bioRxiv preprint doi: https://doi.org/10.1101/556225; this version posted February 21, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

*loc*BLAST v2.0 - an improved PHP library for embedding standalone NCBI BLAST+ program to an interactive graphical user interface

T. Ashok Kumar

Department of Bioinformatics, Noorul Islam College of Arts and Science, Kumaracoil, Thuckalay 629180, Tamil Nadu, INDIA

ashok@biogem.org

Abstract

Background: NCBI BLAST is a most popularly used sequence analysis platform in a broad range of applications. The final release of standalone WWW BLAST server (wwwBLAST v2.2.26) of NCBI was officially released on May 10, 2004, and discontinued its support. Due to the popularity and high demand, the BLAST algorithms were refined with rich features and released as online BLAST, standalone BLAST+, BLAST RESTful service, and cloud BLAST server. *loc*BLAST v2.0 is a lightweight PHP library, designed in preference to WWW BLAST server. It masks command-line BLAST+ programs to a rich graphical user interface (GUI).

Methods: The *loc*BLAST v2.0 library was designed using PHP, CSS, pure JavaScript, and standalone NCBI BLAST+ executables.

*Results: loc*BLAST v2.0 provides an interactive web interface to input query sequences through the web form and provides a graphical report of the sequence alignment result. The graphical overview, tabular summary, and formatted sequence alignment of *loc*BLAST v2.0 result mimic to the native format of online NCBI BLAST result. It allows users to perform both local and online database search.

Availability: Freely available at https://github.com/AshokHub/locBLAST

Keywords: locBLAST, Local BLAST Search, NCBI BLAST+, Sequence Alignment, Embedding BLAST+

1. Introduction

Sequence alignment is one of the best and well-known method in Bioinformatics for identifying the similarity between the two biological sequences through sequence alignment.

bioRxiv preprint doi: https://doi.org/10.1101/556225; this version posted February 21, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Through pairwise sequence alignment, it is possible to predict the structural, functional, and evolutionary relationship of the sequence. The NCBI BLAST (Basic Local Alignment Search Tool) (Altschul et al., 1990) offers different types of programs mainly used for finding local similarity between two biological sequences or a dataset. It provides integrative access to the various biological databases through the web user interface, command-line user interface, and application programming interface. NCBI BLAST+ (Camacho et al., 2009) is a standalone suite of command-line BLAST programs used to compare the query sequence with both online and offline databases. The NCBI BLAST+ executables and source codes (a subset of the NCBI C++ toolkit (Vakatov, 2004)) are freely available at NCBI FTP repository for download.

The standalone NCBI BLAST+ suite consists of various programs and utilities for the different types of sequence alignments, database creation, and data manipulations. The list of NCBI BLAST+ programs which can include with locBLAST v2.0 library is below. All the programs except RPS-BLAST and RPS-TBLASTN support both online and offline sequence database search, and customized BLAST search by changing advanced algorithm parameters (Fig. 1).

- BLASTN program compares the nucleotide query sequence with the nucleotide sequences in the database.
- BLASTP program compares the protein query sequence with the protein sequences in the database.
- BLASTX program translates the nucleotide query sequence and compares with the protein sequences in the database.
- TBLASTN program compares the protein query sequence with the nucleotide sequences in the database, by instantly translating the nucleotide sequences in the database before comparison.
- TBLASTX program compares the translated six reading frames of nucleotide query sequence with the translated six reading frames of nucleotide sequences in the database. The query is translated on the fly before searching sequences in the database.
- DELTA-BLAST (Boratyn et al., 2012) program compares the query protein sequence with the protein sequences in the database by constructing the position-specific scoring matrices (PSSM) from the conserved domain database search results.

bioRxiv preprint doi: https://doi.org/10.1101/556225; this version posted February 21, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

- PSI-BLAST (Altschul et al., 1997) program compares the query protein sequence with the protein sequences in the conserved domain database using PSSM scores by position specific iteration method.
- RPS-BLAST program compares the query protein sequence with the protein sequences in the conserved domain database through profiles. It is a reverse of the PSI-BLAST search.
- RPS-TBLASTN program compares the translated six reading frames of nucleotide query sequence with the protein sequences in the conserved domain database through profiles.

The stand-alone NCBI BLAST+ must be installed and configured in each client machines for individual BLAST search. Moreover, the clients can't access other databases; and specialization is needed to perform the advanced BLAST search. To outcome the issues, *loc*BLAST v2.0 is developed to host a graphical interfaced BLAST server which can accomplish online and offline database search. The default and advanced BLAST algorithm parameters in *loc*BLAST v2.0 web form automatically adjust to the default options according to the BLAST+ program and database on selection. For each BLAST+ program, it has a short title with a hyperlink to the webpage containing a detailed description of the program and algorithm parameters.

2. Methods

The *loc*BLAST v2.0 library is designed using PHP, CSS, and pure JavaScript. It functions under nine different phases: (i) FASTA file format validation, (ii) amino acid or nucleotide sequence validation, (iii) proper BLAST+ program parameters validation, (iv) back-end BLAST+ program execution, (v) BlastXML2 file parser (Ashok *et al*., 2017), (vi) graphical overview of hits of sequence alignments display, (vii) display tabular format short summary of hits of sequence alignments, (viii) formatted three row of sequence alignment, and (ix) save sequence to the disk. The PHP script executes the NCBI BLAST+ programs using exec() function through passing parameters from the HTML form fields. *loc*BLAST v2.0 allows users to submit input sequence either through the text area or file upload. The output of BLAST+ result is generated in plain HTML format or user interactive graphical format. In the graphical output, graphical overview (Fig. 2) and tabular summary are properly hyperlinked with tooltip text to the sequence alignments.

bioRxiv preprint doi: https://doi.org/10.1101/556225; this version posted February 21, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

2.1 Implementation

The *loc*BLAST v2.0 library is designed user-friendly for both beginners and developers. Implementation of *loc*BLAST v2.0 to a local BLAST server consists of two simple stages: (i) Downloading and extracting files, scripts, and BLAST+ binaries from GitHub and NCBI FTP to the web directory, and (ii) Custom database preparation or modification.

