PCTGEN (World Patent Application Biosequences)

Subject Coverage

- Nucleotide and amino acid sequence data as submitted by patent applicants to the World Intellectual Property Organization (WIPO).

File Type

Bibliographic, sequence

Features

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

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

Record Content

- Records contain sequence and patent information as given by the patent applicant.
- Each record includes the actual sequence and additional information on the sequence, e.g. molecule type and organism, and patent information, e.g. application and publication data.

File Size

- More than 20.7 million records (07/2020)
- More than 15.4 million nucleic acid sequences (07/2020)
- More than 5.3 million protein sequences (07/2020)

Coverage

August 2001-present

Updates

Weekly

Language

English

Database Producer

World Intellectual Property Organisation
34, Chemin des Colombettes
1211 Geneva
Switzerland
Phone: +41 22 338 91 11
Fax: +41 22 338 98 20
Copyright Holder

Database Supplier

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

Sources

- Sequence listings submitted by patent applicants as a formal part of WIPO/PCT applications.

User Aids

- Online Helps (HELP DIRECTORY lists all help messages available)
- STNGUIDE

Cluster

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

STN Database Cluster information:
http://www.stn-international.com/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), organism species (ORGN), and molecule type (MTY) fields)	None or /BI	S ANAPHYLATOXIN S PLANT GENE# AND RNA	TI, ORGN, MTY
Accession Number	/AN	S 2002060924.37/AN	AN
Application Country	/AC	S US/AC	AI
Application Date **(1)**	/AD	S 20011129/AD	AI
Application Number **(2)**	/AP	S US2001-809003/AP	AI
Application Year **(1)**	/AY	S 2002/AY	AI
Document Type (code and text)	/DT (or /TC)	S PATENT/DT	DT
Entry Date **(1)**	/ED	S 20021004/ED	ED
Feature Table*	/FEAT	S (RNA AND BINDING)/FEAT S ?COMBINAT?/FEAT	FEAT
File Segment (code and text)	/FS	S PROTEIN/FS S NS/FS	FS
Molecule Type	/MTY	S RNA/MTY	MTY
Organism	/ORGN	S CRASSOSTREA GIGAS/ORGN	ORGN
Patent Assignee **(3)**	/PA (or /CS)	S MOLECULAR DYNAMICS/PA	PA
Patent Country (code and text)	/PC	S WO/PC	PI
Patent Number **(2)**	/PN	S WO 2002074961/PN	PI
Patent Number Group **(2)**	/PATS	S WO 2002074961/PATS	PI
Publication Date **(1)**	/PD	S 20030130/PD	PI
Publication Year **(1)**	/PY	S 2003/PY	PI
Related Application Country	/RLC	S FR/RLC	RLI
Related Application Date **(1)**	/RLD	S 20020208/RLD	RLI
Related Application Number **(2)**	/RLN (or /RLI)	S EP2001-1102050/RLN	RLI
Related Application Year **(1)**	/RLY	S 2000-2001/RLY	RLI
Sequence Identity Number **(1)**	/SEQN	S 337/SEQN	SEQN
Sequence Length **(1)**	/SQL	S 150-175/SQL	SQL
Title	/TI	S HYBRIDIZATION ASSAY#/TI	TI
Update Date **(1)**	/UP	S 20020924/UP	UP

(1) Numeric search field that may be searched using numeric operators or ranges.
(2) Either STN or Derwent format may be used.
(3) Search with implied (S) proximity is available in this field.

Super Search Fields

Enter a super search code to execute a search in one or more fields that may contain the desired information. Super search fields facilitate crossfile and multifile searching. EXPAND may not be used with super search fields. Use EXPAND with the individual field codes instead.

Search Field Name	Search Code	Fields Searched	Search Examples	Display Codes
Application Number Group	/APPS	/AP, /RLN	S US2001-809003/APPS	AI, RLI

Sequence Similarity Searching (BLAST/GETSIM)

The GETSIM and BLAST® run packages are available to search the PCTGEN database for protein and nucleotide sequence data by similarity (homology). BLAST is provided in PCTGEN with the permission of the National Center for Biotechnology Information (NCBI) of the National Library of Medicine (NLM). GETSIM is provided in PCTGEN 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 PCTGEN 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 8.4+. 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.

