

STN[®]

Sequence Basics

Robert Austin – FIZ Karlsruhe


Agenda

- Sequence searchable databases on STN[®]
- BLAST in DGENE, USGENE[®] and PCTGEN
- CAS REGISTRYSM BLAST
- Sequence code match (motif) searching
- Recent enhancements

STN sequence searchable databases

- **DGENE**
 - Thomson Reuters GENESEQ™
 - Value-added patent sequence data from around the globe
- **USGENE**
 - The USPTO Genetic Sequence Database
 - All available sequence data from the USPTO
- **PCTGEN**
 - WIPO/PCT Patent Application Biosequences
 - All available e-published sequence data from WIPO
- **CAS REGISTRY**
 - Chemical Abstracts Service (CAS) REGISTRY
 - Worldwide value-added patent and non-patent sequences

DGENE, USGENE and PCTGEN all offer the same sequence search options

- BLAST similarity 
 - RUN BLAST
- Sequence Code Match (motif) searching
 - RUN GETSEQ
- FASTA similarity
 - RUN GETSIM

Note: this *Sequence Basics* e-Seminar covers RUN BLAST and RUN GETSEQ.

The 7 basic steps of RUN BLAST

- 1) SAVE, UPLOAD, and VERIFY the query (L1)
- 2) RUN the BLAST search (/SQP, /SQN, /TSQN)
- 3) Decide how many answers to keep (L2)
- 4) SORT SCORE in Descending order (L3)
- 5) Review answers in a free-of-charge format
e.g. D L3 TRIAL SCORE ALIGN 1-
- 6) Display selected answers in bibliographic
format, e.g. D L3 BIB ALIGN 1,3,10
- 7) Ensure transcript was captured before logoff

The 7 basic steps of RUN BLAST

Search Question:

Find relevant patent references for this protein sequence:

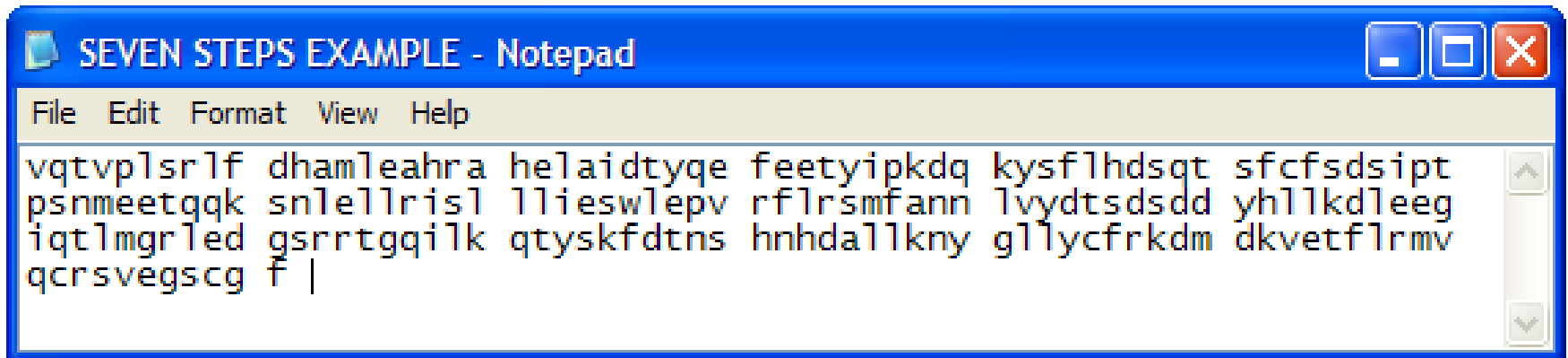
```
1  vqtvplsrlf  dhamleahra  helaidtyqe  feetyipkdq  kysflhdsqt
51  sfcfsdsipt  psnmeetqk  snlellrisl  llieswlep  rflrsmfann
101 lvydtsdsdd  yhllkdleeg  iqtlmgrled  gsrvtgqilk  qtyskfdtns
151 hnhdallkny  gllycfrkdm  dkvetflrmv  qcrsvegscg  f
```

See also: DGENE Workshop Manual:

http://www.stn-international.com/dgene_wm.html

1) SAVE, UPLOAD and VERIFY the query

- Prepare and save the query as a *plain text* file in a suitable text editor, e.g. Windows Notepad

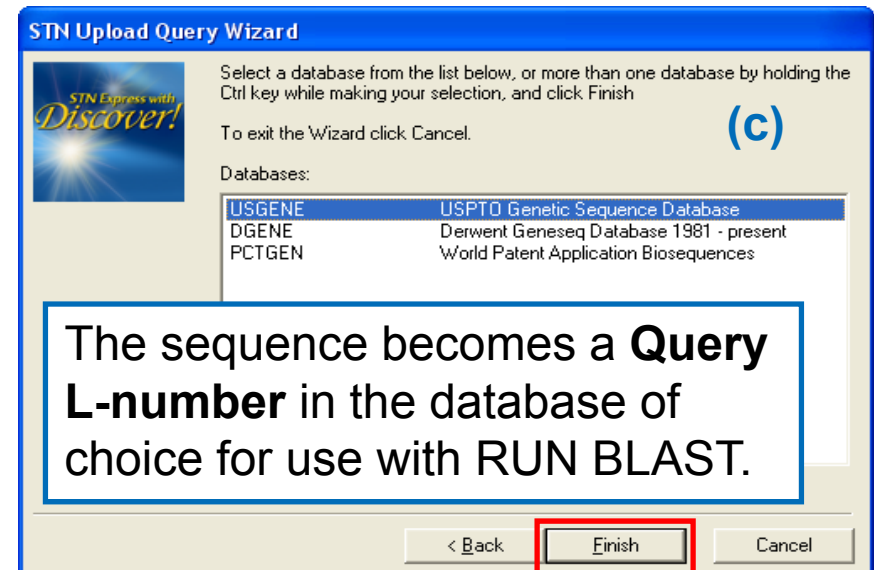
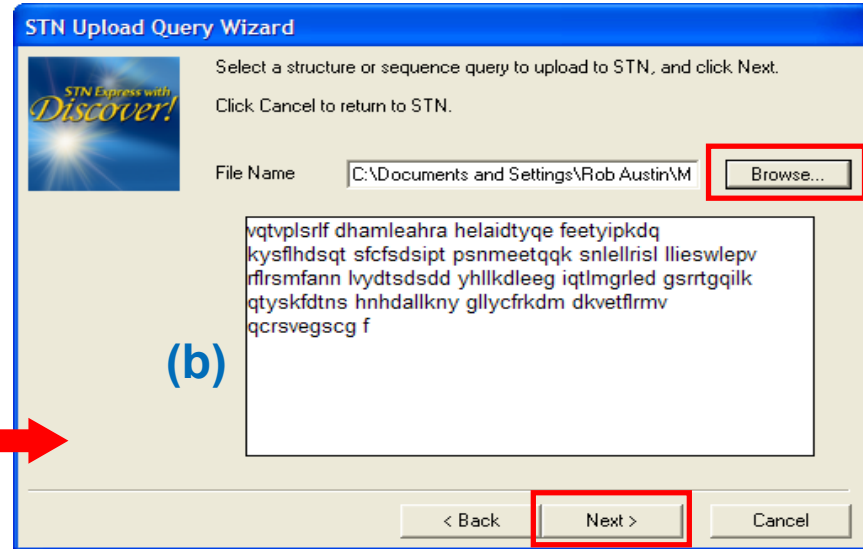
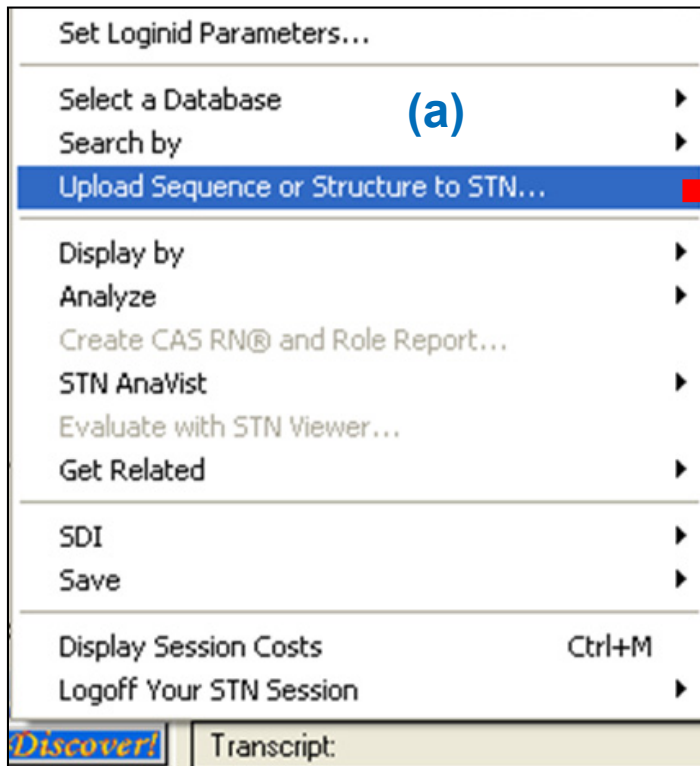


The screenshot shows a Notepad window with the title "SEVEN STEPS EXAMPLE - Notepad". The menu bar includes "File", "Edit", "Format", "View", and "Help". The text content is as follows:

```
vqtvplsr1f dhamleahra helaidtyqe feetyipkdq kysflhdsqt sfcfsdsipt  
psnmeetqk snlellrisl llieswlepv rflrsmfann lvydtsdsdd yhllkdleeg  
iqtlmgrled gsrrtgqilk qtyskfdtns hhhdallkny glycfrkdm dkvetflrmv  
qcrsvegscg f |
```

1) SAVE, UPLOAD and VERIFY the query (cont.)