2.1.1 Installation and configuration

*loc*BLAST v2.0 can be easily installed in any PHP supporting webserver and platform. A brief step-wise instruction for setting *loc*BLAST v2.0 is below:

(1) Create a new directory named 'blast' (optional) in the web-server parent directory (usually named as 'htdocs', 'www', 'wwwroot', or 'webpath').

(2) Download latest BLAST+ binaries from NCBI FTP repository (ftp://ftp.ncbi.nih.gov/blast/executables/blast+/LATEST) based on the platform and extract to the 'blast' directory.

(3) Download *loc*BLAST v2.0 from GitHub (https://github.com/AshokHub/locBLAST) and extract to the 'blast' directory.

(4) Online biological databases can be downloaded from NCBI FTP repository (ftp://ftp.ncbi.nih.gov/blast/db) and moved to 'db' directory to perform database search in offline. A brief protocol for custom database preparation is explained in section 2.1.2.

(5) Add or modify 'database name' and 'path of the database file' codes in 'index.php' and 'footer.php' script for the custom database accordingly (optional).

(6) Open the URL http(s)://<hostname>:<port>/blast (for instance http://localhost/blast) in the web browser to enter the query sequence.

(7) Choose appropriate BLAST program to use and database to search (Fig. 1).

The *loc*BLAST v2.0 provides access to search online databases namely Non-Redundant Protein/Nucleotide Sequences (NR/NT), UniProtKB/Swiss-Prot (SwissProt), Reference Proteins Sequences (RefSeq_Protein), Reference RNA Sequences (RefSeq_RNA), Expressed Sequence Tags (EST), Protein Data Bank (PDB), and Patented Protein Sequences (PAT); but not limited.

bioRxiv preprint doi: https://doi.org/10.1101/556225; this version posted February 21, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

2.1.2 Custom database preparation

The custom database preparation requires little technical knowledge of BLAST+ programs. BLAST+ program supports the nucleotide database, protein database, and conserved domain database search. After database preparation, the form fields to be added or modified to existing database list. A simple protocol for creating own database is below:

- For creating a nucleotide database – execute the following command-line ./**makeblastdb -in** db/test_na.fas **-out** db/test_na **-dbtype nucl -title** Test_NA
- For creating a protein database – execute the following command-line ./**makeblastdb -in** db/test_aa.fas **-out** db/test_aa **-dbtype prot -title** Test_AA
- For creating a RPS/PSI (CDD) database – execute the following command-line ./**makeprofiledb -in** db/CDD/test_CDD_db.pn **-out** db/CDD/test_CDD_db **-dbtype rps -title** Test_CDD_db
- For creating a Delta (CDD) database – execute the following command-line ./**makeprofiledb -in** db/CDD/test_Delta_db.pn **-out** db/CDD/test_Delta_db **-dbtype delta -title** Test_Delta_db

Where, the words represented in bold are keywords and should not be modified. The custom databases namely 'Test_NA', 'Test_AA', 'Test_CDD_db', 'PDTDB', and 'Smart' are included in the *loc*BLAST v2.0 library for testing.

2.1.3 Database manipulation

The BLAST+ applications such as blastdb_aliastool, blast_formatter, blastdbcheck, convert2blastmask, dustmasker, segmasker, blastcmd, makembindex, and windowmasker are used for database manipulation or data retrieval purposes.

3. Results

The result page of *loc*BLAST v2.0 search produces a graphical output which contains well-formatted sequence alignment, color key, and summary table closely similar to the online BLAST result. Moreover, the web form and result page contain hyperlinks linked to the source or anchors, and keywords are annotated with a brief description of the term by tooltip text.

The graphical output consists of five section: (i) program search parameters, (ii) distribution of top BLAST hits in a graph, (iii) summary table of BLAST hits, (iv) row-wise

bioRxiv preprint doi: https://doi.org/10.1101/556225; this version posted February 21, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

alignment for each hits, and (v) Karlin-Altschul statistics (Fig. 3). Moreover, the database accession number and definition line in the graphical output are automatically annotated and hyperlinked to the source database entry based on the database identifiers namely *gb*, *emb*, *dbj*, *pir*, *prf*, *sp*, *pdb*, *pat*, *bbs*, *gnl*, and *ref*. In plain text output, the search result is presented by monospace text and interlinks to the sequence alignment hits (Fig. 4).

*loc*BLAST v2.0 does not depend on any additional third-party scripts or plugins to run the NCBI BLAST+ program. It is platform independent and hence it works in most of the traditional and modern browsers. NCBI BLAST+ executables in the *loc*BLAST v2.0 can be simply updated by downloading current release from NCBI FTP site and replacing the old executables. The current stable release of NCBI BLAST+ is v2.8.1 (as on January 11, 2019), distributed as source codes and compiled binaries for Windows, Linux, and Macintosh operating systems as the installable and compressed files. The *loc*BLAST v2.0 PHP library and the test database files were freely available at GitHub under GNU General Public License v3.0.

