

PCTGEN

COMPLETE HELP TEXT

© Fachinformationszentrum Karlsruhe, July 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 PCTGEN	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.....	17
HELP TLATION.....	21
HELP SBATCH	25
HELP SALERT.....	29
HELP GSEQ	33
HELP SQQ.....	36
HELP QLIMITS.....	39
HELP AAC	40
HELP NUC	42
HELP SQL	43
HELP NCBI	44
HELP ALIGNMENT	45
OTHER GENERAL HELP FOR PCTGEN.....	47
HELP ACCESSION.....	47
HELP FIELDS	47
HELP SFIELDS	48
HELP SRTFIELDS	49
HELP EFIELDS	50
HELP DFIELDS	51
HELP FORMAT	52
HELP CROSSOVER	53
HELP UPDATE/SDI.....	53
HELP RANGE.....	54
HELP HIGHLIGHT	54
HELP (L).....	55
HELP (S).....	55
HELP USAGETERMS	56
HELP COST.....	57
HELP DESK	58

Introduction to PCTGEN

HELP CONTENT

You are currently in the PCTGEN file. PCTGEN covers nucleotide and amino acid sequence information submitted to the World Intellectual Property Organisation (WIPO) by patent applicants as a formal part of the application. The file contains nucleotide and peptide sequences and bibliographic data. Sequence and patent information is compiled in this file as given by the patent applicant.

The file comprises about 5.8 million records from August 2001 to date (07/08) of which 5.0 million are nucleic acid and 800,000 are peptide sequences. The file is updated weekly. Automatic current awareness searches (SDIs) and sequence homology ALERTS will also run weekly.

PCTGEN records contain the following fields: accession number (/AN), title (/TI), document type (/DT), molecule type (/MTY), patent assignee (/PA), patent and application information (/PI, /AI), organism species (/ORGN), sequence length (/SQL), sequence (/SEQ), and feature table (/FEAT).

The basic index (/BI) contains single words from the title (/TI), organism name (/ORGN), and the molecule type (/MTY) fields. Both bibliographic information and sequences are fully searchable and displayable. For direct code match or similarity (homology) sequence searching, the use of one of three RUN package options is required. See HELP GSEQ or HELP SIM (HELP GSIM or HELP BLAST), respectively.

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

PCTGEN database summary sheet:

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

HELP USERAIDS

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

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

PCTGEN summary sheet:

http://www.stn-international.de/stndatabases/sum_sheet/PCTGEN.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 PCTGEN 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 PCTGEN 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 PCTGEN 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 PCTGEN file:

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

Information about Biosequence Searching:

```
HELP SSEARCH   - Sequence searching in PCTGEN
HELP SIM       - Sequence similarity (homology) searching
                 (HELP HOMOMOLOGY)
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       - PCTGEN 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 PCTGEN 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 PCTGEN, 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 PCTGEN:

```
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 PCTGEN database for protein and nucleotide sequence data by similarity (homology). It is also possible to search PCTGEN by similarity using the alternative FASTA-based algorithm (see HELP GSIM).

The BLAST(R) software is provided in PCTGEN 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:

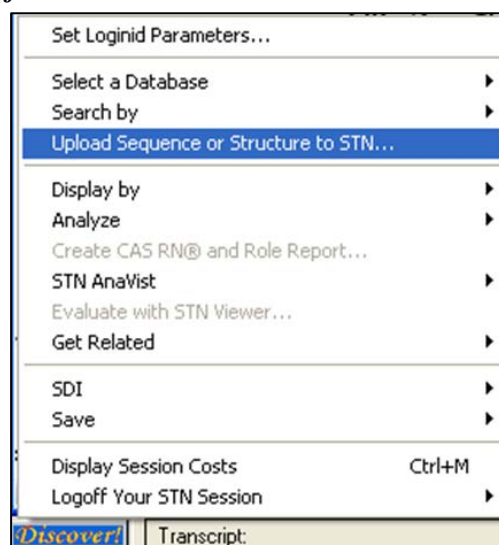
- /SQP** for searching peptide sequences (**BLASTP**) (**default**)
- /SQN** for nucleotide sequences (**BLASTN**)
- /TSQN** for searching peptide sequences translated from PCTGEN nucleotide sequences (**TBLASTN**)

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 DGENE, USGENE or the CAS REGISTRY file.

The minimum length of a sequence query is 5. 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:



The Blast run package also offers the possibility to search for peptides translated from the nucleotide sequences of the database using all three reading frames (/TSQN option). 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 PCTGEN 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). In addition, two values are given, the query self score value defining the maximum score value possible when the query is aligned to itself, and the score value of the best answer of the retrieved answer set. You have three possibilities to select the result answer set.

You can keep:

- 1) the complete answer set (ALL)
- 2) a subset of the complete answer set by specifying a smaller number of just the top scoring answers, or
- 3) you can specify the minimum percentage of the self score value, to keep a subset of the complete answer set, where the answers have a better score than your chosen minimum percentage of the query self score value.

The generated L-number contains all answers or the specified subset of 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 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

```
=> FILE PCTGEN
FILE 'PCTGEN' ENTERED AT 13:23:02 ON 20 MAY 2008
COPYRIGHT (C) 2008 WIPO

=> RUN BLAST gacggcgtggaggtgcataatgccaaagcaaagccgcgaggagcagta/SQN

BLAST Version 2.2

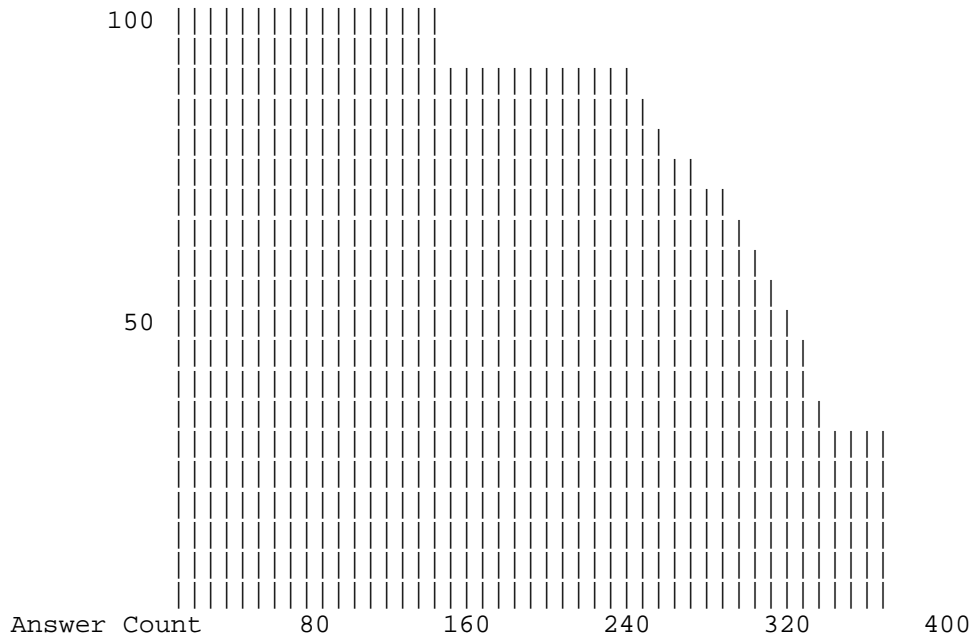
The BLAST software is used herein with permission of the
National Center for Biotechnology Information (NCBI) of
the National Library of Medicine (NLM).....
```

