

USGENE

COMPLETE HELP TEXT

© Fachinformationszentrum Karlsruhe, September 2008

Fachinformationszentrum (FIZ) Karlsruhe
Hermann von Helmholtz Platz 1
76344 Eggenstein-Leopoldshafen
Germany

Tel: +49 7247 808 555

Fax: +49 7247 808 131

Email: helpdesk@fiz-karlsruhe.de

Web: www.fiz-karlsruhe.de

COMPLETE HELP TEXT

Contents

INTRODUCTION TO USGENE	4
HELP CONTENT	4
HELP USERAIDS	5
HELP SSEARCH.....	6
HELP DIRECTORY	7
BIOSEQUENCE SEARCHING	8
HELP SIM.....	8
HELP BLAST	9
HELP OPTIONS	12
HELP GSIM.....	19
HELP TLATION.....	23
HELP SBATCH.....	27
HELP SALERT.....	31
HELP GSEQ	35
HELP SQQ.....	38
HELP QLIMITS.....	41
HELP AAC	42
HELP NUC	44
HELP SQL	45
HELP NCBI	46
HELP ALIGNMENT	47
OTHER GENERAL HELP FOR USGENE	49
HELP ACCESSION.....	49
HELP FIELDS	49
HELP SFIELDS	50
HELP SRTFIELDS	51
HELP EFIELDS.....	52
HELP DFIELDS	53
HELP FORMAT	54
HELP CROSSOVER	55
HELP UPDATE/SDI.....	55
HELP RANGE.....	56
HELP HIGHLIGHT	56
HELP (S)	56
HELP USAGETERMS	58
HELP COST.....	60
HELP DESK	61
USGENE VIA <i>STN ON THE WEB</i>	62

Introduction to USGENE

HELP CONTENT

You are currently in the USGENE file. USGENE covers all peptide and nucleic acid sequences from the published applications and issued patents of the United States Patent and Trademark Office (USPTO). The USGENE database includes extensive bibliographic and text search options, including publication title, abstract, patent assignees at issue, full inventor names plus the complete set of publication, application and parent case WIPO/PCT numbers and dates. Data are typically available within 3 days of publication by the USPTO. Both bibliographic information and sequences are fully searchable and displayable.

USGENE offers three advanced sequence searching methods; NCBI BLAST (R), the FastA-based GETSIM, and GETSEQ for fragment or motif sequence queries. Additional biological data, like organism name, molecule type, sequence length and feature table, are also available.

The file covers data from 1982 to date and is updated weekly. Automatic current awareness searches (SDIs) are run weekly.

The basic index (/BI) contains single words from the title (/TI), abstract (/AB), organism species (/ORGN), and molecule type (/MTY) fields.

For a list of additional messages giving information about the USGENE File, enter HELP DIRECTORY at an arrow prompt (=>).

HELP USERAIDS

For a list of HELP messages available in USGENE type HELP DIRECTORY at the command prompt (=>).

For supplementary information about USGENE please refer to the following list of useful user aids.

USGENE database summary sheet:

http://www.stn-international.de/stndatabases/sum_sheet/USGENE.pdf

BLAST(R) information from the National Center for Biotechnology Information (NCBI):

<http://www.ncbi.nlm.nih.gov/Education/BLASTinfo/information3.html>

HELP SSEARCH

Polypeptide and nucleic acid sequence data are searchable and displayable in the USGENE File. The sequences are searchable using three run package options.

RUN BLAST BLAST(R) sequence similarity searching
From the National Center for Biotechnology Information (NCBI)

RUN GETSIM FASTA based sequence similarity searching
From FIZ Karlsruhe GmbH

RUN GETSEQ Sequence Code Match searching
From FIZ Karlsruhe GmbH
Useful for short and/or highly conserved sequence queries

For information on how to use the RUN command, see HELP RUN. For information on using amino acid or nucleic acid codes to retrieve biosequences in the USGENE File, please consult the following help messages:

```
HELP AAC      - table of the 1- and 3-letter codes for common
                amino acids
HELP EFIELDS  - list of codes that may be used in SELECT
HELP GSEQ     - biosequence searching with GETSEQ
HELP NUC      - codes for nucleic acids
HELP QLIMITS  - limits of sequence queries
HELP SIM      - similarity (homology) searching
HELP SQQ      - GETSEQ variability symbols in sequence queries
```

For information on displaying sequences in the USGENE File, please consult the following help messages:

```
HELP DFIELDS  - list of display field codes
HELP FORMAT   - list of pre-defined formats
HELP HIGHLIGHTING - highlighting information
```

For information on the costs for searching and displaying biosequences, enter HELP COST at an arrow prompt (=>).

HELP DIRECTORY

The following HELP messages are available to obtain information on the USGENE file:

```
HELP ACCESSION - USGENE accession number formats
HELP CONTENT   - general USGENE file description
HELP COST      - price schedule for the USGENE file
HELP CROSSOVER - file crossover searching in USGENE
HELP DESK      - information on USGENE file user assistance
HELP EFIELDS   - list of select fields
HELP FIELDS    - list of field and format help messages for
                 the USGENE file
HELP FORMAT    - predefined formats for display and print
HELP HIGHLIGHT - highlighting in the USGENE file
HELP RANGE     - RANGE parameters for the USGENE file
HELP (S)       - (S) operator use
HELP SFIELDS   - list of search field codes
HELP SRTFIELDS - list of sortable fields in the USGENE file
HELP UPDATE/SDI - manual and automatic update searching
HELP USAGETERMS - use and distribution restrictions applicable
HELP USERAIDS  - useful links to supplementary information
                 on USGENE
```

Information about Biosequence Searching:

```
HELP SSEARCH   - Sequence searching in USGENE
HELP SIM       - Sequence similarity (homology) searching
                 (HELP HOMOLOGY)
HELP BLAST     - BLAST sequence similarity searching
HELP OPTIONS   - BLAST advanced user options
HELP GSIM      - GETSIM (FASTA) sequence similarity searching
HELP TLATION   - TSQN translated peptide options
HELP SBATCH    - Offline BATCH similarity search options
HELP SALERT    - Current awareness ALERT for sequence
                 similarity
HELP GSEQ      - GETSEQ Sequence Code Match searching
HELP SQQ       - GETSEQ variability symbols in sequence queries
HELP QLIMITS   - Limits for sequence queries
HELP AAC       - 1- and 3-letter codes for common amino acids
HELP NUC       - Codes for nucleic acids
HELP SQL       - USGENE Sequence Length field
HELP NCBI     - Links to NCBI documentation on BLAST
HELP ALIGNMENT - Alignment of sequences after a similarity
                 Search
```

For a list of more general help topics such as command usage, enter 'HELP MESSAGES' at an arrow prompt (=>).

Biosequence Searching

HELP SIM

There are two standard methods for searching USGENE by sequence similarity (homology):

RUN BLAST BLAST(R) software
From the National Center for Biotechnology Information (NCBI)

RUN GETSIM FASTA based software
From FIZ Karlsruhe GmbH

Enter HELP BLAST or HELP GSIM for information about each option.

Note: GETSIM is based on FASTA methodology, and consequently will often prove to be more sensitive than BLAST, yielding additional hit sequences especially at the lower end of similarity. If you are conducting a comprehensive patent prior-art search, you should consider using both GETSIM and BLAST algorithms to be certain of comprehensive retrieval. Straightforward Sequence Code Match searching is also available in USGENE, which is often useful for short and/or highly conserved sequence queries. Enter HELP GSEQ for further information.

The following help messages contain details about biosequence searching in USGENE:

```
HELP ALIGNMENT
HELP BLAST
HELP GSIM
HELP GSEQ
HELP AAC
HELP QLIMITS
HELP NUC
HELP SSEARCH
HELP SQQ
```

For information on the costs for searching and displaying biosequences, enter HELP COST at an arrow prompt (=>).

HELP BLAST

The BLAST run package is a tool to search the USGENE database for protein and nucleotide sequence data by similarity (homology). It is also possible to search USGENE by similarity using the alternative FASTA-based algorithm (see HELP GSIM).

The BLAST(R) software is provided in USGENE with the permission of the National Center for Biotechnology Information (NCBI) of the National Library of Medicine (NLM). For further information, please refer to:

<http://www.ncbi.nlm.nih.gov/Education/BLASTinfo/information3.html>

RUN BLAST has a series of advanced customisable search settings, including the option to switch from the default search matrix to several others, as provided to FIZ Karlsruhe by the NCBI. See HELP OPTIONS for further information.

To initiate a BLAST search the following search codes have to be specified:

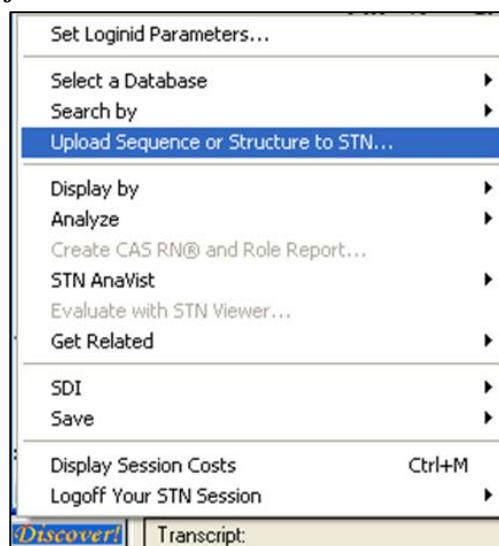
- /SQP** for searching peptide sequences (**BLASTP**) (**default**)
- /SQN** for nucleotide sequences (**BLASTN**)
- /TSQN** for searching a database of peptide sequences translated from USGENE nucleotide sequences (**TBLASTN**)

When BLAST is used online sequences of up to 10,000 characters may be searched. Alternatively, a BLAST search can be run in offline BATCH mode. See HELP SBATCH. Continuously monitoring the patenting of biosequences by BLAST similarity can be conveniently set up with the ALERT feature (see HELP SALERT).

Nucleotide and protein sequences can be subjected to a similarity search involving BLAST in various ways. A query can be prepared with the query command and saved beforehand, it can be entered directly on the command line starting the BLAST package, or it may be uploaded from an ASCII file using the UPLOAD command. Also, the query L-number may derive from a previous sequence search conducted e.g. in the DGENE file.

The minimum length of a sequence query is 5 characters. Sequence queries uploaded from ASCII files (using the UPLOAD command) can be up to 10,000 characters in length. All sequence queries created without using the UPLOAD command have a maximum length of 256 characters (any further characters are ignored). For further information see HELP QLIMITS. Note for STN Express users: to start the upload, select "Online" from the STN Express menu bar and then "Kermit Send". Open the ASCII text file containing the sequence query. Or you may use the Sequence Query Upload Wizard from STN Express version 8.2+ (see figure on the right).

For the Sequence Query Upload Wizard use "Upload Sequence or Structure to STN" from the Discover! button menu:



FIZ Karlsruhe provides a searchable database of peptide sequences which have been translated from USGENE nucleotide sequences. A translation table based on the Universal Genetic Code is used to do this, using all three reading frames of the nucleotide sequences. This translated database is searched when the TSQN option is chosen. The alignment shows the similarity between the query peptide sequence and the translated subject peptide sequence of the answer set. The TSQN search procedure is therefore based on the BLAST peptide homology search algorithm, but the answers retrieved for display are the original USGENE nucleotide sequence records.

When using the SQN or TSQN options it is possible to specify whether single (SIN), complementary (COM) or both (BOTH) strands should be searched. The options can be specified together with the search codes TSQN and SQN, e.g. /SQN COM. If no search option is given, BOTH (both strands) will be used by default. Note that for /TSQN (i.e. /TSQN BOTH) this means that a single polypeptide query will be run six times for the three reading frames of both the single and complementary nucleotide sequences.

Below, an example using the similarity search (SQN) of RUN BLAST for nucleotide sequences is given. A diagram is generated that shows the similarity between the retrieved sequences and the query. The x-axis represents the number of answers with a specific degree of similarity (represented by y-axis). The whole answer set or only the most relevant (at your choice) can be kept. The generated L-number contains these answers, but they are sorted by descending accession number. This L-number may be re-arranged by descending similarity score. Just type "SCORE D" and the corresponding L-number at an arrow prompt.