In addition to the sequence alignment, a special format SEQO is provided in PCTGEN. The display format SEQO shows the corresponding original sequence which might include the nucleotide sequence of a PCTGEN record together with the protein sequence it expresses as given by the patent applicant.

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 a week)
(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
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 similarity search for a protein

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

UPLOAD SUCCESSFULLY COMPLETED
L1   GENERATED

=> D L1 LQUE

L1    ANSWER 1  PCTGEN COPYRIGHT 2008 WIPO on STN
LQUE  MAVMAPRTLL LVLSGVLALT QTWAGSHSMR YFYTSMSRPG RGEPRFFAVG
      YVDDTQFVRFDSDAASQRME PRAPWVEQEG PEYWDRETQN MKAQTQNAPV NLRNLRGYYN
      QSEAGSHTLQTMHGCDLGPD GRLLRGYYQS AYDGKDYFAL NEDLRSWTAA DLAAQNTQRK
      WEAADVAEQIRAYLEGRCVE WLRRYLENGK ETLQRADPPK THVTHHPVSD HEATLRCWAV
      GFYPAEITLTWQRDGEDQTQ DTELMETRPA GDGTFQKWAA VVVPSGKEQR YTCHVQHEGL
      PKPLTLRWEPSSQSTIPIVG IIAGLVLLGA MVIGAVVAAV MWRRKSSDRK GGSYSQAASS
      DSAQGSDVSLTACKV


=> RUN GETSIM L1/SQP

RUN GETSIM AT 16:03:10 ON 04 JUN 2008
COPYRIGHT (C) 2008 FIZ KARLSRUHE GMBH

     70000 SEQUENCES PROCESSED
    100000 SEQUENCES PROCESSED
    ……
    710000 SEQUENCES PROCESSED

  351 ANSWERS FOUND ABOVE A THRESHOLD OF 155
      QUERY SELF SCORE VALUE IS  2497
      BEST ANSWER SCORE VALUE IS 2231

Similarity
Score
     2231 ||
          ||||
          |||||
          |||||||
          |||||||
          |||||||
          |||||||
          |||||||||
          |||||||||
          ||||||||||
     1116 ||||||||||
          ||||||||||||
          ||||||||||||
          ||||||||||||
          ||||||||||||||||||
          ||||||||||||||||||||
          |||||||||||||||||||||||
          |||||||||||||||||||||||||
          |||||||||||||||||||||||||||||||||||||||||||||||||
          |||||||||||||||||||||||||||||||||||||||||||||||||
Answer Count      80       160       240       320       400
```

```
ENTER EITHER THE NUMBER OF ANSWERS YOU WISH TO KEEP
 OR ENTER MINIMUM PERCENT OF SELF SCORE FOLLOWED BY %
 (BEST ANSWER PERCENTAGE IS  89%)
ENTER (ALL) OR ? :80%
L2    RUN STATEMENT CREATED
L2          33 MAVMAPRTLLLVLSGVLALTQTWAGSHSMRYFYTSMSRPGRGEPRFFAVG
               YVDDTQFVRFDSDAASQRMEPRAPWVEQEGPEYWDRETQNMKAQTQNAPV
               NLRNLRGYYNQSEAGSHTLQTMHGCDLGPDGRLLRGYYQSAYDGKDYFAL
               NEDLRSWTAADLAAQNTQRKWEAADVAEQIRAYLEGRCVEWLRRYLENGK
               ETLQRADPPKTHVTHHPVSDHEATLRCWAVGFYPAEITLTWQRDGEDQTQ
               DTELMETRPAGDGTFQKWAAVVVPSGKEQRYTCHVQHEGLPKPLTLRWEP
               SSQSTIPIVGIIAGLVLLGAMVIGAVVAAVMWRRKSSDRKGGSYSQAASS
               DSAQGSDVSLTACKV/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 L2
L3           33 SOR L2 SCORE D

