



Enhancing Sequence Search Efficiency by Using STNext®

Jim Brown (FIZ), John Kratunis (CAS)

STNext

FIZ Karlsruhe
Leibniz Institute for Information Infrastructure

CAS
A DIVISION OF THE
AMERICAN CHEMICAL SOCIETY



Agenda

- Database Overview
- Steps for a CAS Registry BLAST® Sequence Search
- Create a BLAST Alignment Report
- Substructure Search on a Short Sequence
- Step-by-step through a multi-database BLAST search
- Post-processing of a multi-step BLAST search
- Overview of results
- Summary

2

STNext

FIZ Karlsruhe
Leibniz Institute for Information Infrastructure

CAS
A DIVISION OF THE
AMERICAN CHEMICAL SOCIETY

Substance databases containing sequence information

- CAS REGISTRY®
 - Contains >71M biosequences
- DGENE (Derwent GENESQ™)
 - Contains >46M DWPI biosequences
- GENBANK®
 - Contains >230M sequences
- PCTGEN
 - Contains >16M WIPO biosequences
- USGENE®
 - Contains >61M USPTO biosequences

3

STNextFIZ Karlsruhe
Leibniz Institute for Information InfrastructureCAS
A DIVISION OF THE
AMERICAN CHEMICAL SOCIETY

CAS REGISTRY supports searching sequences

- NCBI BLAST similarity searching on STNext
 - Via the BLAST Module for REGISTRY/CaplusSM
 - Available via the command line for USGENE, PCTGEN, and DGENE
- Structure search for shorter sequences
- Text-based searching
 - Sequence code match (motif): exact, family and subsequence
 - Name fragments
 - Chemical modifications

4

STNextFIZ Karlsruhe
Leibniz Institute for Information InfrastructureCAS
A DIVISION OF THE
AMERICAN CHEMICAL SOCIETY

Steps for a REGISTRY BLAST Sequence Search

1. Download the stand-alone BLAST module
2. Logon to BLAST module and complete the BLAST search
3. View Results and Select desired sequences
4. Get STN® Data Script (.scb) and alignment data (.xss)
5. Import into scripts folder on STNext
6. Run script to get sequence records in REGISTRY
7. Cross-file search into CPlusSM
8. Display BIB ABS HITRN
9. Create BLAST alignment report from Transcripts folder

5



CAS Registry BLAST Searching: Example

Search Question:

Locate references that report protein sequences similar to CAS RN 642514-19-0.



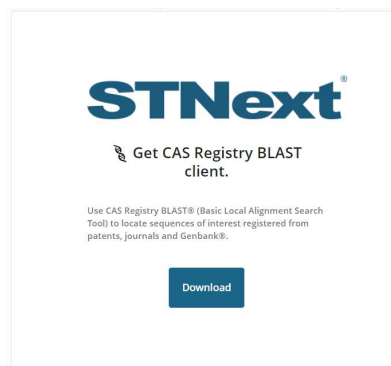
6



NCBI BLAST for CAS REGISTRY / CAplus

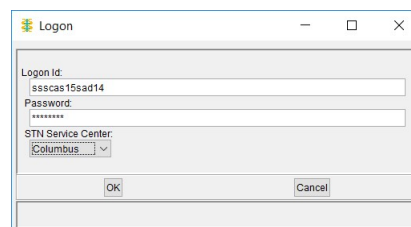
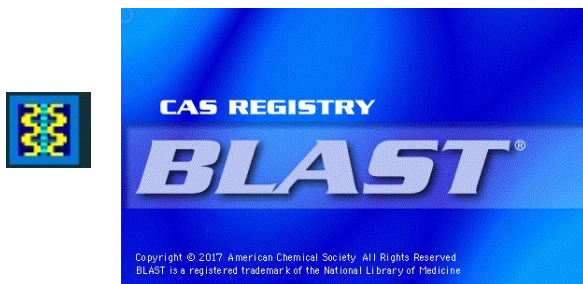
- To download the separate BLAST Module:

- Login to STNext at the following site
<https://next.stn.org/stn/downloads/blast-download.html>
- Click to download



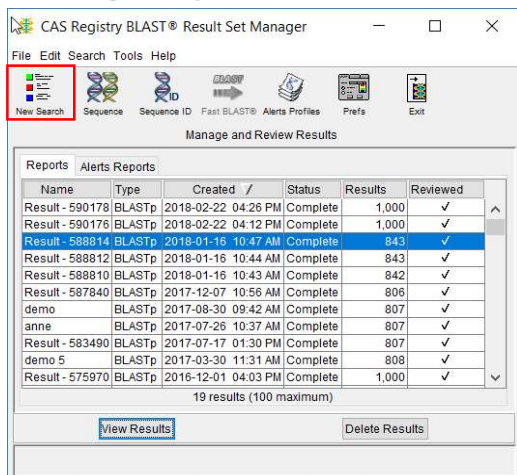
7

Launch CAS Registry BLAST and Logon



8

CAS Registry BLAST Result Set Manager

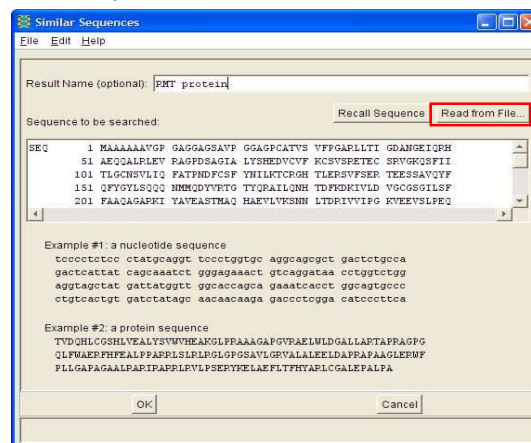
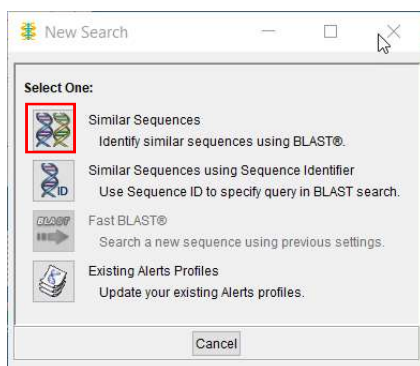


The Result Set Manager is the starting point

- To begin a new sequence search
- To review results of previous sequence searches

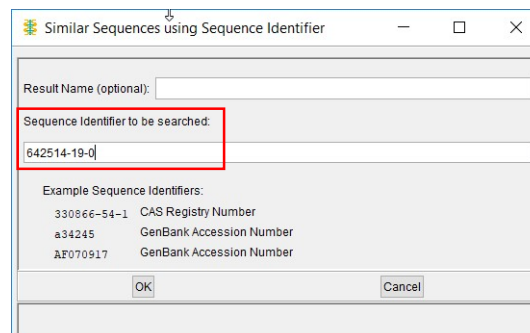
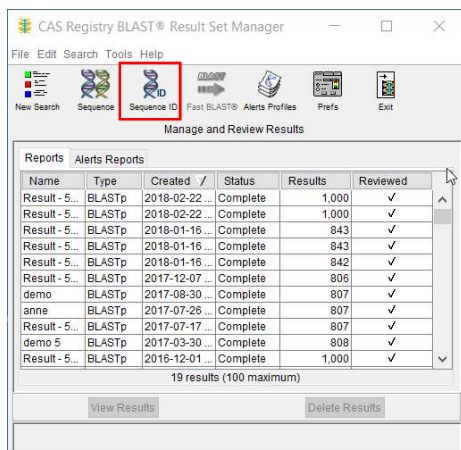
9

Select search option and input the search query Read from File Option (for text files)



