DGENE (Derwent Geneseq™)

Subject Coverage

- All nucleotide sequences of 10 or more bases, all amino acid sequences of 4 or more residues, and probes and primers of any length.

Bibliographic, sequence

For direct code match or similarity (homology) sequence searching, FIZ Karlsruhe provides three specialized RUN package options, GETSEQ, GETSIM and BLAST®

Alerts (SDIs)	Every two weeks		
CAS Registry Number® Identifiers	☐	Page Images	☐
Keep & Share	☑	SLART	☑
Learning Database	☐	Structures	☐

Record Content

- Information on nucleic acid and protein sequences extracted from the original (basic) patent documents published by 52 patent offices worldwide.
- About 49 million sequence records within the database stem from over 340,000 patents from around the globe.
- Records contain a Clarivate Analytics (UK) Limited enhanced title from WPINDEX, a concise sequence description, an English abstract written especially for DGENE by one of Clarivate Analytics (UK) Limited experts, patent information, detailed indexing, a feature table, and sequence data.
- For file crossover to WPINDEX, the DWPI accession number is available in all records.
- DWPI family information can be directly displayed within DGENE using the FAM format.
- Legal status data from INPADOCDB for corresponding DGENE patent sequence documents are available.

File Size

- More than 53.4 million records (07/2020)
- More than 39 million nucleic acid sequences (07/2020)
- More than 14.4 million protein sequences (07/2020)

Coverage

1980-present

Updates

Every two weeks

Language

English

Database Producer

Clarivate
Friars House, 160 Blackfriars Rd.
London SE1 8EZ
United Kingdom

Sources

- Patents from the 41 patent issuing authorities covered by the Derwent World Patents Index® (file WPIDS/WPIX/WPINDEX)

User Aids

- Online Helps (HELP DIRECTORY lists all help messages available)
- STNGUIDE
- DGENE Workshop Manual:
https://stn.products.fiz-karlsruhe.de/en/training-center?query=Sequences

Cluster

- ALLBIB
- AUTHORS
- BIOSCIENCE
- CORPSOURCE
- HPATENTS
- MEDICINE
- PATENTS
- PHARMACOLOGY

STN Database Cluster information:
http://www.stn-international.de/en/customersupport/customer-support#cluster+%7C+subjects+%7C+features

Search and Display Field Codes

Fields that allow left truncation are indicated by an asterisk (*).

General Search Fields

Search Field Name	Search Code	Search Examples	Display Codes
Basic Index* (contains single words from the title (TI), keyword (KW), abstract (AB), description (DESC), and organism name (ORGN) fields)	None or /BI	S F PROMOTER S HUMAN INSULIN (L) PREPARATION S JUNIPERUS VIRGINIANA	TI, KW, AB,DESC, ORGN
Accession Number	/AN	S 95P-R67826/AN	AN
Amino Acid **(3)**	/AA	S 5 S/AA AND 2 T/AA	AA
Amino Acid Count **(1)**	/AA.CNT	S L2 AND AA.CNT>9	AA
Application Country (WIPO code and text)	/AC	S AU/AC	AI
Application Date **(1)**	/AD	S 6 JUN 1994/AD	AI
Application Number **(2)**	/AP	S AU1994-64520/AP S 1994AU-0064520/AP	AI
Application Number Group **(2)**	/APPS	S 1994AU-0064520/APPS	AI, PRAI
Application Year **(1)**	/AY	S 1990/AY	AI
Cross Reference (to related DGENE sequence records)	/CR (or /XR)	S 90N-Q01810/CR	CR
Data Entry Date **(1)**	/DED	S 19901220/DED	DED
Description	/DESC	S DYNORPHIN?/DESC	DESC
Document Type (code and text)	/DT (or /TC)	S PATENT/DT	DT
Entry Date **(1)**	/ED (or /UP)	S ED<JULY 2001	ED
Feature Table * **(3)**	/FEAT	S DISUL?/FEAT	FEAT
File Segment	/FS	S L10 AND PS/FS	FS
Inventor	/IN (or /AU)	S GOSSEN M/IN	IN
Keyword	/KW	S F PROMOTER/KW	KW
Language (ISO code and text)	/LA	S DE/LA S FRENCH/LA	LA
Molecule Type	/MTY	S RNA/MTY	MTY
Nucleic Acid **(3)**	/NA	S 7 C/NA AND 25 A/NA	NA
Nucleic Acid Count **(1)**	/NA.CNT	S L34 AND NA.CNT>10	NA
Organism Name	/ORGN	S ADENOVIRUS/ORGN	ORGN
Other Source (DWPI accession number)	/OS	S 94-151326/OS S 2000-012128/OS	OS
Patent Assignee **(3)**	/PA (or /CS)	S ZAMBON SPA/PA	PA
Patent Assignee Code **(4)**	/PACO	S BADI/PACO	PA
Patent Countries (WIPO code and text)	/PCS	S UNITED KINGDOM/PCS	PI
Patent Country (WIPO code and text)	/PC	S DE/PC S UNITED KINGDOM/PC	PI
Patent Kind Code	/PK	S FRA/PK	PI
Patent Number **(2)**	/PN	S DE4244565/PN S EP-348819/PN	PI
Patent Number Group **(2)**	/PATS	S EP-348819/PATS	PI
Patent Sequence Location	/PSL	S DISCLOSURE/PSL S COL/PSL S FIG/PSL	PSL
Priority Country (WIPO code and text)	/PRC	S US/PRC AND L3 S UNITED STATES/PRC AND L3	PRAI
Priority Date **(1)**	/PRD	S 19930907/PRD	PRAI