=> D 1 33 BIB SCORE ALIGN

L3    ANSWER 1 OF 33  PCTGEN COPYRIGHT 2008 WIPO on STN
AN    2007047796.10068  PRT       PCTGEN
TI    TISSUE- AND SERUM-DERIVED GLYCOPROTEINSAND METHODS OF THEIR USE
PA    Institute for Systems Biology
      Zhang, Hui
      Aebersold, Rudolf H.
PI    WO 2007047796     20070426
AI    PCT      2006-10-17
RLI   US 2005-728044P 20051017
ED    20070427
DT    Patent
SCORE 2231      89% of query self score 2497
ALIGN Smith-Waterman score:  2231
      365 aa overlap starting at 1
      mavmaprtlllvlsgvlaltqtwagshsmryfytsmsrpgrgeprffavgyvddtqfvrf
      ::::::::::::.:::.:::::::::::::::::.::.::::::::::.::::::::::::
      mavmaprtllllllsgalaltqtwagshsmryfftsvsrpgrgeprfiavgyvddtqfvrf
      dsdaasqrmeprapwveqegpeywdretqnmkaqtqnapvnlrnlrgyynqseagshtlq
      ::::::::::::::::.:::::::::.::.:.:::.:.  :.: .:::::::::::::::.:
      dsdaasqrmeprapwieqegpeywdqetrnvkaqsqtdrvdlgtlrgyynqseagshtiq
      tmhgcdlgpdgrllrgyyqsaydgkdyfalnedlrswtaadlaaqntqrkweaadvaeqi
       :.::::.: :::.:::: :.::::::::.:::::::::::::::.::: :.:::::::  :::.
      imygcdvgsdgrflrgyrqdaydgkdyialnedlrswtaadmaaqitkrkweaaheaeql
      raylegrcvewlrrylengketlqradppkthvthhpvsdheatlrcwavgfypaeitlt
      :::::.: ::::::::::::::::::.:::::::.:::::.::::::::::::.::::::::::
      rayldgtcvewlrrylengketlqrtdppkthmthhpisdheatlrcwalgfypaeitlt
      wqrdgedqtqdtelmetrpagdgtfqkwaavvvpsgkeqrytchvqheglpkpltlrwep
      ::::::::::::::.:::::::::::::::::::::::::.::::::::::::::::::::::
      wqrdgedqtqdtelvetrpagdgtfqkwaavvvpsgeeqrytchvqheglpkpltlrwel
      ssqstipivgiiaglvllgamvigavvaavmwrrkssdrkggsysqaassdsaqgsdvsl
      ::: :::::::::::::::.. :::::::::::::::::::::.::::::::::::::::
      ssqptipivgiiaglvllgavitgavvaavmwrrkssdrkggsytqaassdsaqgsdvsl
      tackv
      :::::
      tackv
```