- (a) Click **Upload Sequence**
- (b) Choose the query file
- (c) Select the STN database



From the *Discover!* button menu.

1) SAVE, UPLOAD and VERIFY the query (cont.)

=> **FILE USGENE**

Commands in **red** are automatically run by the STN Express Sequence Query Upload wizard.

=> **UPL R BLAST**

Uploading C:\Documents and Settings\...\SEVEN STEPS EXAMPLE.txt

UPLOAD SUCCESSFULLY COMPLETED

L1 GENERATED

Verify the sequence was uploaded successfully with **D LQUE**.

=> **D L1 LQUE**

L1 ANSWER 1 USGENE COPYRIGHT 2010 SEQUENCEBASE CORP on STN
LQUE vqtvplsrlfdhamleahrahelaidtyqefeetyipkdqkysflhdsqtsfcfsdsi
ptpsnmeetqqksnlellrislllieswlepvrflrsmfannlvdydtsdsddyhllkd
leegiqtlmgrledgsrrtgqilkqtyskfdtnshnhdallknygllycfrkdmdkve
tflrmvqcrsvegscgf

The sequence query is now ready for searching directly in DGENE, USGENE, or PCTGEN using the L-number (**L1**).

The 7 basic steps of RUN BLAST

2) RUN the BLAST search

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

2) RUN the BLAST search

=> FILE DGENE

FILE 'DGENE' ENTERED AT 17:41:21 ON 28 MAY 2010
COPYRIGHT (C) 2010 THOMSON REUTERS

FILE LAST UPDATED: 28 MAY 2010 <20100528/UP>

DGENE CURRENTLY CONTAINS 26,480,336 BIOSEQUENCES

>>> FOR THE LATEST DGENE STN USER DOCUMENTATION, PLEASE VISIT:

http://www.stn-international.com/stn_biosequence_searching_dgene.html

=> RUN BLAST L1 /SQP -F F

BLAST Version 2.2

The BLAST software is used herein with permission of the
National Center for Biotechnology Information (NCBI) of
the National Library of Medicine (NLM). See also,

BLAST SEARCHING

Turn the Low Complexity Filter
off with the syntax: /SQP -F F

RUN BLAST advanced options

Expectation Value (-E)

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

Word Size (-W)

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

RUN BLAST advanced options (cont.)

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

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

Peptide similarity matrices (-M)

For peptide based searches SQP and TSQN the advanced options provide additional scoring matrices to the default BLOSUM62 (next slide).

NCBI guidelines* for selecting the best peptide scoring matrix are as follows:

<u>Query Length</u>	<u>Matrix</u>	<u>Gap/extension costs</u>
<35	PAM-30	(9,1)
35 – 50	PAM-70	(10,1)
50 – 85	BLOSUM-80	(10,1)
>85	BLOSUM-62	(11,1) (BLAST default)

Tip: type **HELP OPTIONS** in DGENE for more information on using BLAST advanced options.

* http://www.ncbi.nlm.nih.gov/BLAST/matrix_info.html

The 7 basic steps of RUN BLAST

3) Decide how many answers to keep (L2)

- After the BLAST search, STN provides a chart summarizing the results, and asks this question:

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 nnn%)

ENTER (ALL) OR ? :

- General recommendation: Keep **ALL** answers, or use BATCH mode* to enable multiple retrievals

* See page 115-119: http://www.stn-international.com/usgene_wm.html

The 7 basic steps of RUN BLAST

4) SORT by SCORE descending (L3)

- Sort the BLAST results answer set:
=> **SOR L2 SCORE D**
- Option: limit using text terms and/or dates (L4)
- Remember to => **SORT L4 SCORE D !! (L5)**

3) Decide how many answers to keep

4002 ANSWERS FOUND BELOW EXPECTATION VALUE OF 10.0

QUERY SELF SCORE VALUE IS 390
BEST ANSWER SCORE VALUE IS 387

The Query Self Score is the ideal score for a perfect answer match.

The Best Answer Score is also given (in this example there isn't a perfect answer match in DGENE).

Similarity
Score

387

194

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

Answer Count

810

1620

2430

3240

4050

(Cont . . .)

4) SORT by SCORE descending

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 99%)

ENTER (ALL) OR ? :85%

In this example, 85% of the *Query Self Score* is used to select out just the most relevant results (L2).

L2 RUN STATEMENT CREATED

L2 640 VQTVPLSRLFDHAMLEAHRAHEL
SFCFSDSIPTSPNMEETQQKSNLELLRISLLLLIESWLEPVRFRLRSMFANN
LVYDTSDDYHLLKDLLEGIQTLMGRLEDGSRRTGQILKQTYSKFDTNS
HNHDALLKNYGLLYCFRKDMDKVETFLRMVQCRSVEGSCGF/SQP.-F F

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

=> **SOR SCORE D**

PROCESSING COMPLETED FOR L2

L3 640 SOR L2 SCORE D

Use **SORT SCORE D** to sort by descending BLAST score.

The 7 basic steps of RUN BLAST

- 5) Review answers using a *free-of-charge* format including alignment (ALIGN), while “parked” in the STNGUIDESM file
 - D L3 TRIAL SCORE ALIGN 1-
 - FILE STNGUIDE

Note: the SCORE display field also includes the percentage of the [Query Self Score](#) (maximum possible BLAST score).

5) Review answers with a free-of-charge format including alignment

=> D L3 TRIAL SCORE ALIGN 1-150; FILE STNGUIDE

L3 ANSWER 1 OF 640 DGENE COPYRIGHT 2010 THOMSON REUTERS on STN
AN AXQ10596 protein DGENE
TI Diagnosing active pre-eclampsia comprises testing in a maternal serum sample obtained from the pregnant female subject the level of proteins of interest relative to the level in normal maternal serum.
DESC Human Chorionic somatomammotropin hormone protein, SEQ ID 10.
KW Chorionic somatomammotropin hormone; Chorionic somatomammotropin ligand; cardiovascular-gen.; diagnostic test; genetic marker; gynecological; hypertension; hypotensive; immuno-diagnosis; immunoassay; pre-eclampsia; protein detection; screening.

SQL 217

SCORE 387

99% of query self score 390

BLASTALIGN

Query = 191 letters

Length = 217

Score = 387 bits (995), Expect

Identities = 189/191 (98%), Positives = 191/191 (99%)

Query: 1 VQTVPLSRLFDHAMLEAHRAHELAIPTYQEFEEETYIPKDQKYSFLHDSQTSFCFSDSIPT
VQTVPLSRLFDHAML+AHRAH+LAIDTYQEFEEETYIPKDQKYSFLHDSQTSFCFSDSIPT
Sbjct: 27 VQTVPLSRLFDHAMLQAHRAHQLAIDTYQEFEEETYIPKDQKYSFLHDSQTSFCFSDSIPT

. . . .

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

5) Review answers with a free-of-charge format including alignment (cont.)

```
L3 ANSWER 5 OF 640 DGENE COPYRIGHT 2010 THOMSON REUTERS on STN
AN AWC86667 protein DGENE
TI New stabilized prolactin receptor antagonist, comprising one or more
engineered linkers that are formed between two amino acid residues,
useful e.g. for the prophylaxis or treatment of cancers of e.g.
breast, prostate, cervix and lung.
DESC Cancer treatment related human placental lactogen (PL), SEQ ID 3.
KW protein engineering; therapeutic; prophylactic to disease;
hyperproliferation; neoplasm; cytostatic; cancer; Chorionic
somatomammotropin ligand; placental lactogen.
SQL 191
SCORE 387 99% of query self score 390
BLASTALIGN
  Query = 191 letters
  Length = 191
  Score = 387 bits (995), Expect = e-113
  Identities = 189/191 (98%), Positives = 191/191 (99%)
Query: 1 VQTVPLSRLFDHAMLEAHRAHELAIPTYQEFEEETYIPKDQKYSFLHDSQTSFCFSDSIPT
          VQTVPLSRLFDHAML+AHRAH+LAIDTYQEFEEETYIPKDQKYSFLHDSQTSFCFSDSIPT
Sbjct: 1 VQTVPLSRLFDHAMLQAHRAHQLAIDTYQEFEEETYIPKDQKYSFLHDSQTSFCFSDSIPT
Query: 61 PSNMEETQQKSNLELLRISLLLIESWLEPVRFLRSMFANNLVYDTSDDSDDYHLL . . .
          PSNMEETQQKSNLELLRISLLLIESWLEPVRFLRSMFANNLVYDTSDDSDDYHLL
Sbjct: 61 PSNMEETQQKSNLELLRISLLLIESWLEPVRFLRSMFANNLVYDTSDDSDDYHLL . . .
```

BLAST alignment
details are explained
on the next slide. . . .

Understanding BLAST alignments

Query	the length of the query sequence
Length	the length of the answer sequence
Score	a relative score assigned by BLAST
Expect	Expectation Value – a value representing the chance that an answer is a random hit. The closer to zero, the less likely the hit is random
Identities	the number of exact letter matches between query and answer within the displayed local alignment. The amino acid letter is repeated* in the display
Positives	a combination of identities and amino acid family matches shown with + (plus) in the alignment
Gaps	shown as dashes - where BLAST must break the query or answer to maintain an alignment

(* For nucleic acid searches a vertical bar is used to indicate nucleotide identities in the alignment display.)

