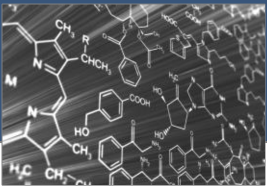
Biosequence Searching in

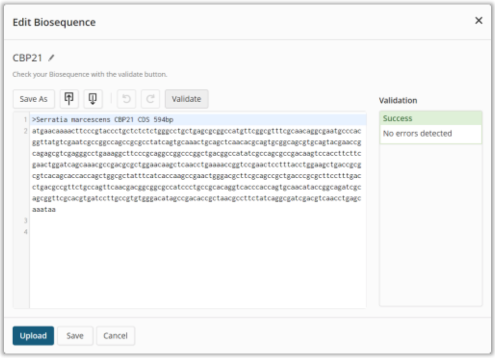
CAS STNext

Access essential sequence databases for comprehensive searches

- CAS REGISTRY^{SM}
- DGENE
- PCTGEN
- USGENE®

(19) United States
(12) Patent Application Publication
Andreini et al.
(54) 3-AMINO-5-PHENYL-5,6-DIHYDRO-2H-[1,4] OXAZINES
Inventors





Edit Biosequence ×

CBP21

Check your Biosequence with the validate button.

```
>Serratia marcescens CBP21 CDS 594bp
atgaacaaaacttcccgtaccctgctctctctgggcctgctgagcgcggccatgttcggcgtttcgcaacaggcgaatgcccac
ggttatgtcgaatcgccggccagccgcgcctatcagtgcaaactgcagctcaacacgcagtgcggcagcgtgcagtacgaaccg
cagagcgtcgagggcctgaaaggcttcccgcaggccggcccggctgacggccatatcgccagcgccgacaagtccaccttcttc
gaactggatcagcaaacgccgacgcgctggaacaagctcaacctgaaaaccggtccgaactcctttacctggaagctgaccgcg
cgtcacagcaccaccagctggcgctatttcatcaccaagccgaactgggacgcttcgcagccgctgacccgcgcttcctttgac
ctgacgccgttctgccagttcaacgacggcggcgccatccctgccgcacaggtcacccaccagtgcaacataccggcagatcgc
agcggttcgcacgtgatccttgccgtgtgggacatagccgacaccgctaacgccttctatcaggcgatcgacgtcaacctgagc
aaataa
```

Validation

Success

No errors detected

FIZ Karlsruhe
Leibniz Institute for Information Infrastructure

CAS
A division of the
American Chemical Society

Authority Sources in STNext

STNext allows you to combine and de-dup your results from the following sequence searchable databases:

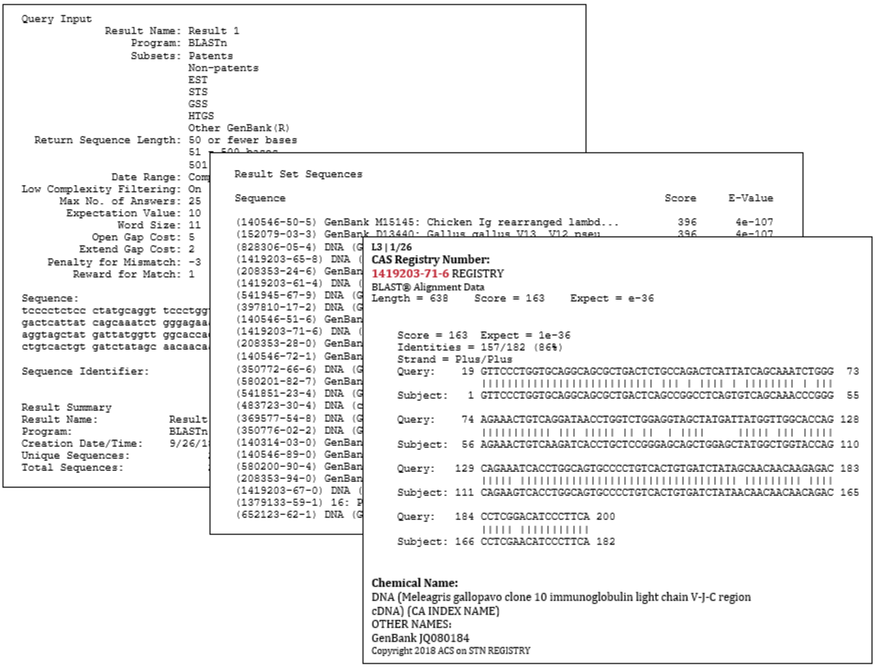
CAS REGISTRY*

DGENE (Derwent Geneseq^{TM})

PCTGEN

USGENE

By using a combination of these databases, you can locate even "hard-to-find" sequences. And, with STN you can gather them all into one, combined report complete with alignment and score matching details.



```
Query Input
             Result Name: Result 1
                 Program: BLASTn
                 Subsets: Patents
                          Non-patents
                          EST
                          STS
                          GSS
                          HTGS
                          Other GenBank(R)
  Return Sequence Length: 50 or fewer bases
                          51 - 500 bases
                          501
              Date Range: Com
Low Complexity Filtering: On
      Max No. of Answers: 25
       Expectation Value: 10
               Word Size: 11
           Open Gap Cost: 5
         Extend Gap Cost: 2
    Penalty for Mismatch: -3
        Reward for Match: 1

Sequence:
tcccctctcc ctatgcaggt tccctgg
gactcattat cagcaaatct gggagaa
aggtagctat gattatggtt ggcacca
ctgtcactgt gatctatagc aacaaca

Sequence Identifier:


Result Summary
Result Name:            Result
Program:                BLASTn
Creation Date/Time:     9/26/1
Unique Sequences:
Total Sequences:
```

```
Result Set Sequences

Sequence                                                     Score     E-Value

(140546-50-5) GenBank M15145: Chicken Ig rearranged lambd...   396     4e-107
(152079-03-3) GenBank D13440: Gallus gallus V13, V12 pseu      396     4e-107
(828306-05-4) DNA (G
(1419203-65-8) DNA (
(208353-24-6) GenBan
(1419203-61-4) DNA (
(541945-67-9) DNA (G
(397810-17-2) DNA (G
(140546-51-6) GenBan
(1419203-71-6) DNA (
(208353-28-0) GenBan
(140546-72-1) GenBan
(350772-66-6) DNA (G
(580201-82-7) GenBan
(541851-23-4) DNA (G
(483723-30-4) DNA (c
(369577-54-8) DNA (G
(350776-02-2) DNA (G
(140314-03-0) GenBan
(140546-89-0) GenBan
(580200-90-4) GenBan
(208353-94-0) GenBan
(1419203-67-0) DNA (
(1379133-59-1) 16: P
(652123-62-1) DNA (G
```

L3 | 1/26
CAS Registry Number:
1419203-71-6 REGISTRY
BLAST® Alignment Data

```
Length = 638    Score = 163    Expect = e-36


    Score = 163  Expect = 1e-36
    Identities = 157/182 (86%)
    Strand = Plus/Plus
    Query:    19 GTTCCCTGGTGCAGGCAGCGCTGACTCTGCCAGACTCATTATCAGCAAATCTGGG  73
                 |||||||||||||||||||||||||| ||| | |||| | |||||||| | |||
    Subject:   1 GTTCCCTGGTGCAGGCAGCGCTGACTCAGCCGGCCTCAGTGTCAGCAAACCCGGG  55

    Query:    74 AGAAACTGTCAGGATAACCTGGTCTGGAGGTAGCTATGATTATGGTTGGCACCAG 128
                 ||||||||||| ||| ||||| || ||  | ||||     ||||| ||| |||||
    Subject:  56 AGAAACTGTCAAGATCACCTGCTCCGGGAGCAGCTGGAGCTATGGCTGGTACCAG 110

    Query:   129 CAGAAATCACCTGGCAGTGCCCCTGTCACTGTGATCTATAGCAACAACAAGAGAC 183
                 ||||| |||||||||||||||||||||||||||||||||| ||||||||| ||||
    Subject: 111 CAGAAGTCACCTGGCAGTGCCCCTGTCACTGTGATCTATAACAACAACAACAGAC 165

    Query:   184 CCTCGGACATCCCTTCA 200
                 ||||| |||||||||||
    Subject: 166 CCTCGAACATCCCTTCA 182
```

Chemical Name:
DNA (Meleagris gallopavo clone 10 immunoglobulin light chain V-J-C region cDNA) (CA INDEX NAME)
OTHER NAMES:
GenBank JQ080184
Copyright 2018 ACS on STN REGISTRY

*Biosequence searching in the CAS REGISTRY database leverages the CAS Registry BLAST® client – which is a freely available utility.

Before You Start Your Search

Save your sequence data in .TXT files. The sequence formats supported by STNext include: plain text, FASTA, GENBANK, and EMBL.

```
MSSPSLKWCF TLNYSSAAER ENFLSLLKEE DVHYAVVGDE VAPATGQKHL
QGYLSLKKRI RLGGLKKKYG SRAHWEIARG TDEENSKYCS KGTLILELGF
PVVNGSNKRK ISEMVARSPD RMKIEQPEIF HRYQSVNKLK KFKEEFVHPC
LDSPWQIQLT EAIDEEPDDR SIIWVYGPYG NEGKSTYAKS LIKKDWFYTR
```

```
>gi|5524211|gb|AAD44166.1| cytochrome b [Elephas
maximus maximus]
LCLYTHIGRNIYYGSYLYSETWNTGIMLLLITMATAFMGYVLPWGQMSFWGATV
ITNLFSAIPYIGTNLVEWIWGGFSVDKATLNRFFAFHFILPFTMVALAGVHLTF
```

```
  1 acaagatgcc attgtccccc ggcctcctgc tgctgctgct
 41 ctccggggcc acggccaccg ctgccctgcc cctggagggt
 81 ggccccaccg gccgagacag cgagcatatg caggaagcgg
121 caggaataag gaaaagcagc ctcctgactt tcctcgcttg
```

```
acaagatgcc attgtccccc ggcctcctgc tgctgctgct      40
ctccggggcc acggccaccg ctgccctgcc cctggagggt      80
ggccccaccg gccgagacag cgagcatatg caggaagcgg      120
caggaataag gaaaagcagc ctcctgactt tcctcgcttg      160
```

NOTE: Any spaces and numbers at the beginning or end of a line in the imported .TXT file will be stripped out when the sequence is uploaded into the session.

Prior to searching biosequences in CAS REGISTRY, download and install the CAS Registry BLAST client.

