

DGENE

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 DGENE	4
HELP CONTENT	4
HELP USERAIDS	5
HELP SSEARCH.....	6
HELP DIRECTORY	7
BIOSEQUENCE SEARCHING	8
HELP SIM.....	8
HELP BLAST	9
HELP OPTIONS	13
HELP GSIM.....	17
HELP TLATION.....	21
HELP SBATCH	25
HELP SALERT.....	29
HELP GSEQ	33
HELP SQQ.....	35
HELP QLIMITS.....	38
HELP AAC	39
HELP NUC	41
HELP SQL	42
HELP NCBI	43
HELP ALIGNMENT	44
OTHER GENERAL HELP FOR DGENE	46
HELP ACCESSION.....	46
HELP FIELDS	46
HELP SFIELDS	47
HELP SRTFIELDS	48
HELP EFIELDS.....	49
HELP DFIELDS	50
HELP FORMAT	51
HELP CROSSOVER	52
HELP UPDATE/SDI.....	52
HELP RANGE	53
HELP RCODE	53
HELP THESAURUS	53
HELP HIGHLIGHT	54
HELP KIND	55
HELP (L).....	57
HELP (S).....	57
HELP USAGETERMS	58
HELP COST.....	59
HELP DESK	60

Introduction to DGENE

HELP CONTENT

You are currently in the DGENE (Derwent Geneseq) file. The DGENE file is a unique database of nucleic and peptide sequences which have been extracted from basic patent documents of the 41 issuing authorities covered by the Derwent World Patents Index (file WPINDEX/WPIDS/WPIX). DGENE includes nucleotide sequences of 10 or more bases, all amino acid sequences of 4 or more residues and nucleic probes and primers of any length. More than half of the sequence data that appears in DGENE is not available in any other public sequence database. The file covers patent literature from 1981 to present, and comprises about 11.5 million sequence records (07/08) of which 8.1 million are nucleic and 3.4 million are peptide sequences. DGENE is updated every two weeks.

DGENE records contain a Thomson Reuters (Scientific) Ltd. enhanced title from WPINDEX, a concise Sequence Description, an English abstract written especially for DGENE by one of Thomson Reuters (Scientific) Ltd. experts, patent information, detailed indexing, a feature table and sequence data. The basic index (/BI) contains single words from the title (/TI), keyword (/KW), abstract (/AB), description (/DESC), and organism name (/ORGN) fields. Both text and sequence data 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 file crossover to the Derwent World Patents Index (files WPIDS/WPINDEX/WPIX), the DWPI accession number is available in all DGENE records. Derwent WPI family information can also be directly displayed within DGENE using the FAM format.

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

DGENE database summary sheet:

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

HELP USERAIDS

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

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

The DGENE Workshop Manual:

http://www.stn-international.com/training_center/bioseq/dgene_wm.pdf

DGENE frequently asked questions about GETSIM/BLAST:

<http://www.stn-international.de/service/faq/dgenefaq.pdf>

DGENE complete online HELP text:

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

Derwent GENESEQ User Guide (Derwent):

http://www.stn-international.com/training_center/bioseq/genesequg.pdf

Independent uniqueness comparison of Geneseq to other sequence databases (Derwent):

http://www.stn-international.com/training_center/bioseq/uniqueness.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 DGENE 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
Useful for short and/or highly conserved sequence queries
From FIZ Karlsruhe GmbH

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 DGENE 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 subsequence queries
```

For information on displaying sequences in the DGENE 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 DGENE file:

```
HELP ACCESSION - DGENE accession number formats
HELP CHANGE    - changes in DGENE
HELP CONTENT   - general DGENE file description
HELP COST      - price schedule for the DGENE file
HELP CROSSOVER - file crossover searching in DGENE
HELP DESK      - information on DGENE 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 DGENE file
HELP FORMAT    - predefined formats for display and print
HELP HIGHLIGHT - highlighting in the DGENE file
HELP KIND      - Patent Kind Codes covered in DGENE
HELP (L)       - (L) operator use in the DGENE file
HELP RANGE     - RANGE parameters for the DGENE file
HELP RCODE     - relationship codes available in the
                 thesaurus in DGENE
HELP (S)       - (S) operator use in the DGENE file
HELP SFIELDS   - list of search field codes
HELP SRTFIELDS - list of sortable fields in the DGENE file
HELP THESAURUS - description of thesaurus in the DGENE file
HELP UPDATE/SDI - manual and automatic update searching
HELP USAGETERMS - use and distribution restrictions applicable
HELP USERAIDS  - useful links to supplementary information
                 on DGENE
```

Information about Biosequence Searching:

```
HELP SSEARCH   - Sequence searching in DGENE
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   - TSQL 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 subsequence
                 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       - DGENE 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 DGENE 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 DGENE, 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 DGENE:

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

The BLAST(R) software is provided in DGENE 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:

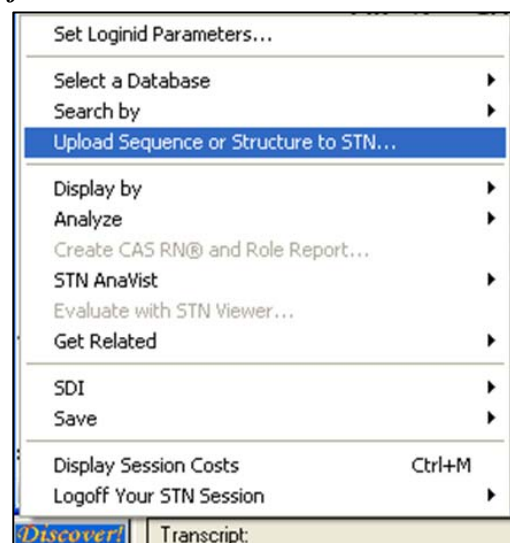
/SQP for searching peptide sequences (**BLASTP**) (**default**)
/SQN for nucleotide sequences (**BLASTN**)
/TSQN for searching peptide sequences translated from DGENE
nucleotide sequences (**TBLASTN**)

When BLAST is used online a query will typically take about 1 minute to complete. 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 USGENE, PCTGEN 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 DGENE 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 DGENE

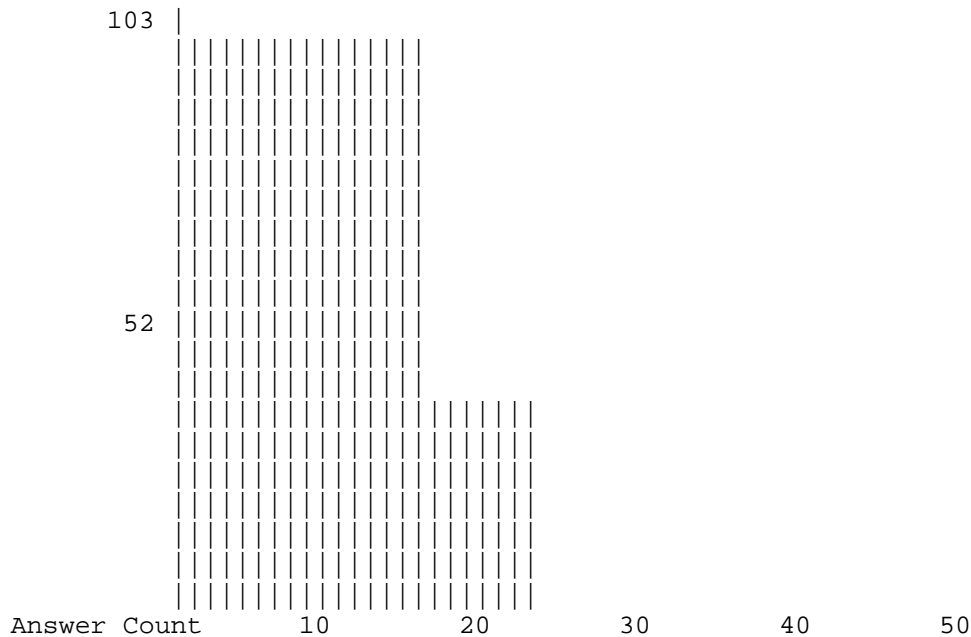
FILE 'DGENE' ENTERED AT 11:12:45 ON 14 MAY 2008
COPYRIGHT (C) 2008 THOMSON REUTERS