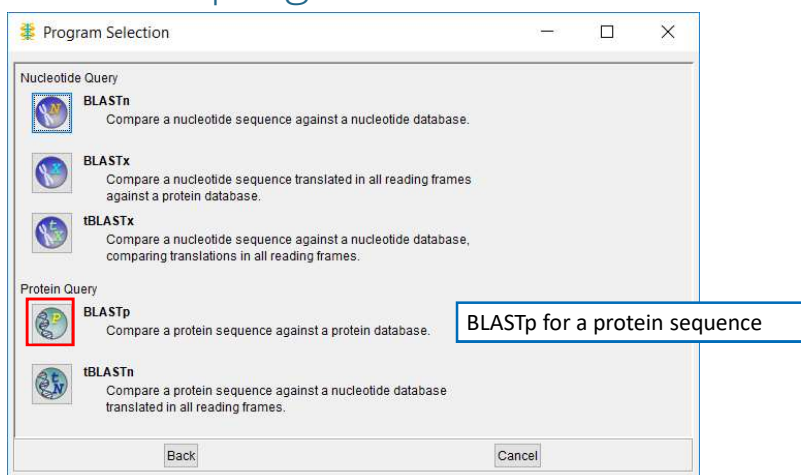
10

Alternatively: Search via the Sequence Identifier



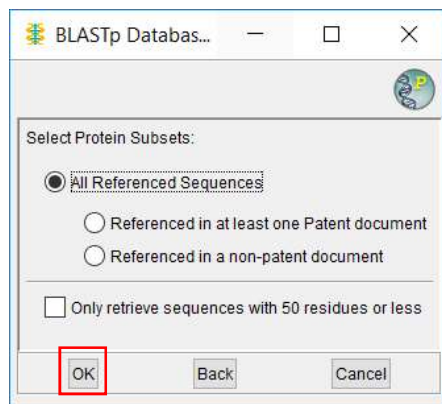
11

Select the BLAST program



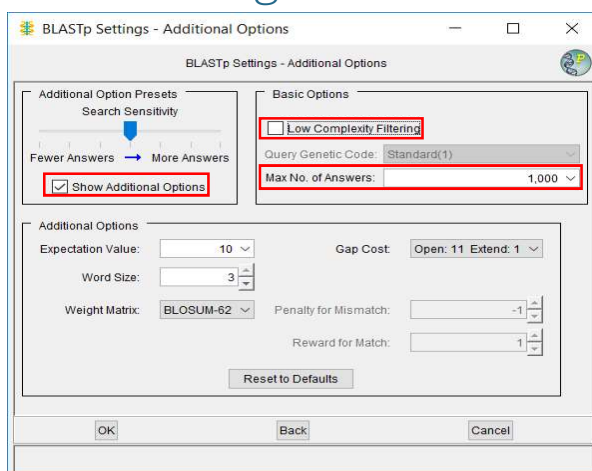
12

Select the segment of the database to search



13

Verify BLAST settings; Launch the search



Recommended settings for patent searches:

- Low Complexity Filtering – unchecked
- Max No. of Answers - 1000

14

NCBI recommended settings* for searching small sequence queries

Peptide sequences

- E-value: 20,000
- Word size: 2
- Matrix: PAM-30
- Gap cost: 9 and 1

NCBI definition of "small" is less than 16 peptides, or less than 20 nucleotides.

Nucleotide sequences

- E-value: 1,000
- Word size: 7
- Matrix: Leave as is
- Gap cost: n/a

* <http://www.ncbi.nlm.nih.gov/blast/Why.shtml>

15

View results

Highlight the result set to be viewed, and click on View Results.

CAS Registry BLAST® Result Set Manager

File Edit Search Tools Help

New Search Sequence Sequence ID Fast BLAST® Alerts Profiles Prefs Exit

Manage and Review Results

Name	Type	Created /	Status	Results	Reviewed
Result - 596408	BLASTp	2018-10-08 03:33 PM	Complete	810	
Result - 590178	BLASTp	2018-02-22 04:26 PM	Complete	1,000	✓
Result - 588814	BLASTp	2018-01-16 10:47 AM	Complete	843	✓
Result - 588812	BLASTp	2018-01-16 10:44 AM	Complete	843	✓
Result - 588810	BLASTp	2018-01-16 10:43 AM	Complete	842	✓
Result - 587840	BLASTp	2017-12-07 10:56 AM	Complete	806	✓
demo	BLASTp	2017-08-30 09:42 AM	Complete	807	✓
anne	BLASTp	2017-07-26 10:37 AM	Complete	807	✓
Result - 583490	BLASTp	2017-07-17 01:30 PM	Complete	807	✓
demo 5	BLASTp	2017-03-30 11:31 AM	Complete	808	✓
Result - 575970	BLASTp	2016-12-01 04:03 PM	Complete	1,000	✓

19 results (100 maximum)

View Results Delete Results

16

Evaluate the alignment report

Click on + to show details including sequence length, score, and percent identity.

CAS Registry BLAST® Report - Result - 596408

File Edit View Search Tools Help

Unique Sequences: 810 Redundant: 319 Selected Results: 69

Alignment Scores

Alignment Summary

Alignment Details

Seq ID	Score	Expect	Ident	Positives	Query	Subject
1266	0.0	(434009-21-9) 5	PN: W00244358 FIGURE: 4A-4B unclaimed sequence		1 MAAAAAAAAVFGAGGASAVFGAGFCATVSVFPGALLITGDANGEIQRREQQA 55	MAAAAAAAAVFGAGGASAVFGAGFCATVSVFPGALLITGDANGEIQRREQQA 55
1265	0.0	(1191194-79-2) 1931	PN: US7608413 SEQID: 2630 unclaimed protein		56 IRLVRAFGPSAGIALYSBEDVCFKCYSRRETECSRVRGQSFIIILGCSNVLIQ 110	IRLVRAFGPSAGIALYSBEDVCFKCYSRRETECSRVRGQSFIIILGCSNVLIQ 110

Get STN Data Script Cancel

Result complete.

17

STNext

FIZ Karlsruhe
Leibniz Institute for Information Infrastructure

CAS
A DIVISION OF THE
AMERICAN CHEMICAL SOCIETY

Select sequences of interest

Sequences can be selected:

- In groups, using a bar in the Alignment Scores
- Individually, by selecting the check box

To transfer the sequence data to STN, click the Get STN Data Script button.

CAS Registry BLAST® Report - Result - 596408

File Edit View Search Tools Help

Unique Sequences: 810 Redundant: 319 Selected Results: 69

Alignment Scores

Alignment Summary

Alignment Details

Seq ID	Score	Expect	Ident	Positives	Query	Subject
1266	0.0	(434009-21-9) 5	PN: W00244358 FIGURE: 4A-4B unclaimed sequence		1 MAAAAAAAAVFGAGGASAVFGAGFCATVSVFPGALLITGDANGEIQRREQQA 55	MAAAAAAAAVFGAGGASAVFGAGFCATVSVFPGALLITGDANGEIQRREQQA 55
1265	0.0	(1191194-79-2) 1931	PN: US7608413 SEQID: 2630 unclaimed protein		56 IRLVRAFGPSAGIALYSBEDVCFKCYSRRETECSRVRGQSFIIILGCSNVLIQ 110	IRLVRAFGPSAGIALYSBEDVCFKCYSRRETECSRVRGQSFIIILGCSNVLIQ 110
1264	0.0	(433617-62-4)	Drug-metabolizing enzyme DME-7 (human B-lycye clone 7486212CD1)			
1264	0.0	(642514-19-8)	Protein 27420 (human)			
1237	0.0	(942169-05-3) 69	PN: US20070141652 SEQID: 69 unclaimed protein			
1232	0.0	(696682-62-9)	Transcription factor SRC-2 (steroid receptor coactivator-2) (mouse)			
1226	0.0	(696686-43-8) 3	PN: US6743614 SEQID: 3 unclaimed protein			
1213	0.0	(867594-03-4)	Protein (Mus musculus strain C57BL/6J clone K230313H13 592-amino acid)			
1206	0.0	(1191194-79-1) 1930	PN: US7608413 SEQID: 2629 unclaimed protein			
1177	0.0	(863541-93-9)	Methyltransferase, transcriptional coactivator protein (arginine) (Rattus norv)			
1159	0.0	(863541-89-3)	Methyltransferase, transcriptional coactivator protein (arginine) (Rattus norv)			
1152	0.0	(300669-71-6)	Protein ORF X (human clone W0059473_SEQID_6160)			
1104	0.0	(863541-91-7)	Methyltransferase, transcriptional coactivator protein (arginine) (Rattus norv)			
1044	0.0	(1402873-20-4)	Protein arginine methyltransferase (Citellus punctatus gene prmt4)			
1041	0.0	(843011-10-9)	Genbank CAID0831			

Get STN Data Script Cancel

18

STNext

FIZ Karlsruhe
Leibniz Institute for Information Infrastructure

CAS
A DIVISION OF THE
AMERICAN CHEMICAL SOCIETY

Get STN Data (.scb file) and Save Alignments (.xss file)

The first screenshot shows the 'Get STN Data Sc...' dialog box. Under 'Retrieve the following data:', the 'Sequence Records' option is selected, which 'Retrieves Sequences from CAS Registry'. A checkbox for 'Transfer all alignment data for postprocessing' is also checked. The second screenshot shows a 'Save File As' dialog box where the file name is 'Result - 596408.scb' and the file type is 'Script files (*.scb)'. The third screenshot shows another 'Save File As' dialog box where the file name is 'Result - 596408' and the file type is 'STNnext Saved Sequences'. A note at the bottom of the third dialog states: 'Note: Use of the saved information is subject to copyright and data use restrictions.'