References

- Altschul, S.F., Gish, W., Miller, W., Myers, E.W., Lipman, D.J., 1990. Basic Local Alignment Search Tool. J. Mol. Biol. 215(3), 403–410. doi: 10.1016/S0022-2836(05)80360-2
- Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., Madden, T.L., 2009. BLAST+: Architecture and Applications. BMC Bioinformatics. 10, 421. doi: 10.1186/1471-2105-10-421
- Vakatov, D., 2004. The NCBI C++ Toolkit Book. National Center for Biotechnology Information, USA.
- Boratyn, G.M., Schäffer, A.A., Agarwala, R., Altschul, S.F., Lipman, D.J., & Madden, T. L., 2012. Domain enhanced lookup time accelerated BLAST. Biology Direct. 7(1), 12. doi: 10.1186/1745-6150-7-12
- Altschul, S.F., Madden, T.L., Schäffer, A.A., Zhang, J., Zhang, Z., Miller, W., Lipman, D.J., 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 25(17), 3389–3402. doi: 10.1093/nar/25.17.3389
- Ashok, K.T., Rajagopal, B., 2009. xmlBLASTparser v1.1 — A PHP Based NCBI BLAST XML Output Parser. Int. J. Adv. Res. Comput. Sci. 8(8), 230–232. doi: 10.26483/ijarcs.v8i8.4648

bioRxiv preprint doi: https://doi.org/10.1101/556225; this version posted February 21, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

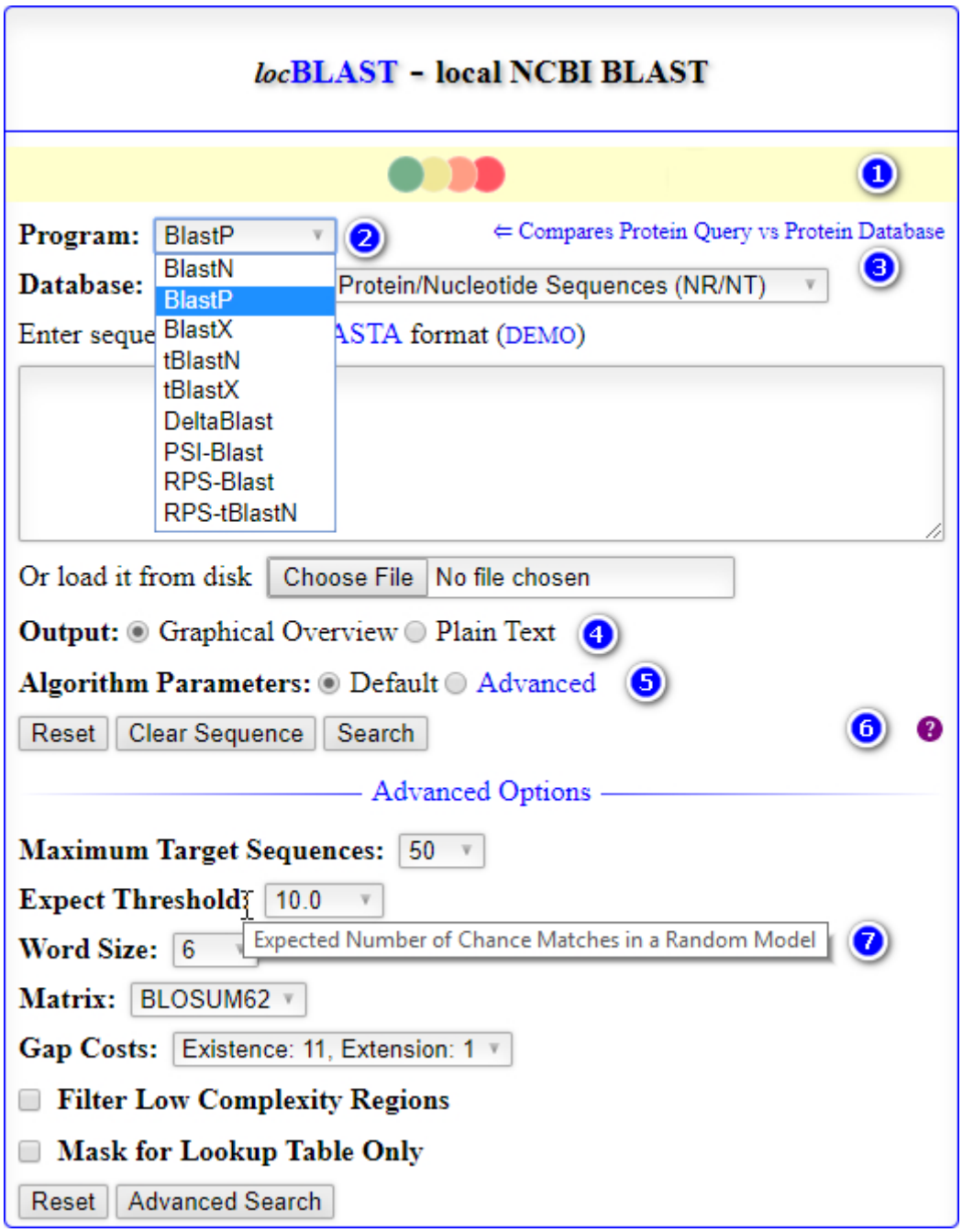


Fig. 1: Homepage of *loc*BLAST v2.0 web form with default and advanced BLAST+ search options. (1) progress bar and status bar, (2) advanced BLAST+ programs deltablast, psiblast, rpsblast, and rpstblastn, (3) hyperlink to detailed program description page, with short title, (4) option to graphical or plain text output, (5) option to advanced search algorithm parameters, (6) tutorial page, and (7) mouse over tooltip text description of the keyword.

bioRxiv preprint doi: https://doi.org/10.1101/556225; this version posted February 21, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