=> run blast ggguuuaggagugguaggucuuacgaugccagcuguaaugccuaccggataa/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).....
```

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 16 GGGUUUAGGAGUGGUAGGUCUUACGAUGCCAGCUGUAAUGCCUACCGGAT
AA/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 L1 SCORE D**

PROCESSING COMPLETED FOR L1

L2 16 SOR L1 SCORE D

=> **D 1 16 ALIGN SCORE SEQ**

L2 ANSWER 1 OF 16 DGENE COPYRIGHT 2008 THOMSON REUTERS on STN
BLASTALIGN

Query = 52 letters
Length = 52
Score = 103 bits (52), Expect = 1e-28
Identities = 52/52 (100%)
Strand = Plus / Plus

Query: 1 gggtttaggagtggttaggtcttacgatgccagctgtaatgcctaccggataa 52
||||||||||||||||||||||||||||||||||||||||||||||||||||||||

Sbjct: 1 gggtttaggagtggttaggtcttacgatgccagctgtaatgcctaccggataa 52
SCORE 103 100% of query self score 103

SEQ

1 ggguuuagga gugguagguc uuacgaugcc agcuguaaug ccuaccggat
51 aa

L2 ANSWER 16 OF 16 DGENE COPYRIGHT 2008 THOMSON REUTERS on STN
BLASTALIGN

Query = 52 letters
Length = 4578
Score = 97.6 bits (49), Expect = 7e-25
Identities = 49/49 (100%)
Strand = Plus / Plus

HELP OPTIONS

RUN BLAST Advanced User Options

For introductory instructions on using RUN BLAST in DGENE 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: expectation value is set to a value different from the default

```
=> run blast L1/sqn -e 0.1
```

```
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).....
```

```
.....
```

```
Number of sequences better than 1.0e-01: 68
```

```
.....
```

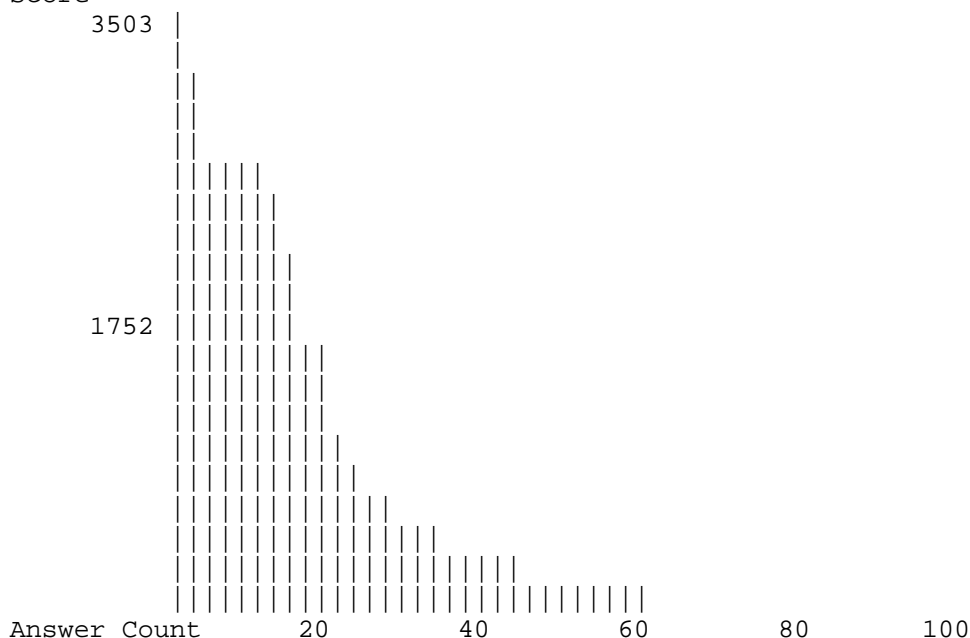
```
68 ANSWERS FOUND BELOW EXPECTATION VALUE OF 1.0e-01
```

```
QUERY SELF SCORE VALUE IS 3503
```

```
BEST ANSWER SCORE VALUE IS 3503
```

```
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%
```

```
L2 RUN STATEMENT CREATED
```

```
L2 4 ATGGCCCTGAAGAATGATGAGATAATAGATGCCACTCAAAAAGGAAATTG  
CTCTCGTTTCATGAATCACAGCTGTGAACCAAATTGTGAAACCCAAAAAT
```

```
.....
```

```
CATTAAGAAGTACATGCAGAAGTTTGGGGCTGTTTACAAACCCAAAGAGG
```

```
ACACTGAATTAGAGTGA/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 L5
```

```
L3 4 SOR L5 SCORE D
```

```
=> d 1 4 score align
```

```
L3 ANSWER 1 OF 4 DGENE COPYRIGHT 2008 THE THOMSON CORP on STN
```

```
SCORE 3503      100% of query self score 3503
BLASTALIGN
  Query  = 1767 letters
  Length = 2510
  Score  = 3503 bits (1767), Expect = 0.0
  Identities = 1767/1767 (100%)
  Strand = Plus / Plus
```

```
Query:1  atggcctgaagaatgatgagataatagatgccactcaaaaaggaaattgctctc
          |||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct:85 atggcctgaagaatgatgagataatagatgccactcaaaaaggaaattgctctc
```

```
.....
Query:1681aaacacaaaaccaaggagtacattaagaagtacatgcagaagtttggggctggtt
          |||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct:1765aaacacaaaaccaaggagtacattaagaagtacatgcagaagtttggggctggtt
```

```
Query:1741cccaaagaggacactgaattagagtga 1767
          |||||||||||||||||||||||||||
Sbjct:1825cccaaagaggacactgaattagagtga 1851
```

```
L3      ANSWER 4 OF 4  DGENE  COPYRIGHT 2008 THE THOMSON CORP on STN
SCORE 3140      89% of query self score 3503
```

```
BLASTALIGN
  Query  = 1767 letters
  Length = 6652
  Score  = 3140 bits (1584), Expect = 0.0
  Identities = 1590/1592 (99%)
  Strand = Plus / Plus
```

```
Query:1  atggcctgaagaatgatgagataatagatgccactcaaaaaggaaattgctctc
          |||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct:3381atggcctgaagaatgatgagataatagatgccactcaaaaaggaaattgctctc
```

```
Query:61  atgaatcacagctgtgaaccaaattgtgaaacccaaaaatggactgtgaacggac
          |||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct:3441atgaatcacagctgtgaaccaaattgtgaaacccaaaaatggactgtgaacggac
```

```
.....
Query:1501aaagtacgaattaaagaccgcaataaactttctacagaggaacgccggaagttgt
          |||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct:4881aaagtacgaattaaagaccgcaataaactttctacagaggaacgccggaagttgt
```

```
Query:1561caagaggtggctcaacgggaggctcagaaaca 1592
          |||||||||||||||||||||||||||
Sbjct:4941caagaggtggctcaacgggaggctcagaaaca 4972
```

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 DGENE database for protein and nucleotide sequence data by similarity (homology). GETSIM is provided in DGENE by FIZ Karlsruhe GmbH and is based upon the FASTA algorithm. It is also possible to search DGENE 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 DGENE 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 in USGENE, PCTGEN 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 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 DGENE 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. Generic codes indexed in DGENE using the International Union of Biochemistry and Molecular Biology (IUBMB) symbols are taken into consideration. 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 DGENE 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 peptide 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 other values are given, the query self score 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