Option: refine BLAST results with additional text and/or date search terms

```
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 99%)
ENTER (ALL) OR ? :85%
```

In this example, 85% of the **Query Self Score** is used to select out just the most relevant results (**L2**).

```
L2 RUN STATEMENT CREATED
L2 640 VQTVPLSRLFDHAMLEAHRAHELAI
SFCFSDSIPTPSNMEETQQKSNLELLKISLILLIESWLEPVKFLKSMFANN
LVYDTSDDYHLLKDLLEGIQTLMGRLEDGSRRTGQILKQTYSKFDTNS
HNHDALLKNYGLLYCFRKDMDKVETFLRMVQCRSVEGSCGF/SQP.-F F
```

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

```
=> SOR SCORE D
```

```
PROCESSING COMPLETED FOR L2
L3 640 SOR L2 SCORE D
```

```
=> S L2 AND PRY<2001
```

```
L4 74 L2 AND PRY<2001
```

```
=> SOR SCORE D
```

```
PROCESSING COMPLETED FOR L4
L5 74 SOR L4 SCORE D
```

The BLAST search (**L2**) is further refined to sequences with a priority date earlier than 2001 (**L4**).

If you limit using text and/or date terms remember to **SORT SCORE D** again (**L5**).

The 7 basic steps of RUN BLAST

- 6) Display selected relevant answers in a bibliographic format including alignment
 - E.g. => **D L5 BIB SCORE ALIGN 1,3,10**

6) Display selected answers in a preferred bibliographic format

=> D BIB SCORE ALIGN 1,3

```
L5 ANSWER 1 OF 74 DGENE COPYRIGHT 2010 THOMSON REUTERS on STN
AN AAW92262 Protein DGENE
TI New anti-angiogenic peptides - comprise N-terminal fragments of
human placental lactogen, human growth hormone, growth hormone
variant or human prolactin
IN Martial J A; Struman I; Taylor R; Weiner R I
PA (REGC) UNIV CALIFORNIA.
PI WO 9851323 A1 19981119
AI WO 1998-US9691 19980512
PRAI US 1997-46394 19970513
PSL Example 3; Page 47
DT Patent
LA English
OS 1999-045192 [04]
CR N-PSDB: AAX01702
DESC Human anti-angiogenic peptide hPL Met-1Phe191.
SCORE 387 99% of query self score 390
BLASTALIGN
Query = 191 letters
Length = 192
Score = 387 bits (995), Expect = e-113
Identities = 189/191 (98%), Positives = 191/191 (99%)
. . . .
```

This sequence comes from a PCT (WO) published application, with priority date earlier than 2001.

6) Display selected answers in a preferred bibliographic format (cont.)

```
L5 ANSWER 3 OF 74 DGENE COPYRIGHT 2010 THOMSON REUTERS on STN
AN ABP43620 Protein DGENE
TI New polypeptides and their encoded proteins, useful as nutritional
sources or supplements, or in gene therapy, particularly for
treating wounds, Alzheimer's disease, amyotrophic lateral sclerosis,
cancer or inflammation -
IN Tang Y T; Liu C; Zhou P; Asundi V; Zhang J; Zhao Q A; Ren F; Xue A
J; Yang Y; Wehrman T; Drmanac R T
PA (HYSE-N) HYSEQ INC.
PI WO 2002031111 A2 20020418
AI WO 2001-US27760 20011011
PRAI US 2000-687527 20001012
PSL Claim 20; SEQ ID # 523
DT Patent
LA English
OS 2002-426278 [45]
CR N-PSDB: ABQ60864
DESC Chronic somatostatin hormone 1 clone MGC:3714.
SCORE 385 98% of query self score 390
BLASTALIGN
Query = 191 letters
Length = 224
Score = 385 bits (989), Expect = e-112
Identities = 188/190 (98%), Positives = 190/190 (99%)
. . . .
```

This sequence also comes from a PCT (WO) published application, with priority date earlier than 2001.

7) Ensure your STN Express session transcript was captured and then logoff

The screenshot shows the STN Online and Results interface. A red circle highlights the 'Capture Session' icon in the toolbar. The 'Capture Session' dialog box is open, showing a file list with 'seven steps example.tm' selected. The 'Capture retrospectively' checkbox is checked and circled in red. The 'Select Discover! Wizard' window is also visible, showing a search history table.

Search history	
L1	6223 S BANANA
L2	640 RUN GETBATCH SEVEN
L3	640 SOR SCORE D
L4	74 S L2 AND PRY<2001
L5	74 SOR SCORE D

Note: if you wish to save everything done prior to choosing "Capture Session", click the "Capture retrospectively" box, before clicking the "Open" button.

The importance of using the correct BLAST advanced options

```
=> RUN BLAST GSSFLSPEHQR/SQP
```

```
. . . .
```

```
NO ANSWERS FOUND BELOW EXPECTATION VALUE OF 10.0
```

```
=> RUN BLAST GSSFLSPEHQR/SQP -M PAM30 -W 2 -E 20000 -F F
```

```
. . . .
```

```
6766 ANSWERS FOUND BELOW EXPECTATION VALUE OF 20000.0
```

```
QUERY SELF SCORE VALUE IS 38
```

```
BEST ANSWER SCORE VALUE IS 38
```

```
. . . .
```

```
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 ? :70%
```

```
L1 RUN STATEMENT CREATED
```

```
L1 1372 GSSFLSPEHQR/SQP.-M PAM30 -W 2 -E 20000 -F F
```

Changing BLAST options is especially important for short sequence queries.

In this example, 70% of the **Query Self Score** is used to select relevant results (L1).

The importance of using the correct BLAST advanced options (cont.)

=> **SOR L1 SCORE D**

```
PROCESSING COMPLETED FOR L1
L2          1372 SOR L1 SCORE D
```

Correct use of BLAST options
finds relevant sequence hits.

=> **D TRI SCORE ALIGN**

```
L2      ANSWER 1 OF 1372  DGENE  COPYRIGHT 2010 THOMSON REUTERS on STN
AN      AXX56579  peptide          DGENE
TI      New amide compounds are ghrelin O-acyltransferase inhibitors, useful
        for treating and preventing obesity, type II diabetes, membrane
        bound O-acyltransferase associated disease, and irritable bowel
        syndrome.
DESC    GOAT acyltransferase assay related ghrelin-27-biotin peptide sequence
KW      anorectic; antidiabetic; gastric motility disorder; . . . .
SQL     29
SCORE   38          100% of query self score 38
BLASTALIGN
  Query = 11 letters
  Length = 29
  Score = 37.5 bits (81), Expect = 2e-09
  Identities = 11/11 (100%), Positives = 11/11 (100%)
Query: 1  GSSFLSPEHQR 11
          GSSFLSPEHQR
Sbjct: 1  GSSFLSPEHQR 11
```

NCBI recommended settings* for searching small sequence queries

Peptide sequences

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

Nucleotide sequences

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

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

Review: 7 steps of RUN BLAST

- 1) SAVE, UPLOAD, and VERIFY the query (L1)
- 2) RUN the BLAST search (/SQP, /SQN, /TSQN)
- 3) Decide how many answers to keep (L2)
- 4) SORT SCORE in Descending order (L3)
- 5) Review answers in a free-of-charge format, e.g. D L3 TRIAL SCORE ALIGN 1-
- 6) Display selected answers in bibliographic format, e.g. D L3 BIB ALIGN 1,3,10
- 7) Ensure transcript was captured before logoff

CAS REGISTRY BLAST searching

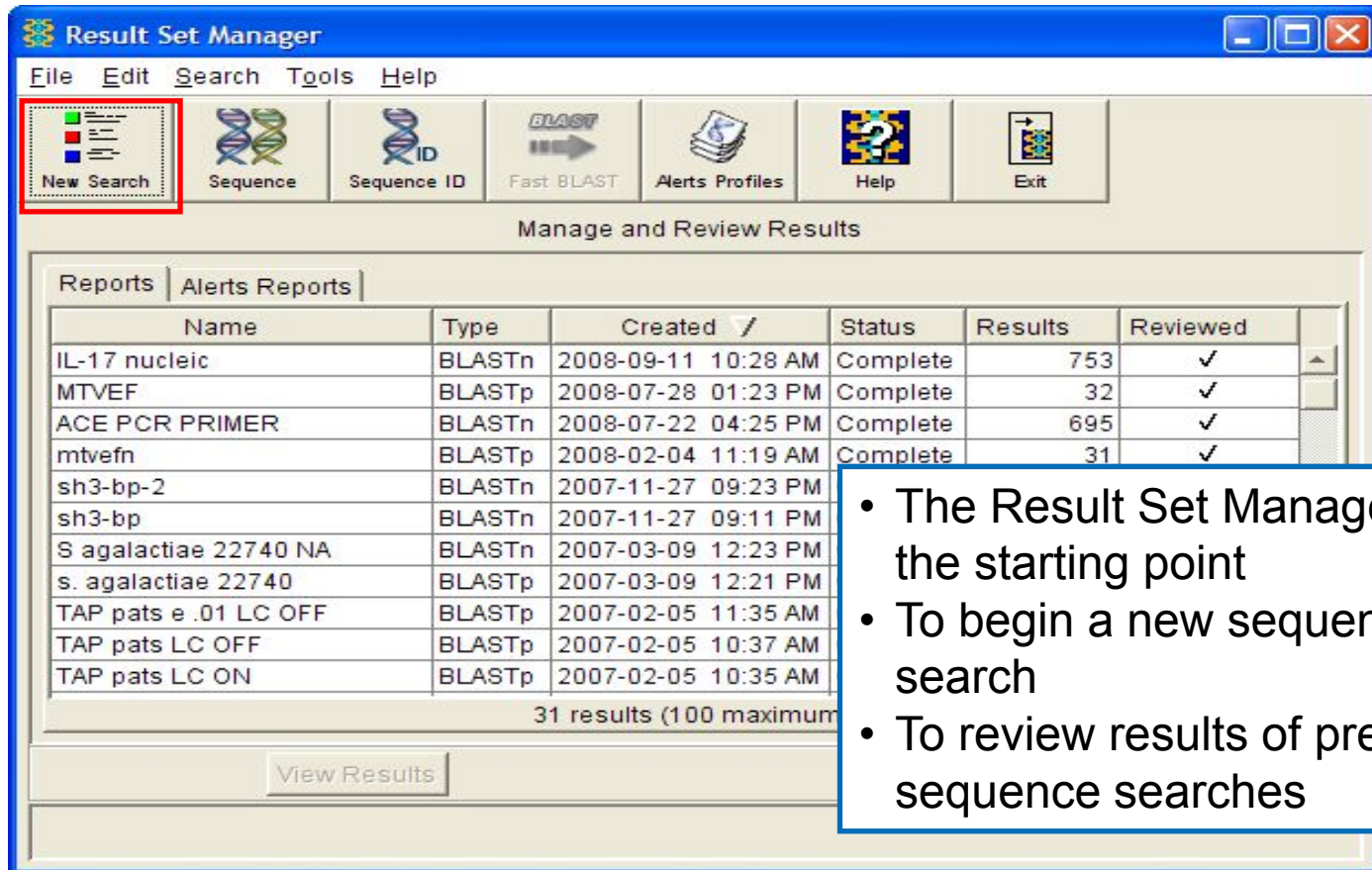
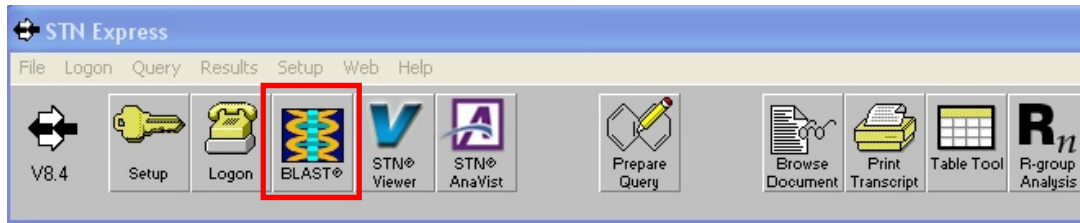
Search Question:

Locate references to Arginine Methyltransferase (RMT) protein sequence.

CAS REGISTRY BLAST search steps

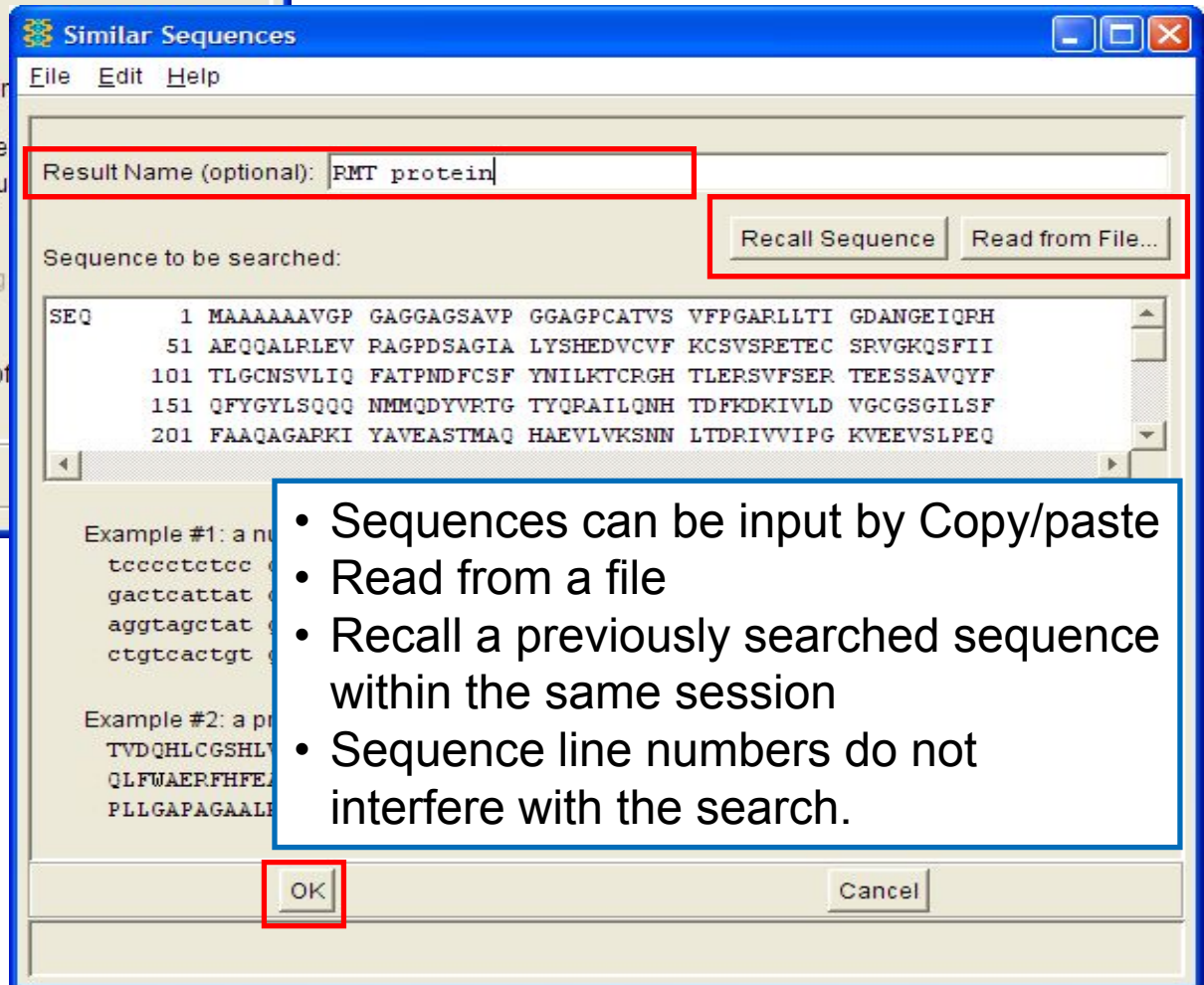
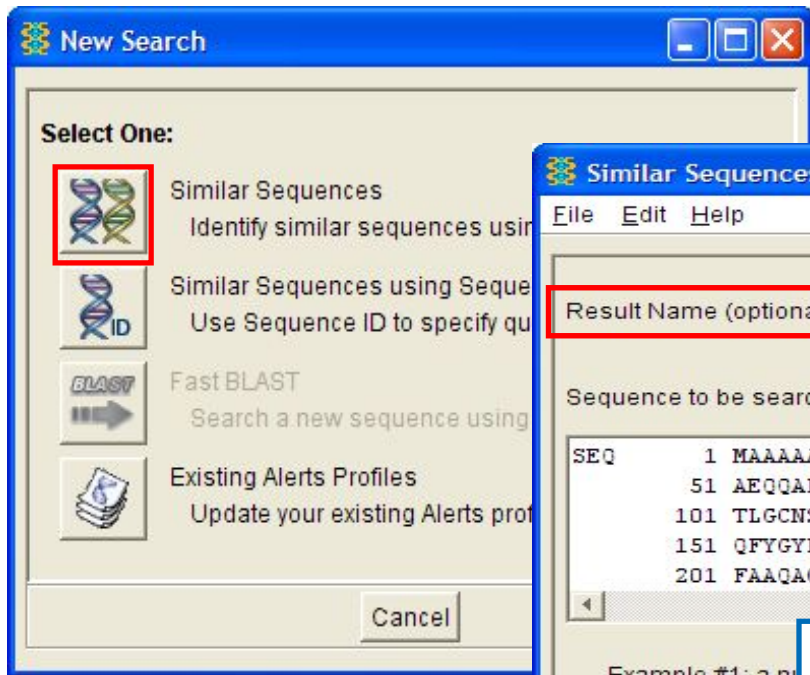
1. Launch BLAST
2. Search the sequence
3. Examine and evaluate alignment/relevance of sequence answers
4. Display STN data on sequences – REGISTRY
5. Display STN data on sequences – CAplusSM
 - Limit CAplus results, if necessary
 - Display CAplus data (references and HITRN)
6. Post-process BLAST alignment data