It is possible to see the alignment between the retrieved sequence and the query sequence with the display format ALIGN. The top line is the query sequence and the bottom line the hit sequence. In this display format the information about the degree of similarity between query sequence and answer subject is indicated as follows: a line represents identical nucleotides, and a blank occurs if there is no match. Gaps inserted in the query or answer sequence for alignment purposes are shown with an underscore. (See HELP ALIGNMENT)

Example : BLAST /SQN search option

```
=> FIL USGENE

=> UPLOAD R BLAST
START LOCAL KERMIT TRANSMIT PROCESS

UPLOAD SUCCESSFULLY COMPLETED
L1 GENERATED

=> D LQUE

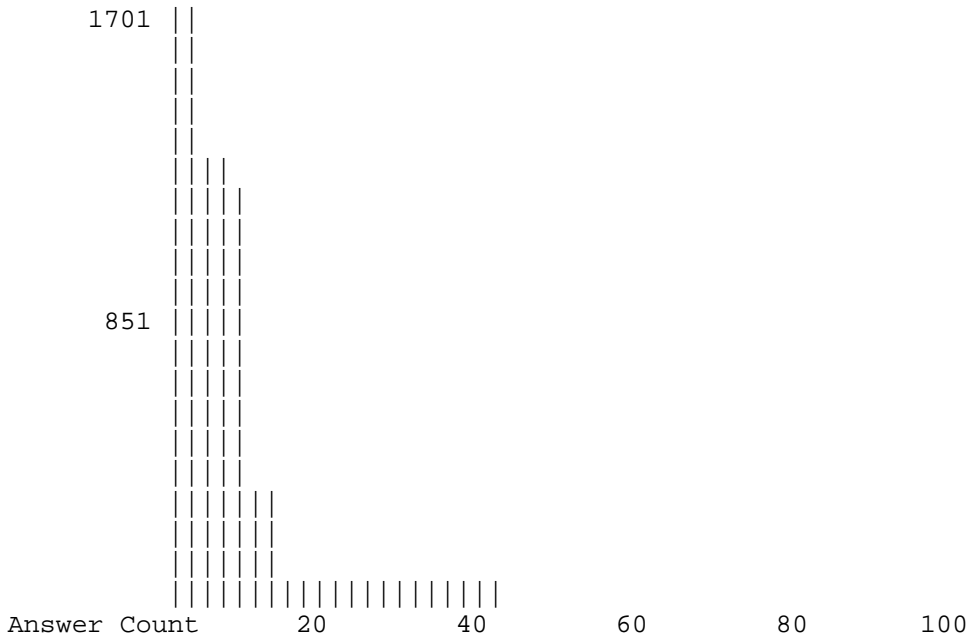
L1 ANSWER 1 USGENE COPYRIGHT 2008 SEQUENCEBASE CORP on STN
LQUE taatgtttag tttattaaca gtaagttcgt atatcaatgt ttagtgctcc
.....
      ggatactctg tttgtgtcct cctgactca ttagatgaag gtgcaaatat

=> RUN BLAST L1/SQN
BLAST Version 2.2
The BLAST software is used herein with permission of the
National Center for Biotechnology Information (NCBI) ...

96 ANSWERS FOUND BELOW EXPECTATION VALUE OF 10.0
```

QUERY SELF SCORE VALUE IS 1701
BEST ANSWER SCORE VALUE IS 1701

Similarity
Score



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      4 TAATGTTTAGTTTATTAACAGTAAGTTCGTATATCAATGTTTAGTGCTCC
      CAAAATTGAAGTTTGAATTTTAAAAGCATCTTGTAGAATTTAGTTGTAT
      .....
      GGATACTCTGTTTGTGTCTTCCCTGACTCATTAGATGAAGGTGCAAATAT
      /SQN. -E 10.0
```

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 4 SOR L2 SCORE D

=> D 1 4 TRIAL SCORE ALIGN

L3 ANSWER 1 OF 4 USGENE COPYRIGHT 2008 SEQUENCEBASE CORP...

TI Staphylococcus aureus polynucleotides and sequences
(PublishedApplication)

MTY nucleic acid

SQL 900

SCORE 1701 100% of query self score 1701

BLASTALIGN

Query = 900 letters

Length = 900

Score = 1701 bits (858), Expect = 0.0

Identities = 886/900 (98%)

Strand = Plus / Plus

```
Query:1 taatgtttagtttattaacagtaagttcgtatatcaatgtttagtgctccccaaaa
      |||
Sbjct:1 taatgtttagtttattaacagtaagttcgtatatcaatgtttagtgctccccaaaa
```

```

Query:61 agtttgaatttttaaagcatcttgtagaatttagttgtannnnnnncaaagaaatt
          ||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct:61 agtttgaatttttaaagcatcttgtagaatttagttgtatttttttcaaagaaatt
          .....
Query:841 tccattagaaggataactctgtttgtgtcttcctgactcattagatgaaggtgca
          ||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct:841 tccattagaaggataactctgtttgtgtcttcctgactcattagatgaaggtgca

```

```

L3      ANSWER 4 OF 4  USGENE COPYRIGHT 2008 SEQUENCEBASE CORP ...
TI      Staphylococcus aureus polynucleotides and sequences(Patent)
MTY     DNA
SQL     900
SCORE  1701          100% of query self score 1701
BLASTALIGN

```

```

Query   = 900 letters
Length  = 900
Score   = 1701 bits (858), Expect = 0.0
Identities = 886/900 (98%)
Strand  = Plus / Plus

```

```

Query:1  taatgtttagtttattaacagtaagttcgtatatcaatgtttagtgctccccaaa
          ||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct:1  taatgtttagtttattaacagtaagttcgtatatcaatgtttagtgctccccaaa
          .....
Query:841 tccattagaaggataactctgtttgtgtcttcctgactcattagatgaaggtgca
          ||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct:841 tccattagaaggataactctgtttgtgtcttcctgactcattagatgaaggtgca

```

.....

The alignment for protein sequences shows the degree of similarity in a third line between the query sequence and the answer subject: Identical amino acids are indicated with the one letter code for the corresponding amino acid, equivalent amino acids (of the same amino acid family) are represented by a plus. No similarity is indicated by a blank. Gaps inserted in the query or answer sequence for alignment purposes are shown with an underscore. (See HELP ALIGNMENT)

HELP OPTIONS

RUN BLAST Advanced User Options

For introductory instructions on using RUN BLAST in USGENE please see HELP BLAST. For the experienced user of BLAST(R), a variety of options is available via the STN command line. Altering these parameters will have a profound effect on the outcome of the search. FIZ Karlsruhe strongly recommends that users are completely familiar with NCBI documentation before embarking on customising any of these settings. For further information:

<http://www.ncbi.nlm.nih.gov/Education/BLASTinfo/information3.html>

See also HELP NCBI

The advanced user options are specified with a single letter code preceded by a hyphen and followed by a blank and the required value, e.g. RUN BLAST L1/SQN -E 0.1.

Advanced User Options

Option	Switch	Values
1. Filter	-f	Values: T (true), F (false), C (default value is T). If T is set, for peptides the SEG, and for nucleotides the DUST filter is employed. C symbolises the 'coiled' filter for nucleotides.
2. Expectation Value	-e	Values: floating point number (default is 10)
3. Word Size	-w	Values: 11 (default) or 7-23 for nucleotides 3 (default) or 2 for peptides
4. Strand	-s	Values: 1 (sin), 2 (com) or 3 (both) default value is 3 (both)
5. Matrix	-m	Values: BLOSUM62 (default), BLOSUM80, BLOSUM45, PAM30 or PAM70
6. Gap Penalty	-g	Default Values: 11 (peptides) 5 (nucleotides)
7. Gap Extension	-x	Default Values: 1 (peptides) 2 (nucleotides)
8. Penalty for nucleotide mismatch	-q	Default Value: -3
9. Reward for nucleotide match	-r	Default Value: 1

Matrix settings (for option 5.)

Please note that for a certain matrix only a restricted set of possible gap and gap extension values is possible. The settings available to each matrix are summarised in the table below. Default settings are indicated in the table. Any different combinations will be rejected by the system and a warning message issued.

Matrix	Gap	Gap Extension
BLOSUM62	9	2
	8	2
	7	2
	12	1
	11	1 (default)
	10	1
BLOSUM80	8	2
	7	2
	6	2
	11	1
	10	1 (default)
	9	1

```

-----
BLOSUM45    13      3
             11      3
             12      3
             9       3
             15      2 (default)
             14      2
             13      2
             12      2
             19      1
             18      1
             17      1
             16      1

```

```

-----
PAM30       7       2
             6       2
             5       2
            10      1
             8       1
             9       1 (default)

```

```

-----
PAM70       8       2
             7       2
             6       2
            11      1
            10      1 (default)
             9       1

```

Example: a short peptide search with the expectation value increased from the default

```
=> RUN BLAST GLYSPNDIAVLSQER/SQP -M PAM30 -W 2 -E 1000
```

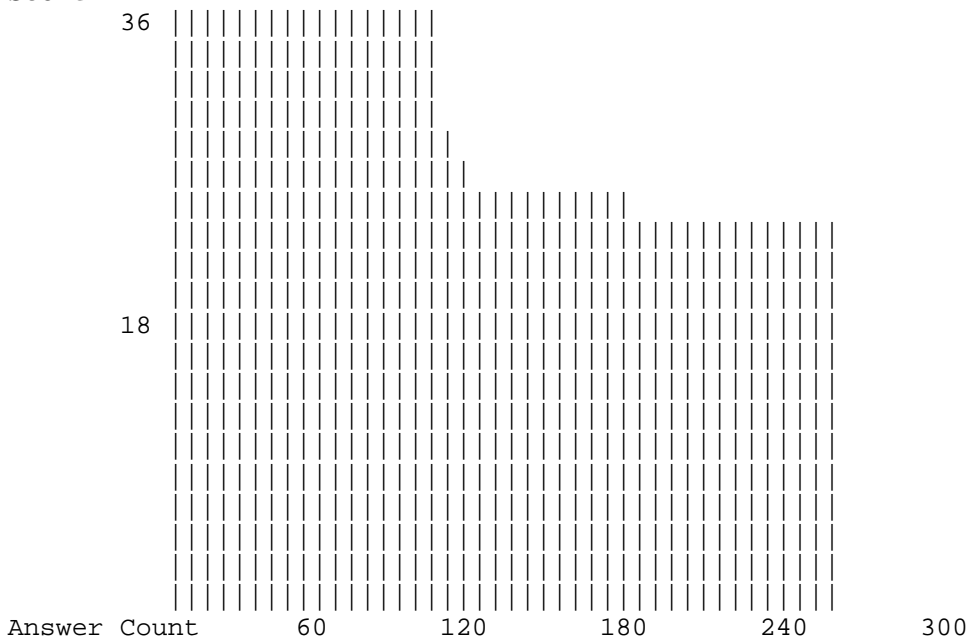
```
BLAST Version 2.2
```

```
The BLAST software is used herein with permission of the  
National Center for Biotechnology Information (NCBI) . . .
```

```
251 ANSWERS FOUND BELOW EXPECTATION VALUE OF 1000.0
```

```
QUERY SELF SCORE VALUE IS    50  
BEST ANSWER SCORE VALUE IS   36
```

```
Similarity  
Score
```



```
ENTER EITHER THE NUMBER OF ANSWERS YOU WISH TO KEEP  
OR ENTER MINIMUM PERCENT OF SELF SCORE FOLLOWED BY %  
(BEST ANSWER PERCENTAGE IS 72%)
```

```
ENTER (ALL) OR ? :70%
```

```
L3 RUN STATEMENT CREATED
```

```
L3 101 GLYSPNDIAVLSQER/SQP.-M PAM30 -W 2 -E 1000
```

```
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 L3
```

```
L4 101 SOR L3 SCORE D
```

```
=> D 1 101 TRIAL SCORE ALIGN
```

```
L4 ANSWER 1 OF 101 USGENE COPYRIGHT 2008 SEQUENCEBASE CORP...
```

```
TI Peptides for treatment and diagnosis of bone diseases  
(PublishedApplication)
```

```
MTY Protein
```

```
SQL 1614
```

```
SCORE 36 72% of query self score 50
```

```
BLASTALIGN
```

```
Query = 15 letters
```

```
Length = 1614
```

```
Score = 36.3 bits (78), Expect = 1e-07
Identities = 12/14 (85%), Positives = 12/14 (85%)
Query: 2   LYSPNDIAVLSQER 15
          LYSP DI VLSQER
Sbjct: 277 LYSPMDIQVLSQER 290
```

```
L4      ANSWER 101 OF 101 USGENE COPYRIGHT 2008 SEQUENCEBASE CORP...
TI      LDL-receptor (Patent)
MTY     Protein
SQL     1615
SCORE 36      72% of query self score 50
BLASTALIGN
  Query = 15 letters
  Length = 1615
  Score = 36.3 bits (78), Expect = 1e-07
  Identities = 12/14 (85%), Positives = 12/14 (85%)
  Query: 2   LYSPNDIAVLSQER 15
            LYSP DI VLSQER
  Sbjct: 278 LYSPMDIQVLSQER 291
```

Example: the same short peptide search (see above), but using default BLAST options finds no answers. Correct use of BLAST options is essential.

=> FILE USGENE

=> RUN BLAST GLYSPNDIAVLSQER/SQP

BLAST Version 2.2

The BLAST software is used herein with permission of the National Center for Biotechnology Information (NCBI) . . .

NO ANSWERS FOUND BELOW EXPECTATION VALUE OF 10.0

Note: For the calculation of the query self score value all parameters changed with the BLAST search will be applied. This means that each parameter changed from the default may also affect the query self score value.

Example: a nucleotide search with e-value decreased from the default

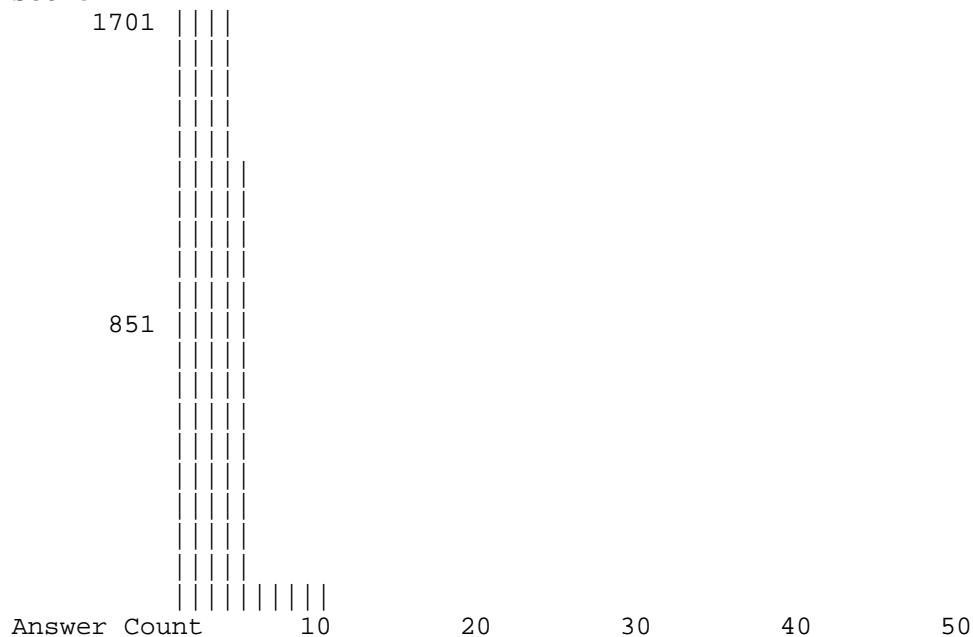
```
=> RUN BLAST L7/SQN -E 0.1
```

