Sequence Search on STN

May 21, 2019

Eunyoung Kim
Patent Examiner / Ph.D.

Biotechnology Examination Division
Korean Intellectual Property Office

Contents

1 Introduction to STN

2 Exact and pattern Searching

3 BLAST Search

Resources

Basic STN Command
(https://www.cas.org/sites/default/files/documents/basic.pdf)

CAS REGISTRY^{SM}: Exact and pattern searching of nucleic acid sequences
(https://www.cas.org/sites/default/files/documents/nucleic.pdf)

CAS REGISTRY^{SM}: Exact and pattern searching of protein sequences
(https://www.cas.org/sites/default/files/documents/protseq.pdf)

Sequence Motif Searches in CAS REGISTRY^{SM}
(https://www.stninternational.org/uploads/tx_ptgsarelatedfiles/20130730_Sequence_Motif_Searches_in_REGISTRY.pdf)

CAS REGISTRY^{SM}: BLAST® similarity searching via STN Express®
(https://www.cas.org/sites/default/files/documents/blast_0.pdf)

Introduction to STN

Why sequence searching on STN

- Comprehensiveness: largest collection of sequence data
  - DGENE (Derwent Geneseq^{TM})
  - CAS REGISTRY^{SM}
  - USGENE
  - PCTGEN

STN Express Version 8.6

STN THE CHOICE OF PATENT EXPERTS

STNext

- Reliability: value-added data
- Flexibility:
  - Different search algorithms (BLAST, GETSIM, GETSEQ)
  - Combination with text searching in DWPI, CAplus, and INPADOC
  - STN’s sophisticated search language including script language
- Embedded in STNext: modern, web-based interface

DGENE (Derwent Geneseq™)

- 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
- 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), etc.
- Legal status data from INPADOCDB (D LS or LS2) directly displayable

CAS REGISTRY^{SM}

- Value-added database produced by Chemical Abstracts Service (CAS)
- Most comprehensive collection of sequences from life science journals and basic patents from CAplus^{SM} from 63 authorities
  - 60 M nucleic sequences
  - 11 M protein sequences
- Sequence details include sequence type, sequence length, nucleic acid type, 1 and 3 letter amino acid codes
- Unique sequence types covered and searchable (e.g. multi-chain, cyclic peptides, peptide-metal complexes, etc.)
- Sequences linked to value-added CAplus records by RNs

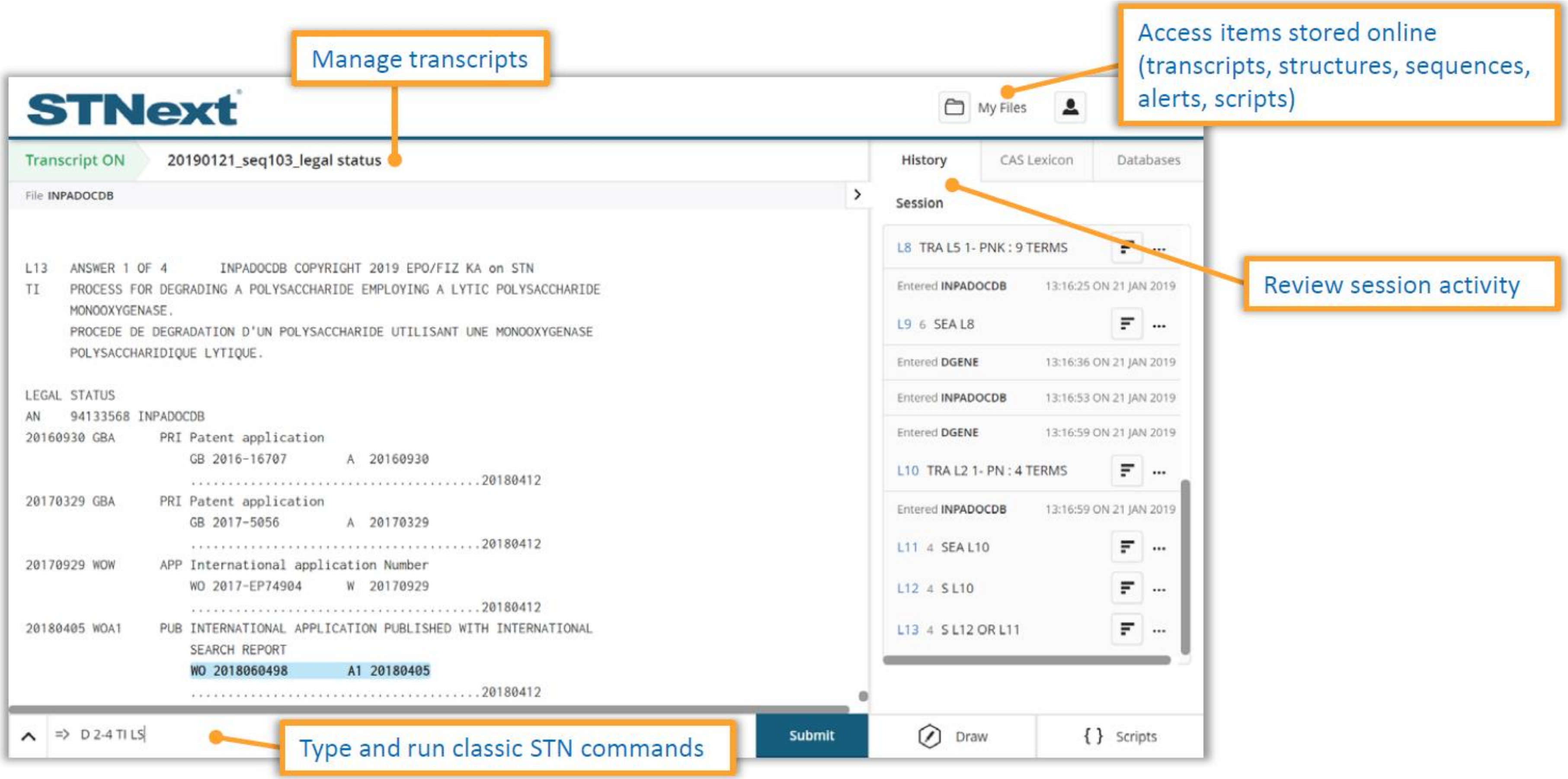
PCTGEN

- All peptide and nucleic acid sequences electronically submitted to WIPO, 2001 to present
- Records created from image format sequence listings are clearly marked („… created by using OCR…”)
- Updated weekly, within 1 day of publication
- 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

USGENE

- All available peptide and nucleic acid sequences from published applications and issued patents of USPTO
- Protein (>3 aa) and nucleic acid (>9 nt) sequences
- USPTO consolidates four sources (/SSO)
- 1981 – present, updated weekly, within 3 days of publication
- 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)

STNext



Manage transcripts

Access items stored online (transcripts, structures, sequences, alerts, scripts)

Review session activity

Type and run classic STN commands

```
L13   ANSWER 1 OF 4       INPADOCDB COPYRIGHT 2019 EPO/FIZ KA on STN
TI    PROCESS FOR DEGRADING A POLYSACCHARIDE EMPLOYING A LYTIC POLYSACCHARIDE
      MONOOXYGENASE.
      PROCEDE DE DEGRADATION D'UN POLYSACCHARIDE UTILISANT UNE MONOOXYGENASE
      POLYSACCHARIDIQUE LYTIQUE.

LEGAL STATUS
AN    94133568 INPADOCDB
20160930 GBA        PRI Patent application
                        GB 2016-16707        A  20160930
                        ......................................20180412
20170329 GBA        PRI Patent application
                        GB 2017-5056         A  20170329
                        ......................................20180412
20170929 WOW        APP International application Number
                        WO 2017-EP74904      W  20170929
                        ......................................20180412
20180405 WOA1       PUB INTERNATIONAL APPLICATION PUBLISHED WITH INTERNATIONAL
                        SEARCH REPORT
                        WO 2018060498        A1 20180405
                        ......................................20180412
```

```
=> D 2-4 TI LS
```

L8 TRA L5 1- PNK : 9 TERMS

Entered **INPADOCDB** 13:16:25 ON 21 JAN 2019

L9 6 SEA L8

Entered **DGENE** 13:16:36 ON 21 JAN 2019

Entered **INPADOCDB** 13:16:53 ON 21 JAN 2019

Entered **DGENE** 13:16:59 ON 21 JAN 2019

L10 TRA L2 1- PN : 4 TERMS

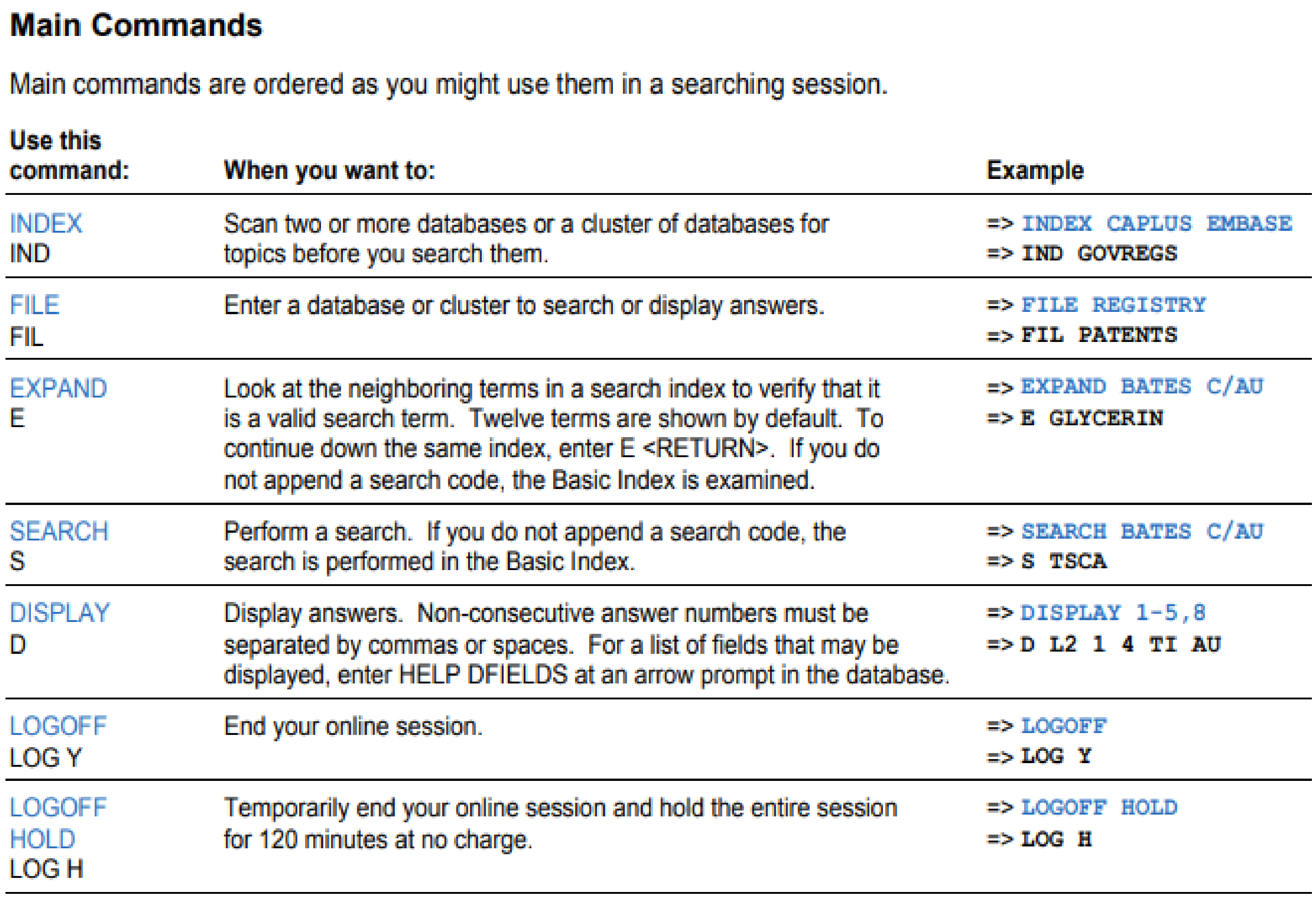
Entered **INPADOCDB** 13:16:59 ON 21 JAN 2019

L11 4 SEA L10

L12 4 S L10

L13 4 S L12 OR L11

Basic Commands



Main Commands

Main commands are ordered as you might use them in a searching session.

Use this command:	When you want to:	Example

INDEX
IND
	Scan two or more databases or a cluster of databases for topics before you search them.	`=> INDEX CAPLUS EMBASE`
		`=> IND GOVREGS`

FILE
FIL
	Enter a database or cluster to search or display answers.	`=> FILE REGISTRY`
		`=> FIL PATENTS`

EXPAND
E
	Look at the neighboring terms in a search index to verify that it is a valid search term. Twelve terms are shown by default. To continue down the same index, enter E <RETURN>. If you do not append a search code, the Basic Index is examined.	`=> EXPAND BATES C/AU`
		`=> E GLYCERIN`

SEARCH
S
	Perform a search. If you do not append a search code, the search is performed in the Basic Index.	`=> SEARCH BATES C/AU`
		`=> S TSCA`

DISPLAY
D
	Display answers. Non-consecutive answer numbers must be separated by commas or spaces. For a list of fields that may be displayed, enter HELP DFIELDS at an arrow prompt in the database.	`=> DISPLAY 1-5,8`
		`=> D L2 1 4 TI AU`

LOGOFF
LOG Y
	End your online session.	`=> LOGOFF`
		`=> LOG Y`

LOGOFF HOLD
LOG H
	Temporarily end your online session and hold the entire session for 120 minutes at no charge.	`=> LOGOFF HOLD`
		`=> LOG H`

Basic Commands

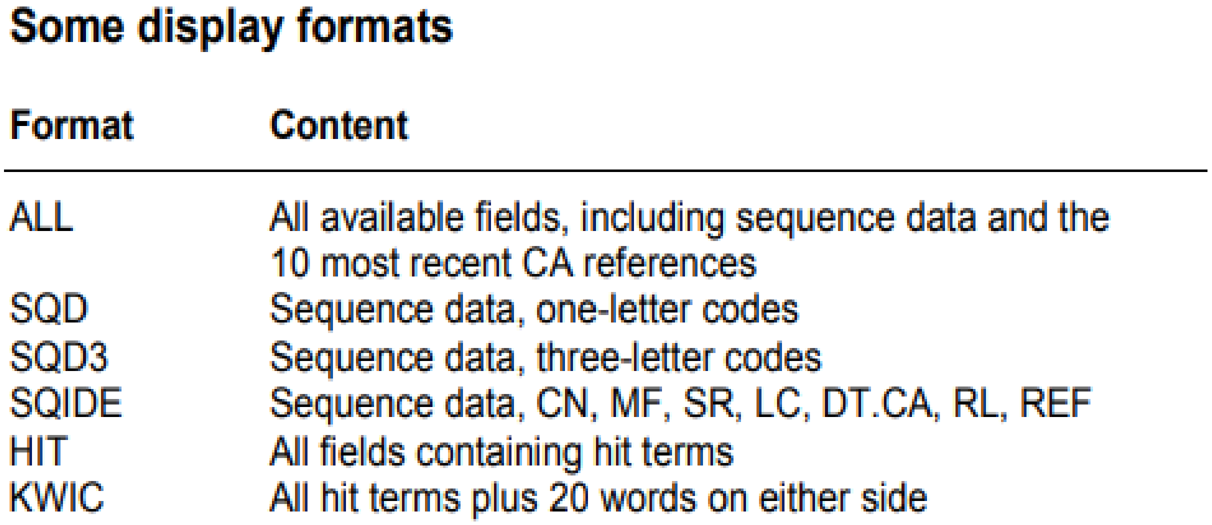
Display options

To display answers in REGISTRY, enter the DISPLAY (or D) command followed by the L-number resulting from a search, answer numbers or a range of numbers, and display fields or formats.