.....

353 ANSWERS FOUND BELOW EXPECTATION VALUE OF 10.0

QUERY SELF SCORE VALUE IS 100
BEST ANSWER SCORE VALUE IS 100

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%

L1 RUN STATEMENT CREATED

L1 233 GACGGCGTGGAGGTGCATAATGCCAAGACAAAGCCGCGGGAGGAGCAGTA
/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 L1

L2 233 SOR L1 SCORE D

=> **d 1 233 SCORE ALIGN SEQ**

L2 ANSWER 1 OF 233 PCTGEN COPYRIGHT 2008 WIPO on STN

SCORE 100 100% of query self score 100

BLASTALIGN

Query = 50 letters

Length = 2034

Score = 99.6 bits (50), Expect = 8e-26

Identities = 50/50 (100%)

Strand = Plus / Plus

Query: 1 gacggcgtggaggtgcataatgccaagacaaagccgcgaggagcagta 50

|||||

Sbjct: 1431 gacggcgtggaggtgcataatgccaagacaaagccgcgaggagcagta1480
SEQ

1 aagcttgccg ccaccatgga atggagctgg gtgttctgt tctttctgtc
51 cgtgaccaca ggcgtgcatt ctgaggtgca gctggtgcag tctggagcag
101 aggtgaaaaa gcccgaggag tctctgaaga tctctgtca gagttttgga

```

.
.
.
1901 tcttcctcta tagcaagctc accgtggaca agagcaggtg gcagcagggg
1951 aacgtcttct catgctccgt gatgcatgag gctctgcaca accactacac
2001 gcagaagagc ctctccctgt ctccgggtaa atag

L2      ANSWER 233 OF 233  PCTGEN COPYRIGHT 2008 WIPO on STN
SCORE 86      86% of query self score 100
BLASTALIGN
  Query   = 50 letters
  Length  = 2475
  Score   = 85.7 bits (43), Expect = 1e-21
  Identities = 43/43 (100%)
  Strand  = Plus / Plus

Query: 8      tggaggtgcataatgccaaagacaaagccgcgggaggagcagta 50
             |||
Sbjct: 1574 tggaggtgcataatgccaaagacaaagccgcgggaggagcagta 1616
SEQ
    1 tttgtacaaa aaagcaggct ttaattaagg aggttaacac catggggcca
    51 accgccatcc tcgcctcct cctggctggt ctccaaggag tctgtgccga
   101 ggtgcagctg gtgcagtctg gacagaggt gaaaaagccc ggggagctct
    .
    .
   1351 tcaccgtgga caagagcagg tggcagcagg ggaacgtctt ctcatgctcc
   1401 gtgatgcatg aggctctgca caaccactac acgcagaaga gcctctcctt
   1451 gtccccgggt aatgagtgac gacaccgagg tacctgctag ggtaaatacc
   1501 cagctttctt gtacaaagtt ggcattataa gaaa

```

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 PCTGEN 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 coil' filter.
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 adapted matrix

```
=> run blast glyspndiavlsqer/sqp -m pam30
```

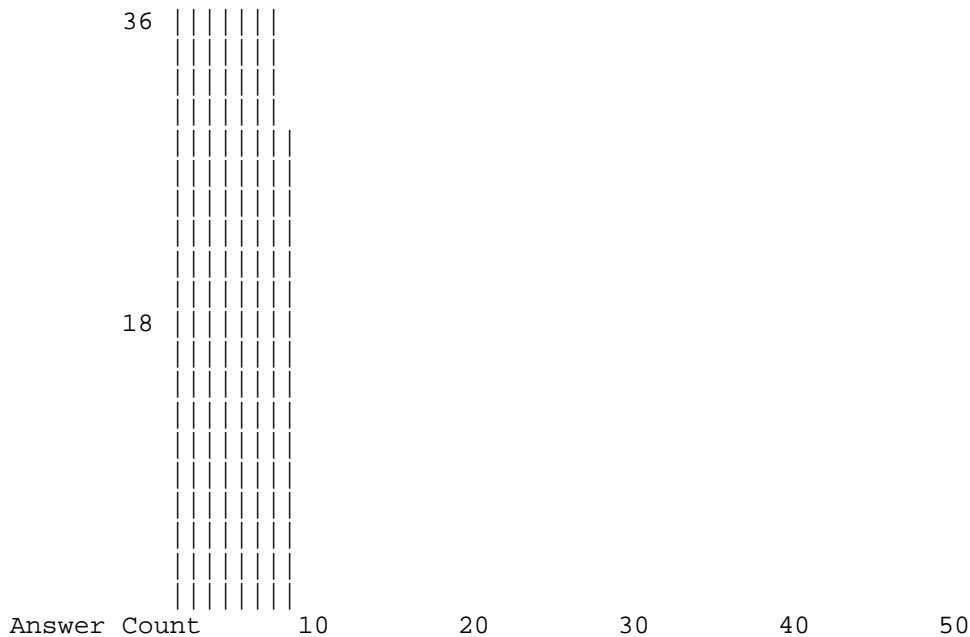
```
BLAST Version 2.2
```

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

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

```
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 ? :65%
```

```
L1 RUN STATEMENT CREATED
```

```
L1 7 GLYSPNDIAVLSQER/SQP.-M PAM30
```

```
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 7 SOR L1 SCORE D
```

```
=> d 1 7 trial score align
```

```
L2 ANSWER 1 OF 7 PCTGEN COPYRIGHT 2008 WIPO on STN
```

```
TI Methods of Diagnosis of Soft Tissue Sarcoma,  
Compositions and Methods for Screening for Soft  
Tissue Sarcoma Modulators
```

```
MTY PRT
```

```
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

L2      ANSWER 7 OF 7  PCTGEN COPYRIGHT 2008 WIPO on STN
TI      HBM Variants that Modulate Bone Mass and Lipid Levels
MTY     PRT
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
```

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.