```
BLAST Version 2.2
```

```
The BLAST software is used herein with permission of the .....
```

```
10 ANSWERS FOUND BELOW EXPECTATION VALUE OF 1.0e-01
```

```
Similarity  
Score
```



```
ENTER EITHER THE NUMBER OF ANSWERS YOU WISH TO KEEP  
OR ENTER MINIMUM PERCENT OF SELF SCORE FOLLOWED BY %  
(BEST ANSWER PERCENTAGE IS 72%)
```

```
ENTER (ALL) OR ? :ALL
```

```
L8 RUN STATEMENT CREATED
```

```
L8 10 TAATGTTTAGTTTATTAACAGTAAGTTCGTATATCAATGTTTAGTGCTCC
```

```
.....  
GGATACTCTGTTTGTGTCTTCCCTGACTCATTAGATGAAGGTGCAAATAT  
/SQN.-E 0.1
```

```
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 L8
```

```
L9 10 SOR L8 SCORE D
```

```
=> D BRIEF ALIGN
```

```
L9 ANSWER 1 OF 10 USGENE COPYRIGHT 2007 SEQUENCEBASE CORP on STN
```

```
AN 20070020746.327 nucleic acid USGENE
```

```
TI Staphylococcus aureus polynucleotides and sequences  
(PublishedApplication)
```

```
IN Kunsch Charles A. (Norcross, GA); Choi Gil H. (Rockville, MD); Barash  
Steven C. (Rockville, MD); Dillon Patrick J. (Carlsbad, CA); Fannon  
Michael R. (Silver Spring, MD); Rosen Craig A. (Laytonsville, MD)
```

PI US 20070020746 A1 20070125
AI US 2004-807556 20040324
ED 20070331
DT Patent

AB The present invention provides polynucleotide sequences of the genome of Staphylococcus aureus, polypeptide sequences encoded by the polynucleotide sequences, corresponding polynucleotides and polypeptides, vectors and hosts comprising the polynucleotides, and assays and other uses thereof. The present invention further provides polynucleotide and polypeptide sequence information stored on computer readable media, and computer-based systems and methods which facilitate its use.

ECLM US20070020746 A1: 1. An isolated protein-encoding nucleic acid fragment of the Staphylococcus aureus genome, wherein said fragment consists of the nucleotide sequence of any one of the fragments of SEQ ID NOS:1-5,191 depicted in Tables 2 and 3, or a degenerate variant thereof.

SSO NUCLEIC; PSIPS; APPLICATION

ORGN Not provided

SQL 900

SEQ

```
1 taatgttttag tttattaaca gtaagttcgt atatcaatgt ttagtgctcc
51 ccaaaattga agtttgaatt taaaagcat cttgtagaat ttagttgtat
.....
801 cagcagcact actttcagca gggcttaaca gagaaaaatc tccattagaa
851 ggatactctg tttgtgtctt ccctgactca ttagatgaag gtgcaaatat
```

BLASTALIGN

```
Query = 900 letters
Length = 900
Score = 1701 bits (858), Expect = 0.0
Identities = 886/900 (98%)
Strand = Plus / Plus
```

```
Query: 1 taatgttttagtttattaacagtaagttcgtatatcaatgttttagtgctccccaaaattga
|||||
Sbjct: 1 taatgttttagtttattaacagtaagttcgtatatcaatgttttagtgctccccaaaattga
.....
Query: 841 tccattagaaggatactctgtttgtgtcttccctgactcattagatgaagggtgcaaatat
|||||
Sbjct: 841 tccattagaaggatactctgtttgtgtcttccctgactcattagatgaagggtgcaaatat
```

HELP GSIM

The GETSIM run package is a tool to search the USGENE database for protein and nucleotide sequence data by similarity (homology). GETSIM is provided in USGENE by FIZ Karlsruhe GmbH and is based upon the FASTA algorithm. It is also possible to search USGENE by similarity using the alternative BLAST algorithm (see HELP BLAST).

To initiate a GETSIM search the following search codes have to be specified:

- /SQP** for searching peptide sequences (**default**)
- /SQN** for nucleotide sequences
- /TSQN** for searching a database of peptide sequences translated from USGENE nucleotide sequences

When GETSIM is used online sequences of up to 500 or 750 characters may be searched (500 characters for nucleotides and 750 for peptides). Alternatively, a GETSIM search can be run in offline BATCH mode where the query limit for the sequence length is raised to 2000 characters. See HELP QLIMITS and also HELP SBATCH. Continuously monitoring the patenting of biosequences by GETSIM similarity can be conveniently set up with the ALERT feature (see HELP SALERT).

Nucleotide and protein sequences can be subjected to a similarity search involving GETSIM in various ways. A query can be prepared with the query command and saved beforehand it can be entered directly on the command line starting the GETSIM package, or it may be uploaded from an ASCII file using the UPLOAD command. Also, the query L-number may derive from a previous sequence search conducted e.g. in the DGENE file or the CAS REGISTRY file.

The minimum length of a sequence query is 5 characters. Sequence queries uploaded from ASCII files (using the UPLOAD command) can be up to 500 or 750 characters in length (500 for nucleotides, 750 for peptides). All sequence queries created without using the UPLOAD command have a maximum length of 256 characters (any further characters are ignored). For further information see HELP QLIMITS. Note for STN Express users: to start the upload, select "Online" from the STN Express menu bar and then "Kermit Send". Open the ASCII text file containing the sequence query. Or you may use the Sequence Query Upload Wizard from STN Express version 8.2+ (see figure on the right).

For the Sequence Query Upload Wizard use "Upload Sequence or Structure to STN" from the Discover! button menu:



FIZ Karlsruhe provides a searchable database of peptide sequences which have been translated from USGENE nucleotide sequences. A translation table based on the Universal Genetic Code is used to do this, using all three reading frames of the nucleotide sequences. This translated database is searched when the TSQN option is chosen. The alignment shows the similarity between the query peptide sequence and the translated subject peptide sequence of the answer set.

The TSQN search procedure is therefore based on the peptide homology search algorithm, but the answers retrieved for display are the original USGENE nucleotide sequence records.

When using the SQN or TSQN options it is possible to specify whether single (SIN), complementary (COM) or BOTH strands should be searched. The options can be specified together with the search codes TSQN and SQN, e.g. /SQN COM. If no search option is given, SIN (single) will be used by default. Note that for /TSQN BOTH this means that a single polypeptide query will be run six times for the three reading frames of both the single and the complementary nucleotide sequences.

Below, an example using the similarity search (SQP) of RUN GETSIM for amino acid sequences is given. A diagram is generated that shows the similarity between the retrieved sequences and the query. The x-axis represents the number of answers with a specific degree of similarity (represented by y-axis). The whole answer set or only the most relevant (at your choice) can be kept. The generated L-number contains these answers, but they are sorted by descending accession number. This L-number may be re-arranged by descending similarity score. Just type "SOR SCORE D" and the corresponding L-number at an arrow prompt.

It is possible to see the alignment between the retrieved sequence and the query sequence with the display format ALIGN. The top line is the query sequence and the bottom line the hit sequence. In this display format a line between the two sequences gives the information about the degree of similarity: two dots represent identical nucleotides/peptides, and a blank occurs if there is no match. One dot indicates a chemical "family" match. Gaps inserted in the query or answer sequence for alignment purposes are shown with an underscore. (See HELP ALIGNMENT)

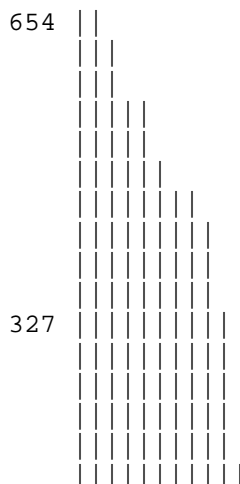
Example : GETSIM /SQP search option

```
=> RUN GETSIM ccldpqrvaikslterlyvggpltnsrgencgyrrcrasgvlttscgntltcyikaraa  
craagldctmlvcgddlvcicesagvqedaaslrifte/sqp
```

```
RUN GETSIM AT 13:24:41 ON 23 MAY 2008  
COPYRIGHT (C) 2008 FIZ KARLSRUHE GMBH  
.....
```

```
4476 ANSWERS FOUND ABOVE A THRESHOLD OF 66  
  QUERY SELF SCORE VALUE IS 656  
  BEST ANSWER SCORE VALUE IS 654
```

Similarity
Score



HELP TLATION

With both homology (similarity) search options (RUN GETSIM and RUN BLAST) a translated search is possible with /TSQN (see HELP GSIM and HELP BLAST). Via this search a peptide query sequence can be searched against a database of translated nucleotide sequences. For this search option FIZ Karlsruhe provides a searchable database of peptide sequences which have been translated from USGENE nucleotide sequences. A translation table based on the Universal Genetic Code is used to do this, using all three reading frames of the corresponding nucleotide sequences. This translated database is searched when the TSQN option is chosen. The alignment shows the similarity between the query peptide sequence and the translated subject peptide sequence of the answer set. The TSQN search procedure is therefore based on the peptide homology search algorithm, but the answers retrieved for display are the original USGENE nucleotide sequence records.

When using the SQN or TSQN options in homology search it is possible to specify whether the single strand (SIN), the complementary strand (COM) or both strands (BOTH) should be searched. For specification of strands in homology search with BLAST see HELP OPTIONS. These search options are used together with the search codes TSQN and SQN, e.g. /TSQN COM. Note that if no search option is given the defaults for Getsim search and Blast search are different. In Getsim translated search SIN (single) will be used by default whereas in Blast translated search BOTH (both) is the default setting. For /TSQN BOTH a single polypeptide query will be run six times for the three reading frames of both the single and the complementary nucleotide sequences.

Below, examples using the translated search of both homology search options (RUN GETSIM and RUN BLAST) for a peptide query sequence are given. A diagram is generated that shows the similarity between the retrieved (translated) sequences and the query sequence. The x-axis represents the number of answers with a specific degree of similarity (represented by y-axis). The whole answer set or only the most relevant (at your choice) can be kept. The generated L-number contains these answers, but they are sorted by descending accession number. This L-number may be re-arranged by descending similarity score. Just type "SOR SCORE D" and the corresponding L-number at an arrow prompt.

It is possible to see the alignment between the retrieved sequence from the translated database and the query sequence with the display format ALIGN. See HELP GSIM, HELP BLAST and HELP ALIGNMENT for more information.

Example : GETSIM /TSQN search option

```
=> RUN GETSIM LPKELLRRIFSFLDIVTLRCRAQISKAWNILALDGSNW/TSQN BOTH
```

```
RUN GETSIM AT 15:19:11 ON 23 MAY 2008  
COPYRIGHT (C) 2008 FIZ KARLSRUHE GMBH
```

```
320000 SEQUENCES PROCESSED
```

```
.....
```

```
54050000 SEQUENCES PROCESSED
```

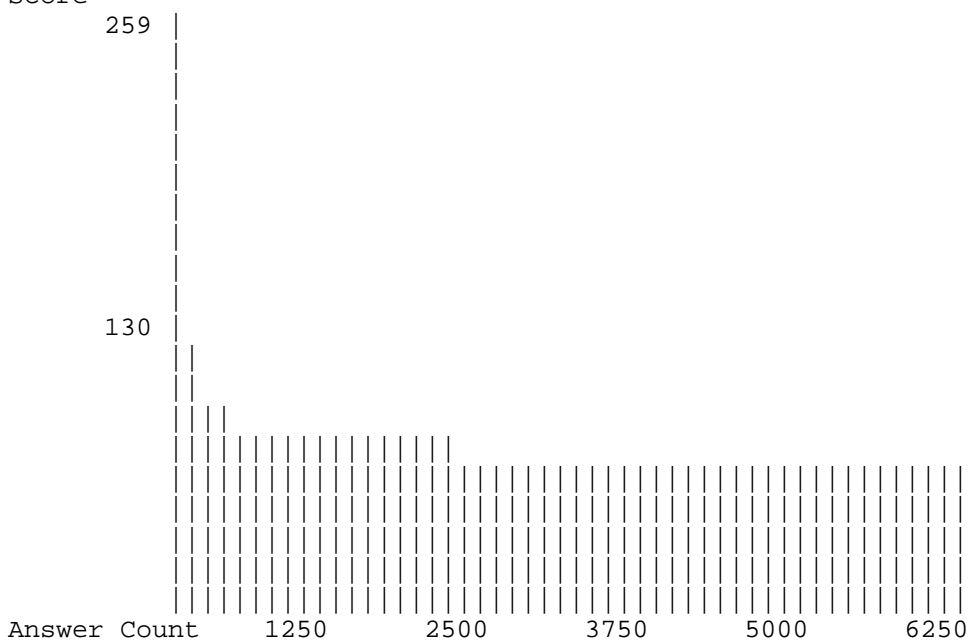
```
6223 ANSWERS FOUND ABOVE A THRESHOLD OF 63
```

