameter should be used to specify the threshold of the E-values; for our example only the sequences with E-value < 0.0005 will be listed. Finally, the "-a" option specifies that alignment backtraces should be used. Total of 40 hits are found in SWISS-PROT for the human arrestin. The output result.out file can be converted into BLAST output format using the following command:

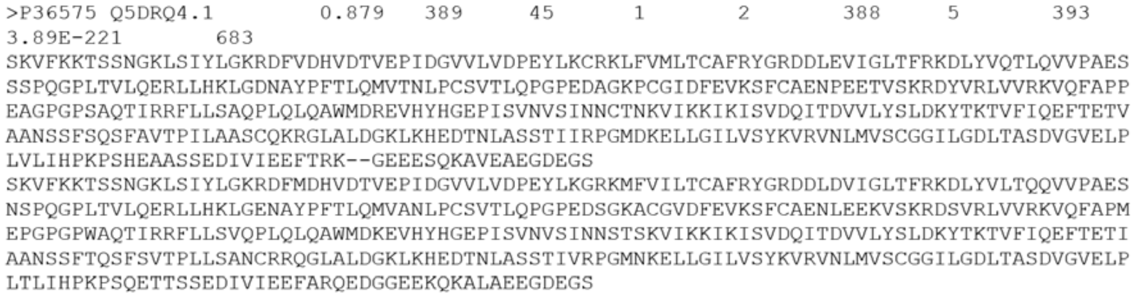
```
mmseqs convertalis arrestin swissprot result.out result.m8
```

The BLAST output-formatted results are saved in the results.m8 file (Figure 15). This text file is formatted as a tab-separated list with 12 columns: (1) query id, (2) hit accession id, (3) identity between query and hit sequences, (4) the length of alignment, (5) number of mismatches, (6) opening gap number, (7-10) the start position and end position of alignment in query and hit sequence, (11) E-value, and (12) alignment bit score. Using the same command and adding option "-format-mode 1" gives the alignment of the query sequence and hit sequences in the FASTA format. Figure 16 shows the alignment between the query chain and the sequence of the Q5DRQ4.1 protein, which is the second hit in Figure 15. The first line in Figure 16 gives the same information and statistics about the alignment that are given in Figure 15 while the second line and third lines give the aligned query and hit sequences, respectively.



```
Query  Hit_id   Identity Alignment Mismatches opening query_start query_end Hit_start  Hit_end   E-value   Bit_
_id                      _length   _num       gap_num _position   _position _position  _position           score
P36575 P36575.2 1.000    388       0          0       1           388       1          388       7.03E-253 774
P36575 Q5DRQ4.1 0.879    389       45         1       2           388       5          393       3.89E-221 683
P36575 Q7YS78.2 0.867    391       49         1       1           388       1          391       3.20E-219 677
P36575 Q9N0H5.2 0.843    389       60         1       1           388       1          389       7.98E-211 653
P36575 Q9EQP6.1 0.841    372       55         1       1           368       1          372       6.80E-200 621
P36575 P51482.1 0.650    369       128        1       2           370       5          372       3.98E-161 509
P36575 P51481.1 0.644    369       130        1       2           370       5          372       4.52E-160 506
P36575 P51483.1 0.628    369       136        1       2           370       5          372       1.47E-152 484
P36575 P17870.1 0.588    389       139        2       2           370       6          393       8.46E-148 471
P36575 P49407.2 0.586    389       140        2       2           370       6          393       1.69E-147 469
......
```

