

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 AB ALIGN 1,3,10
- 7) Ensure transcript was captured before logoff

The 7 basic steps of RUN BLAST

Search Question:

Find relevant U.S. published application and patent references for this protein sequence:

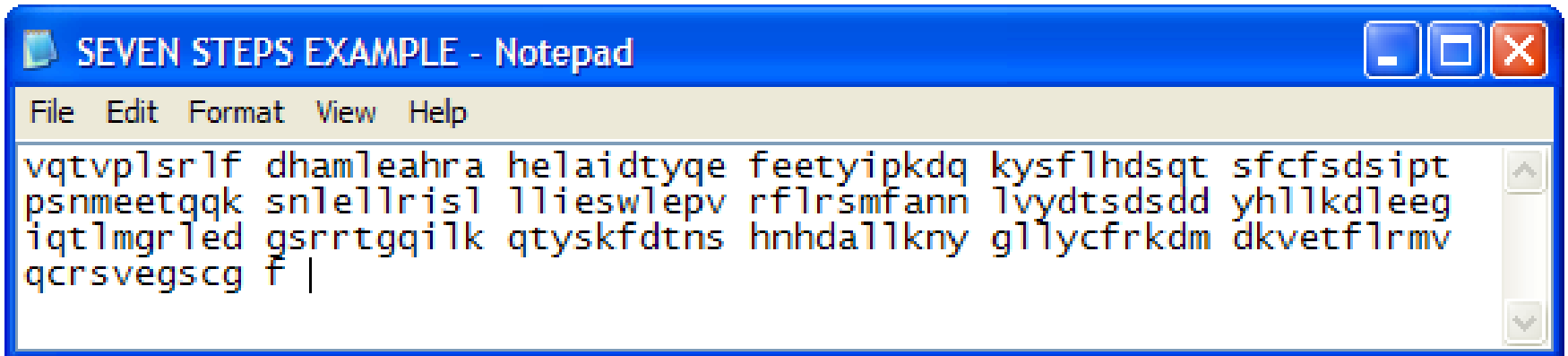
```
1 vqtvplsr1f dhamleahra helaidtyqe feetyipkdq kysflhdsqt
51 sfcfsdsipt psnmeetqk snlellrisl llieswlepv rflrsmfann
101 lvydtsdsdd yhllkdleeg iqtlmgrled gsrrtgqilk qtyskfdtns
151 hnhdallkny gllycfrkdm dkvetflrmv qcrsvegscg f
```

See also: USGENE STN Workshop Manual:

http://www.stn-international.com/usgene_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

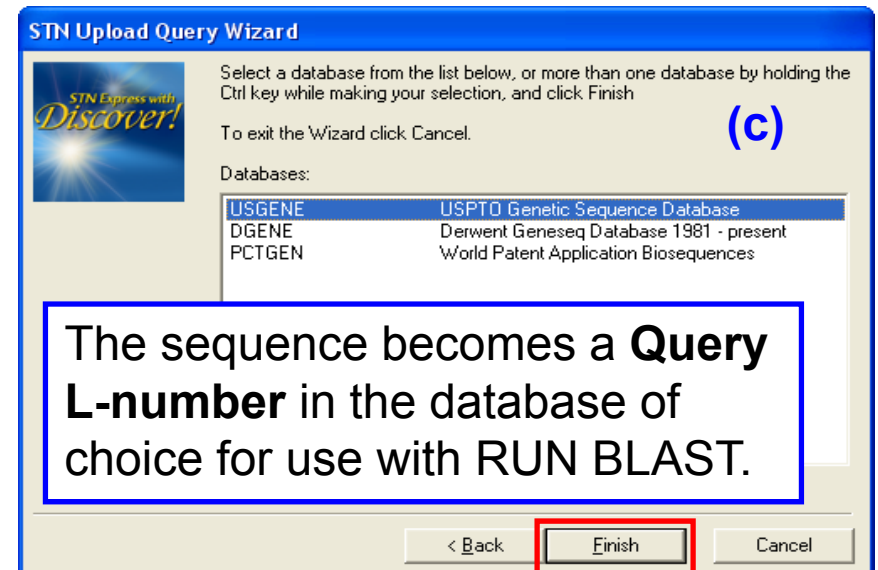
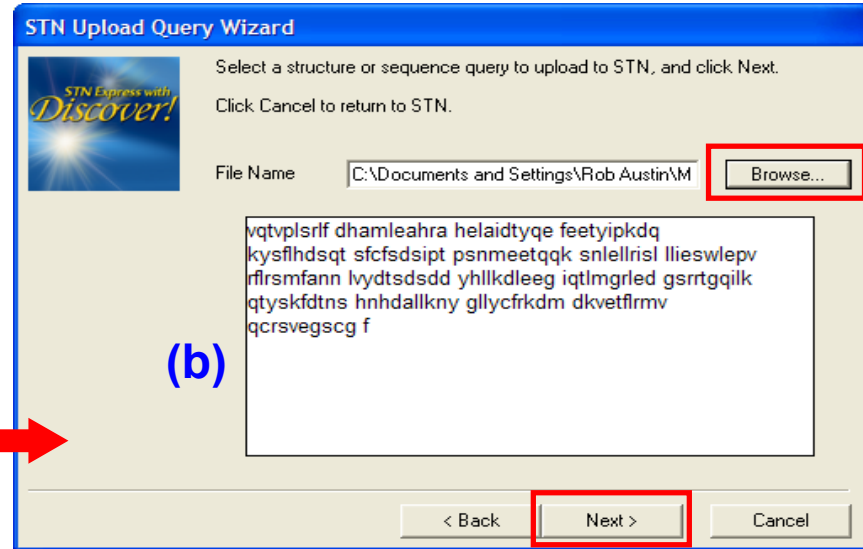
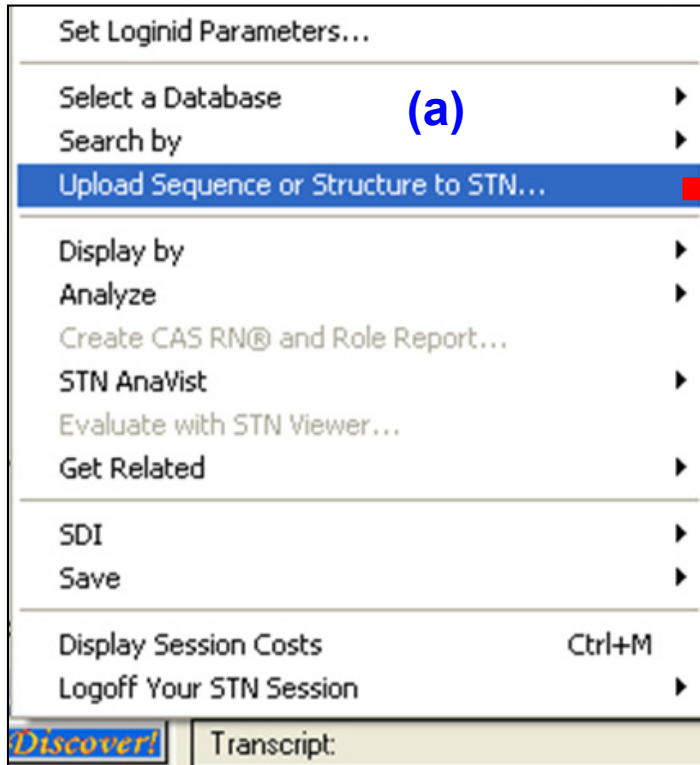