Launch CAS REGISTRY BLAST

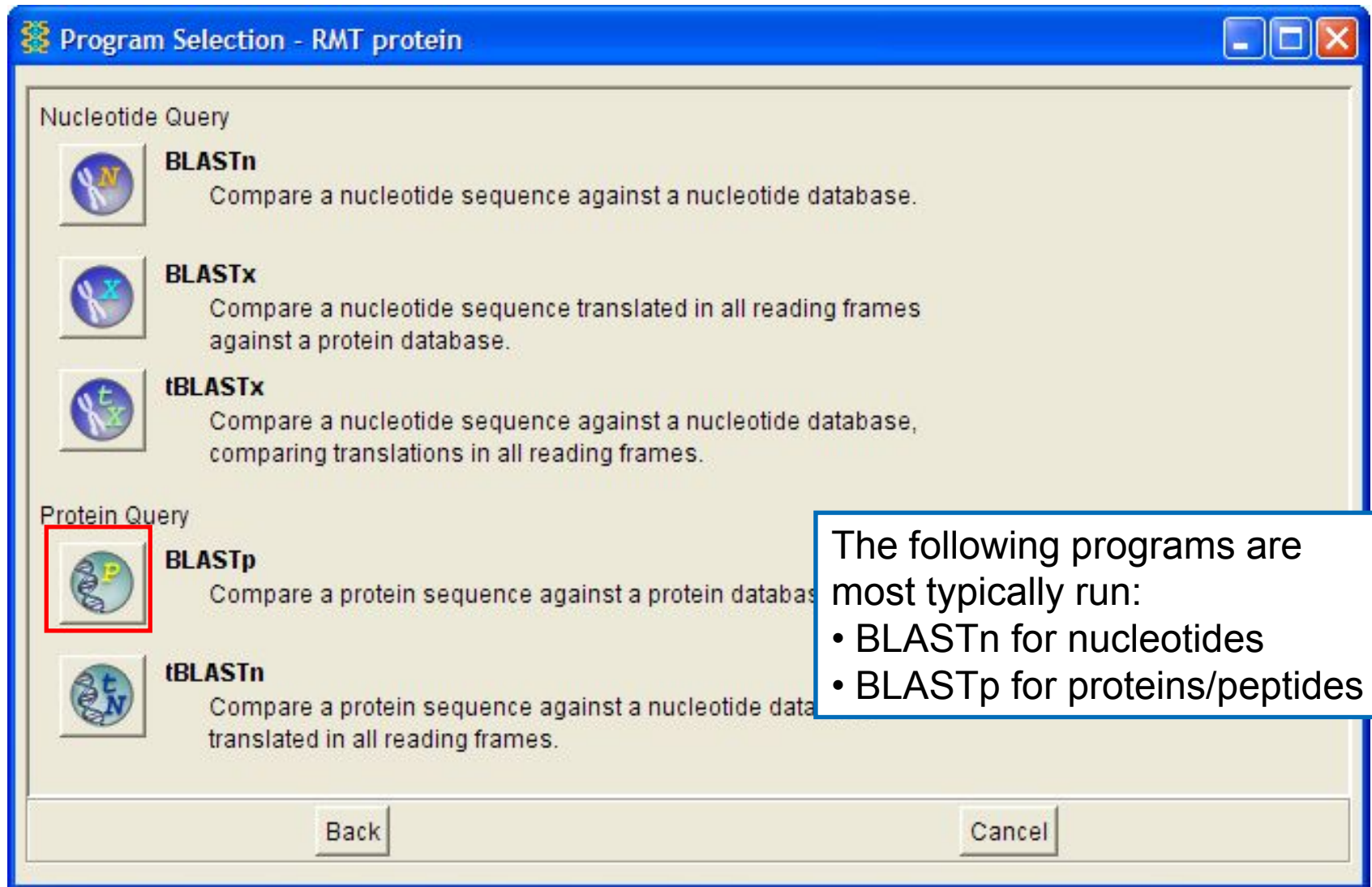


- The Result Set Manager is the starting point
- To begin a new sequence search
- To review results of previous sequence searches

Input the search query



Select the BLAST program



The following programs are most typically run:

- BLASTn for nucleotides
- BLASTp for proteins/peptides

Verify BLAST settings

BLASTp Settings - Additional Options - RMT protein

BLASTp Settings - Additional Options - RMT protein

Additional Option Presets

Search Sensitivity

Fewer Answers → More Answers

Show Additional Options

Basic Options

Low Complexity Filtering

Query Genetic Code: Standard(1)

Max No. of Answers: 1,000

Additional Options

Expectation Value: 10

Gap Cost: Open: 11 Extend: 1

Word Size: 3

Weight Matrix: BLOSUM-62

Penalty for Mismatch:

Reward for Match:

Reset to Defaults

OK Back Ca

Default values have been set to optimize sequence searches for researchers.

Recommended settings for patent searches:

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

View results

Result Set Manager

File Edit Search Tools Help

New Search Sequence Sequence ID Fast BLAST Alerts Profiles Help Exit

Manage and Review Results

Reports Alerts Reports

Name	Type	Created	Status	Results	Reviewed
RMT protein	BLASTp	2008-12-19 04:01 PM	Complete	809	✓
IL-17 nucleic	BLASTn	2008-09-11 10:28 AM	Complete	753	✓
MTVEF	BLASTp	2008-07-28 01:23 PM	Complete	32	✓
ACE PCR PRIMER	BLASTn	2008-07-22 04:25 PM	Complete	695	✓
mtvefn	BLASTp	2008-02-04 11:19 AM	Complete	31	✓
sh3-bp-2	BLASTn	2007-11-27 09:23 PM	Complete	994	✓
sh3-bp	BLASTn	2007-11-27 09:11 PM	Complete	289	✓
S agalactiae 22740 NA	BLASTn	2007-03-09 12:23 PM	Complete	28	✓
s. agalactiae 22740	BLASTp	2007-03-09 12:21 PM	Complete	3	✓
TAP pats e .01 LC OFF	BLASTp	2007-02-05 11:35 AM	Complete	4	✓
TAP pats LC OFF	BLASTp	2007-02-05 10:37 AM	Complete	21	✓

32 results (100 maximum)

View Results Delete Results

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

Evaluate the alignment report

The screenshot displays the CAS Registry BLAST Report for the RMT protein. The interface includes a menu bar (File, Edit, View, Search, Tools, Help) and a status bar at the bottom indicating "Result complete." The main content area is divided into several sections:

- Alignment Scores:** A horizontal bar chart showing score ranges: <40 (black), 40-50 (blue), 50-80 (green), 80-200 (magenta), and >=200 (red).
- Alignment Summary:** A horizontal bar chart showing sequence positions from 1 to 608. A red bar indicates the full length of the sequence.
- Alignment Details:** A section showing the alignment of the query sequence (1225) with the subject sequence (0.0). The details are expanded, showing the following information:
 - Length = 608
 - Score = 1225 Expect = 0.0
 - Identities = 608/608 (100%) Positives = 608/608 (100%)

The alignment details section also displays the query and subject sequences in a tabular format, showing the alignment of the query sequence (1225) with the subject sequence (0.0). The sequences are shown in a tabular format with positions and scores.

Query: 1 MAAAAA AVGP GAGGAGS AVPGGAGPCATVSVFPGARLLTIGDANGEIQRHAEQQA 55
MAAAAA AVGP GAGGAGS AVPGGAGPCATVSVFPGARLLTIGDANGEIQRHAEQQA
Subject: 1 MAAAAA AVGP GAGGAGS AVPGGAGPCATVSVFPGARLLTIGDANGEIQRHAEQQA 55

Query: 56 LRLEVRAGPDSAGIALYSHEDVCFKCSVSRETECSRVGKQSFIIITLGCNSVLIQ 1
LRLEVRAGPDSAGIALYSHEDVCFKCSVSRETECSRVGKQSFIIITLGCNSVLIQ 1
Subject: 56 LRLEVRAGPDSAGIALYSHEDVCFKCSVSRETECSRVGKQSFIIITLGCNSVLIQ 1

Query: 111 FATPNDFCSFYNILKTCRGHTLERSVFSERTEESSAVQYFQFYGYLSQQQNMMD 1
FATPNDFCSFYNILKTCRGHTLERSVFSERTEESSAVQYFQFYGYLSQQQNMMD 1
Subject: 111 FATPNDFCSFYNILKTCRGHTLERSVFSERTEESSAVQYFQFYGYLSQQQNMMD 1

Query: 166 YVRTGTYQRAILQNHTDFKDKIVLDVCGSGILSFFAAQAGARKIYAVEASTMAQ 2
YVRTGTYQRAILQNHTDFKDKIVLDVCGSGILSFFAAQAGARKIYAVEASTMAQ 2
Subject: 166 YVRTGTYQRAILQNHTDFKDKIVLDVCGSGILSFFAAQAGARKIYAVEASTMAQ 2

Buttons: Get STN Data, Cancel

The negative sign represents that the alignment details are shown. Detail information such as the sequence length, score, percent identity are available.

Select sequences of interest

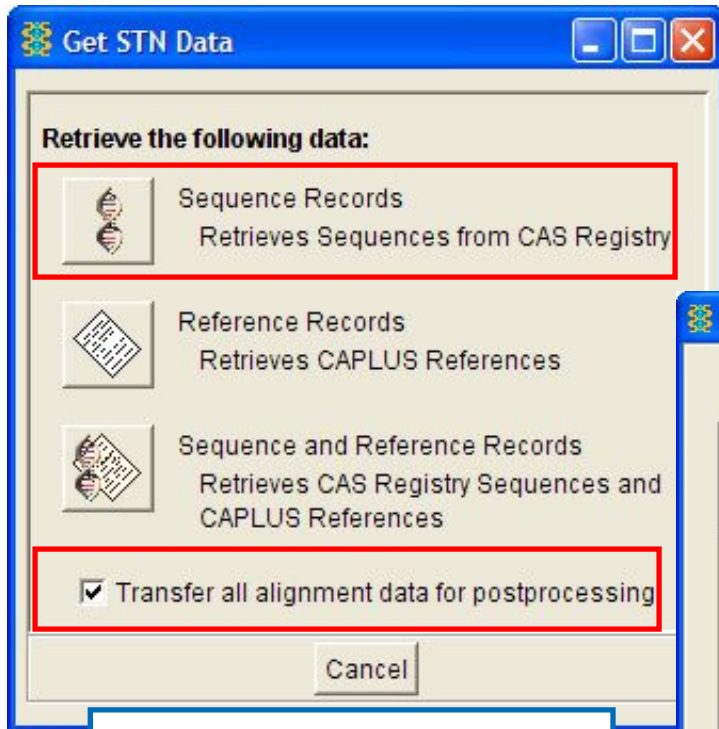
The screenshot displays the CAS Registry BLAST Report for RMT protein. The interface includes a menu bar (File, Edit, View, Search, Tools, Help) and a status bar at the top showing 'Unique Sequences: 809', 'Redundant: 348', and 'Selected Results: 61'. The 'Alignment Scores' section features a color-coded bar with five segments: black (<40), blue (40-50), green (50-80), magenta (80-200), and red (>=200). The 'Alignment Summary' section shows a red bar representing the protein sequence from position 1 to 608. The 'Alignment Details' section is a table with columns for selection status, score, and sequence description. A red box highlights the selection checkboxes for several entries. At the bottom, a 'Get STN Data' button is also highlighted with a red box.