```
=> file dgene

FILE 'DGENE' ENTERED AT 17:02:44 ON 14 MAY 2008
COPYRIGHT (C) 2008 THOMSON REUTERS

=> run getsim LDHILQKTERGVRLHPL...NLSGVNNLEHGLFPQLSAIA/sqp

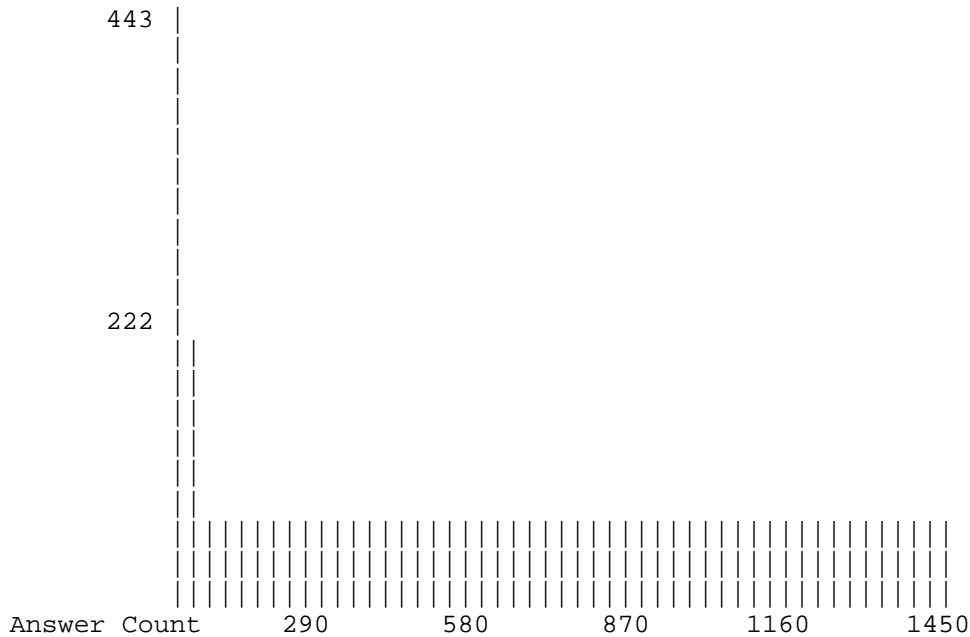
RUN GETSIM AT 17:03:06 ON 14 MAY 2008
COPYRIGHT (C) 2008 FIZ KARLSRUHE GMBH

    120000 SEQUENCES PROCESSED
    .....
    3190000 SEQUENCES PROCESSED

1415 ANSWERS FOUND ABOVE A THRESHOLD OF 57
```

QUERY SELF SCORE VALUE IS 443
BEST ANSWER SCORE VALUE IS 443

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 23 LDHILQKTERGVRLHPLARTAKVKNEVNSFKAALSSLAKHGEYAPFARLL
NLSGVNNLEHGLFPQLSAIA/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 23 SOR L1 SCORE D

=> **d score align seq 1 10 23**

L2 ANSWER 1 OF 23 DGENE COPYRIGHT 2008 THOMSON REUTERS on STN
SCORE 443 100% of query self score 443

ALIGN Smith-Waterman score: 443

70 aa overlap starting at 253

ldhilqktergvrlhplartakvknevnsfkaalsslakhgeyapfarllnlsgvnnleh

.....

ldhilqktergvrlhplartakvknevnsfkaalsslakhgeyapfarllnlsgvnnleh

glfpqlsaia

.....

glfpqlsaia

SEQ

1 rsmdsrpqki wmapshtesd mdyhkiltag lsvqqgivrq rvipvyqvnn

51 leeicqliiq afeagvdfqe sadsfllmlc lhhayqgdyk lflesgavky

101 leghgrfev kkrdgvkrle ellpavssgk nikrtlaamp eetteanag

151 qflsfasfl pklvgegak lekvqrqiv haegqliqyp tawqsvghmm

201 vifrlmrtnf likfllihgg mhmvaghdan davisnsvaq arfsgllivk

251 tvldhilqkt ergvrlhpla rtakvknevn sfkaalssla khgeyapfar

301 llnlsgvnnl ehglfpqlsa ialgvatahg stlagvnvge qyqqldreaat

351 eaekqlqya esreldhlg ddqekkiln fhqkkneisf qqtamvtlr

401 kerlakltea itaaslpkts ghydddddip fpgpinddn pghqdddptd

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 DGENE nucleotide sequences. A translation table based on the Universal Genetic Code is employed, using all three reading frames of the corresponding nucleotide sequences. Generic codes indexed in DGENE using the International Union of Biochemistry and Molecular Biology (IUBMB) symbols are taken into consideration. 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 DGENE 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 14:28:22 ON 15 MAY 2008  
COPYRIGHT (C) 2008 FIZ KARLSRUHE GMBH
```

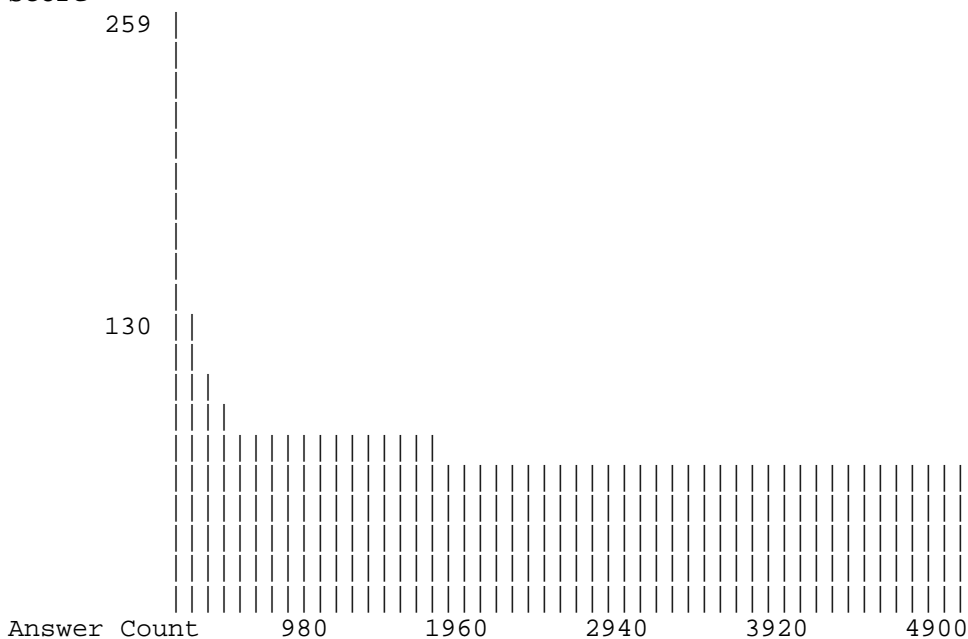
```
160000 SEQUENCES PROCESSED
```

```
.....
```

```
42430000 SEQUENCES PROCESSED
```

```
4895 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%
```

```
L8 RUN STATEMENT CREATED
```

```
L8 40 LPKELLRIFSFLLDIVTLRCAQISKAWNILALDGSNW/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 L8
```

```
L9 40 SOR L8 SCORE D
```

```
=> d 1 40 trial score align seq
```

```
L9 ANSWER 1 OF 40 DGENE COPYRIGHT 2008 THE THOMSON CORP on STN
```

```
AN AER53329 cDNA DGENE
```

```
TI New polypeptide, useful in preparing a composition for  
diagnosing or treating autoimmune disorders or .....
```

```
DESC Human 5' expressed sequence tag, SEQ ID 424.
```

```
KW Cytostatic; Immunosuppressive; Gene therapy; autoimmune  
diseas cancer; expressed sequence tag; EST; ss.
```

```
SQL 228
```

```

SCORE 259          100% of query self score 259
ALIGN Smith-Waterman score: 259
      38aa overlap starting at 37(Frame1-114na overlap starting
      at10
      lpkelllrifsfldivtlcrcaqiskawnilaldgsnw
      ::::::::::::::::::::::::::::::::::::::::::::
      lpkelllrifsfldivtlcrcaqiskawnilaldgsnw
SEQ
      1 aaggacaacg ggcgctgcmr ggcgcgtgtg acttcgggct gtgggctcgc
      .....
      151 atagtaactt tgtgccgatg tgcacagatt tccaaggctt ggaacatctt
      201 agccctggat ggaagcaact ggcagggg

L9 ANSWER 40 OF 40 DGENE COPYRIGHT 2008 THE THOMSON CORP on STN
AN ABL14675 cDNA DGENE
TI New isolated nucleic acid detection reagent for detecting
   1000 or more genes from Drosophila .....
DESC Drosophila melanogaster expressed polynucleotide SEQ
     ID NO 385
KW Drosophila; developmental biology; cell signalling;
   insecticid
SQL 1518
SCORE 225          86% of query self score 259
ALIGN Smith-Waterman score: 225
      38 aa overlap starting at 97(Frame 1-114na overlap starting at
      289
      lpkelllrifsfldivtlcrcaqiskawnilaldgsnw
      :::::::::::::::::::::::::::: : ::::::::::::::
      lpkevllrvfsyldvvsllcrcaqvckywnvlaldgssw
SEQ
      1 aacaacaatc acagcagcaa catcattagc ggcttttgca gcaccatttg
      .....
      201 catggccggc agcgctcaag atcagtcaga ggatcagtcc caaacattcc
      .....