The screenshot shows a Notepad window with the title bar 'SEVEN STEPS EXAMPLE - Notepad'. The menu bar includes 'File', 'Edit', 'Format', 'View', and 'Help'. The text content is as follows:

```
vqtvplsr1f dhamleahra helaidtyqe feetyipkdq kysflhdsqt sfcfsdsipt  
psnmeetqqk snlellrisl llieswlepv rflrsmfann lvydtsdsdd yhllkdleeg  
iqtlmgrled gsrrtgqilk qtyskfdtns hnhdallkny 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
ptpsnmeetqqksnlellrislllieswlepvrflrsmfannlvydtsdsddyhllkd
leegiqtlmgrledgsrrtgqilkqtyskfdtnshnhdallknygllycfrkdmdkve
tflrmvqcrsvegscgf

The sequence query is now ready for searching directly in USGENE 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 USGENE

FILE 'USGENE' ENTERED AT 17:40:40 ON 31 M
COPYRIGHT (C) 2010 SEQUENCEBASE CORP

USGENE is updated within 3 days
of publication by the USPTO.

FILE LAST UPDATED: 28 MAY 2010 <20100528/UP>
MOST RECENT PUBLICATION DATE: 27 MAY 2010 <20100527/PD>
FILE COVERS 1981 TO DATE

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

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

=> RUN BLAST L1 /SQP -F F

Turn the Low Complexity Filter
off with the syntax: /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

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 USGENE 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

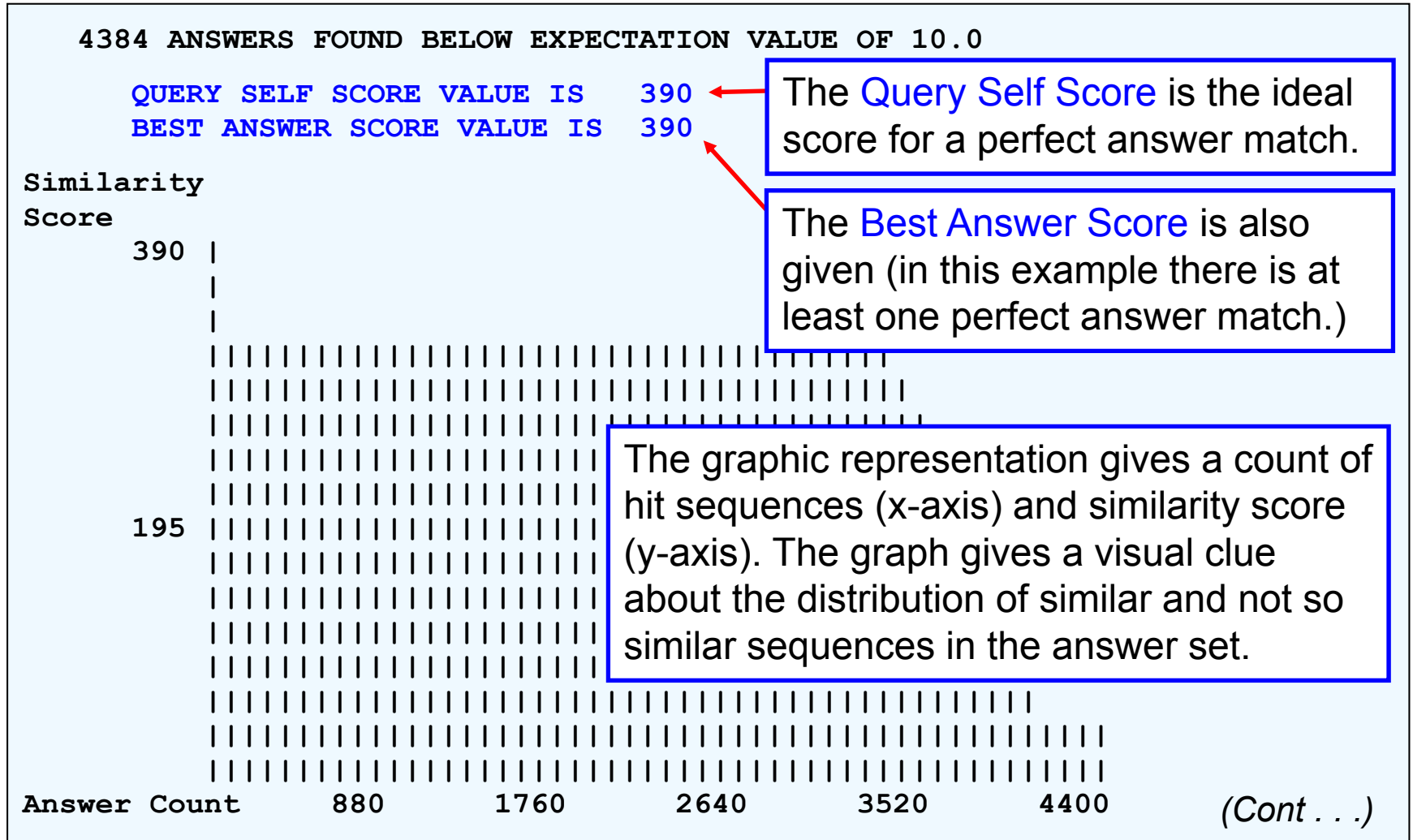
* See page 115-119: http://www.stn-international.com/usgene_wm.html

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



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 100%)

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 951 VQTVPLSRLFDHAMLEAHRAHEL
SFCFSDSIPTPSNMEETQQKSNLELLRISLLLLIESWLEPVRFRLRSMFANN
LVYDTSDDYHLLKDLLEGIQTLMGRLEDGSRRTGQILKQTYSKFDTNS
HNHDALLKNYGLLYCFRKMMDKVETFLRMVQCRSVEGSCGF/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 951 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 951 USGENE COPYRIGHT 2010 SEQUENCEBASE CORP on STN

TI Recombinant DNA transfer vectors (*Patent*)

MTY Protein

SQL 191

SCORE 390

100% of query self score 390

This perfect match top hit comes from a U.S. issued patent.

BLASTALIGN

Query = 191 letters

Length = 191

Score = 390 bits (1001), Expect = e-113

Identities = 191/191 (100%), Positives = 191/191 (100%)

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

Query: 1 VQTVPLSRLFDHAMLEAHRAHEL AIDTYQEF EETYIPKDQKYSFLHDSQTSFCFSDSIPT

VQTVPLSRLFDHAMLEAHRAHEL AIDTYQEF EETYIPKDQKYSFLHDSQTSFCFSDSIPT

Sbjct: 1 VQTVPLSRLFDHAMLEAHRAHEL AIDTYQEF EETYIPKDQKYSFLHDSQTSFCFSDSIPT

Query: 61 PSNMEETQQKSNLELLRISLLLI ESWLEPVRFLRSMFANNLVYDTSDDSDDYHLLKDLEEG

PSNMEETQQKSNLELLRISLLLI ESWLEPVRFLRSMFANNLVYDTSDDSDDYHLLKDLEEG

Sbjct: 61 PSNMEETQQKSNLELLRISLLLI ESWLEPVRFLRSMFANNLVYDTSDDSDDYHLLKDLEEG

Query: 121 IQTLMGRLEDGSRRTGQILKQTY SKFDTN SHNHDALLKNYGLLYCFRKDM

IQTLMGRLEDGSRRTGQILKQTY SKFDTN SHNHDALLKNYGLLYCFRKDM

Sbjct: 121 IQTLMGRLEDGSRRTGQILKQTY SKFDTN SHNHDALLKNYGLLYCFRKDM

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