Alignment data needs to be transferred for post-processing.

19

STNext

FIZ Karlsruhe
Leibniz Institute for Information Infrastructure

CAS
A DIVISION OF THE
AMERICAN CHEMICAL SOCIETY

Logon to STNext

Select "Scripts" from the "My Files" menu

The screenshot shows the STNext web interface in a browser window. The URL is https://next.stn.org/stn/#/. The 'My Files' menu is open, showing options for 'Hist', 'Alerts', 'Databases', 'Transcripts', 'Structures', and 'Scripts'. The 'Scripts' option is highlighted with a red box. The main content area displays a transcript for '2018_0118_Transcript' with various news items and a 'Submit' button at the bottom.

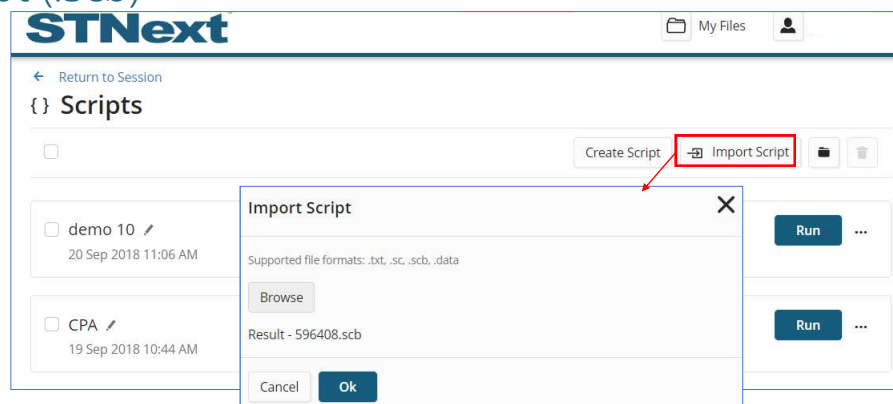
20

STNext

FIZ Karlsruhe
Leibniz Institute for Information Infrastructure

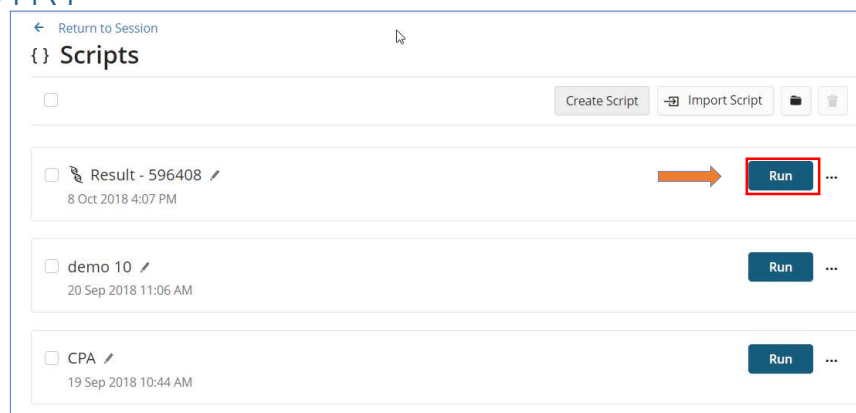
CAS
A DIVISION OF THE
AMERICAN CHEMICAL SOCIETY

Select Import Script and Browse to find Sequence Data Script (.scb)



21

Select RUN; the script will find sequences in REGISTRY



22

Search retrieves the CAS Registry Numbers® of the selected sequences

STNext

Transcript ON 2018_0116_Transcript

File REGISTRY

```

1 734466-64-9/RN
1 487816-23-9/RN
1 623060-39-9/RN
1 482525-15-5/RN
1 778247-89-5/RN
1 666530-59-2/RN
1 622914-88-9/RN
1 486892-00-6/RN
1 487298-31-7/RN
1 913790-09-7/RN
1 481808-09-7/RN
1 1208214-62-3/RN
1 809407-18-9/RN
1 736452-21-4/RN
1 809407-07-6/RN
1 811264-65-0/RN
1 1803235-92-8/RN

```

L7 69 L1 OR L2 OR L3 OR L4 OR L5 OR L6

enter command Submit Draw Scripts

23

STNext

FIZ Karlsruhe
Leibniz Institute for Information Infrastructure

CAS
A DIVISION OF THE
AMERICAN CHEMICAL SOCIETY

Search last L-number in CAPLUS and display references using the BIB ABS HITRN formats

STNext

Transcript ON 2018_0118_Transcript

File CAPLUS

USPTO MANUAL OF CLASSIFICATIONS THESAURUS ISSUE DATE: Dec 2015

CAPLUS now includes the comprehensive Cooperative Patent Classification (CPC). See [HELP CPC](#) for details.

CAS Information Use Policies apply and are available at:
<http://www.cas.org/legal/infopolicy>

This file contains CAS Registry Numbers for easy and accurate substance identification.

=> S L7

L8 39 L7

=> D BIB ABS HITRN 1-39

enter command Submit Draw Scripts

Session

```

L2 QUE (863541-95-1 OR 240488-50-0 OR 696682-62-9 OR 478329-39-4 O
L3 QUE (300609-71-6 OR 863541-91-7 OR 1402873-20-4 OR 843011-10-9
L4 QUE (1402873-22-6 OR 431384-20-2 OR 263494-05-9 OR 623634-37-7
L5 QUE (734466-64-9 OR 487816-23-9 OR 623060-39-9 OR 482525-15-5 O
L6 QUE (736452-21-4 OR 809407-07-6 OR 811264-65-0 OR 1803235-92-8)
L7 69 L1 OR L2 OR L3 OR L4 OR L5 OR L6

