GENESEQ – UNIQUE ADDED VALUE

Brian Larner – Solution Consultant
Thomson Reuters
THOMSON REUTERS GENESEQ

• World’s largest source of patented sequences
  – more than 218,000 unique patents from around the world
  – from Thomson Reuters, experts in patent data for 50 years.

• Covers every sequence in a patent
  – even hard-to-find sequences embedded in Tables or Figures so that you get the full picture.

• Keeps pace with increased patenting
  – in traditional geographies and in those seeing significant growth such as China and Korea.
GENESEQ COVERAGE (JUNE 2013)

• All sequences in a patent are captured.

• Around 36 million sequences from 218 000 *basic patents
  – All Nucleic Acids (10 or more bases in length)
  – All Amino Acids (4 or more residues in length)
  – All PCR primers and probes


• From 1981 to the present for all *basic patents

*Basic patent = First member of a Derwent Patent Family to publish.
Highly trained GENESEQ editorial staff extract valuable information from each patent for every GENESEQ record, including:

- Bibliographic data (Enhanced titles, Priority and Pub. data)
- Sequence location within the document.
- Consistently applied keywords
- Original source organism
- Biologically significant regions of the sequence highlighted
- English language abstract, even from non-English sources, to give biological and IP context of each sequence
- Sequence provided in EMBL-like format to ease integration into in-house systems
OTHER USEFUL FIELDS

- **DWPI ACCESION NUMBER (/OS)** – Facilitates the transfer of data between the DGENE file and the main DWPI files (WPINDEX, WPIDS & WPIX or even DPCI for citation data)

- **RELATED RECORDS (/CR)** – Links to related sequence records in the database

- **FEATURE TABLE (/FEAT)** – Describes chemical modifications to the sequence and other important information related to the sequence such as identifying key domains

- **PATENT ASSIGNEE INFORMATION (/PA & /PACO)** – Cleaned assignee data useful for competitive intelligence
## TYPICAL FEATURE TABLE

<table>
<thead>
<tr>
<th>Key</th>
<th>Location</th>
<th>Qualifier</th>
</tr>
</thead>
<tbody>
<tr>
<td>Peptide</td>
<td>1..33</td>
<td>label</td>
</tr>
<tr>
<td>Protein</td>
<td>34..259</td>
<td>&quot;Human PRO221 mature protein&quot;</td>
</tr>
<tr>
<td>Modified-site</td>
<td>37..43</td>
<td>note</td>
</tr>
<tr>
<td>Modified-site</td>
<td>45..51</td>
<td>&quot;N-myristoylation site&quot;</td>
</tr>
<tr>
<td>Modified-site</td>
<td>47..51</td>
<td>&quot;N-glycosylation site&quot;</td>
</tr>
<tr>
<td>Modified-site</td>
<td>94..98</td>
<td>&quot;N-glycosylation site&quot;</td>
</tr>
<tr>
<td>Modified-site</td>
<td>110..116</td>
<td>note</td>
</tr>
<tr>
<td>Modified-site</td>
<td>162..166</td>
<td>&quot;casein kinase II phosphorylation site&quot;</td>
</tr>
<tr>
<td>Modified-site</td>
<td>175..179</td>
<td>&quot;casein kinase II phosphorylation site&quot;</td>
</tr>
<tr>
<td>Modified-site</td>
<td>199..203</td>
<td>&quot;cAMP and cGMP dependent protein kinase phosphorylation site&quot;</td>
</tr>
<tr>
<td>Domain</td>
<td>205..220</td>
<td>&quot;Transmembrane domain&quot;</td>
</tr>
</tbody>
</table>
The Patent Assignee code (/PACO)

• This is a 4 letter code applied by Thomson Reuters to all patents issued by a company including all its regional subsidiaries
  – These codes can be looked up on the Thomson Reuters Website
  – Only larger companies* have meaningful PACO’s
  – Can be used in common with normal text search times to limit results to a particular organisation
  – S GETH/PACO and tyrosine kinase receptor

