Taking command – effective use of sequence search options in USGENE®, DGENE and PCTGEN

Robert Austin – FIZ Karlsruhe

Agenda

- Sequence searchable databases on STN
- Introduction to USGENE
- Command line sequence searching
  - BLAST searching
  - Refining BLAST searches with text and date terms
  - BLAST advanced options
  - GETSIM (FASTA) searching
  - Offline BATCH search mode
  - Small fragment searches using GETSEQ

STN® sequence searchable databases

- CAS REGISTRY℠
  - Chemical Abstracts Service (CAS) REGISTRY File
- USGENE
  - The USPTO Genetic Sequence Database
- DGENE
  - Thomson Scientific GENESEQ™
- PCTGEN
  - WIPO/PCT Patent Application Biosequences

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

- BLAST similarity 

  - RUN BLAST
- FASTA similarity 

  - RUN GETSIM
- Sequence Code Match (SCM)
  - RUN GETSEQ
- Offline BATCH and ALERT options

Note: in this e-Seminar these techniques will be demonstrated in USGENE.

A new feature to refine USGENE, DGENE and PCTGEN sequence search results

- A new command line feature is now available to refine BLAST® or FASTA (GETSIM) sequence search results by percent (%) match
- The feature enables patent sequence searchers to focus results immediately to the most relevant answers and reduce analysis and review time
- **Search tip:** using offline BATCH search mode, flexible re-retrieval of search results at various percent levels is possible over an 8 day period

Today's e-Seminar reviews both this new feature and other useful tips for effective STN command line sequence searching.

Agenda

- Sequence searchable databases on STN
- Introduction to USGENE
- Command line sequence searching
  - BLAST searching
  - Refining BLAST searches with text and date terms
  - BLAST advanced options
  - GETSIM (FASTA) searching
  - Offline BATCH search mode
  - Small fragment searches using GETSEQ

USGENE is the USPTO Genetic Sequence Database

- Sequences captured from all relevant USPTO published patent applications and granted (issued) patents
- Assignee and full inventor names; publication, application and parent case PCT numbers and dates; original publication **title, abstract,** and **claims**
- Organism name, sequence length, Molecule Type, SEQ ID, and feature tables for features/annotations
- Produced by the SequenceBase Corporation
- Updated weekly – within **3 days** of publication
- 1982 – present

USGENE sequence records are available within 3 days of publication by the USPTO


```
L1    ANSWER 1 OF 1  USGENE COPYRIGHT 2008 SEQUENCEBASE CORP on STN
AN    20080069867.4  Protein       USGENE
TI    Diagnostic and/or Remedy for Ovarian Cancer (Pu
IN    Shimada Kaoru (Chiba, JP); Matsumura Yasuhiro (
      Toshiaki (Tokyo, JP)
PA    Mitsubishi Pharma Corporation (Osakashi JP)
PI    US 20080069867      A1   20080320
AI    US 2005-575406           20050819
RLI   WO 2005-JP15119          20050819
ED    20080321
DT    Patent
AB    It was found tha                                    nces
      described in the                                    n
      antibody having                                     the
      cancer species known so far such as gastric cancer, colon cancer and
      breast cancer, and was useful as an ovarian cancer tissue-specific
      diagnostic and/or therapeutic agent.
ECLM  US20080069867 A1: 1-14. (canceled) 15: A diagnostic and/or
      therapeutic agent for ovarian cancer, which com rises an antibody
      containing the amino acid sequence of any one of SEQ ID NOS: 1 to 6
      described in the sequence listing.
SSO   PROTEIN; USPTO; APPLICATION
ORGN  Homo Sapiens
SQL   17
SEQ
        1 kssqsvlyns nnkkyla
```

AN 20080069867.4 is SEQ ID NO: 4 from US20080069867.

Published application sequences are typically available within **1 day** of publication.

20080069867.4 is displayed here in **BRIEF** format, which includes the Exemplary Claim (ECLM).

USGENE is an essential additional tool for tackling business critical searches

- DGENE provides curated and indexed patent sequence data from the DWPI *basic* publication
  - 61% of *basics* are WIPO/PCT published applications
  - Updated biweekly, typically 65 days from publication
- USGENE provides all available sequence data from the USPTO as a single merged resource
  - Both **U.S. patents** and **U.S. published applications**
  - Updated weekly, within **3 days** of USPTO publication
- Sequence listing variation often occurs between PCT and U.S. granted patent publication stages
  - Especially important, e.g. for freedom-to-operate

Sequence listing variation often occurs between PCT and U.S. granted patent stage


```
L1   ANSWER 1 OF 1  WPINDEX COPYRIGHT 2008       THOMSON REUTERS on STN
AN   1994-358278 [44]   WPINDEX
TI   New polynucleotide(s) specific for hepatitis C virus types 4, 5 and 6  -
     and related antigenic peptide(s) and antibodies, useful in vaccines,
     diagnosis, HCV typing and treatment
DC   B04; D16; S03
IN   PIKE I H; SIMMONDS P; YAP P L
PA   (COMM-N) COMMON SERVICES AGENCY; (MURE-N) MUREX DIAGNOSTICS INT INC; . . .
PI   WO 9425602       A1 19941110 (199444)* EN  70[5]
     AU 9465797       A  1994
     FI 9505224       A  1995
     EP 698101        A1 1996
     JP 09500009      W  1997
     AU 695259        B  1998
     EP 698101        B1 2004
     DE 69434116      E  2004
     US 20050032047   A1 20050210 (200512)  EN
     US 6881821       B2 20050419 (200527)  EN
     . . . . .
ADT  WO 9425602 A1 WO 1994-GB957 19940505 . . . .
PRAI GB 1994-263 19940107
     GB 1993-9237 19930505
```

In this example the patent family has:

- 9 sequences from WO9425602 in DGENE
- 50 sequences from US20050032047 in USGENE
- 58 sequences from US6881821 in USGENE

STN

FIZ Karlsruhe

Agenda

- Sequence searchable databases on STN
- Introduction to USGENE
- Command line sequence searching
  - BLAST searching
  - Refining BLAST searches with text and date terms
  - BLAST advanced options
  - GETSIM (FASTA) searching
  - Offline BATCH search mode
  - Small fragment searches using GETSEQ

The 7 basic steps of USGENE BLAST

1) SAVE, UPLOAD, and VERIFY the query (L1)
2) RUN the BLAST search (/SQP or /SQN)
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 TRI ORGN SCORE ALIGN 1-
6) Display selected answers in bibliographic
   format, e.g. D L3 BIB AB ECLM ALIGN 1,3,10
7) Ensure transcript was captured and Logoff

The 7 basic steps of USGENE BLAST

Search Question:

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

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

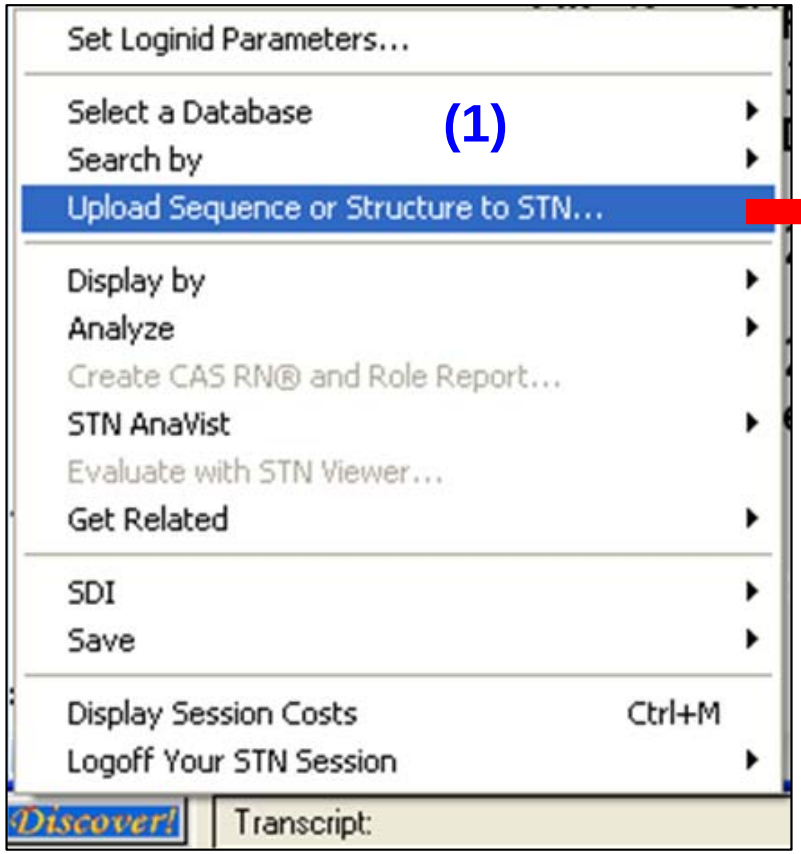
The 7 basic steps of USGENE BLAST

1) SAVE, UPLOAD, and VERIFY the sequence query text file (L1)
   - Upload options
     - STN Express®: Use UPLOAD command or Upload Query Wizard (STN Express 8.x)
     - STN® on the Web℠: Use Upload feature or Sequence Assistant (link below)
   - Verify the sequence with D LQUE

STN on the Web Sequence Search Assistant:
http://www.stn-international.com/training_center/bioseq/seq_se_ass.pdf

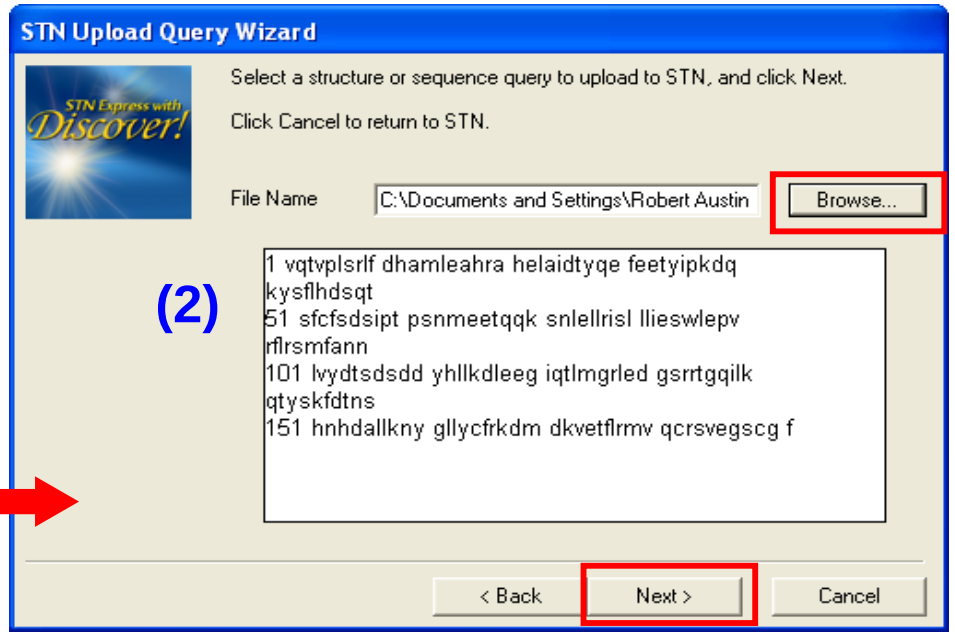
UPLOAD the sequence via STN Express

(1) Click ***Upload Sequence***.
(2) Choose file of interest.
(3) Select database.