```

Entered CAPLUS 10:52:48 ON 09 OCT 2018

L8 39 S L7

24

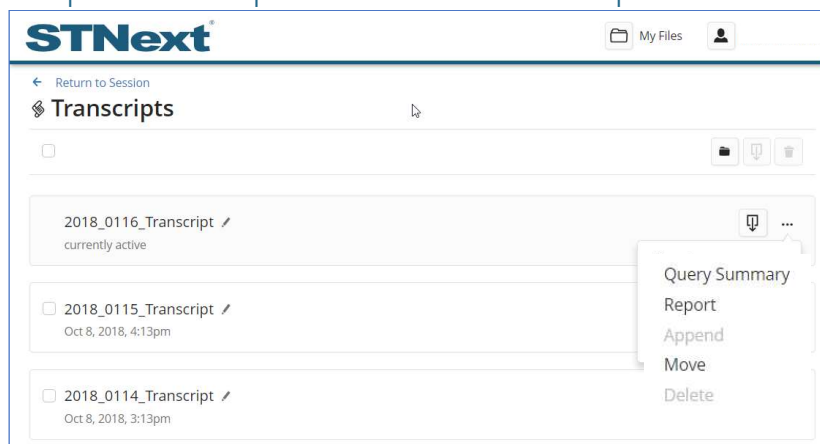
STNext

FIZ Karlsruhe
Leibniz Institute for Information Infrastructure

CAS
A DIVISION OF THE
AMERICAN CHEMICAL SOCIETY

Create BLAST Alignment Report

Step 1: Request a Report in the Transcripts Menu



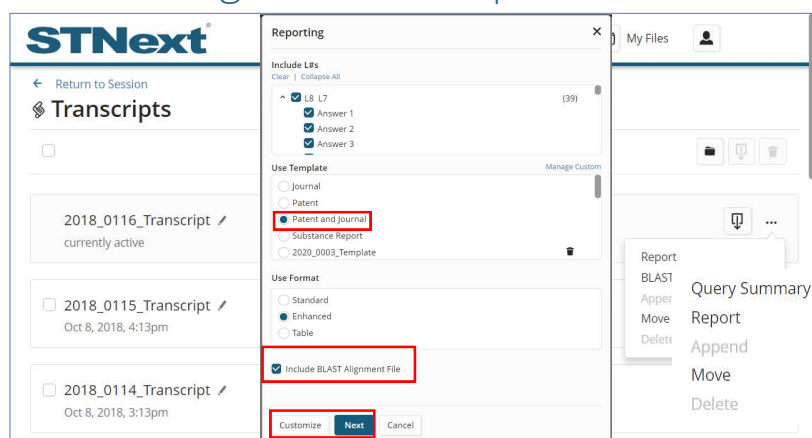
25

STNext

FIZ Karlsruhe
Leibniz Institute for Information Infrastructure

CAS
A DIVISION OF THE
AMERICAN CHEMICAL SOCIETY

Step 2: Select the L#, Template, Format, and check the "Include BLAST Assignment File" option



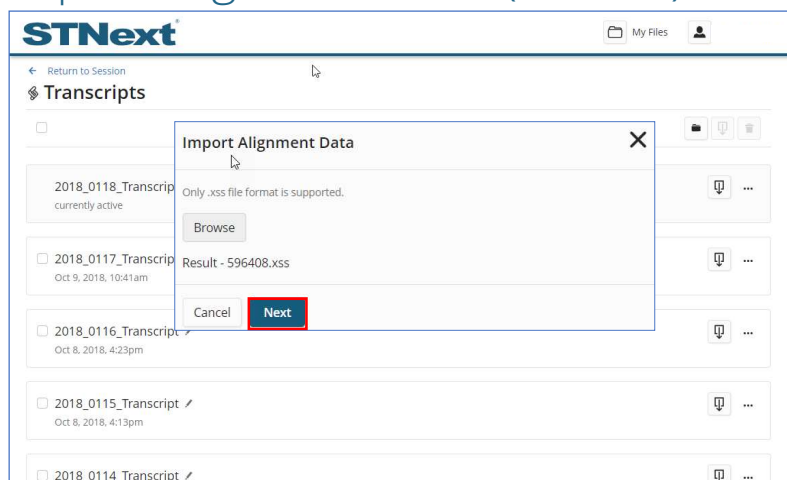
26

STNext

FIZ Karlsruhe
Leibniz Institute for Information Infrastructure

CAS
A DIVISION OF THE
AMERICAN CHEMICAL SOCIETY

Step 3: Import Alignment Data (.xss file)

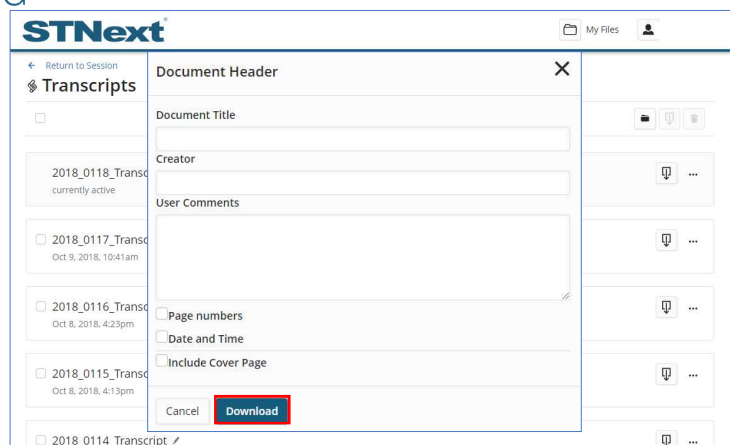


STNext

FIZ Karlsruhe
Leibniz Institute for Information Infrastructure

CAS
A DIVISION OF THE
AMERICAN CHEMICAL SOCIETY

Step 4: Add Document Header Options, and Download



STNext

FIZ Karlsruhe
Leibniz Institute for Information Infrastructure

CAS
A DIVISION OF THE
AMERICAN CHEMICAL SOCIETY

The BLAST alignment report contains the bibliographic record and sequence alignment

L8 Patent | English | 4/39

Crystal structure of human CARM1 methyltransferase for use in identification of modulators of CARM1 activity and rational drug design

[PatentPak PDF](#) | [Full-text](#)

Accession Number: 2008:1278320 CAPLUS

Inventor Name: Foreman, Kenneth William; Shaaban, Salam; Park, Frances E.; Sauder, Michael

Patent Assignee: OSI Pharmaceuticals, Inc., USA; SGX Pharmaceuticals, Inc.

Document Number: 149:488720

Family Accession Number Count: 1

Source: PCT Int. Appl., 324pp.

CODEN: PIXXD2

PatentPak Information:

Patent No.	Kind	Date	Language	Patent
WO 2008128050	A2	20081023	English	PDF

Patent Information:

Patent No.	Kind	Date	Application No.
WO 2008128050	A2	20081023	WO 2008-US60043
WO 2008128050	A3	20090226	
US 20080312298	A1	20081218	US 2008-101631

Index Terms and Role:

1072557-06-2D, complex with S-adenosyl-L-homocysteine 1072557-07-3D, complex with S-adenosyl-L-homocysteine

1072557-07-3

BLAST@ Alignment Data

Length = 353 Score = 745 Expect = 0.0

Score = 745 Expect = 0.0

Identities = 353/353 (100%) Positives = 353/353 (100%)

Query: 128 RHTLERSVFSERTEESSAVQYFQFYGYLSQQQNMMDYVRTGTYQRAILQNHDT 182

RHTLERSVFSERTEESSAVQYFQFYGYLSQQQNMMDYVRTGTYQRAILQNHDT

Subject: 1 RHTLERSVFSERTEESSAVQYFQFYGYLSQQQNMMDYVRTGTYQRAILQNHDT...55

Query: 183 FKDKIVLDVGCSSGILSFFPAQAGARKIYAVEASTMAQHAEVLVKSNLTDRIIV 237

FKDKIVLDVGCSSGILSFFPAQAGARKIYAVEASTMAQHAEVLVKSNLTDRIIV

Subject: 56 FKDKIVLDVGCSSGILSFFPAQAGARKIYAVEASTMAQHAEVLVKSNLTDRIIV 110

Query: 238 IPGKVEEVSLEPQVDIIISEPMGYMLFNERMLESYLHAKKYLKPSGNMFPTIGDV 292

IPGKVEEVSLEPQVDIIISEPMGYMLFNERMLESYLHAKKYLKPSGNMFPTIGDV

Subject: 111 IPGKVEEVSLEPQVDIIISEPMGYMLFNERMLESYLHAKKYLKPSGNMFPTIGDV 165

Query: 293 HLAFFTDEQLYMEQFTKANFWYQPSFHGVDLSALRGAAVDEYFRQFVVDTFDIRI 347

HLAFFTDEQLYMEQFTKANFWYQPSFHGVDLSALRGAAVDEYFRQFVVDTFDIRI

Subject: 166 HLAFFTDEQLYMEQFTKANFWYQPSFHGVDLSALRGAAVDEYFRQFVVDTFDIRI 220

29

STNext

FIZ Karlsruhe
Leibniz Institute for Information Infrastructure

CAS
A DIVISION OF THE
AMERICAN CHEMICAL SOCIETY

Substructure search on a short sequence

- Another search option, distinct from percent match or sequence match
- Many short sequences are in REGISTRY as drawn structures
- You can search a known sequence to find overlapping matches
- Use the drawing tool to indicate specific variations in a known sequence that you'd like to search