Figure 15. The MMseqs2 output for the human arrestin search again the SWISS-PROT database. We shows the 10 hits with the lowest E-values from the complete set of 40 hits.



```
>P36575 Q5DRQ4.1         0.879   389     45      1       2        388     5        393
3.89E-221        683
SKVFKKTSSNGKLSIYLGKRDFVDHVDTVEPIDGVVLVDPEYLKCRKLFVMLTCAFRYGRDDLEVIGLTFRKDLYVQTLQVVPAES
SSPQGPLTVLQERLLHKLGDNAYPFTLQMVTNLPCSVTLQPGPEDAGKPCGIDFEVKSFCAENPEETVSKRDYVRLVVRKVQFAPP
EAGPGPSAQTIRRFLLSAQPLQLQAWMDREVHYHGEPISVNVSINNCTNKVIKKIKISVDQITDVVLYSLDKYTKTVFIQEFTETV
AANSSFSQSFAVTPILAASCQKRGLALDGKLKHEDTNLASSTIIRPGMDKELLGILVSYKVRVNLMVSCGGILGDLTASDVGVELP
LVLIHPKPSHEAASSEDIVIEEFTRK--GEEESQKAVEAEGDEGS
SKVFKKTSSNGKLSIYLGKRDFMDHVDTVEPIDGVVLVDPEYLKGRKMFVILTCAFRYGRDDLDVIGLTFRKDLYVLTQQVVPAES
NSPQGPLTVLQERLLHKLGENAYPFTLQMVANLPCSVTLQPGPEDSGKACGVDFEVKSFCAENLEEKVSKRDSVRLVVRKVQFAPM
EPGPGPWAQTIRRFLLSVQPLQLQAWMDKEVHYHGEPISVNVSINNSTSKVIKKIKISVDQITDVVLYSLDKYTKTVFIQEFTETI
AANSSFTQSFSVTPLLSANCRRQGLALDGKLKHEDTNLASSTIVRPGMNKELLGILVSYKVRVNLMVSCGGILGDLTASDVGVELP
LTLIHPKPSQETTSSEDIVIEEFARQEDGGEEKQKALAEEGDEGS
```

Figure 16. Alignment of query sequence and Q5DRQ4.1 in FASTA-format for the human arresting query that are produced by MMSeq2.

Users who want faster (less sensitive) searches can adjust sensitivity of MMseq2 with the "-s" parameter. The sensitivity parameter can be set to 1.0 (much faster), 4.0 (fast), 5.7 (default) and 7.5 (slower and more sensitive). MMseqs2 also offers iterative searching similar to PSI-BLAST in order to satisfy the requests for more sensitive searches. The command for the iterative search is the same as a regular search with the addition of the "--num-iterations" parameter. Like PSI-

BLAST, MMseqs2 also generates PSSM for the query sequence. The following commands should be used to obtain PSSM for the human arrestin:

```
mmseqs search arrestin swissprot result.out tmp –e 0.0005 -a --num-iterations 5
mmseqs result2profile arrestin swissprot result.out result.profile
mmseqs profile2pssm result.profile result.pssm
```

The first command produces results for the MMSeq search with 5 iterations. The second command converts the search results into a profile format that is stored in the result.profile file. The third command converts this profile into PSSM that is saved in the result.pssm file. The multiple alignment between the query and hit sequences can be generated with the result2msa command from the generated above result.out file as follows:

```
mmseqs result2msa arrestin swissprot result.out result.msa
```

The result.msa file stores the multiple alignments.

LITERATURE CITED

Altschul, S.F. 1991. Amino acid substitution matrices from an information theoretic perspective. *J Mol Biol* 219:555-565.

Altschul, S.F. 1993. A Protein Alignment Scoring System Sensitive at All Evolutionary Distances. *J Mol Evol* 36:290-300.

Altschul, S.F., Gertz, E.M., Agarwala, R., Schaffer, A.A., and Yu, Y.K. 2009. PSI-BLAST pseudocounts and the minimum description length principle. *Nucleic Acids Res* 37:815-824.

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. 1990. Basic local alignment search tool. *J Mol Biol* 215:403-410.

Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D.J. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* 25:3389-3402.

Bairoch, A. and Apweiler, R. 2000. The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Res* 28:45-48.