```
L3    ANSWER 33 OF 33  PCTGEN COPYRIGHT 2008 WIPO on STN
AN    2007047796.10103  PRT       PCTGEN
TI    TISSUE- AND SERUM-DERIVED GLYCOPROTEINSAND METHODS OF THEIR USE
PA    Institute for Systems Biology
      Zhang, Hui
      Aebersold, Rudolf H.
PI    WO 2007047796     20070426
AI    PCT      2006-10-17
RLI   US 2005-728044P 20051017
ED    20070427
DT    Patent
SCORE 2050      82% of query self score 2497
ALIGN Smith-Waterman score:  2050
      366 aa overlap starting at 1
      mavmaprtlllvlsgvlaltqtwagshsmryfytsmsrpgrgeprffavgyvddtqfvrf
      : :::::.:::.::: ::::.::: ::::::: :..:::::::::..:::::::::::
      mrvmaprallllsgglaltetwacshsmryfdtavsrpgrgeprfisvgyvddtqfvrf
      dsdaasqrmeprapwveqegpeywdretqnmkaqtqnapvnlrnlrgyynqseagshtlq
      :::::: : :::::::::::::::::::.: :.:   :.:::::::::::: ::::::
      dsdaasprgeprapwveqegpeywdretqnykrqaqadrvslrnlrgyynqsedgshtlq
      tmhgcdlgpdgrllrgyyqsaydgkdyfalnedlrswtaadlaaqntqrkweaadvaeqi
       :.:::::::::::::: :::::::::.:::::::::::::: ::: :::: ::: .:::.
      rmygcdlgpdgrllrgydqsaydgkdyialnedlrswtaadtaaqitqrkleaaraaeql
      raylegrcvewlrrylengketlqradppkthvthhpvsdheatlrcwavgfypaeitlt
      :::::: ::::::::::::::::::.:::::::::.:::::::::.::::::::::
      raylegtcvewlrrylengketlqraeppkthvthhplsdheatlrcwalgfypaeitlt
      wqrdgedqtqdtelmetrpagdgtfqkwaavvvpsgkeqrytchvqheglpkpltlrwep
      ::::::::::::::.:::::::::::::::::::::::.::::::.::::: .:::: :::
      wqrdgedqtqdtelvetrpagdgtfqkwaavvvpsgqeqrytchmqheglqepltlswep
      ssqstipivgiiaglvllgam_vigavvaavmwrrkssdrkggsysqaassdsaqgsdvs
      ::: ::::.::.:::..: .. :.::::.:.: :::::  :::: :::: :.:::::: :
      ssqptipimgivaglavlvvlavlgavvtammcrrkssggkggscsqaacsnsaqgsdes
      ltackv
      : .::.
      litcka
```

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

```
=> RUN BLAST
CATGGTGGTTAAACTTACCTCATTAGCAGCATCCCTCTACAAGGTGCATTTAACTATAAGTATACT/SQN -E 100

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)……


 120 ANSWERS FOUND BELOW EXPECTATION VALUE OF 100.0

      QUERY SELF SCORE VALUE IS   131
      BEST ANSWER SCORE VALUE IS  131

Similarity
Score
      131 |
          |
          |||
          |||
          |||
          |||
          |||
          |||
          |||
          |||
       66 |||
          |||
          |||
          |||
          |||||
          |||||||||||||||||||||||||||||||||||||||
          |||||||||||||||||||||||||||||||||||||||
          |||||||||||||||||||||||||||||||||||||||
          |||||||||||||||||||||||||||||||||||||||
          |||||||||||||||||||||||||||||||||||||||
Answer Count      30        60        90       120       150

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%
L4    RUN STATEMENT CREATED
L4           7 CATGGTGGTTAAACTTACCTCATTAGCAGCATCCCTCTACAAGGTGCATT
               TAACTATAAGTATACT/SQN.-E 100

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

=> SOR SCORE D
PROCESSING COMPLETED FOR L4
L5            7 SOR L4 SCORE D

=> D 1 7 TI PA PI SCORE ALIGN

L5    ANSWER 1 OF 7  PCTGEN COPYRIGHT 2008 WIPO on STN
TI    RNS-sekretierende Bakterien
```

```
PA    Bachmann, Till; Villatte, Francois
PI    WO 2002024904     20020328
SCORE 131      100% of query self score 131
BLASTALIGN
   Query  = 66 letters
   Length = 452
   Score  =  131 bits (66), Expect = 7e-36
   Identities = 66/66 (100%)
   Strand = Plus / Plus

  Query: 1   catggtggttaaacttacctcattagcagcatccctctacaaggtgcatttaactataag
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
  Sbjct: 201 catggtggttaaacttacctcattagcagcatccctctacaaggtgcatttaactataag

  Query: 61  tatact 66
             ||||||
  Sbjct: 261 tatact 266

L5    ANSWER 7 OF 7  PCTGEN COPYRIGHT 2008 WIPO on STN
TI    Gene Expression Profiles in Breast Tissue
PA    Orr, Michael S.
      Nation, Michele
      Diggans, James C.
      Zeng, Wen
      Gene Logic, Inc.
PI    WO 2002059271     20020801
SCORE 117      89% of query self score 131
BLASTALIGN
   Query  = 66 letters
   Length = 576
   Score  =  117 bits (59), Expect = 1e-31
   Identities = 66/67 (98%), Gaps = 1/67 (1%)
   Strand = Plus / Plus