GETH is the PACO for GENENTECH
GENESEQ CONTENT VS PUBLIC DOMAIN

• GENESEQ captures all sequences in a patent
  – Some Patent offices exclude information from their electronic format publications
  – Our editorial staff capture every sequence
  – Individual records created for mutated sequences

• See Examples given on the following slides...
GENESEQ FILLS THE GAPS

• 2012 Analysis of WIPO Patents – Part 1:
  – 5306 patents had Sequence Listing files published by WIPO on the day of patent publication (ftp://ftp.wipo.int/pub/published_pct_sequences/publication/)
    • 4194 sequence listings were in searchable text format
    • 1112 sequence listings were in PDF or TIFF format.
  – In GENESEQ we published 6491 WO patents for 2012.
GENESEQ FILLS THE GAPS

• 2012 Analysis of WIPO Patents – Part 2:
  – In GENESEQ we published 6491 WO patents for 2012, all with sequences in searchable text format.
  – The extra 1185 patents (18%) without a sequence listing would be inaccessible without GENESEQ’s manual, Intellectual processes.
MANUAL SEQUENCE CAPTURE

- In 2012, GENESEQ captured 18 million sequence characters via our manual indexing approach.
- So far in 2013, more than 10 million characters have been keyed.

<table>
<thead>
<tr>
<th></th>
<th>Total Number of patents</th>
<th>Total Number of characters keyed</th>
</tr>
</thead>
<tbody>
<tr>
<td>2012</td>
<td>1997</td>
<td>17,991,313</td>
</tr>
<tr>
<td>2013 (YTD July)</td>
<td>1127</td>
<td>10,202,005</td>
</tr>
</tbody>
</table>
EXTRA SEQUENCES IN GENESEQ

• Secondary structure sequences
  – For example, WO2013006688 and EP1985709.
  – The sequences in the EP patent are not repeated elsewhere in the patent, and there is no Sequence Listing present.
EXTRA SEQUENCES IN GENESEQ

• Individual records created for mutated sequences
• Individual records are created for claimed SEQ ID fragments where the whole sequence has not been claimed.

What is claimed is:

1. An isolated mutant AlkB enzyme comprising the amino acid sequence of SEQ ID NO: 2 wherein said mutant consists of at least one mutation selected from the group consisting of: V129M, L132V, I233V, and combinations thereof.

2. The isolated mutant AlkB enzyme of claim 1, wherein the mutation is a L132V point mutation.

3. An isolated nucleic acid that encodes a polypeptide with at least 95% identity to the amino acid sequence of SEQ ID NO: 2 and having a point mutation of L132V, wherein the polypeptide can hydroxylate a small chain alkane at the terminal position.

4. The isolated nucleic acid of claim 3, wherein the polypeptide is at least 99% identical to the sequence in SEQ
SEQUENCE SEARCHING IN DGENE

• The following types of sequence searching are available in the DGENE file on STN

• Direct sequence code matching using the GETSEQ command
  – Includes the ability to do family matches and include gaps in a sequence being searched

• Homology searching
  – BLAST searches
  – FASTA searches
Using GETSEQ

• Good for direct sequence code match of peptide and nucleic acid sequences

• Ideal for short and/or highly conserved sequence queries where similarity is not required

• The query L-number can be derived form a previous sequence code match search carried out in PCTGEN or the USGENE file

• System limits
  – 256 characters per line
  – 2000 characters in total (sequences of more than 256 characters must be uploaded form a text file)
GETSEQ for polypeptide sequences

- Exact Sequence Search for Polypeptides (/SQEP) – retrieves sequences that exactly match the search query
- Exact Family Sequence Search of Polypeptides (/SQEFP) – As SQEP but also retrieves hits in which family equivalent substitution is present
- Subsequence Search of polypeptides (/SQSP) – retrieves exact answers plus sequences in which the query sequence is embedded
- Subsequence Family Search of Polypeptides (/SQSFP) – As SQSP plus family member substitution
Families of amino acid equivalents

- P A G S T (neutral, weakly hydrophobic)
- Q N E D B Z (hydrophilic, acid amine)
- H K R (hydrophilic, basic)
- L I V M (hydrophobic)
- F Y W (hydrophobic, aromatic)
- C (Cross-link forming)
GETSEQ for Nucleic Acid Sequences