```

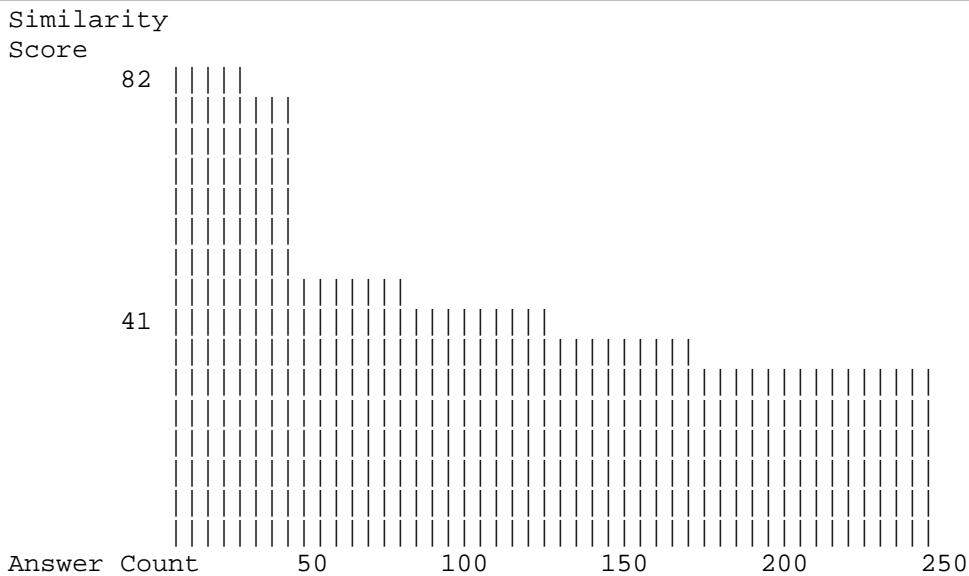
Example : BLAST /TSQN search option

```

=> run blast lpkelllrifsfldivtlcrcaqiskawnilaldgsnw/tsqn -f f

238 ANSWERS FOUND BELOW EXPECTATION VALUE OF 10.0
      QUERY SELF SCORE VALUE IS      82
      BEST ANSWER SCORE VALUE IS     82

```



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%**

L6 RUN STATEMENT CREATED

L6 40 LPKELLRIFSFLLDIVTLRCRAQISKAWNILALDGSNW/TSQN.-F F

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

=> **sor score d**

PROCESSING COMPLETED FOR L6

L7 40 SOR L6 SCORE D

=> **d 1 trial score align seq**

L7 ANSWER 1 OF 40 DGENE COPYRIGHT 2008 THE THOMSON CORP on STN

AN AER53329 cDNA DGENE

TI New polypeptide, useful in preparing a composition for
diagnosing or treating

DESC Human 5' expressed sequence tag, SEQ ID 424.

KW Cytostatic; Immunosuppressive; Gene therapy;.....

SQL 228

SCORE 82 100% of query self score 82

BLASTALIGN

Query = 38 letters

Length = 228

Score = 81.6 bits (200), Expect = 7e-22

Identities = 38/38 (100%), Positives = 38/38 (100%)

Frame = +1

Query: 1 LPKELLRIFSFLLDIVTLRCRAQISKAWNILALDGSNW 38

LPKELLRIFSFLLDIVTLRCRAQISKAWNILALDGSNW

Sbjct: 109 LPKELLRIFSFLLDIVTLRCRAQISKAWNILALDGSNW 222

SEQ

1 aaggacaacg ggcgctgcmr gcgccgtgtg acttcgggct gtgggctcgc

.....

201 agccctggat ggaagcaact ggcagggg

HELP SBATCH

Similarity Batch Search

The GETSIM and BLAST run packages are tools for searching DGENE 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. An average GETSIM batch search will usually be finished within one hour, but may take longer if the system load is high.

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. 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 PROC
```

```
UPLOAD SUCCESSFULLY COMPLETED
```

```
L1 GENERATED
```

```
=> RUN GETSIM L1/SQN BATCH
```

```
PLEASE ENTER BATCH IDENTIFIER (MAX. 8 CHARS):ACTIN2
```

```
RUN GETSIM AT 14:08:45 ON 15 MAY 2008
```

```
COPYRIGHT (C) 2008 FIZ KARLSRUHE GMBH
```

```
BATCH PROCESSING STARTED FOR ACTIN2
```

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 L5/SQN BOTH BATCH
```

```
PLEASE ENTER BATCH IDENTIFIER (MAX. 8 CHARS):aquapor
```

```
RUN GETSIM AT 14:10:09 ON 15 MAY 2008
```

```
COPYRIGHT (C) 2008 FIZ KARLSRUHE GMBH
```

```
PREVIOUS BATCH REQUEST STILL RUNNING
```

```
BATCH PROCESSING QUEUED FOR AQUAPOR
```

