Sequence Basics

Robert Austin – FIZ Karlsruhe
Agenda

• Sequence searchable databases on STN®
• BLAST in DGENE, USGENE® and PCTGEN
• CAS REGISTRY℠ BLAST
• Sequence code match (motif) searching
• Resources

See also: Multifile Patent Sequence Searching on STN e-Seminar:
http://www.stn-international.com/multifile_patent_sequence_search.html
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
CAS REGISTRY/CAplus offers two sequence search modes

- NCBI BLAST similarity
  - Using a separate Graphic User Interface

- Sequence Code Match (motif) searching
  - Using the Search (=> S) command
DGENE, USGENE and PCTGEN offer three sequence search modes

• NCBI BLAST similarity
  => RUN BLAST

• FASTA-based similarity
  => RUN GETSIM

• Sequence Code Match (motif) search
  => RUN GETSEQ

Learn more in the DGENE Workshop Manual:
http://www.stn-international.com/dgene_wm.html
The 7 basic steps of RUN BLAST

1) SAVE, UPLOAD, and VERIFY the query
2) RUN the BLAST search (/SQP, /SQN, /TSQN)
3) Decide how many answers to keep
4) SORT SCORE in Descending order
5) Review answers in a free-of-charge format
   e.g. D L5 TRIAL SCORE ALIGN 1-
6) Display selected answers in bibliographic format, e.g. D L5 BIB ALIGN 1,3,10
7) Ensure transcript was captured before logoff
The 7 basic steps of RUN BLAST

Search Question:
Find relevant references for this human growth hormone sequence (NCBI: AAK69708):

MFPTIPLSRLFDNAMLRAHRLHQLAFDTYQEFEYAYIPEKQKYSFLQNPQTSLCFSSESIPTPSNREETQKSNLELLRISLLLIQSWLEPVQFLRVSFLVYVGAQDSDNYVYDLLKDEEGIGQTLGRLEDGSPRTGQIFKQTYSKFDTNSHNDARLLKNYGLYCVRKDMKVFETFLRIVQCRSVEGSCGF

See also: DGENE Workshop Manual (page 20):
http://www.stn-international.com/dgene_wm.html
Log in to STN using STN Express®
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
1) SAVE, UPLOAD and VERIFY the query (cont.)

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

The sequence becomes a **Query L-number** in the database of choice for use with RUN BLAST.
1) SAVE, UPLOAD and VERIFY the query (cont.)

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

The Query (L1) is uploaded and can be verified with D LQUE.
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

=> 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, Altschul, Stephen F., Thomas L. Madden, Alejandro A. Schaffer, Jinghui Zhang, Zheng Zhang, Webb Miller, and David J. Lipman (1997), "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs." Nucleic Acids Res. 25:3389-3402

Database DGENE DG2_AA

Posted date:  Oct 14, 2011  9:10 AM

. . . .

Note: BLAST searches typically take only seconds to complete (e.g. this one took less than 6 seconds).
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 PERCENTAGE 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 79: http://www.stn-international.com/dgene_wm.html
3) Decide how many answers to keep

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.

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).
3) Decide how many answers to keep (cont.)

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 ? : **ALL**

In this example, ALL of the results are chosen (**L2**).

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

=>
The 7 basic steps of RUN BLAST

4) SORT by SCORE descending

➢ Simply sort the BLAST results answer set:
  => SOR L2 SCORE D   (L3)

➢ Alternatively:
  i. Refine the results using text terms (L3), and/or patent-related terms such as the priority date (L4)
  ii. Then Sort the refined BLAST results answer set
  => SORT L3 SCORE D   (L5)

Tip: Sorting by BLAST local percent identity (IDENT) is also possible.
4) SORT by SCORE descending

Using DGENE value-added text, the BLAST search (L2) is refined to sequences of renal or intestinal interest (L3).

The results (L3) are further refined to sequences with a priority filing date earlier than 2006, and to sequences referred to in the patent claims (L4).