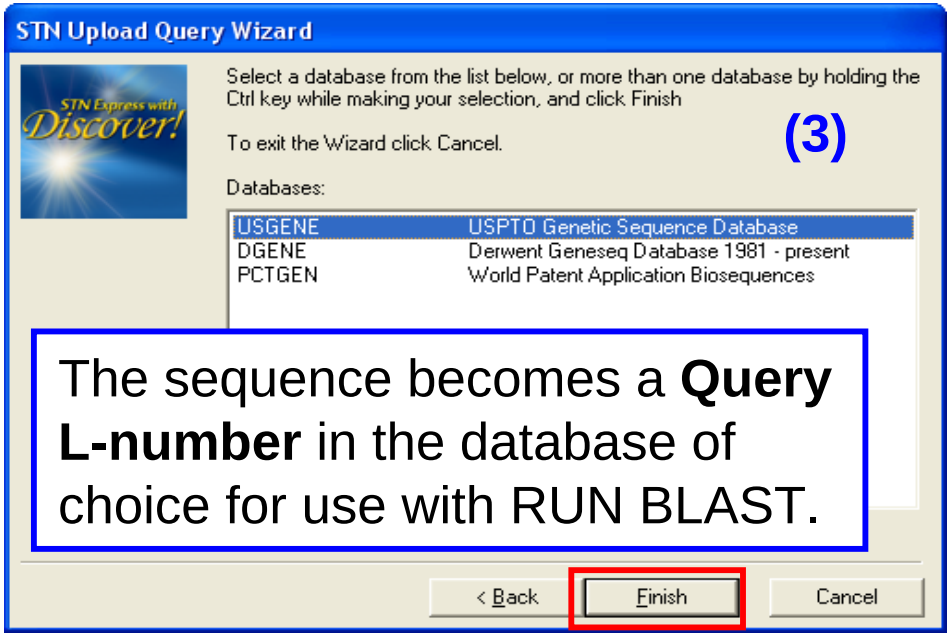


(1)

From the *Discover!* button menu.



(2)



(3)

The sequence becomes a **Query L-number** in the database of choice for use with RUN BLAST.

STN®

FIZ Karlsruhe

1) SAVE, UPLOAD and VERIFY (cont.)


```
=>  FILE USGENE

=>  UPL R BLAST
```

These commands are automatically run by the STN Express Sequence Query Upload wizard.

```
UPLOAD SUCCESSFULLY COMPLETED
L1   GENERATED

=> D L1 LQUE
```

Verify the sequence was uploaded successfully with D LQUE.

```
L1    ANSWER 1  USGENE COPYRIGHT 2008 SEQUENCEBASE CORP on STN
LQUE  vqtvplsrlfdhamleahrahelaidtyqefeetyipkdqkysflhdsqtsfcfsdsi
      ptpsnmeetqqksnlellrisllllieswlepvrflrsmfannlvydtsdsddyhllkd
      leegiqtlmgrledgsrrtgqilkqtyskfdtnshnhdallknygllycfrkdmdkve
      tflrmvqcrsvegscgf

=>
```

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

The 7 basic steps of USGENE 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 USGENE BLAST search


```
=> FILE USGENE

FILE 'USGENE' ENTERED AT 12:09:16 ON 02
COPYRIGHT (C) 2008 SEQUENCEBASE CORP

FILE LAST UPDATED:              2 MAY 2008       <20080502/UP>
MOST RECENT PUBLICATION DATE:   1 MAY 2008       <20080501/PD>

FILE COVERS 1982 TO DATE

>>> SIMULTANEOUS LEFT AND RIGHT TRUNCATION (SLART) IS AVAILABLE
    IN THE BASIC INDEX (/BI) AND FEATURE TABLE (/FEAT) FIELDS <<<

=> 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 . . . .
```

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

Turn the Low Complexity Filter off with the syntax… /SQP –F F

RUN BLAST command syntax

Similarity Searching with BLAST (protein/polypeptides)

=> RUN BLAST L1 (sequence or L-number)

/SQP (protein) (default)

-e (Expect-value)

-f (Filter) (on by default)

-w (Word size)

-m (Matrix)

-g (Gap penalty)

-x (Gap extension)

BATCH (offline)

ALERT (Alert/SDI)

RUN BLAST command syntax

Similarity Searching with BLAST (Nucleotide sequences)

=> RUN BLAST L1 (sequence or L-number)

/SQN (nucleotide)

SIN (single strand)

COM (complementary strand)

BOTH (both strands) (default)

-e (Expect-value)

-f (Filter)

-w (Word size)

-g (Gap penalty)

-x (Gap extension)

-q (penalty for mismatch)

-r (reward for match)

BATCH (offline)

ALERT (Alert/SDI)

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)

Guidelines from NCBI on the use of Advanced Settings for peptide sequence searching are as follows:

Query Length	Matrix	Gap costs
<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.

The 7 basic steps of USGENE 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 IS nnn%)

  ENTER (ALL) OR ? :

- General recommendation: Keep **ALL** answers*

(* Or use BATCH mode to enable multiple retrievals – more on that later in the e-Seminar!)

The 7 basic steps of USGENE 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

New !

```
   2693 ANSWERS FOUND BELOW EXPECTATION VALUE OF 10.0

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

Similarity
Score
      390 |
          |
          |
          |||||||||||||||||||||||||||||||||||||||
          |||||||||||||||||||||||||||||||||||||||||
          |||||||||||||||||||||||||||||||||||||||||||
          |||||||||||||||||||
          |||||||||||||||||||
      195 |||||||||||||||||||
          |||||||||||||||||||
          |||||||||||||||||||
          |||||||||||||||||||
          |||||||||||||||||||||||||||||||||||||||||||||
          |||||||||||||||||||||||||||||||||||||||||||||||
          |||||||||||||||||||||||||||||||||||||||||||||||
Answer Count      540       1080       1620       2160       2700
```

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

The Best Answer Score is also given (in this example there is at least one perfect answer match.)