HELP GSIM

The GETSIM run package is a tool to search the PCTGEN database for protein and nucleotide sequence data by similarity (homology). GETSIM is provided in PCTGEN by FIZ Karlsruhe GmbH and is based upon the FASTA algorithm. It is also possible to search PCTGEN 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 PCTGEN 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 2,000 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 DGENE, USGENE 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 PCTGEN 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 PCTGEN 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 number of retrieved answers is shown as well as the threshold value. In addition, two values are given, the query self score value defining the maximum score value possible when the query is aligned to itself, and the score value of the best answer of the retrieved answer set. You have three possibilities to select the result answer set.

You can keep:

- 1) the complete answer set (ALL)
- 2) a subset of the complete answer set by specifying a smaller number of just the top scoring answers, or
- 3) you can specify the minimum percentage of the self score value, to keep a subset of the complete answer set, where the answers have a better score than your chosen minimum percentage of the query self score value.

The generated L-number contains all answers or the specified subset of 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)

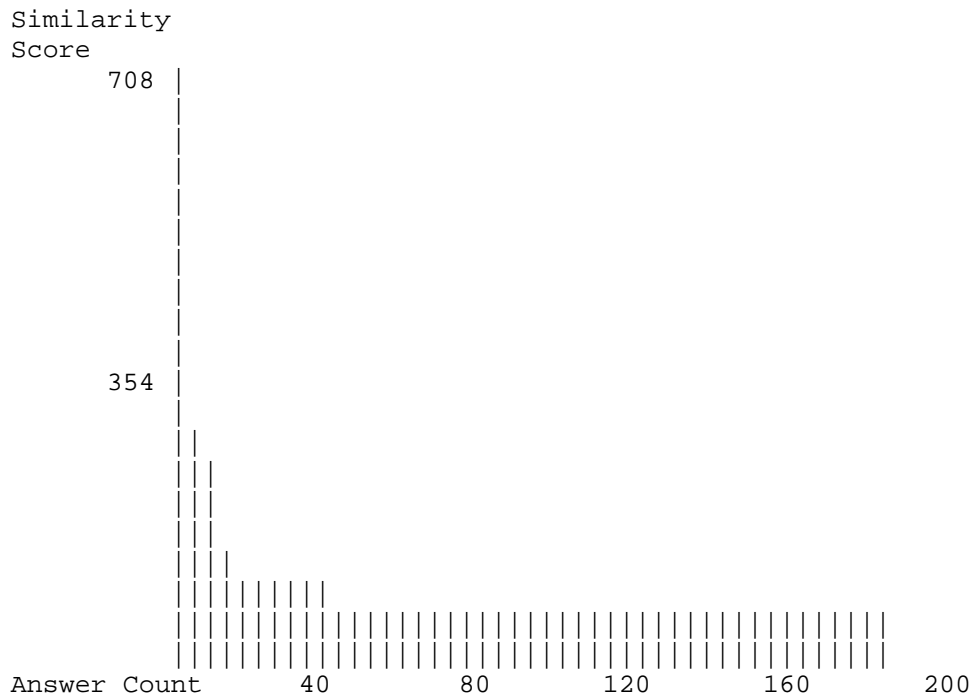
Example : GETSIM /SQP search option

```
=> run getsim aclqeclpkpkpvcgvcrsalapgvavelerqiestetschgcrknff
      lskirshvatcskyqnyimegvkatikdaslqprnvpnrtyfpcpypcpekn/sqp
```

```
RUN GETSIM AT 13:43:29 ON 20 MAY 2008
COPYRIGHT (C) 2008 FIZ KARLSRUHE GMBH
```

```
      80000 SEQUENCES PROCESSED
.....
      730000 SEQUENCES PROCESSED
```

```
179 ANSWERS FOUND ABOVE A THRESHOLD OF 67
      QUERY SELF SCORE VALUE IS 708
      BEST ANSWER SCORE VALUE IS 708
```



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%**

L1 RUN STATEMENT CREATED

L1 4 ACLQECLKPKKPVCGVCRSALAPGVRAVELERQIESTETSCHGCRKNFFL
SKIRSHVATCSKYQNYIMEGVKATIKDASLQPRNVPNRYTFPCPYCPEKN
/SQP

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 4 SOR L1 SCORE D

=> **d 1 4 bib score seq align**

L2 ANSWER 1 OF 4 PCTGEN COPYRIGHT 2008 WIPO on STN
AN 2003009814.455 PRT PCTGEN
TI NOVEL GENES, COMPOSITIONS, KITS, ANDMETHODS FOR
IDENTIFICATION, ASSESSMENT, PREVENTION, AND THERAPY
OF PROSTATE CANCER
PA Millennium Pharmaceuticals, Inc. et al.
PI WO 2003009814 20030206
RLI US 2001-307982P 20010725; US 2001-314356P 20010822;
US 2001-325020P 20010925; US 2001-341746P 20011212;
US 2002-362158P 20020305
ED 20030207
DT Patent
SCORE 708 100% of query self score 708
SEQ

```

1 maaqqrdcgg aaqlagpaae adplgrftcp vclevyekpv qvpcghvfcs
51 aclqeclkpk kpcgvcrsa lapgvavel erqiestets chgcrknffl
101 skirshvatc skyqnyimeg vkatikdasl qprnvpnryt fpcpycpekn
151 fdqeglvehc klfhstdtks vvcpicasmp wgdpnysan frehiqrrhr
201 fsydtfvdyd vdeedmmnqv lqrsiidq