```
L3      ANSWER 5 OF 951  USGENE COPYRIGHT 2010 SEQUENCEBASE CORP on STN
TI      METHOD OF IMPROVING EFFICACY OF BIOLOGICAL RESPONSE-MODIFYING
        PROTEINS AND THE EXEMPLARY MUTEINS (PublishedApplication)
DESC    Artificial Protein; PL 10th, 31st, 44th, 52nd, 54th, 92nd, 97th,
        146th, 166th, 176th or 191st Phe is replaced by Val; sequence 23 of
        65
MTY     Protein
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
Query: 1  VQTVPLSRLFDHAMLEAHRAHELAIPTYQEFEEETYIPKDQKYSFLHDSQTSFCFSDSIPT
          VQTVPLSRLFDHAML+AHRAH+LAIDTYQEFEEETYIPKDQKYSFLHDSQTSFCFSDSIPT
Sbjct: 1  VQTVPLSRLFDHAMLQAHRAHQLAIDTYQEFEEETYIPKDQKYSFLHDSQTSFCFSDSIPT
Query: 61 PSNMEETQQKSNLELLRISLLLIESWLEPVRFLRSMFANNLVYDTSDDY . . . .
          PSNMEETQQKSNLELLRISLLLIESWLEPVRFLRSMFANNLVYDTSDDY
Sbjct: 61 PSNMEETQQKSNLELLRISLLLIESWLEPVRFLRSMFANNLVYDTSDDY . . . .
```

The 5th from top hit comes from a U.S. published application.

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 100%)
```

```
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      951 VQTVPLSRLFDHAMLEAHRAHELAI
        SFCFSDSIPTSPNMEETQQKSNLELLKRTSLLLTESWLEPVRFLKSMFANN
        LVYDTSDDYHLLKDLLEGIQTLMGRLEDGSRRTGQILKQTYSKFDTNS
        HHNDALLKNYGLLYCFRKDMDKVETFLRMVQCRSVEGSCGF/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      951 SOR L2 SCORE D
```

The BLAST search (L2) is further refined to sequences from U.S. granted patents, with application year prior to 2001 (L4).

```
=> S L2 AND AY<2001 AND GRANTED/SSO
```

```
L4      20 L2 AND AY<2001 AND GRANTED/SSO
```

```
=> SOR SCORE D
```

```
PROCESSING COMPLETED FOR L4
```

```
L5      20 SOR L4 SCORE D
```

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 AB SCORE ALIGN 1,3,10**

6) Display selected answers in a preferred bibliographic format

=> D BIB AB SCORE ALIGN 1 4

L5 ANSWER 1 OF 20 USGENE COPYRIGHT 2010 SEQUENCEBASE CORP on STN
AN 4363877.1 Protein USGENE
TI Recombinant DNA transfer vectors (Patent)
IN Goodman Howard M. (San Francisco, CA); Shi
CA); Seeburg Peter H. (San Francisco, CA)
PA The Regents of the University of California
PI US 4363877 A 19821214
AI US 1978-897710 19780419
AB Recombinant DNA transfer vectors containing codons for human
somatomammotropin and for human growth hormone.

This sequence comes from a U.S. granted patent, with an application date prior to 2001.

SCORE 390 100% of query self score 390

BLASTALIGN

Query = 191 letters

Length = 191

Score = 390 bits (1001), Expect = e-113

Identities = 191/191 (100%), Positives = 191/

Query: 1 VQTVPLSRLFDHAMLEAHRAHELAIPTYQEFEEETYIPKDQKYSFLHDSQTSF
VQTVPLSRLFDHAMLEAHRAHELAIPTYQEFEEETYIPKDQKYSFLHDSQTSF
Sbjct: 1 VQTVPLSRLFDHAMLEAHRAHELAIPTYQEFEEETYIPKDQKYSFLHDSQTSF

Note: this USGENE record is an example of one which is not present in DGENE or REGISTRY.

6) Display selected answers in a preferred bibliographic format

```
L5 ANSWER 4 OF 20 USGENE COPYRIGHT 2010 SEQUENCEBASE CORP on STN
AN 6010999.4 protein USGENE
TI Stabilization of fibroblast growth factors by modification of
cysteine residues (Patent)
IN Daley Michael Joseph (Yardley, PA); Buckwold Robert (Yardley, PA);
PA Cady Susan Mancini (Yardley, PA); Shih Hsiang-Chun (Yardley, PA);
BO Bohlen Peter (Peekskill, NY); Seddon James (Peekskill, NY);
PA American Cyanamid Company (Madison NJ)
PI US 6010999 A 20000104
AI US 1995-459906 19950602
DT Patent
AB The present invention relates to physiologically-active derivatized
natural and recombinant mammalian and human proteins and . . .
SCORE 331 84% of query self score 390
BLASTALIGN
Query = 191 letters
Length = 194
Score = 331 bits (849), Expect = 4e-96
Identities = 162/189 (85%), Positives = 174/189
Query: 3 TVPLSRLFDHAMLEAHRAHELAIPTYQEFEEYIPKDKYSFLHDSQTSFCF . . . .
T+PLSRLFD+AML AHR H+LA DTYQEFEE YIPK+QKYSFL + QTS CF
Sbjct: 6 TIPLSRLFDNAMLRAHRLHQLAFDPTYQEFEEAYIPKEQKYSFLQNPQTS LCF . . . .
```

This sequence also comes from a U.S. granted patent, with an application date prior to 2001.

Note: this USGENE record is an example of one which is not present in the NCBI or EMBL-EBI patent divisions.