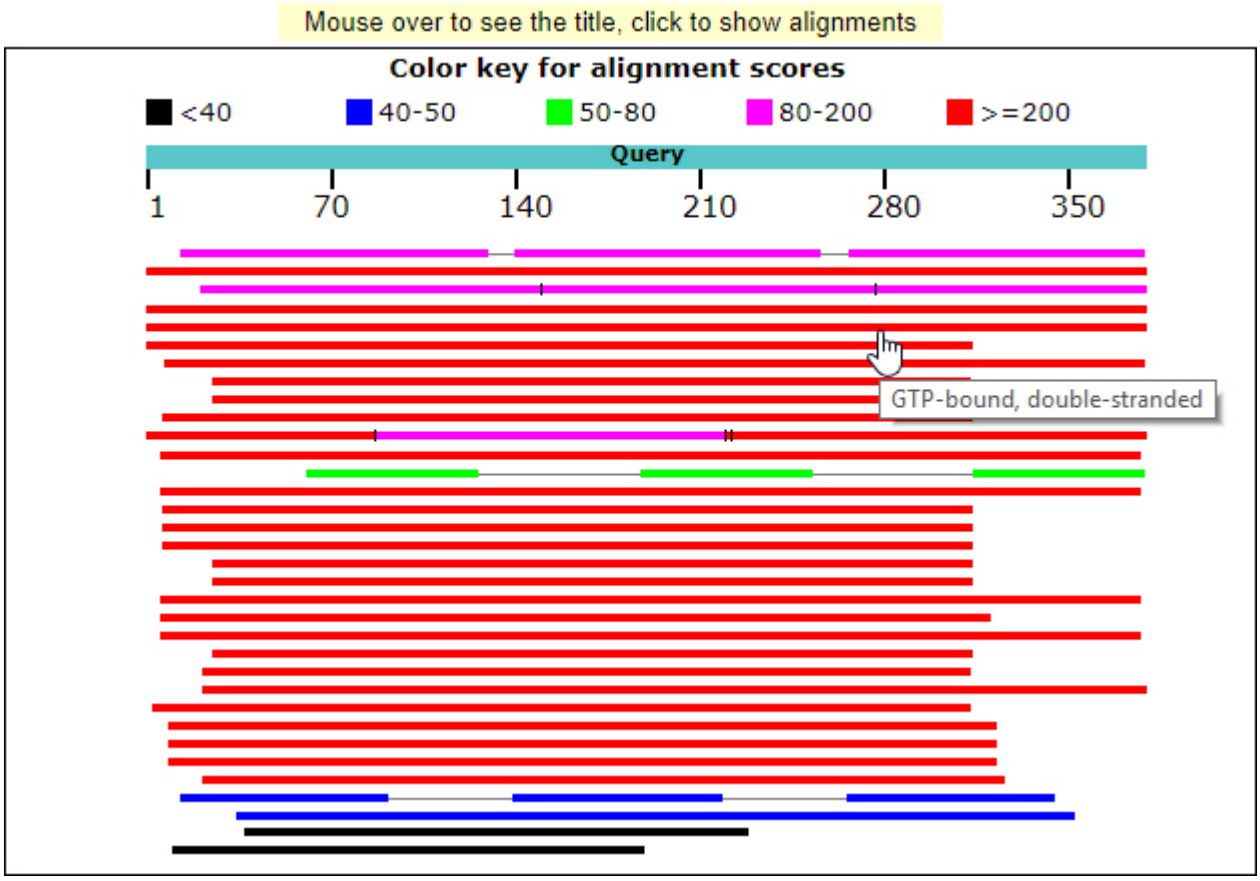


Fig. 2: Graphical overview of hits of sequence alignment sorted by scores (color keys), similar to native NCBI BLAST result.

bioRxiv preprint doi: https://doi.org/10.1101/556225; this version posted February 21, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