General Search Fields (cont'd)

Search Field Name	Search Code	Search Examples	Display Codes
Priority Date First **(1)**	/PRDF	S 17 JUNE 1993/PRDF	PRAI
Priority Number **(2)**	/PRN	S 1993US-0078471/PRN S US1993-78471/PRN	PRAI
Priority Year **(1)**	/PRY	S 1993/PRY	PRAI
Priority Year, First **(1)**	/PRYF	S 1993/PRYF	PRAI
Publication Date **(1)**	/PD	S 18 JAN 1995/PD	PI
Publication Year **(1)**	/PY	S 1999/PY	PI
Sequence Length **(1)**	/SQL	S 70-90/SQL	SQL
Title*	/TI	S HUMAN INSULIN/TI	TI

(1) Numeric search field that may be searched using numeric operators or ranges.
(2) Either STN format or Derwent format may be used.
(3) Search with implied (S) proximity is available in this field.
(4) The list of Clarivate Analytics (UK) Limited-assigned company codes for patent assignees matched with company names is available in this field. See page 14.

SEQUENCE SIMILARITY SEARCHING (BLAST/GETSIM)

The GETSIM and BLAST® run packages are available to search the DGENE database for protein and nucleotide sequence data by similarity (homology). BLAST is provided in DGENE with the permission of the National Center for Biotechnology Information (NCBI) of the National Library of Medicine (NLM). GETSIM is provided in DGENE by FIZ Karlsruhe GmbH, and is based upon the FASTA algorithm.

To initiate a BLAST or GETSIM search the following search codes have to be specified: SQP for searching peptide sequences (default), SQN for nucleotide sequences, or TSQN for searching peptide sequences translated from DGENE nucleotide sequences. The GETSIM or BLAST search can be run in offline BATCH mode or used as the basis of a current-awareness ALERT. The offline search mode offers an email notification option which allows users to see when batch search results are available for download. When using the SQN option 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 code, e.g., /SQN COM. If no search option is given, SIN (single) will be used by default for GETSIM, and BOTH (both) will be used by BLAST. Note that for the TSQN option generally both strands will be searched, i.e., for a single polypeptide query, the TSQN option will cover all six possible translations (three reading frames of both the single and the complementary nucleotide sequences). Nucleotide and protein sequences can be subjected to a similarity search in various ways. A query can be prepared with the query command and saved beforehand, it can be entered directly on the command line using RUN GETSIM/BLAST, or it may be uploaded from an ASCII file using the UPLOAD command. You may also use the Sequence Query Upload Wizard from STN Express version 8.3+. 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 either:
1) Keep the complete answer set (ALL)
2) Keep a subset of the complete answer set by specifying a smaller number of just the top scoring answers.
3) 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 or descending percent identity. Just type "SOR SCORE D" to sort by descending similarity score or SOR IDENT D" to sort by descending percent identity 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 (for GETSIM or for BLAST). The top line is the query sequence and the bottom line the hit sequence. The BLAST ALIGN format follows the standard convention for NCBI alignment displays. The GETSIM ALIGN format uses two dots to represent identical nucleotides/peptides, a blank if there is no match, and one dot to indicate a chemical "family" match. Gaps inserted in the query or answer sequence for alignment purposes are shown with an underscore.

GETSIM / BLAST: TYPES OF SEARCHES

Description	Search Code	Search Example (4)
Peptide Homology	/SQP	RUN BLAST L1 /SQP
Nucleotide Homology	/SQN	RUN BLAST L1 /SQN
Single Strand		RUN GETSIM L1 /SQN SIN **(1)**
Complementary Strand		RUN GETSIM L1 /SQN COM
Both Strands		RUN BLAST L1 /SQN BOTH **(2)**
Translated Peptide Homology	/TSQN	RUN BLAST L1 /TSQN
		RUN GETSIM L1 /TSQN
Offline BATCH search	/SQP BATCH	RUN BLAST L1 /SQP BATCH
	/SQN BATCH	RUN GETSIM L1 /SQN BOTH BATCH
	/TSQN BATCH	RUN BLAST L1 /TSQN BATCH
Current-awareness ALERT **(3)**	/SQP ALERT	RUN BLAST L1 /SQP ALERT
	/SQN ALERT	RUN GETSIM L1 /SQN BOTH ALERT
	/TSQN ALERT	RUN BLAST L1/TSQN ALERT

(1) GETSIM default setting
(2) BLAST default setting
(3) Homology ALERT search, which runs every update of the database (once every two weeks).
(4) Where L1 is a sequence query generated using the UPLOAD or QUERY command

ADVANCED USER OPTIONS FOR BLAST

For the experienced user of BLAST®, 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 customizing any of these settings. For further information:
http://blast.ncbi.nlm.nih.gov/Blast.cgi?CMD=Web&PAGE_TYPE=BlastDocs

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

BLAST 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

BLAST Matrix settings (for option 5.) (cont'd)