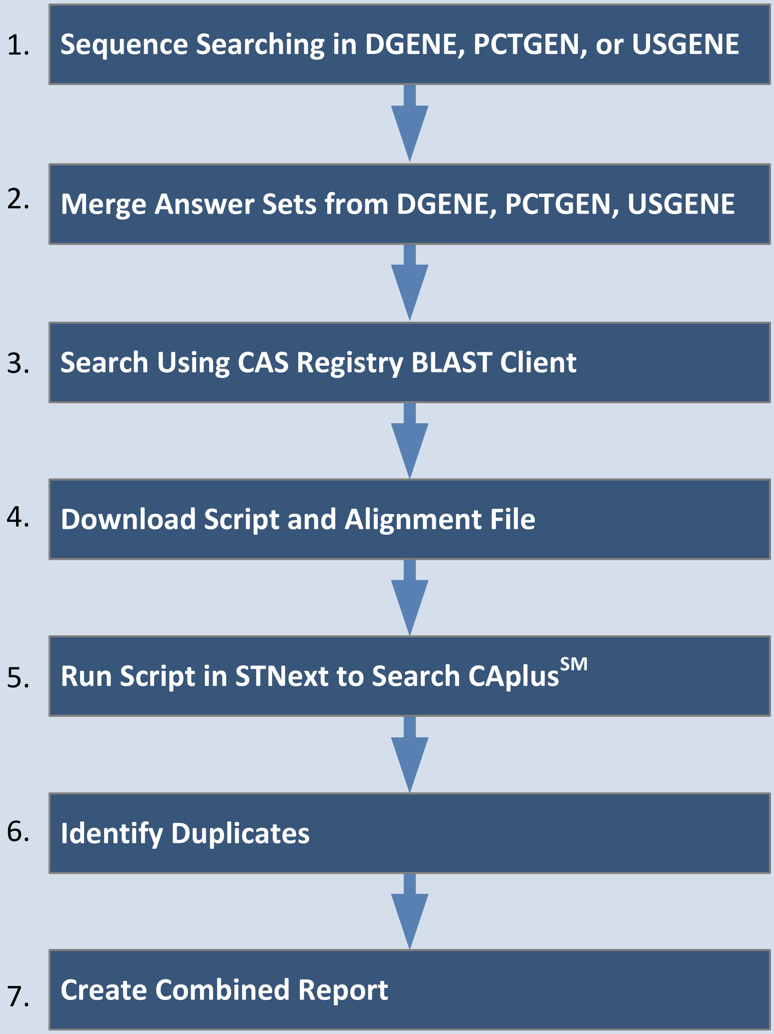
https://next.stn.org/stn/downloads/blast-download.html

NOTE: Your system administrator may need to assist you if you do not have installation privileges on your PC.

Multi-database BLAST Search Strategy

Due to differences across databases, there are two different procedures for sequence searching in STNext. This example strategy walks through the steps of a comprehensive search that leverages all of the sequence searchable databases in STNext.

Example Search Scenario:
Find all **patents** disclosing the gene CBP1 from the soil bacterium *Serratia marcescens* with a minimum overall homology of 80%.

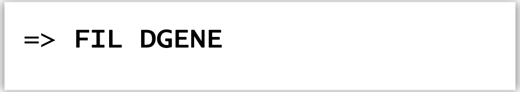


1. Sequence Searching in DGENE, PCTGEN, or USGENE
2. Merge Answer Sets from DGENE, PCTGEN, USGENE
3. Search Using CAS Registry BLAST Client
4. Download Script and Alignment File
5. Run Script in STNext to Search CAplus^{SM}
6. Identify Duplicates
7. Create Combined Report

Sequence Searching in DGENE, PCTGEN, or USGENE

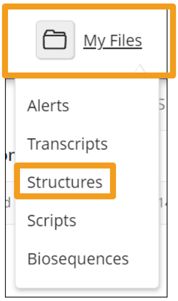
Login to STNext to get started. Use your STN login credentials at next.stn.org (Contact your STN helpdesk if you need assistance.)

1. File into DGENE.

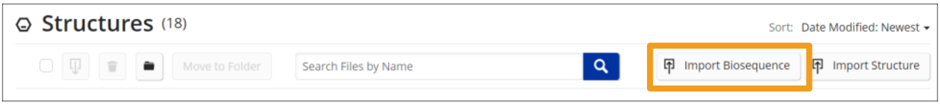


```
=> FIL DGENE
```

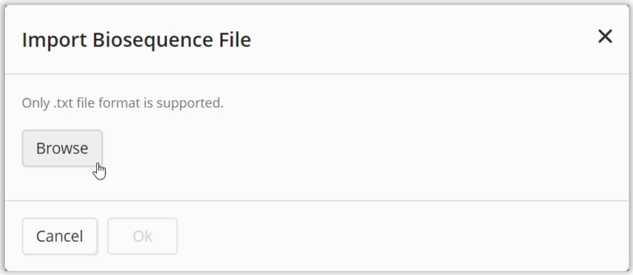
2. Select **Structures** from the My Files menu.



3. Click the **Import Biosequence** button.



4. Browse to locate the .TXT file for the sequence



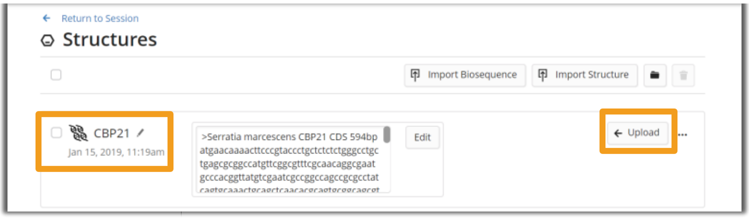
Import Biosequence File

Only .txt file format is supported.

5. Click **OK** to import the file.

Sequence Searching in DGENE, PCTGEN, USGENE

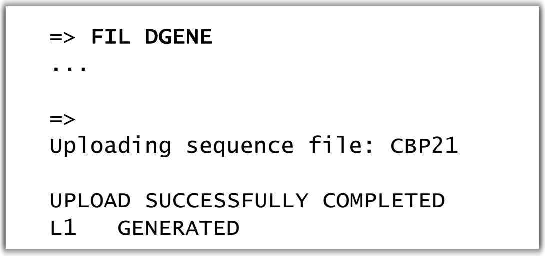
Sequences are indicated by the sequence icon and stored under My Files/Structures. Uploaded sequence queries may be up to 10,000 characters in length for BLAST search.



```
>Serratia marcescens CBP21 CDS 594bp
atgaacaaaacttcccgtaccctgctctctgggcctgc
tgagcgcggccatgttcggcgtttcgcaacaggcgaat
gcccacggttatgtcgaatcgccggccagccgcgcctat
cagtgcaaactgcagctcaacacgcagtgcggcagcgt
```

6. Click the **Upload** button.

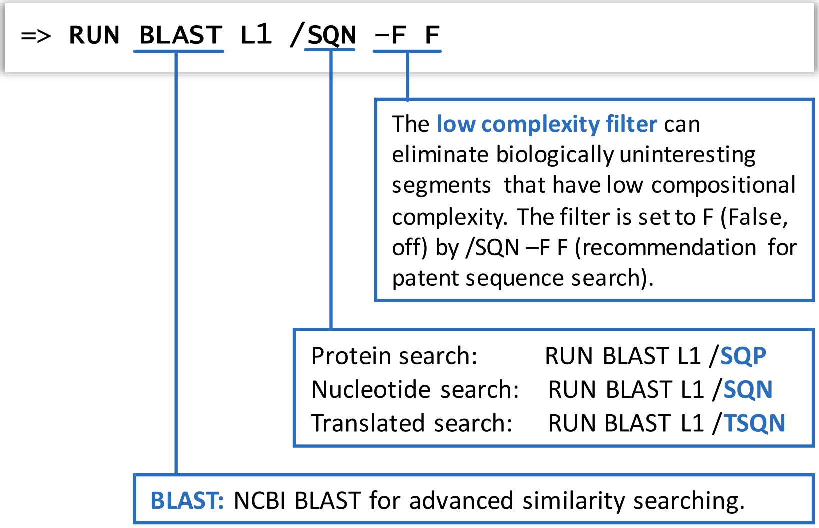
   A sequence query L-number is automatically generated.



```
=> FIL DGENE
...

=>
Uploading sequence file: CBP21

UPLOAD SUCCESSFULLY COMPLETED
L1    GENERATED
```

(Option to verify your sequence using the D LQUE at this point.)

7. Enter the BLAST command using the sequence query L-number.



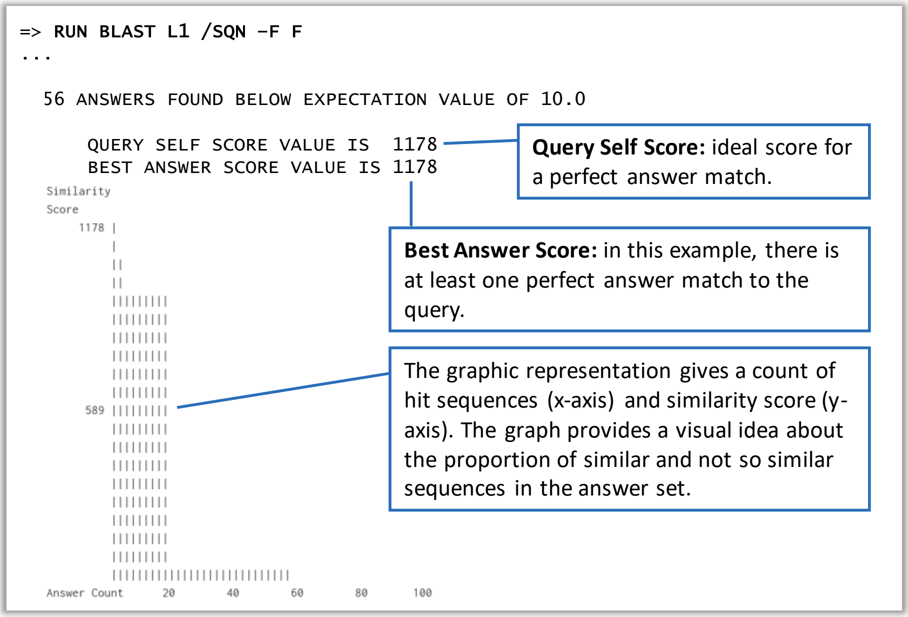
```
=> RUN BLAST L1 /SQN -F F
```

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 (recommendation for patent sequence search).

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.

8. Evaluate the answer set.



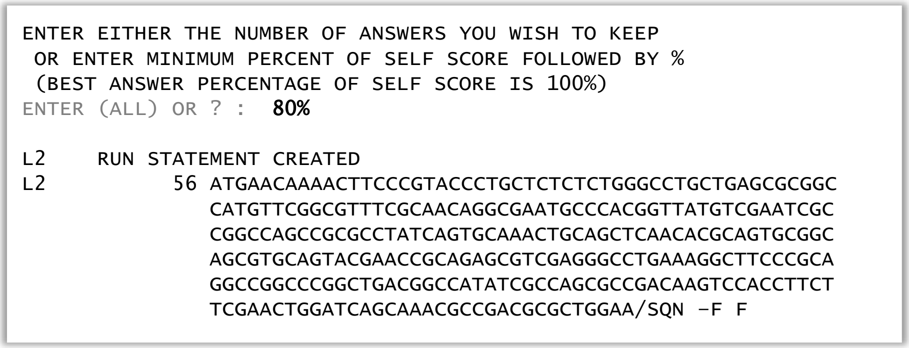
```
=> RUN BLAST L1 /SQN -F F
...

  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.

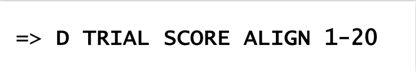
The graphic representation gives a count of hit sequences (x-axis) and similarity score (y-axis). The graph provides a visual idea about the proportion of similar and not so similar sequences in the answer set.

9. Decide how many answers to keep. You can choose a number of answers, all answers or a minimum percent self score.
(This example uses 80% minimum score.)