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. A 'Capture Session' dialog box is open, showing a file list in the 'Look in: Tmscript' folder. The file list includes AAC03542, ACETYLCYSTEINE, AGENTS MULTIFILE 2, AGENTS MULTIFILE 2A, ALICE SCRIPT TEST, ALIGN EXAMPLE, aligns, ALLSTR EXAMPLE, another multifile exam, and ASPIRIN. The 'File name' field contains 'seven steps example tm'. The 'Files of type' dropdown is set to 'Transcript Files (*.tm)'. The 'Capture retrospectively' checkbox is checked and circled in red. A 'Select Discover! Wizard' dialog box is also open, showing a search history table:

Search history	
L2	951 RUN BLAST L1 /SQP -F
L3	951 SOR SCORE D
L4	20 S L2 AND AY<2001 AND
L5	20 SOR SCORE D

The main window displays a transcript with the following text:

```
Answer Count
ENTER EITHER THE
OR ENTER MINIMUM
(BEST ANSWER PER
ENTER (ALL) OR ?
L2 RUN STATEME
L2 951 VQ
SF
LV
HN
Answer set arrang
similarity score,
=> SOR SCORE D
PROCESSING COMPLE
L3 951 S
=> S L2 AND AY<20
1952506 AY
(AY<2001)
5545036 GRANTED/SSO
L4 20 L2 AND AY<2001 AND GRANTED/SSO
=> SOR SCORE D
PROCESSING COMPLETED FOR L4
L5 20 SOR L4 SCORE D
=>
```

The status bar at the bottom shows 'Discover!', 'Transcript:', 'USGENE', 'INS', 'Hold Off', 'Print Off', 'Online', and '00:09:44'.

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 settings 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
```

```
. . . . .
```

```
8518 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 2019 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 2019 SOR L1 SCORE D

Correct use of BLAST options
finds relevant sequence hits.

=> D TRI SCORE ALIGN

L2 ANSWER 1 OF 2019 USGENE COPYRIGHT 2010 SEQUENCEBASE CORP on STN

TI Anti-pro6094 antibodies (Patent)

DESC Homo sapiens Protein; sequence 442 of 550

MTY Protein

SQL 117

SCORE 38 100% of query self score 38

BLASTALIGN

Query = 11 letters

Length = 117

Score = 37.5 bits (81), Expect = 4e-09

Identities = 11/11 (100%), Positives = 11/11 (100%)

Query: 1 GSSFLSPEHQR 11

GSSFLSPEHQR

Sbjct: 24 GSSFLSPEHQR 34

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 AB ALIGN 1,3,10
- 7) Ensure transcript was captured before logoff

CAS REGISTRY BLAST searching

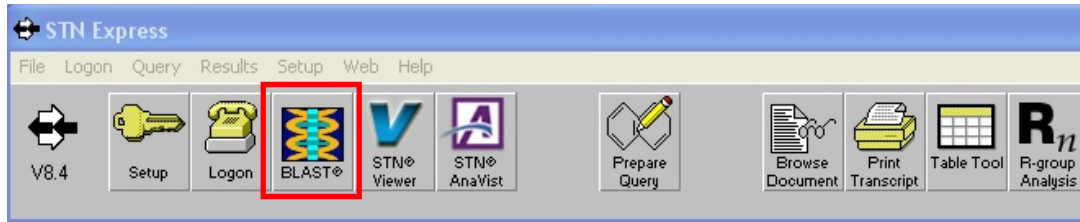
Search Question:

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 screenshot shows the Result Set Manager software interface. The menu bar includes File, Edit, Search, Tools, and Help. The toolbar contains several icons: New Search (highlighted with a red box), Sequence, Sequence ID, Fast BLAST, Alerts Profiles, Help, and Exit. Below the toolbar is a section titled 'Manage and Review Results' which contains a table of search results.

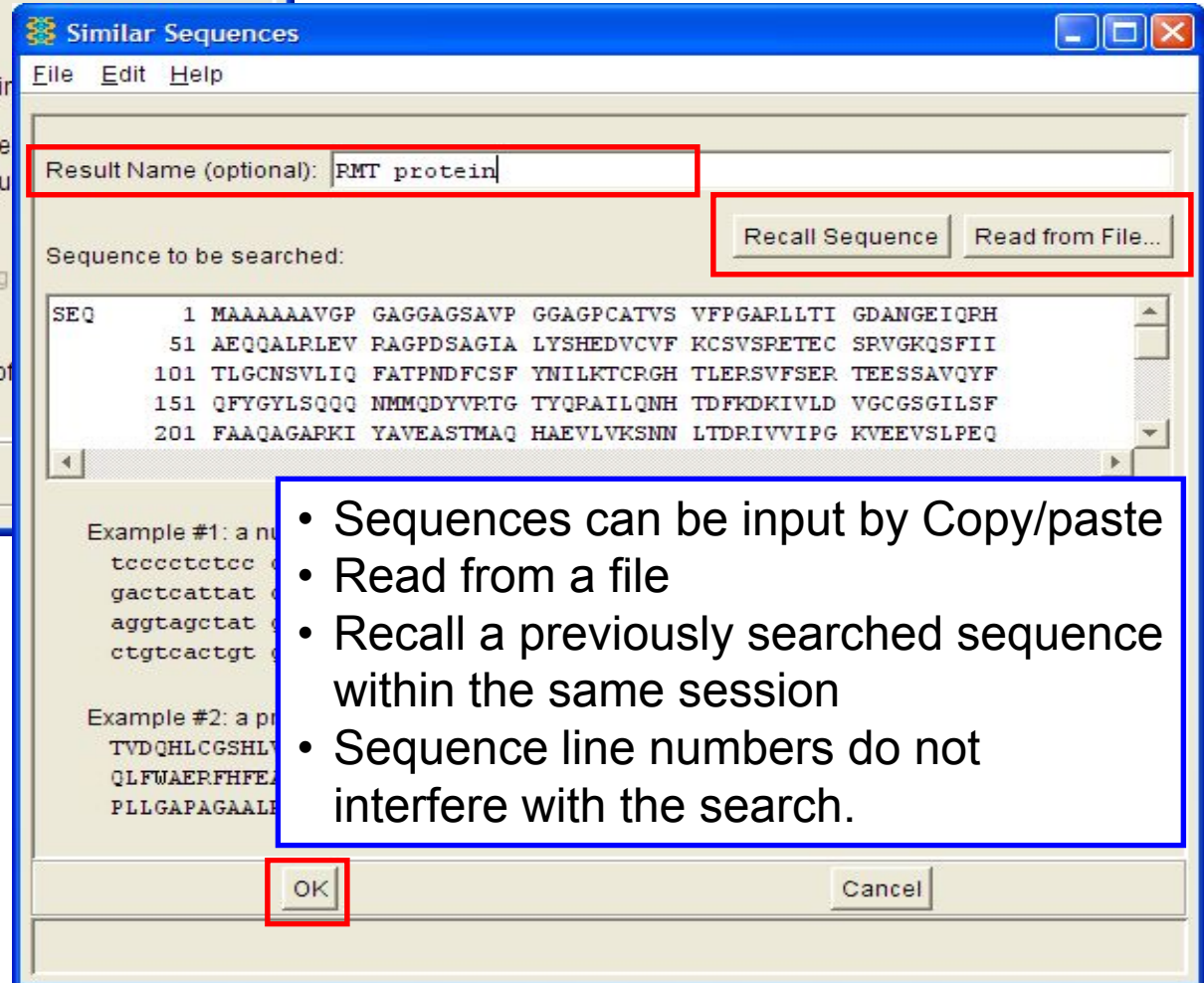
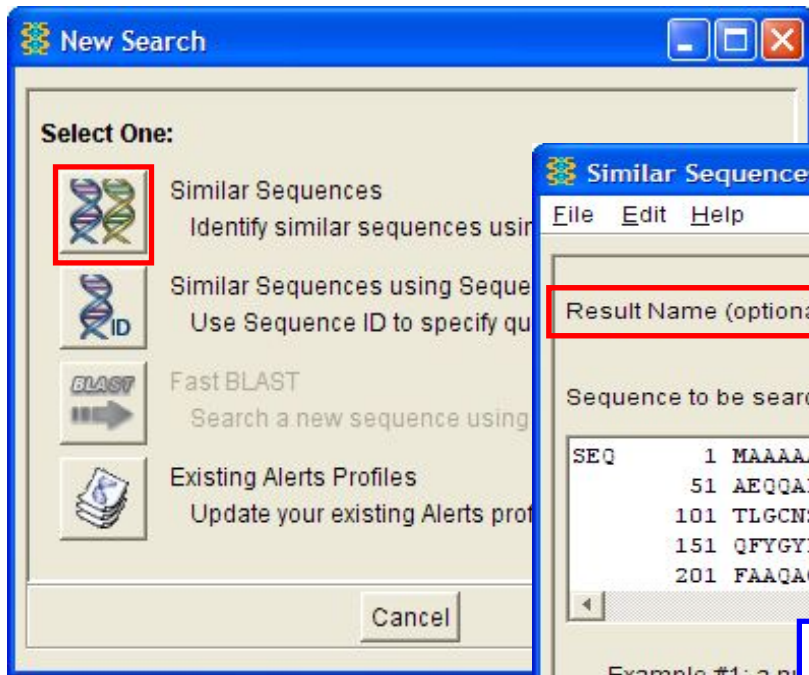
Name	Type	Created /	Status	Results	Reviewed
IL-17 nucleic	BLASTn	2008-09-11 10:28 AM	Complete	753	✓
MTVEF	BLASTp	2008-07-28 01:23 PM	Complete	32	✓
ACE PCR PRIMER	BLASTn	2008-07-22 04:25 PM	Complete	695	✓
mtvefn	BLASTp	2008-02-04 11:19 AM	Complete	31	✓
sh3-bp-2	BLASTn	2007-11-27 09:23 PM			
sh3-bp	BLASTn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