```
QUERY SELF SCORE VALUE IS 259
```

```
BEST ANSWER SCORE VALUE IS 259
```

Similarity

Score



```
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%
```

```
L4 RUN STATEMENT CREATED
```

```
L4 39 LPKELLRRIFSFLDIVTLRCRAQISKAWNILALDGSNW/TSQN.BOTH
```

```
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 L4
```

```
L5 39 SOR L4 SCORE D
```

```
=> D 1 TI SCORE ALIGN SEQ
```

```
L5 ANSWER 1 OF 39 USGENE COPYRIGHT 2008 SEQUENCEBASE CORP...
```

```
TI Novel nucleic acids and polypeptides  
(PublishedApplication)
```

```
SCORE 259 100% of query self score 259
```

```
ALIGN Smith-Waterman score: 259
```

```
38 aa overlap starting at 30
```

```

(Frame 1 - 114 na overlap starting at 88)
lpkelllrifsfldivtlcrcaqiskawnilaldgsnw
:.....:
lpkelllrifsfldivtlcrcaqiskawnilaldgsnw
SEQ
  1 tttcgtgtga cttcgggctg tgggctcgct cgcggctctt cggccatggt
 51 tttctcaaac aatgatgaag gccttattaa caaaaagtta cccaaagaac
  .....
2301 ttagactcat aaaattgaat aaaccgattg caatgcttta aaaaaaatta
2351 aaaaaaaaa

```

Example : BATCH /TSQN search option

```
=> RUN BLAST LPKELLRIFSFLDIVTLCRCAQISKAWNILALDGSNW/TSQN
```

```
BLAST Version 2.2
```

```
The BLAST software is used herein with permission of the
National Center for Biotechnology Information (NCBI) of .....
```

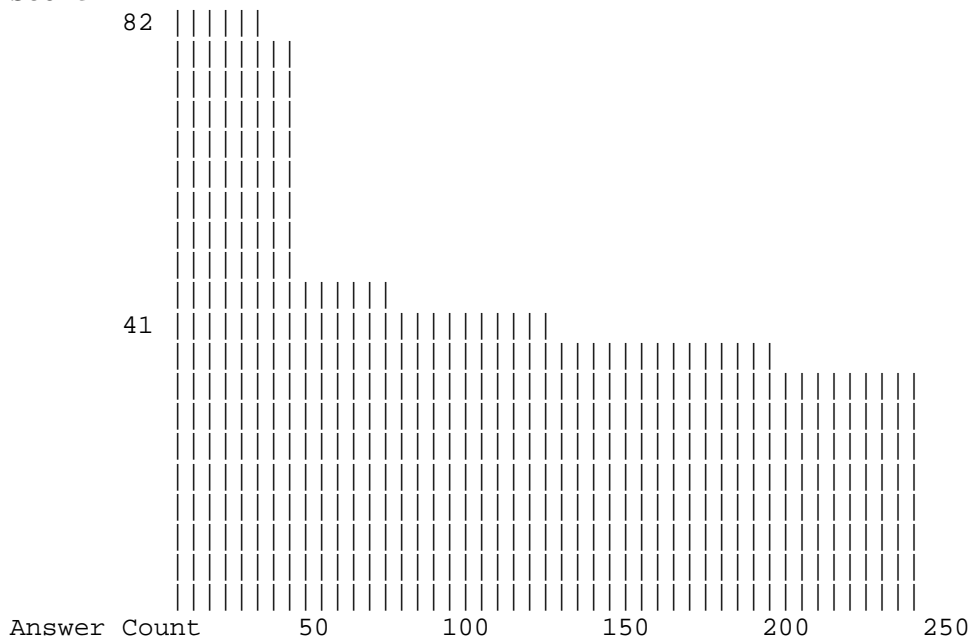
```
233 ANSWERS FOUND BELOW EXPECTATION VALUE OF 10.0
```

```

QUERY SELF SCORE VALUE IS      82
BEST ANSWER SCORE VALUE IS     82

```

```
Similarity
Score
```



```

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 39 LPKELLRIFSFLDIVTLCRCAQISKAWNILALDGSNW/TSQN.
-E 10.0

```

```
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 39 SOR L2 SCORE D

=> D 1 39 TRIAL SCORE ALIGN

L3 ANSWER 1 OF 39 USGENE COPYRIGHT 2008 SEQUENCEBASE CORP...
TI Novel nucleic acids and polypeptides
(PublishedApplication)

MTY DNA
SQL 2358
SCORE 82 100% of query self score 82

BLASTALIGN

Query = 38 letters
Length = 2358
Score = 81.6 bits (200), Expect = 4e-21
Identities = 38/38 (100%), Positives = 38/38 (100%)
Frame = +1

Query: 1 LPKELLRRIFSFLLDIVTLRCRAQISKAWNILALDGSNW 38
LPKELLRRIFSFLLDIVTLRCRAQISKAWNILALDGSNW
Sbjct: 88 LPKELLRRIFSFLLDIVTLRCRAQISKAWNILALDGSNW 201

L3 ANSWER 39 OF 39 USGENE COPYRIGHT 2008 SEQUENCEBASE CORP...
TI Drosophila sequences (Patent)

MTY DNA
SQL 1306
SCORE 70 85% of query self score 82

BLASTALIGN

Query = 38 letters
Length = 1306
Score = 70.1 bits (170), Expect = 6e-18
Identities = 28/38 (73%), Positives = 36/38 (94%)
Frame = -3

Query: 1 LPKELLRRIFSFLLDIVTLRCRAQISKAWNILALDGSNW 38
LPKE+LLR+FS+LD+V+LCRCAQ+ K WN+LALDGS+W
Sbjct: 1019 LPKEVLLRVFSYLDVVSLLRCRAQVCKYWNVLALDGSSW 906

HELP SBATCH

Similarity Batch Search

The GETSIM and BLAST run packages are tools for searching USGENE polypeptide and nucleotide sequence data by similarity (homology). See HELP SIM. The BATCH option provides a facility to run similarity searches offline, especially those which would otherwise take a long time to complete in online mode. The BATCH option is therefore especially useful for the FASTA based GETSIM package which generally takes much longer to run online than BLAST. In BATCH mode the query is processed without the need to stay on-line. The results can be collected in a later session.

Using the offline BATCH option with GETSIM also allows longer search queries to be used. Queries can be up to 2,000 characters. See HELP QLIMITS.

Initiation of Similarity Batch Search

To initiate a similarity batch search, enter at the arrow prompt RUN GETSIM (or RUN BLAST) followed by the L-number of your sequence query qualified (/SQN, /SQP, or /TSQN) and BATCH, e.g. RUN GETSIM L4/SQN BATCH .

The system will then prompt you for a batch request identifier (name) of your choice which may consist of up to 8 letters or digits, e.g. PROJECT1 or PRJ17.

The query L-number used in a GETSIM/BLAST BATCH search will usually have been created by an UPLOAD of an ASCII file containing your sequence query, as sequences longer than 256 characters can only be entered into the system with the UPLOAD command. The processing of your request will commence immediately unless you have already another job in the queue.

Collection of Results

To collect the results or check the status of your GETSIM/BLAST batch search, enter RUN GETBATCH at an arrow prompt. The following options are available with RUN GETBATCH:

- a) enter the batch identifier to collect the batch result , e.g.
RUN GETBATCH PROJECT1. An L-numbered answer set is automatically created and the batch result file receives the status "retrieved". The status of a request is reported with "queued", "running", "completed" or "retrieved".
- b) enter # to see the list and status of your current batch requests
- c) enter * to see the identifier and status of the first of your current batch requests
- d) enter - followed by the batch identifier to cancel the queued or running batch search or to delete the batch result file.
- e) enter END to leave the RUN GETBATCH subcommand level and return to an arrow prompt.

Note: A "retrieved" batch request is deleted automatically one week after the first retrieval. During this time it is possible to retrieve the same request several times and process the answer set.

Costs of a Batch Search

Please note that a special fee is charged for the similarity batch search (for prices see HELP COST). This fee consists of two components:

- a) for the initiation of the batch search, i.e. when RUN GETSIM BATCH L# (or BLAST BATCH L#) is entered, and
- b) for the collection of the results of a completed batch search, i.e. when the batch search completed and when the RUN GETBATCH identifier is entered.

This second component (b) is not charged if the (GETSIM) batch search result is incomplete. Incomplete (GETSIM) batch results are caused by sequence queries which are too unspecific and retrieve more than 10,000 answers with the same score. Only the first retrieval of a batch request will be charged. Batch results are deleted seven days after the first retrieval. During this period subsequent repeat retrievals of the batch result will be free of charge.

Example using GETSIM:

Part 1: Upload and Initiation of GETSIM Batch search

```
=> UPLOAD
IS THIS DATA A QUERY, OR FOR A RUN PACKAGE? Q/R/(END):r
ENTER NAME OF RUN PACKAGE, END OR (?):GETSIM
START LOCAL KERMIT TRANSMIT PROCESS

UPLOAD SUCCESSFULLY COMPLETED
L1 GENERATED

=> RUN GETSIM L1/SQN BOTH BATCH

PLEASE ENTER BATCH IDENTIFIER (MAX. 8 CHARS):ACTIN1

RUN GETSIM AT 18:07:09 ON 13 JUL 2007
COPYRIGHT (C) 2007 FIZ KARLSRUHE GMBH

BATCH PROCESSING STARTED FOR ACTIN1
```

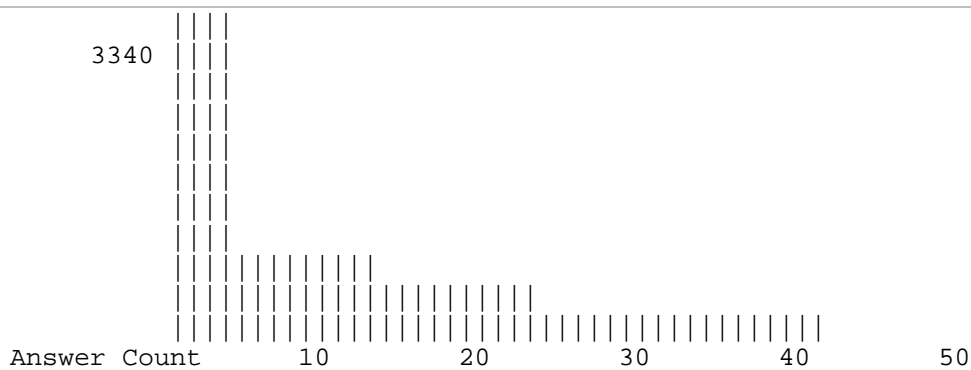
Entering a second Batch search:

```
=> RUN GETSIM L2/SQN COM BATCH

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

RUN GETSIM AT 18:08:00 ON 13 JUL 2007
COPYRIGHT (C) 2007 FIZ KARLSRUHE GMBH

PREVIOUS BATCH REQUEST STILL RUNNING
BATCH PROCESSING QUEUED FOR AQUAPOR
```

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

ENTER (ALL) OR ? :ALL

L1 RUN STATEMENT CREATED

```
L1      41 GCACCCGGCAGCGGTCTCAGGCCAAGCCCCCTGCCAGCATGGCCAGCGAG
      TTCAAGAAGAAGCTCTTCTGGAGGGCAGTGGTGGCCGAGTTCCTGGCCAC
      .
      .
      .
      TGGGTCCAGAAGACGTGGTCTAGACCAGGGCTGCTCTTTCCACTTGCCCT
      GTGTTCTTTCCCCAGGGGCATGACTGTGCGCCACACGCCTCTGCATATATG
      TCTCTTTGGAGTTGGAATTTCAATTATATGTTAAGAAAATAAAGGAAAATG
      ACTTGTAAGGTC/SQN.BOTH
```

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

Batch result files remaining:

Batch result files remaining:

```
ACTIN1   Retrieved (getsim)
PRJ17    Completed (blast)
AQUAPOR  Running   (getsim)
```

HELP SALERT

Similarity (homology) Current Awareness Searching

Continuously monitoring the patenting of peptide or nucleotide sequences by similarity (homology) can be conveniently achieved using the ALERT feature of the USGENE file. Once set up as a Current Awareness search (ALERT search), a biosequence query is routinely run against the updates of the database. The results can be collected online at any time up to three months from the Alert run. Up to sixteen simultaneous tasks are allowed for the Alert option and up to 192 result sets can be stored per loginID. Collected ALERT result sets will stay in the queue till the next update, unless they have been deleted by the customer.

The ALERT option is available for GETSIM as well as for BLAST similarity searches. There is no charge for initiating and executing an Alert, but the result set will be subject to a charge on collection. Uncollected or incomplete (GETSIM) answer sets will not be charged for. Empty answer sets from ALERT will be clearly marked in the output queue.