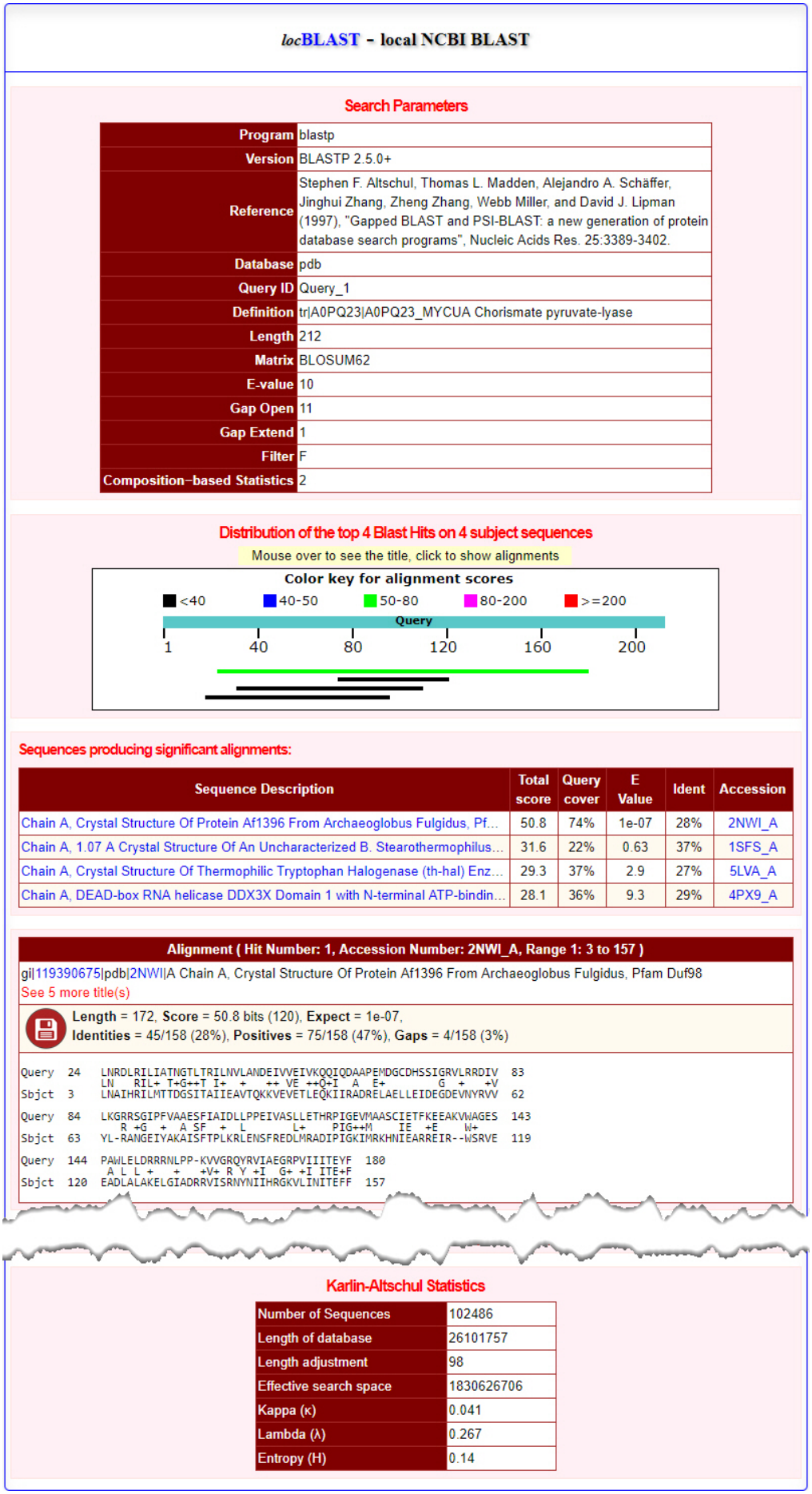
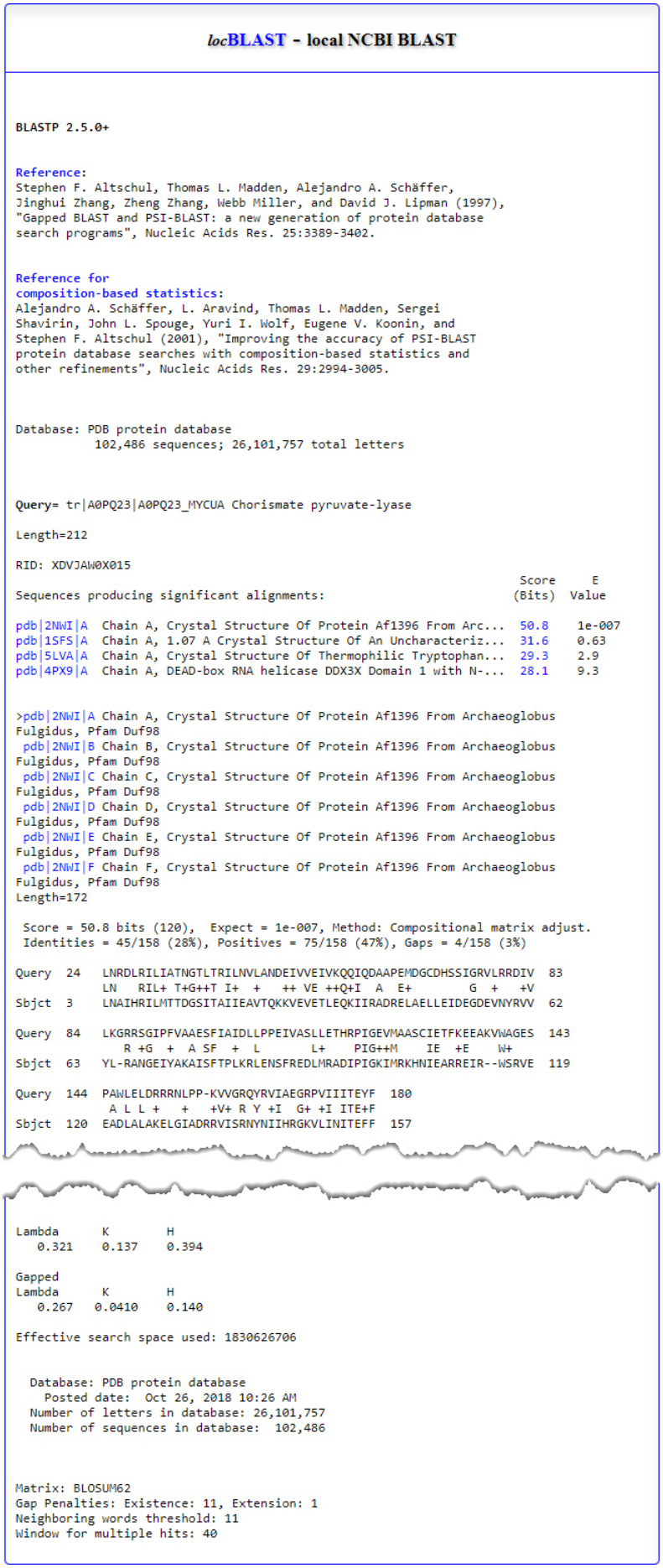


Fig. 3: Graphical output page of *loc*BLAST v2.0 sequence search result (cropped large page).

bioRxiv preprint doi: https://doi.org/10.1101/556225; this version posted February 21, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.