```
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 ? :  80%

L2     RUN STATEMENT CREATED
L2            56 ATGAACAAAACTTCCCGTACCCTGCTCTCTCTGGGCCTGCTGAGCGCGGC
                 CATGTTCGGCGTTTCGCAACAGGCGAATGCCCACGGTTATGTCGAATCGC
                 CGGCCAGCCGCGCCTATCAGTGCAAACTGCAGCTCAACACGCAGTGCGGC
                 AGCGTGCAGTACGAACCGCAGAGCGTCGAGGGCCTGAAAGGCTTCCCGCA
                 GGCCGGCCCGGCTGACGGCCATATCGCCAGCGCCGACAAGTCCACCTTCT
                 TCGAACTGGATCAGCAAACGCCGACGCGCTGGAA/SQN -F F
```

10. Use a Display command to review sample records.



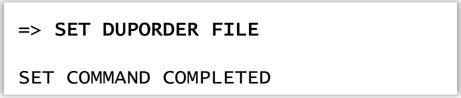
```
=> D TRIAL SCORE ALIGN 1-20
```

11. Repeat this sequence search in PCTGEN and USGENE.

Merge Answer Sets from DGENE, PCTGEN, USGENE

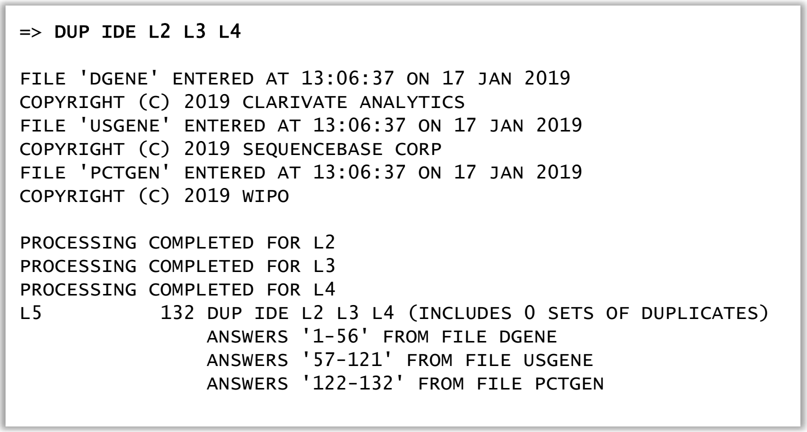
When you have L-numbers for the DGENE, PCTGEN and USGENE sequence search results, merge those results into a single answer set.

1. Enter the SET DUPORDER FILE setting command to specify the answer retrieval follows the order the files were searched.



```
=> SET DUPORDER FILE

SET COMMAND COMPLETED
```

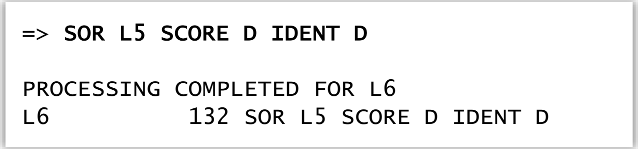
2. Enter the DUPLICATE IDENTIFY (DUP IDE) command to create a new L-number from the multiple answer sets.



```
=> 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
```

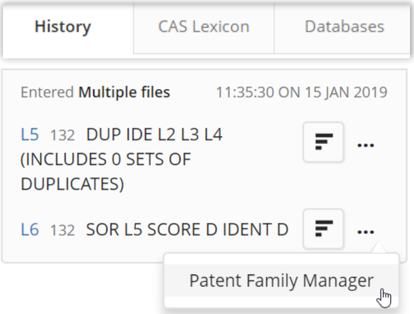
Leveraging the DUP IDE command in this way is simply getting all answers into a single answer set.

3. Sort the results by descending similarity score and identity using the SOR SCORE D IDENT D command. Sorting is useful when there are many answers to review.



```
=> SOR L5 SCORE D IDENT D

PROCESSING COMPLETED FOR L6
L6          132 SOR L5 SCORE D IDENT D
```

4. Click the ellipsis (. . .) button and select the Patent Family Manager to display the results. (This example specifically targets patents.)

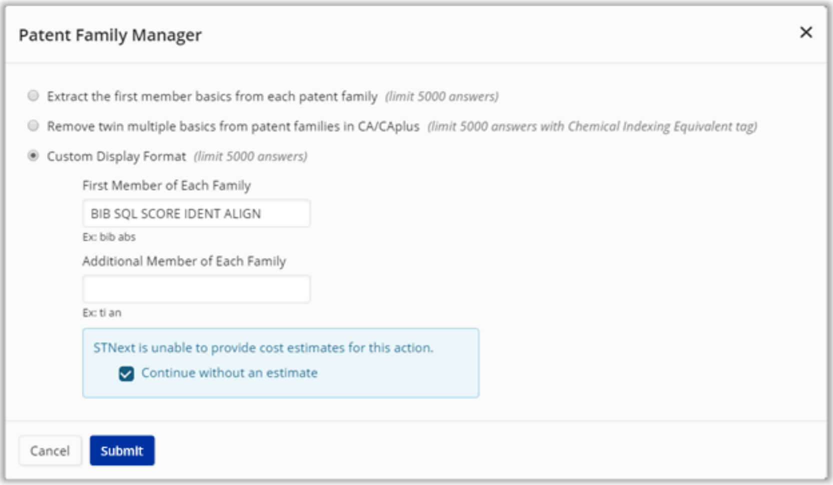


Entered **Multiple files** 11:35:30 ON 15 JAN 2019

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

L6 132 SOR L5 SCORE D IDENT D

5. Choose a patent family display option.



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

Ex: bib abs

Additional Member of Each Family:

Ex: ti an

STNext is unable to provide cost estimates for this action.

☑ Continue without an estimate

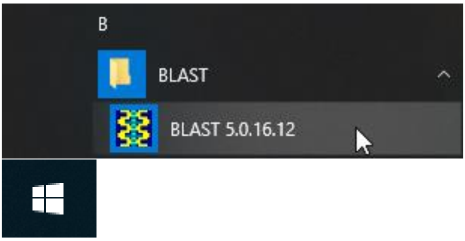
This example uses BIB SQL SCORE IDENT ALIGN for the First Member of Each Family.

6. Click the **Submit** button.

STNext's Patent Family Manager automatically starts with an FSORT to place records in extended patent families – generating a new L-number and then proceeding with the record display.

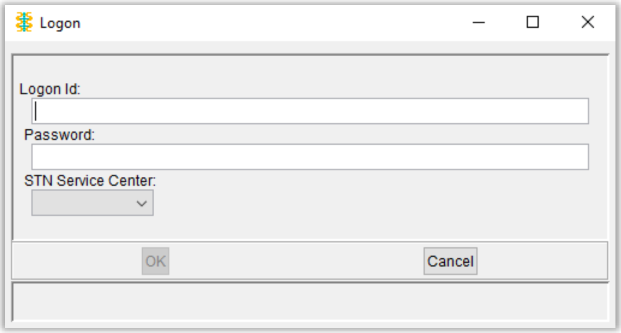
Search Using CAS Registry BLAST Client

After installing the CAS Registry BLAST client on your PC, locate the link/icon under the Start menu.

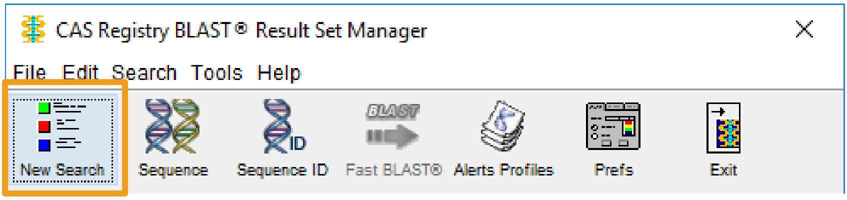


1. Click the **BLAST** icon.

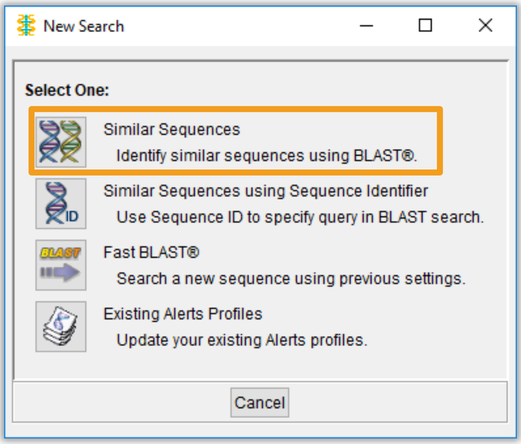
2. Use your STN login credentials and choose your local STN Service Center from the drop-down. Click **OK** to login.



3. Click the **New Search** button.


CAS Registry BLAST® Result Set Manager
File Edit Search Tools Help
New Search
Sequence
Sequence ID
Fast BLAST®
Alerts Profiles
Prefs
Exit

4. Click the **Similar Sequences** button.



Select One:

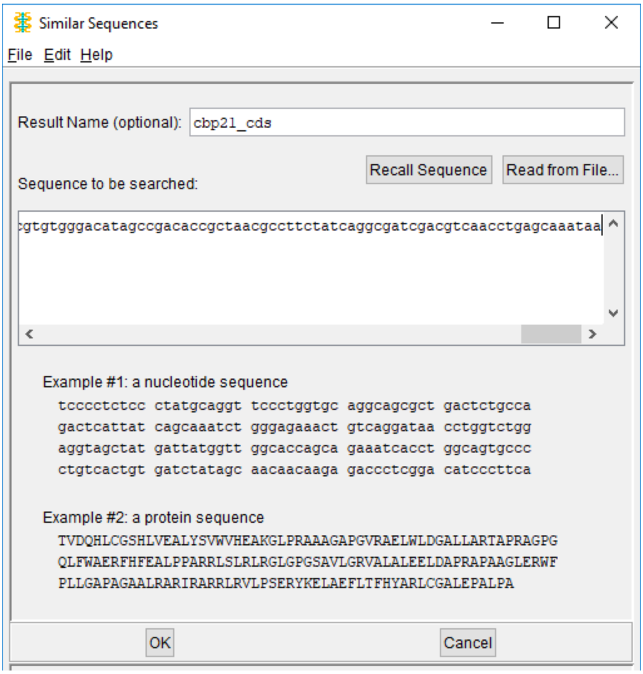
Similar Sequences
Identify similar sequences using BLAST®.

Similar Sequences using Sequence Identifier
Use Sequence ID to specify query in BLAST search.

Fast BLAST®
Search a new sequence using previous settings.

Existing Alerts Profiles
Update your existing Alerts profiles.

5. Name the result and enter the sequence by copy/paste or reading data from a file.



Result Name (optional): cbp21_cds

Sequence to be searched:

```
gtgtgggacatagccgacaccgctaacgccttctatcaggcgatcgacgtcaacctgagcaaataa
```

Example #1: a nucleotide sequence

```
tcccctctcc ctatgcaggt tccctggtgc aggcagcgct gactctgcca
gactcattat cagcaaatct gggagaaact gtcaggataa cctggtctgg
aggtagctat gattatggtt ggcaccagca gaaatcacct ggcagtgccc
ctgtcactgt gatctatagc aacaacaaga gaccctcgga catcccttca
```

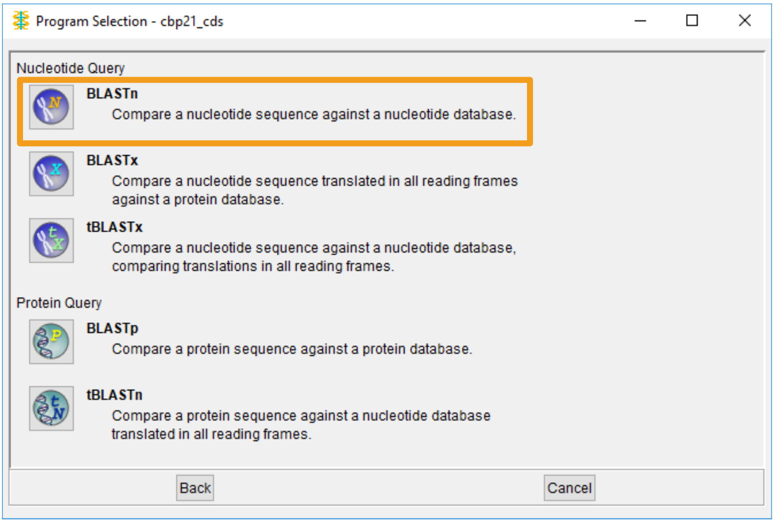
Example #2: a protein sequence

```
TVDQHLCGSHLVEALYSVWVHEAKGLPRAAAGAPGVRAELWLDGALLARTAPRAGPG
QLFWAERFHFEALPPARRLSLRLRGLGPGSAVLGRVALALEELDAPRAPAAGLERWF
PLLGAPAGAALRARIRARRLRVLPSERYKELAEFLTFHYARLCGALEPALPA
```

6. Click the **OK** button.

Search Using CAS Registry BLAST Client

7. Click the appropriate BLAST query type. (This example uses BLASTn.)



Program Selection - cbp21_cds

Nucleotide Query

BLASTn
Compare a nucleotide sequence against a nucleotide database.

BLASTx
Compare a nucleotide sequence translated in all reading frames against a protein database.

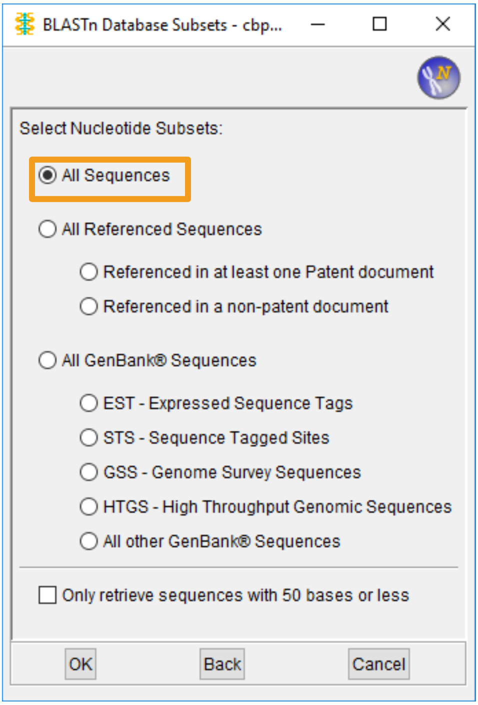
tBLASTx
Compare a nucleotide sequence against a nucleotide database, comparing translations in all reading frames.

Protein Query

BLASTp
Compare a protein sequence against a protein database.

tBLASTn
Compare a protein sequence against a nucleotide database translated in all reading frames.

8. Choose the nucleotide subsets desired. (This example uses All.)



BLASTn Database Subsets - cbp...

Select Nucleotide Subsets:

● All Sequences

○ All Referenced Sequences

○ Referenced in at least one Patent document

○ Referenced in a non-patent document

○ All GenBank® Sequences

○ EST - Expressed Sequence Tags

○ STS - Sequence Tagged Sites

○ GSS - Genome Survey Sequences

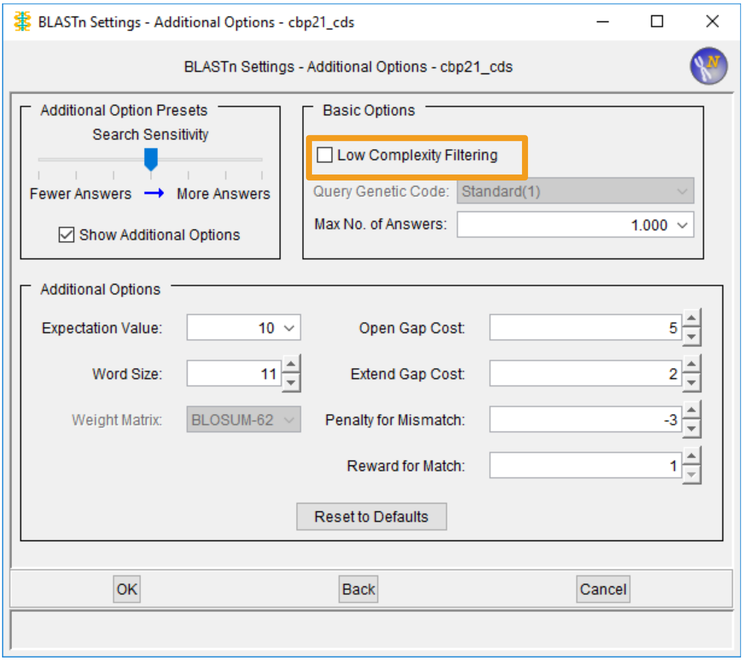
○ HTGS - High Throughput Genomic Sequences

○ All other GenBank® Sequences

□ Only retrieve sequences with 50 bases or less

9. Click the **OK** button to save the subset selection.

10. Adjust the BLAST settings. (This example is patent specific, so uncheck the Low Complexity Filtering option.)



BLASTn Settings - Additional Options - cbp21_cds

Additional Option Presets

☑ Show Additional Options

Basic Options

☐ Low Complexity Filtering

Query Genetic Code: Standard(1)

Max No. of Answers: 1.000

Additional Options

Expectation Value: 10

Word Size: 11

Weight Matrix: BLOSUM-62

Open Gap Cost: 5

Extend Gap Cost: 2

Penalty for Mismatch: -3

Reward for Match: 1

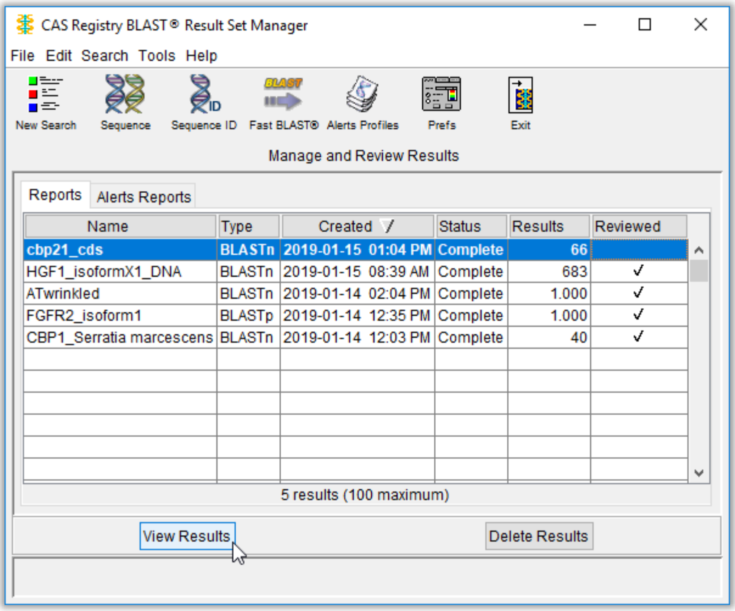
11. Click the **OK** button to start the search.

The Result Set Manager will display the sequence search and status. You can submit additional sequences while the searches are running. Up to 100 results sets can be kept in the Results Set Manager.

When a report is complete, you can view results.

Download Script and Alignment File from CAS Registry

When your BLAST query is complete, you can evaluate the results and select which answers to download for use in STNext.

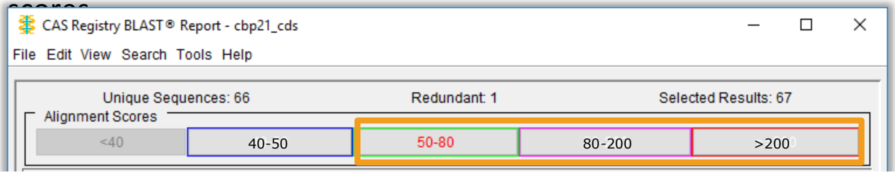


Name	Type	Created	Status	Results	Reviewed
cbp21_cds	BLASTn	2019-01-15 01:04 PM	Complete	66	
HGF1_isoformX1_DNA	BLASTn	2019-01-15 08:39 AM	Complete	683	✓
ATwrinkled	BLASTn	2019-01-14 02:04 PM	Complete	1.000	✓
FGFR2_isoform1	BLASTp	2019-01-14 12:35 PM	Complete	1.000	✓
CBP1_Serratia marcescens	BLASTn	2019-01-14 12:03 PM	Complete	40	✓

5 results (100 maximum)

1. Select the results set from the **Reports** tab and click the **View Results** button.

2. Review key sequence statistics such as total unique and redundant sequences. There is also a grouping of results based on alignment scores


CAS Registry BLAST® Report - cbp21_cds
File Edit View Search Tools Help
Alignment Scores
<40
40-50
50-80
80-200
>200

Unique Sequences: 66 Redundant: 1 Selected Results: 67

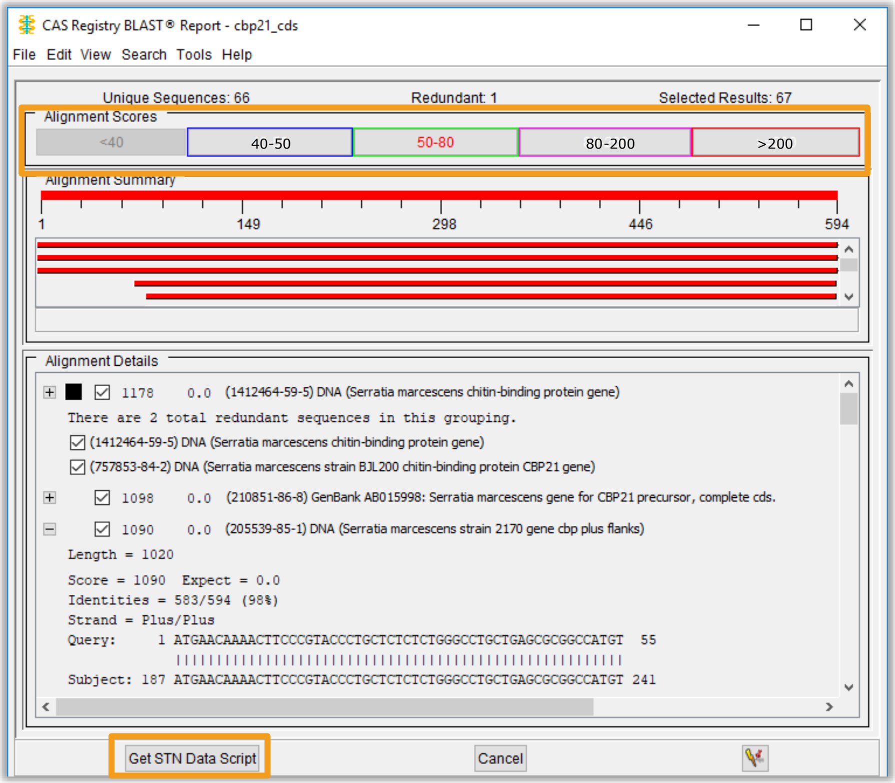
Higher Alignment Scores will have better alignment and match over the length of the query.

BLAST Client

3. Click the **Alignment Score** buttons to select groups of sequences to include in the STN Data Script.

4. Preview the alignment detail for individual results to verify the letter by letter alignment details of the selected results, if desired.

5. Click the **Get STN Data Script** button.

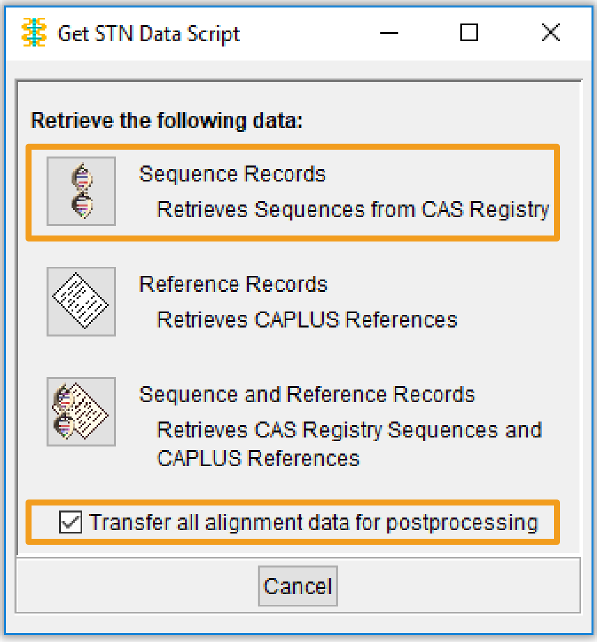


Unique Sequences: 66 Redundant: 1 Selected Results: 67