Initiating the ALERT searches:

```
=> RUN GETSIM
PLEASE ENTER SEQUENCE QUERY OR ?:L1/SQN BOTH ALERT
PLEASE ENTER ALERT IDENTIFIER (MAX. 8 CHARS):AQUAPNG
PLEASE ENTER ALERT TITLE (MAX. 40 CHARS):AQUAPORINNGETSIM

RUN GETSIM AT 17:29:24 ON 13 JUL 2007
COPYRIGHT (C) 2007 FIZ KARLSRUHE GMBH

NEW ALERT CREATED

or

=> RUN GETSIM L1/SQN BOTH ALERT
PLEASE ENTER ALERT IDENTIFIER (MAX. 8 CHARS):AQUAPNG
PLEASE ENTER ALERT TITLE (MAX. 40 CHARS):AQUAPORINNGETSIM

RUN GETSIM AT 17:29:24 ON 13 JUL 2007
COPYRIGHT (C) 2007 FIZ KARLSRUHE GMBH

NEW ALERT CREATED
```

Entering a second ALERT:

Up to 16 ALERT tasks can be set up per login ID.

```
=> RUN BLAST L1/SQN ALERT
PLEASE ENTER ALERT IDENTIFIER (MAX. 8 CHARS):AQUAPNB
PLEASE ENTER ALERT TITLE (MAX. 40 CHARS):AQUAPORINNBLAST

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).....

NEW ALERT CREATED
```

Query Check:

```
=> RUN ALERT
  Enter "R" to process alert results
  or "Q" to process alert queries
  or alert id to retrieve results
  or enter . for (end)
ALERT REQUEST:Q

  CURRENT ACTIVE ALERT QUERIES
NO. NAME          INSTALLED SEARCH TITLE
1) AQUAPNB       20030213 BLAST  AQUAPORINNBLAST
2) AQUAPNG       20030213 GETSIM AQUAPORINNGETSIM
3) CLONTECB      20030128 BLAST  CLONTECPBLAST
4) CLONTECG      20030128 GETSIM CLONTECPGETSIM
-----

Enter No. of query to be displayed
  or "R" to process alert results
  or enter . for (end)
QUERY REQUEST:1

ALERT NAME: AQUAPNB
INSTALLED : 20030213
TITLE      : AQUAPORINNBLAST
gcacccggcagcggtctcaggccaagccccctgccagcatggccagcgag
ttcaagaagaagctcttctggagggcagtggtggccgagttcctggccac
gacctctttgtcttcacatcagcatcggttctgccctgggcttcaaatacc
.....
tgggtccagaagacgtggtctagaccagggctgctctttccacttgcct
gtgttctttcccagggcatgactgtcgcacacgcctctgcatatag
tctcttggagttggaatttcattatatgttaagaaaataaaggaaaatg
acttgaaggtc/sqn. -e 10.0
-----

Enter "T" to change query title
  or "-" to delete alert query
  or "R" to process alert results
  or enter . for (end)
QUERY REQUEST:.
```

Status check and collection of ALERT Search Results

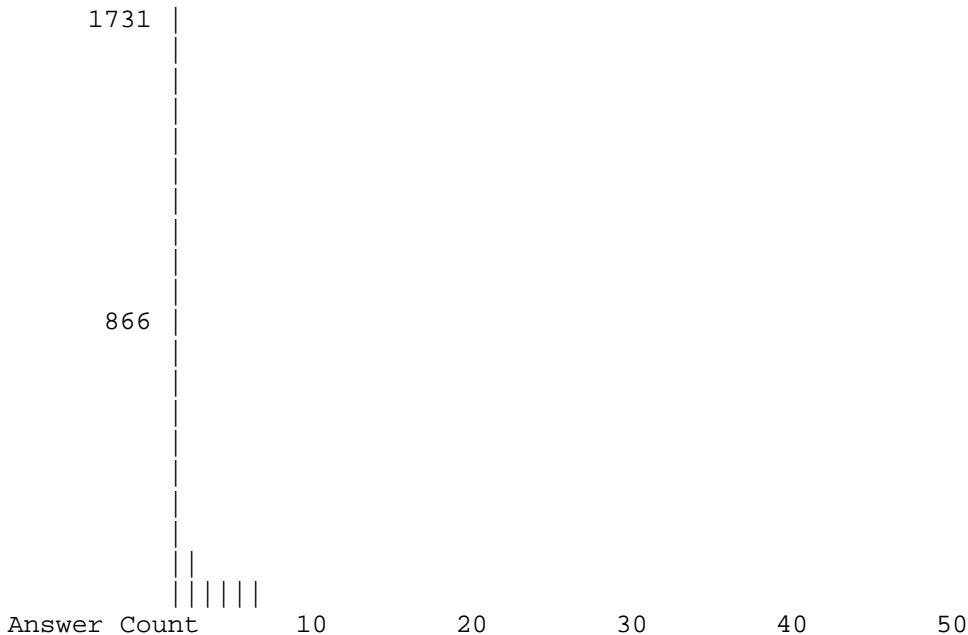
```
=> RUN ALERT
  Enter "R" to process alert results
  or "Q" to process alert queries
  or alert id to retrieve results
  or enter . for (end)
ALERT REQUEST:R

  CURRENT RESULTS AVAILABLE
  NAME          RUN DATE
1) AQUAPNB      20030214 (blast)
2) AQUAPNG      20030214 (getsim)
3) CLONTECB     20030214 (blast)
4) CLONTECG     20030214 (No answers - getsim)
-----

Enter No. of result to be selected
  or "-" before No. to delete result
  or "Q" to process alert queries
  or enter . for (end)
RESULT REQUEST:1
.....
```

38 ANSWERS FOUND BELOW EXPECTATION VALUE OF 10.0

Similarity
Score



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

L16 RUN STATEMENT CREATED

```
L16      38 GCACCCGGCAGCGGTCTCAGGCCAAGCCCCCTGCCAGCATGGCCAGCGAG
      .....
      GTGTTCTTTCCCCAGGGGCATGACTGTCGCCACACGCCTCTGCATATATG
      TCTCTTTGGAGTTGGAATTTTCATTATATGTTAAGAAAATAAAGGAAAATG
      ACTTGTAAGGTC/sqn. -e 10.0
```

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

CURRENT RESULTS AVAILABLE

NAME	RUN DATE	
1) AQUAPNB	20070214	(blast)
2) AQUAPNG	20070214	(getsim)
3) CLONTECB	20070214	(blast)
4) CLONTECG	20070214	(No answers - getsim)

Enter No. of result to be selected
or "-" before No. to delete result
or "Q" to process alert queries
or enter . for (end)

RESULT REQUEST: **END**

If no answers are available this will be clearly marked in the output queue. No answer set is created, hence no answer collection charge is incurred.

=> RUN ALERT

Enter "R" to process alert results
or "Q" to process alert queries
or alert id to retrieve results

```

    or enter . for (end)
ALERT REQUEST:R

    CURRENT RESULTS AVAILABLE
      NAME          RUN DATE
1) AQUAPNB      20070214  (blast)
2) AQUAPNG      20070214  (getsim)
3) CLONTECB     20070214  (blast)
4) CLONTECG     20070214  (No answers - getsim)
-----

Enter No. of result to be selected
  or "-" before No. to delete result
  or "Q" to process alert queries
  or enter . for (end)
RESULT REQUEST:4

NO ANSWERS FOUND ABOVE A THRESHOLD OF 303
  QUERY SELF SCORE VALUE IS 7210

    CURRENT RESULTS AVAILABLE
      NAME          RUN DATE
1) AQUAPNB      20070214  (blast)
2) AQUAPNG      20070214  (getsim)
3) CLONTECB     20070214  (blast)
4) CLONTECG     20070214  (No answers - getsim)
-----

Enter No. of result to be selected
  or "-" before No. to delete result
  or "Q" to process alert queries
  or enter . for (end)
RESULT REQUEST:END

```

Please note that the score threshold for GETSIM ALERT searches has been lowered compared to the standard procedures. This reflects the smaller number of sequences searched and has the benefit of higher selectivity.

HELP GSEQ

The GETSEQ run package is a tool to search the USGENE database for a direct sequence code match of peptide and nucleic acid sequences. This method is ideal for short and/or highly conserved sequence queries where similarity (homology) searching is not required. When using GETSEQ, note that the query L-number can be derived from a previous sequence code match search carried out in the DGENE, PCTGEN or the CAS REGISTRY file. Maximum length of sequence queries are listed in HELP QLIMITS. For information on similarity searching see HELP SIM.

Below, the different approaches to use RUN GETSEQ are shown.

```
=> QUE MRRGARCRNARVRPSFGGEG/SQSP
L1  QUE MRRGARCRNARVRPSFGGEG/SQSP

=> RUN GETSEQ L1/SQSP

RUN GETSEQ AT 09:59:41 ON 16 JUL 2007
COPYRIGHT (C) 2007 FIZ KARLSRUHE GMBH

L2  RUN STATEMENT CREATED
L2          4 MRRGARCRNARVRPSFGGEG/SQSP

=> RUN GETSEQ
PLEASE ENTER SEQUENCE PATTERN OR ? :MRRGARCRNARVRPSFGGEG
TYPE OF SEARCH ? (SQSP) :SQSP

RUN GETSEQ AT 10:00:09 ON 16 JUL 2007
COPYRIGHT (C) 2007 FIZ KARLSRUHE GMBH

L3  RUN STATEMENT CREATED
L3          4 MRRGARCRNARVRPSFGGEG/SQSP

=> RUN GETSEQ
PLEASE ENTER SEQUENCE PATTERN OR ? :MRRGARCRNARVRPSFGGEG/SQSP

RUN GETSEQ AT 10:00:31 ON 16 JUL 2007
COPYRIGHT (C) 2007 FIZ KARLSRUHE GMBH

L4  RUN STATEMENT CREATED
L4          4 MRRGARCRNARVRPSFGGEG/SQSP

=> RUN GETSEQ MRRGARCRNARVRPSFGGEG/SQSP

RUN GETSEQ AT 10:00:45 ON 16 JUL 2007
COPYRIGHT (C) 2007 FIZ KARLSRUHE GMBH

L5  RUN STATEMENT CREATED
L5          4 MRRGARCRNARVRPSFGGEG/SQSP

=> D HIT

L5  ANSWER 1 OF 4  USGENE COPYRIGHT 2007 SEQUENCEBASE CORP on STN
SEQ
      1 lswvwwvscp mrrgarcrna rvrpsfggeg drgsihsavd pspnrrqdyh
      =====
HITS AT:  11-30
```

GETSEQ for polypeptide sequences

Four options are available in the GETSEQ run package for searching polypeptide sequences using amino acid codes. Each requires the corresponding field qualifier described below. The sequence query is input using 1- and/or 3-letter codes for the amino acids. Enter HELP AAC at an arrow prompt (=>) in the USGENE file for a list of codes for the common amino acids. Enter HELP SQQ at an arrow prompt for information on symbols used to allow for variability in sequence queries.

Exact Sequence Search of Polypeptides (/SQEP) retrieves sequences that exactly match the search query. Variability searching is possible.

Exact Family Sequence Search of Polypeptides (/SQEFP) retrieves answers that exactly match the query and answers in which family-equivalent substitution of the query amino acids occurs. Variability searching is possible.

Subsequence Search of Polypeptides (/SQSP) retrieves exact answers plus sequences in which the query sequence is embedded. Variability searching is possible.

Subsequence Family Search of Polypeptides (/SQSFP) retrieves exact sequences, subsequences, and answers in which family-equivalent substitution of the query amino acids occurs. For example, the query ADHIFC/SQSFP retrieves the equivalent fragment ...PQKLYC..

The families of amino acid equivalents retrieved in polypeptide family searches are:

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)

A GETSEQ polypeptide sequence query (i.e. a query consisting of one or more of these fields: /SQEP, /SQSP, /SQEFP, /SQSFP) may be combined directly in a single search with only the following fields: /FS, /UP. However, any sequence L-numbered answer set from RUN GETSEQ may be combined with any search field in the USGENE File (e.g. => S L10 AND ARTIFICIAL SEQUENCE/ORGN, where L10 represents the answer set from a RUN GETSEQ operation).

GETSEQ for Nucleic Acid Sequences

Two options are available in the GETSEQ run package for searching nucleic acid sequences using 1-letter codes. Each requires the corresponding field qualifier described below. Enter HELP NUC at an arrow prompt in the USGENE file for a list of codes for nucleic acids. Enter HELP SQQ for information on symbols used to allow for variability in sequence queries.

Exact Sequence Search of Nucleic Acids (/SQEN) retrieves sequences that exactly match the search query. Ambiguity codes for nucleic acids are allowed. Variability searching is possible.

Subsequence Search of Nucleic Acids (/SQSN) retrieves exact answers plus sequences in which the query sequence is embedded. Variability searching is possible.

A GETSEQ nucleic acid sequence query (i.e. a query consisting of one or more of these fields: /SQEN, /SQSN) may be combined directly in a single search with only the following fields: /FS, /UP. However, any sequence field may be combined with any search field in the USGENE file (e.g. => S L10 AND ARTIFICIAL SEQUENCE/ORGN, where L10 represents the answer set from a RUN GETSEQ operation).

HELP SQQ

The following symbols may be used in sequence searches within RUN GETSEQ to allow for variability in residues. These options are not applicable to either RUN BLAST or RUN GETSIM (see HELP SIM).

Symbol(s)	Function	Search Example: what the query retrieves
[]	Specify alternate residues	LGP[VL]/SQSP: LGP followed by either V or L
[-] or the tilde in brackets	Exclude a specific residue or alternate residues	ATTGC[-A]GAAG/SQSN: ATTGC followed by any nucleotide except A followed by GAAG
{ } with a number or range	Repeat the preceding symbol, sequence, or an L-number for a sequence query	(FL){2}/SQSP: FL repeated twice, i.e. FLFL. GG(FL){1-3}/SQSP (or GG(FL){1,3}/SQSP): GGFL, or GGFLFL, or GGFLFLFL. KLK(WD){0,N}/SQSP: KLKN or KLK followed by any number of repetitions of WD followed by N, e.g., KLKWDN, KLKWDWDN, KLKWDWDWDN, etc. CAT(CTG){1,}TATT/SQSN: CAT followed by one or more repetitions of CTG followed by TATT, e.g. CATCTGTATT, CATCTGCTGTATT, CATCTGCTGCTGTATT etc.
?	Repeat the preceding symbol, sequence, or sequence query zero or one time	FLRRI(RP)?K/SQSP is equivalent to FLRRI(RP){0,1}K/SQSP: FLRRIK or FLRRIRPK
*	Repeat the preceding symbol, sequence, or sequence query zero or more times	CAT(CTG)*TATT/SQSN is the same as CAT(CTG){0,}TATT/SQSN: CATTATT or CAT followed by any number of repetitions of CTG followed by TATT, e.g. CATCTGTATT, CATCTGCTGTATT etc.
+	Repeat the preceding symbol, sequence, sequence query one or more times	CAT(CTG)+TATT/SQSN is equivalent or to CAT(CTG){1,}TATT/SQSN: CAT followed by one or more repetitions of CTG followed by TATT, e.g. CATCTGTATT, CATCTGCTGTATT, CATCTGCTGCTGTATT etc.