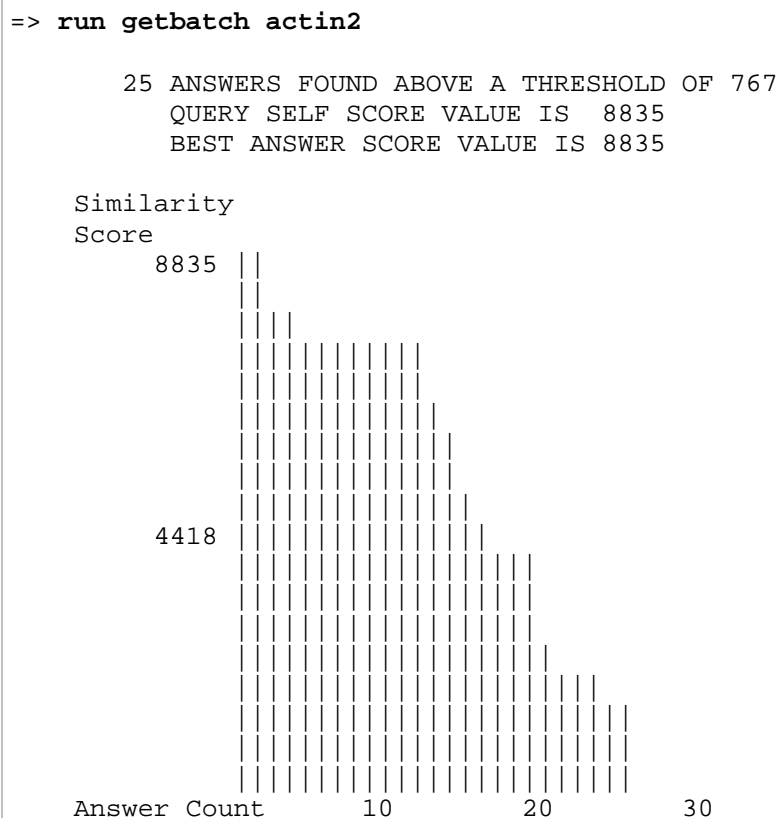
Status Check:

```
=> RUN GETBATCH ACTIN2
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
.....
```

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:#
  ACTIN1    Completed (blast)
  ACTIN2    Running   (getsim)
  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:



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 ? :90%

L6 RUN STATEMENT CREATED

L6 3 ATGGCCCTGAAGAATGATGAGATAATAGATGCCACTCAAAAAGGAAATTG
.....
ACACTGAATTAGAGTGA/SQN.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 L6

L7 3 SOR L6 SCORE D

=> **d 1 3 trial score align**

L7 ANSWER 1 OF 3 DGENE COPYRIGHT 2008 THE THOMSON CORP on STN
AN AAH75515 cDNA DGENE

TI Huntingtin protein interactive protein 65 and polynucleotide
for coding said polypeptide -

DESC Human huntingtin interacting protein 65 encoding cDNA.

KW Human; huntingtin; cancer; HIV; human immunodeficiency
virus; infection; ss.

SQL 2510

SCORE 8835 100% of query self score 8835

ALIGN Smith-Waterman score: 8835

1767 na overlap starting at 85

atggccctgaagaatgatgagataatagatgccactcaaaaaggaaattgctctcgttt

.....

atggccctgaagaatgatgagataatagatgccactcaaaaaggaaattgctctcgttt

.....

cccaaagaggacactgaattagagtga

.....

cccaaagaggacactgaattagagtga

L7 ANSWER 3 OF 3 DGENE COPYRIGHT 2008 THE THOMSON CORP on STN
AN ADX06552 DNA DGENE

TI Biomarkers useful for predicting or determining the response
of a mammal to a cancer treatment comprising administration
of a modulator of cyclin- dependent kinase activity.

DESC Cyclin-dependent kinase modulation biomarker DNA SEQ ID
NO 111

KW cytostatic; cyclin-dependent kinase; cdk; biomarker; gene;
ds.

SQL 7311

SCORE 7964 90% of query self score 8835

ALIGN Smith-Waterman score: 7964

1609 na overlap starting at 3774

atggccctgaagaatgatgagataatagatgccactcaaaaaggaaattgctctcgtttc

.....

atggccctgaagaatgatgagataatagatgccactcaaaaaggaaattgctctcgtttc

.....

caagaggtggctcaacgggaggctcagaacatctggctcgcaagctga

.....

caagaggtggctcaacgggaggctcagaacaacagcaacagatgcaga

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 DGENE 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 96 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):AQUAP1
PLEASE ENTER ALERT TITLE (MAX. 40 CHARS):AQUAPORINNG

RUN GETSIM AT 09:22:30 ON 19 MAR 2002
COPYRIGHT (C) 2002 FIZ KARLSRUHE GMBH

NEW ALERT CREATED

or

=> RUN GETSIM L1/SQN BOTH ALERT
PLEASE ENTER ALERT IDENTIFIER (MAX. 8 CHARS):AQUAP1
PLEASE ENTER ALERT TITLE (MAX. 40 CHARS):AQUAPORINNG

RUN GETSIM AT 09:22:30 ON 19 MAR 2002
COPYRIGHT (C) 2002 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):AQUAP2
PLEASE ENTER ALERT TITLE (MAX. 40 CHARS):AQUAPORINNB

BLAST Version 2.2

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

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)	AQUAP1	20020319	GETSIM	AQUAPORINNG
2)	AQUAP2	20020319	BLAST	AQUAPORINNB

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

QUERY REQUEST:1

ALERT NAME: AQUAP1

INSTALLED : 20020319

TITLE : AQUAPORINNG

ccggggatccacgcgcgcccaccctgcccgcccagcagcgcgcccgc
ctgccccgccatgggtcgacagaaggagctgggtgtcccgctgccccgaga

.....

tcggaaggatctgctattggggaccagagacagggaggcagcctgtcca
tctgtgcataaggagaggaaagttccaggggtgtgtatgttttcaggggcc
ttcacatggaggagctgcagatagatatgtgtttctccggaa/sqn.both

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)	AQUAP1	20080516 (No answers - getsim)
2)	AQUAP2	20080516 (blast)

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

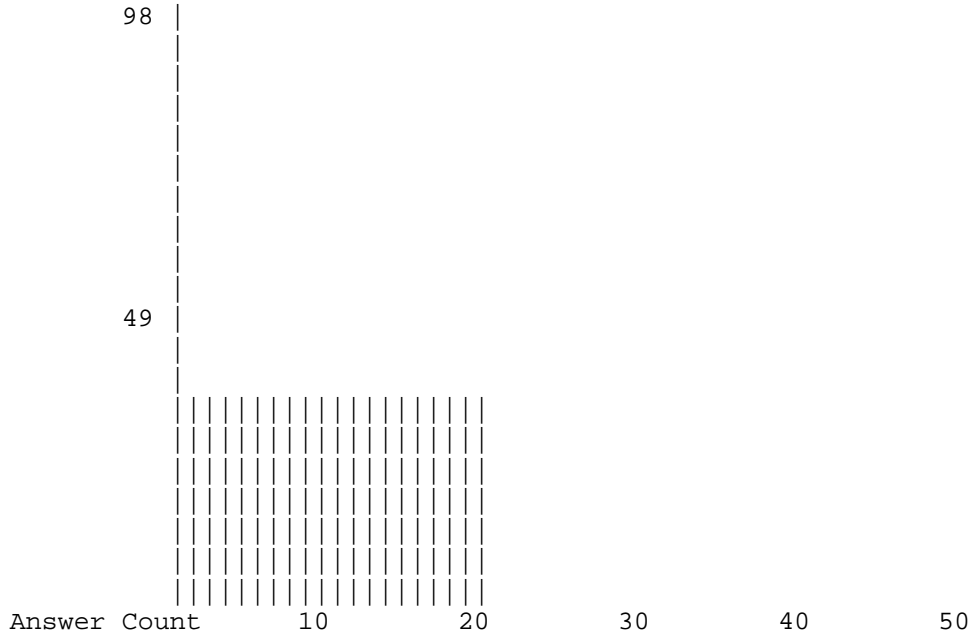
RESULT REQUEST:2

.....