Selection	Score	Description
<input checked="" type="checkbox"/>	1225	(1072557-06-2) Methyltransferase, transcriptional coactivator protein (arginine) (human)
<input checked="" type="checkbox"/>	1223	(453617-62-4) Drug-metabolizing enzyme DME-7 (hu
<input checked="" type="checkbox"/>	1222	(642514-19-0) Protein 27420 (human)
<input checked="" type="checkbox"/>	1222	(434009-21-9) 5; PN: W00244358 FIGURE: 4A-4B u
<input checked="" type="checkbox"/>	1198	(942169-05-3) 69; PN: US20070141652 SEQID: 69 u
<input checked="" type="checkbox"/>	1192	(696682-62-9) Transcription factor SRC-2 (steroid r
<input checked="" type="checkbox"/>	1186	(696686-43-8) 3; PN: US6743614 SEQID: 3 unclai
<input checked="" type="checkbox"/>	1177	(863541-89-3) Methyltransferase, transcriptional co
<input checked="" type="checkbox"/>	1175	(867594-03-4) Protein (Mus musculus strain C57BL/6
<input checked="" type="checkbox"/>	1167	(631936-53-3) Methyltransferase, transcriptional co
<input checked="" type="checkbox"/>	1140	(863541-93-9) Methyltransferase, transcriptional coactivator protein (arginine) (Rattus norvegicus gene CARM1 isoenzyme CARM1-v4)

Sequences can be selected:

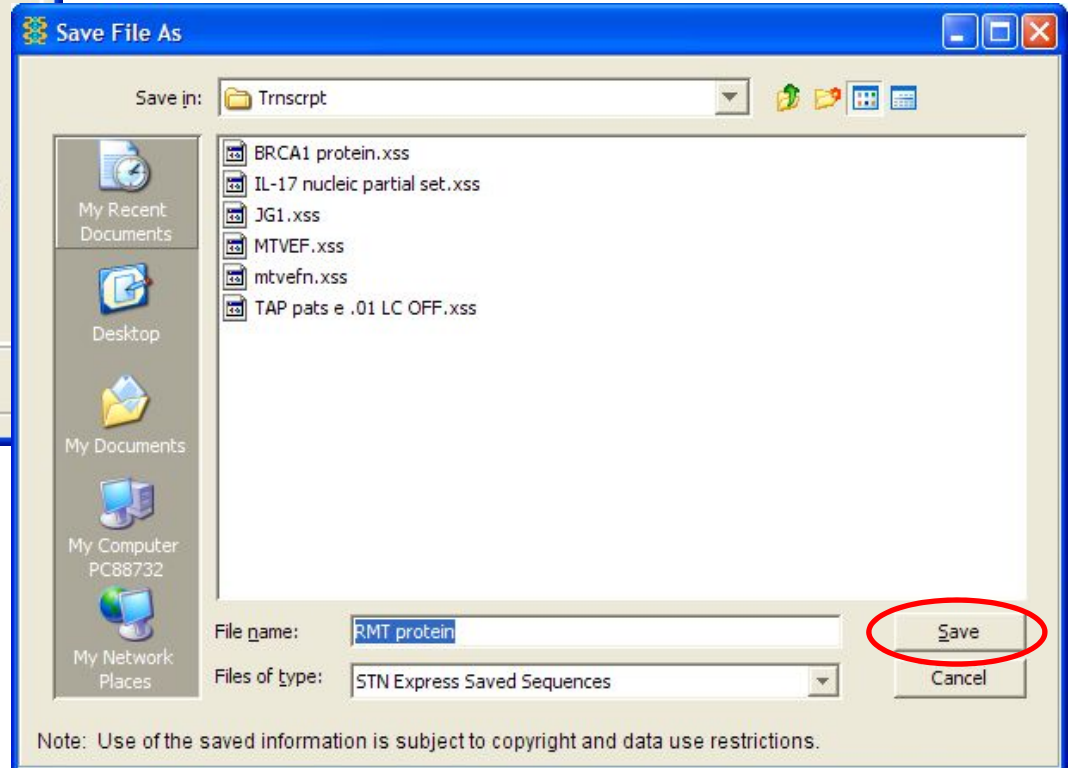
- In groups, using the color bar in the Alignment Scores
- Individually, by selecting the check box
- To transfer the sequence data to STN, click the Get STN Data button.

Get STN Data and Save alignments (.xss)



The alignment data is saved in STN Express Saved Sequences (.xss) format.

Alignment data needs to be transferred for post-processing.



Transfer sequences to STN

The screenshot shows the STN Online and Results - [STN/CAS] interface. The window title is "STN Online and Results - [STN/CAS]". The menu bar includes "File", "Edit", "Online", "Query", "Results", "Preferences!", "Web", "Window", and "Help". The toolbar contains various icons for file operations, search, and navigation. The main display area shows a list of sequence identifiers, each preceded by the number "1":

- 1 [587911-73-7/RN](#)
- 1 [816477-44-8/RN](#)
- 1 [734466-64-9/RN](#)
- 1 [921699-34-5/RN](#)
- 1 [665417-30-1/RN](#)
- 1 [487816-23-9/RN](#)
- 1 [623060-39-9/RN](#)
- 1 [778247-89-5/RN](#)
- 1 [482525-15-5/RN](#)
- 1 [666530-59-2/RN](#)
- 1 [622914-88-9/RN](#)
- 1 [486892-00-6/RN](#)
- 1 [487298-31-7/RN](#)
- 1 [913790-78-0/RN](#)
- 1 [481808-09-7/RN](#)
- 1 [736452-21-4/RN](#)
- 1 [809407-18-9/RN](#)
- 1 [809407-07-6/RN](#)

Below the list, the text "L6" is displayed on the left, and "61 L1 OR L2 OR L3 OR L4 OR L5" is displayed on the right. A red box highlights the command prompt "=> D L6 1 SQIDE". At the bottom of the window, there is a "Discover!" button and the text "Get additional data from STN".

- Logon to STN and a REGISTRY search of the sequences is automatic.
- Results display can be accomplished using either Discover! wizards or command line input.
- Note: Type END or click Cancel to get out of the "Display Wizard". You can turn off the "Display Wizard" in Preferences.

Display sequences if desired.

Crossover to CPlus

=> FILE CAPLUS

=> S L6 AND NONPATENT/DT

L7 14 L6 AND NONPATENT/DT

=> D L7 IBIB ABS HITRN 1-14

=> S L6 AND PATENT/DT

L8 20 L6 AND PATENT/DT

=> FSORT L8

L9 20 FSO L8

 3 Multi-record Families

 Family 1

 Family 2

 Family 3

 13 Individual Records

 0 Non-patent Records

=> D L9 IBIB ABS HITRN 1-20

Additional keyword refinement or other searches can be used in CPlus. In this example, patents and nonpatents were separated in 2 L-numbers.