```
⊞ ■ ☑ 1178    0.0  (1412464-59-5) DNA (Serratia marcescens chitin-binding protein gene)
     There are 2 total redundant sequences in this grouping.
     ☑ (1412464-59-5) DNA (Serratia marcescens chitin-binding protein gene)
     ☑ (757853-84-2) DNA (Serratia marcescens strain BJL200 chitin-binding protein CBP21 gene)
⊞   ☑ 1098    0.0  (210851-86-8) GenBank AB015998: Serratia marcescens gene for CBP21 precursor, complete cds.
⊟   ☑ 1090    0.0  (205539-85-1) DNA (Serratia marcescens strain 2170 gene cbp plus flanks)
     Length = 1020
     Score = 1090  Expect = 0.0
     Identities = 583/594  (98%)
     Strand = Plus/Plus
     Query:     1 ATGAACAAAACTTCCCGTACCCTGCTCTCTCTGGGCCTGCTGAGCGCGGCCATGT   55
                  |||||||||||||||||||||||||||||||||||||||||||||||||||||||
     Subject: 187 ATGAACAAAACTTCCCGTACCCTGCTCTCTCTGGGCCTGCTGAGCGCGGCCATGT 241
```

Download Script and Alignment, continued

6. Check the **Transfer all alignment data for postprocessing** option.



Get STN Data Script

Retrieve the following data:

Sequence Records
Retrieves Sequences from CAS Registry

Reference Records
Retrieves CAPLUS References

Sequence and Reference Records
Retrieves CAS Registry Sequences and CAPLUS References

☑ Transfer all alignment data for postprocessing

7. Click the **Sequence Records** button.

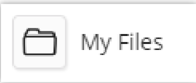
The system will prompt you to name and save a .SCB script file that will be used in STNext.

The system will then prompt you a second time to save a .XSS file that contains the alignment data that will be available for use in your combined report.

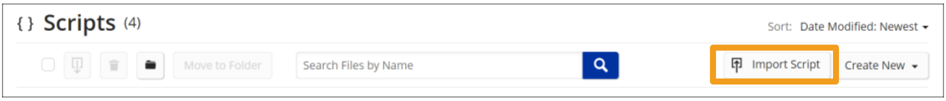
Note the file directory so it is easy to locate and import the .SCB and .XSS files into STNext.

Run Script in STNext to Search CAplus

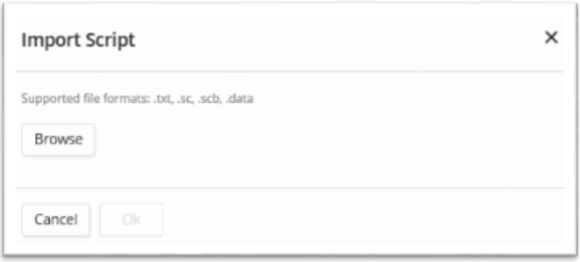
1. Select **Scripts** from the My Files menu.



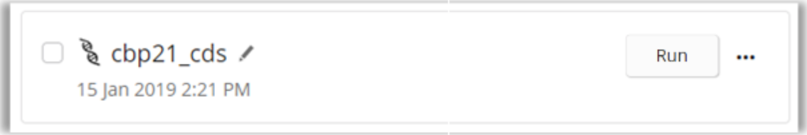
2. Click the **Import Script** button.



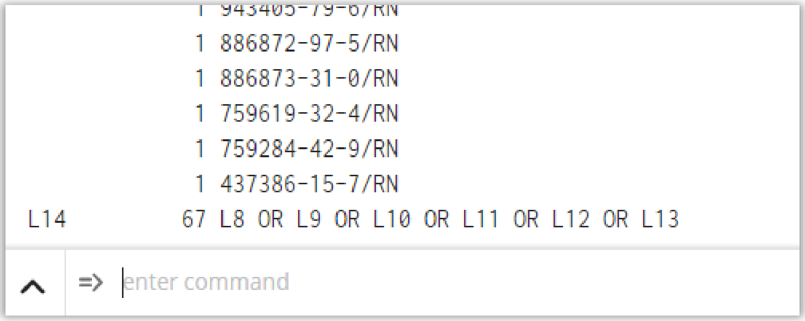
3. Browse to locate the .SCB script file that was created in the CAS Registry BLAST Client. Click the **OK** button.



4. The .SCB file is saved to the Scripts page. Click the **Run** button.



The script will automatically file into CAS REGISTRY and search for the CAS RNs selected in the BLAST client session and combine the RNs in a single L-number.

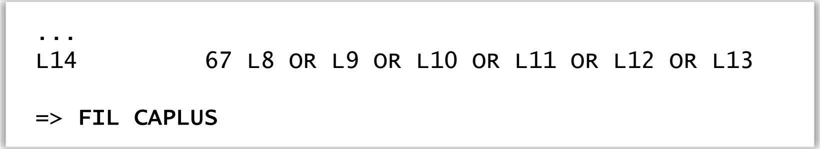


```
                  1 943405-79-6/RN
                  1 886872-97-5/RN
                  1 886873-31-0/RN
                  1 759619-32-4/RN
                  1 759284-42-9/RN
                  1 437386-15-7/RN
L14              67 L8 OR L9 OR L10 OR L11 OR L12 OR L13
```

Identify Duplicates

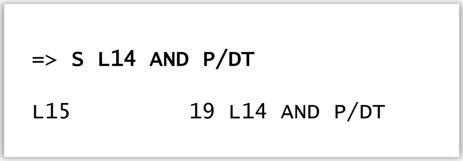
When the .SCB script runs, it finds *SUBSTANCE* records in CAS REGISTRY related to the sequences. To combine these answers with those found in DGENE, PCTGEN or USGENE, we have to retrieve the corresponding *DOCUMENT* records. Then we can compare all the answers and avoid duplicates.

1. File into CAplus or HCAplus.



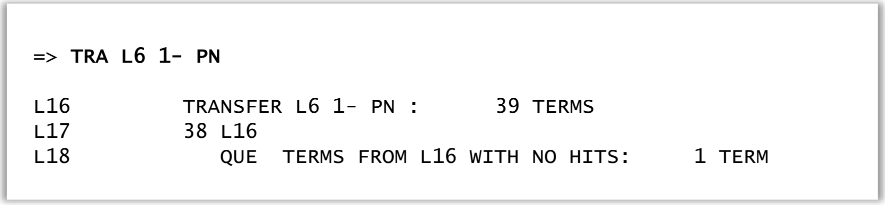
```
...
L14          67 L8 OR L9 OR L10 OR L11 OR L12 OR L13

=> FIL CAPLUS
```

2. Search L14 for corresponding patent records by adding the P/DT command. (Patent/Document Type)



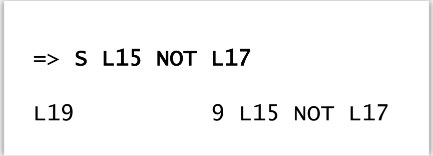
```
=> S L14 AND P/DT

L15          19 L14 AND P/DT
```

3. Transfer patent numbers from the DGENE, PCTGEN and USGENE answers to CAplus.



