GENESEQ (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.

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 or monthly (weekly is the default)		
CAS Registry Number® Identifiers	☐	SLART	☑
Keep & Share	☑	Structures	☐

Record Content

- Information on nucleic acid and protein sequences extracted from the original (basic) patent documents published by 57 patent offices worldwide.
- More than 65 million sequence records within the database stem from over 480,000 patents from around the globe.
- Records contain a Clarivate enhanced title from WPINDEX, a concise sequence description, an English abstract written especially for GENESEQ by one of Clarivate 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.

File Size

- More than 65.7 million records (12/2023)
- More than 45.2 million nucleic acid sequences (12/2023)
- More than 20.5 million protein sequences (12/2023)

Coverage

1980-present

Updates

Weekly

Language

English

Database Producer / Supplier

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

Sources

Patents from the 61 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

Cluster

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

STN Database Cluster information:
https://www.cas.org/support/training/stn/database-clusters

Search and Display Field Codes

General Search Fields

Search Field Name	Search Code	Search Examples	Display Codes
Basic Index **(4)** (contains single words from the title (TI), keyword (KW), description (DESC), organism species (ORGN), molecule type (MTY), and feature table (FEAT) fields)	None or /BI	S ANAPHYLATOXIN S PLANT GENE# AND RNA	TI, KW, DESC, ORGN, MTY, FEAT
Abstract	/AB	S GLUCOSE/AB	AB
Accession Number	/AN	S BJP34154/AN	AN
Amino Acid	/AA	S (T OR M)/AA	AA
Amino Acid Count **(1)**	/AA.CNT	S (T OR M OR F OR H)/AA (S) 50-100/AA.CNT	AA
Amino Acid Percentage **(1)**	/AA.PER	S (T OR M OR F OR H)/AA (S) 25-30/AA.PER	AA
Application Country	/AC	S US/AC	AI
Application Date **(1)**	/AD	S 20011129/AD	AI
Application Number **(2)**	/AP	S WO 2001-BA6/AP	AI
Application Number, Original	/APO	S WOBE000003/APO	APO
Application Year **(1)**	/AY	S 2002/AY	AI
Cross Reference	/CR	S HTTP://WWW.NCBI.NLM.NIH.GOV/GENE/10000/CR/CR	CR
Data Entry Date **(1)**	/DED	S 20190307/DED	DED
Description	/DESC	S SYNTHASE/DESC	DESC
Data Update Date **(1)**	/DUPD	S 20190307/DUPD	DUPD
Document Type (code and text)	/DT (or /TC)	S PATENT/DT	DT
Entry Date **(1)**	/ED	S 20211030/ED	ED
Field Availability	/FA	S AI/FA	FA
Feature Table **(4)**	/FEAT	S (RNA AND BINDING)/FEAT S ?COMBINAT?/FEAT	FEAT
File Segment (code and text)	/FS	S PROTEIN/FS S NS/FS	FS
Inventor	/IN	S MILLER/IN	IN
Language	/LA	S ENGLISH/LA	LA
Keyword	/KW	S MUTEIN/KW	KW
Molecule Type	/MTY	S RNA/MTY	MTY
Nucleic Acid	/NA	S (G OR C)/NA	NA
Nucleic Acid Count **(1)**	/NA.CNT	S (G OR C)/NA (S) 50-100/NA.CNT	NA
Nucleic Acid Percentage **(1)**	/NA.PER	S (G OR C)/NA (S) 60-70/NA.PER	NA
Organism Name **(3,4)**	/ORGN	S CRASSOSTREA GIGAS/ORGN	ORGN
Other Source	/OS	S 2020-A0561Q/OS	OS
Patent Assignee **(3)**	/PA (or /CS)	S MOLECULAR DYNAMICS/PA	PA
Patent Assignee Code	/PACO	S UYHA-N/PACO	PACO
Patent Country (code and text)	/PC	S WO/PC	PI
Patent Information Type	/PIT	S “USA9 CORRECTED PATENT APPLICATION (FROM 2001 ONWARDS)”/PIT	PI
Patent Number **(2)**	/PN	S WO2002074965/PN	PI
Patent Number Kind Code (**2)**	/PNK	S WO2002074965A2/PNK	PI
Patent Number, Original	/PNO	S WO200206834/PNO	PNO
Patent Number Group **(2)**	/PATS	S WO2002074965/PATS	PI
Patent Sequence Location	/PSL	S 6/PSL	PSL
Publication Date **(1)**	/PD	S 20030130/PD	PI
Publication Year **(1)**	/PY	S 2003/PY	PI
Priority Country	/PRC	S FR/PRC	PRAI
Priority Date **(1)**	/PRD	S 20150606/PRD	PRAI
Priority Date, First	/PRDF	S 20150608/PRDF	PRAI
Priority Number **(2)**	/PRN	S EP2001-102050/PRN	PRAI
Priority Number, Original	/PRNO	S DE04447388/PRNO	PRNO