  Query: 1   catggtggttaaa-cttacctcattagcagcatccctctacaaggtgcatttaactataa
             ||||||||||||| ||||||||||||||||||||||||||||||||||||||||||||
  Sbjct: 231 catggtggttaaaacttacctcattagcagcatccctctacaaggtgcatttaactataa

  Query: 60  gtatact 66
             |||||||
  Sbjct: 291 gtatact 297
```

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 'GLP'GY/SQSP 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) 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
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

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.

Format	Content	Examples
AI (AP) **(1)**	Application Information	D 21 2 AI
AIO **(1)**	Application Information, Original	DIS AIO
AN	Accession Number	D AN TI
DT (TC)	Document Type	
ED	Entry Date	D AN ED
FEAT	Feature Table	D 1 5 10 FEAT
FS **(2)**	File Segment	
IDENT **(2,3)**	Percent Identity	D IDENT
MTY	Molecule Type	DIS L5 1-10 MTY
ORGN	Organism Name	D ORGN
PA (CS)	Patent Assignee	D 1-25 PA
PI (PN) **(1)**	Patent Information	D 1-15 PA PI
RLI **(1)**	Related Application Information	DIS RLI
RLIO	Related Application Information, Original	D RLIO
SCORE **(2,4)**	Similarity Score	D TI SCORE
SEQ **(5)**	Sequence (one-letter codes)	D 1-3 TI SEQ
SEQ3 **(5)**	Sequence (three-letter codes)	D 1 5 10 TI SEQ3
SEQN	Sequence Identity Number	D SEQN
SEQO **(5)**	Original Sequence (alignment of nucleotide sequence and peptide sequence it expresses when given)	D SEQO
SQL	Sequence Length	D 1-20 SQL
TI	Title	D L7 1-25 TI
UP	Update Date	D AN TI UP

PREDEFINED DISPLAY AND PRINT FORMATS

Format	Content	Examples
ALIGN **(4)**	Alignment between query and retrieved sequence in a similarity search (RUN GETSIM or RUN BLAST)	D ALIGN
ALL **(1)**	AN, MTY, TI, PA, PI, AI, RLI, DT, ORGN, SQL, SEQ, FEAT	D ALL
APPS **(1)**	AI, RLI	D APPS
CFAM **(1)**	Condensed family format (from INPADOCDB)	D CFAM
BIB **(1)**	AN, MTY, TI, PA, PI, AI, RLI, DT (BIB is the default)	D BIB
FAM **(1)**	AN, table of patent family information (from INPADOCDB)	D FAM
IBIB **(1)**	BIB, indented with text labels	D IBIB ALIGN
FASTA	FASTA format	D FASTA
FASTA2	FASTA format, header truncated	D FASTA2
IALL **(1)**	ALL, indented with text labels	D L2 1-5 IALL
LS **(1)**	Legal Status (from INPADOCDB)	D LS
LS2 **(1)**	Legal Status (from INPADOCDB), detailed version with display headers	D LS2
SQIDE **(5)**	TI, SQL, SEQ, FEAT	D SQIDE
SCAN **(6)**	TI (random display without answer numbers)	D SCAN
SQ3IDE **(5)**	TI, SQL, SEQ3, FEAT	D SQ3IDE
TRIAL (TRI,SAM, SAMPLE, FREE)	TI, MTY, SQL	D 1-20 TRI
HIT	Hit term(s) and field(s)	D HIT
KWIC	Up to 50 words before and after hit term(s) (KeyWord-In-Context)	D KWIC
OCC	Number of occurrences of hit term(s) and field(s) in which they occur	

(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) Custom display only.
(3) Use RUN BLAST first. See page 4, Similarity Search.
(4) Use RUN GETSIM or RUN BLAST first. See page 4, Similarity Search.
(5) Sequences in PCTGEN are given according to WST.25 of the WIPO.
(6) SCAN must be specified on the command line, e.g., 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
Accession Number	AN	N	Y
Application Country	AC	Y **(2)**	Y
Application Date	AD	Y **(2)**	Y
Application Information, Original	AIO	Y	Y
Application Number	AP (AI)	Y	Y
Application Number and Related Application Number	APPS	Y	N
Application Year	AY	Y **(2)**	Y
Document Type	DT (TC)	Y **(2)**	Y
Entry Date	ED	Y **(2)**	Y
Feature Table	FEAT	Y	N
File Segment	FS	Y	Y
Molecule Type	MTY	Y	Y
Organism Name	ORGN	Y	Y

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

Field Name	Field Code	ANALYZE/ SELECT (1)	SORT
Patent Assignee	PA	Y	Y
Patent Country	PC	Y	Y
Patent Number	PN (PI)	Y	Y
Percent Identity	IDENT	N	Y
Publication Date	PD	Y	Y
Publication Year	PY	Y	Y
Related Application Country	RLC	Y	Y
Related Application Date	RLD	Y	Y
Related Application Information, Original	RLIO	Y	Y
Related Application Number	RLN (RLI)	Y	Y
Related Application Year	RLY	Y	Y
Sequence (one-letter codes)	SEQ	Y **(2,3)**	N
Sequence (three-letter codes)	SEQ3	Y **(2,3)**	N
Sequence Identity Number	SEQN	Y	Y
Sequence Length	SQL	Y	Y
Similarity Score	SCORE **(4)**	N	Y
Title	TI	Y (default)	Y
Update Date	UP	Y **(2)**	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 PA.
(2) SELECT HIT and ANALYZE HIT are not valid with this field.
(3) Appends /SQSP to the terms created by SELECT.
(4) Used with a L-number created with BLAST and GETSIM.

Sample Records

DISPLAY TRIAL