Display fields

Code	Content
RN	CAS Registry Number
CN	Chemical Name
PNTE	Patent Annotation
FS	File Segment
SQL	Sequence Length
NTE	Sequence Annotation
SEQ	Sequence (one-letter codes)
SEQ3	Sequence (three-letter codes)
MF	Molecular Formula
CI	Substance Class Identifier
SR	Source of Registration
LC	CAS Registry Number Locator
DT.CA	CAplus Document Type
RL	CAplus Super Roles
RL.NP	CAplus Super Roles from Non-patents
RL.P	CAplus Super Roles from Patents

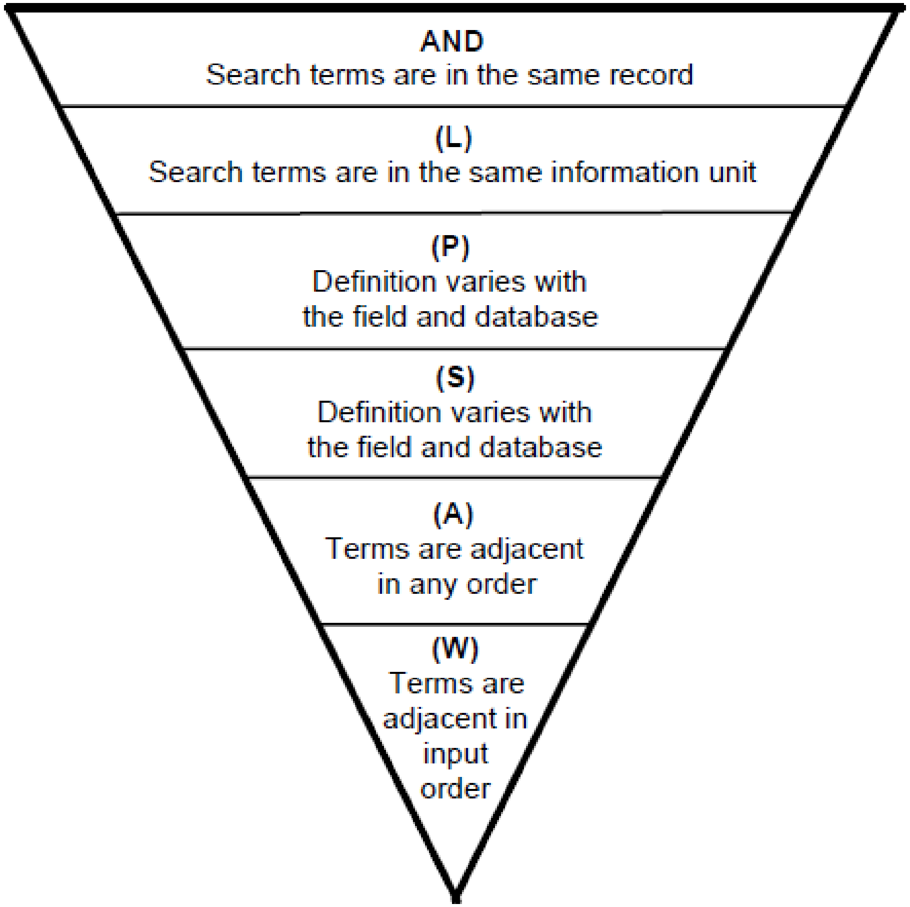
Basic Commands



Some display formats

Format	Content
ALL	All available fields, including sequence data and the 10 most recent CA references
SQD	Sequence data, one-letter codes
SQD3	Sequence data, three-letter codes
SQIDE	Sequence data, CN, MF, SR, LC, DT.CA, RL, REF
HIT	All fields containing hit terms
KWIC	All hit terms plus 20 words on either side

Proximity operators



AND
Search terms are in the same record

(L)
Search terms are in the same information unit

(P)
Definition varies with the field and database

(S)
Definition varies with the field and database

(A)
Terms are adjacent in any order

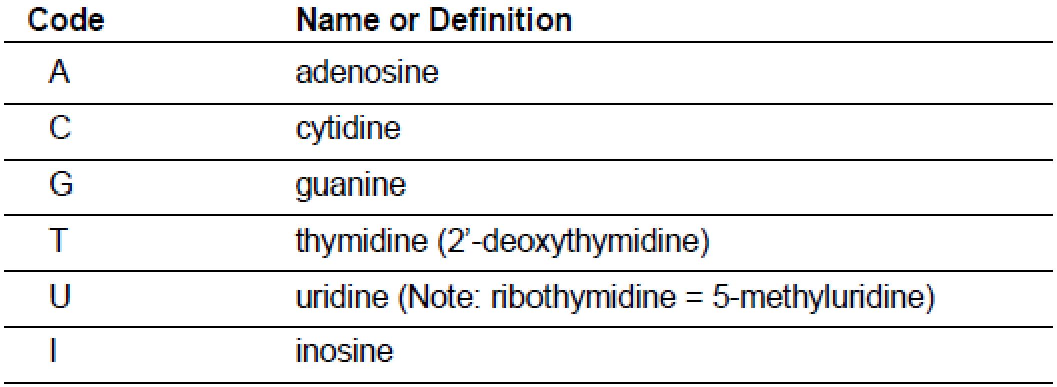
(W)
Terms are adjacent in input order

Exact and pattern Searching

Nucleic acid sequences

Searching exact sequences

To find an exact sequence of a nucleotide in REGISTRY, enter the sequence in the Exact Sequence Search (**/SQEN**) field.



Code	Name or Definition
A	adenosine
C	cytidine
G	guanine
T	thymidine (2’-deoxythymidine)
U	uridine (Note: ribothymidine = 5-methyluridine)
I	inosine

Using SEQLINK

The SEQLINK EXACT command is used to locate additional nucleic acid sequences that match a sequence that has already been retrieved from REGISTRY.

Nucleic acid sequences

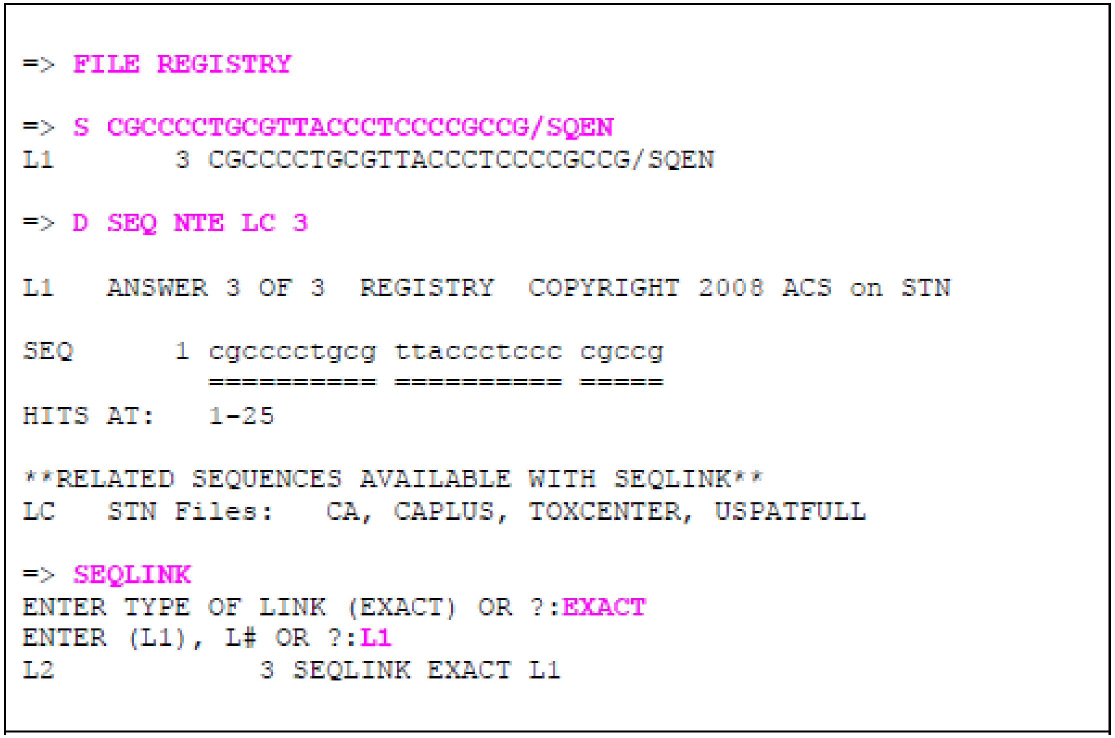
Find literature or patents on a diagnostic probe with the sequence CGCCCCTGCGTTACCCTCCCCGCCG.

1 *Enter REGISTRY.*

2 *Use the SEARCH (or S) command to search the exact sequence in the /SQEN field.*

3 *Display the sequence (SEQ), annotation (NTE), and the Locator (LC) field listing the databases containing references to the CAS Registry Number®.*

4 *Use the SEQLINK command (free of charge) to find related sequences, if any.*



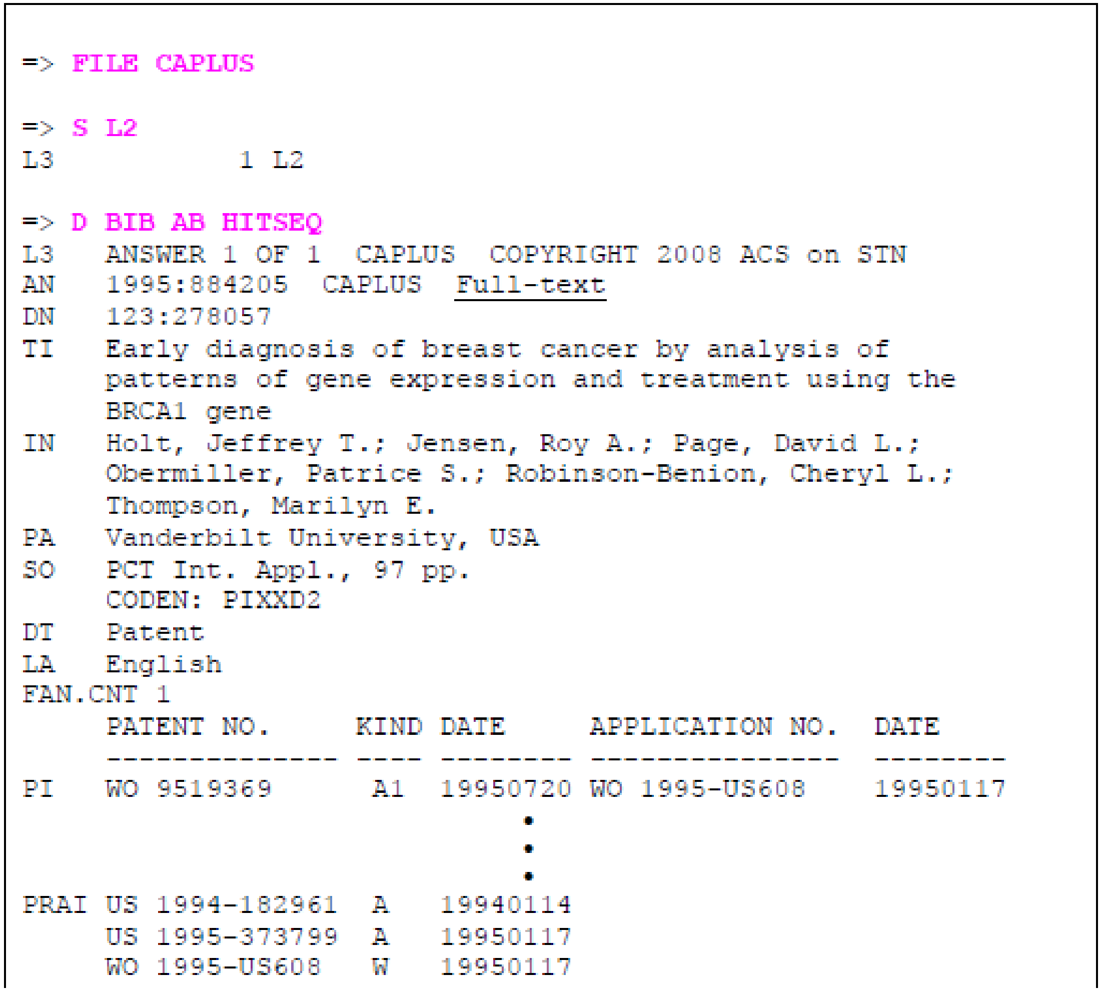
```
=> FILE REGISTRY

=> S CGCCCCTGCGTTACCCTCCCCGCCG/SQEN
L1       3 CGCCCCTGCGTTACCCTCCCCGCCG/SQEN

=> D SEQ NTE LC 3

L1   ANSWER 3 OF 3  REGISTRY  COPYRIGHT 2008 ACS on STN

SEQ      1 cgcccctgcg ttaccctccc cgccg
           ========== ========== =====

HITS AT:   1-25

**RELATED SEQUENCES AVAILABLE WITH SEQLINK**
LC   STN Files:   CA, CAPLUS, TOXCENTER, USPATFULL

=> SEQLINK
ENTER TYPE OF LINK (EXACT) OR ?:EXACT
ENTER (L1), L# OR ?:L1
L2            3 SEQLINK EXACT L1
```

Nucleic acid sequences

5 *Enter one or more of the databases containing the CAS Registry Number.*

6 *Search the REGISTRY L-number (L2).*

7 *Display the bibliographic information (BIB), abstract (AB), and index entry for the hit sequence (HITSEQ).*



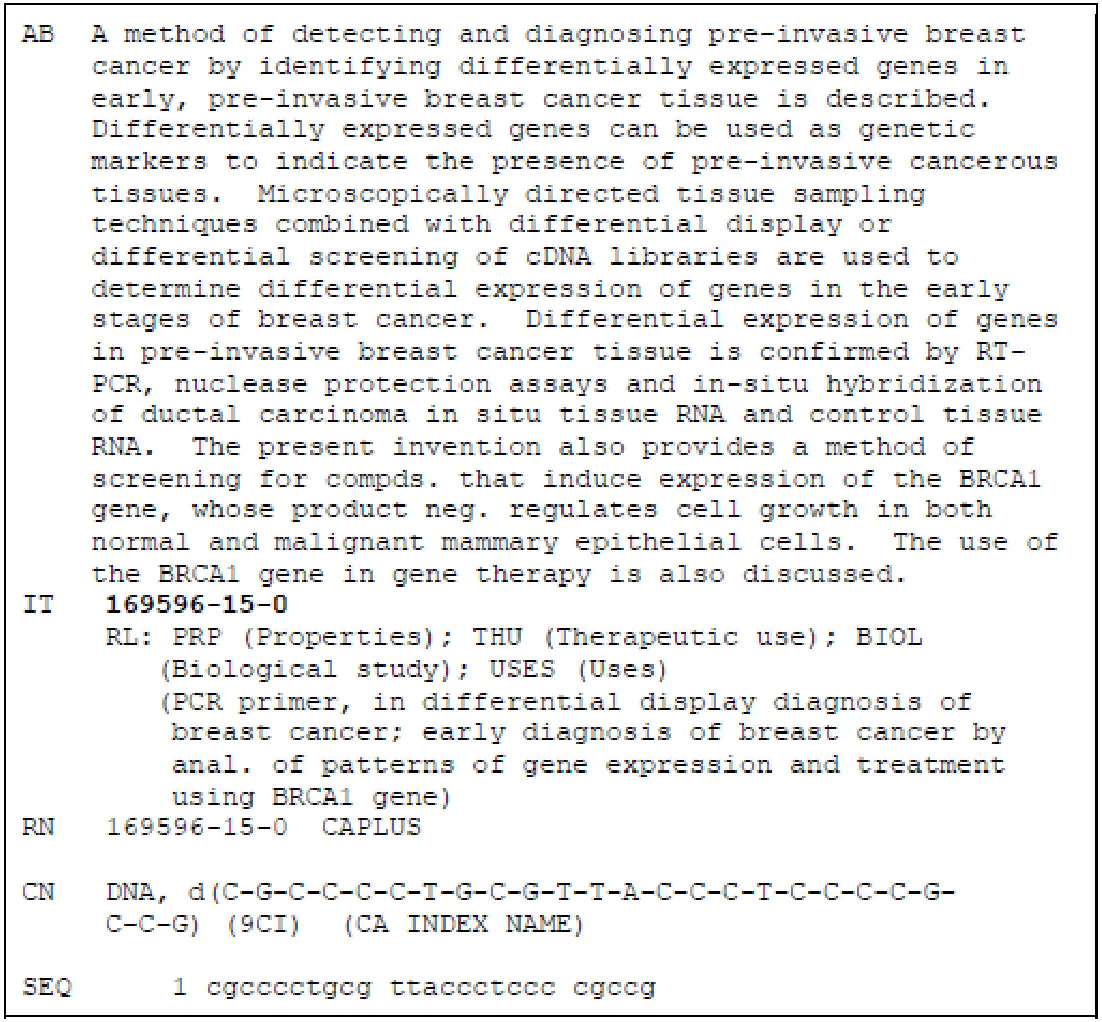
```
=> FILE CAPLUS

=> S L2
L3           1 L2

=> D BIB AB HITSEQ
L3   ANSWER 1 OF 1  CAPLUS  COPYRIGHT 2008 ACS on STN
AN   1995:884205  CAPLUS  Full-text
DN   123:278057
TI   Early diagnosis of breast cancer by analysis of
     patterns of gene expression and treatment using the
     BRCA1 gene
IN   Holt, Jeffrey T.; Jensen, Roy A.; Page, David L.;
     Obermiller, Patrice S.; Robinson-Benion, Cheryl L.;
     Thompson, Marilyn E.
PA   Vanderbilt University, USA
SO   PCT Int. Appl., 97 pp.
     CODEN: PIXXD2
DT   Patent
LA   English
FAN.CNT 1
     PATENT NO.     KIND DATE     APPLICATION NO.   DATE
     -------------- ---- -------- ---------------   --------
PI   WO 9519369      A1  19950720 WO 1995-US608     19950117
                              •
                              •
                              •
PRAI US 1994-182961  A   19940114
     US 1995-373799  A   19950117
     WO 1995-US608   W   19950117
```