```
=> TRA L6 1- PN

L16          TRANSFER L6 1- PN :       39 TERMS
L17          38 L16
L18             QUE  TERMS FROM L16 WITH NO HITS:     1 TERM
```

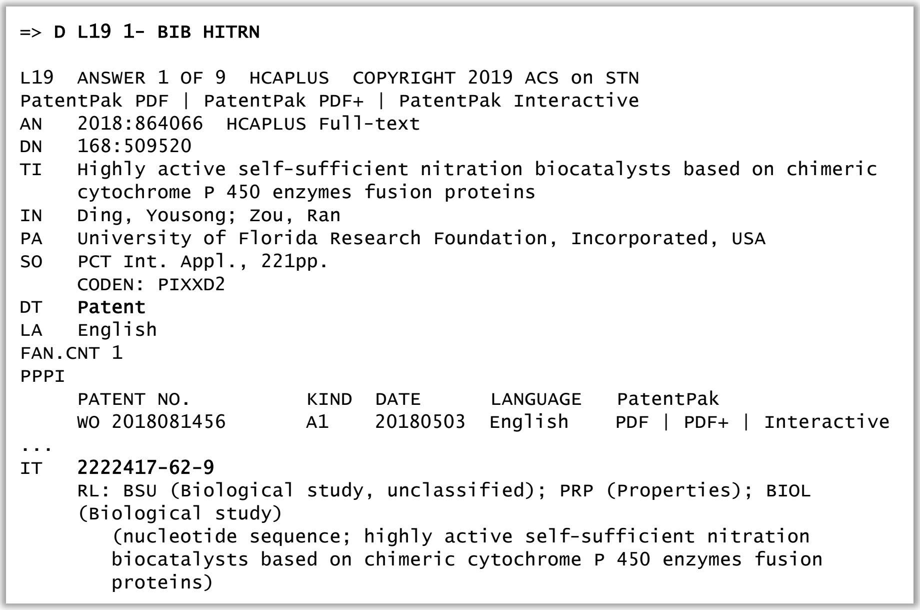
4. Search both the CAplus and DGENE/PCTGEN/USGENE answer sets using the NOT operator to identify the unique records from the CAplus search.



```
=> S L15 NOT L17

L19            9 L15 NOT L17
```

In this example, there were 9 additional records found.

5. Display those unique records using the HITRN command so the CAS Registry BLAST alignments can be included in the combined report.



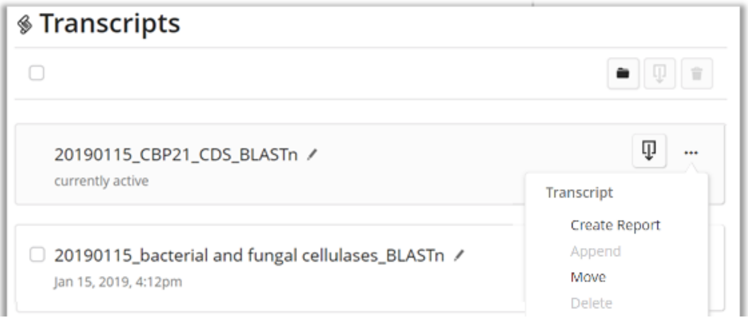
```
=> D L19 1- BIB HITRN

L19  ANSWER 1 OF 9  HCAPLUS  COPYRIGHT 2019 ACS on STN
PatentPak PDF | PatentPak PDF+ | PatentPak Interactive
AN   2018:864066  HCAPLUS Full-text
DN   168:509520
TI   Highly active self-sufficient nitration biocatalysts based on chimeric
     cytochrome P 450 enzymes fusion proteins
IN   Ding, Yousong; Zou, Ran
PA   University of Florida Research Foundation, Incorporated, USA
SO   PCT Int. Appl., 221pp.
     CODEN: PIXXD2
DT   Patent
LA   English
FAN.CNT 1
PPPI
     PATENT NO.          KIND  DATE       LANGUAGE   PatentPak
     WO 2018081456       A1    20180503  English     PDF | PDF+ | Interactive
...
IT   2222417-62-9
     RL: BSU (Biological study, unclassified); PRP (Properties); BIOL
     (Biological study)
        (nucleotide sequence; highly active self-sufficient nitration
        biocatalysts based on chimeric cytochrome P 450 enzymes fusion
        proteins)
```

Create Combined Report

After collecting answers from all the databases, we can create a single report that also includes the alignment data.

1. Select **Transcripts** from the My Files menu.

2. Click the ellipsis (. . .) button and select **Create Report**.



Transcripts

□

20190115_CBP21_CDS_BLASTn
currently active

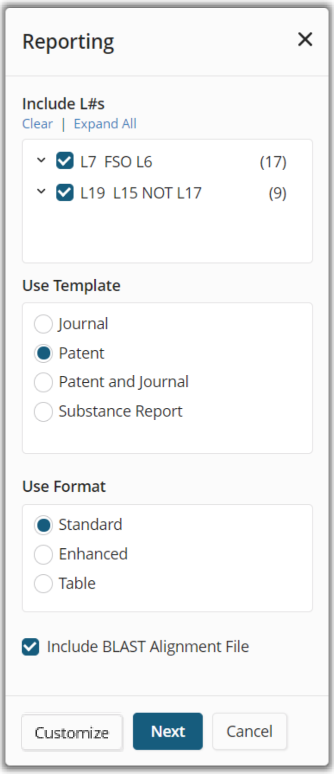
□ 20190115_bacterial and fungal cellulases_BLASTn
Jan 15, 2019, 4:12pm

3. Select the L7 records sorted by patent family, and L19 records that were unique to the CAplus search.

4. This example scenario was patent-specific, so use the **Patent** template and **Standard** format.

5. Check the **Include BLAST Alignment File** option.

6. Click the **Next** button.



Reporting ×

Include L#s
Clear | Expand All

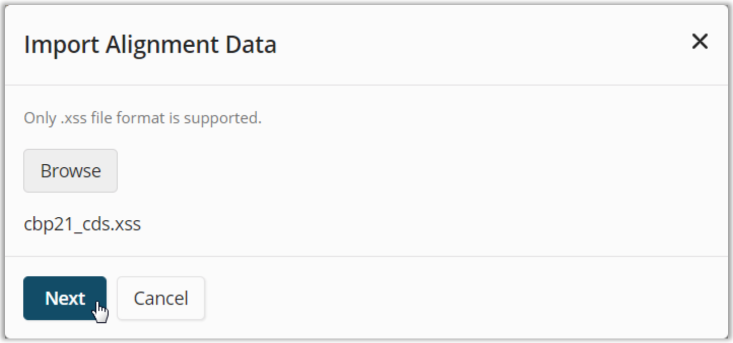
☑ L7 FSO L6 (17)
☑ L19 L15 NOT L17 (9)

Use Template
○ Journal
● Patent
○ Patent and Journal
○ Substance Report

Use Format
● Standard
○ Enhanced
○ Table

☑ Include BLAST Alignment File

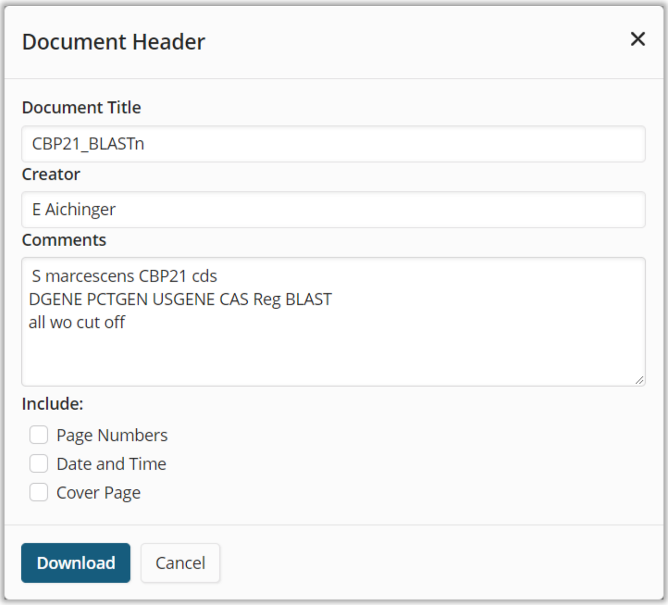
7. Browse to locate the .XSS alignment file. Click the **Next** button.



Only .xss file format is supported.

cbp21_cds.xss

8. Enter Report Header information as desired.



Document Title: CBP21_BLASTn

Creator: E Aichinger

Comments:
S marcescens CBP21 cds
DGENE PCTGEN USGENE CAS Reg BLAST
all wo cut off

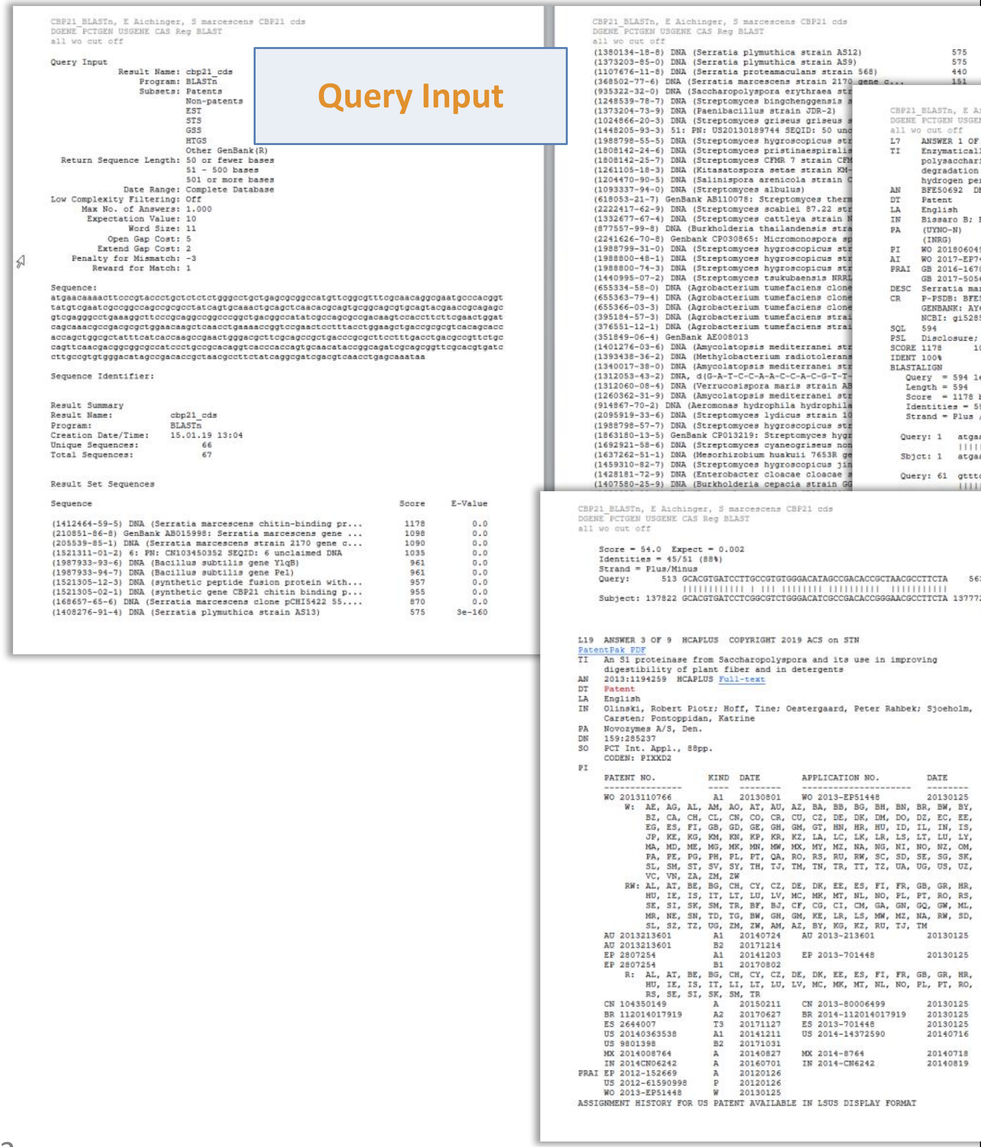
Include:
□ Page Numbers
□ Date and Time
□ Cover Page

9. Click the **Download** button.

Summary

STNext offers four essential patent sequence databases with unrivaled coverage for comprehensive searching.

The cross-over capability allows you to combine data from several sources into one report that includes query information, record details and BLAST alignment data.



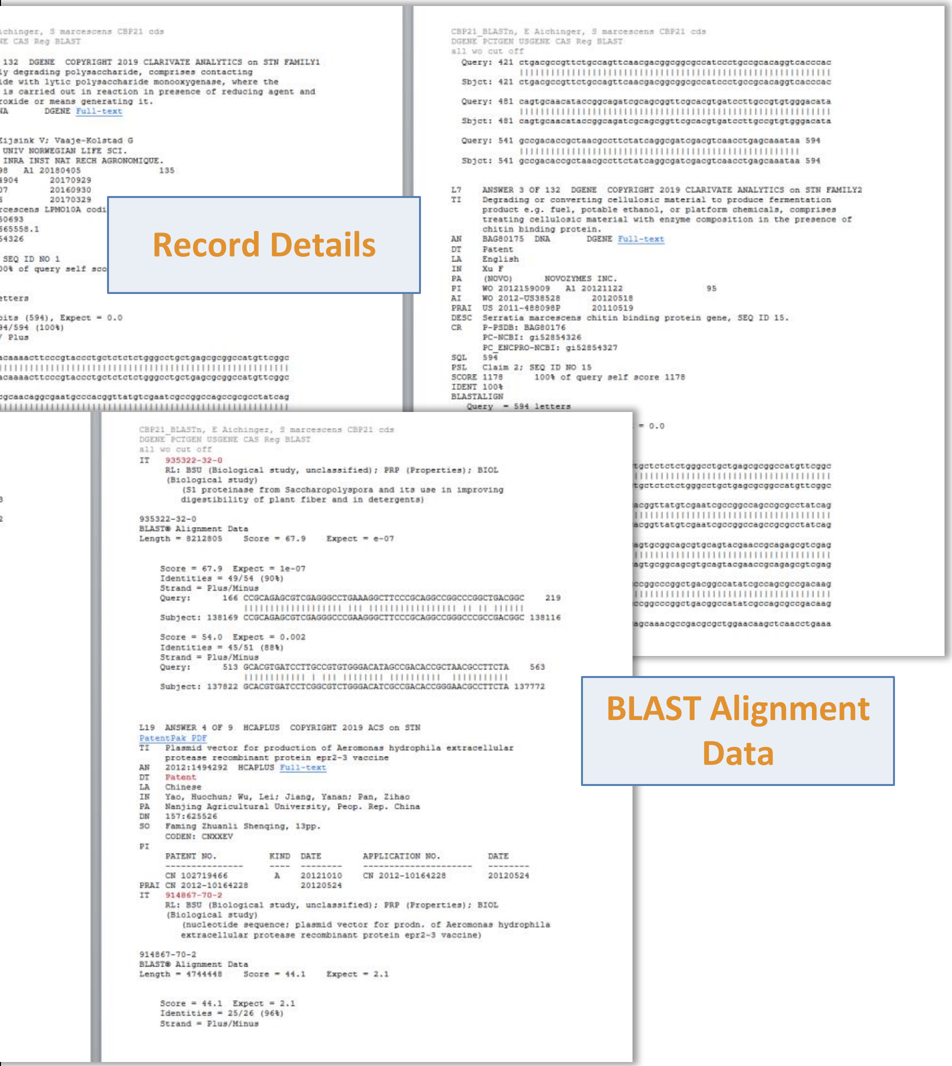
```
CBP21_BLASTn, E Aichinger, S marcescens CBP21 cds
DGENE PCTGEN USGENE CAS Reg BLAST
all wo cut off