Matrix	Gap	Gap Extension
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: Online GETSIM homology search for a protein

```
=> run getsim
LDHILQKTERGVRLHPLARTAKVKNEVNSFKAALSSLAKHGEYAPFARLLNLSGVNNLEHGLFPQLSAIA/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
      443 |
          |
          |
          |
          |
          |
          |
          |
          |
          |
      222 |
          ||
          ||
          ||
          ||
          ||
          ||
          ||||||||||||||||||||||||||||||||||||||||||||||||||
          ||||||||||||||||||||||||||||||||||||||||||||||||||
          ||||||||||||||||||||||||||||||||||||||||||||||||||
Answer Count     290       580       870      1160      1450
```

DGENE

```
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 Clarivate Analytics 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 wmapsltesd mdyhkiltag lsvqqgivrq rvipvyqvnn
       51 leeicqliiq afeagvdfqe sadsfllmlc lhhayqgdyk lflesgavky
      101 leghgfrfev kkrdgvkrle ellpavssgk nikrtlaamp eeetteanag
      151 qflsfaslfl pklvvgekac lekvqrqiqv haeqgliqyp tawqsvghmm
      201 vifrlmrtnf likfllihqg mhmvaghdan davisnsvaq arfsgllivk
      251 tvldhilqkt ergvrlhpla rtakvknevn sfkaalssla khgeyapfar
      301 llnlsgvnnl ehglfpqlsa ialgvatahg stlagvnvge qyqqlreaat
      351 eaekqlqqya esreldhlgl ddqekkilmn fhqkkneisf qqtnamvtlr
      401 kerlakltea itaaslpkts ghydddddip fpgpindddn pghqdddptd
      451 sqdttipdvv vdpddgsyge yqsysengmn apddlvlfdl deddedtkpv
      .......

L2    ANSWER 10 OF 23  DGENE  COPYRIGHT 2008 Clarivate Analytics on STN
SCORE 439      99% of query self score 443
ALIGN Smith-Waterman score:  439
      70 aa overlap starting at 251
      ldhilqktergvrlhplartakvknevnsfkaalsslakhgeyapfarllnlsgvnnleh
      ::::::::.:::::::::::::::::::::::::::::::::::::::::::::::::::
      ldhilqktdrgvrlhplartakvknevnsfkaalsslakhgeyapfarllnlsgvnnleh
      glfpqlsaia
      ::::::::::
      glfpqlsaia
SEQ
        1 mdsrpqkiwm xpsltesdmd yhkiltagls vqqgivrqkv ipvyqvnnle
       51 eicqliiqaf eagidfqesa dsfllmlclh hayqgdyklf lesgavkyle
      101 ghgfrfevkk rdgvkrleel lpavssgkni krtlaalpee etteanagqf
      151 lsfaslflpk lvvgekacle kvqrqiqvha eqgliqypta wqsvghmmvi
      201 frlmrtnfli kfllihqgmh mvaghdanda visnsvaqar fsgllivktv
      251 ldhilqktdr gvrlhplart akvknevnsf kaalsslakh geyapfarll
      301 nlsgvnnleh glfpqlsaia lgvatahgst lagvnvgeqy qqlreaatea
      351 ekqlqqyaes reldhlgldd qekkilmnfh qkkn
```

```
L2    ANSWER 23 OF 23  DGENE  COPYRIGHT 2008 Clarivate Analytics on STN
SCORE 367      82% of query self score 443
ALIGN Smith-Waterman score:  367
      70 aa overlap starting at 233
      ldhilqktergvrlhplartakvknevnsfkaalsslakhgeyapfarllnlsgvnnleh
      :. ::::::. :: ::::.::.::::::: ::: :::.::.::::::::::.::::::::
      lefilqktdsgvtlhplvrtskvknevasfkqalsnlarhgeyapfarvlnlsginnleh
      glfpqlsaia
      ::.::::::::
      glypqlsaia
SEQ
        1 mdlhsllelg tkptaphvrn kkvilfdtnh qvsicnqiid ainsgidlgd
       51 lleggllltlc vehyynsdkd kfntspvaky lrdagyefdv iknadatrfl
      101 dvspnephys plilalktle stesqrgrig lflsfcslfl pklvvgdras
      151 iekalrqvtv hqeqgivtyp nhwlttghmk vifgilrssf ilkfvlihqg
      201 vnlvtghday dsiisnsvgq trfsgllivk tvlefilqkt dsgvtlhplv
      251 rtskvkneva sfkqalsnla rhgeyapfar vlnlsginnl ehglypqlsa
      301 ialgvatahg stlagvnvge qyqqlreaah daevklqrrh ehqeiqaiae
      ......
```