The refined results (L4) are sorted into descending BLAST similarity score order using SORT SCORE D (L5).
Understanding DGENE search fields

• The Basic Index (default search) comprises
  – Enhanced patent title for the invention
  – Concise one-line description of the sequence
  – Controlled-term keyword indexing
  – Enhanced sequence-focused abstract

• Additional search indexes include
  – Patent sequence location (PSL) (claim, example, etc.)
  – Feature table (FEAT) for detailed annotations
  – Publication, application, and priority date or year

Learn more in this recent STN News article (09/2011):
http://www.stn-international.com/dgene_keyword_search.html
The 7 basic steps of RUN BLAST

5) Review answers using a free-of-charge format including alignment (ALIGN), while “parked” in the STNGUIDE℠ file

=> D L5 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

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

=> D L5 TRIAL SCORE ALIGN 1-150; FILE STNGUIDE

L5 ANSWER 1 OF 509 DGENE COPYRIGHT 2011 THOMSON REUTERS on STN
AN AEB96178 protein DGENE
TI Novel mutant polypeptide having O-linked glycosylation site that does not exist in wild-type polypeptide, useful for providing granulocyte colony stimulating factor therapy, hormone therapy and interferon therapy.
DESC Human growth hormone, hGH, polypeptide SEQ ID NO 19.
KW antianemic; nephrotropic; neuroprotective; nootropic; cerebroprotective; anti-HIV; antiinflammatory; virucide; hepatotropic; cytostatic; immunosuppressive; respiratory-gen; muscular-gen; immunomodulator; human growth hormone; hGH; . . . .
SQL 192
SCORE 389
100% of query self score 389
BLASTALIGN
Query: 1 MFPTIPLSRLFDNAMLRAHRLHQALFDTQEFEEAYIPKEQKYSFLQNPQTSLCFSESIP
Sbjct: 1 MFPTIPLSRLFDNAMLRAHRLHQALFDTQEFEEAYIPKEQKYSFLQNPQTSLCFSESIP
Query: 61 TPSNREETQKSNLELLRISLLLIQSWLEPVQFLRSVFANSLVYGASDNSVYDL . . .
Sbjct: 61 TPSNREETQKSNLELLRISLLLIQSWLEPVQFLRSVFANSLVYGASDNSVYDL . . .
Understanding BLAST alignments

<table>
<thead>
<tr>
<th>Term</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Query</td>
<td>the length of the query sequence</td>
</tr>
<tr>
<td>Length</td>
<td>the length of the answer sequence</td>
</tr>
<tr>
<td>Score</td>
<td>a relative score assigned by BLAST</td>
</tr>
<tr>
<td>Expect</td>
<td>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</td>
</tr>
<tr>
<td>Identities</td>
<td>the number of exact letter matches between query and answer within the displayed local alignment. The amino acid letter is repeated* in the display</td>
</tr>
<tr>
<td>Positives</td>
<td>a combination of identities and amino acid family matches shown with + (plus) in the alignment</td>
</tr>
<tr>
<td>Gaps</td>
<td>shown as dashes - where BLAST must break the query or answer to maintain an alignment</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>=&gt; D L5 BIB SCORE ALIGN 1,3,10</th>
</tr>
</thead>
<tbody>
<tr>
<td>L5</td>
</tr>
<tr>
<td>AN</td>
</tr>
<tr>
<td>TI</td>
</tr>
<tr>
<td>IN</td>
</tr>
<tr>
<td>PA</td>
</tr>
<tr>
<td>PI</td>
</tr>
<tr>
<td>AI</td>
</tr>
<tr>
<td>PRAI</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>PSL</td>
</tr>
<tr>
<td>DT</td>
</tr>
<tr>
<td>LA</td>
</tr>
<tr>
<td>OS</td>
</tr>
<tr>
<td>CR</td>
</tr>
<tr>
<td>DESC</td>
</tr>
<tr>
<td>SCORE</td>
</tr>
<tr>
<td>BLASTALIGN</td>
</tr>
</tbody>
</table>
| Query  = 192 letters . . . .

This sequence comes from a WIPO/PCT published patent application, with a priority date earlier than 2006.