Query Input
                 Result Name: cbp21_cds
                     Program: BLASTn
                     Subsets: Patents
                              Non-patents
                              EST
                              STS
                              GSS
                              HTGS
                              Other GenBank(R)
      Return Sequence Length: 50 or fewer bases
                              51 - 500 bases
                              501 or more bases
                  Date Range: Complete Database
   Low Complexity Filtering: Off
         Max No. of Answers: 1.000
          Expectation Value: 10
                  Word Size: 11
              Open Gap Cost: 5
            Extend Gap Cost: 2
       Penalty for Mismatch: -3
           Reward for Match: 1

Sequence:
atgaacaaaacttcccgtaccctgctctctctgggcctgctgagcgcggccatgttcggcgtttcgcaacaggcgaatgcccacggt
tatgtcgaatcgccggccagccgcgcctatcagtgcaaactgcagctcaacacgcagtgcggcagcgtgcagtacgaaccgcagagc
gtcgagggcctgaaaggcttcccgcaggccggcccggctgacggccatatcgccagcgccgacaagtccaccttcttcgaactggat
cagcaaacgccgacgcgctggaacaagctcaacctgaaaaccggtccgaactcctttacctggaagctgaccgcgcgtcacagcacc
accagctggcgctatttcatcaccaagccgaactgggacgcttcgcagccgctgacccgcgcttcctttgacctgacgccgttctgc
cagttcaacgacggcggcgccatccctgccgcacaggtcacccaccagtgcaacataccggcagatcgcagcggttcgcacgtgatc
cttgccgtgtgggacatagccgacaccgctaacgccttctatcaggcgatcgacgtcaacctgagcaaataa

Sequence Identifier:


Result Summary
Result Name:            cbp21_cds
Program:                BLASTn
Creation Date/Time:     15.01.19 13:04
Unique Sequences:             66
Total Sequences:              67


Result Set Sequences

Sequence                                                          Score     E-Value

(1412464-59-5) DNA (Serratia marcescens chitin-binding pr...      1178         0.0
(210851-86-8) GenBank AB015998: Serratia marcescens gene ...      1098         0.0
(205539-85-1) DNA (Serratia marcescens strain 2170 gene c...      1090         0.0
(1521311-01-2) 6: PN: CN103450352 SEQID: 6 unclaimed DNA          1035         0.0
(1987933-93-6) DNA (Bacillus subtilis gene YlqB)                   961         0.0
(1987933-94-7) DNA (Bacillus subtilis gene Pel)                    961         0.0
(1521305-12-3) DNA (synthetic peptide fusion protein with...       957         0.0
(1521305-02-1) DNA (synthetic gene CBP21 chitin binding p...       955         0.0
(168657-65-6) DNA (Serratia marcescens clone pCHI5422 55....       870         0.0
(1408276-91-4) DNA (Serratia plymuthica strain AS13)               575      3e-160
```

```
CBP21_BLASTn, E Aichinger, S marcescens CBP21 cds
DGENE PCTGEN USGENE CAS Reg BLAST
all wo cut off
(1380134-18-8) DNA (Serratia plymuthica strain AS12)                    575
(1373203-85-0) DNA (Serratia plymuthica strain AS9)                     575
(1107676-11-8) DNA (Serratia proteamaculans strain 568)                 440
(368502-77-6) DNA (Serratia marcescens strain 2170 gene c...            151
(935322-32-0) DNA (Saccharopolyspora erythraea str
(1248539-78-7) DNA (Streptomyces bingchenggensis s
(1373204-73-9) DNA (Paenibacillus strain JDR-2)
(1024866-20-3) DNA (Streptomyces griseus griseus s
(1448205-93-3) 51: PN: US20130189744 SEQID: 50 unc
(1988798-55-5) DNA (Streptomyces hygroscopicus str
(1808142-24-6) DNA (Streptomyces pristinaespiralis
(1808142-25-7) DNA (Streptomyces CFMR 7 strain CFM
(1261105-18-3) DNA (Kitasatospora setae strain KM-
(1204470-90-5) DNA (Salinispora arenicola strain C
(1093337-94-0) DNA (Streptomyces albulus)
(618053-21-7) GenBank AB110078: Streptomyces therm
(2222417-62-9) DNA (Streptomyces scabiei 87.22 str
(1332677-67-4) DNA (Streptomyces cattleya strain N
(877557-99-8) DNA (Burkholderia thailandensis stra
(2241626-70-8) Genbank CP030865: Micromonospora sp
(1988799-31-0) DNA (Streptomyces hygroscopicus str
(1988800-48-1) DNA (Streptomyces hygroscopicus str
(1988800-74-3) DNA (Streptomyces hygroscopicus str
(1440995-07-2) DNA (Streptomyces tsukubaensis NRRL
(655334-58-0) DNA (Agrobacterium tumefaciens clone
(655363-79-4) DNA (Agrobacterium tumefaciens clone
(655366-03-3) DNA (Agrobacterium tumefaciens clone
(395184-57-3) DNA (Agrobacterium tumefaciens strai
(376551-12-1) DNA (Agrobacterium tumefaciens strai
(351849-06-4) GenBank AE008013
(1401276-03-6) DNA (Amycolatopsis mediterranei str
(1393438-36-2) DNA (Methylobacterium radiotolerans
(1340017-38-0) DNA (Amycolatopsis mediterranei str
(1312053-43-2) DNA, d(G-A-T-C-C-A-A-C-C-A-C-G-T-T-
(1312060-08-4) DNA (Verrucosispora maris strain AB
(1260362-31-9) DNA (Amycolatopsis mediterranei str
(914867-70-2) DNA (Aeromonas hydrophila hydrophila
(2095919-33-6) DNA (Streptomyces lydicus strain 10
(1988798-57-7) DNA (Streptomyces hygroscopicus str
(1863180-13-5) GenBank CP013219: Streptomyces hygr
(1692921-58-6) DNA (Streptomyces cyaneogriseus non
(1637262-51-1) DNA (Mesorhizobium huakuii 7653R ge
(1459310-82-7) DNA (Streptomyces hygroscopicus jin
(1428181-72-9) DNA (Enterobacter cloacae cloacae s
(1407580-25-9) DNA (Burkholderia cepacia strain GG
```

```
CBP21_BLASTn, E Ai
DGENE PCTGEN USGEN
all wo cut off
L7     ANSWER 1 OF
TI     Enzymaticall
       polysacchari
       degradation
       hydrogen per
AN     BFE50692  DN
DT     Patent
LA     English
IN     Bissaro B; F
PA     (UYNO-N)
       (INRG)
PI     WO 201806049
AI     WO 2017-EP74
PRAI   GB 2016-1670
       GB 2017-5056
DESC   Serratia mar
CR     P-PSDB: BFE5
       GENBANK: AY6
       NCBI: gi5285
SQL    594
PSL    Disclosure;
SCORE 1178        10
IDENT 100%
BLASTALIGN
   Query  = 594 le
   Length = 594
   Score  = 1178 b
   Identities = 59
   Strand = Plus /