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.

(Cont . . .)

4) SORT by SCORE descending

New !

```
ENTER EITHER THE NUMBER OF ANSWERS YOU WISH TO KEEP
 OR ENTER MINIMUM PERCENT OF SELF SCORE FOLLOWED BY %
 (BEST ANSWER PERCENTAGE IS 100%)
ENTER (ALL) OR ? : 85%

L2    RUN STATEMENT CREATED
L2         153 VQTVPLSRLFDHAMLEAHRAHELAIDTYQEFEETYIPKDQKYSFLHDSQT
               SFCFSDSIPTPSNMEETQQKSNLELLRISLLLIESWLEPVRFLRSMFANN
               LVYDTSDSDDYHLLKDLEEGIQTLMGRLEDGSRRTGQILKQTYSKFDTNS
               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          153 SOR L2 SCORE D
```

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

Use SORT SCORE D to sort by descending BLAST score.

The 7 basic steps of USGENE BLAST

5) Review answers using a *free-of-charge* format including alignment (ALIGN), while “parked” in the STNGUIDE^{SM} file
   - D L3 TRI ORGN SCORE ALIGN 1-
   - FILE STNGUIDE

New !

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 TRI ORGN SCORE ALIGN 1-153; FILE STNGUIDE

L3    ANSWER 1 OF 153  USGENE COPYRIGHT 2008 SEQUENCEBASE CORP on STN
TI    Recombinant DNA transfer vectors (Patent)
MTY   Protein
SQL   191
ORGN  Unknown
SCORE 390
BLASTALIGN
   Query  = 191 letters
   Length = 191
   Score  =  390 bits (1001), Expect = e-113
   Identities = 191/191 (100%), Positives = 191/191 (100%)
  Query: 1   VQTVPLSRLFDHAMLEAHRAHELAIDTYQEFEETYIPKDQKYSFLHDSQTSFCFSDSIPT
              VQTVPLSRLFDHAMLEAHRAHELAIDTYQEFEETYIPKDQKYSFLHDSQTSFCFSDSIPT
  Sbjct: 1   VQTVPLSRLFDHAMLEAHRAHELAIDTYQEFEETYIPKDQKYSFLHDSQTSFCFSDSIPT
  Query: 61  PSNMEETQQKSNLELLRISLLLIESWLEPVRFLRSMFANNLVYDTSDSDDYHLLKDLEEG
              PSNMEETQQKSNLELLRISLLLIESWLEPVRFLRSMFANNLVYDTSDSDDYHLLKDLEEG
  Sbjct: 61  PSNMEETQQKSNLELLRISLLLIESWLEPVRFLRSMFANNLVYDTSDSDDYHLLKDLEEG
              . . . .
```

New !

100% of query self score 390

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

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

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


```
L3    ANSWER 5 OF 153  USGENE COPYRIGHT 2008 SEQUENCEBASE CORP on STN
TI    Novel antiangiogenic peptide agents and their therapeutic and
      diagnostic use (PublishedApplication)
MTY   Protein
SQL   192
ORGN  Homo Sapiens
SCORE 387      99% of query self score 390
BLASTALIGN
   Query  = 191 letters
   Length = 192
   Score  =  387 bits (995), Expect = e-1
   Identities = 189/191 (98%), Positives
  Query: 1   VQTVPLSRLFDHAMLEAHRAHELAIDTY
              VQTVPLSRLFDHAML+AHRAH+LAIDTYQEFEETYIPKDQKYSFLHDSQTSFCFSDSIPT
  Sbjct: 2   VQTVPLSRLFDHAMLQAHRAHQLAIDTYQEFEETYIPKDQKYSFLHDSQTSFCFSDSIPT
  Query: 61  PSNMEETQQKSNLELLRISLLLIESWLEPVRFLRSMFANNLVYDTSDSDDYHLLKDLEEG
              PSNMEETQQKSNLELLRISLLLIESWLEPVRFLRSMFANNLVYDTSDSDDYHLLKDLEEG
  Sbjct: 62  PSNMEETQQKSNLELLRISLLLIESWLEPVRFLRSMFANNLVYDTSDSDDYHLLKDLEEG
              . . . .
```

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 USGENE 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 IS 100%)
ENTER (ALL) OR ? : 85%
L2    RUN STATEMENT CREATED
L2          153 VQTVPLSRLFDHAMLEAHRAHELAID
                SFCFSDSIPTPSNMEETQQKSNLELL
                LVYDTSDSDDYHLLKDLEEGIQTLMGRLEDGSRRTGQILKQTYSKFDTNS
                HNHDALLKNYGLLYCFRKDMDKVETFLRMVQCRSVEGSCGF/SQP.-F F

Answer set arranged by accession n
similarity score, enter at an arro

=> SOR SCORE D
PROCESSING COMPLETED FOR L2
L3          153 SOR L2 SCORE D

=> S L2 AND SOMATOMAMMOTROPIN/BI,ECLM AND AY<1996 AND GRANTED/SSO
L4            2 L2 AND SOMATOMAMMOTROPIN/BI,ECLM AND AY<1996 AND GRANTED/SSO

=> SOR SCORE D
PROCESSING COMPLETED FOR L4
L5            2 SOR L4 SCORE D
```

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

The BLAST search (L2) is further refined to sequences from granted patents, with application year prior to 1996, and to a specific text search term (L4).

If you limit using text and/or date terms remember to SORT SCORE D again!