- Exact sequence search of nucleic acids (/SQEN) – retrieves sequences which exactly match the search query
- Subsequence search of nucleic acids (/SQSN) – retrieves exact answers plus sequences in which the query sequence is embedded
BLAST searches

• BLAST is the most commonly used similarity search tool

• Sequences of up to 10000 characters may be searched (256 characters per line maximum)

• Sequences of more than 256 characters must be entered via the UPLOAD command

• Minimum sequence length 5 characters
BLAST search codes

• To initiate a BLAST search you must specify one of the following search codes
  – /SQP searches for peptide sequences (this is the default so if no option is specified it assumes this is what you want to do)
  – /SQN searches for nucleotide sequences
  – /TSQN for searching a database of peptide sequences translated from DGENE nucleotide sequences

• BLAST searches can be typically completed in one minute or less on line
GETSIM searches

• GETSIM uses a FASTA algorithm
• This is much slower than BLAST and also tends to be less useful for nucleotide sequence searching
• On-line searches are restricted to 500 – 750 characters
• In batch mode a sequence lengths of up to 2000 characters can be run
• Limit of 256 characters per line
• Sequences of 257 or more characters must be entered using the UPLOAD command
# Comparison of BLAST® and FASTA algorithms

## General differences between FASTA and BLAST® algorithms

<table>
<thead>
<tr>
<th></th>
<th>BLAST®</th>
<th>FASTA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Faster than FASTA</td>
<td>Slower than BLAST</td>
<td></td>
</tr>
<tr>
<td>Equivalent for highly similar sequences</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Misses some less similar sequences</td>
<td>Better for less similar sequences</td>
<td></td>
</tr>
<tr>
<td>Comparison of shorter sequence parts</td>
<td>Comparison of entire sequence length</td>
<td></td>
</tr>
<tr>
<td>Less sensitive when using default settings</td>
<td>More sensitive, misses less homologs</td>
<td></td>
</tr>
<tr>
<td>Less separation between true homologs and random hits</td>
<td>More separation between true homologs and random hits</td>
<td></td>
</tr>
<tr>
<td>Calculates probabilities</td>
<td>Calculates significance “on the fly” from the given dataset</td>
<td></td>
</tr>
</tbody>
</table>
Useful URLs

• Web site
  – http://scientific.thomsonreuters.com/

• User guides
  – http://scientific.thomsonreuters.com/training/geneseq/#materials

• Patentee code look up facility
Thank you for listening

• Any questions please contact

• Brian.Larner@thomsonreuters.com

• Tel +44 0207 433 4656

• www.thomsonscientific.com
Exploring GENESEQ™ (DGENE) on STN®

FIZ Karlsruhe
Agenda

• BLAST search
  – Polypeptide BLAST searches
    • BLAST parameters
  – Nucleic acid BLAST search
    • BATCH option
• Sequence Code Match (GETSEQ) search
  – Exact, subsequence and family searches
  – Variability options

See also: Exploring PCTGEN and USGENE® on STN:
http://www.stn-international.com/usgene_seminar.html
Agenda

• BLAST search
  – Polypeptide BLAST searches
    • BLAST parameters
  – Nucleic acid BLAST search
    • BATCH option

• Sequence Code Match (GETSEQ) search
  – Exact, subsequence and family searches
  – Variability options
The 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 basic steps of RUN BLAST

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

MFPTIPLSRLFDNAMLRAHRLHQLAFDITYQEFEEAYIPKEQKYSFLQNPQTSLCEFSESIP
TPSNREETQQKSNLELLRISLLLIIQSWSLEPVQFLRSVFANSLVYGASDNSNVYDLDLKDLEE
GIQTLMGRLEDGSPRTGQIFKQTYSKFDTNSHNDALLKNYGLLYCFRDMDKVEFLRI
VQCRSVEGSCGF

See also: DGENE Workshop Manual (page 20):
http://www.stn-international.com/dgene_wm.html
1) SAVE, UPLOAD and VERIFY the query