  Query: 1   atgaa
             |||||
  Sbjct: 1   atgaa

  Query: 61  gtttc
             |||||
```

```
CBP21_BLASTn, E Aichinger, S marcescens CBP21 cds
DGENE PCTGEN USGENE CAS Reg BLAST
all wo cut off

    Score = 54.0  Expect = 0.002
    Identities = 45/51 (88%)
    Strand = Plus/Minus
    Query:      513 GCACGTGATCCTTGCCGTGTGGGACATAGCCGACACCGCTAACGCCTTCTA    563
                    |||||||||||| | ||| |||||||| ||||||||||  |||||||||||
    Subject: 137822 GCACGTGATCCTCGGCGTCTGGGACATCGCCGACACCGGGAACGCCTTCTA 137772


L19  ANSWER 3 OF 9  HCAPLUS  COPYRIGHT 2019 ACS on STN
PatentPak PDF
TI   An S1 proteinase from Saccharopolyspora and its use in improving
     digestibility of plant fiber and in detergents
AN   2013:1194259  HCAPLUS Full-text
DT   Patent
LA   English
IN   Olinski, Robert Piotr; Hoff, Tine; Oestergaard, Peter Rahbek; Sjoeholm,
     Carsten; Pontoppidan, Katrine
PA   Novozymes A/S, Den.
DN   159:285237
SO   PCT Int. Appl., 88pp.
     CODEN: PIXXD2
PI
     PATENT NO.          KIND  DATE        APPLICATION NO.         DATE
     ---------------     ----  --------    ---------------------   --------
     WO 2013110766        A1   20130801    WO 2013-EP51448         20130125
         W:  AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY,
             BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DK, DM, DO, DZ, EC, EE,
             EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IS,
             JP, KE, KG, KM, KN, KP, KR, KZ, LA, LC, LK, LR, LS, LT, LU, LY,
             MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM,
             PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SC, SD, SE, SG, SK,
             SL, SM, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ,
             VC, VN, ZA, ZM, ZW
         RW: AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR,
             HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, RS,
             SE, SI, SK, SM, TR, BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML,
             MR, NE, SN, TD, TG, BW, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SD,
             SL, SZ, TZ, UG, ZM, ZW, AM, AZ, BY, KG, KZ, RU, TJ, TM
     AU 2013213601        A1   20140724    AU 2013-213601          20130125
     AU 2013213601        B2   20171214
     EP 2807254           A1   20141203    EP 2013-701448          20130125
     EP 2807254           B1   20170802
         R:  AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR,
             HU, IE, IS, IT, LI, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO,
             RS, SE, SI, SK, SM, TR
     CN 104350149         A    20150211    CN 2013-80006499        20130125
     BR 112014017919      A2   20170627    BR 2014-112014017919    20130125
     ES 2644007           T3   20171127    ES 2013-701448          20130125
     US 20140363538       A1   20141211    US 2014-14372590        20140716
     US 9801398           B2   20171031
     MX 2014008764        A    20140827    MX 2014-8764            20140718
     IN 2014CN06242       A    20160701    IN 2014-CN6242          20140819
PRAI EP 2012-152669       A    20120126
     US 2012-61590998     P    20120126
     WO 2013-EP51448      W    20130125
ASSIGNMENT HISTORY FOR US PATENT AVAILABLE IN LSUS DISPLAY FORMAT
```

Use your STN login credentials to access STNext at next.stn.org



```
ichinger, S marcescens CBP21 cds
NE CAS Reg BLAST

 132  DGENE  COPYRIGHT 2019 CLARIVATE ANALYTICS on STN FAMILY1
ly degrading polysaccharide, comprises contacting
ide with lytic polysaccharide monooxygenase, where the
 is carried out in reaction in presence of reducing agent and
roxide or means generating it.
NA       DGENE Full-text


Eijsink V; Vaaje-Kolstad G
 UNIV NORWEGIAN LIFE SCI.
 INRA INST NAT RECH AGRONOMIQUE.
98   A1 20180405                135
4904       20170929
07         20160930
6          20170329
rcescens LPMO10A codi
50693
565558.1
54326

 SEQ ID NO 1
00% of query self sco

etters

bits (594), Expect = 0.0
94/594 (100%)
/ Plus

acaaaacttcccgtaccctgctctctctgggcctgctgagcgcggccatgttcggc
||||||||||||||||||||||||||||||||||||||||||||||||||||||||
acaaaacttcccgtaccctgctctctctgggcctgctgagcgcggccatgttcggc

cgcaacaggcgaatgcccacggttatgtcgaatcgccggccagccgcgcctatcag
||||||||||||||||||||||||||||||||||||||||||||||||||||||||
```

```
CBP21_BLASTn, E Aichinger, S marcescens CBP21 cds
DGENE PCTGEN USGENE CAS Reg BLAST
all wo cut off
  Query: 421 ctgacgccgttctgccagttcaacgacggcggcgccatccctgccgcacaggtcacccac
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
  Sbjct: 421 ctgacgccgttctgccagttcaacgacggcggcgccatccctgccgcacaggtcacccac

  Query: 481 cagtgcaacataccggcagatcgcagcggttcgcacgtgatccttgccgtgtgggacata
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
  Sbjct: 481 cagtgcaacataccggcagatcgcagcggttcgcacgtgatccttgccgtgtgggacata

  Query: 541 gccgacaccgctaacgccttctatcaggcgatcgacgtcaacctgagcaaataa 594
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||
  Sbjct: 541 gccgacaccgctaacgccttctatcaggcgatcgacgtcaacctgagcaaataa 594


L7    ANSWER 3 OF 132  DGENE  COPYRIGHT 2019 CLARIVATE ANALYTICS on STN FAMILY2
TI    Degrading or converting cellulosic material to produce fermentation
      product e.g. fuel, potable ethanol, or platform chemicals, comprises
      treating cellulosic material with enzyme composition in the presence of
      chitin binding protein.
AN    BAG80175  DNA       DGENE Full-text
DT    Patent
LA    English
IN    Xu F
PA    (NOVO)      NOVOZYMES INC.
PI    WO 2012159009   A1 20121122                95
AI    WO 2012-US38528       20120518
PRAI  US 2011-488098P       20110519
DESC  Serratia marcescens chitin binding protein gene, SEQ ID 15.
CR    P-PSDB: BAG80176
      PC-NCBI: gi52854326
      PC_ENCPRO-NCBI: gi52854327
SQL   594
PSL   Claim 2; SEQ ID NO 15
SCORE 1178      100% of query self score 1178
IDENT 100%
BLASTALIGN
   Query  = 594 letters

                                              = 0.0


                                              tgctctctctgggcctgctgagcgcggccatgttcggc
                                              ||||||||||||||||||||||||||||||||||||||
                                              tgctctctctgggcctgctgagcgcggccatgttcggc

                                              acggttatgtcgaatcgccggccagccgcgcctatcag
                                              ||||||||||||||||||||||||||||||||||||||
                                              acggttatgtcgaatcgccggccagccgcgcctatcag

                                              agtgcggcagcgtgcagtacgaaccgcagagcgtcgag
                                              ||||||||||||||||||||||||||||||||||||||
                                              agtgcggcagcgtgcagtacgaaccgcagagcgtcgag

                                              ccggcccggctgacggccatatcgccagcgccgacaag
                                              ||||||||||||||||||||||||||||||||||||||
                                              ccggcccggctgacggccatatcgccagcgccgacaag

                                              agcaaacgccgacgcgctggaacaagctcaacctgaaa
```

```
CBP21_BLASTn, E Aichinger, S marcescens CBP21 cds
DGENE PCTGEN USGENE CAS Reg BLAST
all wo cut off
IT   935322-32-0
     RL: BSU (Biological study, unclassified); PRP (Properties); BIOL
     (Biological study)
        (S1 proteinase from Saccharopolyspora and its use in improving
        digestibility of plant fiber and in detergents)

935322-32-0
BLAST® Alignment Data
Length = 8212805    Score = 67.9    Expect = e-07


    Score = 67.9  Expect = 1e-07
    Identities = 49/54 (90%)
    Strand = Plus/Minus
    Query:      166 CCGCAGAGCGTCGAGGGCCTGAAAGGCTTCCCGCAGGCCGGCCCGGCTGACGGC    219
                    ||||||||||||||||||| ||| ||||||||||||||||| || || ||||||
    Subject: 138169 CCGCAGAGCGTCGAGGGCCCGAAGGGCTTCCCGCAGGCCGGGCCCGCCGACGGC 138116

    Score = 54.0  Expect = 0.002
    Identities = 45/51 (88%)
    Strand = Plus/Minus
    Query:      513 GCACGTGATCCTTGCCGTGTGGGACATAGCCGACACCGCTAACGCCTTCTA    563
                    |||||||||||| | ||| |||||||| ||||||||||  |||||||||||
    Subject: 137822 GCACGTGATCCTCGGCGTCTGGGACATCGCCGACACCGGGAACGCCTTCTA 137772


L19  ANSWER 4 OF 9  HCAPLUS  COPYRIGHT 2019 ACS on STN
PatentPak PDF
TI   Plasmid vector for production of Aeromonas hydrophila extracellular
     protease recombinant protein epr2-3 vaccine
AN   2012:1494292  HCAPLUS Full-text
DT   Patent
LA   Chinese
IN   Yao, Huochun; Wu, Lei; Jiang, Yanan; Pan, Zihao
PA   Nanjing Agricultural University, Peop. Rep. China
DN   157:625526
SO   Faming Zhuanli Shenqing, 13pp.
     CODEN: CNXXEV
PI
     PATENT NO.          KIND  DATE       APPLICATION NO.        DATE
     ---------------     ----  --------   ---------------------  --------
     CN 102719466         A    20121010   CN 2012-10164228       20120524
PRAI CN 2012-10164228          20120524
IT   914867-70-2
     RL: BSU (Biological study, unclassified); PRP (Properties); BIOL
     (Biological study)
        (nucleotide sequence; plasmid vector for prodn. of Aeromonas hydrophila
        extracellular protease recombinant protein epr2-3 vaccine)

914867-70-2
BLAST® Alignment Data
Length = 4744448    Score = 44.1    Expect = 2.1


    Score = 44.1  Expect = 2.1
    Identities = 25/26 (96%)
    Strand = Plus/Minus
```

Record Details

BLAST Alignment Data

Download the CAS Registry BLAST® client

Prior to searching biosequences in CAS REGISTRY, download and install the CAS Registry BLAST client.

With the CAS REGISTRY^{SM} database, the CAS Registry BLAST® client must be used to locate the sequence data as CAS Registry Numbers® and export a script that will help load the Registry Numbers into your session on STNext.

Download the CAS Registry BLAST client at:
https://next.stn.org/stn/downloads/blast-download.html

Please note, your system administrator may need to assist you if you do not have installation privileges on your PC.

For more information...

CAS help@cas.org	FIZ Karlsruhe helpdesk@fiz-karlsruhe.de
Support & Training: www.cas.org/support	Support & Training: www.stn-international.de