20 ANSWERS FOUND BELOW EXPECTATION VALUE OF 10.0

QUERY SELF SCORE VALUE IS 2278
BEST ANSWER SCORE VALUE IS 98

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 4%)
ENTER (ALL) OR ? :3%

```
L1 RUN STATEMENT CREATED
L1 1 CCGGGATCCACGCGCGCCGCCACCCCTGCCCGCCCGACAGCGCCGCCGC
    CTGCCCGCCCATGGGTTCGACAGAAGGAGCTGGTGTCCCGCTGCGGGGAGA
    .....
    TCTGTGCATAAGGAGAGGAAAGTTCCAGGGTGTGTATGTTTTTCAGGGGCC
    TTCACATGGAGGAGCTGCAGATAGATATGTGTTTCTCCGGAA/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) AQUAP1	20080516 (No answers - getsim)
2) AQUAP2	20080516 (blast)

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) AQUAP1     20080516 (No answers - getsim)
2) AQUAP2     20080516 (blast)
-----

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:1

NO ANSWERS FOUND ABOVE A THRESHOLD OF 442
    QUERY SELF SCORE VALUE IS 7210

CURRENT RESULTS AVAILABLE
NAME          RUN DATE
1) AQUAP1     20080516 (No answers - getsim)
2) AQUAP2     20080516 (blast)
-----

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 96 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 has been lowered for GETSIM ALERT searches 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 DGENE 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 USGENE, 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 YADAIF/SQSP
L1  QUE YADAIF/SQSP

=> RUN GETSEQ L1
L2  RUN STATEMENT CREATED
L2  637 YADAIF/SQSP

=> RUN GETSEQ
PLEASE ENTER SEQUENCE QUERY : YADAIF
TYPE OF SEQUENCE ? (N OR P): P
L3  RUN STATEMENT CREATED
L3  637 YADAIF/SQSP

=> RUN GETSEQ
PLEASE ENTER SEQUENCE QUERY : YADAIF/SQSP
L4  RUN STATEMENT CREATED
L4  637 YADAIF/SQSP

=> RUN GETSEQ YADAIF/SQSP
L5  RUN STATEMENT CREATED
L5  637 YADAIF/SQSP

=> D HIT

L5  ANSWER 1 OF 637  DGENE  COPYRIGHT 2008 THOMSON REUTERS on STN
SEQ
      1 yadaiftnsy rkvlqqlsar kllqdimr
      =====
HITS AT: 1-6
```

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 DGENE 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 DGENE File (e.g. => S L10 AND US/PC, 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 DGENE 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 DGENE file (e.g. => S L10 AND US/PC, 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 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

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.

If you want to search longer sequences, the UPLOAD command needs to be used. The maximum length for uploaded sequence queries used for RUN GETSEQ is 2,000 characters. For RUN GETSIM, sequence queries uploaded from ASCII files may have a maximum length of 500 characters for /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 (similarity search), also an ALERT feature and BATCH mode are available that allows for sequences as long as 2,000 characters. See also HELP SALERT and HELP SBATCH.

HELP AAC

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 'XXX'). Details about uncommon amino acids in a sequence can be found in the corresponding feature table (FEAT). There are standard abbreviations for the most non-standard amino acids as those proposed by the US Patent and Trade Mark Office (US Official Gazette May 16th 1989).

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	Xxx	Uncommon
Y	Tyr	Tyrosine
Z	Glx	Glutamic acid or Glutamine

The codes B and Z may be used only in subsequence searches (/SQSP and /SQSFP). 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
Xxx	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

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

Modified bases are represented by the parent base or by N with the annotations added in the feature table. There are standard abbreviations for the most common modified bases available.

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 (/SQSN) 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

DGENE has a fully numerically range searchable Sequence Length (SQL) field. In addition, individual amino acid residues and nucleotides are also numerically range searchable in the amino acid AA and nucleic acid NA fields respectively.

Definition -----	Search Code -----
Sequence Length	/SQL
Amino Acid	/AA, /AA.CNT
Nucleic Acid	/NA, /NA.CNT

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.

The /AA and /NA fields are also numeric fields and may be searched in combination with the single letter code for a particular amino acid residue or nucleotide. This is shown in the examples below.

Example 1: A search for sequences with 1 to 10 Alanine (A) residues:

```
=> S 1-10 A/AA
      874837 1-10/AA
      887381 A/AA
L1    454446 1-10 A/AA
      (1-10/AA (S) A/AA)

=> D HIT

L1 ANSWER 1 OF 454446 DGENE (C) 2002 THOMSON DERWENT
AA  2 A; 2 R; 0 N; 0 D; 0 B; 1 C; 0 Q; 1 E; 0 Z; 4 G; 1
H; 0 I; 2 L; 0 K; 0 M; 2 F; 0 P; 3 S; 0 T; 1 W; 0 Y;
1 V; 0 Others
```

Example 2: A search for sequences with more than 10 adenine nucleotides:

```
=> S NA>10 (S) A/NA
      1210089 NA>10
      2080283 A/NA
L2    1075849 NA>10(S)A/NA

=> D HIT 1-2

L2 ANSWER 1 OF 1075849 DGENE (C) 2002 THOMSON DERWENT
NA  55 A; 39 C; 42 G; 43 T; 0 other

L2 ANSWER 2 OF 1075849 DGENE (C) 2002 THOMSON DERWENT
NA  105 A; 57 C; 69 G; 65 T; 0 other
```

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 ctgttatggtgcagagagtgtaacattgacaagaggacaaaatacagtcaaggatcagg
          ||||| ||||| ||||| ||||| | || ||||| ||||| ||||| | |||
Sbjct: 207  ctgttacggtgcagaaagtgtaacactctcacgaggacaaaatactgtcaaaattactgg

Query: 3102 gaaagggtggccatagtggttcaacatttaggtggtgcatggggaggactgttcacaaat
          ||||| ||||| ||||| ||||| || || ||||| ||||| || ||||| ||
Sbjct: 147  gaaagggtggccatagtggttcttcatttaaagtctgtcatgggaaagaatgttcatcaac

Query: 3162 tggactccatgctgctgc----caccttgacaaggtaaatagggatttctgagatagaaaa
          ||| ||||| || ||| || || || ||||| || ||||| ||||| |||||
Sbjct: 87   tggcctccaagccagtcaccacatctggataaggtaaataggatctctgagtttagaaaa