General Search Fields (cont'd)

Search Field Name	Search Code	Search Examples	Display Codes
Priority Year **(1)**	/PRY	S 2000-2001/PRY	PRAI
Priority Year, First	/PRYF	S 2015/PRYF	PRAI
Sequence Key	/SEQK	S A000007FBFF70702CD23FD26650417EF67 EF9F464A27334A6217 /SEQK	SEQK
Sequence Identity Number **(1)**	/SEQN	S 337/SEQN	SEQN
Sequence Length **(1)**	/SQL	S 150-175/SQL	SQL
Title (**4)**	/TI	S HYBRIDIZATION ASSAY#/TI	TI
Update Date **(1)**	/UP	S 20211030/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.
(4) Fields that allow left truncation

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 cross-file and multi-file 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, /PRN	S US2001-809003/APPS	AI, PRAI

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
AA	Amino Acid	D AA
AB	Abstract	D AB
AI (AP) **(1)**	Application Information	D AI
AN	Accession Number	D AN
APO (AIO)	Application Number, Original	D APO
CR	Cross Reference	D CR
DED	Data Entry Date	D DED
DESC	Description	D DESC
DUPD	Data Update Date	D DUPD
DT (TC)	Document Type	D TC
ED	Entry Date	D AN ED
FASTA	Sequence (FASTA format)	D FASTA
FEAT	Feature Table	D 1 5 10 FEAT
FS **(2)**	File Segment	D FS
IDENT **(2,3)**	Percent Identity	D IDENT
IN	Inventor	D IN
LA	Language	D LA
KW	Keyword	D KW
MTY	Molecule Type	D MTY
ORGN	Organism Name	D ORGN
OS	Other Source	D OS
PA (CS)	Patent Assignee	D PA
PI (PN) **(1)**	Patent Information	D PI
PIT	Patent Information Type	D PIT
PNO	Patent Number, Original	D PNO
PRAI	Priority Information	D PRAI
PRNO	Priority Number, Original	D PRNP
PSL	Patent Sequence Location	D PSL
SCORE **(2,3)**	Similarity Score	D SCORE
SEQ **(4)**	Sequence (one-letter codes)	D SEQ
SEQ3 **(4)**	Sequence (three-letter codes)	D SEQ3
SEQK	Sequence Key	D SEQK
SEQN	Sequence Identify Number	D SEQN
SQL	Sequence Length	D 1-20 SQL
TI	Title	D L7 1-25 TI
UP	Update Date	D AN TI UP

(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 GETSIM or RUN BLAST first. See page 7, Similarity Search.

(4) Sequences in GENESEQ are given according to WST.25 of the WIPO.