31 results (100 maximum)

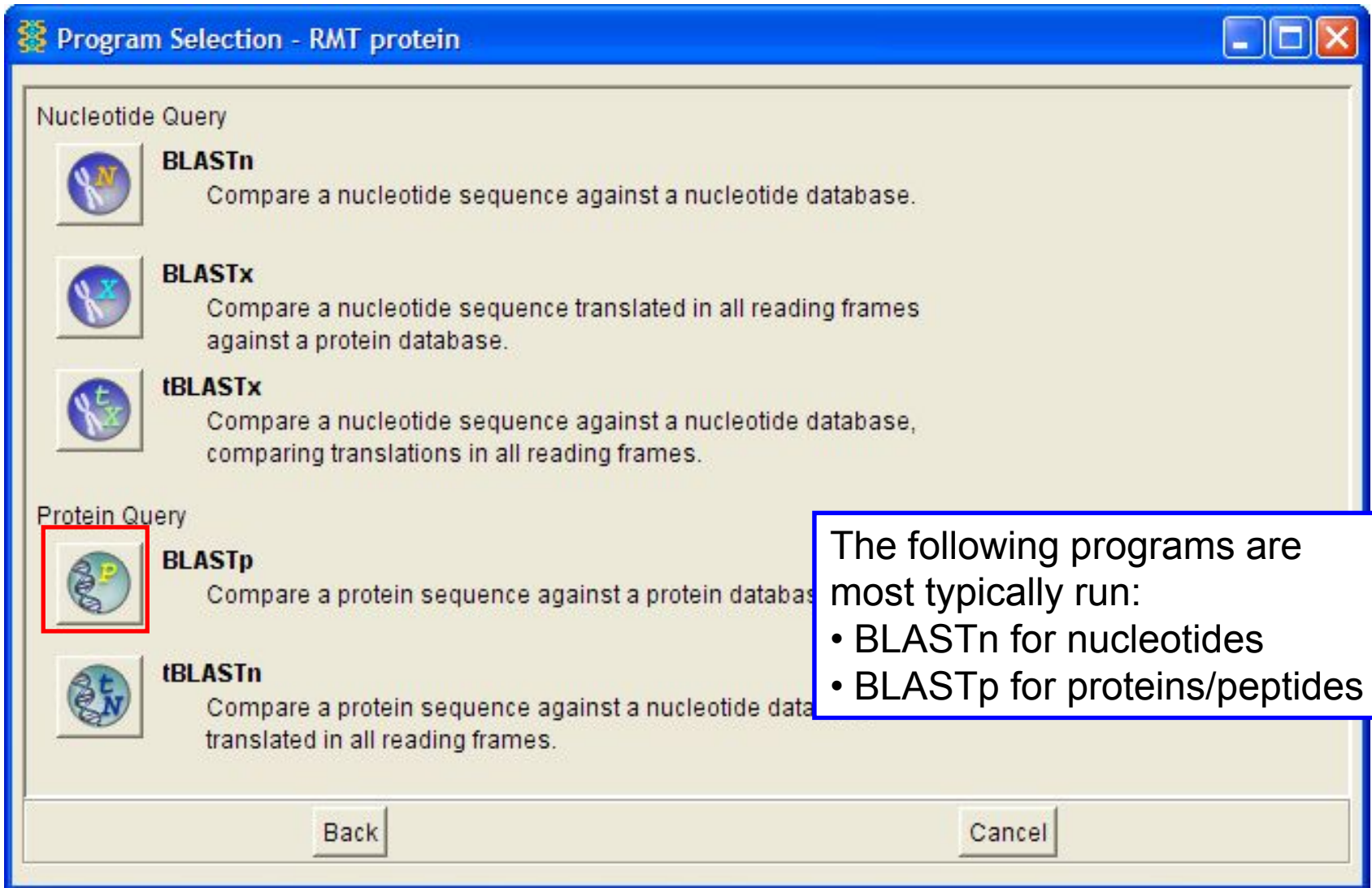
View Results

- 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	20	✓
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