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  QVGKGSVSPNLIHVKLEALEARELIKISILQNC EE-DKQTVAEKISARSGAEIVQVIGRT
          QV + S+  + ++ +K ALEA+EL +  +      +E +K+T          RS + + +
Sbjct: 565 QVQQNSLHRDSLVT LKRALEAKELARQH LRDQLDEVEKET-----RSKLQEIDIFNNQ
Query: 81  IILYKTSV NKQQIK 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
      :: : . .: . .: : ::::: : ::
      agtattcatatttactaacaagccagctggaat      <- 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 nonmatching 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
      ::: :: :: . ::      : :::: .. ::::: .. ::
      vgeygasp lclpyap__pegqpaalgftvalvmmnsfcflvvagayikl      <- Answer
```

The similarity percentage of the Smith-Waterman score is given in the SCORE field and can be displayed with D SCORE. A short, precise description of the corresponding sequence is given in the description field (display with D DESC).

Other General HELP for DGENE

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 format is shown below.

STN output format-----AAR75658

HELP FIELDS

The following messages are available in the DGENE 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 SSEARCH	-	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 DGENE 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 title, abstract, keywords, description, and organism name. The Feature Table field allows simultaneous left and right truncation.

Search Code -----	Definition -----
/AA	Amino Acid
/AA.CNT	Amino Acid Count
/AC	Application Country
/AD	Application Date
/AN	Accession Number
/AP	Application Number
/APPS	Application Number Group
/AY	Application Year
/BI	Basic Index
/CR (XR)	Cross Reference (to related DGENE sequence records)
/DED	Data Entry Date
/DT (TC)	Document Type
/ED	Entry Date
/FEAT	Feature Table
/FS	File Segment
/IN (AU)	Inventor
/KW (ST)	Keyword
/LA	Language
/MTY	Molecule Type
/NA	Nucleic Acid
/NA.CNT	Nucleic Acid Count
/ORGN	Organism Name
/OS	Other Source (DWPI accession number)
/PA (CS)	Patent Assignee
/PACO	Patent Assignee Code
/PATS	Patent Number Group
/PC	Patent Country (WIPO code and text)
/PCS	Patent Countries (WIPO code and text)
/PD	Publication Date
/PK	Patent Kind Code
/PN	Patent Number
/PRC	Priority Country (WIPO code and text)
/PRD	Priority Date
/PRDF	Priority Date, First
/PRN	Priority Number
/PRY	Priority Year
/PRYF	Priority Year, First
/PSL	Patent Sequence Location
/PY	Publication Year
/SQL	Sequence Length
/TI	Title
/UP	Update Date

All fields are text fields except: AA.CNT, AD, AY, DED, ED, NA.CNT, PD, PRD, PRDF, PRY, PRYF, PY and SQL, which are numeric, and can be searched with numeric operators or ranges, e.g. 1990-1991/PRY or 1 APR 90-15 APR 90/PD.

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 DGENE file are listed below.

Sort Code	Definition
----	-----
AC	Application Country
AD	Application Date
AI	Applciation Information
AP	Application Number
AY	Application Year
DED	Data Entry Date
FS	File Segment
IN	Inventor
LA	Language
MTY	Molecule Type
ORGN	Organism Name
OS	Other Source (DWPI accession number)
PA	Patent Assignee
PACO	Patent Assignee Code
PC	Patent Country
PD	Publication Date
PI	Patent Information
PK	Patent Kind Code
PN	Patent Number
PRAI	Priority Information
PRC	Priority Country
PRD	Priority Date
PRDF	Priority Date, First
PRN	Priority Number
PRY	Priority Year
PRYF	Priority Year, First
PSL	Patent Sequence Location
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 DGENE file are listed below.

Display Code -----	Definition -----
AA	Amino Acid
AB	Abstract
AC	Application Country
AD	Application Date
AI	Application Information
AN	Accession Number
AP	Application Number
APPS	Application Number Group
AY	Application Year
CR (XR)	Cross References (to related DGENE records)
DED	Data Entry Date
DT (TC)	Document Type
FEAT	Feature Table
FS	File Segment
IN (AU)	Inventor
KW	Keyword
LA	Language
MTY	Molecule Type
NA	Nucleic Acid
ORGN	Organism Name
OS	Other Source (DWPI accession number)
PA (CS)	Patent Assignee
PACO	Patent Assignee Code
PATS	Patent Number Group
PC	Patent Country
PCS	Patent Countries
PD	Publication Date
PI	Patent Information
PK	Patent Kind Code
PN	Patent Number
PRAI	Priority Information
PRC	Priority Country
PRD	Priority Date
PRDF	Priority Date First
PRY	Priority Year
PRYF	Priority Year First
PSL	Patent Sequence Location
SQL	Sequence Length
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
-----	-----
AA	Amino Acid
AB	Abstract
AI (AP)	Application Information
AN	Accession Number
APPS	Application Number Group
CR (XR)	Cross Reference (to related DGENE sequence records)
DED	Data Entry Date
DESC	Description
DT (TC)	Document Type
FAM	Patent Family Information (from the DWPI database)
FEAT	Feature Table
FS	File Segment
IN (AU)	Inventor
KW (ST)	Keyword
LA	Language
LS	Legal Status Data (from the INPADOCDB database)
LS2	Legal Status Data (from the INPADOCDB database), detailed version with display headers
MTY	Molecule Type
NA	Nucleic Acid
ORGN	Organism Name
OS	Other Source (DWPI accession number)
PA (CS)	Patent Assignee
PATS	Patent Number Group
PI (PN)	Patent Information
PRAI (PRN)	Priority Information
PSL	Patent Sequence Location
SEQ	Sequence (1-letter-codes)
SEQ3	Sequence (3-letter-codes)
SQL	Sequence Length
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 DGENE 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:

```
ABS ----- AN, MTY, AB
ALIGN -----1)----- Alignment between query and retrieved
                        sequence in a similarity search
                        (RUN GETSIM or RUN BLAST)
ALL ----- AN, MTY, TI, IN, PA, PI, AI, PRAI, PSL,
            DED, DT, LA, OS, DESC, KW, ORGN, AB, AA,
            NA, SQL, SEQ, FEAT
IALL ----- As ALL, but indented with text labels
BIB----- AN, MTY, TI, IN, PA, PI, AI, PRAI, DT,
            LA, OS, CR
FAM ----- Family Information from the Derwent World
            Patents Index (PI, ADT, FDT, PRAI)
IBIB ----- As BIB, but indented with text labels
SQIDE ----- AN, MTY, AA, NA, SQL, SEQ, FEAT
SQ3IDE ----- AN, MTY, AA, NA, SQL, SEQ3, FEAT
SCAN ----- TI, AN, DESC
TRIAL ----- AN, MTY, TI, DESC, KW
            (TRI, SAM)

1) Use RUN GETSIM or RUN BLAST first. See HELP SIM, HELP GSIM
   or HELP BLAST
```

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,IN'; 'PI,AI,PRAI'. 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)

=> SEARCH HUMAN INSULIN/TI
      13053 HUMAN/TI
      824  INSULIN/TI
L1      213 HUMAN INSULIN/TI
      ((HUMAN(W)INSULIN)/TI)

(In the DGENE file)

=> SEARCH L1
      14528 HUMAN/TI
      1229  INSULIN/TI
L2      96  HUMAN INSULIN/TI
      ((HUMAN(W)INSULIN)/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, allows L-numbered queries from other STN sequence files, e.g. **USGENE**, **PCTGEN** and the **CAS REGISTRY** file.

HELP UPDATE/SDI

Update searching (also called current-awareness, Alert or SDI searching) can be done manually or automatically in the DGENE 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 DGENE file is updated every two weeks. Automatic SDI's and ALERT sequence searches are run every two weeks. The default print format is **BIB**.

HELP RANGE

Searches in the DGENE 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.

Examples:

```
=> 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 time period.

HELP RCODE

A thesaurus is present in the Patent Assignee Code (/PACO) search field in the DGENE file. To display hierarchies of terms in this thesaurus, use the EXPAND command with a term, followed by a plus symbol, (+), a Relationship Code, and /PACO, e.g. => E GENZYME+ALL/PACO.

To use this thesaurus to automatically include additional other terms in a search, enter the SEARCH command with a term, followed by a plus symbol, (+), a Relationship Code, and /PACO, e.g. => S GENZYME+ALL/PACO.

The following Relationship Codes can be used with the EXPAND and SEARCH commands in the Patent Assignee Code (/PACO) field:

Relationship Code	Description
ALL	All Associated Terms (SELF, DEF, CODE)
CODE	Related Codes (CODE, SELF)
DEF	Definition of the Code (SELF, DEF)

HELP THESAURUS

A thesaurus is present for the Patent Assignee Code (/PACO) field in the DGENE file. When you request an alphabetic EXPAND display of the /PACO field, a column labeled AT (Associated Terms) will be included in which the number indicates the number of terms that are associated with the term in that E-number line in the thesaurus.

You may also use the EXPAND command to request a display of hierarchies of terms in the thesaurus. Use the EXPAND command with a term, followed by a plus symbol, (+), a Relationship Code, /PACO e.g. => E GENZYME+ALL/PACO.

To use this thesaurus to automatically include additional terms in a search, enter the SEARCH command with a term, followed by a plus symbol, (+), a Relationship Code, and /PACO, e.g., => S GENZYME+ALL/PACO.

For a list of Relationship Codes that can be used with the thesaurus in the DGENE file, enter 'HELP RCODE' at an arrow prompt (=>).