***loc*BLAST - local NCBI BLAST**

```
BLASTP 2.5.0+


Reference:
Stephen F. Altschul, Thomas L. Madden, Alejandro A. Schäffer,
Jinghui Zhang, Zheng Zhang, Webb Miller, and David J. Lipman (1997),
"Gapped BLAST and PSI-BLAST: a new generation of protein database
search programs", Nucleic Acids Res. 25:3389-3402.


Reference for
composition-based statistics:
Alejandro A. Schäffer, L. Aravind, Thomas L. Madden, Sergei
Shavirin, John L. Spouge, Yuri I. Wolf, Eugene V. Koonin, and
Stephen F. Altschul (2001), "Improving the accuracy of PSI-BLAST
protein database searches with composition-based statistics and
other refinements", Nucleic Acids Res. 29:2994-3005.



Database: PDB protein database
          102,486 sequences; 26,101,757 total letters



Query= tr|A0PQ23|A0PQ23_MYCUA Chorismate pyruvate-lyase

Length=212

RID: XDVJAW0X015
                                                                    Score     E
Sequences producing significant alignments:                        (Bits)  Value

pdb|2NWI|A  Chain A, Crystal Structure Of Protein Af1396 From Arc...  50.8    1e-007
pdb|1SFS|A  Chain A, 1.07 A Crystal Structure Of An Uncharacteriz...  31.6    0.63
pdb|5LVA|A  Chain A, Crystal Structure Of Thermophilic Tryptophan...  29.3    2.9
pdb|4PX9|A  Chain A, DEAD-box RNA helicase DDX3X Domain 1 with N-...  28.1    9.3


>pdb|2NWI|A Chain A, Crystal Structure Of Protein Af1396 From Archaeoglobus
Fulgidus, Pfam Duf98
 pdb|2NWI|B Chain B, Crystal Structure Of Protein Af1396 From Archaeoglobus
Fulgidus, Pfam Duf98
 pdb|2NWI|C Chain C, Crystal Structure Of Protein Af1396 From Archaeoglobus
Fulgidus, Pfam Duf98
 pdb|2NWI|D Chain D, Crystal Structure Of Protein Af1396 From Archaeoglobus
Fulgidus, Pfam Duf98
 pdb|2NWI|E Chain E, Crystal Structure Of Protein Af1396 From Archaeoglobus
Fulgidus, Pfam Duf98
 pdb|2NWI|F Chain F, Crystal Structure Of Protein Af1396 From Archaeoglobus
Fulgidus, Pfam Duf98
Length=172

 Score = 50.8 bits (120),  Expect = 1e-007, Method: Compositional matrix adjust.
 Identities = 45/158 (28%), Positives = 75/158 (47%), Gaps = 4/158 (3%)

Query  24   LNRDLRILIATNGTLTRILNVLANDEIVVEIVKQQIQDAAPEMDGCDHSSIGRVLRRDIV  83
            LN   RIL+ T+G++T I+  +   ++ VE ++Q+I  A  E+        G  +   +V
Sbjct  3    LNAIHRILMTTDGSITAIIEAVTQKKVEVETLEQKIIRADRELAELLEIDEGDEVNYRVV  62

Query  84   LKGRRSGIPFVAAESFIAIDLLPPEIVASLLETHRPIGEVMAASCIETFKEEAKVWAGES  143
               R +G  +  A SF  +  L       L+    PIG++M    IE  +E    W+
Sbjct  63   YL-RANGEIYAKAISFTPLKRLENSFREDLMRADIPIGKIMRKHNIEARREIR--WSRVE  119

Query  144  PAWLELDRRRNLPP-KVVGRQYRVIAEGRPVIIITEYF  180
             A L L +   +   +V+ R Y +I  G+ +I ITE+F
Sbjct  120  EADLALAKELGIADRRVISRNYNIIHRGKVLINITEFF  157


Lambda      K        H
   0.321    0.137    0.394

Gapped
Lambda      K        H
   0.267   0.0410    0.140

Effective search space used: 1830626706


  Database: PDB protein database
    Posted date:  Oct 26, 2018 10:26 AM
  Number of letters in database: 26,101,757
  Number of sequences in database:  102,486



Matrix: BLOSUM62
Gap Penalties: Existence: 11, Extension: 1
Neighboring words threshold: 11
Window for multiple hits: 40
```

Fig. 4: Plain text output page of *loc*BLAST v2.0 sequence search result (cropped large page).

*loc*BLAST - local NCBI BLAST

1

Program: BlastP 2 ⇐ Compares Protein Query vs Protein Database

- BlastN
- BlastP
- BlastX
- tBlastN
- tBlastX
- DeltaBlast
- PSI-Blast
- RPS-Blast
- RPS-tBlastN

Database: Protein/Nucleotide Sequences (NR/NT) 3

Enter seque ASTA format (DEMO)