Benson, D.A., Cavanaugh, M., Clark, K., Karsch-Mizrachi, I., Ostell, J., Pruitt, K.D., and Sayers, E.W. 2018. GenBank. *Nucleic Acids Res* 46:D41-D47.

Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N., and Bourne, P.E. 2000. The Protein Data Bank. *Nucleic Acids Res* 28:235-242.

Boratyn, G.M., Schaffer, A.A., Agarwala, R., Altschul, S.F., Lipman, D.J., and Madden, T.L. 2012. Domain enhanced lookup time accelerated BLAST. *Biology Direct* 7.

Boutet, E., Lieberherr, D., Tognolli, M., Schneider, M., Bansal, P., Bridge, A.J., Poux, S., Bougueleret, L., and Xenarios, I. 2016. UniProtKB/Swiss-Prot, the Manually Annotated Section of the UniProt KnowledgeBase: How to Use the Entry View. *Methods Mol Biol* 1374:23-54.

Buchfink, B., Xie, C., and Huson, D.H. 2015. Fast and sensitive protein alignment using DIAMOND. *Nat Methods* 12:59-60.

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., and Madden, T.L. 2009. BLAST+: architecture and applications. *BMC Bioinformatics* 10:421.

Choudhuri, S. 2014. Chapter 6 - Sequence Alignment and Similarity Searching in Genomic Databases: BLAST and FASTA*. *In* Bioinformatics for Beginners pp. 133-155. Academic Press, Oxford.

Disfani, F.M., Hsu, W.L., Mizianty, M.J., Oldfield, C.J., Xue, B., Dunker, A.K., Uversky, V.N., and Kurgan, L. 2012. MoRFpred, a computational tool for sequence-based prediction and characterization of short disorder-to-order transitioning binding regions in proteins. *Bioinformatics* 28:i75-83.

Edgar, R.C. 2010. Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* 26:2460-2461.

Hauser, M., Steinegger, M., and Soding, J. 2016. MMseqs software suite for fast and deep clustering and searching of large protein sequence sets. *Bioinformatics*.

Heffernan, R., Paliwal, K., Lyons, J., Dehzangi, A., Sharma, A., Wang, J., Sattar, A., Yang, Y., and Zhou, Y. 2015. Improving prediction of secondary structure, local backbone angles, and solvent accessible surface area of proteins by iterative deep learning. *Sci Rep* 5:11476.

Henikoff, S. and Henikoff, J.G. 1992. Amino-Acid Substitution Matrices from Protein Blocks. *P Natl Acad Sci USA* 89:10915-10919.

Johnson, L.S., Eddy, S.R., and Portugaly, E. 2010. Hidden Markov model speed heuristic and iterative HMM search procedure. *Bmc Bioinformatics* 11.

Johnson, M., Zaretskaya, I., Raytselis, Y., Merezhuk, Y., McGinnis, S., and Madden, T.L. 2008. NCBI BLAST: a better web interface. *Nucleic Acids Res* 36:W5-9.

Jones, D.T. 1999. Protein secondary structure prediction based on position-specific scoring matrices1. *Journal of Molecular Biology* 292:195-202.

Karlin, S. and Altschul, S.F. 1990. Methods for Assessing the Statistical Significance of Molecular Sequence Features by Using General Scoring Schemes. *P Natl Acad Sci USA* 87:2264-2268.

Kulikova, T., Akhtar, R., Aldebert, P., Althorpe, N., Andersson, M., Baldwin, A., Bates, K., Bhattacharyya, S., Bower, L., Browne, P., Castro, M., Cochrane, G., Duggan, K., Eberhardt, R., Faruque, N., Hoad, G., Kanz, C., Lee, C., Leinonen, R., Lin, Q., Lombard, V., Lopez, R., Lorenc, D., McWilliam, H., Mukherjee, G., Nardone, F., Pastor, M.P., Plaister, S., Sobhany, S., Stoehr, P., Vaughan, R., Wu, D., Zhu, W., and Apweiler, R. 2007. EMBL Nucleotide Sequence Database in 2006. *Nucleic Acids Res* 35:D16-20.