```
TI    NOVEL NUCLEIC ACIDS AND POLYPEPTIDES
MTY   DNA
SQL   2093
```

DISPLAY SQIDE

```
AN    2002070737.12103  DNA       PCTGEN
TI    Compositions and Methods Relating to Osteoarthritis
SQL   100
SEQ
        1 agttgngtgc cgttggaccg naggaaaact catagactca tgggagcgtg
       51 aggcttcgag ccgcctaatt ttttaaccct aaatgtcgaa aggcttctgg

FEATURE TABLE:
Key         ¦Location¦
============+========+======================
misc_feature¦6, 21   ¦n = A,T,C or G
```

DISPLAY IALL

```
ACCESSION NUMBER: 2001057272.15599  DNA       PCTGEN
TITLE:            HUMAN GENOME-DERIVED SINGLE EXON NUCLEIC ACID PROBES USEFUL
                  FOR ANALYSIS OF GENE EXPRESSION IN HUMAN PLACENTA
PATENT ASSIGNEE:  Molecular Dynamics, Inc.Penn, Sharron G.Rank, David R.Hanzel,
                  David K.Chen, Wensheng
PATENT INFO:      WO 2001057272     20010809
REL APPL INFO:    US 2000-180312P 20000204; US 2000-207456P 20000526; US
                  2000-632366 20000803; GB 2000-24263 20001003; US 2000-236359P
```

```
                20000927; US 2000-234687P 20000921; US 2000-608408 20000630
ENTRY DATE:   20020923
DOCUMENT TYPE:    Patent
ORGANISM:         Homo sapiens
SEQUENCE LENGTH:  100
SEQUENCE
        1 cccagagatt ctgattctgc aaatcttgag cagcctgaga ttctgcagtt
       51 ctatgaagct tccaggtagt gtcaatgctg gtgctaggct gaccatagta

FEATURE TABLE:
Key       ¦Location¦
==========+========+============================
          ¦        ¦MAP TO AL035448.28
          ¦        ¦EXPRESSED IN PLACENTA, SIGNAL
          ¦        ¦= 1.5
          ¦        ¦NT HIT: U29185.1, EVALUE
          ¦        ¦7.00e-04
          ¦        ¦EST_HUMAN HIT: AA047634.1,
          ¦        ¦EVALUE 2.20e-01
```

DISPLAY SEQO