Example: Online BLAST homology search for a nucleotide with altered parameter

```
=> 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
     3503 |
          |
          ||
          ||
          ||
          ||||||
          |||||||
          |||||||
          ||||||||
          ||||||||
     1752 ||||||||
          ||||||||||
          ||||||||||
          ||||||||||
          ||||||||||||
          ||||||||||||
          ||||||||||||||
          ||||||||||||||||||
          ||||||||||||||||||||||
          ||||||||||||||||||||||||||||||
Answer Count      20        40        60        80       100

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
               GGACTGTGAACGGACAACTGAGGGTTGGGTTTTTTACCACCAAACTGGTT
               .......
               GCAAGCTGACTCACGGTGTTATGAATAAGGAGCTGAAGTACTGTAAGAAT
               CCTGAGGACCTGGAGTGCAATGAGAATGTGAAACACAAAACCAAGGAGTA
               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 Clarivate Analytics 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    atggccctgaagaatgatgagataatagatgccactcaaaaaggaaattgctctcgtttc
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 85   atggccctgaagaatgatgagataatagatgccactcaaaaaggaaattgctctcgtttc

Query: 61   atgaatcacagctgtgaaccaaattgtgaaacccaaaaatggactgtgaacggacaactg
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 145  atgaatcacagctgtgaaccaaattgtgaaacccaaaaatggactgtgaacggacaactg
.
.
.
Query: 1681 aaacacaaaaccaaggagtacattaagaagtacatgcagaagtttggggctgtttacaaa
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 1765 aaacacaaaaccaaggagtacattaagaagtacatgcagaagtttggggctgtttacaaa

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

L3 ANSWER 4 OF 4 DGENE COPYRIGHT 2008 Clarivate Analytics 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    atggccctgaagaatgatgagataatagatgccactcaaaaaggaaattgctctcgtttc
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 3381 atggccctgaagaatgatgagataatagatgccactcaaaaaggaaattgctctcgtttc

Query: 61   atgaatcacagctgtgaaccaaattgtgaaacccaaaaatggactgtgaacggacaactg
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 3441 atgaatcacagctgtgaaccaaattgtgaaacccaaaaatggactgtgaacggacaactg
.
.
.
Query: 1501 aaagtacgaattaaagaccgcaataaactttctacagaggaacgccggaagttgtttgag
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 4881 aaagtacgaattaaagaccgcaataaactttctacagaggaacgccggaagttgtttgag

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

SEARCHING SEQUENCE DATA WITH THE GETSEQ RUN PACKAGE

Sequence information (amino acid and nucleic acid sequences) may be retrieved by using a variety of search fields available with the GETSEQ run package. The query may be first created with the QUERY command, and subsequently searched through the GETSEQ run package specifying the query L-number (e.g., RUN GETSEQ L9, if L9 represents the sequence query). The L-number may also derive from a previous sequence search in another STN database with biosequence search capabilities, e.g., the CAS REGISTRY[SM] file. The query may also be directly entered within the GETSEQ package at a colon prompt after GETSEQ has been initialized with the RUN command (i.e., RUN GETSEQ). Offline sequence searching is also available for GETSEQ searches.

SEQUENCE SEARCH TERMS

Terms	Query Examples
One-letter codes for common amino acids **(1,2)** Three-letter codes for common amino acids **(1,2**) Enclose codes or strings of codes in single quotes. Use dashes to separate codes in strings. One-letter codes for nucleic acids **(3)**	QUE LAGLL/SQSP QUE 'THR-SER-GLY-MET-THR'/SQSFP QUE GLPGY/SQS QUE 'LEU-ARG-ASP-THR'/SQEP QUE ATGAAN/SQEN QUE ATGAAN/SQSN

(1) Enter 'HELP AAC' at an arrow prompt to display a table of the one- and three-letter codes for common amino acids.
(2) 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).
(3) Enter 'HELP NUC' at an arrow prompt to display a table of the codes for nucleic acids.

TYPES OF SEQUENCE SEARCHES

Sequence data for nucleic acid and protein sequences are displayed in the SEQ field with one-letter codes and the SEQ3 field with three-letter codes for proteins only.

Type	Definition	Search Code	Query Examples
Sequence Exact Protein	Search for sequences that match the query**.(2)**	/SQEP	QUE GAPGEK/SQEP QUE 'ALA-PHE-PHE-PHE-PHE'/SQEP
Sequence Exact Family, Protein	Search for sequences that match the query and those in which family-equivalent substitution of the query amino acids occur. **(1,2)**	/SQEFP	QUE YGGFL/SQEFP QUE 'TYR-GLY-GLY-PHE-LEU'/SQEFP
Subsequence, Protein	Search for exact answers plus sequences in which the query sequence is embedded. **(2)**	/SQSP	QUE LAGLL/SQSP QUE 'GLP'GY/SQSP
Subsequence Family, Protein	Search for exact sequences, subsequences, and answers in which family-equivalent substitution of the query amino acids occurs. **(1,2)**	/SQSFP	QUE ATCXAWV/SQSFP QUE 'THR-SER-GLY-MET-THR'/SQSFP
Sequence Exact, Nucleic Acid	Search for sequences that match the query. Ambiguity codes for nucleic acids are allowed. **(2)**	/SQEN	QUE ATGAAN/SQEN

TYPES OF SEQUENCE SEARCHES (cont'd)

Type	Definition	Search Code	Query Examples
Subsequence, Nucleic Acid	Search for exact answers, plus sequences in which the query sequence is embedded. Ambiguity codes for nucleic acids are allowed. **(2)**	/SQSN	QUE ATGAAN/SQSN

(1) The families of amino acid equivalents retrieved in protein 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)
F, Y, W (hydrophobic, aromatic)
L, I, V, M (hydrophobic)
C (cross-link forming)

(2) Variability symbols are allowed.

VARIABILITY SYMBOLS FOR SEQUENCE CODE MATCH SEARCHES (1,2)

Symbol(s)	Function	Query Examples
[ ]	to specify alternate residues	QUE LGP[VL]/SQSP QUE LGP['VAL''LEU''LYS']/SQSP
[-]	to exclude a specific residue or alternate residues	QUE LGP[-H]/SQSP QUE LGP[-'HIS']/SQSPSP QUE LGP[-HL]/SQSP
{m}	to repeat the preceding sequence or sequence query (L#) m times	QUE (FL){2}/SQSP QUE L4{2}/SQSP QUE (CTG){2}/SQSN QUE TAA(TAAA){2}/SQSN
{m,u} or {m-u}	to repeat the preceding sequence or sequence query (L#) m to u times	QUE GG(FL){1,2}/SQSP QUE L3{1,3}/SQSP QUE (CTG){1,3}/SQSN
? or {0,1} or {0-1}	to repeat the preceding sequence or sequence query (L#) zero or	QUE FLRRI(RP)?K/SQSP QUE FLRRI(RP){0,1}K/SQSP QUE L1{-1}NN/SQSP QUE L1{0,1}NN/SQSP QUE CAT(CGA){0,1}GGAC/SQSN
* or {0,} or {0-}	to repeat the preceding sequence or sequence query (L#) zero or more times	QUE KLK(WD){0,}N/SQSP QUE KLK(WD)*N/SQSP QUE L1{0-}NN/SQSP QUE L1{0,}NN/SQSP QUE CAT(CTG){0,}TATT/SQSN
+ or {1,} or {1-}	to repeat the preceding sequence or sequence query (L##) one or more times	QUE KLK(DLE){1,}/SQSP QUE KLK(DLE)+/SQSP QUE L2{1-}/SQSP QUE L2{1,}/SQSP QUE CAT(CTG){1,}TATT/SQSN
&	to join together sequence expressions or queries (L#s)	QUE L1&L3/SQSFP QUE L2&L5{1,3}/SQSP

(1) In addition, the caret (Λ) and the vertical bar (|) may be used. The caret is used at the beginning or at the end of a sequence to search for that sequence at the beginning or end of sequence field. The vertical bar is the symbol for alternation, i.e., it is used to separate alternate sequence queries.