```

ALIGN Smith-Waterman score: 708

100 aa overlap starting at 51

```
aclqeclpkpkpvcgvcrsalapgvavelerqiestetschgcrknfflskirshvat
.....
aclqeclpkpkpvcgvcrsalapgvavelerqiestetschgcrknfflskirshvat
skyqnyimegvmkatikdaslqprnvprnrytfcpcycpekn
.....
skyqnyimegvmkatikdaslqprnvprnrytfcpcycpekn
```

```
L2 ANSWER 4 OF 4 PCTGEN COPYRIGHT 2008 WIPO on STN
AN 2001055322.837 PRT PCTGEN
TI Nucleic Acids, Proteins, and Antibodies
PA Human Genome Sciences, Inc., et al.
PI WO 2001055322 20010802
RLI US 2000-179065P 20000131; US 2000-180628P 20000204;
US 2000-214886P 20000628; US 2000-217487P 20000711;
US 2000-225758P 20000814; US 2000-220963P 20000726;
US 2000-217496P 20000711; US 2000-225447P 20000814;
.
.
US 2000-246611P 20001108; US 2000-230437P 20000906;
US 2000-251990P 20001208; US 2000-251988P 20001205;
US 2000-251030P 20001205; US 2000-251479P 20001206;
Not assigned 2000-12-05; Not assigned 2000-12-01;
US 2000-251989P 20001208; US 2000-250391P 20001201;
Not assigned 2000-12-11
ED 20020923
DT Patent
SCORE 708 100% of query self score 708
SEQ
```

```
1 aremaaqrdr cggaaXlagp aeadplgrf tcpvclevye kpvqvpcghv
51 fcsaclqecl kpkpvcgvc rsalapgvra velerqiest etschgcrkn
101 fflskirshv atcskyqnyi megvmkatikd aslqprnvprn rytfcpcycp
151 eknfdqeglv ehcklfhstd tksvvcpica smpwgdprnyr sanfrehiqr
201 rhrfsydtfv dydvdeedmm nqvlqrsiid q
```

```
ALIGN Smith-Waterman score: 708
100 aa overlap starting at 54
aclqeclpkpkpvcgvcrsalapgvavelerqiestetschgcrknfflskirshvat
.....
aclqeclpkpkpvcgvcrsalapgvavelerqiestetschgcrknfflskirshvat
skyqnyimegvmkatikdaslqprnvprnrytfcpcycpekn
.....
skyqnyimegvmkatikdaslqprnvprnrytfcpcycpekn
```

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 nucleotide sequences which have been translated to all potential derived protein sequences. For the GETSIM TSQN search option FIZ Karlsruhe provides a searchable database of peptide sequences which have been translated from PCTGEN nucleotide sequences. A translation table based on the Universal Genetic Code is employed, using all three reading frames of the corresponding nucleotide sequences. This translated database is searched when the GETSIM TSQN option is chosen. The BLAST TSQN search option uses the general nucleotide sequence database. Here, the algorithm itself translates all nucleotides into potential proteins and searches against these translated sequences. The alignment after a TSQN search 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 PCTGEN 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). In addition, two values are given, the query self score value defining the maximum score value possible when the query is aligned to itself, and the score value of the best answer of the retrieved answer set. You have three possibilities to select the result answer set.

You can keep:

- 1) the complete answer set (ALL)
- 2) a subset of the complete answer set by specifying a smaller number of just the top scoring answers, or
- 3) you can specify the minimum percentage of the self score value, to keep a subset of the complete answer set, where the answers have a better score than your chosen minimum percentage of the query self score value.

The generated L-number contains all answers or the specified subset of 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.

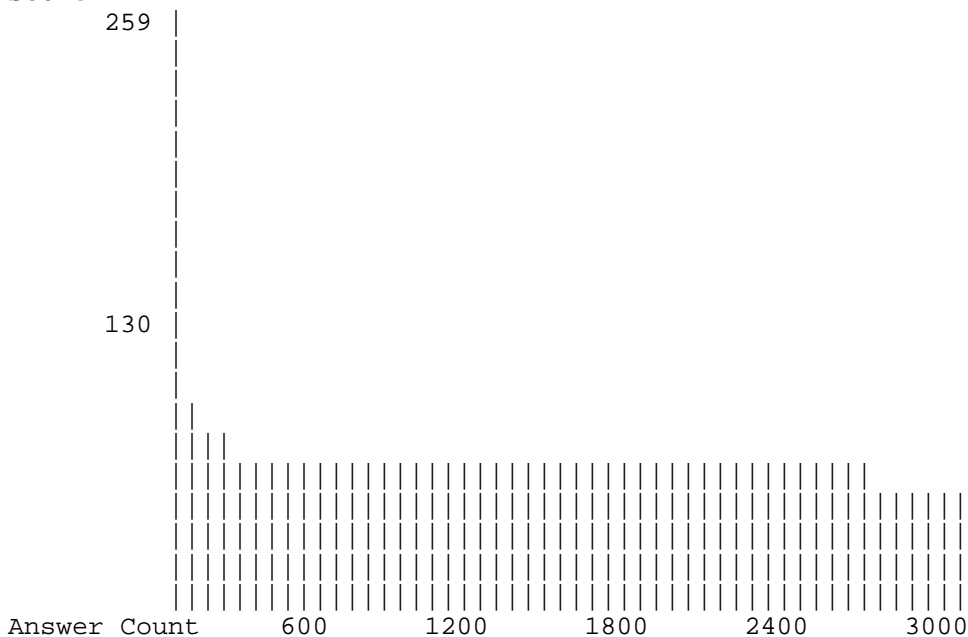
Example : GETSIM /TSQN search option

```
=> run getsim lpkelllrifsfldivtlcrcaqiskawnilaldgsnw/tsqn both
```

```
RUN GETSIM AT 16:02:58 ON 20 MAY 2008  
COPYRIGHT (C) 2008 FIZ KARLSRUHE GMBH
```

```
2998 ANSWERS FOUND ABOVE A THRESHOLD OF 57  
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%
```

```
L3 RUN STATEMENT CREATED
```

```
L3 3 LPKELLRIFSFLDIVTLCRCAQISKAWNILALDGSNW/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 L3
```

```
L4 3 SOR L3 SCORE D
```

```
=> d 1 3 ti score align
```

```
L4 ANSWER 1 OF 3 PCTGEN COPYRIGHT 2008 WIPO on STN
```

```
TI Diagnosis and Prognosis of Breast Cancer Patients
```