Or load it from disk Choose File No file chosen

Output: • Graphical Overview Plain Text 4

Algorithm Parameters: • Default Advanced 5

Reset Clear Sequence Search 6

Advanced Options

Maximum Target Sequences: 50

Expect Threshold 10.0

Word Size: 6 Expected Number of Chance Matches in a Random Model 7

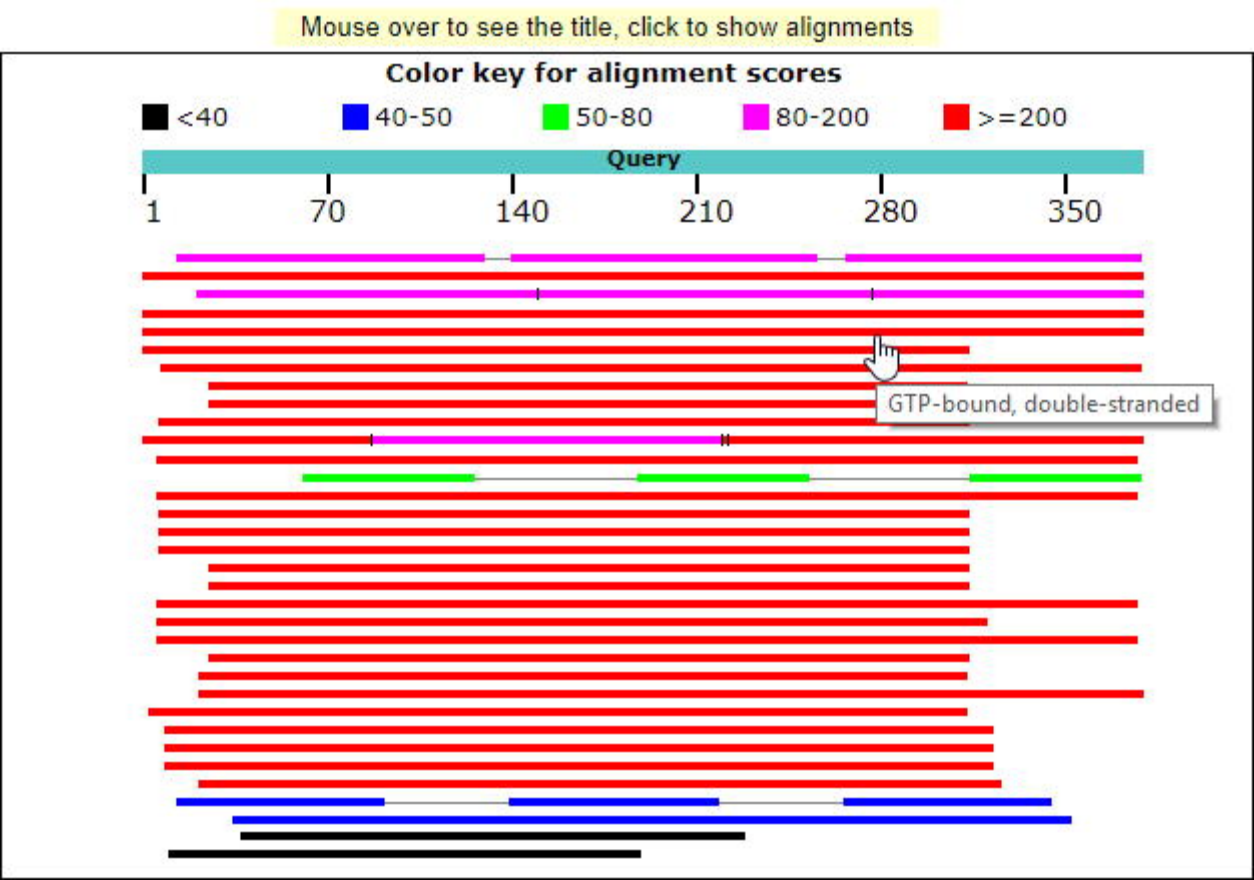
Matrix: BLOSUM62

Gap Costs: Existence: 11, Extension: 1

Filter Low Complexity Regions

Mask for Lookup Table Only

Reset Advanced Search


Mouse over to see the title, click to show alignments
Color key for alignment scores
<40
40-50
50-80
80-200
>=200
Query
1
70
140
210
280
350
GTP-bound, double-stranded

*loc*BLAST - local NCBI BLAST

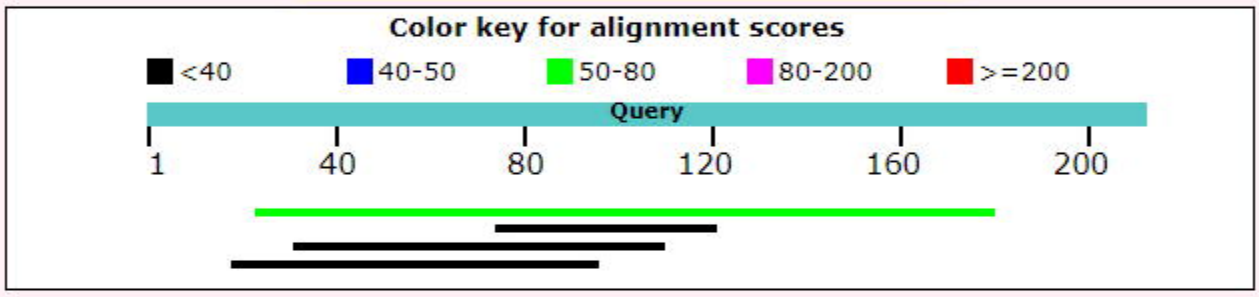
Search Parameters

Program	blastp
Version	BLASTP 2.5.0+
Reference	Stephen F. Altschul, Thomas L. Madden, Alejandro A. Schäffer, Jinghui Zhang, Zheng Zhang, Webb Miller, and David J. Lipman (1997), "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs", Nucleic Acids Res. 25:3389-3402.
Database	pdb
Query ID	Query_1
Definition	tr|A0PQ23|A0PQ23_MYCUA Chorismate pyruvate-lyase
Length	212
Matrix	BLOSUM62
E-value	10
Gap Open	11
Gap Extend	1
Filter	F
Composition-based Statistics	2

bioRxiv preprint doi: https://doi.org/10.1101/556225; this version posted February 21, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Distribution of the top 4 Blast Hits on 4 subject sequences

Mouse over to see the title, click to show alignments



Sequences producing significant alignments:

Sequence Description	Total score	Query cover	E Value	Ident	Accession
Chain A, Crystal Structure Of Protein Af1396 From Archaeoglobus Fulgidus, Pf...	50.8	74%	1e-07	28%	2NWI_A
Chain A, 1.07 A Crystal Structure Of An Uncharacterized B. Stearothermophilus...	31.6	22%	0.63	37%	1SFS_A
Chain A, Crystal Structure Of Thermophilic Tryptophan Halogenase (th-hal) Enz...	29.3	37%	2.9	27%	5LVA_A
Chain A, DEAD-box RNA helicase DDX3X Domain 1 with N-terminal ATP-bindin...	28.1	36%	9.3	29%	4PX9_A

Alignment (Hit Number: 1, Accession Number: 2NWI_A, Range 1: 3 to 157)

gi|119390675|pdb|2NWI|A Chain A, Crystal Structure Of Protein Af1396 From Archaeoglobus Fulgidus, Pfam Duf98

See 5 more title(s)