30

STNext

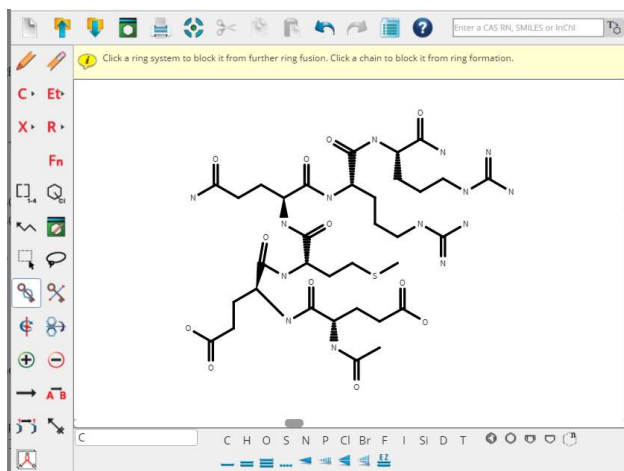
FIZ Karlsruhe
Leibniz Institute for Information Infrastructure

CAS
A DIVISION OF THE
AMERICAN CHEMICAL SOCIETY

Example: Argireline (616204-22-9), a topical mimetic of Botox

616204-22-9

The "text to structure" tool converts CAS Registry Numbers, SMILES and InChi strings into structures



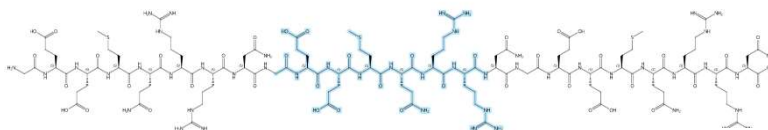
31

STNext

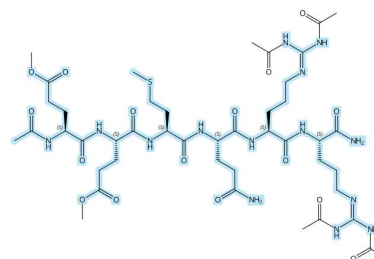
FIZ Karlsruhe
Leibniz Institute for Information Infrastructure

CAS
A DIVISION OF THE
AMERICAN CHEMICAL SOCIETY

A substructure search finds overlapping short sequences



Absolute stereochemistry shown



Absolute stereochemistry shown

32

STNext

FIZ Karlsruhe
Leibniz Institute for Information Infrastructure

CAS
A DIVISION OF THE
AMERICAN CHEMICAL SOCIETY

Agenda

- Intro
- Steps for a CAS Registry BLAST® Sequence Search
- Create a BLAST Alignment Report
- Substructure Search on a Short Sequence
- Database Overview
- Step-by-step through a multi-database BLAST search
- Post-processing of a multi-step BLAST search
- Overview of results
- Summary



33

STNext

FIZ Karlsruhe
Leibniz Institute for Information Infrastructure

CAS
A DIVISION OF THE
AMERICAN CHEMICAL SOCIETY

Sequence searchable databases on STN

- ❖ DGENE (Derwent GENESEQ™)*
- ❖ USGENE
- ❖ PCTGEN
- ❖ CAS REGISTRY *

*Sequences intellectually derived by indexers
(include unique sequences not disclosed in formal listings)

34

STNext

FIZ Karlsruhe
Leibniz Institute for Information Infrastructure

CAS
A DIVISION OF THE
AMERICAN CHEMICAL SOCIETY

DGENE (Derwent Geneseq)

- Value-added patent sequence data produced by Clarivate Analytics
 - Enhanced titles from DWPI
 - Concise one-line description of the sequence
 - Keyword indexing and abstract focused on sequence
 - Abstract providing information on sequence and context
 - Additionally: feature table (FEAT), patent sequence location (PSL),...
- Sequences from 1981 of the basic patents of the Derwent World Patents Index®, covering 47 patent-issuing authorities
- Nucleotides of 10 or more bases, amino acid sequences of 4 or more residues and primers and probes of any length
- Sequences intellectually derived by indexers
- Legal status data from INPADOCDB (D LS or LS2) directly displayable

35




FIZ Karlsruhe
Leibniz Institute for Information Infrastructure



CAS
A DIVISION OF THE
AMERICAN CHEMICAL SOCIETY

USGENE

- Protein (>3 aa) and nucleic acid (>9 nt) sequences from 1981 to date
- All available peptide and nucleic acid sequences from published applications and issued patents of USPTO
- USPTO consolidates four sources (/SSO)
- Bibliographic details including publication and priority details, assignee and inventor names
- Sequence details including one-line description, organism name, length, molecule type, sequence source, feature table and patent sequence location (PSL) from 2005 onwards
- Original title, abstract and claims text (ECLM searchable)
- Updated weekly, within 3 days of publication

36




FIZ Karlsruhe
Leibniz Institute for Information Infrastructure



CAS
A DIVISION OF THE
AMERICAN CHEMICAL SOCIETY

PCTGEN

- All peptide and nucleic acid sequences electronically submitted to WIPO, 2001 to present
- Bibliographic details including publication and application details, assignee and inventor names
- Sequence details include molecule type, organism, sequence length, feature table
- Original published application title
- Records created from image format sequence listings are clearly marked („...created by using OCR...“)
- Updated weekly, within 1 day of publication

37

STNextFIZ Karlsruhe
Leibniz Institute for Information InfrastructureCAS
A DIVISION OF THE
AMERICAN CHEMICAL SOCIETY

CAS REGISTRY

- Produced by CAS
- Sequences from over 3,000 life science journals and 63 patent authorities, including WO US EP JP DE GB FR RU CA IN KR CN
- Sequence details include sequence type, sequence length, nucleic acid type, 1 and 3 letter amino acid code
- Unique sequence types (e.g. cyclic peptides, peptide-metal complexes,...)
- Sequences linked to value-added CPlus records by RNs

38

STNextFIZ Karlsruhe
Leibniz Institute for Information InfrastructureCAS
A DIVISION OF THE
AMERICAN CHEMICAL SOCIETY

Agenda

- Intro
- Steps for a CAS Registry BLAST® Sequence Search
- Create a BLAST Alignment Report
- Substructure Search on a Short Sequence
- Database Overview
- Step-by-step through a multi-database BLAST search
- Post-processing of a multi-step BLAST search
- Overview of results
- Summary



39

STNext

FIZ Karlsruhe
Leibniz Institute for Information Infrastructure

CAS
A DIVISION OF THE
AMERICAN CHEMICAL SOCIETY

Step-by-step through a multi-database BLAST search

Search example:

Find all patents disclosing the gene *CBP1* from the soil bacterium *Serratia marcescens*. CBP21 has been described in degradation of crystalline cellulose and chitin.

NAME: *Chitin-binding protein precursor CBP1*
 GENE MODEL: *AY665558*
 ORGANISM: *Serratia marcescens strain BJL200*
 SEQUENCE TYPE: coding sequence (CDS)
 SEQUENCE LENGTH: 594 bp
 SEQUENCE: ATGAACAAAACCTCCCGTACCCTGCTCTCTCTGGGCTGCTGAGCGCGGCCATGTTCCGGCTTTCGCAACA
 GCGGAATGCCACGGTTATGTCGAATCGCCGGCCAGCCGCGCTATCAGTGCAAACCTGCAGCTCAACACG
 CAGTGCAGCGAGCGTGCAGTACGAACCGCAGAGCGTGCAGGGCCTGAAAGGCTTCCCGCAGGCCGCGCCG
 GCTGACGGCCATATCGCCAGCGCCGACAAGTCCACCTTCTCGAACTGGATCAGCAAACGCCGACGCGCT
 GGAACAAGCTCAACCTGAAAACCGGTCCGAACCTCTTACCTGGAAGCTGACCGCGCTCACAGCACCA...