Answers 1-7

Answers 1-3

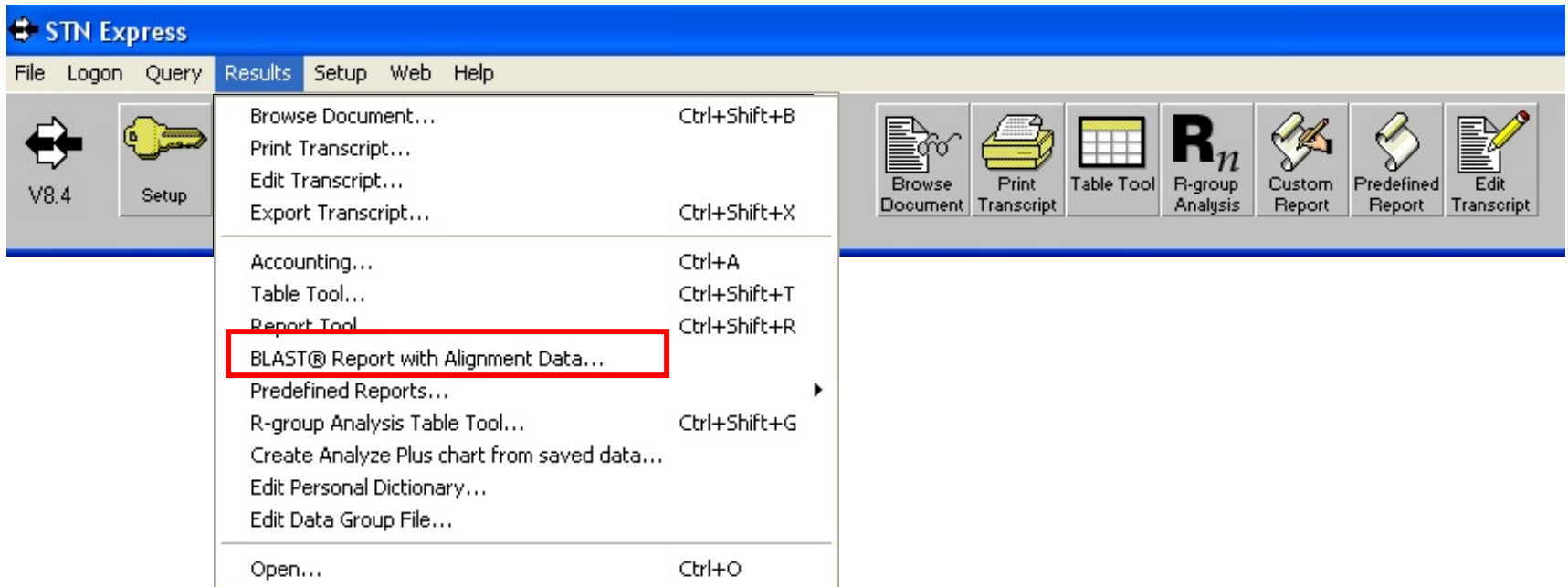
Answers 4-5

Answers 6-7

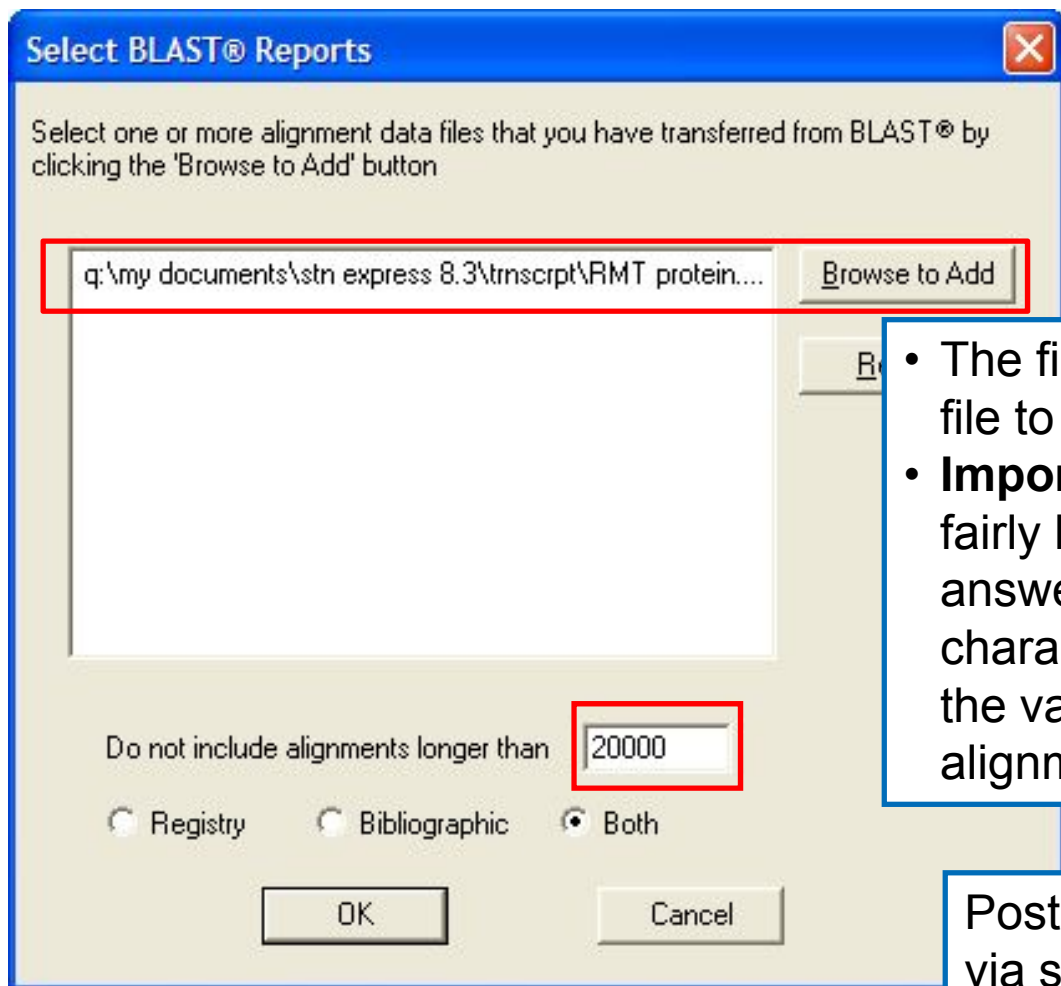
Answers 8-20

Consider SAVE or SAVE TEMP to keep your answer sets.

Post-process BLAST alignments



Select BLAST alignment reports



- The first step is to select the XSS file to include in the BLAST report.
- **Important:** If your BLAST query is fairly long, or a nucleic acid, or the answers may exceed 1000 characters, make sure you change the value in the Do not include alignments longer than box.

Post-processing then continues via standard STN Express *Custom Report Tool* steps.

Review – Search Steps

1. Launch BLAST
2. Search the sequence
3. Examine and evaluate alignment/relevance of sequence answers
4. Display STN data on sequences – REGISTRY
5. Display STN data on sequences – CAplus
 - Limit CAplus results, if necessary
 - Display CAplus data (references and HITRN)
6. Post-process BLAST alignment data

Sequence code match (motif) searching

- GETSEQ is designed to retrieve either exact matches to a sequence query or answers with conservative variation using special symbols
- It can also be used to retrieve exact length matches or subsequence hits, i.e. where the query is a small part of a larger hit sequence
- GETSEQ can prove to be a fast, precise and effective alternative to BLAST for very short sequence queries, e.g. DNA probes and primers
- A Sequence Code Match (SCM) search may be run in REGISTRY, but the SEARCH (=> S) command is used instead of RUN GETSEQ

The RUN GETSEQ command

=> **RUN GETSEQ L1** (sequence or query L-number)

/SQEP (**exact protein**) (**default**)

/SQEFP (exact family protein)

/SQSP (subsequence protein)

/SQSFP (subsequence family protein)

/SQEN (exact nucleotide)

/SQSN (subsequence nucleotide)

Note: an SCM search may also be run in REGISTRY, but the SEARCH (= > **S**) command is used instead of **RUN GETSEQ**.

EXACT (/SQEN) and SUBSEQUENCE (/SQSN) nucleic acid searching

```
=> RUN GETSEQ GCCGCCGT/SQEN
```

```
L1 RUN STATEMENT CREATED
```

```
L1 2 GCCGCCGT/SQEN
```

```
=> D L1 1 SEQ SQL
```

```
L1 ANSWER 1 OF 2 DGENE COPYRIGHT 2010
```

```
SEQ 1 gccgccgt
```

```
=====
```

```
HITS AT: 1-8
```

```
SQL 8
```

The SEQ display in DGENE shows the entire sequence with the hit nucleic acids underlined and identified by "HITS AT".

```
=> RUN GETSEQ ACCCTGCAAATAGCA/SQSN
```

```
L2 RUN STATEMENT CREATED
```

```
L2 49 ACCCTGCAAATAGCA/SQSN
```

```
=> D L2 30 SEQ SQL
```

```
L2 ANSWER 30 OF 49 DGENE COPYRIGHT 2010 THOMSON REUTERS on STN
```

```
SEQ 1 tgtagtcat tatcatcttt gtcatcagct gaagatgaaa taagatgtaa
```

```
51 tcagacgaca caggaagcag attctgctaa taccctgcaa atagcaga
```

```
=====
```

```
HITS AT: 82-96
```

```
SQL 98
```

A **SUBSEQUENCE** search also includes answers which are longer than the query sequence.

EXACT (/SQEP) and SUBSEQUENCE (/SQSP) protein searching

```
=> RUN GETSEQ SMAEP/SQEP
```

```
L3 RUN STATEMENT CREATED
```

```
L3 3 SMAEP/SQEP
```

```
=> D L3 1 SQL SEQ
```

```
L3 ANSWER 1 OF 3 DGENE COPYRIGHT 2010 THOMSON REUTERS on STN
```

```
SQL 5
```

```
SEQ 1 smaep
```

```
=====
```

```
HITS AT: 1-5
```

```
=> RUN GETSEQ KGPSYSLR/SQSP
```

```
L4 RUN STATEMENT CREATED
```

```
L4 102 KGPSYSLR/SQSP
```

```
=> D L4 11 SQL SEQ
```

```
L4 ANSWER 11 OF 102 DGENE COPYRIGHT 2010 THOMSON REUTERS on STN
```

```
SQL 19
```

```
SEQ 1 kgpsyslrst tmmirpldf
```

```
=====
```

```
HITS AT: 1-8
```

In all sequence databases, the typed order of the display fields will be the order that the fields are displayed.

A **SUBSEQUENCE** search also includes answers which are longer than the query sequence.

EXACT (/SQEFP) and SUBSEQUENCE (/SQSFP) FAMILY protein searching

```

=> RUN GETSEQ SMAEP/SQEFP
L5 RUN STATEMENT CREATED
L5 23 SMAEP/SQEFP
    
```

SMAEP/SQEP retrieved 3 records (L3).
SMAEP/SQEFP retrieved 23 records.

```

=> D L5 2-3 SQL SEQ
L5 ANSWER 2 OF 23 DGENE COPYRIGHT 2010 TH

SQL 5
SEQ 1 gites
=====
HITS AT: 1-5
    
```

Possible amino acid family substitutions for SMAEP:

S	M	A	E	P
P	I	G	Q	A
A	L	T	N	G
G	V	P	D	S
T		S	B	T

```

=> RUN GETSEQ KGPSYSLR/SQSFP
L6 RUN STATEMENT CREATED
L6 2384 KGPSYSLR/SQSFP
    
```

KGPSYSLR/SQSP retrieved 102 records (L4).
KGPSYSLR/SQSFP retrieved 2384 records.

```

=> D L6 73 SEQ SQL
L6 ANSWER 73 OF 2384 DGENE

SQL 43
SEQ 1 hfrgkfcgki apppvvssgp flfikfvscy ethgagfsir yei
=====
HITS AT: 33-40
    
```

Amino acid families for RUN GETSEQ SQEFP and QSFP search options

GROUP	AMINO ACIDS
Neutral-Weak Hydrophobics	P, A, G, S, T
Acid Amines-Hydrophilic	Q, N, E, D, B, Z
Basic-Hydrophilic	H, K, R
Hydrophobics	I, M, L, V
Aromatic	F, W, Y
Cross-Linking	C

Special variability symbols allow flexibility in RUN GETSEQ searching

- Variability symbols (pattern matching):
 - Allow users to specify motif patterns that consist of different amino acid(s) at one location of a sequence
 - Provide the ability to specify sequences separated by an unknown number of amino acids (gaps)
 - Provide the ability to search for sequence patterns at either beginning or the end of the sequence
 - Allow users to specify the number or range of repeats for amino acid(s) or gaps

Note: a complete table of all variability symbols, with search examples, is given in the DGENE, USGENE and PCTGEN database summary sheets:

http://www.stn-international.com/stndatabases/databases/onlin_db.html

Variability symbols for RUN GETSEQ

<u>Symbol</u>	<u>Function</u>
[]	Specify alternate residues
[-]	Exclude a specific residue or alternate residues
{ }	Repeat the preceding symbol(s) (number or range)
?	Repeat the preceding symbol(s) zero or one time
*	Repeat the preceding symbol(s) zero or more times
+	Repeat the preceding symbol(s) one or more times
^	Query appears at the beginning or the end of a sequence
	Alternate sequence expressions
.	A gap of one residue
:	A gap of zero or one residues
&	Concatenate (join together) sequence queries

Using RUN GETSEQ variability symbols to search in DGENE and REGISTRY

Search Question:

Find patent references disclosing one or more of the sequences represented by this Markush peptide sequence formula:

LGPX₁QLCX₂LVX₃CAP

X₁ = V or L

X₂ = any amino acid except, G or H

X₃ = any amino acid

RUN GETSEQ SCM search strategy

=> **RUN GETSEQ LGP[VL]QLC[-GH]LV.CAP/SQSP**

– Possible sequence retrieval

- *LGPVQLCALVHCAP*
- *LGPVQLCSLVVCAP*
- *LGPLQLCVLVACAP*
- *LGPLQLCPLVTCAP*

Reminder: an SCM search may also be run in REGISTRY, but the SEARCH (=> S) command is used instead of **RUN GETSEQ**.