The 7 basic steps of USGENE BLAST

6) Display selected relevant answers in a bibliographic format including alignment
   - D L5 BIB AB ECLM SCORE ALIGN 1 5 6
7) Ensure your STN Express session transcript was captured and then logoff

6) Display selected USGENE answers in a preferred bibliographic format


```
=> D BIB AB ECLM ORGN SSO SCORE ALIGN 1-2

L5    ANSWER 1 OF 2  USGENE COPYRIGHT 2008 SEQUENCEBASE CORP on STN
AN    4363877.1  Protein       USGENE
TI    Recombinant DNA transfer vectors (Patent
IN    Goodman Howard M. (San Francisco, CA); Sl
      CA); Seeburg Peter H. (San Francisco, CA
PA    The Regents of the University of Californ
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.

ECLM  US4363877 A: What is claimed is:
      1. A recombinant DNA transfer vector comprising codons for human
      chorionic somatomammotropin comprising the nucleotide  . . . .

ORGN  Unknown
SSO   PROTEIN; EMBL; GRANTED
SCORE 390      100% of query self score 3
BLASTALIGN . . . .
```

This sequence hit comes from a U.S. granted patent, with an application date prior to 1996, and a key concept in the abstract and claims.

Note: this USGENE sequence record, sourced from EMBL, is an example of one which is not indexed in DGENE or REGISTRY.

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 1000 -F F
 . . . .
 1107 ANSWERS FOUND BELOW EXPECTATION VALUE OF 1000.0

      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 IS 100%)
ENTER (ALL) OR ? : ALL
L1    RUN STATEMENT CREATED
L1        1107 GSSFLSPEHQR/SQP.-M PAM30 -W 2 -E 1000 -F F

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

Changing BLAST options is especially important for short sequence queries!

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


```
=> SOR L1 SCORE D
PROCESSING COMPLETED FOR L1
L2          1107 SOR L1 SCORE D

=> D TRI ORGN SCORE ALIGN

L2    ANSWER 1 OF 1107  USGENE COPYRIGHT 2008 SEQUENCEBASE CORP on STN
TI    Antibodies against the PRO1754 polypeptides (Patent)
MTY   Protein
SQL   117
ORGN  Homo Sapiens
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
```

Correct use of BLAST options finds relevant sequence hits.

Reminder: type HELP OPTIONS in USGENE for more information on using BLAST advanced options.

Review: 7 steps of USGENE BLAST

1) SAVE, UPLOAD, and VERIFY the query (L1)
2) RUN the BLAST search (/SQP or /SQN)
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 TRI ORGN SCORE ALIGN 1-
6) Display selected answers in bibliographic format, e.g. D L3 BIB AB ECLM ALIGN 1,3,10
7) Ensure transcript was captured and Logoff

Similarity searching in USGENE using FASTA-based RUN GETSIM

- GETSIM was originally developed by FIZ Karlsruhe for DGENE, and it has since been implemented in both PCTGEN and USGENE
- It is based on the industry standard FASTA methodology, and offers the same basic search modes as BLAST (/SQP, /SQN and /TSQN)
- Since GETSIM requires more computational time than BLAST, it is a usually a good idea to make use of the offline BATCH search mode

RUN GETSIM command syntax

Similarity Searching with GETSIM (protein/polypeptides)

=> RUN GETSIM L1 (sequence or L-number)

/SQP (protein) (default)

BATCH (offline)

ALERT (current awareness)

Note: unlike RUN BLAST, RUN GETSIM does not have any user-defined advanced options to consider. The optimum FASTA search settings are selected automatically by the GETSIM software, depending upon the sequence query.

RUN GETSIM command syntax

Similarity Searching with GETSIM (nucleotide sequences)

=> RUN GETSIM L1 (sequence or L-number)

/SQN (nucleotide)

SIN (single strand) (default)

COM (complementary strand)

BOTH (both strands)

BATCH (offline)

ALERT (current awareness)

Note: to automatically search the nucleotide sequence *and* its complement specify BOTH:

=> RUN GETSIM . . . /SQN BOTH

GETSIM and BLAST similarity searches can both be run offline in BATCH search mode

- Multiple BATCH requests may be queued, to run sequentially one after another
  - A maximum of 16 requests can be queued per STN Login ID
- BATCH request results may be collected in an online session up to 3 months from initiation
  - Results that have been collected may be re-retrieved multiple times at no additional cost, up to 8 days from the initial retrieval
  - For example: multiple times each at a different score percent (%)
- BATCH is most useful for GETSIM queries, as these can take considerable computational time when run online
  - Also a higher query length limit of 2,000 characters is permitted

Similarity searching in USGENE using GETSIM in offline BATCH mode

Search Question:

Find sequences in U.S. published applications and patents which are similar to this specific cholinesterase protein (NCBI: AAA98113):

MPSSVSWGILLLAGLCCLVPVSLAEDPQGDAAQKTDTSHHDQDHPTFNKITPN
LAEFAFSLYRQLASTNIFFSPVSIATAFAMLSLGTKADTHDEILEGLNFNLTE
IPEAQIHEGFQELLRTLNQPDSQLQLTTGNGLFLSEGLKLVDKFLEDVKKLYH
SEAFTVNFGDTEEAKKQINDYVEKGTQGKIVDLVKELDRDTVFALVNYIFFKG
KWERPFEVKDTEEEDFHVDQVTTVKVPMMKRLGMFNIQHCKKLSSWVLLMKYL
GNATAIFFLPDEGKLQHLENELTHDIITKFLENEDRRSASLHLPKLSITGTYD
LKSVLGQLGITKVFSNGADLSGVTEEAPLKLSKAVHKAVLTIDEKGTEAAGAM
FLEAIPMSIPPEVKFNKPFVFLMIEQNTKSPLFMGKVVNPTQK

SAVE, UPLOAD, and VERIFY the query text file for the GETSIM BATCH search


```
=> FILE USGENE

=> UPL R BLAST
```