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

**Note:** Any spaces, numbers or headers in the query, are stripped out by the STN Express Sequence Upload Query Wizard, prior to uploading to STN. Uploaded sequence queries may be up to 10,000 characters in length.
Log in to STN using STN Express®

Welcome to STN International

NEWS 20 Jul 08  TULSA and TULSA2 have been reloaded with enhancements
NEWS 21 Jul 31  New IFIALL Database on STN Increases US Patent
Retrieval Capabilities
NEWS 22 Jul 31  Find the Most Comprehensive and Timely Results when
Searching the Newly Enhanced Embase Alert(TM)
together with Embase(TM)
NEWS 23 Aug 01  New PV Cluster on STN(R) Simplifies Pharmacovigilance
Alerting and Searching
NEWS 24 Aug 08  DWPI Manual Code Revision - submit your suggestions
NEWS 25 Aug 15  PCTFULL documents with Chinese, Japanese, or Korean as
filling language have English machine translations
NEWS 26 Sep 10  CAS Expands Coverage of Philippines Patents
NEWS 27 Sep 13  STN on the web Enhanced with Updated Structure
and BLAST Plug-ins
NEWS 28 Sep 25  Emtree Thesaurus Updated in Embase
NEWS 29 Sep 30  Application Numbers for U.S. Patents in CA/Caplus
and USPATFULL/USPAT2 Enhanced with U.S. Series
Code Information
NEWS 30 Oct 14  Additional Experimental Spectra Now Available
in CAS REGISTRY on SciFinder and in STN
NEWS EXPRESS  23 MAY 2012 CURRENT WINDOWS VERSION IS V8.5.1,
AND CURRENT DISCOVER FILE IS DATED 22 JULY 2013.
NEWS HOURS  STN Operating Hours
NEWS PRICE  Jan 02, 2013  STN International Fees and Prices,
Effective Jan 1, 2013
FILE 'HOME' ENTERED AT 20:11:34 ON 23 OCT 2013
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.

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.)

**=> FIL DGENE**  
FILE 'DGENE' ENTERED AT 20:12:07 ON 23 OCT 2013  
COPYRIGHT (C) 2013 THOMSON REUTERS

FILE LAST UPDATED:       11 OCT 2013 <20131011/UP>  
MOST RECENT PUBLICATION DATE: 6 SEP 2013 <20130906/PD>

**=> UPL R BLAST**  
Uploading C:\Users\Jim\Desktop\Nov e-Seminars\BLAST sequence.txt

UPLOAD SUCCESSFULLY COMPLETED  
L1   GENERATED

**=> D L1 LQUE**  
L1   ANSWER 1  DGENE  COPYRIGHT 2013 THOMSON REUTERS on STN  
LQUE  MFPTIPLSRLFDNAMLRAHRLHQLAFDTYQEFEEAYIPKEQKYSFLQNPQTSLCFSESI  
      PTPSNREETQQKSNLELLRISLLLIQSWLEPVQFLRSVFANSLVYGASDNSVYDLLKDL  
      EEGIQTLMGRLEDGSPRTGQIFKQTYSKFDTNSHNDALLKNYGLLYCFRKDMKVETF  
      LRIVQCRSVEGSCGF

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 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 2.2.20

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 /data/blastdb/DG2_AA

Posted date: Oct 11, 2013 10:01 AM . . . .

Note: BLAST searches typically take only seconds to complete (e.g. this one took about 8 seconds).
The 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 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).

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 . . .)
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).

L2 RUN STATEMENT CREATED

L2 5490 MFPTIPLSRLFDNAMLRAHRLHQLAFDTYQEFEEAYIPKEQKYSFLQNPQ
TSLCFSESIPTPSNREETQQKSNLELLRISLLLQSWLEPVQFLRSVFAN
SLVYGASDNSVYDLLKDLLEEGIQTLMGRLEDGSPRTGQIFKQTYSKFDTN
SHNDDALLKNYGGLLYCFRKDMKDVETFLRIVQCRSVEGSCGF/SQP.-F

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

=>
The basic steps of RUN BLAST

4) **SORT by SCORE descending**
   - Simply sort the BLAST results answer set:
     
     \[ \Rightarrow \text{SOR L2 SCORE D} \quad (L3) \]