In addition, the caret character may be used at the beginning or at the end of a sequence to search for that sequence at the beginning or end of the sequence field.

To require alternate sequence queries, separate the sequence expressions by the vertical bar.

Specifying Gaps

You may specify a gap in a sequence expression using the period (.) for one residue, the colon (:) for zero or one residue or the period (.) followed by an appropriate repeat expression. The following table summarizes all the options for specifying gaps in GETSEQ sequence searches.

Symbol(s)	Function	Query Example: what the query retrieves
.	a gap of one residue	SY.RPG/SQSP: SY followed by one residue followed by RPG
.{m} or .m.	a gap of m residues	SY.{2}RPG/SQSP: SY followed by any 2 residues followed by RPG
.{m,u} or . {m-u}	a gap of m to u residues	GFF.{2,10}LSS/SQSP: GFF followed by a gap of 2 to 10 residues followed by LSS
.? or : or .{0,1} or .{0-1}	a gap of zero or one residue	AGA.?SRI/SQSFP is equivalent to AGA.{0,1}SRI/SQSFP: AGA followed by zero or one residue followed by SRI
.* or . {0,} or . {0-}	A gap of zero or more residues	HLC.*TYG/SQSP is equivalent to HLC.{0,}TYG/SQSP: HLC followed by a gap of zero or more residues followed by TYG
.+ or . {1,} or .{1-}	A gap of one or more residues	SY.+TH/SQSP is equivalent to SY.{1,}TH/SQSP: SY followed by any number of residues followed by TH

Concatenating Queries

In addition to the variability symbols, you may use the & symbol to join together sequences or L-numbered queries. The concatenation symbol may be used in subsequence searches within RUN GETSEQ (/SQSN, /SQSP, /SQSFP) and also in exact sequence searches of proteins or nucleic acids (/SQEP, /SQEFP, /SQEN).

&	Concatenate or join together sequences or queries	L1&L2&L3/SQSN: the sequence in L1 followed by the sequence in L2 followed by the sequence in L3.
---	---	---

Order of Precedence

More than one symbol may be used to create complex sequence queries. For example, the query L2&L5{1,3}/SQSN specifies that the sequence in L2 is to be followed by one to three repetitions of the sequence query in L5. If you do not use parentheses in sequence queries, the operations will be executed in the following order:

1. repeat symbols ? or * or +
2. repeat expressions using curly braces, e.g. {3,6},
3. concatenation symbol &,
4. the vertical bar

HELP QLIMITS

The minimum length of sequence queries is 5 characters.

Sequence queries directly entered or created with the QUERY command to be used for the run commands GETSEQ, GETSIM and BLAST may have a maximum length of 256 characters. Any further characters will be ignored.

When searching sequences longer than 256 characters, the UPLOAD command needs to be used. The maximum length for uploaded sequence queries used for RUN GETSEQ is 2,000 characters. Sequence queries uploaded from ASCII files may have a maximum length of 500 characters for RUN GETSIM /SQN and /TSQN searches, and 750 characters for /SQP searches. For RUN BLAST the maximum length is 10,000 characters. In any case the line length may not exceed 300 characters.

For RUN GETSIM and RUN BLAST, also a BATCH mode and an ALERT feature are available that allow for searching sequences offline (BATCH) and setting up sequence current awareness searches (ALERT). Sequence query maximum lengths are 2,000 characters for GETSIM and 10,000 characters for BLAST BATCH or ALERT searches. See also HELP SALERT and HELP SBATCH.

HELP AAC

Sequences submitted to the United States Patent and Trademark Office (USPTO) are given by patent applicants according to WST.25.

The following table lists the 1- and 3-letter codes that may be used for the common amino acids in sequence searches with RUN GETSEQ. Uncommon amino acids are represented in the sequence either by a related parent amino acid, if available, or by an 'X' (or 'Xaa'). Details about uncommon amino acids in a sequence can be found in the corresponding feature table (FEAT).

1-Letter Code -----	3-Letter Code -----	Name ----
A	Ala	Alanine
B	Asx	Aspartic acid or Asparagine
C	Cys	Cysteine
D	Asp	Aspartic acid
E	Glu	Glutamic acid
F	Phe	Phenylalanine
G	Gly	Glycine
H	His	Histidine
I	Ile	Isoleucine
K	Lys	Lysine
L	Leu	Leucine
M	Met	Methionine
N	Asn	Asparagine
P	Pro	Proline
Q	Gln	Glutamine
R	Arg	Arginine
S	Ser	Serine
T	Thr	Threonine
V	Val	Valine
W	Trp	Tryptophan
X	Xaa	Uncommon
Y	Tyr	Tyrosine
Z	Glx	Glutamic acid or Glutamine

The codes B and Z may be used only in subsequence searches (/SQSP and /QSFP). In family searches B and Z match both the specific amino acids and the generic B and Z in the database.

3-Letter Code -----	1-Letter Code -----	Name ----
Ala	A	Alanine
Arg	R	Arginine
Asn	N	Asparagine
Asp	D	Aspartic acid
Asx	B	Aspartic acid or Asparagine
Cys	C	Cysteine
Gln	Q	Glutamine
Glu	E	Glutamic acid
Glx	Z	Glutamic acid or Glutamine
Gly	G	Glycine
His	H	Histidine
Ile	I	Isoleucine
Leu	L	Leucine
Lys	K	Lysine
Met	M	Methionine
Phe	F	Phenylalanine
Pro	P	Proline
Ser	S	Serine
Thr	T	Threonine
Trp	W	Tryptophan
Tyr	Y	Tyrosine
Val	V	Valine
Xaa	X	Uncommon

The codes Asx and Glx may be used only in subsequence searches (/SQSP and /SQSFP). In family searches Asx and Glx match both the specific amino acids and the generic Asx and Glx in the database. Broad queries can also be composed employing the methods exemplified in HELP SQQ.

HELP NUC

Sequences submitted to the United States Patent and Trademark Office (USPTO) are given by patent applicants according to WST.25.

The following table lists the symbols and ambiguity codes for nucleotides according to the IUPAC system that may be used in nucleic acid sequence searches employing RUN GETSEQ.

Codes -----	Name or Definition -----
A	Adenine
G	Guanine
U	Uracil
R	A or G
S	C or G
K	G or T/U
H	A, C or T/U; not G
B	C, G or T/U; not A
C	Cytosine
T	Thymine
M	A or C
W	A or T/U
Y	C or T/U
V	A, C or G; not T/U
D	A, G or T/U; not C
N	Unknown or Other

Exact Sequence Searches of Nucleic Acids (/SQEN) allow all codes and match the codes in the query exactly against the codes in the database.

Subsequence searches allow the requested sequences to be a subsequence of the sequences in the database.

Broad queries can also be composed employing the methods exemplified in HELP SQQ.

HELP SQL

USGENE has a fully numerically range searchable Sequence Length (SQL) field.

The /SQL field may be searched with numeric operators or ranges, e.g. 100-200/SQL or SQL>400. SQL can also be used with the SORT command, e.g. SORT SQL D would give the longest sequence first and the shortest last.

Example: a search for short peptide sequences with 30-50 amino acid residues

```
=> S PROTEIN/FS AND 30-50/SQL
      1678105 PROTEIN /FS
      403007 30-50/SQL
L1    74758 PROTEIN /FS AND 30-50/SQL

=> D BRIEF

L1    ANSWER 1 OF 74758 USGENE COPYRIGHT 2007 SEQUENCEBASE CORP on STN
AN    6720166.541 Protein USGENE
TI    Non-a, non-b, non-c, non-c, non-d, non-e hepatitis reagents and methods
      for their use (Patent)
IN    Simons John N. (Grayslake, IL); Pilot-Matias Tami J. (Green Oaks, IL);
      Dawson George J. (Libertyville, IL); Schlauder George G. (Skokie, IL);
      Desai Suresh M. (Libertyville, IL); Leary Thomas P. (Kenosha, WI);
      Muerhoff Anthony Scott (Kenosha, WI); Erker James Carl (Hainesville, IL);
      Buijk Sheri L. (Round Lake, IL); Mushahwar Isa K. (Grayslake, IL)
PA    Abbott Laboratories(Abbott Park IL)
PI    US 6720166 B2 20040413
      US 20020119447 A1 20020829
AI    US 1995-424550 19950605
RLI   WO 1995-US2118
DT    Patent
AB    Hepatitis GB Virus (HGBV) nucleic acid and amino acid sequences useful
      for a variety of diagnostic and therapeutic applications, kits for using
      the HGBV nucleic acid or amino acid sequences, HGBV immunogenic
      particles, and antibodies which specifically bind to HGBV. Also provided
      are methods for producing antibodies, polyclonal or monoclonal, from the
      HGBV nucleic acid or amino acid sequences.
ECLM  US6720166 B2: What is claimed is:1. A purified HGBV polynucleotide,
      isolated from a positive stranded HGBV RNA genome, wherein said genome
      encodes a polypeptide, wherein said polypeptide consists essentially of
      an amino acid sequence, at least 48% identical to the polypeptide
      sequence SEQ ID NO:387.
SSO   PROTEIN; EMBL; GRANTED
ORGN  Unknown
SQL   40
SEQ   1 dpllwawypp rayedwsppc ilpfqggvre igrpvlragg

FEATURE TABLE:
Key      |Location|
=====+=====+=====
source   |1..40   |
```

HELP NCBI

BLAST(R) is a product of the U.S. National Center of Biotechnology Information (NCBI) and the U.S. National Library of Medicine (NLM). On NCBI's web sites comprehensive documentation on the algorithm, the basics of similarity searching with BLAST(R), and basic and advanced parameters are provided to the scientific community. For NCBI documentation on BLAST please consult the following NCBI sites:

<http://www.ncbi.nlm.nih.gov/blast/html/blastcghelp.html>

<http://www.ncbi.nlm.nih.gov/Education/BLASTinfo/information3.html>

<http://www.ncbi.nlm.nih.gov/Education/BLASTinfo/similarity.html>

<http://www.ncbi.nlm.nih.gov/About/outreach/glossary.html>

<http://www.ncbi.nlm.nih.gov/Education/BLASTinfo/glossary2.html>

<http://www.ncbi.nlm.nih.gov/Education/BLASTinfo/auxiliary.html>

http://www.ncbi.nlm.nih.gov/Education/BLASTinfo/BLAST_algorithm.html

<http://www.ncbi.nlm.nih.gov/Education/BLASTinfo/Scoring2.html>

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

HELP ALIGNMENT

The basic information about the similarity between two compared sequences is given by the alignment of both (displayed by the command `D ALIGN`). This means a direct comparison is made residue by residue between the two sequences over the area of their similarity. The representation of the extent of the similarity found between query sequence and hit sequence varies for alignments depending on the program (GETSIM or BLAST) producing the alignment. Please note that the exact definition and classification of amino acid families differs slightly in GETSIM and BLAST alignments. For definition of single letter characters see `HELP NUC` and `HELP AAC`.

BLAST alignments of nucleic acid sequences

Similarity in BLAST alignments is given by bars in a line between the two lines representing the query sequence (upper line) and the hit or subject sequence (lower line). A bar marks a full match between two nucleic acid residues and blanks show non-matching residues. Gaps are introduced in the query or the subject sequence for a better alignment of both sequences.

Example:

```
BLASTALIGN
  Query  = 3405 letters
  Length = 412
  Score  = 77.8 bits (39), Expect = 5e-18
  Identities = 158/207 (76%)
  Strand = Plus / Minus

Query: 3042 ctggttatggtgcagagagtgtaacattgacaagaggacaaaatacagtcaaggatcagg
          ||||| ||||| ||||| ||||| ||||| ||||| ||||| ||||| ||||| |||||
Sbjct: 207  ctgttacggtgcagaaagtgtaacactctcacgaggacaaaatactgtcaaaattactgg

Query: 3102 gaaaggtggccatagtggttcaacatttaggtggtgcatggggaggactgttcacaaat
          ||||| ||||| ||||| ||||| ||||| ||||| ||||| ||||| ||||| |||||
Sbjct: 147  gaaaggtggccatagtggttcttcatttaaagtctgtcatgggaaagaatggtcatcaac

Query: 3162 tggactccatgctgctgc----caccttgacaaggtaaatgggatttctgagatagaaaa
          ||| ||||| || ||||| || ||||| ||||| ||||| ||||| ||||| |||||
Sbjct: 87   tggcctccaagccagtgccaccacatctggataaggtaaatggtatctctgagttagaaaa

Query: 3222 tagtaaagtatatgatgatggggcacc 3248
          ||||| ||||| ||||| |||||
Sbjct: 27   cgagaaagtttatgatg----tgcacc 1
```

BLAST alignments of amino acid sequences

Similarity in BLAST alignments of amino acid sequences is given by different characters in a line between the two lines representing the query sequence (upper line) and the hit or subject sequence (lower line). A full match is given by the one letter code of the corresponding matching residue and a plus sign represents a protein family match. Blanks show non-matching residues. Gaps are introduced in the query or the subject sequence for a better alignment of both sequences.