```
SEQO
 cgctcgcagt ctgtgggccc tccgggaggc ggcggaggtc accgcgggga gaggggcggg     60
 cgcagc   atg gca gcc tcc tta cgg ctc ctc gga gct gcc tcc ggt ctc     108
          Met Ala Ala Ser Leu Arg Leu Leu Gly Ala Ala Ser Gly Leu
            1               5                  10
 cgg tac tgg agc cgg cgg ctg cgg ccg gca gcc ggc agc ttt gca gcg      156
 Arg Tyr Trp Ser Arg Arg Leu Arg Pro Ala Ala Gly Ser Phe Ala Ala
  15                  20                  25                  30
 gtg tgt tct agg tca gtg gct tca aag act cca gtt gga ttc att gga      204
 Val Cys Ser Arg Ser Val Ala Ser Lys Thr Pro Val Gly Phe Ile Gly
                  35                  40                  45
 ctg ggc aac atg ggg aat cca atg gca aaa aat ctc atg aaa cat ggc      252
 Leu Gly Asn Met Gly Asn Pro Met Ala Lys Asn Leu Met Lys His Gly
              50                  55                  60
 tat cca ctt att att tat gat gtg ttc cct gat gcc tgc aaa gag ttt      300
 Tyr Pro Leu Ile Ile Tyr Asp Val Phe Pro Asp Ala Cys Lys Glu Phe
          65                  70                  75
 caa gat gca ggt gaa cag gta gta tct tcc cca gca gat gtt gct gaa      348
 Gln Asp Ala Gly Glu Gln Val Val Ser Ser Pro Ala Asp Val Ala Glu
      80                  85                  90
 aaa gct gac aga att att aca atg ctg ccc acc agt atc aat gca ata      396
 Lys Ala Asp Arg Ile Ile Thr Met Leu Pro Thr Ser Ile Asn Ala Ile
  95                  100                 105                 110
 gaa gct tat tcc gga gca aat ggg att cta aaa aaa gtg aag aag ggc      444
 Glu Ala Tyr Ser Gly Ala Asn Gly Ile Leu Lys Lys Val Lys Lys Gly
                 115                 120                 125
 tca tta tta ata gat tcc agc act att gat cct gca gtt tca aaa gaa      492
 Ser Leu Leu Ile Asp Ser Ser Thr Ile Asp Pro Ala Val Ser Lys Glu
             130                 135                 140
 ttg gcc aaa gaa gtt gag aaa atg gga gca gtt ttc atg gat gcc cct      540
 Leu Ala Lys Glu Val Glu Lys Met Gly Ala Val Phe Met Asp Ala Pro
         145                 150                 155
 gtt tct ggt ggt gta gga gct gca cga tct ggg aac ctc acg ttt atg      588
 Val Ser Gly Gly Val Gly Ala Ala Arg Ser Gly Asn Leu Thr Phe Met
     160                 165                 170
 gtg gga gga gtt gaa gat gaa ttt gct gct gcc caa gag ttg ctg ggg      636
 Val Gly Gly Val Glu Asp Glu Phe Ala Ala Ala Gln Glu Leu Leu Gly
 175                 180                 185                 190
```