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

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

```
38aa overlap starting at 28 (Frame 1-114na overlap starting at 82)
```

```
lpkelllrifsfldivtlcrcaqiskawnilaldgsnw
```

```
::::::::::::::::::::::::::::::::::::::::
```

```
lpkelllrifsfldivtlcrcaqiskawnilaldgsnw
```

```
L4 ANSWER 3 OF 3 PCTGEN COPYRIGHT 2008 WIPO on STN
```

```
TI DETECTION KIT, SUCH AS NUCLEIC ACIDARRAYS, FOR DETECTING  
EXPRESSION OF 10,000 OR MORE DROSOPHILA GENES.
```

```

SCORE 225      86% of query self score 259
ALIGN Smith-Waterman score: 225
 38aa overlap starting at 97 (Frame 1-114na overlap starting at 289)
lpkelllrifsfldivtlcrcaqiskawnilaldgsnw
:::::.....:.....:.....:.....:.....:.....:.....:.....:.....:.....:.....:.....:
lpkevllrvfsyldvsvlcrcaqvckywnvlaldgssw

```

Example : BLAST /TSQN search option

```
=> run blast lpkelllrifsfldivtlcrcaqiskawnilaldgsnw/tsqn
```

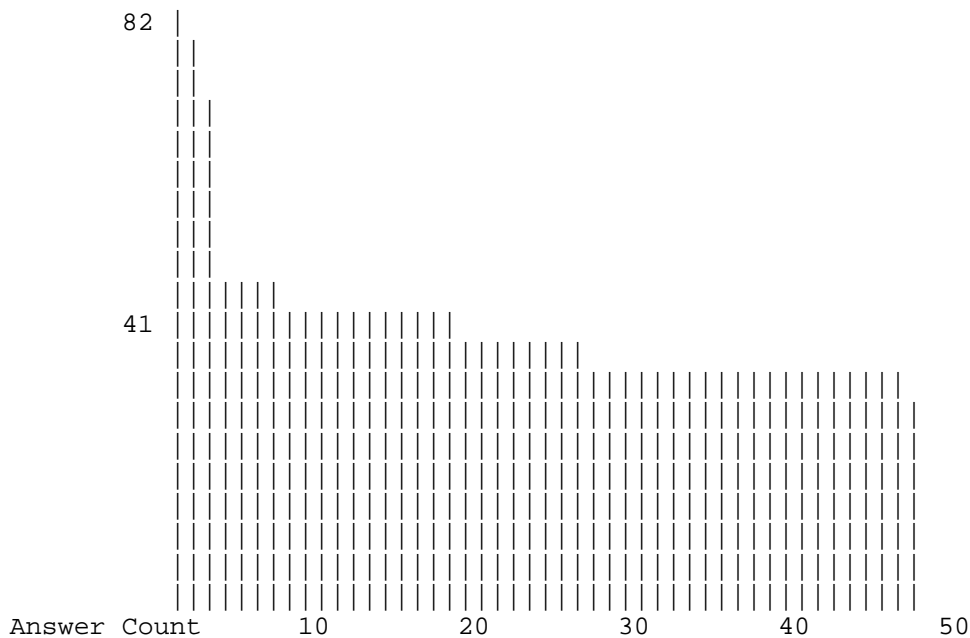
```
47 ANSWERS FOUND BELOW EXPECTATION VALUE OF 10.0
```

```

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 ? :80%
```

```
L1 RUN STATEMENT CREATED
```

```

L1          3 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 L1
```

```
L2          3 SOR L1 SCORE D
```

```
=> d 1 3 ti score align seq
```

```
L2 ANSWER 1 OF 3 PCTGEN COPYRIGHT 2008 WIPO on STN
```

```
TI Diagnosis and Prognosis of Breast Cancer Patients
```

```
SCORE 82      100% of query self score 82
```

BLASTALIGN

Query = 38 letters
Length = 2328
Score = 81.6 bits (200), Expect = 3e-21
Identities = 38/38 (100%), Positives = 38/38 (100%)
Frame = +1

Query: 1 LPKELLRRIFSFLDIVTLRCAQISKAWNILALDGSNW 38
LPKELLRRIFSFLDIVTLRCAQISKAWNILALDGSNW
Sbjct: 82 LPKELLRRIFSFLDIVTLRCAQISKAWNILALDGSNW 195

SEQ

1 gtgacttcgg gctgtgggct cgctcggcgc tcttcggcca tggttttctc
51 aaacaatgat gaaggcctta ttaacaaaaa gttaccctaaa gaacttctgt
.....
2301 aaaattgaat aaaccattgc aatgcttt

L2 ANSWER 3 OF 3 PCTGEN COPYRIGHT 2008 WIPO on STN
TI DETECTION KIT, SUCH AS NUCLEIC ACIDARRAYS, FOR DETECTING EXPRESSION
OF 10,000 OR MORE DROSOPHILA GENES.

SCORE 70 85% of query self score 82

BLASTALIGN

Query = 38 letters
Length = 1518
Score = 70.1 bits (170), Expect = 7e-18
Identities = 28/38 (73%), Positives = 36/38 (94%)
Frame = +1

Query: 1 LPKELLRRIFSFLDIVTLRCAQISKAWNILALDGSNW 38
LPKE+LLR+FS+LD+V+LCRCAQ+ K WN+LALDGS+W
Sbjct: 289 LPKEVLLRVFSYLDVVS LCRCAQVCKYWNVLALDGSSW 402

SEQ

1 aacaacaatc acagcagcaa catcattagc ggcttttgca gcaccatttg
.....
1501 tgctgtgaga ttctctga

HELP SBATCH

Similarity Batch Search

The GETSIM and BLAST run packages are tools for searching PCTGEN 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:

- for the initiation of the batch search, i.e. when RUN GETSIM BATCH L# (or BLAST BATCH L#) is entered, and
- 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. 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 (?):GET
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 15:32:17 ON 20 MAY 2008
COPYRIGHT (C) 2008 FIZ KARLSRUHE GMBH
BATCH PROCESSING STARTED FOR ACTIN1
```

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+.

Entering a second Batch search:

```
=> RUN GETSIM L2/SQN COM BATCH
PLEASE ENTER BATCH IDENTIFIER (MAX. 8 CHARS):AQUAPOR
RUN GETSIM AT 15:33:59 ON 20 MAY 2008
COPYRIGHT (C) 2008 FIZ KARLSRUHE GMBH
PREVIOUS BATCH REQUEST STILL RUNNING
BATCH PROCESSING QUEUED FOR AQUAPOR
```

Status Check:

```
=> RUN GETBATCH ACTIN1
  Please enter your batch identifier
    or enter # for batch id list
    or enter * for batch id at top of list
    or enter - before batch id to delete
    or enter . for (end)
REQUESTED BATCH RESULT FILE STILL RUNNING
Please enter your batch identifier
  or enter # for batch id list
  or enter * for batch id at top of list
  or enter - before batch id to delete
  or enter . for (end)
BATCH REQUEST: end
```

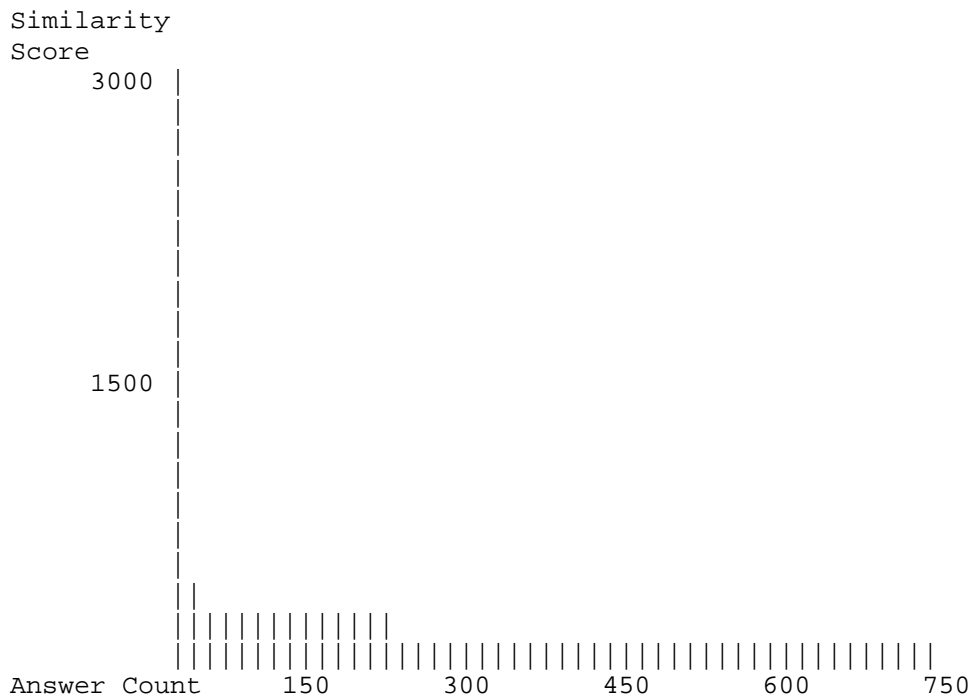
Status Check and Listing of All Open Batch Requests:

```
=> RUN GETBATCH
  Please enter your batch identifier
    or enter # for batch id list
    or enter * for batch id at top of list
    or enter - before batch id to delete
    or enter . for (end)
BATCH REQUEST: #
Batch result files remaining:
  ACTIN1    Running    (getsim)
  PRJ17     Completed (blast)
  AQUAPOR   Queued      (getsim)
-----
Please enter your batch identifier
  or enter # for batch id list
  or enter * for batch id at top of list
  or enter - before batch id to delete
  or enter . for (end)
BATCH REQUEST: end
```

Part 2: Collection of Batch Search Result:

```
=> RUN GETBATCH ACTIN1
.....

  709 ANSWERS FOUND ABOVE A THRESHOLD OF 144
  QUERY SELF SCORE VALUE IS 3000
  BEST ANSWER SCORE VALUE IS 3000
```



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%**

L1 RUN STATEMENT CREATED

```
L1      3 ATGGCCCTGAAGAATGATGAGATAATAGATGCCACTCAAAAAGGAAATTG
CTCTCGTTTTTCATGAATCACAGCTGTGAACCAAATTGTGAAACCCAAAAAT
GGACTGTGAACGGACAACACTGAGGGTTGGGTTTTTTTACCACCAAACCTGGTT
CCTTCAGGCTCAGAGTTAACGTTTGACTATCAGTTCCAGAGATATGGAAA
AGAAGCCCAGAAATGTTTCTGCGGATCAGCCAATTGCCGGGGTTACCTGG
GAGGAGAAAACAGAGTCAGCATCAGAGCAGCAGGAGGGAAAATGAAGAAG
GAACGATCTCGTAAGAAGGATTCAGTGGATGGAGAGCTAGAAGCTCTGAT
GGAAAATGGTGAGGGTCTCTCTGATAAAAACCAGGTGCTCAGCTTATCCC
GGCTAATGGTTAGAATTGAACTTTGGAGCAGAACTTACCTGTCTGGAA
CTCATAACAGAACACACACTCACAGTCCTGCCTGAAGTCCTTTCTGGAACG
TCATGGGCTGTCTTTGTTGTGGATCTGGATGGCAGAGCTAGGTGACGGCC
GGGAAAGTAACCAGAAGCTTCAGGAAGAGATTATAAAGACTTTGGAACAC
/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:

```
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 PCTGEN 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.

Initiation of 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 FEB 2003
COPYRIGHT (C) 2003 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 FEB 2003
COPYRIGHT (C) 2003 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
gacctctttgtcttcacagcatcggttctgccctgggcttcaaatacc
.....
tgggtccagaagacgtggtctagaccagggctgctctttccacttgcct
gtgttctttcccagggcatgactgtcgcacacgcctctgcatatag
tctcttgagttggaatttcattatatgttaagaaaataaaggaaaatg
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

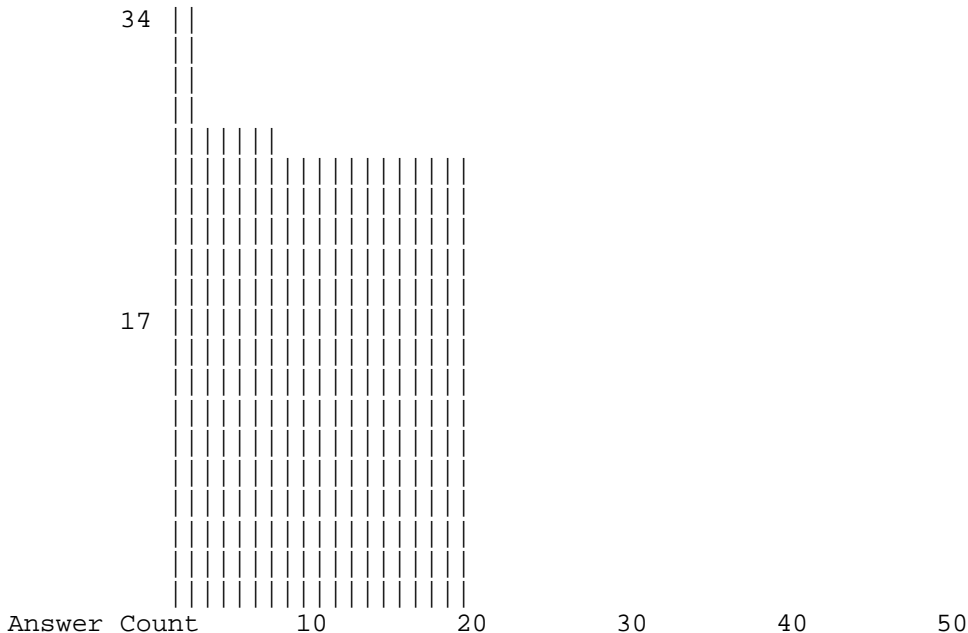
      CURRENT RESULTS AVAILABLE
      NAME          RUN DATE
1) AQUAPNB       20080518 (blast)
2) AQUAPNG       20080518 (No answers - getsim)
3) CLONTECB      20080518 (blast)
4) CLONTECG      20080518 (No answers - getsim)
-----

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
.....
```