Example:

```
BLASTALIGN
  Query = 96 letters
  Length = 2070
  Score = 26.9 bits (58), Expect = 8e-04
  Identities = 19/74 (25%), Positives = 38/74 (50%), Gaps = 8/74 (10%)
Query: 22  QVGKGSVSPNLIHVKALEARELIKISILQNC EE-DKQTVAEKISARSGAEIVQVIGRT
          QV + S+ + ++ +K ALEA+EL + + +E +K+T          RS + + +
Sbjct: 565 QVQQNSLHRDSLVT LKRALEAKELARQH LRDQLDEVEKET-----RSKLQEIDIFNNQ
Query: 81  IILYKTSV NKQ QIK 94
          + + NKQQ++
Sbjct: 618 LKELREIHNKQQLQ 631
```

GETSIM alignments of nucleic acid sequences

Similarity in GETSIM alignments is given by dots in a line between the two lines representing the query sequence (upper line) and the hit sequence (lower line). Two dots mark a full match between two nucleic acid residues and one dot represents the "family" similarity between Uracil (U) in RNA sequences and Thymine (T) in DNA sequences. Blanks show non-matching residues. Underscores in the query or the hit sequence are introduced for a better alignment of both sequences.

Example:

```
ALIGN Smith-Waterman score: 57
 33 na overlap starting at 546
aggagugguaggucuuacgaugccagcuguaau      <- Query
:: : . .: . .: : ::::: : ::
agtattcatattactaacaagccagctggaat      <- Answer
```

GETSIM alignments of amino acid sequences

Similarity for amino acids is also given by dots in GETSIM alignments. Two dots represent a full match and one dot a protein family match. Blanks show non-matching residues. Underscores in the query or the hit sequence are introduced for a better alignment of both sequences.

Example:

```
ALIGN Smith-Waterman score: 80
 49 aa overlap starting at 569
vge_gaiplsigyatllhmdqgvalgrvlpvmvlggltaiiisgclnql      <- Query
::: :: : . : : : : : : : : : : : : : : : : : : : : : : :
vgeygaspplclpyap__pegqpaalgftvalvmmnsfcflvvagayikl      <- Answer
```

The similarity percentage of the Smith-Waterman score is given in the SCORE field and can be displayed with D SCORE.

Other General HELP for USGENE

HELP ACCESSION

The STN output format may be used to input an accession number in the DISPLAY ACC, ORDER ACC, or PRINT ACC command. The USGENE output format is shown below. It consists of three parts: the publication year followed by the application number, and after the dot the sequence identity number of the sequence in the corresponding application document.

STN output format ----- 20030027139.16
STN output format -----7001596.33

HELP FIELDS

The following messages are available in the USGENE file for help with DISPLAY, PRINT, SEARCH, SELECT, and SORT fields and formats.

```
HELP ACCESSION - USGENE accession number formats
HELP CONTENT   - general USGENE file description
HELP COST      - price schedule for the USGENE file
HELP CROSSOVER - file crossover searching in USGENE
HELP DESK      - information on USGENE file user assistance
HELP DFIELDS   - list of display field codes
HELP EFIELDS   - list of fields from which terms may be
                  extracted
HELP FORMAT    - predefined formats for DISPLAY and PRINT
HELP SEQUENCE  - biosequence search and display codes
HELP SFIELDS   - list of search field codes
HELP SRTFIELDS - list of fields in search results that may
                  be used to sort answers in alphabetic or
                  numeric order
```

HELP SFIELDS

The searchable fields in the USGENE file are listed below. If you do not specify a field, your term will be searched in the basic index, which contains single words from the title, molecule type and organism name. The Feature Table field allows simultaneous left and right truncation (SLART).

Search Code	Definition
/AB	Abstract
/AC	Application Country
/AD	Application Date
/AN	Accession Number
/AP (AI)	Application Number
/AY	Application Year
/BI	Basic Index
/DT (TC)	Document Type
/ECLM (MCLM)	Exemplary Claim
/ED (UP)	Entry Date
/FEAT	Feature Table
/FS	File Segment
/IN (AU)	Inventor
/MTY	Molecule Type
/ORGN	Organism Name
/PA (CS)	Patent Assignee
/PC	Patent Country
/PD	Publication Date
/PK	Patent Kind Code
/PN (PATS)	Patent Number
/PY	Publication Year
/RLC	Related Application Country
/RLD	Related Application Date
/RLN (RLI)	Related Application Number
/RLY	Related Application Year
/SEQC	Sequence Number Count
/SEQN	Sequence Identity Number
/SQL	Sequence Length
/SSO	Sequence Source
/TI	Title

All fields are text fields except: AD, AY, ED, RLD, RLY, SEQN, PD, PY and SQL, which are numeric, and can be searched with numeric operators or ranges, e.g., 500-1000/SQL or 2002/AY.

Search and display fields generally have the same field codes. To see a list of display fields, enter 'HELP DFIELDS' at an arrow prompt (=>).

HELP SRTFIELDS

The SORT command is used to rearrange search results in either alphabetic or numeric order of sortable fields. The fields that you may use for sorting answers in the USGENE file are listed below.

Sort Code	Definition
----	-----
AI (AP)--	Application Information
AN -----	Accession Number
DT (TC)--	Document Type
ED (UP)--	Entry Date
FS -----	File Segment
MTY -----	Molecule Type
ORGN ----	Organism Name
PA (CS)--	Patent Assignee
PC -----	Patent Country
PD -----	Publication Date
PI (PN)--	Patent Information
PK -----	Patent Kind Code
PY -----	Publication Year
SCORE ---	Similarity Score
SQL -----	Sequence Length
TI -----	Title

HELP EFIELDS

The SELECT command is used to create E-numbered or L-numbered lists containing terms taken from a specified display field in search answers.

The keyword, HIT, may be used in the SELECT command to restrict the terms extracted from the displayed data to terms which match the search expression used to create the answer set. The HIT keyword functions only if the answer set was created with HIGHLIGHTING ON. The resulting list of terms are the hit terms in the specified field.

The display fields from which terms may be extracted in the USGENE file are listed below.

Display Code -----	Definition -----
AB	Abstract
AC	Application Country
AD	Application Date
AI (AP)	Application Information
AIO	Application Information Original
AN	Accession Number
AY	Application Year
DT (TC)	Document Type
ED	Entry Date
FEAT	Feature Table
FS	File Segment
IN (AU)	Inventor
MTY	Molecule Type
ORGN	Organism Name
PA (CS)	Patent Assignee
PC	Patent Country
PD	Publication Date
PI (PN)	Patent Information
PK	Patent Kind Code
PY	Publication Year
RLC	Related Application Country
RLD	Related Application Date
RLIO	Related Application Information Original
RLN (RLI)	Related Application Number
RLY	Related Application Year
SCORE	Similarity Score
SEQ	Sequence (1-letter codes)
SEQ3	Sequence (3-letter codes)
SEQC	Sequence Number Count
SEQN	Sequence Identity Number
SQL	Sequence Length
SSO	Sequence Source
TI	Title

HELP DFIELDS

The display fields which you may see in records in this file are listed below. You may use these field codes in any combination with the DISPLAY and PRINT commands.

Display Code -----	Definition -----
AB	Abstract
AI (AP)	Application Information
AN	Accession Number
CLM	Claims
DT (TC)	Document Type
ECLM (MCLM)	Exemplary Claim
ED (UP)	Entry Date
FEAT	Feature Table
FS	File Segment
IN (AU)	Inventor
MTY	Molecule Type
ORGN	Organism Name
PA (CS)	Patent Assignee
PI (PN, PATS)	Patent Information
RLI (RLN)	Related Application Information
RLIO	Related Application Information, Original
SCORE	Similarity Score
SEQ	Sequence (1-letter-codes)
SEQ3	Sequence (3-letter-codes)
SEQC	Sequence Number Count
SEQN	Sequence Identity Number
SEQO	Sequence Original
SQL	Sequence Length
SSO	Sequence Source
TI	Title

For more information on displaying individual fields, enter 'HELP FORMAT' at an arrow prompt (=>). To find out about creating search terms from display fields, see 'HELP SELECT'. For information on which display fields may be used in the SELECT command see 'HELP EFIELDS'.

HELP FORMAT

Search results in the USGENE file may be displayed online or printed offline to see one of the predefined formats of fields listed below or a combination of these.

The following predefined formats of fields can be requested:

```
ALIGN -----Alignment between query and retrieved sequence
                in a similarity search (RUN GETSIM or RUN BLAST)
ALL -----AN (MTY), TI, IN, PA, PI, AI, RLI, ED, DT,
                AB, CLM, SSO, ORGN, SQL, SEQ, FEAT
APPS -----AI, RLI
IALL -----ALL, indented with text labels
BIB (STD)---AN (MTY), TI, IN, PA, PI, AI, RLI, DT
                (BIB is the default)
IBIB -----BIB, indented with text labels
BRIEF -----AN (MTY), TI, IN, PA, PI, AI, RLI, ED, DT,
                AB, ECLM, SSO, ORGN, SQL, SEQ, FEAT
IBRIEF -----BRIE, indented with the text labels
SQIDE -----AN, SQL, SEQ, FEAT
SQ3IDE -----AN, SQL, SEQ3, FEAT
SCAN -----TI (random display without answer numbers)
TRIAL -----TI, MTY, SQL
                (TRI, SAM, FREE)
```

- 1) Use RUN GETSIM or RUN BLAST first. See HELP SIM, HELP GSIM or HELP BLAST
- 2) By default, patent numbers, application and priority numbers are displayed in STN format. To display them in Derwent format, enter SET PATENT DERWENT at an arrow prompt. To reset display to STN format, enter SET PATENT STN.
- 3) Sequences in USGENE are given according to WST.25 of the WIPO.

Three special formats are available for use with hit-term highlighting. They can be used alone or with other fields or predefined formats for displaying search results. They are:

```
HIT ----- All fields containing hit terms
KWIC ----- All hit terms plus a maximum of 50 words on either
                side
OCC ----- List of display fields containing hit terms

Hit terms will be highlighted in all display fields.
```

To display a particular field or fields, enter the display field codes. For a list of display field codes, enter 'HELP DFIELDS' at an arrow prompt (=>). Examples of formats include: 'TI'; 'AN, TI, PI'; 'PI, AI, RLI'. Information will be displayed in the same order as your format specification.

The same formats except SCAN may be used in the PRINT command to print search results. All of the formats except for SCAN, HIT, KWIC, and OCC may be used with the DISPLAY ACC command to display the record for a specified accession number, and with the PRINT ACC command to print accession number records offline.

HELP CROSSOVER

The term 'file crossover' refers to the use of an answer set created by a search in one file as a search profile in another file.

If you want to search the same query, use the L-number of an answer set created in another file as a search profile in this file. The query used to create the answer set is searched.

Example:

```
(In another STN file)

=> S DISEASE /TI
      216249 DISEASE/TI
      200876 DISEASES/TI
L1      402358 DISEASE /TI
      ((DISEASE OR DISEASES)/TI)

(In the USGENE file)

=> S L1
      16832 DISEASE/TI
      2946 DISEASES/TI
L2      19778 DISEASE /TI
      ((DISEASE OR DISEASES)/TI)
```

You may also crossover and search a set of terms extracted from an answer set. For more information on crossover of extracted terms, enter **HELP TERM CROSSOVER** at an arrow prompt (=>). The run packages **GETSEQ**, **GETSIM**, and **BLAST**, which are used for sequence searching, allow L-numbered queries from other STN sequence files, e.g. **DGENE**.

HELP UPDATE/SDI

Update searching (also called current-awareness, Alert or SDI searching) can be done manually or automatically in the USGENE file. To do manual searches of this type, use the /ED field. The /ED field contains the date the record was added to the file.

To request a standard automatic update search, enter 'SDI' at an arrow prompt (=>). You will be prompted for all additional information needed for the request. The L# used in the SDI search profile can be generated from any **SEARCH**, **ACTIVATE**, or **QUERY** command but not from any **RUN** command.

To request an automatic update search based on sequence similarity (homology) answer sets created using **RUN BLAST** or **RUN GETSIM**, use the **ALERT** feature. See **HELP SALERT**.

The USGENE file is updated weekly. Automatic SDIs and ALERTs are therefore also run weekly. The default print format is **BIB**.

HELP RANGE

Searches in the USGENE file can be restricted to one of two file segments, protein(p), or nucleic(n). Valid keywords are NUC, N, PROT and P. RANGE parameters are the same for the SET and SEARCH commands.

Example:

```
=> SET RANGE=PROT  
  
=> SEARCH L10 RANGE=N
```

Enter 'HELP SEARCH RANGE' for an explanation of using RANGE in SEARCH. Enter 'HELP SET RANGE' for a method of doing a series of searches in a particular range set.

HELP HIGHLIGHT

Scanning search results in online displays and offline prints can be made easier by hit term highlighting. This feature is available for most display fields in the USGENE file. In the display or print, the hit terms, which are the terms in the document or record that matched your search profile, are either given in bold and red (online display) or preceded and followed by three asterisks. For example, if your search was on 'bone marrow', part of the display might look like this:

```
TI      HUMAN GENOME-DERIVED SINGLE EXON NUCLEIC ACID  
        PROBES USEFUL FOR ANALYSIS OF GENE EXPRESSION IN  
        HUMAN BONE MARROW  
  
or  
  
TI      HUMAN GENOME-DERIVED SINGLE EXON NUCLEIC ACID  
        PROBES USEFUL FOR ANALYSIS OF GENE EXPRESSION IN  
        HUMAN ***BONE MARROW***
```

In addition to the highlighting of hit terms in answer displays in the standard formats, there are also three formats that specifically involve hit term highlighting. They are the HIT, KWIC, and OCC formats. The HIT format shows only the display fields containing hit terms, the KWIC format shows the hit term(s) and a maximum of 50 words on either side, and the OCC format consists of a table of fields containing the hit terms, with the number of occurrences in each field being given.