```
tgc atg ggc tcc aac gtg gtg tac tgt gga gct gtt ggg act ggg cag      684
Cys Met Gly Ser Asn Val Val Tyr Cys Gly Ala Val Gly Thr Gly Gln
                195                 200                 205
gcg gca aag atc tgc aac aac atg ctg tta gct att agt atg att gga      732
Ala Ala Lys Ile Cys Asn Asn Met Leu Leu Ala Ile Ser Met Ile Gly
            210                 215                 220
act gct gaa gct atg aat ctt gga atc agg tta ggg ctt gac cca aaa      780
Thr Ala Glu Ala Met Asn Leu Gly Ile Arg Leu Gly Leu Asp Pro Lys
        225                 230                 235
cta ctg gct aaa atc cta aat atg agc tca gga cgg tgt tgg tca agt      828
Leu Leu Ala Lys Ile Leu Asn Met Ser Ser Gly Arg Cys Trp Ser Ser
    240                 245                 250
gac act tat aat cct gta cct gga gtg atg gat ggc gtt ccc tcg gct      876
Asp Thr Tyr Asn Pro Val Pro Gly Val Met Asp Gly Val Pro Ser Ala
255                 260                 265                 270
aat aac tat cag ggt gga ttt gga aca aca ctc atg gct aag gat ctg      924
Asn Asn Tyr Gln Gly Gly Phe Gly Thr Thr Leu Met Ala Lys Asp Leu
                275                 280                 285
gga ttg gca caa gac tct gct acc agc aca aag agc cca atc ctt ctt      972
Gly Leu Ala Gln Asp Ser Ala Thr Ser Thr Lys Ser Pro Ile Leu Leu
            290                 295                 300
ggc agt ctg gcc cat cag atc tac agg atg atg tgt gca aag ggc tac     1020
Gly Ser Leu Ala His Gln Ile Tyr Arg Met Met Cys Ala Lys Gly Tyr
        305                 310                 315
tca aag aaa gac ttc tca tcc gtg ttc cag ttc cta cga gag gag gag     1068
Ser Lys Lys Asp Phe Ser Ser Val Phe Gln Phe Leu Arg Glu Glu Glu
    320                 325                 330
acc ttc tga gtgtgcc ctttggccac ggacactgtt gggaaccaaa ctctgtcttg     1124
Thr Phe
335
gagcctcctt ttagctcact ccacaagtaa atggatttaa tcaaaggtca cctatctgct   1184
tttgattgtc taggtcacag taatccctag gatttttcac cgcttattct ttttgtcttt   1244
ttaacaaaca tattatccga attttttttc tgcaagccac tgatagtctc tgctaactag   1304
cttaattgac ctttttacaa agtttgatcc ccaagcatcc tcaactaaat cattgaatac   1364
ttcaatcagg atattatctg ctttacttta caaataaaac caaatctttt gtcaacagga   1424
tgaaacccat cttaaaggaa agaaaaggaa ttggtgtgaa gagagaagtt agagaaggga   1484
aatgcagtga attactatct gtgtccatca ggaagtttgt cctgttaacc aaatggttac   1544
tgcactacca gggttactgg tttatttttcc agggagctga taaagcagga gaactgttgc   1604
tgcatgtttt ctatttggac tccgtcacaa tatggtagga tatccctcac caactcccga   1664
cactcagcag acttgttttt atatttttttt ctttcttgtt cattcttact acgtatttttt   1724
tgacttaaga atgacatctt tagatgcatt tcagagccaa tgatgatatt tgctttagat   1784
aattattata ttattataaa tatagccata ttattttgaa ttcaaataaa tttctatact   1844
ggtaaaaaaa aaaaa                                                    1859
```

DISPLAY FASTA

```
=> d fasta

FASTA:
>PCTGEN|2009009830.168|PRT|sequence 168 from WO2009009830
mappqpklprpptrrractpsssarrrssqrrtpsslaswrpswppagsrppprwpqssraqrrtrkhqt
wwtacsgcwrrttcralwrrvrmaaspggtaprrcasspptrtvspwrrsrstrtrsswragttrtrsl
taasrstrrtacrrssttaqtrastacstkrrtirssspsrssssttasrasapsstsaaasappwppsp
httppsrgststsptsspsrrrrsrasptsaatcsrrcpwatpssgssttgvtsivarcsrtattrcqctv
rwcwwsascrtrrpsparrgcststscsrttqaagrgtrgstrrspgelaspassphtstptpgpsssps
```

DISPLAY FASTA2

```
=> d fasta2

FASTA2:
>PCTGEN|PRT
mappqpklprpptrrractpsssarrrssqrrtpsslaswrpswppagsrppprwpqssraqrrtrkhqt
wwtacsgcwrrttcralwrrvrmaaspggtaprrcasssppptrtvspwrrsrstrtrsswragttrtrsl
taasrstrrtacrrssttaqtrastacstkrrtirsssprsssssttasrasapsstsaaasappwppsp
httppsrgststsptssprrrrsrasptsaatcsrrcpwatpssgssttgvtsivarcsrtattrcqctv
rwcwwsascrtrrpsparrgcststscsrttqaagrgtrgstrrspgelaspassphtstptpgpsssps
```

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