Run the GETSEQ SCM search

=> FILE DGENE

=> RUN GETSEQ LGP[VL]QLC[-GH]LV.CAP/SQSP

L1 RUN STATEMENT CREATED
L1 51 LGP[VL]QLC[-GH]LV.CAP/SQSP

51 sequence hits (L1) have been found in DGENE containing the sequence fragment(s) of interest.

=> D TRI SEQ

L1 ANSWER 1 OF 51 DGENE COPYRIGHT 2010 THOMSON REUTERS ON SIN
AN AXF01515 protein DGENE
TI Use of a molecule capable of modifying a tissue level and/or activity of a type of lysyl oxidase for preparing a pharmaceutical composition for modulating angiogenesis in a mammalian tissue.
DESC Human lysyl-oxidase protein, SEQ ID NO:8.
KW angiogenesis modulation; lysyl-oxidase; BOND_PC; lysyl oxidase preproprotein; protein-lysine 6-oxidase
SQL 417
SEQ
1 mrfawtvlll gplqlcalvh cappaagqqq p
= =====
51 ngqvflslsl gsqqpqrqr dpgaavpgaa nasaqqprtp illirdnrta
.
401 rytghhayas gctispy
HITS AT: 10-23

The hit portion of the answer sequence is highlighted with double underlining.

Repeat the DGENE search in REGISTRY and combine all results in CPlusSM

```
=> FILE REGISTRY
=> S L1
L2          41 LGP[VL]QLC[-GH]LV.CAP/SQSP
=> FIL HCAPLUS
=> S L2 AND P/DT
L3          30 L2 AND P/DT
=> TRA PN L1
L4          TRANSFER L1 1- PN :      30 TERMS
L5          82 L4
=> S L3 OR L5
L6          89 L3 OR L5
=> S L6 AND (ANTIBOD### OR IMMUNOGLOBULIN#) AND DIAGNOS? AND
PROSTAT? AND (CANCER? OR TUMOR? OR NEOPLAS?)
L7          4 L6 AND (ANTIBOD### OR IMMUNOGLOBULIN#) AND
PROSTAT? AND
```

To repeat an SCM search in REGISTRY simply **SEARCH** the answer set L-number from DGENE.

L3 = CPlus patent records found using REGISTRY.
L5 = CPlus patent records found using DGENE.
L6 = CPlus records found using both DGENE and REGISTRY in combination.

The CPlus search may be further refined using CAS value-added abstracts and indexing.

Use DGENE and REGISTRY in combination to locate relevant CPlus records

=> D L7 BIB ABS HITIND 2

L7 ANSWER 2 OF 4 HCAPLUS COPYRIGHT 2010 ACS on STN
 AN 2007:463771 HCAPLUS
 DN 146:397152
 TI Detection of tissue-derived glycoproteins
 the **diagnosis** and monitoring of disease
 IN Zhang, Hui; Aebersold, Rudolf H.
 PA Institute for Systems Biology, USA
 SO PCT Int. Appl., 242 pp.
 CODEN: PIXXD2
 DT Patent
 LA English
 FAN.CNT 1

This example CPlus record was retrieved by the unique combination of a DGENE GETSEQ search and CPlus value-added indexing search.

	PATENT NO.	KIND	DATE	APPLICATION NO.	DATE
PI	WO 2007047796	A2	20070426	WO 2006-US40784	20061017 <--
				

Tip: this arrow indicates the family member which was retrieved in the DGENE RUN GETSEQ search (L1).

IT CD antigens
 RL: ANT (Analyte); DGN (Diagnostic)
 BIOL (Biological study); USES (Use)
 (CD109, in serum, as **prostate cancer** marker; detection of tissue-derived glycoproteins shed into blood serum in **diagnosis** and monitoring of disease)

New option to SORT by BLAST percent identity (IDENT) in DGENE, USGENE, and PCTGEN

- Useful for identifying short, highly similar sequences, that have a low overall BLAST similarity score, e.g., probes, primers
- Useful for identifying short, highly similar areas within larger sequences, e.g., motifs, biomarkers
- Option to double-sort in combination with the overall BLAST similarity score
 - User chooses which is the primary sort parameter

Learn more about the new percent identity feature for sorting BLAST answer sets in DGENE, USGENE, and PCTGEN at:
http://www.stn-international.com/percent_identity_sorting.html

Example: SORT by percent identity

=> D IDENT SCORE 1-7

L3 ANSWER 1 OF 446 DGENE COPYRIGHT 2010 THOMSON REUTERS on STN
IDENT 100%
SCORE 496 100% of query self score 496

L3 ANSWER 2 OF 446 DGENE COPYRIGHT 2010 THO
IDENT 100%
SCORE 496 100% of query self score 496

• • •

L3 ANSWER 5 OF 446 DGENE COPYRIGHT 2010 THOMSON REUTERS on STN
IDENT 100%
SCORE 98 19% of query self score 496

L3 ANSWER 6 OF 446 DGENE COPYRIGHT 2010 THO
IDENT 100%
SCORE 98 19% of query self score 496

L3 ANSWER 7 OF 446 DGENE COPYRIGHT 2010 THOMSON REUTERS on STN
IDENT 100%
SCORE 98 19% of query self score 496

Sequences with 100% identity and with 100% overall similarity.

Sequences with areas of 100% identity, but with low overall similarity.

New FASTA and FASTA2 display formats added to USGENE and PCTGEN

- FASTA is a standard sequence format, which enables USGENE and PCTGEN data to be more easily imported for further offline analysis
- The FASTA display comprises the sequence in lines of 70 characters, and a header line providing a unique description of the sequence
- The FASTA2 display is a lower-priced alternative to FASTA, which provides the same sequence information with a simplified header line
- FASTA and FASTA2 may be used with standard formats ALL and BRIEF at no additional cost

Example: FASTA and FASTA2 display formats in USGENE and PCTGEN

=> FILE USGENE

=> S 20100017904.32958/AN

L1 1 20100017904.32958/AN

=> D FASTA

L1 ANSWER 1 OF 1 USGENE COPYRIGHT 2010 SEQUENCEBASE CORP on STN
FASTA:

```
>USGENE|20100017904.32958|Protein|sequence 32958 from US20100017904
mgevvatweateggagvkgpvvvtgasgflgswlvmkllqagytrratvrdpanvvktpkplldlpqater
lslwkadladegsfddairgctgvfhvatpmdfeskdpenevikptvegmmmsimrackeagtvrriivfts
sagtvnieerqrpvydqdnwsdvdfcqrvmkgwmyfvskslaekaamayaaehgldfisiptlvvgpf
lsagmpplitalalvtgnehahysilkqvqfvhlldldahlfhfehpaagryvcsshdatihglaaml
Rdrypeydiperfpgieddlqpvhfsskklldhgftfkytvedmfdairmcrekgliplatagggralp
```

=> D FASTA2

L1 ANSWER 1 OF 1 USGENE COPYRIGHT 2010 SEQUENCEBASE CORP on STN
FASTA2:

```
>USGENE|Protein
mgevvatweateggagvkgpvvvtgasgflgswlvmkllqagytrratvrdpanvvktpkplldlpqater
lslwkadladegsfddairgctgvfhvatpmdfeskdpenevikptvegmmmsimrackeagtvrriivfts
sagtvnieerqrpvydqdnwsdvdfcqrvmkgwmyfvskslaekaamayaaehgldfisiptlvvgpf
lsagmpplitalalvtgnehahysilkqvqfvhlldldahlfhfehpaagryvcsshdatihglaaml
rdrypeydiperfpgieddlqpvhfsskklldhgftfkytvedmfdairmcrekgliplatagggralp
```

AN 20100017904.32958
is SEQ ID NO 32958
from US20100017904.

Summary

- RUN BLAST is available for searching DGENE, USGENE and PCTGEN directly on STN
- CAS REGISTRY BLAST provides BLAST searching options for the REGISTRY database
- Sequence code match searching is available for DGENE, USGENE, PCTGEN and REGISTRY
- DGENE, USGENE, and PCTGEN search results may now be sorted by BLAST percent identity
- USGENE and PCTGEN results may now be displayed in FASTA and FASTA2 format

Resources for sequence searching on STN

- Sequence Searching on STN modular workshop
http://www.stn-international.com/sequence_searching.html
 - STN Sequence Databases
 - Sequence Code Match (SCM) searching
 - Searching DGENE, USGENE, PCTGEN
 - CAS REGISTRY BLAST
 - Multifile searching using DGENE, USGENE and PCTGEN
- USGENE resources, reference materials and FAQ
<http://www.sequencebase.com>
- CAS REGISTRY sequence coverage and resources
<http://www.cas.org/support/stngen/stndoc/sequences.html>

STN[®]

For more information ...

CAS

E-mail: help@cas.org

Support and Training:

www.cas.org

FIZ Karlsruhe

helpdesk@fiz-karlsruhe.de

Support and Training:

www.stn-international.de