19 ANSWERS FOUND BELOW EXPECTATION VALUE OF 10.0

QUERY SELF SCORE VALUE IS 1733
BEST ANSWER SCORE VALUE IS 34

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 1%)
ENTER (ALL) OR ? :.

```
L1 RUN STATEMENT CREATED
L1 19 GCACCCGGCAGCGGTCTCAGGCCAAGCCCCCTGCCAGCATGGCCAGCGAG
    TTCAAGAAGAAGCTCTTCTGGAGGGCAGTGGTGGCCGAGTTCCTGGCCAC
    .....
    GTGTTCTTTCCCCAGGGGCATGACTGTGCGCCACACGCCTCTGCATATATG
    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	20080518 (blast)
2) AQUAPNG	20080518 (No answers - getsim)
3) CLONTECB	20080518 (blast)
4) CLONTECG	20080518 (No answers - getsim)

Enter Number of result to be selected
or "-" before Number 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	20080518 (blast)
2)	AQUAPNG	20080518 (No answers - getsim)
3)	CLONTECB	20080518 (blast)
4)	CLONTECG	20080518 (No answers - getsim)

Enter Number of result to be selected
or "-" before Number to delete result
or "Q" to process alert queries
or enter . for (end)

RESULT REQUEST:**2**

NO ANSWERS FOUND ABOVE A THRESHOLD OF 662
QUERY SELF SCORE VALUE IS 8310

CURRENT RESULTS AVAILABLE

	NAME	RUN DATE
1)	AQUAPNB	20080518 (blast)
2)	AQUAPNG	20080518 (No answers - getsim)
3)	CLONTECB	20080518 (blast)
4)	CLONTECG	20080518 (No answers - getsim)

Enter Number of result to be selected
or "-" before Number to delete result
or "Q" to process alert queries
or enter . for (end)

RESULT REQUEST:**END**

Uncollected ALERT results will be purged from the system after 3 months. Since up to sixteen simultaneous tasks are allowed, up to 192 result sets will be stored. Collected ALERT result sets will stay in the queue till the next update, unless they have been deleted by the customer. 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 PCTGEN 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 DGENE, USGENE 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 MCLHFLVLVICIL/SQSP
L1  QUE MCLHFLVLVICIL/SQSP

=> RUN GETSEQ L1

RUN GETSEQ AT 13:19:47 ON 23 JUL 2008
COPYRIGHT (C) 2008 FIZ KARLSRUHE GMBH

L2  RUN STATEMENT CREATED
L2      8 MCLHFLVLVICIL/SQSP

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

RUN GETSEQ AT 13:20:18 ON 23 JUL 2008
COPYRIGHT (C) 2008 FIZ KARLSRUHE GMBH

L3  RUN STATEMENT CREATED
L3      8 MCLHFLVLVICIL/SQSP

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

RUN GETSEQ AT 13:20:49 ON 23 JUL 2008
COPYRIGHT (C) 2008 FIZ KARLSRUHE GMBH

L4  RUN STATEMENT CREATED
L4      8 MCLHFLVLVICIL/SQSP

=> RUN GETSEQ MCLHFLVLVICIL/SQSP

RUN GETSEQ AT 13:20:59 ON 23 JUL 2008
COPYRIGHT (C) 2008 FIZ KARLSRUHE GMBH

L5  RUN STATEMENT CREATED
L5      8 MCLHFLVLVICIL/SQSP

=> D HIT

L5  ANSWER 1 OF 8  PCTGEN COPYRIGHT 2008 WIPO on STN
SEQ
      1 mftirsrml hflvlvicil recesvcvcv cvcvclwhlg rvv
      === =====
HITS AT: 8-20
```

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 PCTGEN 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.

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.

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

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.. Variability symbols are allowed.

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 L-numbered sequence answer set from RUN GETSEQ may be combined with any search field in the PCTGEN 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 PCTGEN 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.

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

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 L-numbered sequence answer set from RUN GETSEQ may be combined with any search field in the PCTGEN 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 FLRRIKPK
*	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, or sequence query one or more times	CAT(CTG)+TATT/SQSN is equivalent 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 World Intellectual Property Organisation (WIPO) 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 World Intellectual Property Organisation (WIPO) 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

PCTGEN 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
      318180 PROTEIN/FS
      55978 30-50/SQL
L1    51879 PROTEIN/FS AND 30-50/SQL