(2) For more information on specifying variability in sequence code match queries, enter 'HELP SQQ' at an arrow prompt (=>).

SPECIFYING GAPS IN GETSEQ SEQUENCE QUERIES

Symbol(s)	Function	Query Examples
.	a gap of one residue	QUE SY.RPG/SQSP QUE SY..RPG/SQSP QUE AAG...TGC/SQSN
.{m} or [m.]	a gap of m residues	QUE SY.{2}RPG/SQSP QUE SY[2.]RPG/SQSP
.{m,u} or .{m-u}	a gap of m to u residues	QUE GFF.{2,10}LSS/SQSP QUE GFF.{2-10}LSS/SQSP QUE AAG.{2,5}TGC/SQSN
: or .? or .{0,1} or .{0-1}	a gap of zero or one residues	QUE AGA:SRI/SQSFP QUE AGA.?SRI/SQSFP QUE AGA.{0,1}SRI/SQSFP QUE AGA.{0-1}SRI/SQSFP
.* or .{0,} or .{0-}	a gap of zero or more residue	QUE HLC.*TYG/SQSP QUE HLC.{0,}TYG/SQSP QUE HLC.{0-}TYG/SQSP QUE AAGGCAGATG.*GCAA/SQSN
.+ or .{1,} or .{1-}	a gap of one or more residues	QUE SY.+TH/SQSP QUE SY.{1,}TH/SQSP QUE SY.{1-}TH/SQSP QUE TCCTG.+GTGG/SQSN

PATENT ASSIGNEE CODE DICTIONARY

The list of Clarivate Analytics (UK) Limited-assigned company codes for patent assignees matched with company names is available in field /PACO. This feature allows you to easily and comprehensively identify the company names associated with a code, or to identify the code(s) used for a company name. Expanding in field /PACO (Patent Assignee Code) provides the alphabetical list of codes, single words and the full name from the company field (/PA). Each code is listed with its frequency in field /PACO and with the number of associated terms (AT) in the dictionary.

Relationship Code	Content	Example
ALL	All patent assignee code(s) defined for the company name (SELF, CODE)	E GENZYME+ALL/PACO
DEF	All name definitions for the given code (SELF, DEF)	E MYRI-N+DEF/PACO

DISPLAY and PRINT Formats

Any combination of formats may be used to display or print answers. Multiple codes must be separated by spaces or commas, e.g., D L1 1-5 TI AU. The fields are displayed or printed in the order requested.

Hit-term highlighting is available for all fields. Highlighting must be ON during SEARCH to use the HIT, KWIC, and OCC formats.

DISPLAY and PRINT Formats (cont'd)

Format	Content	Examples
AA AB AI (AP) **(1)** AN APPS **(1)** CR (XR) DED DESC DT (TC) FEAT FS IDENT **(2,3)** IN (AU) KW LA MTY NA ORGN OS PA (CS) PATS **(1)** PI (PN) **(1)** PRAI (PRN) **(1)** PSL SCORE **(3,4)** SEQ SEQ3 SQL TI	Amino Acid Abstract Application Information Accession Number Application Number Group Cross Reference Data Entry Date Description Document Type Feature Table File Segment Percent Identity Inventor Keyword Language Molecule Type Nucleic Acid Organism Name Other Source Patent Assignee Patent Number Group Patent Information Priority Information Patent Sequence Location Similarity Score Sequence (one-letter codes) Sequence (three-letter codes) Sequence Length Title	D AA TI KW D AB D AI D AN D APPS D CR D DED D DESC D TC D FEAT D FS D IDENT D IN D KW D LA D MTY D NA D ORGN D OS D PA D PATS D PI D PRAI D PSL D SCORE D SEQ D SEQ3 D SQL D TI
ABS ALIGN **(4)** ALL **(1)** BIB **(1)** FAM **(1)** IALL **(1)** IBIB **(1)** LS **(1)** LS2 **(1)** SCAN **(5)** SQIDE SQ3IDE TRIAL (TRI,SAM)	AN, MTY, AB Alignment between query and retrieved sequence in a similarity search (RUN GETSIM or RUN BLAST) AN, MTY, TI, IN, PA, PI, AI, PRAI, PSL, DED, DT, LA, OS, CR, DESC, KW, ORGN, AB, AA, NA, SQL, SEQ, FEAT AN, MTY, TI, IN, PA, PI, AI, PRAI, PSL, DT, LA, OS, CR, DESC Patent family information from the Derwent World Patents Index (PI, ADT, FDT, PRAI) ALL, indented with text labels BIB, but indented with text labels Legal Status (from INPADOCDB) Legal Status (from INPADOCDB), detailed version with display headers AN, MTY, TI, DESC AN, MTY, AA, NA, SQL, SEQ, FEAT AN, MTY, AA, NA, SQL, SEQ3, FEAT AN, MTY, TI, DESC, KW, SQL	D ABS D ALIGN D ALL D BIB D FAM D IALL D IBIB D LS D LS2 D SCAN D SQIDE ABS D SQ3IDE D TRI 1-10
HIT KWIC OCC	Hit term(s) and field(s) Up to 50 words before and after hit term(s) (KeyWord-In-Context) Number of occurrences of hit term(s) and field(s) in which they occur	D HIT D KWIC D OCC

(1) By default, patent numbers, application and priority numbers are displayed in STN Format. To display them in Derwent format, enter SET PATENT DERWENT at an arrow prompt. To reset display to STN Format, enter SET PATENT STN.
(2) Use RUN BLAST first. See page 4, Similarity Search.
(3) Custom display only.
(4) Use RUN GETSIM or RUN BLAST first. See page 4, Similarity Search.
(5) SCAN must be specified on the command line, i.e., D SCAN or DISPLAY SCAN.

SELECT, ANALYZE, and SORT Fields

The SELECT command is used to create E-numbers containing terms taken from the specified field in an answer set.

The ANALYZE command is used to create an L-number containing terms taken from the specified field in an answer set.

The SORT command is used to rearrange the search results in either alphabetic or numeric order of the specified field(s).

Field Name	Field Code	ANALYZE/ SELECT (1)	SORT
Abstract	AB	Y **(2)**	N
Accession Number	AN	Y	Y
Amino Acid	AA	Y	N
Amino Acid Count	AA.CNT	Y	N
Application Country	AC	Y	Y
Application Date	AD	Y	N
Application Number	AP (AI)	Y	Y
Application Number Group	APPS	Y **(3)**	N
Application Year	AY	Y	N
Cross References (to related DGENE records)	CR (XR)	Y	N
Data Entry Date	DED	Y **(4)**	Y
Description	DESC	Y	Y
Document Type	DT	Y	Y
Entry Date	ED (UP)	Y **(4)**	Y
Feature Table	FEAT	Y **(4)**	N
File Segment	FS	Y	Y
Inventor	IN (AU)	Y	Y
Keyword	KW	Y	N
Language	LA	Y	Y
Molecule Type	MTY	Y	Y
Nucleic Acid	NA	Y	N
Nucleic Acid Count	NA.CNT	Y	N
Organism Name	ORGN	Y	Y
Other Source (DWPI accession number)	OS	Y	Y
Patent Assignee	PA (CS)	Y	Y
Patent Assignee Code	PACO	Y	Y
Patent Country	PC	Y	Y
Patent Countries	PCS	Y	N
Patent Kind Code	PK	Y	Y
Patent Number	PN (PI)	Y	Y
Patent Number Group	PATS	Y	N
Patent Sequence Location	PSL	Y	Y
Percent Identity	IDENT **(5)**	N	Y
Priority Country	PRC	Y	Y
Priority Date	PRD	Y	Y
Priority Date First	PRDF	Y **(4)**	Y
Priority Number	PRN (PRAI)	Y	Y
Priority Year	PRY	Y	Y
Priority Year First	PRYF	Y **(4)**	Y
Publication Date	PD	Y	Y
Publication Year	PY	N	Y

SELECT, ANALYZE, and SORT Fields (cont'd)

Field Name	Field Code	ANALYZE/ SELECT (1)	SORT
Sequence (1-letter codes)	SEQ	Y **(6)**	N
Sequence (3-letter codes)	SEQ3	Y **(6)**	N
Sequence Length	SQL	Y	Y
Similarity Score	SCORE **(7)**	N	Y
Title	TI	Y (default)	Y

(1) HIT may be used to restrict terms extracted to terms that match the search expression used to create the answer set, e.g., SEL HIT TI.
(2) Appends /BI to the terms created by SELECT.
(3) Selects or analyzes application and priority numbers and appends /APPS to the terms created by SELECT.
(4) SELECT HIT and ANALYZE HIT are not valid with this field.
(5) Used with a L-number created by BLAST.
(6) Appends /SQSN or /SQSP to the terms created by SELECT.
(7) Used with a L-number created by BLAST or GETSIM.

Sample Records

DISPLAY TRIAL