40

STNext

FIZ Karlsruhe
Leibniz Institute for Information Infrastructure

CAS
A DIVISION OF THE
AMERICAN CHEMICAL SOCIETY

Two different procedures for sequence searching

- ❖ DGENE (Derwent Geneseq)
- ❖ USGENE
- ❖ PCTGEN



- ❖ CAS REGISTRY



41

STNext

FIZ Karlsruhe
Leibniz Institute for Information Infrastructure

CAS
A DIVISION OF THE
AMERICAN CHEMICAL SOCIETY

Sequence searching in DGENE/USGENE/PCTGEN

DGENE/USGENE/PCTGEN

- (0) FIL DGENE
- (1) Import sequence in STNext
- (2) Validate sequence in Biosequence Editor
- (3) Upload sequence in sequence-database
- (4) Verify if uploaded sequence is correct
- (5) Run BLAST search (and decide how many answers to keep)
- (6) Review search (e.g. D TRIAL ALIGN)
- (7) Run BLAST in USGENE and PCTGEN
- (8) Merge answer sets
- (9) Sort results (SCORE, IDENT)
- (10) Family sorting and display

42

STNext

FIZ Karlsruhe
Leibniz Institute for Information Infrastructure

CAS
A DIVISION OF THE
AMERICAN CHEMICAL SOCIETY

Import sequence for search in DGENE, USGENE and PCTGEN

1 => **FIL DGENE**

STNext

My Files

2 **Structures**

Import Biosequence

Import Structure

Import Biosequence File

Only .txt file format is supported.

Browse

Cancel Ok

3 Prepare the sequence query as a plain text file in a suitable text editor, e.g. Windows Notepad

43

STNext

FIZ Karlsruhe
Leibniz Institute for Information Infrastructure

CAS
A DIVISION OF THE
AMERICAN CHEMICAL SOCIETY

Supported sequence formats

```
MSSPSLKWCF TLNYSSAAER ENFLSLLKEE DVHYAVVGDE VAPATGQKHL
QGYSLLKKRI RLGGLKKKYG SRAHWEIARG TDEENSKYCS KGTLILELGF
PVVNGSNKRR ISEMVARS PD RMKIEQPEIF HRYQSVNKLK KFKEEFVHPC
LDSPWQIQLT EAIDEEPDDR SIIWVYGPYG NEGKSTYAKS LIKKDWFYTR
```

plain text format

```
>gi|5524211|gb|AAD44166.1| cytochrome b [Elephas maximus maximus]
LCLYTHIGRNIYYGSYLYSETWNTGIMLLITMATAFMGVLPFGQMSFWGATVITNLFSAIPYIGTNLVEW
WGGFSDKATLNRFFAFHFILPFTMVALAGVHLTF
```

FASTA format

```
1 acaagatgcc attgtccccc ggccctcctgc tgetgctgct
41 ctccggggcc acggccaaccg ctgcctcctgcc cctggagggt
81 ggccccaccg gccgagacag cgagcatatg caggaagcgg
121 caggaataag gaaaagcagc ctctcgactt tctcogcttg
```

GENBANK format

```
acaagatgcc attgtccccc ggccctcctgc tgetgctgct 40
ctccggggcc acggccaaccg ctgcctcctgcc cctggagggt 80
ggccccaccg gccgagacag cgagcatatg caggaagcgg 120
caggaataag gaaaagcagc ctctcgactt tctcogcttg 160
```

EMBL format

Save as
.TXT-file

44

STNext

FIZ Karlsruhe
Leibniz Institute for Information Infrastructure

CAS
A DIVISION OF THE
AMERICAN CHEMICAL SOCIETY

Sequences are stored under My Files/Structures

STNext

Return to Session

Structures

Import Biosequence Import Structure

CBP21 Jan 15, 2019, 11:19am

```
>Serratia marcescens CBP21 CDS 594bp
atgaacaaaacttccgtaccctgctctctctggcctgc
tgagcggcctatgttcgctttgcaacagggcaat
gcccacggttatgtgaatgccgcccagccgcctat
raateraactcaertraaraceraetecoraerct
```

CBP21 Jan 15, 2019, 11:23am

Validation Errors

```
>Serratia marcescens CBP21 fragment
cgcagccgcccggctgacggccatgccagcgc
gacaagtccac
>Serratia marcescens CBP21 CDS 594bp
ataaacaaraacttccctaccctctctctctgacct
```

45

Validate sequence in Biosequence editor

Edit Biosequence

CBP21

Save As Validate

Validation

Success
No errors detected

Upload Save Cancel

- Multiple sequences detected
 - Validation errors, e.g. invalid characters
- (see STNext Help for examples)

Upload sequence to STNext

Before Upload:
Enter DGENE, USGENE or PCTGEN

A sequence query L-number is automatically generated

```
>Serratia marcescens CBP21 CDS 594bp
atgaacaaaactcccgtacctctctctggcctgc
tgagcggccatgttggcttgcacaagcgaat
gcccacggttatgtcgaatcggccagcgcgctat
caoter.aaator.aotr.aar.aar.aotor.aor.aot
```

=> **FIL DGENE**
...
=> Uploading sequence file: CBP21

UPLOAD SUCCESSFULLY COMPLETED
L1 GENERATED

47

Use D LQUE to verify your sequence query

=> **D L1 LQUE**

```
L1 ANSWER 1 DGENE COPYRIGHT 2019 CLARIVATE ANALYTICS on STN
LQUE atgaacaaaactcccgtacctctctctggcctgctgagcggccatggtcggcgtttcg
caacaggcgaatgccacggttatgtcgaatcggcggccagccgcctatcagtgcactgag
ctcaacacgagtcggcagcgtgagtagcaaccgagagcgtcgagggcctgaaaggctcccg
caggccggcccggctgacggccatcgccagcgcgacaagtcaccttcttgcactggatcag
caaacggcgacgcgctggaacaagctcaacctgaaaacgggtccgaactccttacctggaagctg
accgcgctcacagcaccacagctggcgctattcatcaccaagcgaactgggacgcttcgag
ccgctgacccgcgcttcttgacctgacgcggttctgccagttcaacgacggcggcgcacatccc
gcccacaggtcaccaccagtgcaacataccggcagatcgacggttcgacgtgatccttgcc
gtgtgggacatagccgaccgctaacgccttctatcaggcgtcgacgtcaacctgagcaataa
```

Verify sequence
Any spaces, numbers or headers in the query are stripped out during Upload. Uploaded sequence queries may be up to 10,000 characters in length for BLAST search.

The sequence query is now ready for searching directly in DGENE, PCTGEN and USGENE using the L-number (regardless of database used for upload).

48

The RUN BLAST command

=> RUN **BLAST** **L1** **/SQN** **-F F** **BATCH**

BATCH mode: Enter an email address to receive a notification when search is complete. Use RUN GETBATCH to check on status or to receive results.

The **low complexity filter** can eliminate biologically uninteresting segments that have low compositional complexity. The filter is set to F (False, off) by /SQN -F F (recommended for patent sequence search). The default setting is T (True, on).

Protein search: RUN BLAST L1 /SQP
Nucleotide search: RUN BLAST L1 /SQN
Translated search: RUN BLAST L1 /TSQN

BLAST NCBI BLAST for advanced similarity searching (this seminar)
GETSIM FASTA based search algorithm; comparison of entire sequence length; more computational time required – use BATCH mode
GETSEQ Sequence Code Match (SCM) for simple sequence queries to find exact sequences or with controlled variations

49

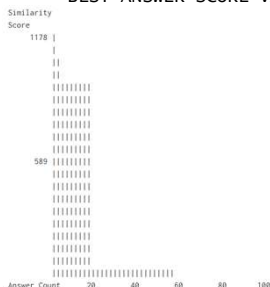
Run BLAST in DGENE

=> RUN BLAST L1 /SQN -F F

...

56 ANSWERS FOUND BELOW EXPECTATION VALUE OF 10.0

QUERY SELF SCORE VALUE IS 1178
BEST ANSWER SCORE VALUE IS 1178



Query Self Score: ideal score for a perfect answer match.

Best Answer Score: in this example, there is at least one perfect answer match to the query.

The graphic representation gives a count of hit sequences (x-axis) and similarity score (y-axis). The graph provides a visual idea about the proportion of similar and not so similar sequences in the answer set.