Nucleic acid sequences

5 Enter one or more of the databases containing the CAS Registry Number.

6 Search the REGISTRY L-number (L2).

7 Display the bibliographic information (BIB), abstract (AB), and index entry for the hit sequence (HITSEQ).

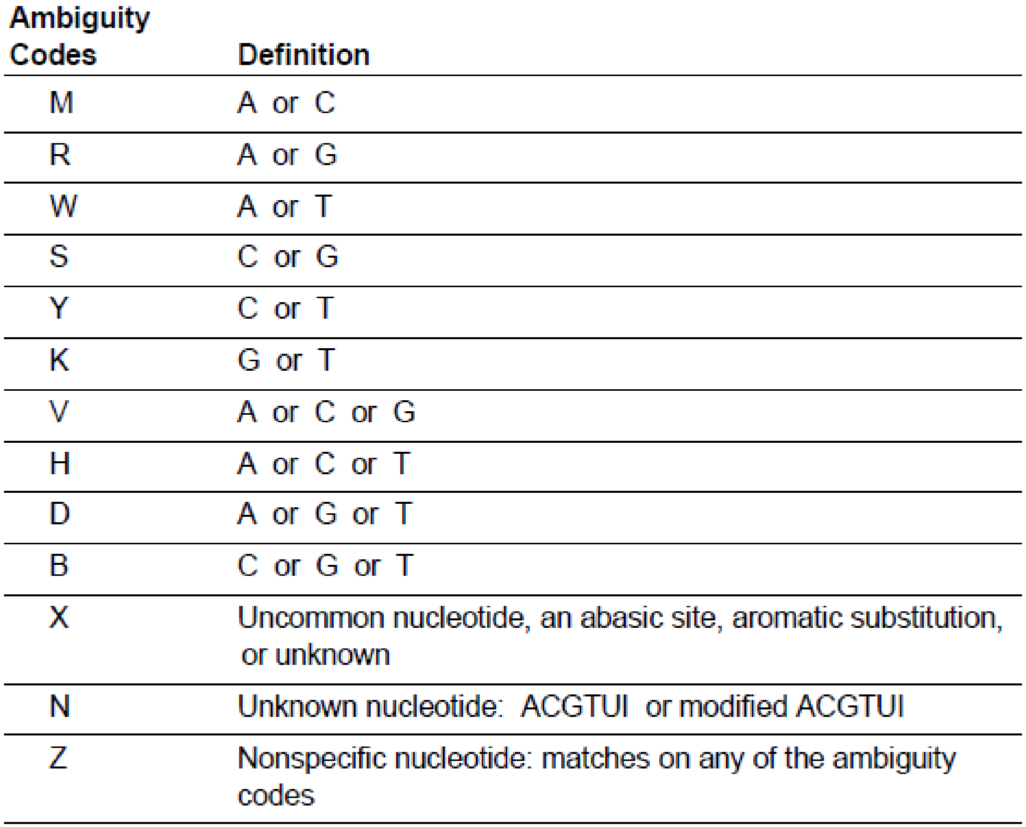


```
AB   A method of detecting and diagnosing pre-invasive breast
     cancer by identifying differentially expressed genes in
     early, pre-invasive breast cancer tissue is described.
     Differentially expressed genes can be used as genetic
     markers to indicate the presence of pre-invasive cancerous
     tissues.  Microscopically directed tissue sampling
     techniques combined with differential display or
     differential screening of cDNA libraries are used to
     determine differential expression of genes in the early
     stages of breast cancer.  Differential expression of genes
     in pre-invasive breast cancer tissue is confirmed by RT-
     PCR, nuclease protection assays and in-situ hybridization
     of ductal carcinoma in situ tissue RNA and control tissue
     RNA.  The present invention also provides a method of
     screening for compds. that induce expression of the BRCA1
     gene, whose product neg. regulates cell growth in both
     normal and malignant mammary epithelial cells.  The use of
     the BRCA1 gene in gene therapy is also discussed.
IT    169596-15-0
      RL: PRP (Properties); THU (Therapeutic use); BIOL
          (Biological study); USES (Uses)
          (PCR primer, in differential display diagnosis of
           breast cancer; early diagnosis of breast cancer by
           anal. of patterns of gene expression and treatment
           using BRCA1 gene)
RN    169596-15-0  CAPLUS

CN    DNA, d(C-G-C-C-C-C-T-G-C-G-T-T-A-C-C-C-T-C-C-C-C-G-
      C-C-G)  (9CI)   (CA INDEX NAME)

SEQ       1 cgcccctgcg ttaccctccc cgccg
```

Nucleic acid sequences

Searching partial sequences

To find partial sequences or sequences with gaps, repeating units, or alternate units, search the partial sequence in the Subsequence Search (**/SQSN**) field in REGISTRY. You can use the codes for specific nucleotides or ambiguity codes.

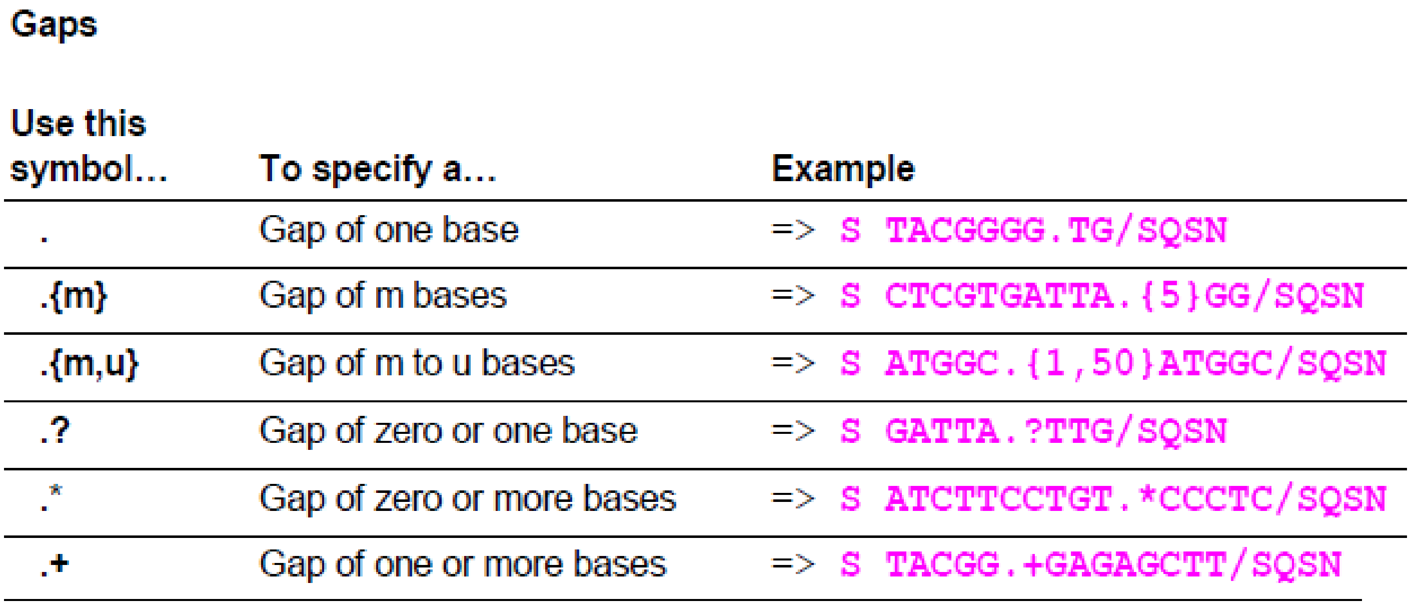


Ambiguity Codes	Definition
M	A or C
R	A or G
W	A or T
S	C or G
Y	C or T
K	G or T
V	A or C or G
H	A or C or T
D	A or G or T
B	C or G or T
X	Uncommon nucleotide, an abasic site, aromatic substitution, or unknown
N	Unknown nucleotide: ACGTUI or modified ACGTUI
Z	Nonspecific nucleotide: matches on any of the ambiguity codes

Nucleic acid sequences

Pattern Searching

Complex pattern searching of nucleic acid subsequences is possible using special notations for gaps, repeating resides, and other types of variability.



Gaps

Use this symbol...	To specify a...	Example
.	Gap of one base	=> S TACGGGG.TG/SQSN
.{m}	Gap of m bases	=> S CTCGTGATTA.{5}GG/SQSN
.{m,u}	Gap of m to u bases	=> S ATGGC.{1,50}ATGGC/SQSN
.?	Gap of zero or one base	=> S GATTA.?TTG/SQSN
.*	Gap of zero or more bases	=> S ATCTTCCTGT.*CCCTC/SQSN
.+	Gap of one or more bases	=> S TACGG.+GAGAGCTT/SQSN

Nucleic acid sequences

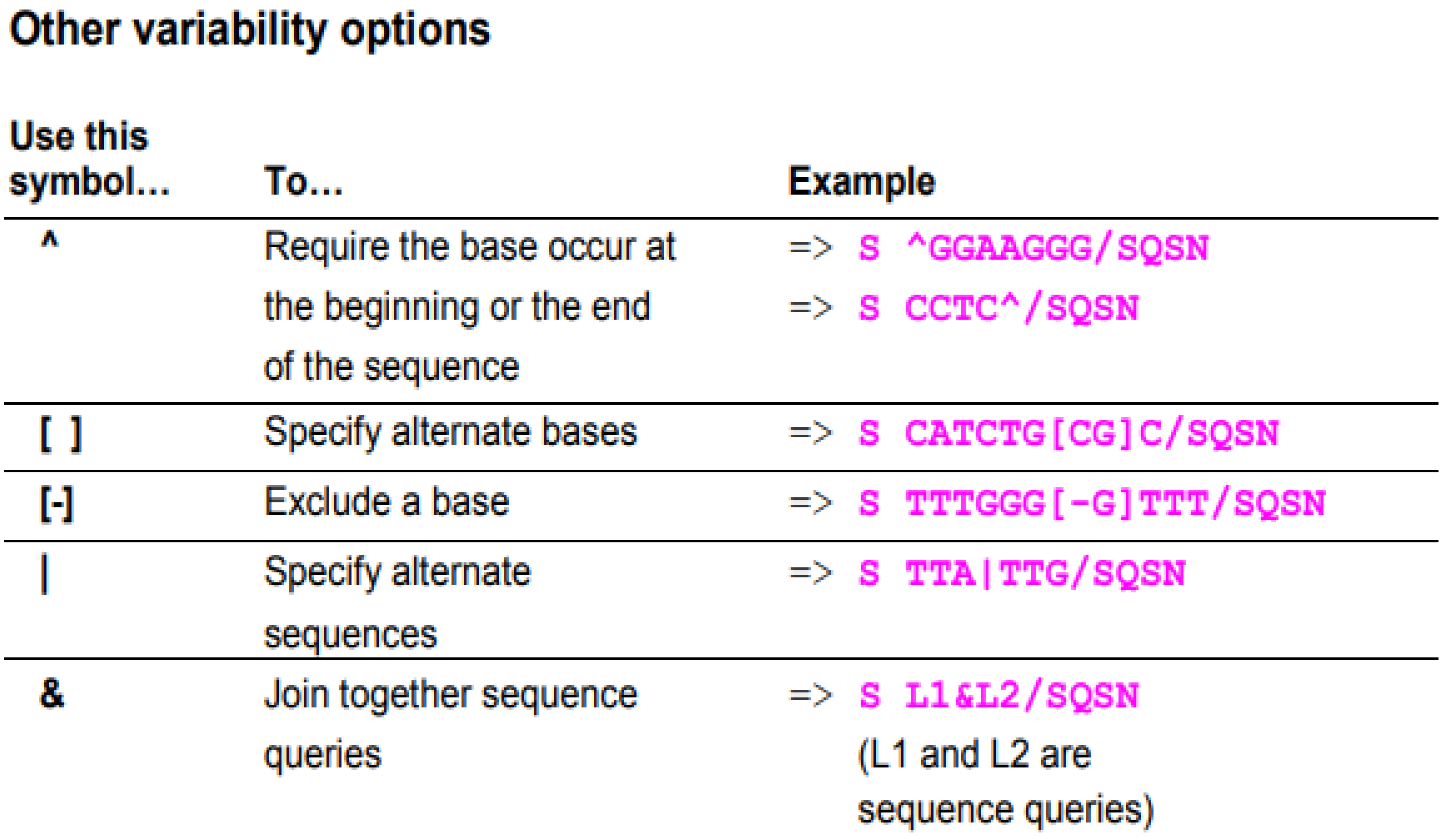
Pattern Searching

Repetition

Use this symbol…	To…	Example
{ } with a number or range	Repeat the preceding unit	=> S GAAT(TAA){2}/SQSN
?	Repeat the preceding unit zero or one time	=> S CAT(CGA)?GGAC/SQSN
*	Repeat the preceding unit zero or more times	=> S CAT(CTG)*TATT/SQSN
+	Repeat the preceding unit one or more times	=> S CAT(CTG)+TATT/SQSN

Nucleic acid sequences

Pattern Searching



Other variability options

Use this symbol…	To…	Example
^	Require the base occur at the beginning or the end of the sequence	=> S ^GGAAGGG/SQSN => S CCTC^/SQSN
[ ]	Specify alternate bases	=> S CATCTG[CG]C/SQSN
[-]	Exclude a base	=> S TTTGGG[-G]TTT/SQSN
|	Specify alternate sequences	=> S TTA|TTG/SQSN
&	Join together sequence queries	=> S L1&L2/SQSN (L1 and L2 are sequence queries)

Nucleic acid sequences

.	Gap of one base
.{m}	Gap of m bases

Pattern searching example

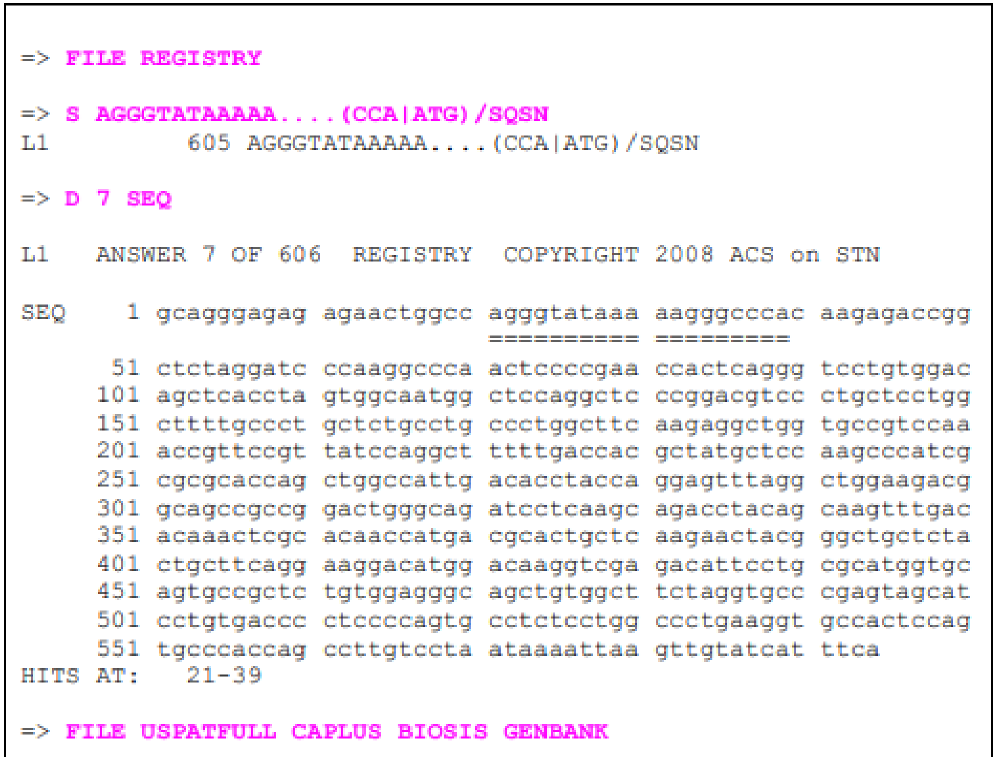
Find patents and literature on the following partial sequence: AGGGTATAAAAA....(CCA|ATG), where .... is a gap of four nucleotides followed by either CCA or ATG.