```
AN    AQC88542  protein       DGENE
TI    Treating cancer e.g. breast cancer, brain cancer and leukemia, comprises
      administering a vector comprising a nucleic acid which operably encodes
      an interferon protein, and an inhibitor of transforming growth factor-
      beta activity.
DESC  Rat TGFRII protein SEQ ID:43.
KW    therapeutic; gene therapy; vector; breast tumor; brain tumor; uterine
      cervix tumor; leukemia; lymphoma; prostate tumor; skin tumor; colon
      tumor; lung tumor; mesothelioma; ovary tumor; pancreas tumor; liver
      tumor; bladder tumor; renal tumor; myeloma; colorectal tumor;
      nasopharyngeal carcinoma; endometroid carcinoma; cytostatic; Transforming
      growth factor-beta receptor II; TGFRII; BOND_PC; transforming growth
      factor-b type II receptor; transforming growth factor, beta receptor 2;
      transforming growth factor beta receptor 2; transforming growth factor
      beta, receptor 2; transforming growth factor beta receptor II; Tgfbr2;
      transforming growth factor-beta type II receptor; TGF-beta 2;
      transforming growth factor, beta receptor II; TGF-beta RII; T beta RII;
      GO166; GO287; GO324; GO1570; GO1666; GO3674; GO4702; GO4713; GO4872;
      GO5026; GO5515; GO5524; GO5737; GO5829; GO5887; GO5901; GO6468; GO6470;
      GO6898; GO7179; GO7182; GO7507; GO7566; GO7584; GO8150; GO8284; GO8285;
      GO9612; GO9749; GO9887; GO9986; GO10033; GO14070; GO16020; GO16021;
      GO16740; GO30145; GO30324; GO31435; GO42060; GO42127; GO42803; GO43415;
      GO43627; GO46982; GO48545; GO48661; GO60044; GO4682; GO4691; GO5024;
      GO7178; GO7180.
SQL   567
```

DISPLAY SQIDE