CAS Registry BLAST® Report - RMT protein

File Edit View Search Tools Help

Unique Sequences: 809 Redundant: 348 Selected Results: 0

Alignment Scores

<40 40-50 50-80 80-200 >=200

Alignment Summary

1 153 305 456 608

(1072557-06-2) Methyltransferase, transcriptional coactivator protein (arginine) (human)

Alignment Details

1225 0.0 (1072557-06-2) Methyltransferase, transcriptional coactivator protein (arginine) (human)

Length = 608
Score = 1225 Expect = 0.0
Identities = 608/608 (100%) Positives = 608/608 (100%)

Query: 1 MAAAAA AVGP GAGGAGS AVPGGAGPCATVSVFPGARLLTIGDANGEIQPHAEQQA 55
MAAAAA AVGP GAGGAGS AVPGGAGPCATVSVFPGARLLTIGDANGEIQPHAEQQA

Subject: 1 MAAAAA AVGP GAGGAGS AVPGGAGPCATVSVFPGARLLTIGDANGEIQPHAEQQA 55

Query: 56 LRLEVRAGPDSAGIALYSHEDVCFKCSVSPETECSRVGKQSFIIITLGCNSVLIQ 110
LRLEVRAGPDSAGIALYSHEDVCFKCSVSPETECSRVGKQSFIIITLGCNSVLIQ

Subject: 56 LRLEVRAGPDSAGIALYSHEDVCFKCSVSPETECSRVGKQSFIIITLGCNSVLIQ 110

Query: 111 FATPND FCSFYNILKTCRGHTLERSVFSERTEESSAVQYFQFYGYLSQQQNMMD 168
FATPND FCSFYNILKTCRGHTLERSVFSERTEESSAVQYFQFYGYLSQQQNMMD

Subject: 111 FATPND FCSFYNILKTCRGHTLERSVFSERTEESSAVQYFQFYGYLSQQQNMMD 168

Query: 166 YVRTGT YQRAILQNHTDFKDKIVLDVCGSGILSFFAAQAGARKIYAVEASTMAQ 222
YVRTGT YQRAILQNHTDFKDKIVLDVCGSGILSFFAAQAGARKIYAVEASTMAQ

Subject: 166 YVRTGT YQRAILQNHTDFKDKIVLDVCGSGILSFFAAQAGARKIYAVEASTMAQ 222

Get STN Data Cancel

Result complete.

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

CAS Registry BLAST® Report - RMT protein

File Edit View Search Tools Help

Unique Sequences: 809 Redundant: 348 Selected Results: 61

Alignment Scores

<40 40-50 50-80 80-200 >=200

Alignment Summary

1 153 305 456 608

Alignment Details

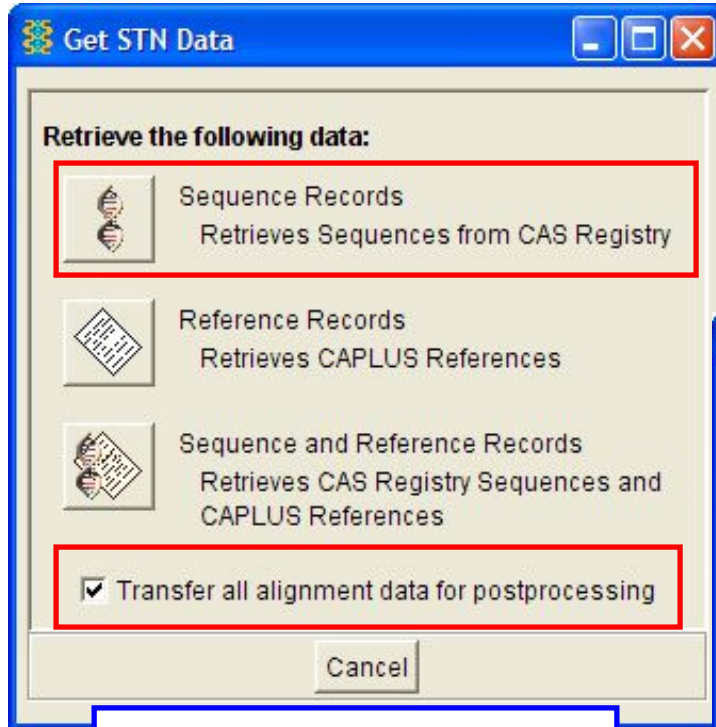
+	█	☑	ID	Score	Description
+	█	☑	1225	0.0	(1072557-06-2) Methyltransferase, transcriptional coactivator protein (arginine) (human)
+	█	☑	1223	0.0	(453617-62-4) Drug-metabolizing enzyme DME-7 (hu
+	█	☑	1222	0.0	(642514-19-0) Protein 27420 (human)
+	█	☑	1222	0.0	(434009-21-9) 5: PN: WO0244358 FIGURE: 4A-4B u
+	█	☑	1198	0.0	(942169-05-3) 69: PN: US20070141652 SEQID: 69 U
+	█	☑	1192	0.0	(696682-62-9) Transcription factor SRC-2 (steroid re
+	█	☑	1186	0.0	(696686-43-8) 3: PN: US6743614 SEQID: 3 unclame
+	█	☑	1177	0.0	(863541-89-3) Methyltransferase, transcriptional co
+	█	☑	1175	0.0	(867594-03-4) Protein (Mus musculus strain C57BL/6
+	█	☑	1167	0.0	(631936-53-3) Methyltransferase, transcriptional co
+	█	☑	1140	0.0	(863541-93-9) Methyltransferase, transcriptional coactivator protein (arginine) (Rattus norvegicus gene CARM1 isoenzyme CARM1-v4)

Get STN Data Cancel

Sequences can be selected:

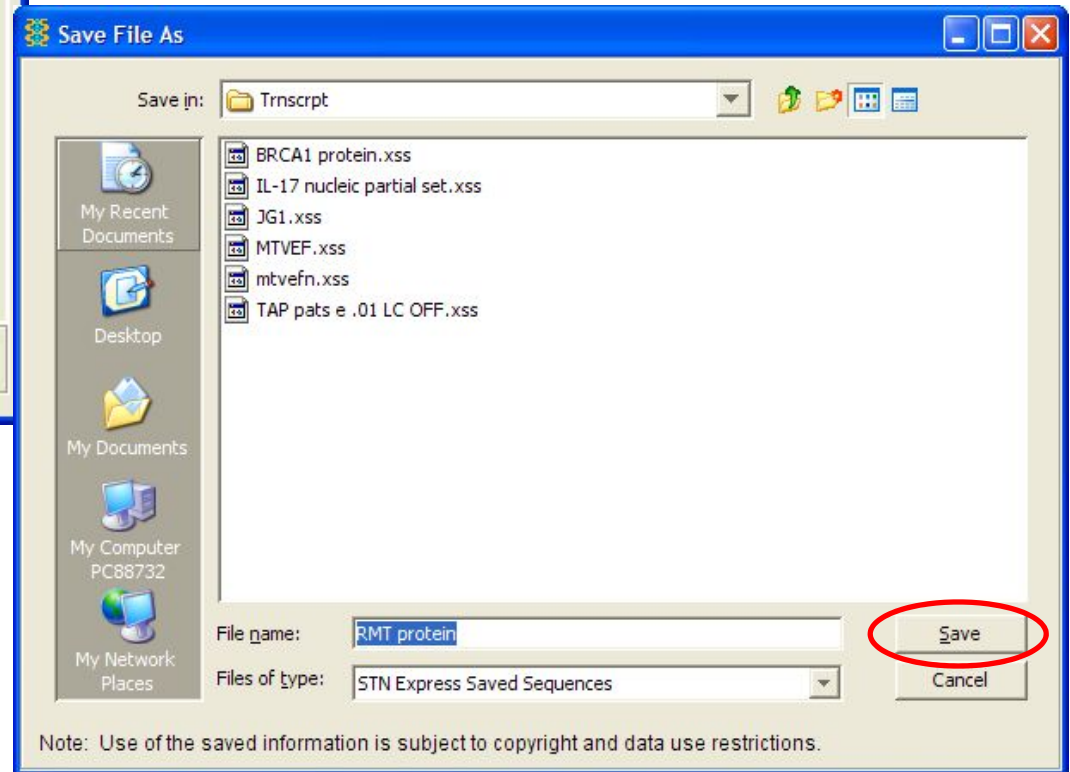
- In groups, using the color bar in the Alignment Scores
- Individually, by selecting the check box
- To transfer the sequence data to STN, click the Get STN Data button.

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] window. The menu bar includes File, Edit, Online, Query, Results, Preferences!, Web, Window, and Help. The toolbar contains various icons for file operations, search, and execution. The main display area shows a list of sequences, each starting with '1' followed by a sequence ID and '/RN'. The sequences are:

- 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. At the bottom, a command prompt shows the command: => D L6 1 SQIDE. A 'Discover!' button is visible at the bottom left, and a status bar at the bottom right says '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 Answers 1-7