Larkin, M.A., Blackshields, G., Brown, N.P., Chenna, R., McGettigan, P.A., McWilliam, H., Valentin, F., Wallace, I.M., Wilm, A., Lopez, R., Thompson, J.D., Gibson, T.J., and Higgins, D.G. 2007. Clustal W and Clustal X version 2.0. *Bioinformatics* 23:2947-2948.

Lipman, D.J. and Pearson, W.R. 1985. Rapid and sensitive protein similarity searches. *Science* 227:1435-1441.

Magnan, C.N. and Baldi, P. 2014. SSpro/ACCpro 5: almost perfect prediction of protein secondary structure and relative solvent accessibility using profiles, machine learning and structural similarity. *Bioinformatics* 30:2592-2597.

Marchler-Bauer, A., Lu, S., Anderson, J.B., Chitsaz, F., Derbyshire, M.K., DeWeese-Scott, C., Fong, J.H., Geer, L.Y., Geer, R.C., Gonzales, N.R., Gwadz, M., Hurwitz, D.I., Jackson, J.D., Ke, Z., Lanczycki, C.J., Lu, F., Marchler, G.H., Mullokandov, M., Omelchenko, M.V., Robertson, C.L., Song, J.S., Thanki, N., Yamashita, R.A., Zhang, D., Zhang, N., Zheng, C., and Bryant, S.H. 2011. CDD: a Conserved Domain Database for the functional annotation of proteins. *Nucleic Acids Res* 39:D225-229.

Meng, F. and Kurgan, L. 2016. Computational Prediction of Protein Secondary Structure from Sequence. *Curr Protoc Protein Sci* 86:2 3 1-2 3 10.

Meng, F., Uversky, V.N., and Kurgan, L. 2017. Comprehensive review of methods for prediction of intrinsic disorder and its molecular functions. *Cell Mol Life Sci* 74:3069-3090.

Mizianty, M.J., Stach, W., Chen, K., Kedarisetti, K.D., Disfani, F.M., and Kurgan, L. 2010. Improved sequence-based prediction of disordered regions with multilayer fusion of multiple information sources. *Bioinformatics* 26:i489-496.

Morgulis, A., Coulouris, G., Raytselis, Y., Madden, T.L., Agarwala, R., and Schaffer, A.A. 2008. Database indexing for production MegaBLAST searches. *Bioinformatics* 24:1757-1764.

Needleman, S.B. and Wunsch, C.D. 1970. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J Mol Biol* 48:443-453.

O'Leary, N.A., Wright, M.W., Brister, J.R., Ciufo, S., Haddad, D., McVeigh, R., Rajput, B., Robbertse, B., Smith-White, B., Ako-Adjei, D., Astashyn, A., Badretdin, A., Bao, Y., Blinkova, O., Brover, V., Chetvernin, V., Choi, J., Cox, E., Ermolaeva, O., Farrell, C.M., Goldfarb, T., Gupta, T., Haft, D., Hatcher, E., Hlavina, W., Joardar, V.S., Kodali, V.K., Li, W., Maglott, D., Masterson, P., McGarvey, K.M., Murphy, M.R., O'Neill, K., Pujar, S., Rangwala, S.H., Rausch, D., Riddick, L.D., Schoch, C., Shkeda, A., Storz, S.S., Sun, H., Thibaud-Nissen, F., Tolstoy, I., Tully, R.E., Vatsan, A.R., Wallin, C., Webb, D., Wu, W., Landrum, M.J., Kimchi, A., Tatusova, T., DiCuccio, M., Kitts, P., Murphy, T.D., and Pruitt, K.D. 2016. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res* 44:D733-745.

Pearson, W.R. and Lipman, D.J. 1988. Improved Tools for Biological Sequence Comparison. *P Natl Acad Sci USA* 85:2444-2448.