```
AN    AAW79650  peptide       DGENE
AA    0  A; 2  R; 0  N; 2  D; 0  B; 2  C; 0  Q; 0  E; 0  Z; 2  G; 0  H; 0  I; 0
      L; 0  K; 0  M; 0  F; 0  P; 2  S; 0  T; 0  W; 0  Y; 0  V; 0 Others
SQL   10
SEQ
          1 crgdsrgdsc

FEATURE TABLE:
Key            |Location|Qualifier|
===============+========+=========+============================
Modified-site |1       |note     |"N-alpha-acetyl-Cys"
Modified-site |2..6    |note     |"optionally one of Arg(2) and
               |        |         |Arg(6) is in the form of
               |        |         |MeArg"
Modified-site |10      |note     |"Cys-NH2"
Disulfide-bond|1..10   |         |
```

DISPLAY SQ3IDE

```
AN    AAW79650  peptide       DGENE
AA    0  A; 2  R; 0  N; 2  D; 0  B; 2  C; 0  Q; 0  E; 0  Z; 2  G; 0  H; 0  I; 0
      L; 0  K; 0  M; 0  F; 0  P; 2  S; 0  T; 0  W; 0  Y; 0  V; 0 Others
SQL   10
SEQ3
        1 Cys-Arg-Gly-Asp-Ser-Arg-Gly-Asp-Ser-Cys


FEATURE TABLE:
Key            |Location|Qualifier|
===============+========+=========+============================
Modified-site |1       |note     |"N-alpha-acetyl-Cys"
Modified-site |2..6    |note     |"optionally one of Arg(2) and
               |        |         |Arg(6) is in the form of
               |        |         |MeArg"
Modified-site |10      |note     |"Cys-NH2"
Disulfide-bond|1..10   |         |
```

DISPLAY IALL