Length = 172, **Score** = 50.8 bits (120), **Expect** = 1e-07,
Identities = 45/158 (28%), **Positives** = 75/158 (47%), **Gaps** = 4/158 (3%)

```
Query  24   LNRDLRILIATNGTLTRILNVLANDEIVVEIVKQQIQDAAPEMDGCDHSSIGRVLRRDIV  83
            LN   RIL+ T+G++T I+  +   ++ VE ++Q+I  A  E+        G  +   +V
Sbjct  3    LNAIHRILMTTDGSITAIIEAVTQKKVEVETLEQKIIRADRELAELLEIDEGDEVNYRVV  62

Query  84   LKGRRSGIPFVAAESFIAIDLLPPEIVASLLETHRPIGEVMAASCIETFKEEAKVWAGES  143
               R +G  +  A SF  +  L        L+    PIG++M    IE  +E    W+
Sbjct  63   YL-RANGEIYAKAISFTPLKRLENSFREDLMRADIPIGKIMRKHNIEARREIR--WSRVE  119

Query  144  PAWLELDRRRNLPP-KVVGRQYRVIAEGRPVIIITEYF  180
             A L L +   +   +V+ R Y +I  G+ +I ITE+F
Sbjct  120  EADLALAKELGIADRRVISRNYNIIHRGKVLINITEFF  157
```

Karlin-Altschul Statistics

Number of Sequences	102486
Length of database	26101757
Length adjustment	98
Effective search space	1830626706
Kappa (κ)	0.041
Lambda (λ)	0.267
Entropy (H)	0.14

BLASTP 2.5.0+

Reference:
Stephen F. Altschul, Thomas L. Madden, Alejandro A. Schäffer,
Jinghui Zhang, Zheng Zhang, Webb Miller, and David J. Lipman (1997),
"Gapped BLAST and PSI-BLAST: a new generation of protein database
search programs", Nucleic Acids Res. 25:3389-3402.

**Reference for
composition-based statistics**:
Alejandro A. Schäffer, L. Aravind, Thomas L. Madden, Sergei
Shavirin, John L. Spouge, Yuri I. Wolf, Eugene V. Koonin, and
Stephen F. Altschul (2001), "Improving the accuracy of PSI-BLAST
protein database searches with composition-based statistics and
other refinements", Nucleic Acids Res. 29:2994-3005.

Database: PDB protein database
102,486 sequences; 26,101,757 total letters

bioRxiv preprint doi: https://doi.org/10.1101/556225; this version posted February 21, 2019. The copyright holder for this pr
was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuit
available under aCC-BY-NC-ND 4.0 International license.

Query= tr|A0PQ23|A0PQ23_MYCUA Chorismate pyruvate-lyase

Length=212

RID: XDVJAW0X015

Sequences producing significant alignments:	Score (Bits)	E Value
pdb|2NWI|A Chain A, Crystal Structure Of Protein Af1396 From Arc...	50.8	1e-007
pdb|1SFS|A Chain A, 1.07 A Crystal Structure Of An Uncharacteriz...	31.6	0.63
pdb|5LVA|A Chain A, Crystal Structure Of Thermophilic Tryptophan...	29.3	2.9
pdb|4PX9|A Chain A, DEAD-box RNA helicase DDX3X Domain 1 with N-...	28.1	9.3

>pdb|2NWI|A Chain A, Crystal Structure Of Protein Af1396 From Archaeoglobus
Fulgidus, Pfam Duf98
 pdb|2NWI|B Chain B, Crystal Structure Of Protein Af1396 From Archaeoglobus
Fulgidus, Pfam Duf98
 pdb|2NWI|C Chain C, Crystal Structure Of Protein Af1396 From Archaeoglobus
Fulgidus, Pfam Duf98
 pdb|2NWI|D Chain D, Crystal Structure Of Protein Af1396 From Archaeoglobus
Fulgidus, Pfam Duf98
 pdb|2NWI|E Chain E, Crystal Structure Of Protein Af1396 From Archaeoglobus
Fulgidus, Pfam Duf98
 pdb|2NWI|F Chain F, Crystal Structure Of Protein Af1396 From Archaeoglobus
Fulgidus, Pfam Duf98
Length=172

 Score = 50.8 bits (120), Expect = 1e-007, Method: Compositional matrix adjust.
 Identities = 45/158 (28%), Positives = 75/158 (47%), Gaps = 4/158 (3%)

```
Query  24   LNRDLRILIATNGTLTRILNVLANDEIVVEIVKQQIQDAAPEMDGCDHSSIGRVLRRDIV  83
            LN   RIL+ T+G++T I+  +   ++ VE ++Q+I  A  E+         G  +   +V
Sbjct  3    LNAIHRILMTTDGSITAIIEAVTQKKVEVETLEQKIIRADRELAELLEIDEGDEVNYRVV  62

Query  84   LKGRRSGIPFVAAESFIAIDLLPPEIVASLLETHRPIGEVMAASCIETFKEEAKVWAGES  143
               R +G  +  A SF  +  L        L+    PIG++M    IE  +E    W+
Sbjct  63   YL-RANGEIYAKAISFTPLKRLENSFREDLMRADIPIGKIMRKHNIEARREIR--WSRVE  119

Query  144  PAWLELDRRRNLPP-KVVGRQYRVIAEGRPVIIITEYF  180
             A L L +   +   +V+ R Y +I  G+ +I ITE+F
Sbjct  120  EADLALAKELGIADRRVISRNYNIIHRGKVLINITEFF  157
```

Lambda   K   H
   0.321   0.137   0.394

Gapped
Lambda   K   H
   0.267   0.0410   0.140

Effective search space used: 1830626706

  Database: PDB protein database
    Posted date: Oct 26, 2018 10:26 AM
  Number of letters in database: 26,101,757
  Number of sequences in database: 102,486

Matrix: BLOSUM62
Gap Penalties: Existence: 11, Extension: 1
Neighboring words threshold: 11
Window for multiple hits: 40