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 DGENE 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 'PRODUCTION OF CARBOCYCLIC NUCLEOSIDES', part of the display might look like this:

```
AB ... for their ability to resolve racemic 4-amino-cyclopentane-
carboxylic acid methyl ester derivatives, for possible use as
intermediates in production of carbocyclic nucleosides.
```

or

```
AB ... for their ability to resolve racemic 4-amino-cyclopentane-
carboxylic acid methyl ester derivatives, for possible use as
intermediates in ***production of carbocyclic nucleosides***.
```

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 DGENE 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 KIND

Searching for patent information often requires knowledge of patent kind codes and country codes of corresponding patent issuing authorities. The following table summarises kind codes covered in DGENE in the Patent Kind Code field (/PK).

Abbreviations used in the table:

CC - Country Code
NTIS - National Technical Information Service
OPI - Open for Public Inspection
PCT - Patent Cooperation Treaty
PK - Patent Kind Code
SR - Search Report
UM - Utility Model

CC	PK	Patent Kind covered in DGENE (field /PK)
--	--	-----
AT	A	OPI application
AU	A	OPI application
	B	Examined and accepted patent (from 9308)
BE	A	Patent of invention
BR	A	OPI application
CA	A	Examined granted patent before 01.10.89 (old law) or OPI application from 01.10.89 (new law)
	A1	Examined granted patent before 01.10.89 (old law) or OPI application from 01.10.89 (new law)
	C	Granted patent (old and new law)
CH	A	Granted unexamined patent or examined application
	A5	Granted without examination
CN	A	OPI application
CZ	A3	Unexamined application (from 9417)
DD	A	Examined granted patent
DE	A	OPI application before examination (from 1968)
	A1	Unexamined patent application (from 9301)
	C	Granted patent from 1981 (from 8138)
	C1	Examined patent - first publication (from 9252)
	U	Utility Model (from 9626)
	U1	Utility Model (from 9626)
DK	A	OPI application which has been (i) neither searched nor examined, or (II) searched, but not examined (from 1978)
EP	A	OPI application
	A1	OPI application with SR (from 9220)
	A2	OPI application without SR (from 9221)
ES	A	Patent granted without examination (pre-1987)
	A	Patent application published with search report
	A1	Patent application published with search report
FI	A	OPI application
FR	A	OPI application (from 1969)
	A1	OPI application (first application)
GB	A	Examined granted specification (<2000000)
	A	OPI application (2000000+)
HU	A	OPI application - examination requested
	A	OPI application - examination deferred
	T	English language abstracts of Hungarian patent specifications (from 9223)

IL	A	Application of patent for invention
JP	A	OPI application
	B	Application published after examination (Tokkyo koho) (up to 9626)
	B1	Registered patent without published application
KR	A	Patent application
	B	Examined patent specification
	B1	Examination patent specification (from 9252)
MX	A	Patent of Invention (from 9816)
	A1	Patent application
NL	A	OPI application
	C	Granted patent
	C2	20-years New Law granted patent (from 9608)
NO	A	OPI application
NZ	A	Examined application (from 9301)
PT	A	Application for patent of invention
RD	A	Research Disclosure (1978-2001)
RU	C	Granted patent (from 9406)
	C1	Granted patent of invention
	C2	Granted patent of invention
SE	A	OPI application
	B	Examined accepted specification (from 8701)
SG	A	Registration via GB or EP designating GB
	A1	Patent application (from 9631)
SU	A	Examined granted patent
	A1	Inventor's Certificate
	B	Reissued patent
TP	A	International Technology Disclosure (1984-1993)
TW	A	Unexamined patent application
US	A	Examined granted patent (prior to 02.01.2001)
	A1	Utility Patent Application (from 02.01.2001)
	B	Re-examination Certificate (prior to 02.01.2001)
	B1	Utility Patent Grant no pre-grant publication (from 02.01.2001)
	B1	Re-examination Certificate, 1st reexamination (prior to 02.01.2001)
	H	Defensive Publication
	H	Statutory Invention Registration (S.I.R.)
	N	NTIS published invention application (1983-1996)
WO	A	OPI application
	A1	OPI application with SR (from 9220)
	A2	OPI application without SR (from 9220)
ZA	A	Unexamined accepted specification

HELP (L)

The link operator, (L), is used in the DGENE 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, KW, AB, DESC, ORGN). Terms from each separate field have been linked with the (L) operator.

```
Example: => SEARCH FUSION PROTEIN#(L)HOST CELL#
```

HELP (S)

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

When searching in /BI, (S) works like (L) proximity.

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 Thomson Reuters (Scientific) Ltd Databases Terms of Agreement apply to your use of any Thomson Reuters (Scientific) Ltd database on STN International. For review of these Terms go to

http://www.stn-international.de/service/tsd_ta.html

Any other use of the data without the express written permission of Thomson Reuters (Scientific) Ltd is strictly prohibited.

Thomson Reuters (Scientific) Ltd also participates in the STN Information Keep & Share Program, with additional conditions outlined in Section 1, paragraph iii of the Thomson Reuters (Scientific) Ltd Databases Terms of Agreement.

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.

HELP COST

STN International Fees and Prices, Effective Jan 1, 2008

DGENE File	Euro
-----	-----
Connect Hour Fee (per hour) .	127,00
SDI Search Fee (every 2 weeks)	29,90
SDI PACKAGE Component Fee 1)	29,90
SDI PACKAGE Component Frequency: every 2 weeks	
Display Fee (per answer) . . .	
- BIB, IBIB	3,05
- ABS	3,05
- SQIDE, SQ3IDE	6,20
- ALL, IALL	12,30
- FAM	9,20
- LS, LS2	0,87
(from the INPADOCDB file)	
- TRIAL (TRI, SAM), SCAN . .	FREE
Print Fee (per answer)	
- BIB, IBIB	3,20
- ABS	3,20
- SQIDE, SQ3IDE	7,55
- ALL, IALL	13,95
- FAM	11,25
- LS, LS2	0,87
(from the INPADOCDB file)	
- TRIAL (TRI, SAM)	0,82
Offline Print Postage Fee	
(additional per answer) . .	0,16
Sequence Search	
- Sequence Search per RUN GETSEQ	17,30
- Homology Search per RUN GETSIM	22,80
GETSIM Batch Initiation Fee	8,55
GETSIM Batch Collection Fee	28,10
GETSIM Alert Collection Fee	8,55
- Homology search per RUN BLAST	22,80
BLAST Batch Initiation Fee	8,55
BLAST Batch Collection Fee	28,10
BLAST Alert Collection Fee	8,55
ARCHIVE Per Record Surcharge	
1-25 Users	6,15
26-200 Users	24,60
201-500 Users	61,50
501-1000 Users	86,10
1001+ Users	110,70
REDISTRIBUTE Per Record Surcharge	
2-25 Users	6,15
26-200 Users	24,60
201-500 Users	61,50
501-1000 Users	86,10
1001+ Users	110,70

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 or the Thomson Reuters (Scientific) Ltd Customer Technical Support desks which are listed below. Enter 'HELP STN' for a list of Service Centers.

Thomson Reuters (Scientific) Ltd Customer Technical Support:

Americas	Telephone: +1 800 336 4474 or +1 215 386 0100 Fax: +1 215 386 6362
Europe, Middle East, & Africa	Telephone: +44 20 7433 4999 Fax: +44 20 7433 4001
Asia/Pacific (Singapore)	Telephone: +65 6879 4118 Fax: +65 6223 2634
Japan	Telephone: 0800 888 8855 or +81 3 5218 6500 Fax: +81 3 5218 6536
China	Telephone: +86 10 8286 2099 Fax: +86 (10) 8286 2088
Korea	Telephone: 080 010 8100 or +82 2 2076 8100 Fax: +82 2 2076 8122
Australia & New Zealand	Telephone: Australia 1800 007 214 New Zealand 0800 443 162 Fax: +65 6223 2634

Thomson Reuters (Scientific) Ltd Customer Technical Support Contact Form
<http://thomsonscientific.com/support/techsupport/>

© 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