   - **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
     
     \[ \Rightarrow \text{SORT L4 SCORE D} \quad (L5) \]

**Tip:** Sorting by BLAST local percent identity \((\text{IDENT})\) 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 basic steps of RUN BLAST

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

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

L5  ANSWER 1 OF 506  DGENE  COPYRIGHT 2013 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
BLASTALIGN
Query = 192 letters
Length = 192
Score = 389 bits (998), Expect = e-113
Identities = 192/192 (100%), Positives = 192/192 (100%)
Query: 1 MFPTIPLSRLFDNAMLRAHRLHQFLAFDTYQEFEEAYIPKEQKYSFLQNPQTSLCLFSESIP
Sbjct: 1 MFPTIPLSRLFDNAMLRAHRLHQFLAFDTYQEFEEAYIPKEQKYSFLQNPQTSLCLFSESIP
Query: 61 TPSNREETQQKSNLELLRISLLLIQSWLEPVQFLRVSFANSLVYGASDSNVYDL . . .
Sbjct: 61 TPSNREETQQKSNLELLRISLLLIQSWLEPVQFLRVSFANSLVYGASDSNVYDL . . .

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

```plaintext
=> FILE DGENE; D L5 BIB SCORE ALIGN 1,3,10 . . .

L5  ANSWER 1 OF 506  DGENE  COPYRIGHT 2013 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.
IN  Defrees S; Zopf D A; Wang Z; Clausen H
PA  (NEOS-N)  NEOSE TECHNOLOGIES INC.
PI  WO 2005070138 A2 20050804
AI  WO 2005-US799  20050110
PRAI US 2004-535284P  20040108
   US 2004-544411P  20040212
   US 2004-546631P  20040220
   US 2004-555813P  20040323
   US 2004-570891P  20040512
PSL Claim 26; SEQ ID NO 19
DT  Patent
LA  English
OS  2005-597714 [61]
CR  PC-NCBI: gi208475
DESC Human growth hormone, hGH, polypeptide SEQ ID NO 19.
SCORE 389  100% of query self score 389
BLASTALIGN
   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.
Review: 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

• BLAST search
  – Polypeptide BLAST searches
    • BLAST parameters
  – Nucleic acid BLAST search
    • BATCH option

• Sequence Code Match (GETSEQ) search
  – Exact, subsequence and family searches
  – Variability options
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.
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 using Advanced Settings for peptide sequence searching

<table>
<thead>
<tr>
<th>Query Length</th>
<th>Matrix</th>
<th>Gap costs</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt;35</td>
<td>PAM30</td>
<td>(9,1)</td>
</tr>
<tr>
<td>35-50</td>
<td>PAM70</td>
<td>(10,1)</td>
</tr>
<tr>
<td>50-85</td>
<td>BLOSUM80</td>
<td>(10,1)</td>
</tr>
<tr>
<td>&gt;85</td>
<td>BLOSUM62</td>
<td>(11,1) (BLAST default)</td>
</tr>
</tbody>
</table>

**Tip:** Type HELP OPTIONS in USGENE 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

The importance of using the correct BLAST advanced options

=> FILE DGENE

FILE 'DGENE' ENTERED AT 20:50:49 ON 23 OCT 2013
COPYRIGHT (C) 2013 THOMSON REUTERS

FILE LAST UPDATED: 11 OCT 2013 <20131011/UP>
MOST RECENT PUBLICATION DATE: 6 SEP 2013 <20130906/PD>

DGENE CURRENTLY CONTAINS 35,956,153 BIOSEQUENCES

=> RUN BLAST GSSFLSPEHQR/SQP

BLAST Version 2.2.20

... 

NO ANSWERS FOUND BELOW EXPECTATION VALUE OF 10.0

Changing BLAST options is REQUIRED for short sequence queries.
The importance of using the correct BLAST advanced options (cont.)

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

BLAST Version 2.2.20

... 

5718 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 ? : ALL
L1 RUN STATEMENT CREATED