1 Enter REGISTRY.

2 Search the partial sequence in the /SQSN field.

3 Display the sequence (SEQ).

4 Enter the reference databases containing CAS Registry Numbers for the sequences.



```
=> FILE REGISTRY

=> S AGGGTATAAAAA....(CCA|ATG)/SQSN
L1           605 AGGGTATAAAAA....(CCA|ATG)/SQSN

=> D 7 SEQ

L1   ANSWER 7 OF 606  REGISTRY  COPYRIGHT 2008 ACS on STN

SEQ    1 gcagggagag agaactggcc agggtataaa aagggcccac aagagaccgg
                               ========== =========
      51 ctctaggatc ccaaggccca actccccgaa ccactcaggg tcctgtggac
     101 agctcaccta gtggcaatgg ctccaggctc ccggacgtcc ctgctcctgg
     151 cttttgccct gctctgcctg ccctggcttc aagaggctgg tgccgtccaa
     201 accgttccgt tatccaggct ttttgaccac gctatgctcc aagcccatcg
     251 cgcgcaccag ctggccattg acacctacca ggagtttagg ctggaagacg
     301 gcagccgccg gactgggcag atcctcaagc agacctacag caagtttgac
     351 acaaactcgc acaaccatga cgcactgctc aagaactacg ggctgctcta
     401 ctgcttcagg aaggacatgg acaaggtcga gacattcctg cgcatggtgc
     451 agtgccgctc tgtggagggc agctgtggct tctaggtgcc cgagtagcat
     501 cctgtgaccc ctccccagtg cctctcctgg ccctgaaggt gccactccag
     551 tgcccaccag ccttgtccta ataaaattaa gttgtatcat ttca
HITS AT:   21-39

=> FILE USPATFULL CAPLUS BIOSIS GENBANK
```

Nucleic acid sequences

5 Enter SET MSTEPS ON to create an L-number for a search in each database.

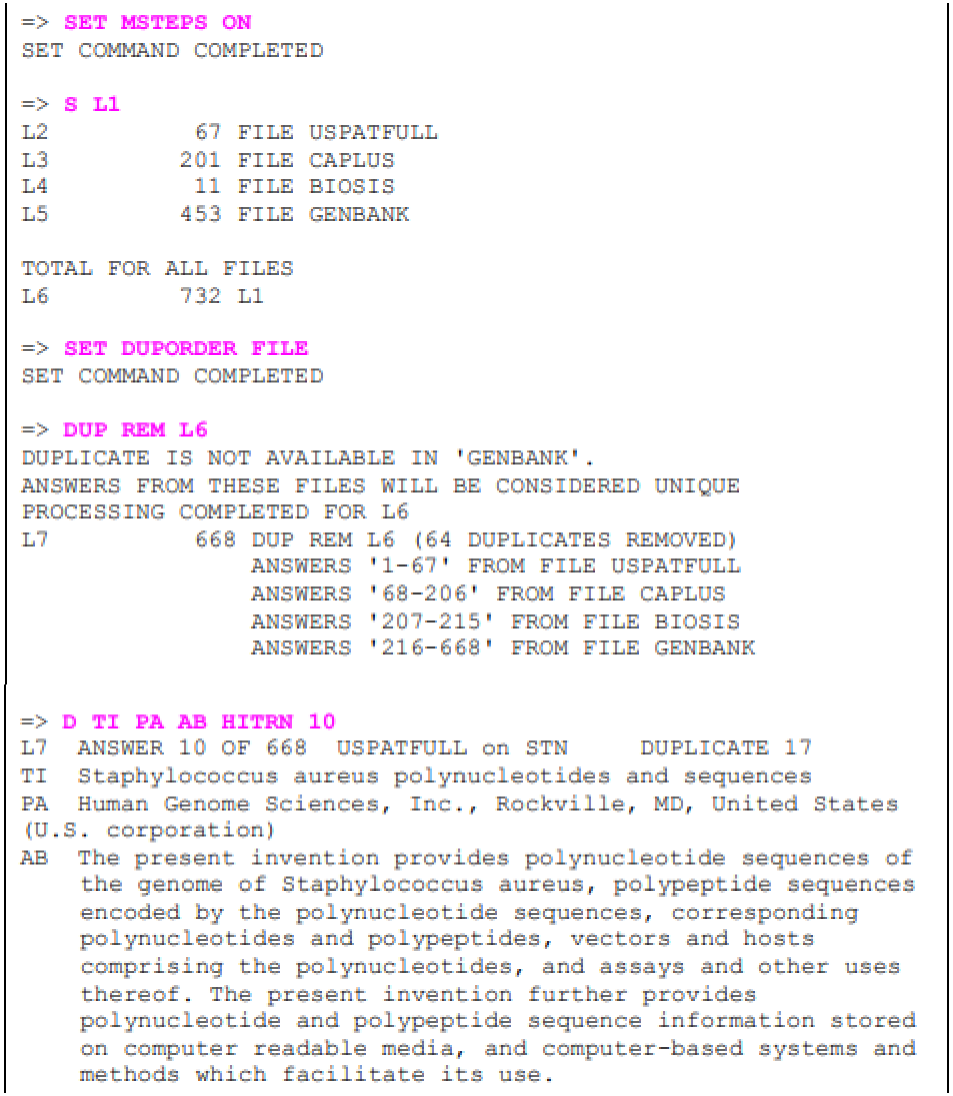
6 Search the REGISTRY L-number (L1). Each database is searched, and an L-number answer set is created in each database. A composite L-number (L6) with all references is created.

7 Set the arrangement of answers in database order in the process of duplicate identification or elimination.

8 Remove duplicates. Answers are arranged in database order.

9 Display references from selected databases.

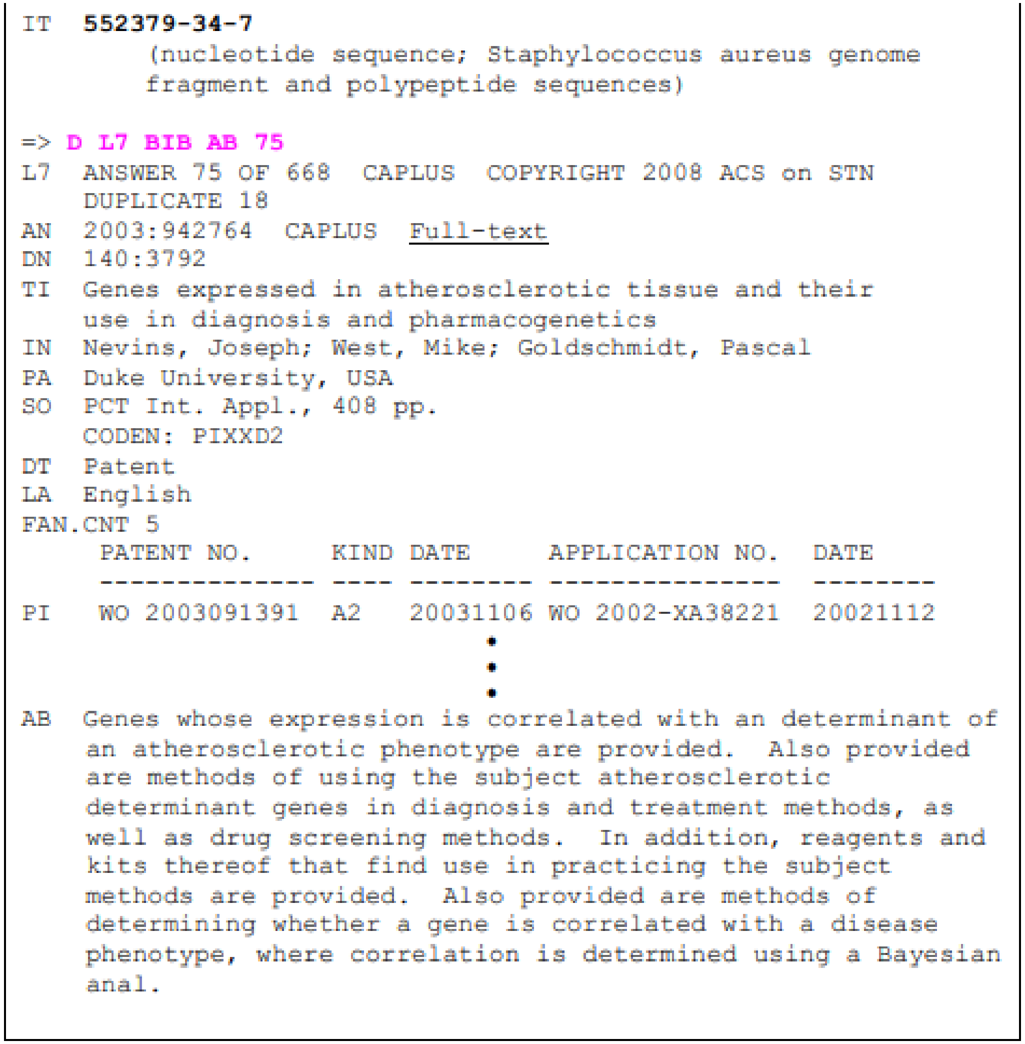
Answer 10 is from USPATFULL.



```
=> SET MSTEPS ON
SET COMMAND COMPLETED

=> S L1
L2            67 FILE USPATFULL
L3           201 FILE CAPLUS
L4            11 FILE BIOSIS
L5           453 FILE GENBANK

TOTAL FOR ALL FILES
L6           732 L1

=> SET DUPORDER FILE
SET COMMAND COMPLETED

=> DUP REM L6
DUPLICATE IS NOT AVAILABLE IN 'GENBANK'.
ANSWERS FROM THESE FILES WILL BE CONSIDERED UNIQUE
PROCESSING COMPLETED FOR L6
L7            668 DUP REM L6 (64 DUPLICATES REMOVED)
                  ANSWERS '1-67' FROM FILE USPATFULL
                  ANSWERS '68-206' FROM FILE CAPLUS
                  ANSWERS '207-215' FROM FILE BIOSIS
                  ANSWERS '216-668' FROM FILE GENBANK

=> D TI PA AB HITRN 10
L7  ANSWER 10 OF 668  USPATFULL on STN     DUPLICATE 17
TI  Staphylococcus aureus polynucleotides and sequences
PA  Human Genome Sciences, Inc., Rockville, MD, United States
(U.S. corporation)
AB  The present invention provides polynucleotide sequences of
    the genome of Staphylococcus aureus, polypeptide sequences
    encoded by the polynucleotide sequences, corresponding
    polynucleotides and polypeptides, vectors and hosts
    comprising the polynucleotides, and assays and other uses
    thereof. The present invention further provides
    polynucleotide and polypeptide sequence information stored
    on computer readable media, and computer-based systems and
    methods which facilitate its use.
```

Nucleic acid sequences

Answer 75 is from CAplus.

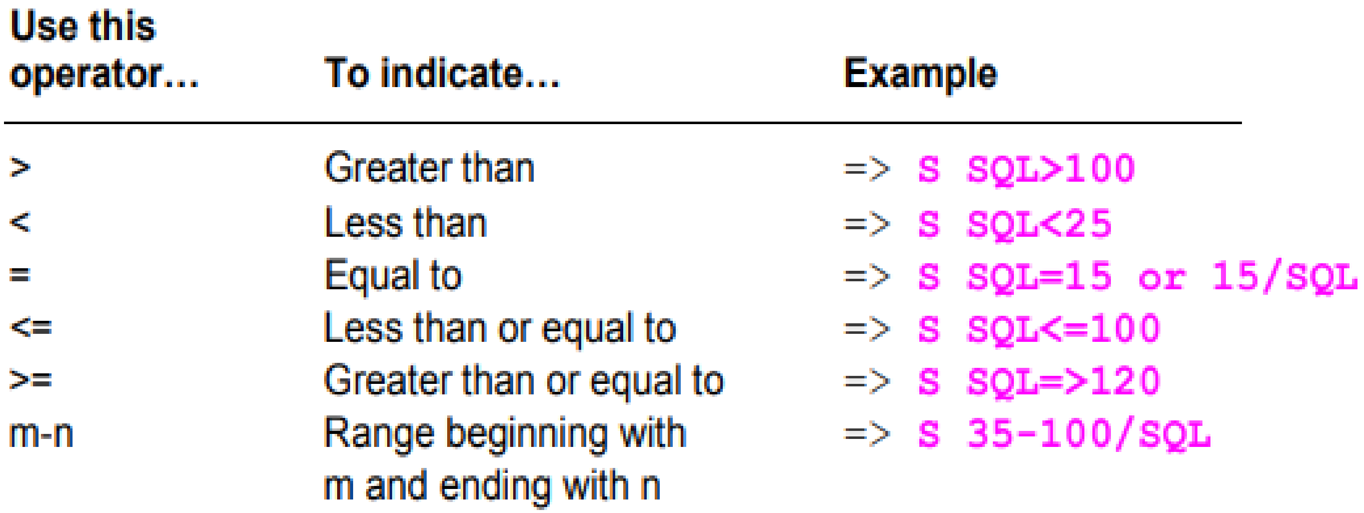


```
IT   552379-34-7
          (nucleotide sequence; Staphylococcus aureus genome
          fragment and polypeptide sequences)

=> D L7 BIB AB 75
L7   ANSWER 75 OF 668  CAPLUS  COPYRIGHT 2008 ACS on STN
     DUPLICATE 18
AN   2003:942764  CAPLUS  Full-text
DN   140:3792
TI   Genes expressed in atherosclerotic tissue and their
     use in diagnosis and pharmacogenetics
IN   Nevins, Joseph; West, Mike; Goldschmidt, Pascal
PA   Duke University, USA
SO   PCT Int. Appl., 408 pp.
     CODEN: PIXXD2
DT   Patent
LA   English
FAN.CNT 5
      PATENT NO.      KIND DATE     APPLICATION NO.  DATE
      --------------- ---- -------- ---------------  --------
PI    WO 2003091391   A2   20031106 WO 2002-XA38221  20021112
                                 •
                                 •
                                 •
AB   Genes whose expression is correlated with an determinant of
     an atherosclerotic phenotype are provided.  Also provided
     are methods of using the subject atherosclerotic
     determinant genes in diagnosis and treatment methods, as
     well as drug screening methods.  In addition, reagents and
     kits thereof that find use in practicing the subject
     methods are provided.  Also provided are methods of
     determining whether a gene is correlated with a disease
     phenotype, where correlation is determined using a Bayesian
     anal.
```

Nucleic acid sequences

Searching length

You can refine a sequence search by combining it with a search of sequence length in the Sequence Length (/SQL) field.



Use this operator...	To indicate...	Example
>	Greater than	=> S SQL>100
<	Less than	=> S SQL<25
=	Equal to	=> S SQL=15 or 15/SQL
<=	Less than or equal to	=> S SQL<=100
>=	Greater than or equal to	=> S SQL=>120
m-n	Range beginning with m and ending with n	=> S 35-100/SQL

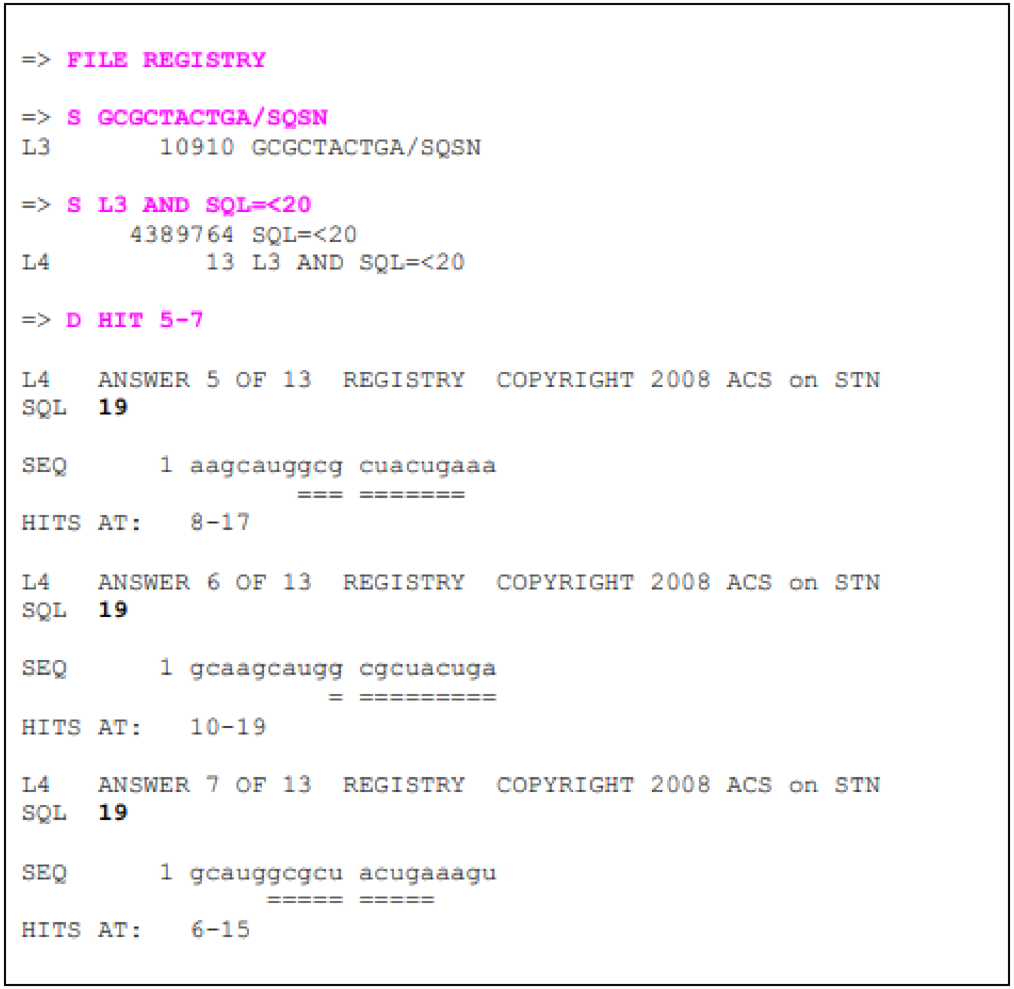
Nucleic acid sequences

Find GCGCTACTGA containing sequences with 20 or fewer nucleotides.