The sequence is SEQ ID 19, and appears in Claim 26.
7) Ensure your STN Express session transcript was captured before logoff

**Note**: if you wish to save everything done prior to choosing “Capture Session”, click the “Capture retrospectively” box, before clicking the “Open” button.
RUN BLAST advanced options

Expected 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.
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:

<table>
<thead>
<tr>
<th>Query Length</th>
<th>Matrix</th>
<th>Gap/extension costs</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt;35</td>
<td>PAM-30</td>
<td>(9,1)</td>
</tr>
<tr>
<td>35 – 50</td>
<td>PAM-70</td>
<td>(10,1)</td>
</tr>
<tr>
<td>50 – 85</td>
<td>BLOSUM-80</td>
<td>(10,1)</td>
</tr>
<tr>
<td>&gt;85</td>
<td>BLOSUM-62</td>
<td>(11,1) (BLAST default)</td>
</tr>
</tbody>
</table>

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

NCBI recommended settings* for searching small sequence queries

**Peptide sequences**
- E-value: 20,000
- Word size: 2

**Nucleotide sequences**
- E-value: 1,000
- Word size: 7

- Matrix: PAM-30
- Gap cost: 9 and 1
- Matrix: Leave as is
- Gap cost: n/a

The importance of using the correct BLAST advanced options

=> RUN BLAST TCGCAGGTGACA /SQN
   . . . .
   NO ANSWERS FOUND BELOW EXPECTATION VALUE OF 10.0

=> RUN BLAST TCGCAGGTGACA /SQN -E 1000 -W 7 -F F
   . . . .
   522 ANSWERS FOUND BELOW EXPECTATION VALUE OF 1500.0

   QUERY SELF SCORE VALUE IS    24
   BEST ANSWER SCORE VALUE IS   24

   . . . .

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 ?: ALL

L1   RUN STATEMENT CREATED
L1   522 TCGCAGGTGACA/SQN -E 1000 -W 7 -F F

Changing BLAST options is especially important for short sequence queries.

In this example, ALL of the results are chosen (L1).
The importance of using the correct BLAST advanced options (cont.)

Correct use of BLAST options finds relevant sequence hits.

```
=> SOR L1 SCORE D
PROCESSING COMPLETED FOR L1
L2 522 SOR L1 SCORE D

=> D L2 TRIAL SCORE ALIGN 1

L2 ANSWER 1 OF 522 DGENE COPYRIGHT 2011 THOMSON REUTERS on STN
AN AKW07364 RNA DGENE
TI New functional and hyperfunctional siRNA directed against bcl-2, the
  siRNA comprising an antisense and a sense strand, useful for gene
  silencing.
DESC Human siRNA molecule SEQ ID NO:286895.
KW RNA interference; gene silencing; short interfering RNA; siRNA; ss.
SQL 19
BLASTALIGN
  Query = 12 letters
  Length = 19
  Score = 24.3 bits (12), Expect = 7e-06
  Identities = 12/12 (100%)
  Strand = Plus / Minus

Query: 1 tcgcaggtgaca 12
  "|"""""""""""""""""""""""""
Sbjct: 15 tcgcaggtgaca 4
```

“Plus/Minus” indicates that this is a hit via the BLAST complementary sequence.
Review: 7 basic steps of RUN BLAST

1) SAVE, UPLOAD, and VERIFY the query
2) RUN the BLAST search (/SQP, /SQN, /TSQN)
3) Decide how many answers to keep
4) SORT SCORE in Descending order
5) Review answers in a free-of-charge format, e.g. D L5 TRIAL SCORE ALIGN 1-
6) Display selected answers in bibliographic format, e.g. D L5 BIB ALIGN 1,3,10
7) Ensure transcript was captured before logoff
Agenda

• Sequence searchable databases on STN
• BLAST in DGENE, USGENE and PCTGEN
• CAS REGISTRY BLAST
• Sequence code match (motif) searching
• Recent enhancements
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 – CAplus\textsuperscript{SM}
   - 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