L1 5718 GSSFLSPEHQR/SQP.-M PAM30 -W 2 -E 20000 -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 L1
L2 5718 SOR L1 SCORE D

Changing BLAST options is REQUIRED for short sequence queries.
The importance of using the correct BLAST advanced options (cont.)

=> D TRIAL ORGN SCORE ALIGN

L2  ANSWER 1 OF 5718  DGENE  COPYRIGHT 2013 THOMSON REUTERS on STN
AN  BAR97112 peptide  DGENE
TI  New ghrelin analogs for treating e.g. cachexia, low body weight, loss of body weight, low or decreased appetite, low muscle mass, effects of gastrectomy, anorexia, bulimia, chronic obstructive pulmonary disease and cardiovascular indication.
DESC  Human ghrelin agonist #69.
KW  GHSR1a agonist; Ghrelin ligand agonist; antiarthritic; antiasthmatic; antiinflammatory; appetite loss; arthritis; bulimia nervosa; cachexia; cardiac reperfusion injury; cardiovascular disease; . . .
SQL  28
ORGN  Homo sapiens. Synthetic.
SCORE  38   100% of query self score 38
BLASTALIGN
  Query  = 11 letters
  Length  = 28
  Score  = 37.5 bits (81), Expect = 2e-09
  Identities  = 11/11 (100%), Positives  = 11/11 (100%)
Query: 1  GSSFLSPEHQR 11  
        GSSFLSPEHQR
Sbjct: 1  GSSFLSPEHQR 11

Type HELP OPTIONS in DGENE for more information on using BLAST advanced options.
Agenda

• BLAST search
  – Polypeptide BLAST searches
    • BLAST parameters
  – Nucleic acid BLAST search
    • BATCH option

• Sequence Code Match (GETSEQ) search
  – Exact, subsequence and family searches
  – Variability options
Enhanced batch search options for DGENE

- BATCH searching allows for concurrent searching, saving time and money
- BATCH searching will send an email when the search is completed
  - Available for all three sequence search modes: BLAST, GETSIM (FASTA), and GETSEQ
Example: BLAST search with BATCH option

=> FILE DGENE  

=> RUN BLAST AUACUGUCCCUUUUAAGCA/SQN -E 1000 -W 7 -F F BATCH

PLEASE ENTER BATCH IDENTIFIER (MAX. 8 CHARs): DGENENOV

TO BE NOTIFIED WHEN THIS BATCH SEARCH IS COMPLETE, PLEASE ENTER YOUR EMAIL ADDRESS (MAX. 50 CHARs) OR "NONE"

INPUT: OR (END): JIM.BROWN@FIZ-K.COM

BLAST Version 2.2.20

BATCH PROCESSING STARTED FOR DGENENOV
Email is sent when search has run to completion.
=> FILE DGENE
...
=> RUN GETBATCH DGENENOV
...
4152 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 OF SELF SCORE IS 100%)
ENTER (ALL) OR ? : ALL
L1 RUN STATEMENT CREATED
L1  4152 AUACUGUCCCUUUUAAGCA/SQN.-E 1000 -W 7 -F F

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

Tip: Use RUN GETBATCH #
to see a list of all BATCH
searches and their status.
BLAST search with BATCH option (cont.)

=> SOR SCORE D
PROCESSING COMPLETED FOR L1
L2  4152 SOR L1 SCORE D

=> D TRIAL SCORE ALIGN 1

L2  ANSWER 1 OF 4152  DGENE  COPYRIGHT 2013 THOMSON REUTERS on STN
AN  BAP78389  DNA  DGENE
TI  New oligonucleotide capable of binding with apolipoprotein C-III gene, useful in agent for treating hyperlipidemia and arteriosclerosis, comprises sugar-modified nucleoside.
DESC  Human apolipoprotein C-III (ApoC3) gene, SEQ ID 1.
KW  APOC3 gene; Apolipoprotein C-III; Apolipoprotein C3; antilipemic; antisense therapy; ds; gene silencing; hyperlipidemia; . . .
SQL 533
SCORE 38  100% of query self score 38
BLASTALIGN
   Query  = 19 letters
   Length = 533
   Score  = 38.2 bits (19), Expect = 2e-08
   Identities = 19/19 (100%)
   Strand = Plus / Minus
   Query: 1  atactgtcccttttaagca 19
          |||||||||||||||||||||