=> d all

L1    ANSWER 1 OF 51879 PCTGEN (C) 2003 WIPO
AN    2003004623.516 PRT PCTGEN
TI    Human Secreted Proteins
PA    Human Genome Sciences, Inc.
PI    WO 2003004623 20030116
RLI   US 2001-278650P 20010327; US 2001-950082 20010912; US
      2001-950083 20010912
DT    Patent
ORGN  Homo sapiens
SQL   43
SEQ   1 mftirsrml hflvlvicil recesvcvcv cvcvclwhlg rvv
```

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/blastcgihelp.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 ctggttatggtgcagagagtgtaacattgacaagaggacaaaatacagtcaaggtatcagg
          ||||| ||||| ||||| ||||| | || ||||| ||||| ||||| | |||
Sbjct: 207  ctggttacggtgcagaaagtgtaacactctcacgaggacaaaatactgtcaaaattactgg

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.

Other General HELP for PCTGEN

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 PCTGEN output format is shown below. It consists of three parts: the publication year followed by the publication number of the WIPO/PCT application, and after the dot the sequence identity number from the sequence listing belonging to the corresponding application document.

STN output format ----- 2002102994.12

HELP FIELDS

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

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 PCTGEN 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 -----
/AC	Application Country
/AD	Application Date
/AN	Accession Number
/AP (AI)	Application Number
/AY	Application Year
/BI	Basic Index
/DT (TC)	Document Type
/ED	Entry Date
/FEAT	Feature Table
/FS	File Segment
/MTY	Molecule Type
/ORGN	Organism Name
/PA (CS)	Patent Assignee
/PATS	Patent Number Group
/PC	Patent Country
/PD	Publication Date
/PN	Patent Number
/PY	Publication Year
/RLC	Related Application Country
/RLD	Related Application Date
/RLN (RLI)	Related Application Number
/RLY	Related Application Year
/SEQN	Sequence Identity Number
/SQL	Sequence Length
/TI	Title
/UP	Update Date

All fields are text fields except: AD, AY, ED, RLD, RLY, SEQN, SQL, and UP 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 PCTGEN file are listed below.

Sort Code	Definition
----	-----
AC	Application Country
AD	Application Date
AI	Application Information
AN	Accession Number
AP	Application Number
AY	Application Year
DT (TC)--	Document Type
ED	Entry Date
FS	File Segment
MTY	Molecule Type
ORGN	Organism Name
PA	Patent Assignee
PC	Patent Country
PD	Publication Date
PI	Patent Information
PN	Patent Number
PY	Publication Year
RLC-----	Related Application Country
RLD-----	Related Application Date
RLN-----	Related Application Number
RLY-----	Related Application Year
SEQN	Sequence Identity number
SQL	Sequence Length
SCORE	Similarity Score (GETSIM/BLAST run package)
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 PCTGEN file are listed below.

Display Code -----	Definition -----
AC	Application Country
AD	Application Date
AI	Application Information
AIO	Application Information Original
AP (AI)	Application Number
APPS	Application Number and Related Application Number
AY	Application Year
DT (TC)	Document Type
ED	Entry Date
FS	File Segment
MTY	Molecule Type
ORGN	Organism Name
PA (CS)	Patent Assignee
PC	Patent Country
PD	Publication Date
PN	Patent Number
PY	Publication Year
RLC	Related Application Country
RLD	Related Application Date
RLI	Related Application Information
RLIO	Related Application Information Original
RLN	Related Application Number
RLY	Related Application Year
SEQ	Sequence (1-letter codes)
SEQ3	Sequence (3-letter codes)
SEQN	Sequence Identity Number
SQL	Sequence Length
TI	Title
UP	Update Date

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 -----
AI (AP)	Application Information
AIO	Application Information Original
AN	Accession Number
DT (TC)	Document Type
ED	Entry Date
FEAT	Feature Table
FS	File Segment
MTY	Molecule Type
ORGN	Organism Name
PA (CS)	Patent Assignee
PI (PN, PATS)	Patent Information
RLI	Related Patent Information
RLIO	Related Patent Information Original
SCORE	Similarity Score
SEQ	Sequence (1-letter-codes)
SEQ3	Sequence (3-letter-codes)
SEQN	Sequence Identity Number
SEQO	Sequence Original
SQL	Sequence Length
TI	Title
UP	Update Date

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 PCTGEN 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 -----1)---- Alignment between query and retrieved
                        sequence in a similarity search
                        (RUN GETSIM or RUN BLAST)
ALL -----2)---- AN, MTY, TI, PA, PI, AI, RLI, DT, ORGN,
                        SQL, SEQ, FEAT
IALL -----2)---- ALL, indented with text labels
APPS -----2)---- AI, RLI
BIB-----2)---- AN, MTY, TI, PA, PI, AI, RLI, DT
IBIB -----2)---- BIB, indented with text labels
SQIDE -----3)---- TI, SQL, SEQ, FEAT
SQ3IDE -----3)---- TI, SQL, SEQ3, FEAT
SCAN -----3)---- TI (random display without answer numbers)
TRIAL -----3)---- 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 PCTGEN 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 PCTGEN 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 PCTGEN 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 PCTGEN file is updated weekly. Automatic SDIs and ALERTs are also run weekly. The default print format is **BIB**.

HELP RANGE

Searches in the PCTGEN 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 PCTGEN 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 PCTGEN 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 (L)

The link operator, (L), is used in the PCTGEN file to specify that two terms must occur within the same information unit. Terms with different search field codes, belonging to the same 'unit' of information may be linked. The Basic Index contains information from several fields (TI, ORGN, MTY). Terms from each separate field have been linked with the (L) operator.

Example:

```
=> SEARCH OLIGODENDROGLIOMA TUMOR CELL# (L) DIFFERENTIAT?
```

HELP (S)

The (S) proximity operator is used in the PCTGEN 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.

HELP USAGETERMS

The following database producer's special conditions for use of his database(s) apply to your use of the PCTGEN 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.

HELP COST

STN International Fees and Prices, Effective Jan 1, 2008

PCTGEN File	Euro
-----	-----
Connect Hour Fee (per hour) .	96,00
SDI Search Fee (weekly)	8,60
SDI PACKAGE Component Fee 1)	8,60
SDI PACKAGE Component Frequency: weekly	
Display Fee (per answer)	
- BIB, IBIB	1,06
- SQIDE, SQ3IDE	1,28
- ALL, IALL	2,34
- TRIAL (TRI, SAM), SCAN . .	FREE
Print Fee (per answer)	
- BIB, IBIB	1,06
- SQIDE, SQ3IDE	1,28
- ALL, IALL	2,34
- TRIAL (TRI, SAM)	0,10
Offline Print Postage Fee (additional per answer) . .	0,16
Sequence Search	
- Sequence Search per RUN GETSEQ	9,90
- Homology Search per RUN GETSIM	12,50
GETSIM Batch Initiation Fee	4,00
GETSIM Batch Collection Fee	14,00
GETSIM Alert Collection Fee	8,50
- Homology search per RUN BLAST	12,50
BLAST Batch Initiation Fee	4,30
BLAST Batch Collection Fee	14,00
BLAST Alert Collection Fee	8,50
ARCHIVE Per Record Surcharge	
1-25 Users	1,17
26-200 Users	4,68
201-500 Users	11,70
501-1000 Users	16,38
1001+ Users	21,06
REDISTRIBUTE Per Record Surcharge	
2-25 Users	1,17
26-200 Users	4,68
201-500 Users	11,70
501-1000 Users	16,38
1001+ Users	21,06

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 the 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.

© Fachinformationszentrum Karlsruhe, July 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