Rose, P.W., Prlic, A., Altunkaya, A., Bi, C., Bradley, A.R., Christie, C.H., Costanzo, L.D., Duarte, J.M., Dutta, S., Feng, Z., Green, R.K., Goodsell, D.S., Hudson, B., Kalro, T., Lowe, R., Peisach, E., Randle, C., Rose, A.S., Shao, C., Tao, Y.P., Valasatava, Y., Voigt, M., Westbrook, J.D., Woo, J., Yang, H., Young, J.Y., Zardecki, C., Berman, H.M., and Burley, S.K. 2017. The RCSB protein data bank: integrative view of protein, gene and 3D structural information. *Nucleic Acids Res* 45:D271-D281.

Schwarz, R. and Dayhoff, M. 1979. Matrices for detecting distant relationships. *In* Atlas of protein sequences (M. Dayhoff, ed.) pp. 353-358. National Biomedical Research Foundation.

Smith, T.F. and Waterman, M.S. 1981. Identification of common molecular subsequences. *J Mol Biol* 147:195-197.

Steinegger, M. and Soding, J. 2017. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nat Biotechnol* 35:1026-1028.

The UniProt, C. 2017. UniProt: the universal protein knowledgebase. *Nucleic Acids Res* 45:D158-D169.

Thompson, J.D., Higgins, D.G., and Gibson, T.J. 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res* 22:4673-4680.

Valdar, W.S. 2002. Scoring residue conservation. *Proteins* 48:227-241.

Wang, S., Peng, J., Ma, J.Z., and Xu, J.B. 2016. Protein Secondary Structure Prediction Using Deep Convolutional Neural Fields. *Sci Rep-Uk* 6.

Yan, J., Friedrich, S., and Kurgan, L. 2016. A comprehensive comparative review of sequence-based predictors of DNA- and RNA-binding residues. *Brief Bioinform* 17:88-105.

Zhang, J. and Kurgan, L. 2017. Review and comparative assessment of sequence-based predictors of protein-binding residues. *Brief Bioinform*.

Zhang, J., Ma, Z., and Kurgan, L. 2017. Comprehensive review and empirical analysis of hallmarks of DNA-, RNA- and protein-binding residues in protein chains. *Brief Bioinform*.

Zhang, Z., Schaffer, A.A., Miller, W., Madden, T.L., Lipman, D.J., Koonin, E.V., and Altschul, S.F. 1998. Protein sequence similarity searches using patterns as seeds. *Nucleic Acids Research* 26:3986-3990.

Zhao, Y., Tang, H., and Ye, Y. 2012. RAPSearch2: a fast and memory-efficient protein similarity search tool for next-generation sequencing data. *Bioinformatics* 28:125-126.

KEY REFERENCES

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. 1990. Basic local alignment search tool. *J Mol Biol* 215:403-410.
Introduces the BLAST algorithm

Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D.J. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* 25:3389-3402.
Describes the PSI-BLAST algorithm

Johnson, M., Zaretskaya, I., Raytselis, Y., Merezhuk, Y., McGinnis, S., and Madden, T.L. 2008. NCBI BLAST: a better web interface. *Nucleic Acids Res* 36:W5-9.
Presents the WEB BLAST suite

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., and Madden, T.L. 2009. BLAST+: architecture and applications. *BMC Bioinformatics* 10:421.
Describes the BLAST+ suite

Steinegger, M. and Soding, J. 2017. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nat Biotechnol* 35:1026-1028.
Introduces the MMseqs2 algorithm

INTERNET RESOURCES

ftp://ftp.ncbi.nlm.nih.gov/blast/executables/blast+/LATEST/
BLAST+

ftp://ftp.ncbi.nlm.nih.gov/blast/db/
BLAST databases

https://github.com/soedinglab/MMseqs2
MMSeq2

https://blast.ncbi.nlm.nih.gov/Blast.cgi
WEB BLAST