1 *Enter REGISTRY and search the sequence.*

2 *Search SQL<=20 to retrieve only sequences with 20 or fewer residues.*

3 *Display some answers in the HIT format.*



```
=> FILE REGISTRY

=> S GCGCTACTGA/SQSN
L3       10910 GCGCTACTGA/SQSN

=> S L3 AND SQL=<20
       4389764 SQL=<20
L4          13 L3 AND SQL=<20

=> D HIT 5-7

L4   ANSWER 5 OF 13  REGISTRY  COPYRIGHT 2008 ACS on STN
SQL  19

SEQ      1 aagcauggcg cuacugaaa
                  === =======
HITS AT:   8-17

L4   ANSWER 6 OF 13  REGISTRY  COPYRIGHT 2008 ACS on STN
SQL  19

SEQ      1 gcaagcaugg cgcuacuga
                    = =========
HITS AT:   10-19

L4   ANSWER 7 OF 13  REGISTRY  COPYRIGHT 2008 ACS on STN
SQL  19

SEQ      1 gcauggcgcu acugaaagu
                ===== =====
HITS AT:   6-15
```

Protein sequences

Common amino acids

1-Letter Code	3-Letter Code	Name
A	Ala	Alanine
B	Asx	Aspartic acid or Asparagine
C	Cys	Cysteine
D	Asp	Aspartic acid
E	Glu	Glutamic acid
F	Phe	Phenylalanine
G	Gly	Glycine
H	His	Histidine
I	Ile	Isoleucine
J	Xle	Isoleucine or Leucine
K	Lys	Lysine
L	Leu	Leucine
M	Met	Methionine
N	Asn	Asparagine
O	Pyl	Pyrrolysine
P	Pro	Proline
Q	Gln	Glutamine
R	Arg	Arginine
S	Ser	Serine
T	Thr	Threonine
U	Scy	Selenocysteine
V	Val	Valine
W	Trp	Tryptophan
X	Xxx	Uncommon or Unspecified
Y	Tyr	Tyrosine
Z	Glx	Glutamic acid or Glutamine

Uncommon amino acids

3-Letter Code	Name
Aaa	α-amino acid
Aad	2-aminoadipic acid (2-aminohexanedioic acid)
Aan	α-asparagine
Abu	2-aminobutanoic acid
Aca	2-aminocapric acid (2-aminodecanoic acid)
Agn	α-glutamine
Aib	α-aminoisobutyric acid (α-methylalanine)
Apm	2-aminopimelic acid (2-aminoheptanedioic acid)

Note: The codes B, J, and Z may be used only in subsequence searches (/SQSP and /SQSFP).

Protein sequences

Tips

- Use 1-letter codes for common resides
- Use 3-letter codes for uncommon residues

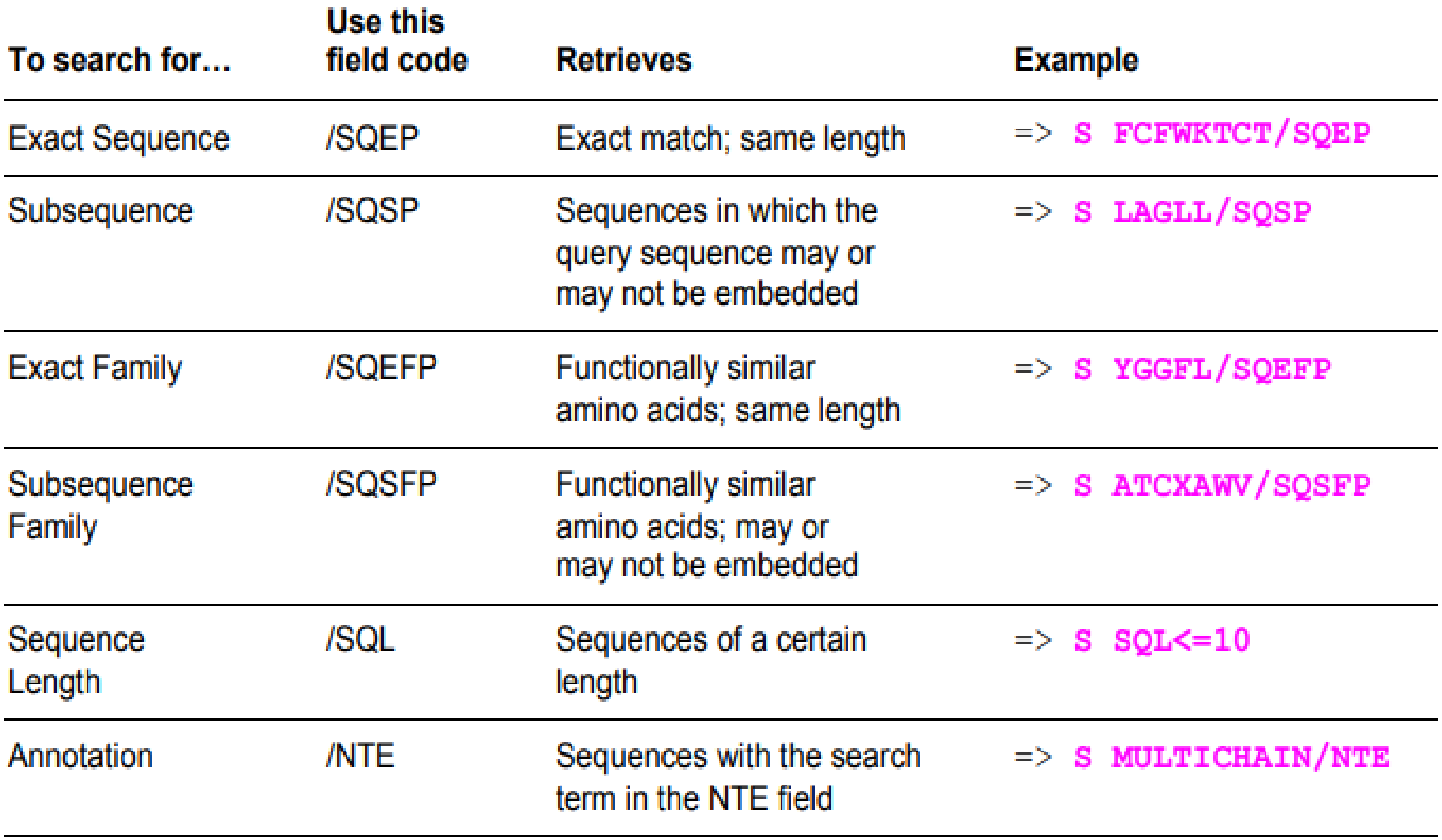
  - Enclose 3-letter codes in single quotes
- 1-letter and 3-letter codes can be mixed

  e.g. => S ‘AIB’A’ABU’’PIP’/SQSP
- Search shortcuts for Blocking groups in the Notes (NTE) field

  e.g. => S BOC/NTE

Protein sequences

Search options



To search for...	Use this field code	Retrieves	Example
Exact Sequence	/SQEP	Exact match; same length	=> S FCFWKTCT/SQEP
Subsequence	/SQSP	Sequences in which the query sequence may or may not be embedded	=> S LAGLL/SQSP
Exact Family	/SQEFP	Functionally similar amino acids; same length	=> S YGGFL/SQEFP
Subsequence Family	/SQSFP	Functionally similar amino acids; may or may not be embedded	=> S ATCXAWV/SQSFP
Sequence Length	/SQL	Sequences of a certain length	=> S SQL<=10
Annotation	/NTE	Sequences with the search term in the NTE field	=> S MULTICHAIN/NTE

Protein sequences

Searching for exact sequence strings

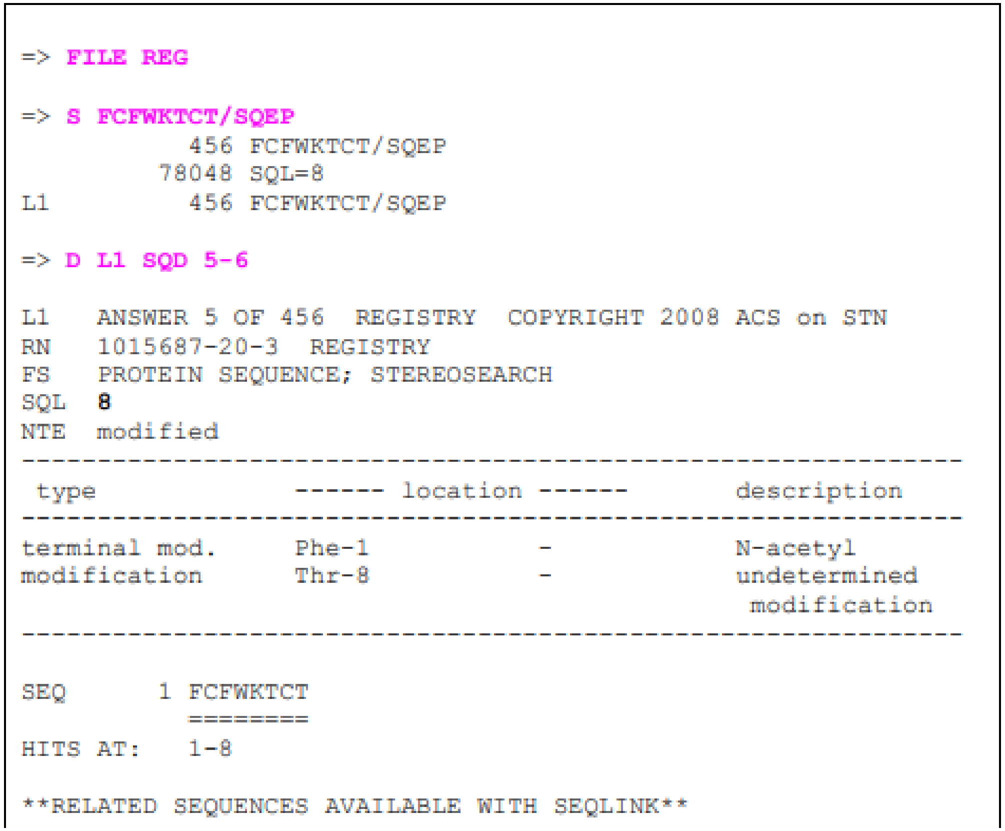
1 Enter REGISTRY.

2 Enter S (SEARCH) and the exact sequence in the /SQEP field. You can use one-letter codes for common amino acids.

An L-number answer set (L1) is created. The number of sequences retrieved (456) is displayed.

3 Display sequence data by entering D (DISPLAY), the L-number, the format, and the answer numbers. The SQD format includes the CAS Registry Number® and sequence data using one-letter codes.

Find analogs of the drug Sandostatin with the sequence FCFWKTCT.

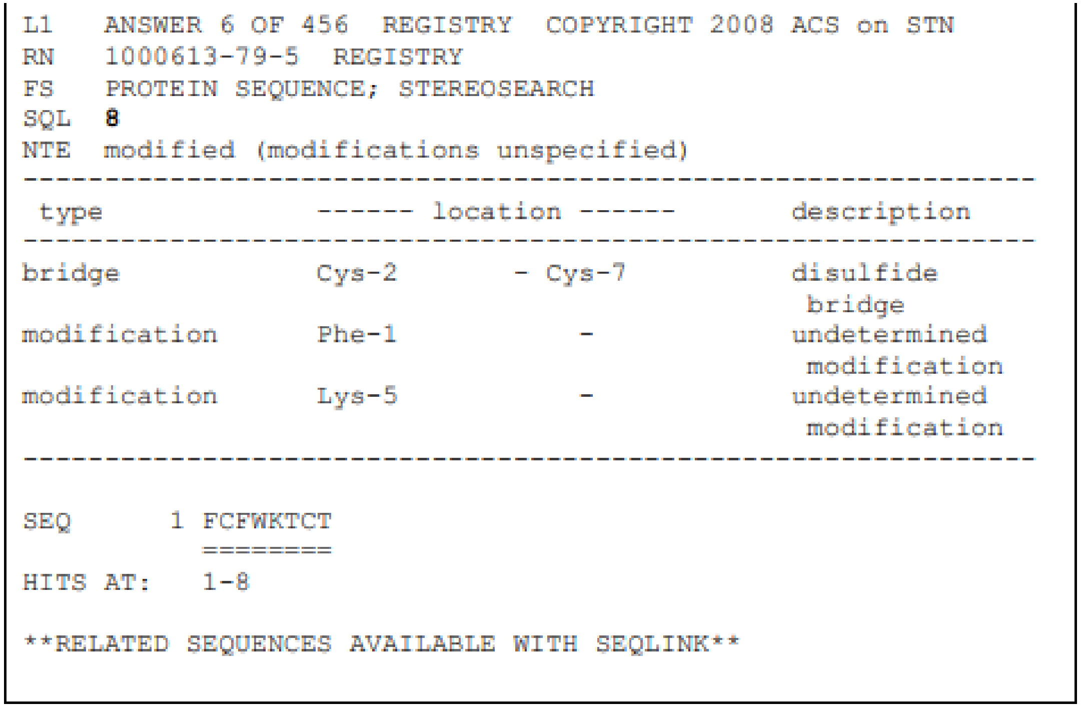


```
=> FILE REG

=> S FCFWKTCT/SQEP
            456 FCFWKTCT/SQEP
          78048 SQL=8
L1          456 FCFWKTCT/SQEP

=> D L1 SQD 5-6

L1    ANSWER 5 OF 456   REGISTRY   COPYRIGHT 2008 ACS on STN
RN    1015687-20-3   REGISTRY
FS    PROTEIN SEQUENCE; STEREOSEARCH
SQL   8
NTE   modified
--------------------------------------------------------------
 type                  ------ location ------        description
--------------------------------------------------------------
terminal mod.      Phe-1              -             N-acetyl
modification       Thr-8              -             undetermined
                                                     modification
--------------------------------------------------------------

SEQ        1 FCFWKTCT
             ========
HITS AT:   1-8

**RELATED SEQUENCES AVAILABLE WITH SEQLINK**
```

Protein sequences

Searching for exact sequence strings (cont.)

The answers have the same sequence and length, but they differ in chemical annotation in the NTE field.



```
L1   ANSWER 6 OF 456  REGISTRY  COPYRIGHT 2008 ACS on STN
RN   1000613-79-5  REGISTRY
FS   PROTEIN SEQUENCE; STEREOSEARCH
SQL  8
NTE  modified (modifications unspecified)
--------------------------------------------------------------
 type                 ------ location ------       description
--------------------------------------------------------------
bridge                Cys-2        - Cys-7         disulfide
                                                    bridge
modification          Phe-1            -           undetermined
                                                    modification
modification          Lys-5            -           undetermined
                                                    modification
--------------------------------------------------------------

SEQ       1 FCFWKTCT
            ========
HITS AT:    1-8

**RELATED SEQUENCES AVAILABLE WITH SEQLINK**
```

Protein sequences

Searching Subsequences

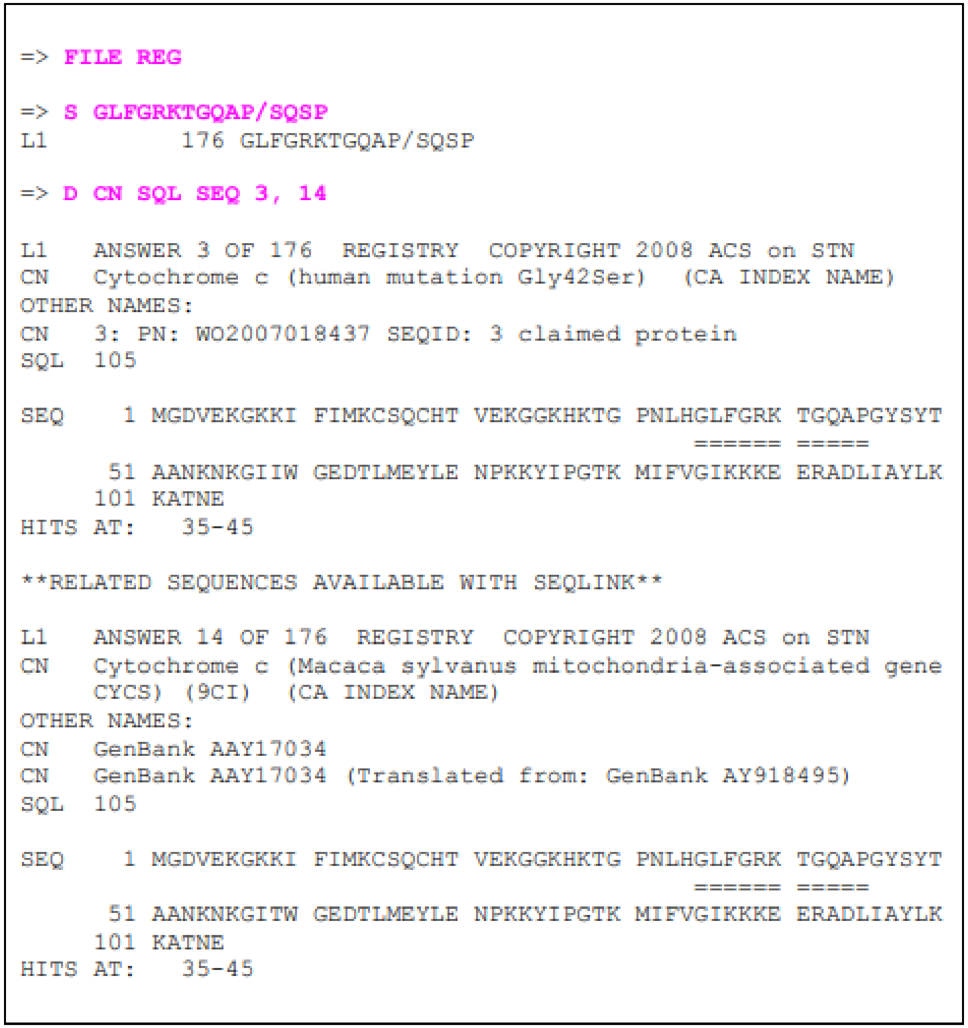
1 Enter REGISTRY.

2 Search the subsequence in the /SQSP field. You can use one-letter codes for common amino acids.

3 Display chemical names (CN), sequence length (SQL), and sequences using one-letter codes (SEQ).

Notice the different chemical names and variable sequence length. The query subsequence is highlighted.