These commands are automatically run by the STN Express Sequence Query Upload wizard (slides 14-16).

```
UPLOAD SUCCESSFULLY COMPLETED
L1   GENERATED

=> D L1 LQUE

L1    ANSWER 1  USGENE COPYRIGHT 2008 SEQUENCEBASE CORP on STN
LQUE  MPSSVSWGILLLAGLCCLVPVSLAEDPQGDAAQKTDTSHHDQDHPTFNKITPNLAE
      FAFSLYRQLASTNIFFSPVSIATAFAMLSLGTKADTHDEILEGLNFNLTEIPEAQI
      HEGFQELLRTLNQPDSQLQLTTGNGLFLSEGLKLVDKFLEDVKKLYHSEAFTVNFG
      DTEEAKKQINDYVEKGTQGKIVDLVKELDRDTVFALVNYIFFKGKWERPFEVKDTE
      EEDFHVDQVTTVKVPMMKRLGMFNIQHCKKLSSWVLLMKYLGNATAIFFLPDEGKL
      QHLENELTHDIITKFLENEDRRSASLHLPKLSITGTYDLKSVLGQLGITKVFSNGA
      DLSGVTEEAP                                          KPFVF
      LMIEQNTKSP

=>
```

Verify the sequence was uploaded successfully with D LQUE.

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

RUN the GETSIM search in BATCH mode


```
=> FILE USGENE

FILE 'USGENE' ENTERED AT 17:32:27 ON 08
COPYRIGHT (C) 2008 SEQUENCEBASE CORP

FILE LAST UPDATED:             6 JUN 2008       <20080606/UP>
MOST RECENT PUBLICATION DATE: 5 JUN 2008        <20080605/PD>

FILE COVERS 1982 TO DATE

=> RUN GETSIM L1 /SQP BATCH

PLEASE ENTER BATCH IDENTIFIER (MAX. 8 CHARS): EXAMPLE4

RUN GETSIM AT 17:32:48 ON 08 JUN 2008
COPYRIGHT (C) 2008 FIZ KARLSRUHE GMBH

PREVIOUS BATCH REQUEST STILL RUNNING
BATCH PROCESSING QUEUED FOR  EXAMPLE4

=> LOG H
SESSION WILL BE HELD FOR 120 MINUTES
STN INTERNATIONAL SESSION SUSPENDED AT 17:33:06 ON 08 JUN 2008
```

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

Add BATCH for BATCH mode.

Name the BATCH search.

In this example, there is already a BATCH search running, so this request has been Queued.

Use RUN GETBATCH to retrieve and manage the results of BATCH searches


```
* * * * * * RECONNECTED TO STN INTERNATIONAL * *
SESSION RESUMED IN FILE 'USGENE' AT 18:06:27 ON 0
FILE 'USGENE' ENTERED AT 18:06:27 ON 08 JUN 2008

=> RUN GETBATCH
Please enter your batch identifier
    or enter # for batch id list
    or enter * for batch id at top of list
    or enter - before batch id to delete
    or enter . for (end)
BATCH REQUEST: #
Batch result files remaining:
  EXAMPLE1  Retrieved (blast)
  EXAMPLE2  Retrieved (getsim)
  EXAMPLE3  Completed (blast)
  EXAMPLE4  Completed (getsim)
----------------------
Please enter your batch identifier
    or enter # for batch id list
    or enter * for batch id at top of list
    or enter - before batch id to delete
    or enter . for (end)
BATCH REQUEST: EXAMPLE4
```

Login with 2 hours if you want to reconnect to your previous STN session.

Enter # for a BATCH ID list.

BATCH results file status can be: Queued, Running, Completed or Retrieved.

Enter the name of the BATCH search results to retrieve.

Decide how many answers to keep

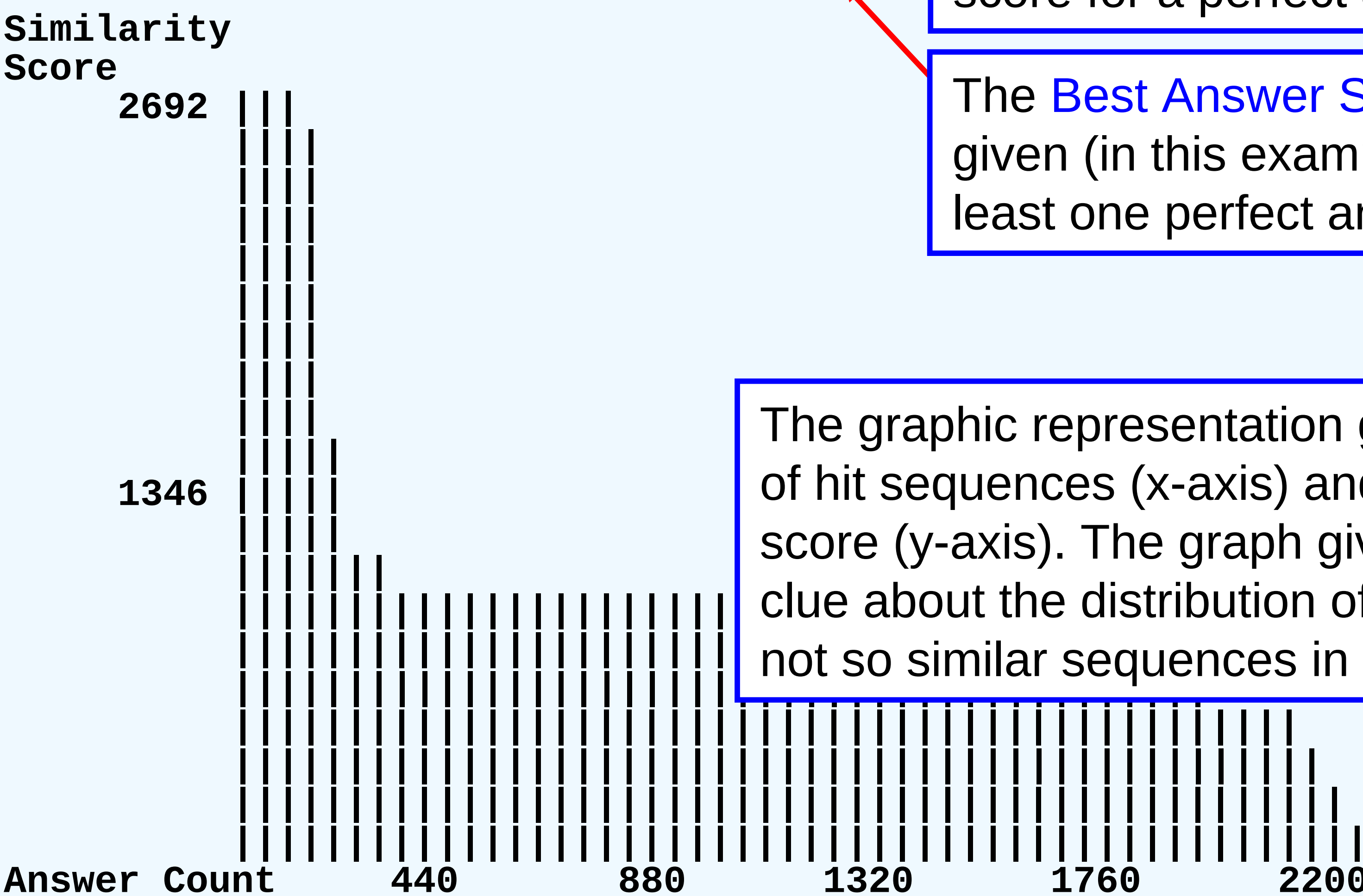
New !