50

Decide how many answers to keep

ENTER EITHER THE NUMBER OF ANSWERS YOU WISH TO KEEP
OR ENTER MINIMUM PERCENT OF SELF SCORE FOLLOWED BY %
(BEST ANSWER PERCENTAGE OF SELF SCORE IS 100%)
ENTER (ALL) OR ? : **ALL**

```
L2 RUN STATEMENT CREATED
L2 56 ATGAACAAAACCTCCCGTACCCTGCTCTCTCTGGGCCTGCTGAGCGCGGC
CATGTTGGCGGTTTTGCAACAGGCGAATGCCACGGTTATGTCGAATCGC
CGGCCAGCCGCGCCTATCAGTGCAAATGCAGCTAACACGCAAGTGGCGG
AGCGTGCACTACGAACCGCAGAGCGTCGAGGGCCTGAAAGGCTTCCCGCA
GGCCGGCCCGGCTGACGGCCATATCGCCAGCGCCGACAAGTCCACCTTCT
TCGAACTGGATCAGCAAACGCCGACGCGCTGGAACAAGCTCAACCTGAAA
ACCGGTCGAACTCCTTTACCTGGAAGCTGACCGCGCTCACAGCACCAC
CAGCTGGCGCTATTTTCATACCAAGCCGAACTGGGACGCTTCGACGCGC
TGACCCGCGCTTCTTTGACCTGACGCGCTTTCGCAAGTTCAACGACGCG
GGCCCATCCCTGCCGCAAGTCAACCCAGTCAACATACCGGAGAG
TCGACGCGGTTTCGACGCTGATCCTTGCCGTGGGACATAGCCGACACCG
CTAACGCCTTCTATCAGGCGATCGACGCTCAACCTGAGCAAATAA/SQN.-
F F
```

Answer set arranged by accession number; to sort by descending similarity score, enter at an arrow prompt (=>) "sor score d".

In this example all results are kept. Preselect by setting cut off, e.g. 80% of the *Query Self Score* ($0.8 \times 1178 = 942$) can be used to select out the most relevant results. For this, type "80%" or "942".

For reviewing the DGENE results, sort before displaying (see next slide). Alternatively, refine search (e.g. priority date) and then sort and review.

51

FIZ Karlsruhe
Leibniz Institute for Information Infrastructure

CAS
A DIVISION OF THE
AMERICAN CHEMICAL SOCIETY

Review DGENE records

=> **SOR L2 SCORE D**
...

=> **D TRIAL SCORE ALIGN 1-20**

```
AN BFE50692 DNA DGENE
TI Enzymatically degrading polysaccharide, comprises contacting
polysaccharide with lytic polysaccharide monoxygenase, where the
degradation is carried out in reaction in presence of reducing agent
and hydrogen peroxide or means generating it.
DESC Serratia marcescens LPM010A coding DNA, SEQ ID 1.
KW Chitin-binding protein; LPM010A gene; Lytic polysaccharide
monoxygenase 10A; Lytic polysaccharide monoxygenase auxiliary
activity family 10A; alcohol; biofuel; degradation; ds; ethanol;
fermentation; gene; polysaccharide.
SQL 594
SCORE 1178 100% of query self score 1178
BLASTALIGN
Query = 594 letters
Length = 594
Score = 1178 bits (594), Expect = 0.0
Identities = 594/594 (100%)
Strand = Plus / Plus
```

```
Query: 1 atgaacaaaacttcccgtagcctgctctctctgggctgctgagcgcgccatgttcggc
|||||
Sbjct: 1 atgaacaaaacttcccgtagcctgctctctctgggctgctgagcgcgccatgttcggc...
```

The TRIAL display format is a free-of-charge display format. Based on the graphic representation of the similarity score, 20 records are selected for display.

The SCORE display field includes the percentage of the *Query Self Score*.

52

FIZ Karlsruhe
Leibniz Institute for Information Infrastructure

CAS
A DIVISION OF THE
AMERICAN CHEMICAL SOCIETY

Repeat sequence search in PCTGEN and USGENE

```

=> FIL USGENE
=> RUN BLAST L1 /SQN -F F

L3  RUN STATEMENT CREATED
L3  65 ATGAACAAAACCTCCCGTACCCTGCTCTCTCTGGGCCTGCTGAGCGCGGC
    CATGTTCCGGCGTTTCGCAACAGGCGAATGCCACGGTTATGTGCAATCGC
    ...

=> FIL PCTGEN
=> RUN BLAST L1 /SQN -F F


L4  RUN STATEMENT CREATED
L4  11 ATGAACAAAACCTCCCGTACCCTGCTCTCTCTGGGCCTGCTGAGCGCGGC
    CATGTTCCGGCGTTTCGCAACAGGCGAATGCCACGGTTATGTGCAATCGC
    ...

```

53

Sequence searching in DGENE/USGENE/PCTGEN

DGENE/USGENE/PCTGEN

- (0) FIL DGENE
- (1) Import sequence in STNext
- (2) Validate sequence in Biosequence Editor
- (3) Upload sequence in sequence-database
- (4) Verify if uploaded sequence is correct
- (5) Run BLAST search (and decide how many answers to keep)
- (6) Review search (e.g. D TRIAL ALIGN)
- (7) Run BLAST in USGENE and PCTGEN
- (8) Merge answer sets
- (9) Sort results (SCORE, IDENT) (> L6) 
- (10) Family sorting and display (>L7)

54

Merge sequence results from DGENE, USGENE and PCTGEN

=> SET DUPORDER FILE

SET COMMAND COMPLETED

=> DUP IDE L2 L3 L4

FILE 'DGENE' ENTERED AT 13:06:37 ON 17 JAN 2019
 COPYRIGHT (C) 2019 CLARIVATE ANALYTICS
 FILE 'USGENE' ENTERED AT 13:06:37 ON 17 JAN 2019
 COPYRIGHT (C) 2019 SEQUENCEBASE CORP
 FILE 'PCTGEN' ENTERED AT 13:06:37 ON 17 JAN 2019
 COPYRIGHT (C) 2019 WIPO

PROCESSING COMPLETED FOR L2

PROCESSING COMPLETED FOR L3

PROCESSING COMPLETED FOR L4

L5 132 DUP IDE L2 L3 L4 (INCLUDES 0 SETS OF DUPLICATES)
 ANSWERS '1-56' FROM FILE DGENE
 ANSWERS '57-121' FROM FILE USGENE
 ANSWERS '122-132' FROM FILE PCTGEN

=> SOR L5 SCORE D IDENT D

PROCESSING COMPLETED FOR L6

L6 132 SOR L5 SCORE D IDENT D

SET DUPORDER FILE first to retrieve answers according file entries. By default, the answers are arranged in reverse chronological order.

DUPLICATE IDENTIFY (DUP IDE) is used here to create a single multi-file L-number in preferred file order

The multi-file answer set is sorted by descending similarity score and identity. Alternatively, the answer set can be sorted by BLAST identity (SOR IDENT D) or similarity score (SOR SCORE D) alone. (D...Descending)

55

FIZ Karlsruhe
Leibniz Institute for Information Infrastructure

CAS
A DIVISION OF THE
AMERICAN CHEMICAL SOCIETY

Optional: Review results with an alignment (1)

=> D L6 BIB SCORE ALIGN

L6 ANSWER 1 OF 132 DGENE COPYRIGHT 2019 CLARIVATE ANALYTICS on STN
 AN BFE50692 DNA DGENE Full-text
 TI Enzymatically degrading polysaccharide, comprises contacting polysaccharide with lytic polysaccharide monoxygenase, where the degradation is carried out in reaction in presence of reducing agent and hydrogen peroxide or means generating it.
 IN Bissaro B; Eijsink V; Vaaje-kolstad G
 PA (UYNO-N) UNIV NORWEGIAN LIFE SCI. (INRG) INRA INST NAT RECH AGRONOMIQUE.
 PI WO 2018060498 A1 20180405 135
 AI WO 2017-EP74904 20170929
 PRAI GB 2016-16707 20160930
 GB 2017-5056 20170329
 PSL Disclosure; SEQ ID NO 1
 DT Patent
 LA English
 OS 2018-26087T [28]
 CR P-PSDB: BFE50693
 GENBANK: AY665558.1
 NCBI: gi52854326
 DESC Serratia marcescens LPM010A coding DNA, SEQ ID 1.