- Sequences can be input by Copy/paste
- Read from a file
- Recall a previously searched sequence within the same session
- Sequence line numbers do not interfere with the search.
Select the BLAST program

The following programs are most typically run:
- BLASTn for nucleotides
- BLASTp for proteins/peptides
Verify BLAST settings

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

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

- 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

- 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 CAplus

=> 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 CAplus. 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 – 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
Agenda

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

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 2011
SEQ 1 gccgccccgt
=====
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 DGENE COPYRIGHT 2011 THOMSON REUTERS on STN
SEQ 1 tgtagttcat tatcatcttt gtcatacgct gaagatgaaa taggatgtaa
tccctgcaa atagcaga
tcagacgaca caggaagcag attctgctaa
HITS AT: 82-96
SQL 98

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

A SUBSEQUENCE search also includes answers which are longer than the query sequence.
EXACT (/SQEP) and SUBSEQUENCE (/SQSP) protein searching

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

=> D L5 2-3 SQL SEQ
L5  ANSWER 2 OF 23  DGENE  COPYRIGHT 2011 THOMSON REUTERS on STN
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  DGENE
SQL     43
SEQ     1 hfrgkfcgki apppvvssgp flfikfvsdy ethgagfsir yei
HITS AT: 33-40

SMAEP/SQEP retrieved 3 records (L3).
SMAEP/SQEP 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/SQSP retrieved 2384 records.
# Amino acid families for RUN GETSEQ SQEFP and SQSFP search options

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

A complete table of all variability symbols, with search examples, is given in the DGENE database summary sheet: http://www.stn-international.com/dgene.html
## Variability symbols for RUN GETSEQ

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Function</th>
</tr>
</thead>
<tbody>
<tr>
<td>[ ]</td>
<td>Specify alternate residues</td>
</tr>
<tr>
<td>[ ]</td>
<td>Exclude a specific residue or alternate residues</td>
</tr>
<tr>
<td>{}</td>
<td>Repeat the preceding symbol(s) (number or range)</td>
</tr>
<tr>
<td>?</td>
<td>Repeat the preceding symbol(s) zero or one time</td>
</tr>
<tr>
<td>*</td>
<td>Repeat the preceding symbol(s) zero or more times</td>
</tr>
<tr>
<td>+</td>
<td>Repeat the preceding symbol(s) one or more times</td>
</tr>
<tr>
<td>^</td>
<td>Query appears at the beginning or the end of a sequence</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td>.</td>
<td>A gap of one residue</td>
</tr>
<tr>
<td>:</td>
<td>A gap of zero or one residues</td>
</tr>
<tr>
<td>&amp;</td>
<td>Concatenate (join together) sequence queries</td>
</tr>
</tbody>
</table>
Search Question:
Find patent references disclosing one or more of the sequences represented by this Markush peptide sequence formula:

\[ \text{LGX}_1\text{LQX}_2\text{CX}_3\text{CAP} \]

\( X_1 = \text{Q or P} \)
\( X_2 = \text{any amino acid except, G or H} \)
\( X_3 = \text{either } -\text{ALVH} - \text{ or } -\text{PLL}\text{R} - \)
RUN GETSEQ SCM search strategy