2194 ANSWERS FOUND ABOVE A THRESHOLD OF 172

QUERY SELF SCORE VALUE IS 2692
BEST ANSWER SCORE VALUE IS 2692

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

The Best Answer Score is also given (in this example there is at least one perfect answer match.)



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.

(Cont . . .)

After BATCH collection all search, sort and display options are the same as in online search mode

New !

```
ENTER EITHER THE NUMBER OF ANSWERS YOU WISH TO KEEP
 OR ENTER MINIMUM PERCENT OF SELF SCORE FOLLOWED BY %
 (BEST ANSWER PERCENTAGE IS 100%)
ENTER (ALL) OR ? :80%

L2    RUN STATEMENT CREATED
L2         142 MPSSVSWGILLLAGLCCLVPVSLAE
               TPNLAEFAFSLYRQLASTNIFFSPVSIATAFAMLSLGTKADTHDEILEGL
               NFNLTEIPEAQIHEGFQELLRTLNQPDS
               LEDVKKLYHSEAFTVNFGDTEEAKKQIN
               VFALVNYIFFKGKWERPFEVKDTEEEDF
               HCKKLSSWVLLMKYLGNATAIFFLPDEG
               RRSASLHLPKLSITGTYDLKSVLGQLGI
               KAVHKAVLTIDEKGTEAAGAMFLEAIPM
               SPLFMGKVVNPTQK/SQP

Answer set arranged by accession number; to sort by descending
similarity score, enter at an arrow p

=> SOR SCORE D
PROCESSING COMPLETED FOR L2
L3          142 SOR L2 SCORE D
```

In this example, 80% of the Query Self Score is used to select out just the most relevant results (L2).

Reminder: BATCH results that have been collected, may be re-retrieved multiple times at no additional cost, up to 8 days from the initial retrieval.

As with a BLAST search, the initial GETSIM search answer set should be sorted by similarity score descending, to bring the best answers to the top.

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


```
=> D L3 TRI ORGN SCORE ALIGN 1-142; FILE STNGUIDE

L3    ANSWER 1 OF 142  USGENE COPYRIGHT 2008 SEQUENCEBASE CORP on STN
TI    Inhibitors of serine protease activity, methods and compositions for
      treatment of viral infections (Patent)
MTY   Protein
SQL   414
SCORE 2692        100% of query self score 2692
ORGN  Unknown
ALIGN Smith-Waterman score:  2692
      414 aa overlap starting at 1
      mpssvswgilllaglcclvpvslaedpqgdaaqktdtshhdqdhptfnkitpnlaefafs
      ::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::
      mpssvswgilllaglcclvpvslaedpqgdaaqktdtshhdqdhptfnkitpnlaefafs
       . . . .

L3    ANSWER 142 OF 142  USGENE COPYRIGHT 2(
TI    Anti-Cd28 Antibody (PublishedApplicati
MTY   Protein
SQL   669
SCORE 2344       87% of query self score 269
ORGN  ARTIFICIAL SEQUENCE
ALIGN Smith-Waterman score:  2344
      372 aa overlap starting at 277
      fnkitpnlaefafslyrqla____stniffspvsiata
      ::::::::::::::::::    ::::::::::::
      fnkitpnlaefafslyrqlahqsnstniffspvsiata
       . . . .
```

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

The GETSIM ALIGN display:

- First line: portion of query with similarity
- Second line: similarity (identical- 2 dots, no match- blank, one dot- family match)
- Third line: portion of retrieved sequence with similarity

Display selected USGENE answers in a preferred bibliographic format


```
=> D L3 BIB AB ECLM ORGN SQL SCORE ALIGN

L3    ANSWER 1 OF 142  USGENE COPYRIGHT 2008 SEQUENCEBASE CORP on STN
AN    6849605.19  Protein       USGENE
TI    Inhibitors of serine protease activity, methods and compositions for
      treatment of viral infections (Patent)
IN    Shapiro Leland (Denver, CO)
PA    The Trustees of University Technology Corporation(Boulder CO)
PI    US 6849605          B1   20050201
AI    US 2000-518098           20000303
DT    Patent
AB    A novel method of treating and preventing
      provided. In particular a method of blocki
      facilitated by a serine proteolytic (SP) a
      consists of administering to a subject suf
      from viral infection a therapeutically effective amount of . . .
ECLM  US6849605 B1: What is claimed is:1. A method for inhibiting human
      immunodeficiency virus (HIV) replication in a patient harboring said
      HIV comprising administering to the patient a combination
      comprising:at least one first compound exhibiting a1-antitrypsin
      (AAT)-like protease inhibiting activiity, wherein said compound . . .
ORGN  Unknown
SQL   414
SCORE 2692
ALIGN Smith-Waterman score:  2692
      414 aa overlap starting at 1
      mpssvswgilllaglcclvpvslaedpqgda
      ::::::::::::::::::::::::::::::
      mpssvswgilllaglcclvpvslaedpqgda
```

100% of query self score 2692

USGENE records can be displayed in a wide variety of customized formats.