Find proteins containing the sequence string GLFGRKTGQAP from the human cytochrome c.



```
=> FILE REG

=> S GLFGRKTGQAP/SQSP
L1         176 GLFGRKTGQAP/SQSP

=> D CN SQL SEQ 3, 14

L1   ANSWER 3 OF 176  REGISTRY  COPYRIGHT 2008 ACS on STN
CN   Cytochrome c (human mutation Gly42Ser)  (CA INDEX NAME)
OTHER NAMES:
CN   3: PN: WO2007018437 SEQID: 3 claimed protein
SQL  105

SEQ    1 MGDVEKGKKI FIMKCSQCHT VEKGGKHKTG PNLHGLFGRK TGQAPGYSYT
                                              ====== =====
      51 AANKNKGIIW GEDTLMEYLE NPKKYIPGTK MIFVGIKKKE ERADLIAYLK
     101 KATNE
HITS AT:   35-45

**RELATED SEQUENCES AVAILABLE WITH SEQLINK**

L1   ANSWER 14 OF 176  REGISTRY  COPYRIGHT 2008 ACS on STN
CN   Cytochrome c (Macaca sylvanus mitochondria-associated gene
     CYCS) (9CI)  (CA INDEX NAME)
OTHER NAMES:
CN   GenBank AAY17034
CN   GenBank AAY17034 (Translated from: GenBank AY918495)
SQL  105

SEQ    1 MGDVEKGKKI FIMKCSQCHT VEKGGKHKTG PNLHGLFGRK TGQAPGYSYT
                                              ====== =====
      51 AANKNKGITW GEDTLMEYLE NPKKYIPGTK MIFVGIKKKE ERADLIAYLK
     101 KATNE
HITS AT:   35-45
```

Protein sequences

Searching for functionally similar sequences

To search for functionally similar sequences, use the “family” search options:

- Family Exact Sequence Search (/SQEFP)
- Family Subsequence Search (/SQSFP)

In family searches, each common amino acid in the query has to match either the exact amino acid or a functionally similar “equivalent”.

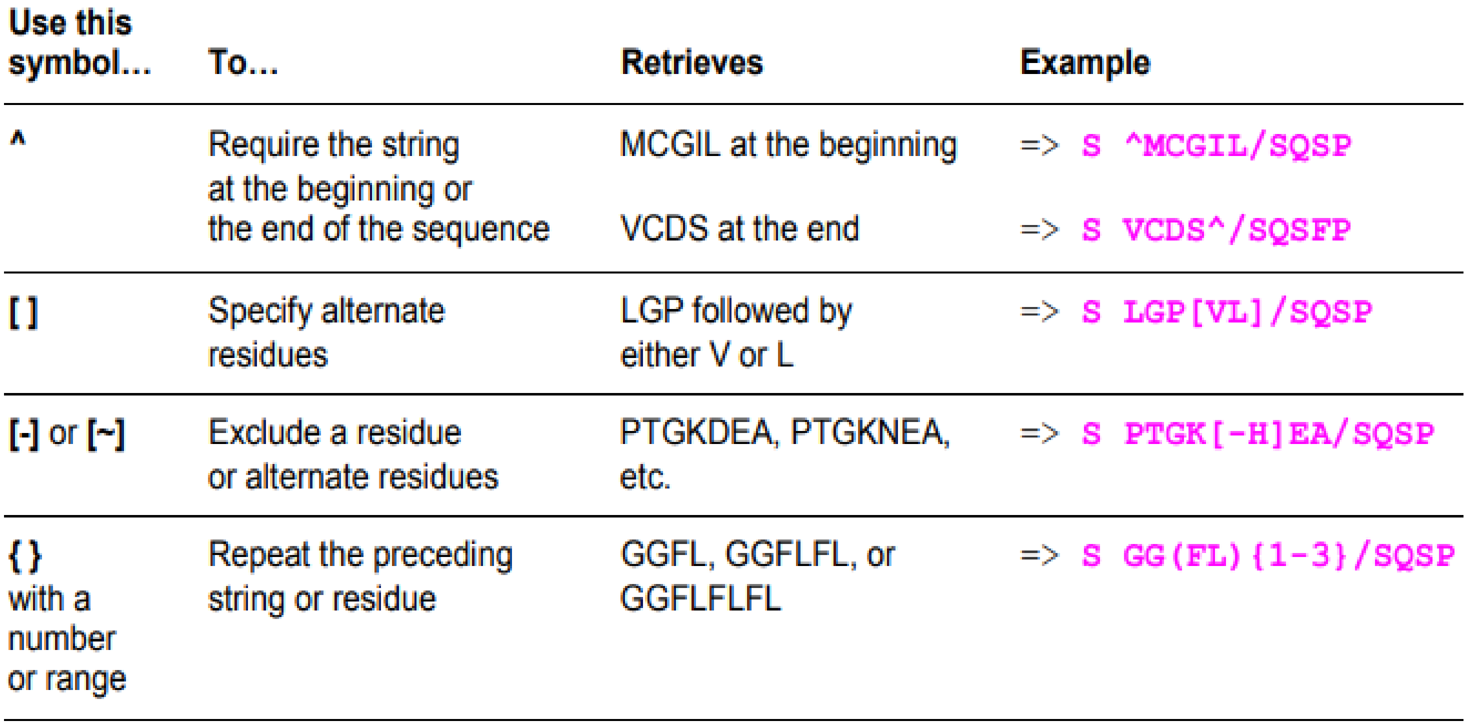
Possible family substitutions for KTDS_VCDS:

K	T	D	S	V	C	D	S
H	A	B	A	I	C	B	A
R	G	E	G	M		E	G
	P	N	P	L		N	P
	S	Q	T			Q	T

Property	Functionally Similar Amino Acids
Neutral-Weakly Hydrophobic	Ala,Gly,Pro,Ser,Thr (A, G, P, S, T)
Hydrophilic-Acid Amine	Asn,Asp,Gln,Glu (N, D, Q, E)
Hydrophilic-Basic	Arg,His,Lys (R, H, K)
Hydrophobic	Ile,Met,Leu,Val (I, M, L, V)
Hydrophobic-Aromatic	Phe,Trp,Tyr (F, W, Y)
Cross-linking	Cys (C)

Protein sequences

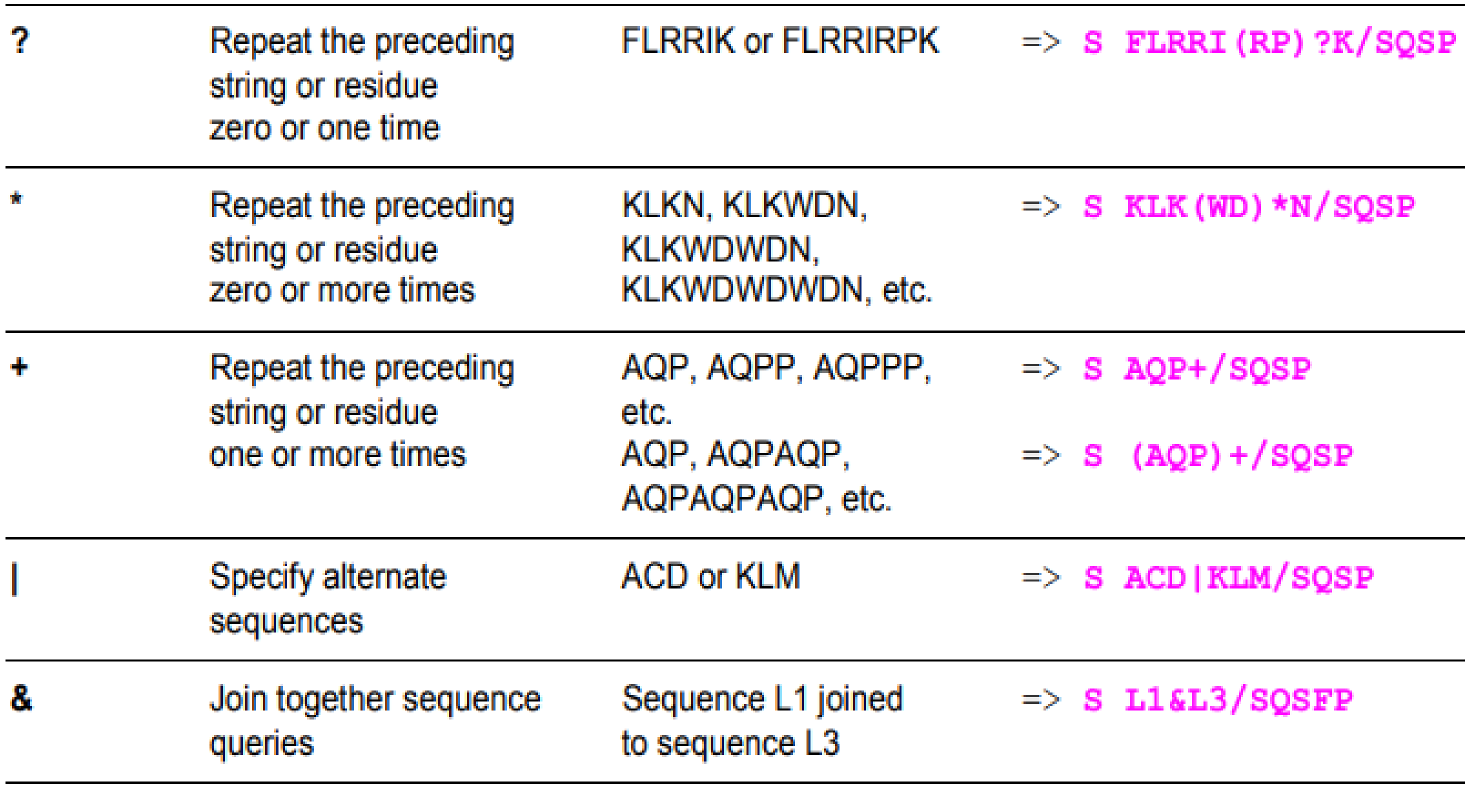
Searching motifs and patterns



Use this symbol…	To…	Retrieves	Example
^	Require the string at the beginning or the end of the sequence	MCGIL at the beginning VCDS at the end	=> S ^MCGIL/SQSP => S VCDS^/SQSFP
[ ]	Specify alternate residues	LGP followed by either V or L	=> S LGP[VL]/SQSP
[-] or [~]	Exclude a residue or alternate residues	PTGKDEA, PTGKNEA, etc.	=> S PTGK[-H]EA/SQSP
{ } with a number or range	Repeat the preceding string or residue	GGFL, GGFLFL, or GGFLFLFL	=> S GG(FL){1-3}/SQSP

Protein sequences

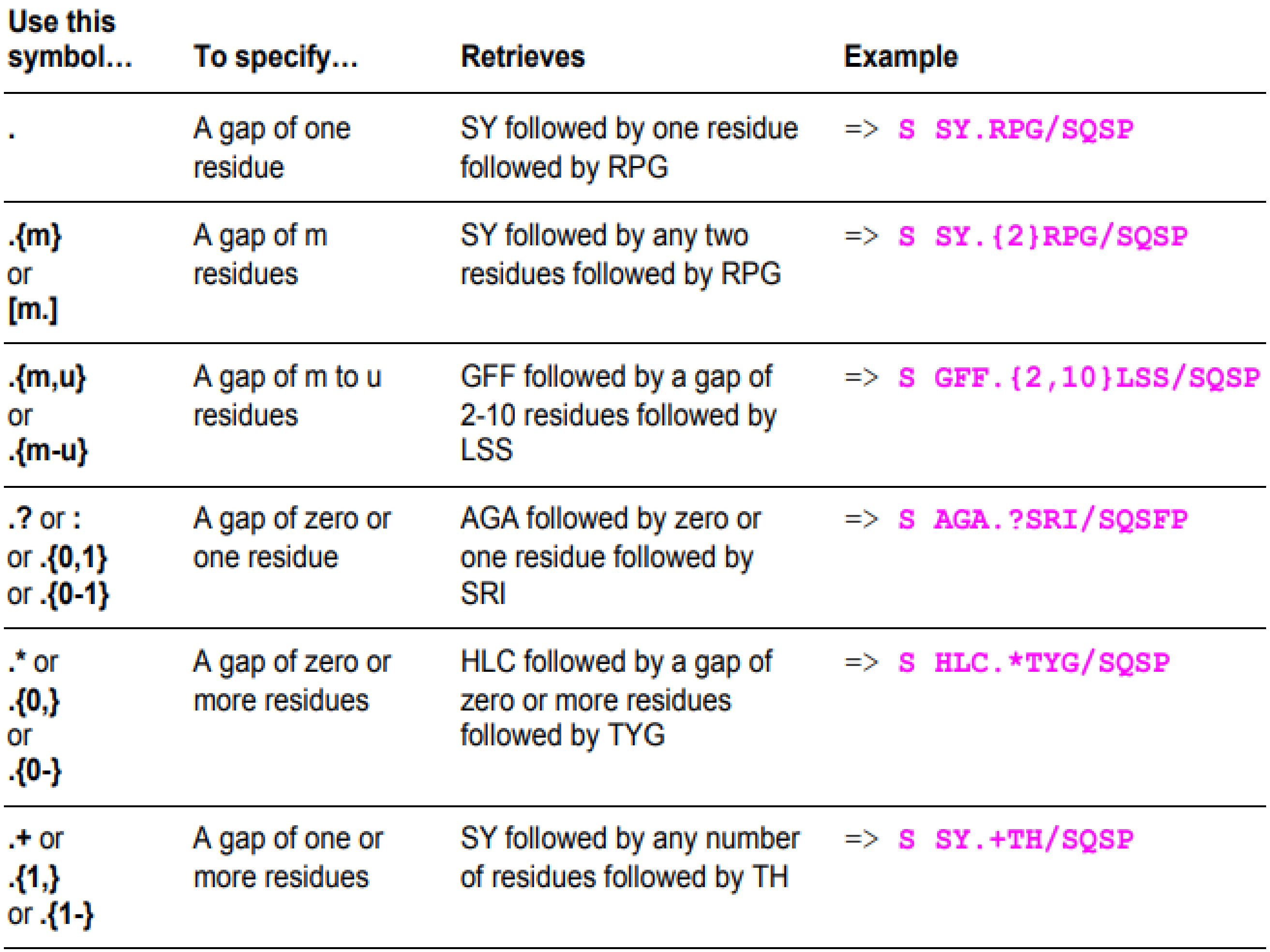
Searching motifs and patterns



?	Repeat the preceding string or residue zero or one time	FLRRIK or FLRRIRPK	=> S FLRRI(RP)?K/SQSP
*	Repeat the preceding string or residue zero or more times	KLKN, KLKWDN, KLKWDWDN, KLKWDWDWDN, etc.	=> S KLK(WD)*N/SQSP
+	Repeat the preceding string or residue one or more times	AQP, AQPP, AQPPP, etc. AQP, AQPAQP, AQPAQPAQP, etc.	=> S AQP+/SQSP => S (AQP)+/SQSP
|	Specify alternate sequences	ACD or KLM	=> S ACD|KLM/SQSP
&	Join together sequence queries	Sequence L1 joined to sequence L3	=> S L1&L3/SQSFP

Protein sequences

Gaps



Use this symbol...	To specify...	Retrieves	Example
.	A gap of one residue	SY followed by one residue followed by RPG	=> S SY.RPG/SQSP
.{m} or [m.]	A gap of m residues	SY followed by any two residues followed by RPG	=> S SY.{2}RPG/SQSP

.{m,u}
or
.{m-u}
	A gap of m to u residues	GFF followed by a gap of 2-10 residues followed by LSS	=> S GFF.{2,10}LSS/SQSP

.? or :
or .{0,1}
or .{0-1}
	A gap of zero or one residue	AGA followed by zero or one residue followed by SRI	=> S AGA.?SRI/SQSFP

.* or
.{0,}
or
.{0-}
	A gap of zero or more residues	HLC followed by a gap of zero or more residues followed by TYG	=> S HLC.*TYG/SQSP

.+ or
.{1,}
or .{1-}
	A gap of one or more residues	SY followed by any number of residues followed by TH	=> S SY.+TH/SQSP

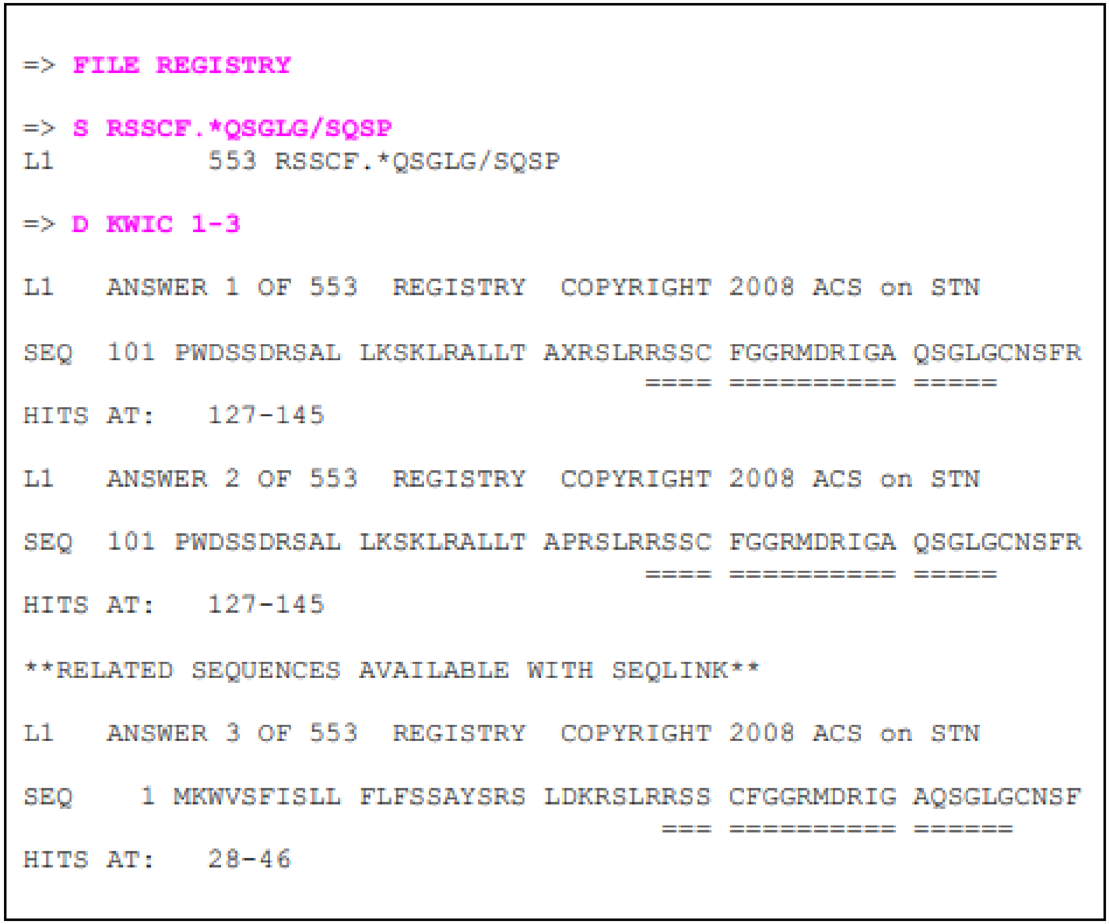
Protein sequences

Find atriopeptin analogs containing RSSCF and QSGLG, separated by a gap of zero or any number of amino acids.