When you enter the USGENE file, HIGHLIGHTING is SET ON by default. If you do not wish to have hit terms highlighted, you may enter SET HIGHLIGHT OFF at an arrow prompt. However, remember that answers from searches done while highlighting is set to OFF cannot be highlighted even if you set it back to ON. After SET HIGHLIGHT OFF is entered, the information that is necessary for highlighting is not saved with the answers.

HELP (S)

The (S) proximity operator is used in the USGENE file to specify that two terms must occur in the same sentence, in any order. The meaning of 'sentence' depends on the field.

Using (S) proximity is especially recommended when searching in PA. In this field (S)-implied proximity is implemented. Search terms are automatically combined with (S) proximity. Please note that you can avoid the use of (S)-implied proximity by putting the whole search expression in quotation marks. Then the expression is searched as a fixed string.

Examples:

```
=> S (MERCK CO)/PA
      6928 MERCK/PA
      196672 CO/PA
L1      5724 (MERCK CO)/PA
          ((MERCK(S)CO)/PA)

=> S (SMITH(S)JOHN)/IN
      535078 SMITH/IN
      440649 JOHN/IN
L2      3430 (SMITH(S)JOHN)/IN

=> S WO/PC(S)PY=2002
      75689 WO/PC
      259568 PY=2002
          (PY=2002)
L3      1329 WO/PC(S)PY=2002
```

HELP USAGETERMS

The following database producer's special conditions for use of his database(s) apply to your use of the USGENE file on STN.

I. General Part

1. Scope

Section 2 to 4 of these conditions apply to all databases offered via STN Karlsruhe as far as no differing regulations are specified under II. Special Part.

2. Customers

A customer is an individual or an institution (i.e. a legal body such as a university, public authority, company) for whom online access has been ordered.

3. Search Results

All rights are reserved. Search results delivered online or offline are only for internal (own) use of the customer. No written permission of the database producer is required if search results delivered to the customer in computer-readable form are used only for internal purposes of the customer, i.e. printed, processed, modified, or linked with other data (e.g. for creating a database). The customer must observe the copyright of the database producer.

Enter HELP SHARETERMS at an arrow prompt (=>) or visit

<http://www.stn-international.de/stndatabases/keepshare/index.html>

for detailed information on the STN Information Keep & Share Program, which allows Recipients to purchase the right to archive and / or redistribute search results from STN databases for internal re-use.

Results from searches carried out by the customer on request, or on behalf of individuals or institutions (see under 2.) outside his own institution (third parties) may only be given to them for explicit internal use. The transmitted search results must include the database producer's copyright. For safety purposes, the customer may keep a copy of the results obtained from a search carried out for third parties.

The customer must obtain database producer's specific written permission for any further uses of search results obtained for third parties, particularly for the transmission of search results in electronic form or their distribution in hardcopy, e.g. sale, loan, license, or free charge.

The customer must do his/her best efforts in preventing a theft or inadvertent illicit dissemination of the records.

4. Warranty and Liability

The database producers shall use their best efforts to deliver correct information in their databases, however, they do not accept warranty and liability for completeness, accuracy and timeliness unless set out differently in II. Special Part.

II. Special Part

1. Database producers, vendors and online information service providers are strictly prohibited from searching, analyzing or displaying USGENE data, for the purpose of direct or indirect comparisons of USGENE database content or coverage to other databases, without the prior written consent of the SequenceBase Corporation.

2. It is strictly prohibited to use USGENE data for the creation of a database for public use, without the prior written consent of the SequenceBase Corporation and FIZ Karlsruhe.

HELP COST

STN International Fees and Prices, Effective Jul 29, 2007

USGENE File	Euro
-----	-----
Connect Hour Fee (per hour) .	83,00
SDI Search Fee (weekly)	11,00
SDI PACKAGE Component Fee 1)	11,00
SDI PACKAGE Component Frequency: weekly	
Display Fee (per answer)	
- AB	0,83
- BIB, IBIB	1,30
- SQIDE, SQ3IDE	3,00
- ECLM	0,46
- CLM	0,92
- ALL, IALL	6,05
- BRIEF, IBRIEF	5,59
- TRIAL (TRI, SAM), FREE, SCAN	FREE
Print Fee (per answer)	
- AB	0,83
- BIB, IBIB	1,30
- SQIDE, SQ3IDE	3,00
- ECLM	0,46
- CLM	0,92
- ALL, IALL	6,05
- BRIEF, IBRIEF	5,59
- TRIAL (TRI, SAM), FREE . .	0,17
Offline Print Postage Fee (additional per answer) . .	0,16
Select Fees (per record)	
- AN, AP, PN, RLN	0,17
Sequence Search	
- Sequence Search per RUN GETSEQ	15,84
- Homology Search per RUN GETSIM	15,84
GETSIM Alert Collection Fee	5,42
GETSIM Batch Initiation Fee	5,42
GETSIM Batch Collection Fee	18,33
- Homology search per RUN BLAST	15,84
BLAST Alert Collection Fee	5,42
BLAST Batch Initiation Fee	5,42
BLAST Batch Collection Fee	18,33
ARCHIVE Per Record Surcharge	
1-25 Users	3,03
26-200 Users	12,10
201-500 Users	30,25
501-1000 Users	42,35
1001+ Users	54,45
REDISTRIBUTE Per Record Surcharge	
2-25 Users	3,03
26-200 Users	12,10
201-500 Users	30,25
501-1000 Users	42,35
1001+ Users	54,45

1) SDI PACKAGE cost is variable. The total monthly fee is a summation of each SDI package component run during the month (plus any associated search term charges and display charges). See HELP COST in each component file for cost and frequency information. Charges are incurred only for the SDI package component runs that complete by the last day of month.

HELP DESK

For detailed help on database content and search strategy, you may contact the nearest STN Service Center. Enter 'HELP STN' for a list of Service Centers.

USGENE via *STN on the Web*

If you are using [STN on the Web](#) to search USGENE you have two options:

1) Use the Sequence Search Assistant for easy menu driven sequence searching without STN command language. Click on Search Assistants and select Sequence Assistant for access to its full functionality. For more details on the Sequence Search Assistant for DGENE, PCTGEN and USGENE see

http://www.stn-international.de/training_center/bioseq/dgene_ssa.pdf

STN on the web

- Help
- News
- Search Assistants
 - Search Assistant
 - Patent Search
 - Alert (SDI) Asst.
 - Structure Query
 - Sequence Asst.
 - Upload Sequence
 - Upload Cmd. File
- Results Assistant
- Transcript Assistant
- => Command Line
- Logoff Hold
- Logoff
- Feedback
- Send Break

Transcript: ON

STN Command List
File-Specific Help List

Sequence Search Assistant

CAS Registry BLAST®:

[Cost Information](#) [\(Plug-in Required\)](#)

Launch CAS Registry BLAST Launch sequence searching and review BLAST Reports [\(Help\)](#) [\(Security\)](#)

Select STN session status following Launch of CAS Registry BLAST:

- Session will time out in 20 minutes
- LOGOFF HOLD (session may be resumed within 120 minutes)
- LOGOFF

Retrieve RNs from BLAST Retrieve a set of RNs previously transferred from a BLAST Report [\(Help\)](#)

DGENE/PCTGEN/USGENE Sequence Searches:

Cost Information: [DGENE](#) [PCTGEN](#) [USGENE](#) [\(No Plug-in Required\)](#)

Launch **BLAST**

in DGENE PCTGEN USGENE

Select the Type of Search:

Continue Conduct a menu-driven sequence search using the selected search algorithm in the selected database [\(Help\)](#)

Show Batch Status Show Status Information about Offline Batch Searches [\(Help\)](#)

Show Alert Status Show Status Information about Alert(SDI) Searches [\(Help\)](#)

BLAST is a registered trademark of the National Library of Medicine

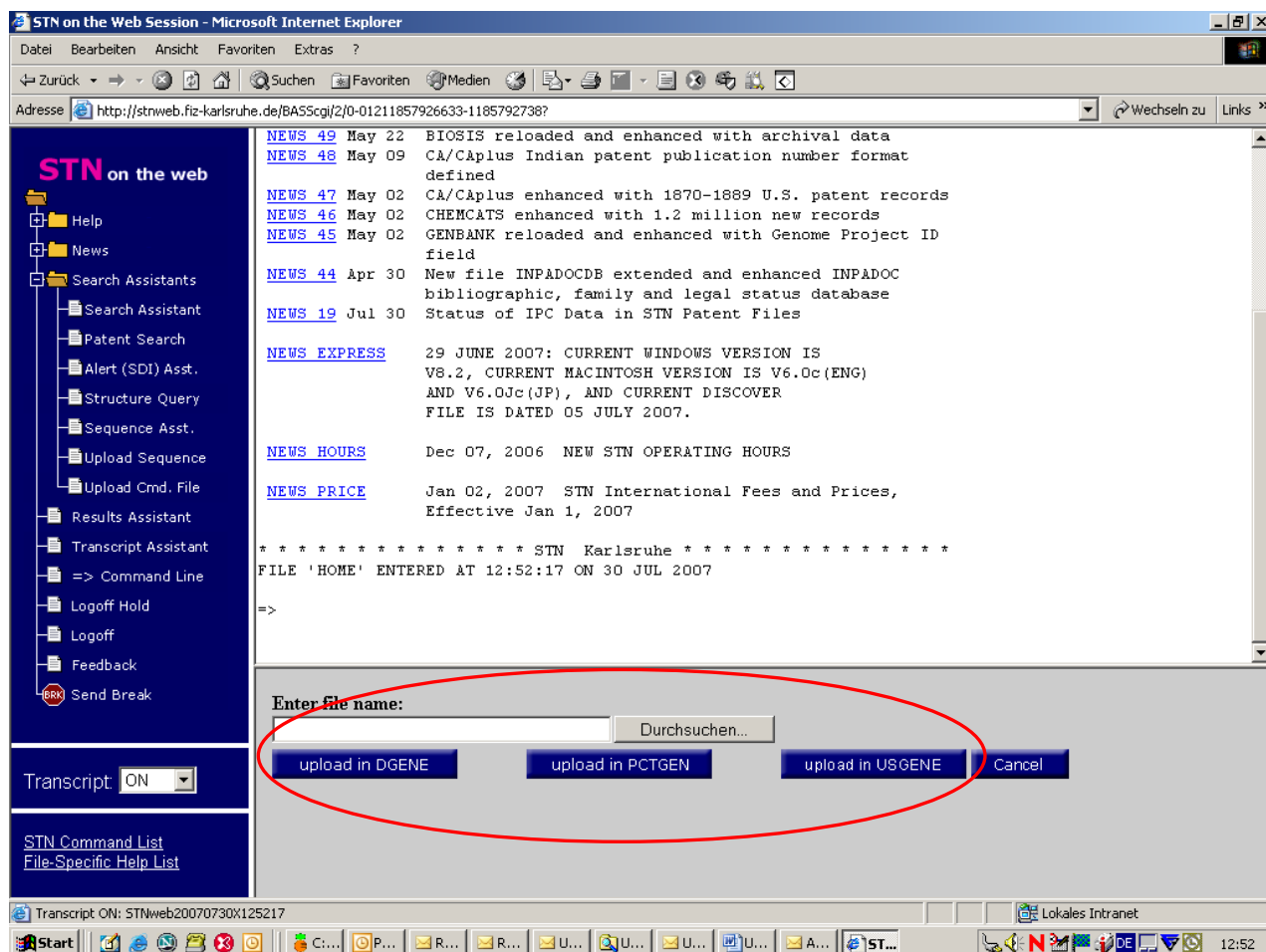
Cancel

Transcript ON: STNweb20070730X115024

Lokales Intranet

11:53

2) You may also search USGENE on STN on the Web in command-line mode. For uploading a sequence query use Upload Sequence under the Search Assistants. Browse your directories for the sequence query file and decide in which database (DGENE, PCTGEN or USGENE) to upload the sequence. Please note that a sequence uploaded in one of the three databases may also be searched in the other using the corresponding L-number.



Use D LQUE to determine that the sequence has UPLOADED correctly, and conduct your sequence search.

STN on the web

FILE LAST UPDATED: 24 JUL 2007 <20070724/UP>
LAST PUBLICATION DATE <20070712/PD>
FILE COVERS 1982 TO DATE

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

>>> See presentation at: <http://www.fiz-k.com/usgene>

>>> DOWNLOAD RUN BLAST/GETSIM FREQUENTLY ASKED QUESTIONS:
<http://www.stn-international.de/service/faq/dgenefaq.pdf> <<<

>>> Introductory offer from August to September 2007 - see NEWS <<<

*** YOU HAVE NEW MAIL ***

=>
UPLOAD SUCCESSFULLY COMPLETED
L1 GENERATED

=> d lque

L1 ANSWER 1 USGENE COPYRIGHT 2007 SEQUENCEBASE CORP on STN
LQUE msspslkwcf tlnyssaaer enflsilkee dvhyavvge vapatgqkhlqgyisikkri
rlggllkkyg srahweiarg tdeenskycs kgtlilelqfpvngsnkrk isemvarspd
rmkieqpeif hryqsvnlk kfkeefvhpcldsqwqilt eaideepddr siiwvygpyg
negkstyaks likkdfytrggkknifls yvdegskhi vfdiprcnqd ylnydvieal
kdrviestkykpikivelgk invivmanfm pdfckisedr ikiiy

=>

Submit Hide session output Show session output

© Fachinformationszentrum Karlsruhe, September 2008

Fachinformationszentrum (FIZ) Karlsruhe
Hermann von Helmholtz Platz 1
76344 Eggenstein-Leopoldshafen
Germany

Tel: +49 7247 808 555

Fax: +49 7247 808 131

Email: helpdesk@fiz-karlsruhe.de

Web: www.fiz-karlsruhe.de