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

Sequence code match (SCM) searching in USGENE using RUN GETSEQ

- 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
- Remember that an SCM search may also be run in REGISTRY, but the SEARCH (=> S) command is used instead of RUN GETSEQ

RUN GETSEQ command syntax

Sequence Code Match (SCM) searching with GETSEQ

=> 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)

Reminder: USGENE, DGENE and PCTGEN all use the same search command for SCM: **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 2008

SEQ      1 gccgccgt
           ========
HITS AT:   1-8
SQL    8

=> RUN GETSEQ ACCCTGCAAATAGCA/SQSN
L2   RUN STATEMENT CREATED
L2          49 ACCCTGCAAATAGCA/SQSN

=> D L2 30 SEQ SQL
L2 ANSWER 30 OF 49 USGENE COPYRIGHT 2008 SEQUENCEBASE CORP on STN

SEQ      1 tgtagttcat tatcatcttt gtcatcagct gaagatgaaa taggatgtaa
        51 tcagacgaca caggaagcag attctgctaa taccctgcaa atagcaga
                                            ========= ======

HITS AT:   82-96
SQL    98
```

The SEQ display in USGENE shows the entire sequence with the hit nucleic acids underlined and identified by “HITS AT”.

A SUBSEQENCE 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 2008 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 2008 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 SUBSEQENCE 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

=> D L5 2-3 SQL SEQ
L5 ANSWER 2 OF 23  USGENE COPYRIGHT 2008 SE

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

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

=> D L6 73 SEQ SQL
L6 ANSWER 73 OF 2384 USGENE C

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

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

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

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

Amino acid families for RUN GETSEQ SQEFP and SQSFP 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 sequence motif 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 USGENE, DGENE and PCTGEN database summary sheets: **http://www.stn-international.com/stndatabases/databases/onlin_db.html**

Variability symbols for RUN GETSEQ sequence code match searches

Symbol	Function
[ ]	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

Case study: using SCM variability symbols to search USGENE* and REGISTRY

Search Question:

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

LGPX_1QLCX_2LVX_3CAP

X_1 = V or L
X_2 = any amino acid except, G or H
X_3 = any amino acid

(* DGENE and PCTGEN should also be included, but have been omitted simply to save on presentation time.)

RUN GETSEQ SCM search strategy

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

- Possible sequence retrieval
  - *LGP**V**QLC**A**LV**H**CAP*
  - *LGP**V**QLC**S**LV**V**CAP*
  - *LGP**L**QLC**V**LV**A**CAP*
  - *LGP**L**QLC**P**LV**T**CAP*

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

RUN GETSEQ AT 21:42:25 ON 13 MAY 2008
COPYRIGHT (C) 2008 FIZ KARLSRUHE GMBH

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

=> D TRI SEQ

L1    ANSWER 1 OF 32  USGENE COPYRIGHT 2008 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 cappaaqqqq
                    = ========== ===
       51 ngqvfsllsl gsqyqpqrrr dpgaavpgaa nasaqqprtp illirdnrta
          . . . .
      401 rytghhayas gctispy
HITS AT:  10-23
```

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

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

Repeat the USGENE search in REGISTRY and combine all results in CAplus[SM]


```
=> 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 (ANTIBO
               PROSTAT? AND
```

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

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

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

Use USGENE and REGISTRY in combination to locate relevant CAplus records


```
=> D L7 BIB ABS HITIND

L7   ANSWER 1 OF 4  HCAPLUS  COPYRIGHT
AN   2007:463771  HCAPLUS
TI   Detection of tissue-derived glyco
     diagnosis and monitoring of disea
IN   Zhang, Hui; Aebersold, Rudolf H.
PA   Institute for Systems Biology, USA
     . . . .
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)
     . . . .
```

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

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

Summary

- RUN BLAST, RUN GETSIM (FASTA) and RUN GETSEQ (SCM) command line search options are available for DGENE, USGENE and PCTGEN
- A new command line feature to refine BLAST and GETSIM answer sets by percent (%) is now available
- USGENE is a vital tool for business critical patent searches, providing a complete collection of all available U.S. patent and published application sequence data within 3 days of publication by the USPTO
- DGENE, USGENE, PCTGEN and REGISTRY are all required for a comprehensive patent sequence search

Resources for sequence searching on STN

- More on the new percent option for BLAST & GETSIM
  http://www.stn-international.com/New_sequence_search.html
- *Sequence Searching on STN* modular workshop
  http://www.fiz-k.com/bostonsequenceworkshop
  - Sequence Code Match (SCM) searching
  - DGENE, USGENE, PCTGEN content and searching
  - CAS REGISTRY and REGISTRY BLAST
  - Multifile searching using USGENE and DGENE
- USGENE resources, reference materials and FAQ
  http://www.fiz-k.com/usgene
  http://www.sequencebase.com
- CAS REGISTRY sequence coverage and resources
  http://www.cas.org/support/stngen/stndoc/sequences.html

Taking command – effective use of sequence search options in USGENE®, DGENE and PCTGEN

www.stn-international.com