AN (Accession number) and MTY (molecule type), here DNA

Enhanced title from DWPI

(PACO) Patent Assignee Code

PSL (Patent sequence location): claim, disclosure or example

OS (Other source): accession number of corresponding DWPI record

CR: Internal cross reference (e.g. BFE50693 was disclosed as the protein encoded by BFE50694) and external cross references

DESC: Concise, one-line description

56

FIZ Karlsruhe
Leibniz Institute for Information Infrastructure

CAS
A DIVISION OF THE
AMERICAN CHEMICAL SOCIETY

Optional: Review results with an alignment (1)

```

...
SCORE 1178      100% of query self score 1178
BLASTALIGN
  Query = 594 letters
  Length = 594
  Score = 1178 bits (594), Expect = 0.0
  Identities = 594/594 (100%)
  Strand = Plus / Plus

```

Query Self Score and percentage

BLAST Percent Identity (IDENT)

```

Query: 1   atgaacaaaacttcccgtagcctctctctgggctgagcgcggccatgttcggc
          |||
Sbjct: 1   atgaacaaaacttcccgtagcctctctctgggctgagcgcggccatgttcggc

Query: 61  gtttcgcaacaggcgaatgccacggttatgtcgaatcgccggccagccgcctatcag
          |||
Sbjct: 61  gtttcgcaacaggcgaatgccacggttatgtcgaatcgccggccagccgcctatcag

Query: 121 tgcaaaactgcagctcaacacgcagtgccgagcgtgcagtacgaaccgcagagcgtcgag
          |||
Sbjct: 121 tgcaaaactgcagctcaacacgcagtgccgagcgtgcagtacgaaccgcagagcgtcgag

Query: 181 ggctgaaaggcttcccgaggccggcccggctgacggccatatcgccagccgacaag
          |||
Sbjct: 181 ggctgaaaggcttcccgaggccggcccggctgacggccatatcgccagccgacaag

```

57

STNext

FIZ Karlsruhe
Leibniz Institute for Information Infrastructure

CAS
A DIVISION OF THE
AMERICAN CHEMICAL SOCIETY

Use the Patent Family Manager to display the results

History CAS Lexicon Databases

Entered Multiple files 11:35:30 ON 15 JAN 2019

L5 132 DUP IDE L2 L3 L4
(INCLUDES 0 SETS OF
DUPLICATES)

L6 132 SOR L5 SCORE D IDENT D

1 Patent Family Manager

- The first record of each family will be displayed.
- Optional: Choose different display format for other family members, e.g. TRIAL SCORE ALIGN

Patent Family Manager

- Extract the first member basics from each patent family (limit 5000 answers)
- Remove twin multiple basics from patent families in CA/CAPlus (limit 5000 answers with Chemical Indexing Equivalent tag)
- Custom Display Format (limit 5000 answers)
 - First Member of Each Family
 - BIB SQL SCORE IDENT ALIGN 2
 - Ex: bib abs
 - Additional Member of Each Family
 - 3
 - Ex: ti an

STNext is unable to provide cost estimates for this action.

Continue without an estimate

Cancel Submit 4

58

STNext

FIZ Karlsruhe
Leibniz Institute for Information Infrastructure

CAS
A DIVISION OF THE
AMERICAN CHEMICAL SOCIETY

The Patent Family Manager organizes results in extended patent families

=> FSORT L6

...
L7

132 FSO L6

16 Multi-record Families	Answers 1-131
Family 1	Answers 1-2
Family 2	Answers 3-10
Family 3	Answers 11-15
Family 4	Answers 16-28

1 Individual Record	Answer 132
0 Non-patent Records	

The Patent Family Manager starts automatically with an FSORT, which means all patent records will be sorted in extended patent families.

The user does not have to type in these commands!

=> DIS L7 PFAM=1 1 BIB SQL SCORE IDENT ALIGN

```
L7 ANSWER 1 OF 132 DGENE COPYRIGHT 2019 CLARIVATE ANALYTICS ON STN
FAMILY1
AN BFE50692 DNA DGENE Full-text
TI Enzymatically degrading polysaccharide, comprises contacting
polysaccharide with lytic polysaccharide monoxygenase, where the
degradation is carried out in reaction in presence of reducing agent
and hydrogen peroxide or means generating it.
IN Bissaro B; Eijsink V; Vaaje-Kolstad G
```

16 extended patent families and 1 individual record.

The first record of each family will be displayed. This record (answer 1) is from Family 1 containing 2 records.

59

STNext

FIZ Karlsruhe
Leibniz Institute for Information Infrastructure

CAS
A DIVISION OF THE
AMERICAN CHEMICAL SOCIETY

Expectation Value (-E)

Expectation value (E value) is the statistical significance threshold for reporting matches against a sequence database. The E value can be any positive number, and the default value is 10. This means that 10 matches may be expected to be found merely by chance. In general E value is lowered to make the search more precise and raised to retrieve more answers.

Word Size (-W)

Word Size is the length of the character string fragments of a sequence query which are used as the basis for a BLAST search. For SQN the default is 11 and the range 7-23. For all other BLAST searches the default is 3 and the range 2-3. For short search queries, reducing the default word size can give improved search results.

60

STNext

FIZ Karlsruhe
Leibniz Institute for Information Infrastructure

CAS
A DIVISION OF THE
AMERICAN CHEMICAL SOCIETY

Low Complexity Filtering (on by default) (-F)

The low complexity filter can eliminate biologically uninteresting segments that have low compositional complexity and are statistically significant, as determined by specific programs for peptide or nucleotide sequences in nature. Filtering is applied to the query sequence and is indicated by a series of Xs for peptide sequences and Ns for nucleotide sequences. Low Complexity Filtering can be turned off (i.e., set to F - false).

Peptide similarity matrices (-M)

For peptide based searches SQP and TSQN the advanced options provide additional scoring matrices to the default BLOSUM62

61

NCBI guidelines for using Advanced Settings for peptide sequence searching

Query Length	Matrix	Gap costs
<35	PAM30	(9,1)
35-50	PAM70	(10,1)
50-85	BLOSUM80	(10,1)
>85	BLOSUM62	(11,1) (BLAST default)

62

BLAST search with different parameters

Search example:

Find all patents disclosing the sequence –

iskdgst

63

BLAST search in DGENE with default settings

Transcript ON Changing BLAST values

File DGENE

```
=> run blast iskdgst/sqp -f f
```

...

BLAST Version 2.2.20

X2: 38 (14.6 bits)

X3: 64 (24.7 bits)

S1: 44 (21.6 bits)

S2: 68 (30.8 bits)

Total Execution Time: 2.5974

NO ANSWERS FOUND BELOW EXPECTATION VALUE OF 10.0

64

SORT and DISPLAY

```
=> sor score d ident d
```

```
PROCESSING COMPLETED FOR L4
L5          12 SOR L4 SCORE D IDENT D
```

```
=> d kwic align score 1-3
```

```
L5 ANSWER 1 OF 12 DGENE COPYRIGHT 2020 CLARIVATE ANALYTICS on STN
AB The present invention relates to a novel dual binding moiety comprising a
non-complementarity determining region (CDR) loop and a cleavable
linker, where the moiety is capable of interacting with a domain by
specific binding or by...
BLASTALIGN
Query = 7 letters
Length = 117
Score = 23.5 bits (48), Expect = 7e-05
Identities = 7/7 (100%), Positives = 7/7 (100%)
Query: 1 ISKDGST 7
      ISKDGST
Sbjct: 51 ISKDGST 57
SCORE 24      100% of query self score 24
```

67

FIZ Karlsruhe
Leibniz Institute for Information Infrastructure

CAS
A DIVISION OF THE
AMERICAN CHEMICAL SOCIETY

Summary

- STNext offers excellent combination of value-added and first level sequence data
- DGENE and REGISTRY are the “industry-standard” databases
- STNext ideally supports data cross over to other databases and combine data from several databases in one report
- DGENE, REGISTRY, USGENE and PCTGEN are all required for a comprehensive search

68

FIZ Karlsruhe
Leibniz Institute for Information Infrastructure

CAS
A DIVISION OF THE
AMERICAN CHEMICAL SOCIETY

For more information ...

CAS

help@cas.org

Support and Training:

www.cas.org

FIZ Karlsruhe

helpdesk@fiz-karlsruhe.de

Support and Training:

www.stn-international.de

STNext

 **FIZ Karlsruhe**
Leibniz Institute for Information Infrastructure

 **CAS**
A DIVISION OF THE
AMERICAN CHEMICAL SOCIETY