   Sbjct: 447 atactgtcccttttaagca 429

"Plus/Minus" indicates that this is a hit via the BLAST complementary sequence.
Agenda

• BLAST search
  – Polypeptide BLAST searches
    • BLAST parameters
  – Nucleic acid BLAST search
    • BATCH option

• Sequence Code Match (GETSEQ) search
  – Exact, subsequence and family searches
  – Variability options
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.

See also: DGENE Workshop Manual (page 101):
http://www.stn-international.com/dgene_wm.html
The RUN GETSEQ command

=> RUN GETSEQ L1 (sequence or query L-number)
   /SQEP (exact protein)
   /SQSP (subsequence protein)
   /SQEFP (exact family protein)
   /SQSFP (subsequence family protein)
   /SQEN (exact nucleotide)
   /SQSN (subsequence nucleotide)
EXACT (/SQEP) and SUBSEQUENCE (/SQSP) protein searching

=> RUN GETSEQ SMAEP/SQEP
L3   RUN STATEMENT CREATED
L3     4 SMAEP/SQEP

=> D L3 1 SQL SEQ
L3  ANSWER 1 OF 3  DGENE  COPYRIGHT 2013 THOMSON REUTERS on STN
SQL   5
SEQ    1 smaep

HITS AT:  1-5

=> RUN GETSEQ KGPSYSLR/SQSP
L4   RUN STATEMENT CREATED
L4     163 KGPSYSLR/SQSP

=> D L4 1 SQL SEQ
L4  ANSWER 1 OF 163  DGENE  COPYRIGHT 2013 THOMSON REUTERS on STN
SQL   88
SEQ
    1 tagkqsslil hgadfstkda dndnomckca lmltggwwfd acgpsnlngm
      51 fytagqnhgk lngikwhyfk gsyslrsstt mmirpldf

HITS AT:  70-77

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

SMAEP/SQEFP retrieved 4 records (L3).
SMAEP/SQEFP retrieved 98 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 163 records (L4).
KGPSYSLR/SQSFP retrieved 6017 records.
EXACT (/SQEN) and SUBSEQUENCE (/SQSN) nucleic acid searching

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.
Agenda

• BLAST search
  – Polypeptide BLAST searches
    • BLAST parameters
  – Nucleic acid BLAST search
    • BATCH option

• Sequence Code Match (GETSEQ) search
  – Exact, subsequence and family searches
  – Variability options
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:

\[ LGX_1LQX_2CX_3CAP \]

\( X_1 = Q \text{ or } P \)
\( X_2 = \text{any amino acid except, G or H} \)
\( X_3 = \text{either } -ALVH- \text{ or } -PLLR- \)
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  58 LG[QP]LQ[−GH]C(ALVH|PLLR)CAP/SQSP

=> D TRI ALIGN 3

L1  ANSWER 1 OF 58 DGENE COPYRIGHT 2013 THOMSON REUTERS on STN
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

<table>
<thead>
<tr>
<th>lgplqlcalv hcap</th>
</tr>
</thead>
</table>

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.
Summary

• BLAST search
  – Polypeptide BLAST searches
    • BLAST parameters
  – Nucleic acid BLAST search
    • BATCH option

• Sequence Code Match (GETSEQ) search
  – Exact, subsequence and family searches
  – Variability options
Resources for sequence searching on STN

• Recorded e-Seminars (audio/video)
  http://www.stn-international.com/recorded_events.html
  – Sequence Basics (all databases)
  – Multifile patent sequence searching (all databases)

• DGENE Workshop Manual
  http://www.stn-international.com/dgene_wm.html
Companion sequence searching on STN e-Seminar

- *Exploring PCTGEN and USGENE on STN*
  Date: November 21, 2013
  Times: 9:00 am EDT and 2:00 pm EDT

- To Register for this e-Seminar, please visit:
For more information …

CAS
E-mail: help@cas.org
Support and Training:
www.cas.org

FIZ Karlsruhe
helpdesk@fiz-karlsruhe.de
Support and Training:
www.stn-international.de