=> RUN GETSEQ LG[QP]LQ[-GH]C(ALVH|PLLRCAP/SQSP

– Possible sequence retrieval
  
  • LGQLQACALVHCAP
  
  • LGQLQSCPPLLRCAP
  
  • LGPLQVCALVHCAP
  
  • LGPLQPCPLLRCAP
Run the GETSEQ SCM search

=> FILE DGENE

=> RUN GETSEQ LG[QP]LQ[-GH]C(ALVH|PLLR)CAP/SQSP

L1  RUN STATEMENT CREATED
L1          56 LG[QP]LQ[-GH]C(ALVH|PLLR)CAP/SQSP

=> D TRI ALIGN

L1  ANSWER 1 OF 56 DGENE COPYRIGHT 2012 THOMSON REUTERS
AN  AZF00879  protein  DGENE
TI  New isolated nucleic acid molecule useful for producing polypeptide as agent for modulating cellular processes and diagnosing, screening and treating diseases e.g. cancer.
DESC  Human lysyl oxidase (LOX) protein, SEQ ID 14.
KW  LOX protein; antiarthritic; antidiabetic; antiinflammatory; cancer; cytostatic; diabetes mellitus; diagnostic test; dna detection; drug screening; endocrine-gen.; gastrointestinal-gen.; hepatotropic;
hypertension; hypotensive; ischemia; liver fibrosis; lysyl oxidase;
metabolic-gen.; osteoarthritis; osteopathic; osteoporosis; protein
detection; protein production; therapeutic; transgenic animal;
vasotropic; vulnerary; wound healing.
SQL  417
SEQ  lglplqlcalv hcap

HITS AT:  10-23

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

The free-of-charge ALIGN display shows just the hit portion of the answer sequence.
Repeat the DGENE search in REGISTRY and combine all results in CAplusSM

=> FILE REGISTRY
=> S L1
L2 66 LGP[VL]QLC[-GH]LV.CAP/SQSP

=> FIL HCAPLUS

=> S L2 AND P/DT
L3 34 L2 AND P/DT

=> TRA PN L1
L4 TRANSFER L1 1- PN : 37 TERMS
L5 85 L4

=> S L3 OR L5
L6 94 L3 OR L5

=> S L6 AND (ANTIBOD### OR IMMUNOGLOBULIN#) AND DIAGNOS? AND PROSTAT? AND (CANCER? OR TUMOR? OR NEOPLAS?)
L7 5 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 = CAplus patent records found using REGISTRY.
L5 = CAplus patent records found using DGENE.
L6 = CAplus records found using both DGENE and REGISTRY in combination.

The CAplus search may be further refined using CAS value-added abstracts and indexing.
Use DGENE and REGISTRY in combination to locate relevant CAplus records

```plaintext
=> D L7 BIB ABS HITIND 3

L7  ANSWER 3 OF 5  HCAPLUS  COPYRIGHT 2012 ACS on STN
AN  2007:463771  HCAPLUS
DN  146:397152
TI  Detection of tissue-derived glycoproteins in the diagnosis and monitoring of disease
IN  Zhang, Hui; Aebersold, Rudolf H.
PA  Institute for Systems Biology, USA
CODEN: PIXXD2
DT  Patent
LA  English
FAN.CNT 1

<table>
<thead>
<tr>
<th>PATENT NO.</th>
<th>KIND</th>
<th>DATE</th>
<th>APPLICATION NO.</th>
<th>DATE</th>
</tr>
</thead>
<tbody>
<tr>
<td>WO 2007047796</td>
<td>A2</td>
<td>20070426</td>
<td>WO 2006-US40784</td>
<td>20061017</td>
</tr>
</tbody>
</table>

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

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

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

- Sequence searchable databases on STN
- BLAST in DGENE, USGENE and PCTGEN
- CAS REGISTRY BLAST
- Sequence code match (motif) searching
- Resources

See also: Multifile Patent Sequence Searching on STN e-Seminar: http://www.stn-international.com/multifile_patent_sequence_search.html
Resources for sequence searching on STN

- DGENE Workshop Manual
  http://www.stn-international.com/dgene_wm.html
- USGENE Workshop Manual
  http://www.stn-international.com/usgene_wm.html
- CAS REGISTRY sequence searching resources
  http://www.cas.org/support/stngen/stndoc/sequences.html
- Multifile BLAST searching (step-by-step guide)
  http://www.stn-international.com/usgene_wm_mfs.html
Recorded STN e-Seminars are available to watch at your own pace….

- FIZ Karlsruhe recorded e-Seminars:
  http://www.stn-international.com/recorded_events.html
  - Sequence Basics (all databases)
  - Multifile patent sequence searching (all databases)

- CAS recorded e-Seminars:
  http://www.cas.org/support/stngen/stntraining/recorded.html
  - Sequence motif searching (all databases)
  - Processing sequence data (REGISTRY)
  - Unmasking the World of Antibodies (REGISTRY)