1 *Enter REGISTRY.*

2 *Search the sequence pattern in the /SQSP field. The symbol .* indicates a gap of any number of amino acids, including zero.*

3 *Use the KWIC format to display the hit subsequence in context.*



```
=> FILE REGISTRY

=> S RSSCF.*QSGLG/SQSP
L1          553 RSSCF.*QSGLG/SQSP

=> D KWIC 1-3

L1   ANSWER 1 OF 553  REGISTRY  COPYRIGHT 2008 ACS on STN

SEQ  101 PWDSSDRSAL LKSKLRALLT AXRSLRRSSC FGGRMDRIGA QSGLGCNSFR
                                     ==== ========== =====
HITS AT:   127-145

L1   ANSWER 2 OF 553  REGISTRY  COPYRIGHT 2008 ACS on STN

SEQ  101 PWDSSDRSAL LKSKLRALLT APRSLRRSSC FGGRMDRIGA QSGLGCNSFR
                                     ==== ========== =====
HITS AT:   127-145

**RELATED SEQUENCES AVAILABLE WITH SEQLINK**

L1   ANSWER 3 OF 553  REGISTRY  COPYRIGHT 2008 ACS on STN

SEQ    1 MKWVSFISLL FLFSSAYSRS LDKRSLRRSS CFGGRMDRIG AQSGLGCNSF
                                      === ========== ======
HITS AT:   28-46
```

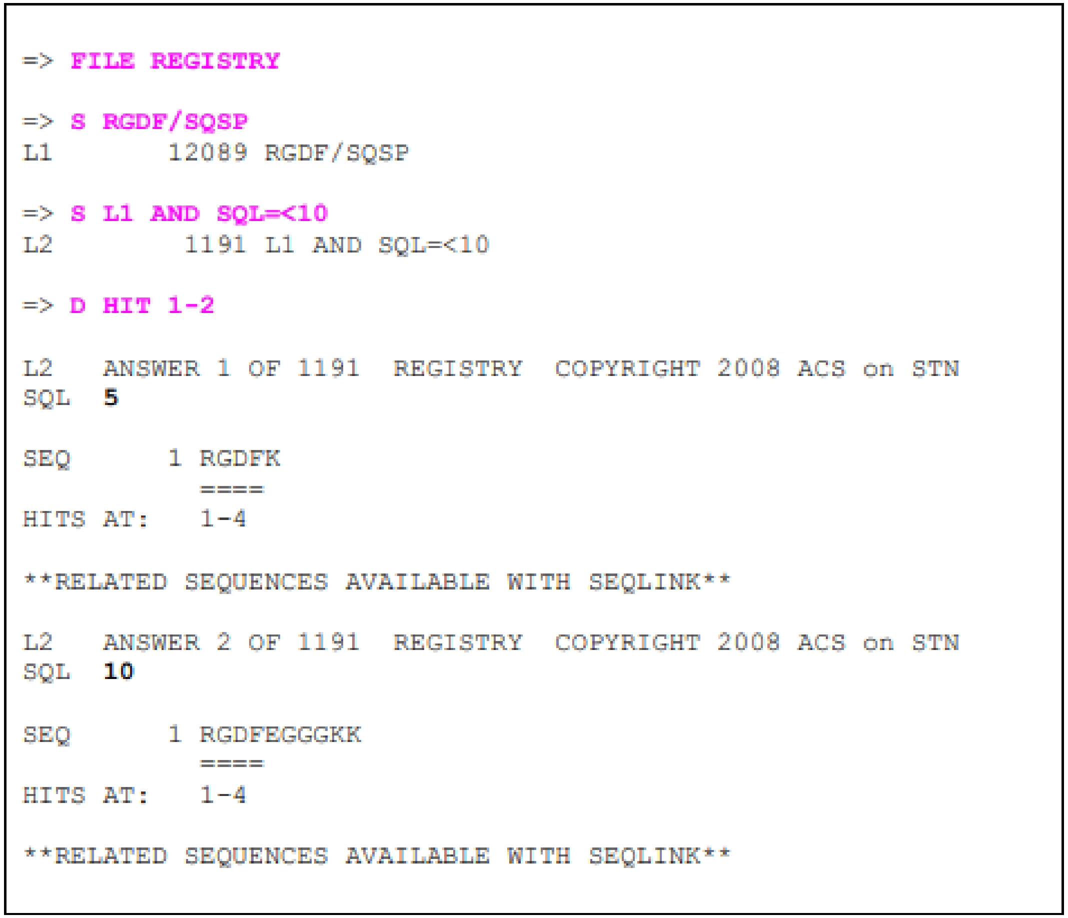
KIPO
KOREAN INTELLECTUAL PROPERTY OFFICE

Protein sequences

1 *Enter REGISTRY and search the sequence.*

2 *Search SQL<=10 to retrieve only sequences with 10 or fewer residues.*

Find RGDF containing peptides with 10 or fewer amino acids.



```
=> FILE REGISTRY

=> S RGDF/SQSP
L1         12089 RGDF/SQSP

=> S L1 AND SQL=<10
L2          1191 L1 AND SQL=<10

=> D HIT 1-2

L2   ANSWER 1 OF 1191  REGISTRY  COPYRIGHT 2008 ACS on STN
SQL  5

SEQ        1 RGDFK
             ====
HITS AT:     1-4

**RELATED SEQUENCES AVAILABLE WITH SEQLINK**

L2   ANSWER 2 OF 1191  REGISTRY  COPYRIGHT 2008 ACS on STN
SQL  10

SEQ        1 RGDFEGGGKK
             ====
HITS AT:     1-4

**RELATED SEQUENCES AVAILABLE WITH SEQLINK**
```

STN Blast Search

STN BLAST Search

Two different procedures depending on database

- ❖ DGENE (Derwent Geneseq^{TM})
- ❖ USGENE
- ❖ PCTGEN

STNext®

- ❖ CAS REGISTRY^{SM}

CAS REGISTRY BLAST®

STNext®

STN BLAST Search

Two different procedures depending on database

- DGENE (Derwent Geneseq^{TM})
- USGENE
- PCTGEN

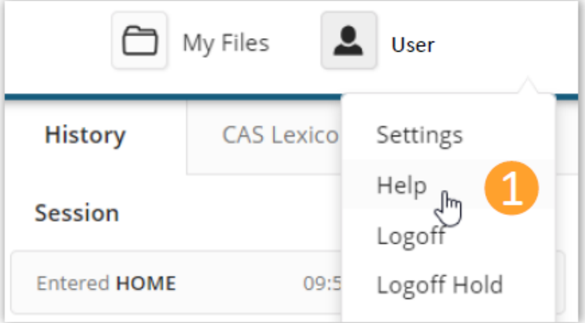
(1) Import sequence in STNext
(2) Validate sequence in Biosequence Editor
(3) Upload sequence in sequence-database
(4) Verify if uploaded sequence is correct
(5) Run BLAST search (and decide how many answers to keep)
(6) Review search (e.g. D TRIAL ALIGN)
(7) Run BLAST in other databases
(8) Merge answer sets
(9) Sort results (SCORE, IDENT)
(10) Display in STNext
(11) Report with STNext

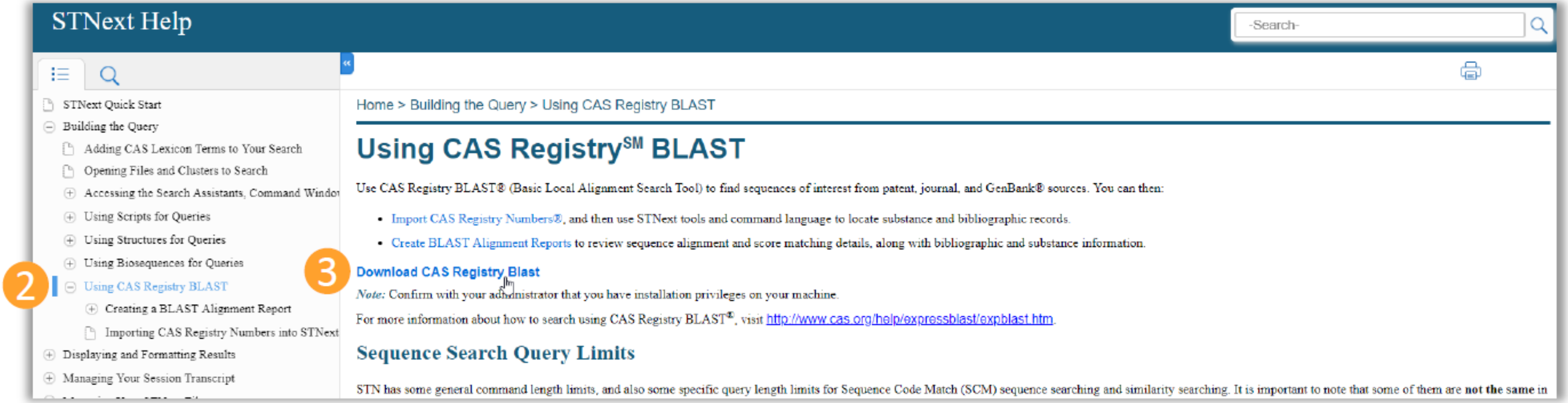
- CAS REGISTRY^{SM}

(1) Start CAS Registry BLAST client
(2) Start new search
(3) Paste sequence or load sequence file
(4) Select BLAST mode
(5) Adjust BLAST settings
(6) Review and select results
(7) Download two files: Script and alignment file
(8) From now STNext: Import script in STNext
(9) Start script and retrieve RNs
(10) Search for patents in CAplus
(11) Display in STNext
(12) Report with STNext (including alignment file)

CAS Registry BLAST Search

Install CAS Registry BLAST




STNext Help
-Search-
STNext Quick Start
Building the Query
Adding CAS Lexicon Terms to Your Search
Opening Files and Clusters to Search
Accessing the Search Assistants, Command Window
Using Scripts for Queries
Using Structures for Queries
Using Biosequences for Queries
Using CAS Registry BLAST
Creating a BLAST Alignment Report
Importing CAS Registry Numbers into STNext
Displaying and Formatting Results
Managing Your Session Transcript
2
3

Home > Building the Query > Using CAS Registry BLAST

Using CAS Registry^{SM} BLAST

Use CAS Registry BLAST® (Basic Local Alignment Search Tool) to find sequences of interest from patent, journal, and GenBank® sources. You can then:

- Import CAS Registry Numbers®, and then use STNext tools and command language to locate substance and bibliographic records.
- Create BLAST Alignment Reports to review sequence alignment and score matching details, along with bibliographic and substance information.

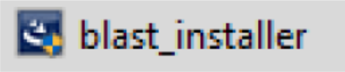
Download CAS Registry Blast

Note: Confirm with your administrator that you have installation privileges on your machine.

For more information about how to search using CAS Registry BLAST®, visit http://www.cas.org/help/expressblast/expblast.htm.

Sequence Search Query Limits

STN has some general command length limits, and also some specific query length limits for Sequence Code Match (SCM) sequence searching and similarity searching. It is important to note that some of them are **not the same** in

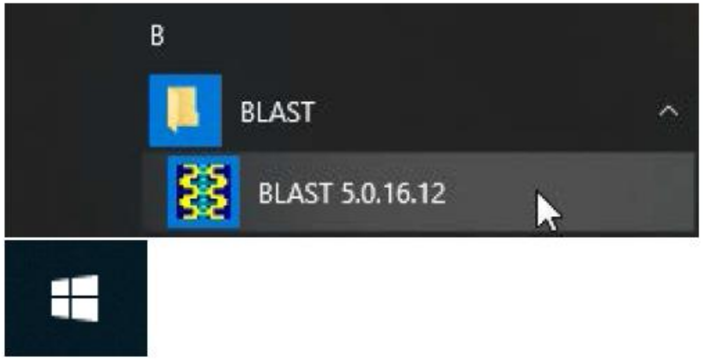
4 

Install CAS Registry BLAST

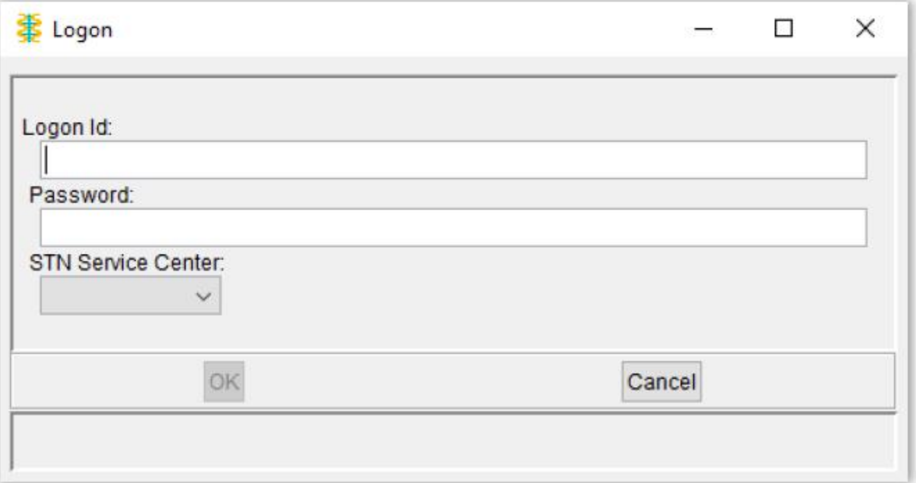
CAS Registry BLAST Search

Launch CAS Registry BLAST client

1



2

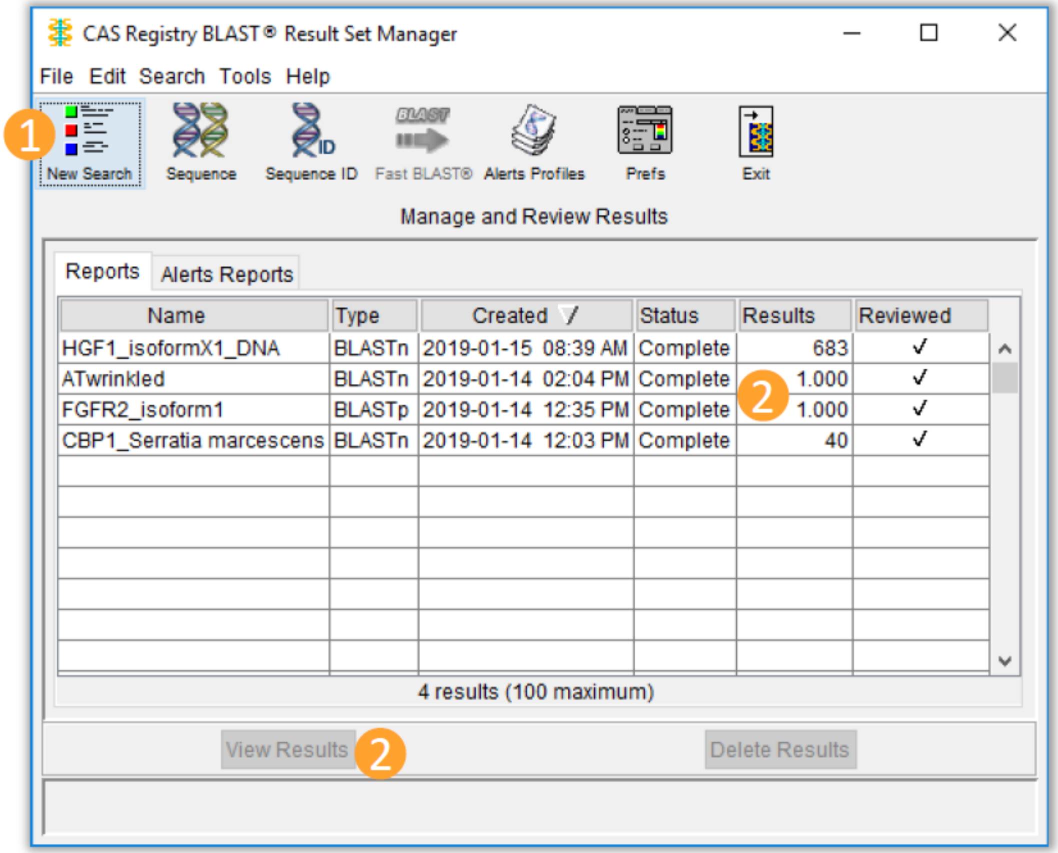

Logon

Logon Id:

Password:

STN Service Center:

CAS Registry BLAST Result Set Manager



Name	Type	Created	Status	Results	Reviewed
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	✓

4 results (100 maximum)

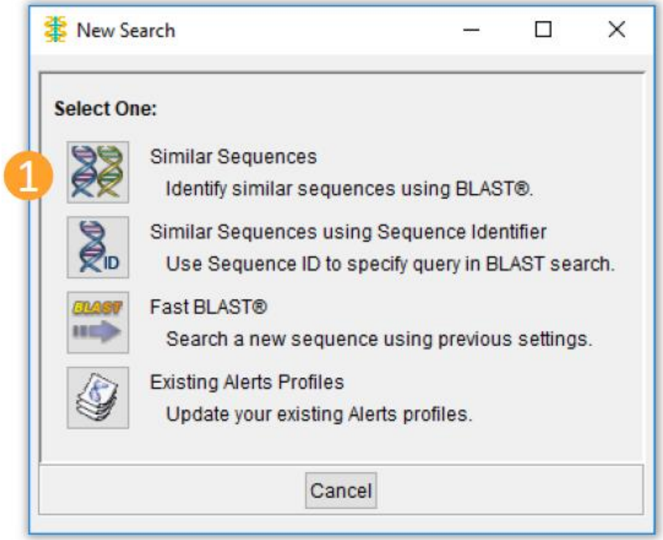
The Result Set Manager is the starting point:

1 to begin a new sequence search

2 to review results of previous sequence searches (double click or select)

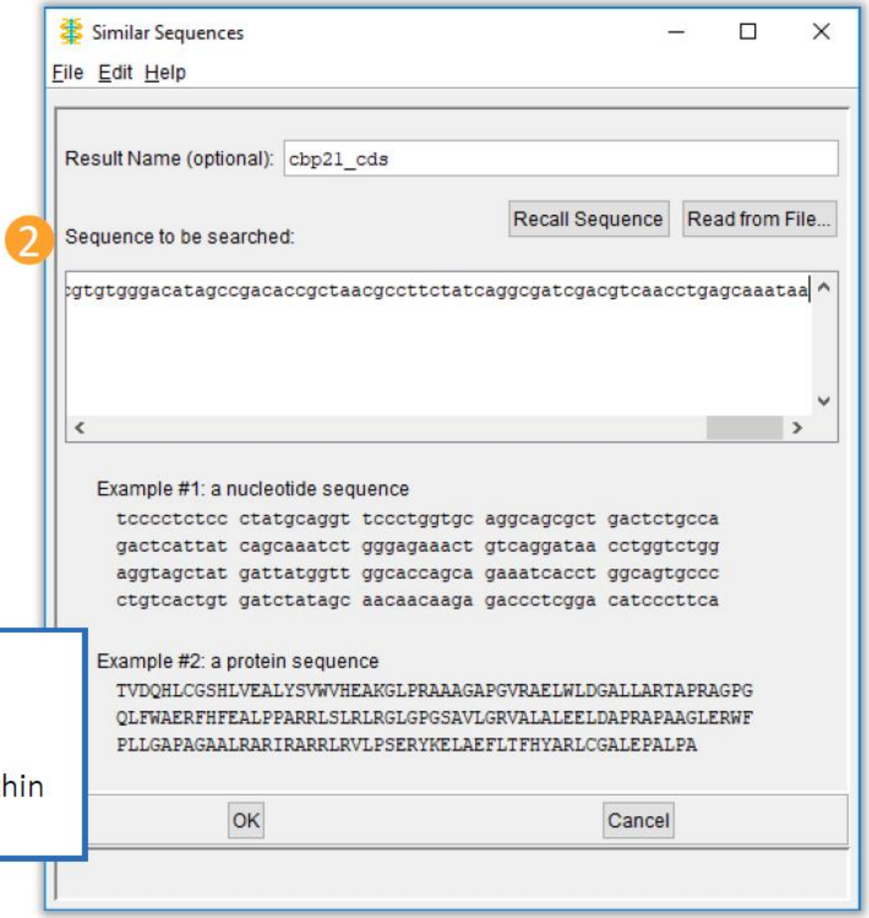
CAS Registry BLAST Search

Input the search query



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.



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

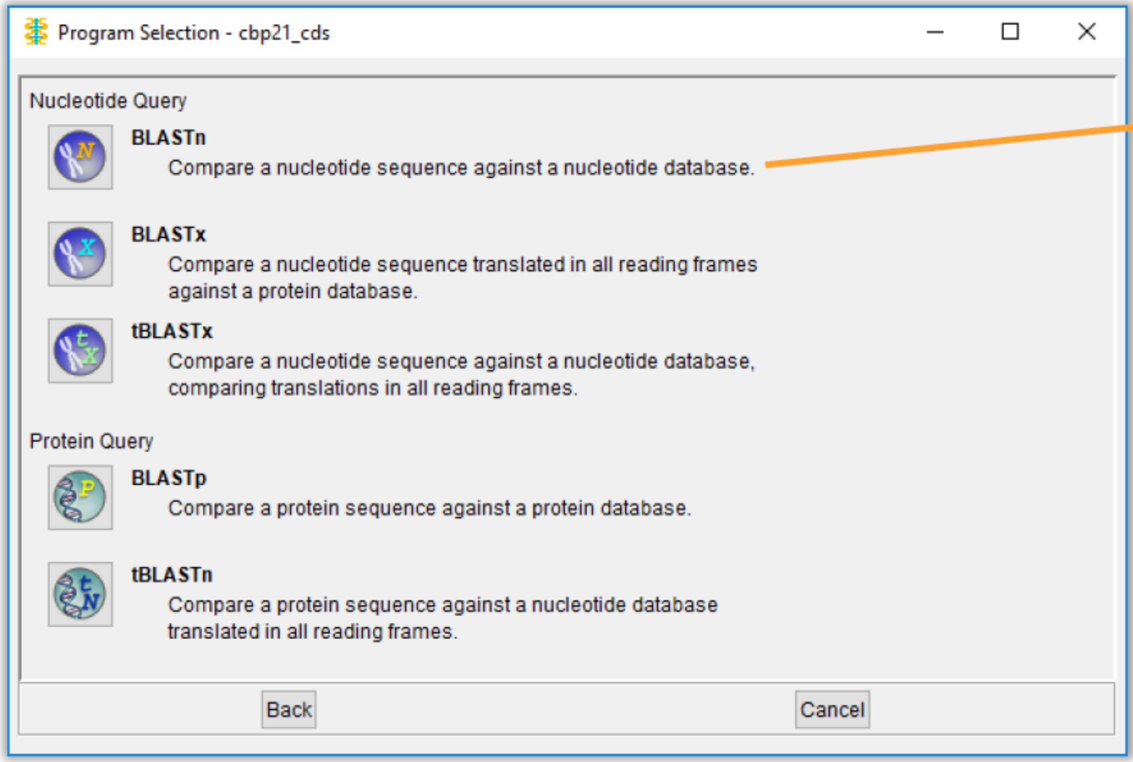
```
TVDQHLCGSHLVEALYSVWVHEAKGLPRAAAGAPGVRAELWLDGALLARTAPRAGPG
QLFWAERFHFEALPPARRLSLRLRGLGPGSAVLGRVALALEELDAPRAPAAGLERWF
PLLGAPAGAALRARIRARRLRVLPSERYKELAEFLTFHYARLCGALEPALPA
```

Sequences can be:

- Input by copy/paste
- Read from a file
- Recall from a previous search within the same session

CAS Registry BLAST Search

Start your BLAST search



Program Selection - cbp21_cds

Nucleotide Query

BLASTn
Compare a nucleotide sequence against a nucleotide database.

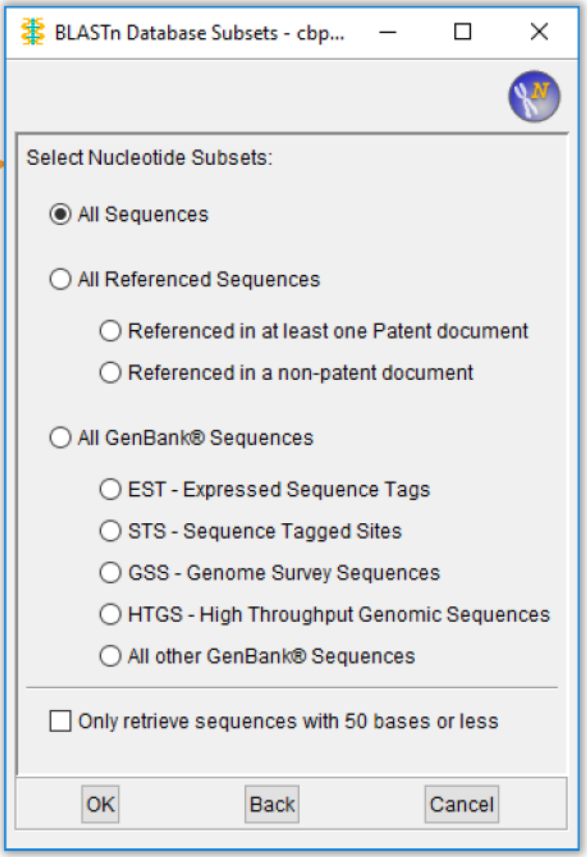
BLASTx
Compare a nucleotide sequence translated in all reading frames against a protein database.

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.



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

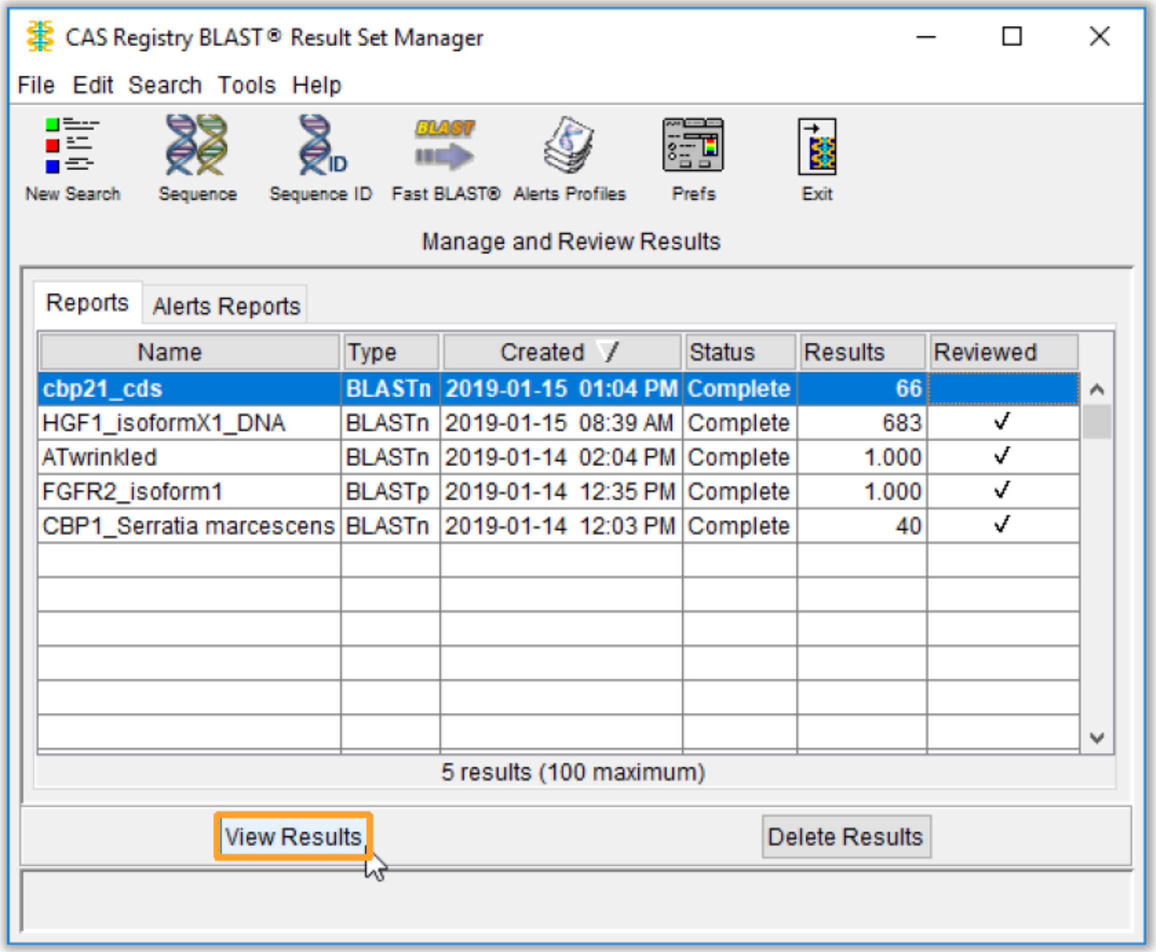
○ All GenBank® Sequences

- ○ 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

CAS Registry BLAST Search

View results of your completed search



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
Reports
Alerts Reports

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)

CAS Registry BLAST Search

Evaluate and select

Sequence Statistics

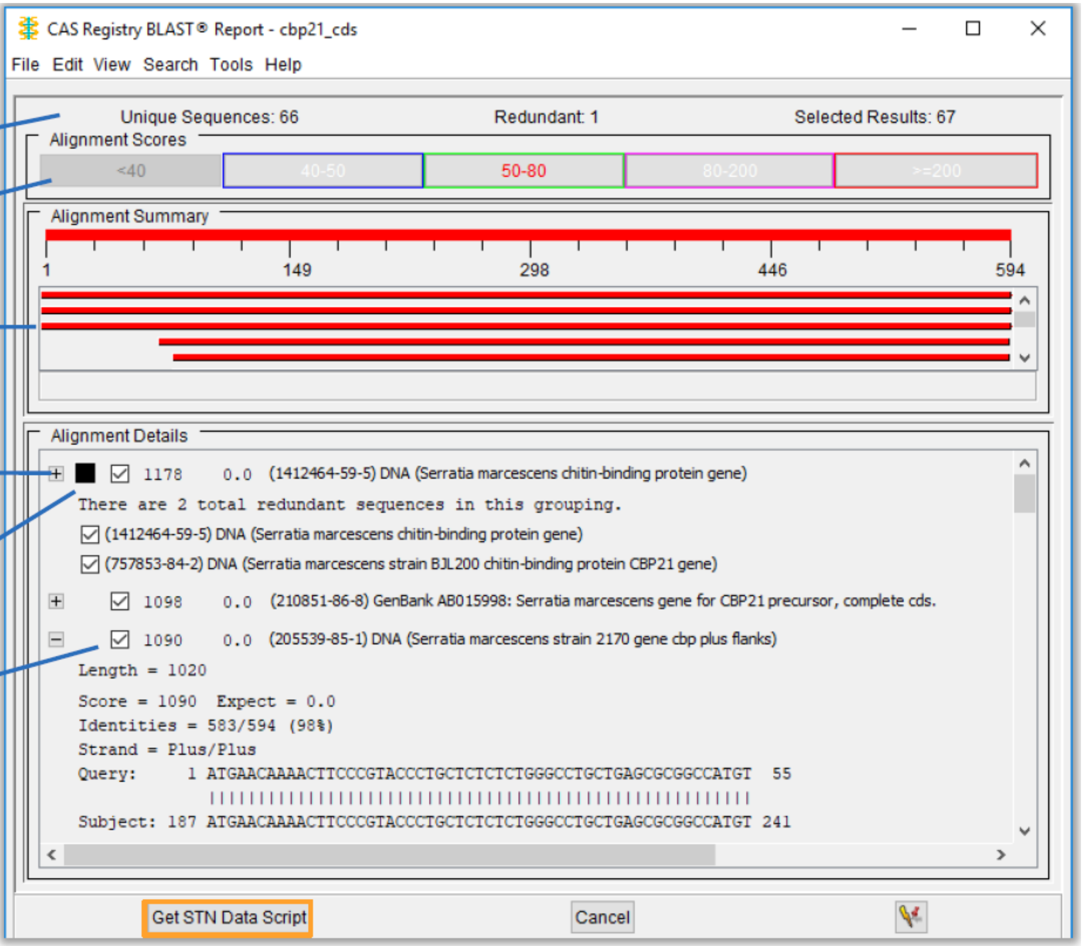
Alignment Score (click to select groups of sequences)

Alignment Summary for overview

Toggle +/- to show or hide the BLAST alignment details

Click on box to see “redundant” sequences (black = selected)

Select sequences of interest to retrieve STN data (or Alignment Score)



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
```

THANK YOU!