 Family 1 Answers 1-3

 Family 2 Answers 4-5

 Family 3 Answers 6-7

 13 Individual Records Answers 8-20

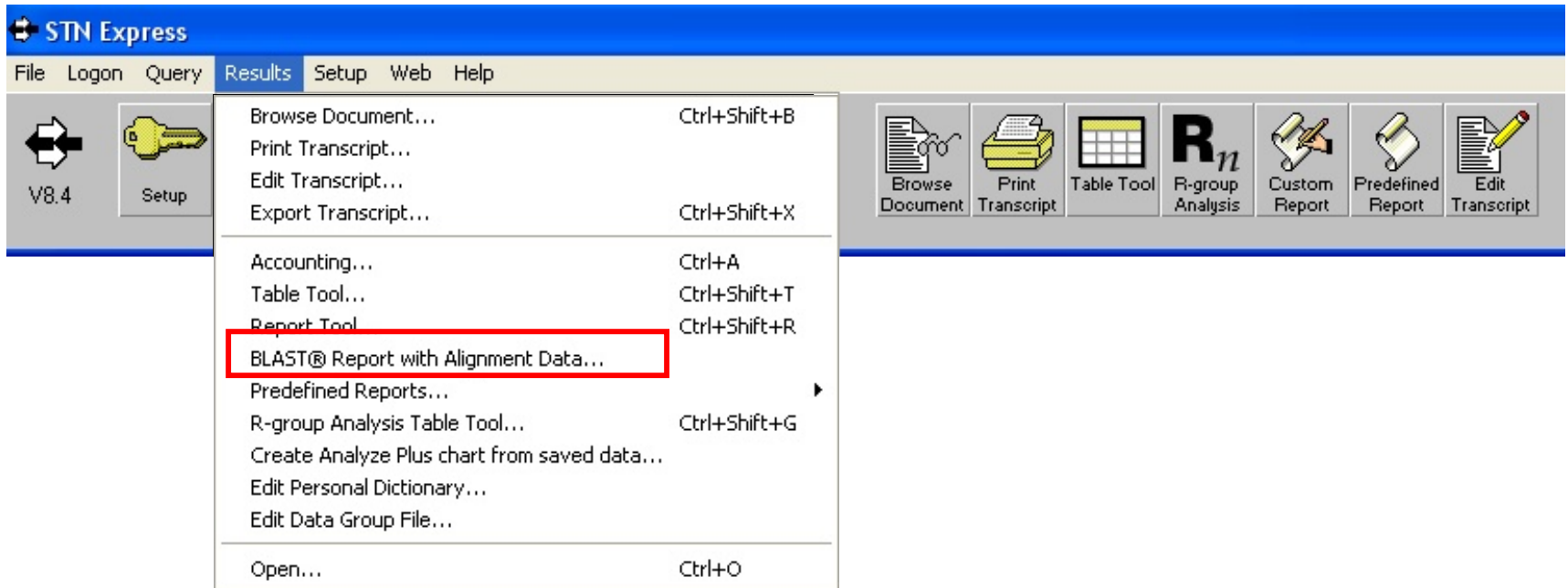
 0 Non-patent Records

=> **D L9 IBIB ABS HITRN 1-20**

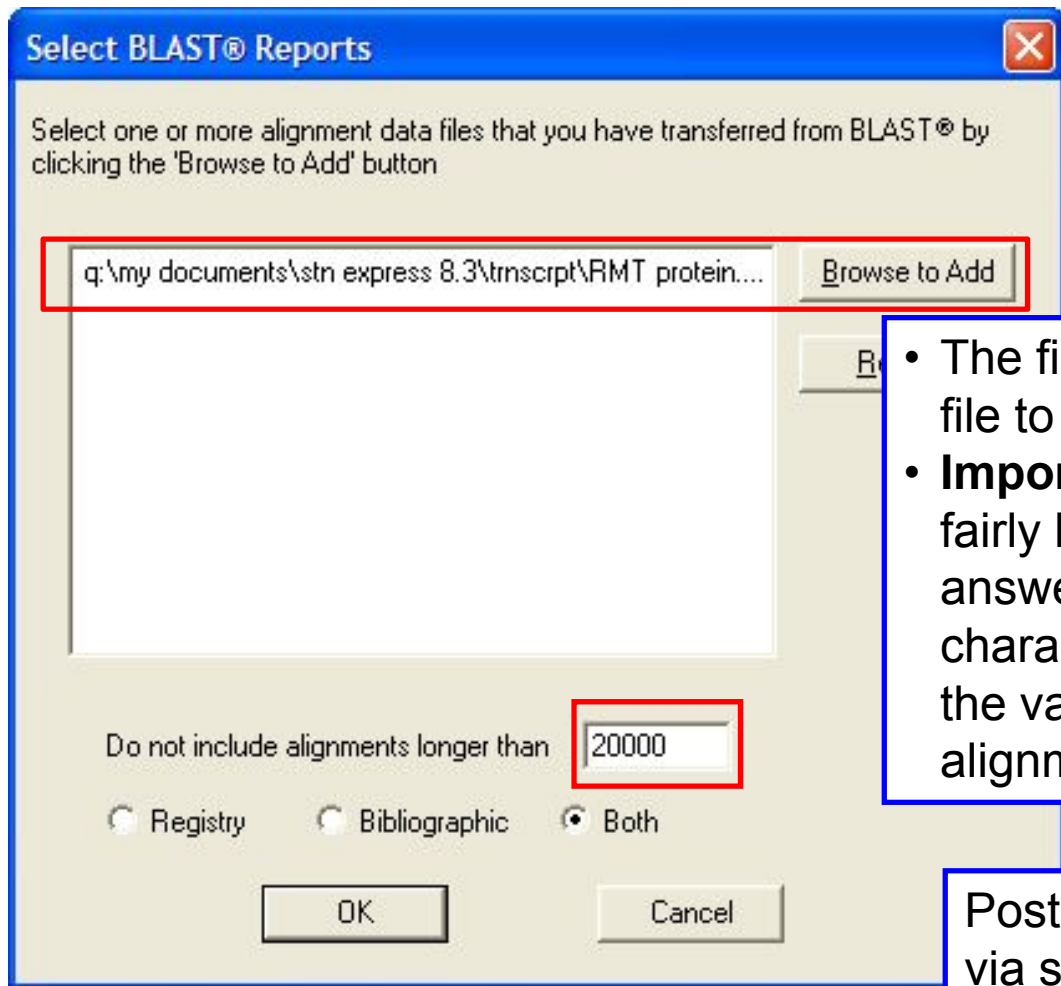
Additional keyword refinement or other searches can be used in CPlus. In this example, patents and nonpatents were separated in 2 L-numbers.

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 – Similarity Search Strategy

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 USGENE COPYRIGHT 2010
```

```
SEQ 1 gccgccgt
```

```
=====
```

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

```
SQL 8
```

The SEQ display in USGENE 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 USGENE COPYRIGHT 2010 SEQUENCEBASE CORP on STN
```

```
SEQ 1 tgtagttcat tatcatcttt gtcacagct gaagatgaaa taogatgtaa
```

```
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 USGENE COPYRIGHT 2010 SEQUENCEBASE CORP 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 USGENE COPYRIGHT 2010 SEQUENCEBASE CORP 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 USGENE COPYRIGHT 2010 SE
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 USGENE C
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 USGENE 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 USGENE GETSEQ SCM search

=> FILE USGENE

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

L1 RUN STATEMENT CREATED

L1 32 LGP[VL]QLC[-GH]LV.CAP/SQSP

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

=> D TRI SEQ

L1 ANSWER 1 OF 32 USGENE COPYRIGHT 2010 SEQUENCEBASE CORP on STN

TI Nucleotide and amino acid sequences, and assays and methods of use thereof for diagnosis of prostate cancer (Patent)

MTY Protein

SQL 417

SEQ

1 mrfawtvlll gplqlcalvh cappaagqqq |

= ===== =

51 ngqvfsllsl gsqyqpqrrr dpgaavpgaa nasaqqprtp illirdnrta

. . . .

401 rytghhayas gctispy

HITS AT: 10-23

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

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

=> FILE REGISTRY

=> S L1

L2 38 LGP[VL]QLC[-GH]LV.CAP/SQSP

=> FIL HCAPLUS

=> S L2 AND P/DT

L3 28 L2 AND P/DT

=> TRA PN L1

L4 TRANSFER L1 1- PN : 30 TERMS

L5 65 L4

=> S L3 OR L5

L6 75 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#)
PROSTAT? AND (CANCER? OR TUMOR? OR NEOPLAS?)

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

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

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

Use USGENE and REGISTRY in combination to locate relevant CPlus records

=> D L7 BIB ABS HITIND

L7 ANSWER 1 OF 4 HCAPLUS COPYRIGHT
AN 2007:463771 HCAPLUS
TI Detection of tissue-derived glycoproteins in blood serum in **diagnosis** and monitoring of disease
IN Zhang, Hui; Aebersold, Rudolf H.
PA Institute for Systems Biology, USA

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

FAN.CNT 1

	PATENT NO.	KIND	DATE	APPLICATION NO.	DATE
PI	WO 2007047796	A2	20070426	WO 2006-US40784	20061017
	US 20070099251	A1	20070503	US 2006-582861	20061017 <--
PRAI	US 2005-728044P	P	20051017		

AB A method of detecting tissue-derived glycoproteins in blood serum that is useful in the **diagnosis** of disease and in monitoring

IT Bladder, **neoplasm**
Ovary, **neoplasm**
Prostate gland, disease
Prostate gland, **neoplasm**

(glycoprotein shedding into blood in **diagnosis** of; detection of tissue-derived glycoproteins shed into blood serum in diagnosis and monitoring of disease)

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

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 USGENE COPYRIGHT 2010 SEQUENCEBASE CORP on STN
IDENT 100%
SCORE 496 100% of query self score 496

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

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

• • •

L3 ANSWER 5 OF 446 USGENE COPYRIGHT 2010 SEQUENCEBASE CORP on STN
IDENT 100%
SCORE 98 19% of query self score 496

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

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

L3 ANSWER 7 OF 446 USGENE COPYRIGHT 2010 SEQUENCEBASE CORP on STN
IDENT 100%
SCORE 98 19% of query self score 496

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
mgevvatweateggagvkgpvvvtgasgflgswlvmkllqagytvratvrdpanvvktpkplldlpgater
lslwkadladegsfddairgctgvfhvatpmdfeskdpenevikptvegmmmsimrackeagtvrriivfts
sagtvnieerqrpvydqdnwsdvdvcqrvmkgwmyfvskslaekaamayaaehgldfisiptlvvgpf
lsagmpplitalalvtgneahysilkqvqfvhlldldahflfehpaagryvcsshdatihglaaml
Rdrypeydiperfpgieddllqpvhfsskklldhgftfkytvedmfdairmcrekgliplatagggralp

=> D FASTA2

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

>USGENE|Protein
mgevvatweateggagvkgpvvvtgasgflgswlvmkllqagytvratvrdpanvvktpkplldlpgater
lslwkadladegsfddairgctgvfhvatpmdfeskdpenevikptvegmmmsimrackeagtvrriivfts
sagtvnieerqrpvydqdnwsdvdvcqrvmkgwmyfvskslaekaamayaaehgldfisiptlvvgpf
lsagmpplitalalvtgneahysilkqvqfvhlldldahflfehpaagryvcsshdatihglaaml
rdrypeydiperfpgieddllqpvhfsskklldhgftfkytvedmfdairmcrekgliplatagggralp

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[®]

Sequence Basics

www.stn-international.com