```
ACCESSION NUMBER: AAD05559  cDNA       DGENE
TITLE:            Isolated nucleic acid molecule encoding a human secreted
                  protein is used in preventing, treating or ameliorating a
                  medical condition  -
INVENTOR:         Soppet D R; Komatsoulis G; Shi Y; Olsen H S; Ruben S M
PATENT ASSIGNEE:  (HUMA-N)HUMAN GENOME SCI INC.
PATENT INFO:      WO-200134767 A2 20010517              540p
APPLICATION INFO: 2000WO-US30036  20001101
PRIORITY INFO:    1999US-0163576  19991105
                  2000US-0221366  20000727
PAT. SEQ. LOC:    Claim 1; Page 459
DATA ENTRY DATE:  18 JUL 2001  (first entry)
DOCUMENT TYPE:    Patent
LANGUAGE:         English
OTHER SOURCE:     2001-316492 ã33î
CROSS REFERENCES: P-PSDB: AAE01738
```

DESCRIPTION: Human secreted protein-encoding gene 19 cDNA clone HYASC80, SEQ ID NO:78.
KEYWORD: Human; secreted protein; proliferative disorder; cancer; tumour; foetal abnormality; developmental abnormality; haematopoietic disorder; immune system disorder; AIDS; autoimmune disease; rheumatoid arthritis; inflammation; allergy; neurological disorder; Alzheimer's disease; Parkinson's disease; cognitive disorder; schizophrenia; asthma; skin disorder; psoriasis; sepsis; diabetes; atherosclerosis; cardiovascular disorder; angiogenic disorder; kidney disorder; gastrointestinal disorder; pregnancy-related disorder; gene therapy; endocrine disorder; infection; wound healing; vulnerary; cell culture; chemotaxis; food additive; binding partner identification; ss.
ORGANISM: Homo sapiens.
ABSTRACT:
AAD05492-AAD05564 represent cDNAs corresponding to 22 human secreted protein genes, and AAE01672-AAE01743 represent the proteins they encode. AAE01744-AAE01763 represent human secreted protein fragments or variants. The secreted proteins and their genes are useful for preventing, treating or ameliorating medical conditions, e.g., by protein or gene therapy. Pathological conditions can be diagnosed by determining the amount of the new protein in a sample or by determining the presence of mutations in the new genes. Specific uses are described for each of the 22 genes, based on the tissues in which they are most highly expressed, and include developing products for the diagnosis or treatment of proliferative disorders, cancer, tumours, foetal and developmental abnormalities, haematopoietic disorders, diseases of the immune system, AIDS, autoimmune diseases (e.g., rheumatoid arthritis), inflammation, allergies, neurological disorders (e.g., Alzheimer's disease, Parkinson's disease), cognitive disorders, schizophrenia, asthma, skin disorders (e.g., psoriasis), sepsis, diabetes, atherosclerosis, cardiovascular disorders, angiogenic disorders, kidney disorders, gastrointestinal disorders, pregnancy-related disorders, endocrine disorders, and infections. The proteins can also be used to aid wound healing and epithelial cell proliferation, to prevent skin aging due to sunburn, to maintain organs before transplantation, for supporting cell culture of primary tissues, to regenerate tissues, to identify their cognate ligands or binding partners, and in chemotaxis, and can be used as a food additive or preservative to modify storage properties. Antibodies specific for a protein of the invention can be used in alleviating symptoms associated with the disorders mentioned above, and in diagnostic immunoassays e.g., radioimmunoassay or enzyme linked immunosorbent assay (ELISA). The present sequence represents a human secreted protein-encoding cDNA of the invention.
NUCLEIC ACID COUNTS: 8 A; 16 C; 20 G; 14 T; 2 other
SEQUENCE LENGTH: 60
SEQUENCE
1 ggggatggga acccaaggct gtccacatcc cagctggctg mtmcttctgg
51 ggctgtcttg

```
FEATURE TABLE:
Key       |Location|Qualifier    |
==========+========+=============+============================
CDS       |5..58   |*tag=  a     |
          |        |product      |"Human secreted protein
          |        |             |precursor"
          |        |transl_except|(pos:41..43, aa:Xaa)
          |        |note         |"Xaa = any of the naturally
          |        |             |occurring L-amino acids. CDS
          |        |             |does not include stop codon"
          |        |             |/partial
```

DISPLAY FAM

```
PI    WO 2001034767   A2 20010517 (200133)* EN  540[0]
      AU 2001017537   A  20010606 (200152)  EN
      JP 2003520033   W  20030702 (200352)  JA  639
      EP 1556401      A2 20050727 (200549)  EN
ADT   WO 2001034767 A2 WO 2000-US30036 20001101; EP 1556401 A2 EP 2000-980250
      20001101; JP 2003520033 W WO 2000-US30036 20001101; EP 1556401 A2 WO
      2000-US30036 20001101; AU 2001017537 A AU 2001-17537 20001101; JP
      2003520033 W JP 2001-537464 20001101
FDT   AU 2001017537 A Based on WO 2001034767 A; JP 2003520033 W Based on WO
      2001034767 A; EP 1556401 A2 Based on WO 2001034767 A
PRAI  US 2000-221366P      20000727
      US 1999-163576P      19991105
```

In North America
CAS
STN North America
P.O. Box 3012
Columbus, Ohio 43210-0012 U.S.A.

CAS Customer Center:
Phone: 800-753-4227 (North America)
614-447-3700 (worldwide)
Fax: 614-447-3751
Email: help@cas.org
Internet: www.cas.org

In Europe
FIZ Karlsruhe
STN Europe
P.O. Box 2465
76012 Karlsruhe
Germany
Phone: +49-7247-808-555
Fax: +49-7247-808-259
Email: helpdesk@fiz-karlsruhe.de
Internet: www.stn-international.com

In Japan
JAICI (Japan Association for
International Chemical Information)
STN Japan
Nakai Building
6-25-4 Honkomagome, Bunkyo-ku
Tokyo 113-0021, Japan
Phone: +81-3-5978-3601 (Technical Service)
+81-3-5978-3621 (Customer Service)
Fax: +81-3-5978-3600
Email: support@jaici.or.jp (Technical Service)
customer@jaici.or.jp (Customer Service)
Internet: www.jaici.or.jp