Predefined Display and Print Formats

Format	Content	Examples
ABS	AN, AB	D ABS
ALIGN **(1)**	Alignment as text between query and retrieved sequence in a similarity search (RUN GETSIM, RUN BLAST, or RUN GETSEQ)	D ALIGN
ALIGNG **(1)**	Alignment as image between query and retrieved sequence in a similarity search (RUN GETSIM, RUN BLAST, or RUN GETSEQ)	D ALIGNG
ALL	AN, ED, UP, DED, DUPD, TI, IN, PA, PACO, LA, DT, PI, PIT, AI, PRAI, FS, CR, OS, MTY, PSL, DESC, KW, ORGN, AB, SEQN, SQL, SEQK, SEQ, AA or NA, FEAT	D ALL
IALL	ALL, indented with text labels	D L2 1-5 IALL
APPS	AI, PRAI	D APPS
BIB	AN, ED, UP, DED, DUPD, TI, IN, PA, PACO, LA, DT, PI, PIT, AI, PRAI, FS, CR, OS, MTY, PSL, DESC (BIB is the default)	D BIB
IBIB	BIB, indented with text labels	D IBIB ALIGN
FASTA	FASTA format	D FASTA
SCAN	ED, UP, DED, DUPD, TI, MTY, DESC (random display without answer numbers)	D SCAN
SQIDE	AN, ED, UP, DED, DUPD, MTY, ORGN, SEQN, SQL, SEQK, SEQ, AA or NA, FEAT	D SQIDE
SQ3IDE	AN, ED, UP, DED, DUPD, MTY, ORGN, SEQN, SQL, SEQK, SEQ3, AA or NA, FEAT	D SQ3IDE
TRIAL (TRI, SAM, SAMPLE, FREE)	AN, TI, MTY, DESC, KW, 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	D OCC

(1) Use RUN GETSIM, RUN BLAST or RUN GETSEQ first.

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	Y
Accession Number	AN	N	Y
Amino Acid,	AA	Y	N
Amino Acid, Count	AA.CNT	Y	N
Amino Acid, Percentage	AA.PER	Y	N
Application Country	AC	Y	Y
Application Date	AD	Y	Y
Application Number	AP (AI)	Y	Y
Application Number, Original	APO (AIO)	Y	Y
Application Number and Related Application Number	APPS	Y	N
Application Year	AY	Y	Y
Cross Reference	CR	Y	Y
Data Entry Date	DED	Y	Y
Data Update Date	DUPD	Y	Y
Description	DESC	Y	Y
Document Type	DT (TC)	Y	Y
Entry Date	ED	Y	Y
Feature Table	FEAT	Y	N
File Segment	FS	Y	Y
Inventor	IN	Y	Y
Language	LA	Y	Y
Keyword	KW	Y	Y
Molecule Type	MTY	Y	Y
Nucleic Acid	NA	Y	N
Nucleic Acid, Count	NA.CNT	Y	N
Nucleic Acid, Percentage	NA.PER	Y	N
Other Source	OS	Y	Y
Organism Name	ORGN	Y	Y
Patent Assignee	PA	Y	Y
Patent Country	PC	Y	Y
Patent Information Type	PIT	Y	Y
Patent Number	PN (PI)	Y	Y
Patent Number Group	PATS	Y	Y
Percent Identity	IDENT	N	Y
Priority Country	PRC	Y	Y
Priority Date	PRD	Y	Y
Priority Date, First	PRDF	Y **(2)**	Y
Priority Number	PRN	Y	Y
Priority Number, Original	PRNO	Y	Y
Priority Year	PRY	Y	Y
Priority Year, First	PRYF	Y **(2)**	Y
Patent Sequence Location	PSL	Y	Y
Publication Date	PD	Y	Y
Publication Year	PY	Y	Y
Sequence Identity Number	SEQN	Y	Y
Sequence Key	SEQK	Y	Y
Sequence Length	SQL	Y	Y
Similarity Score	SCORE **(3)**	N	Y
Title	TI	Y (default)	Y
Update Date	UP	Y	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) Used with a L-number created with BLAST and GETSIM.

Sequence Similarity Searching (BLAST/GETSIM)

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

Nucleotide and protein sequences can be subjected to a similarity search as a query entered directly on the command line using RUN GETSIM/BLAST or they may be uploaded via the “Structures” page. See details here. The uploaded sequence can be displayed with D LQUE.

To initiate a BLAST or GETSIM search with the command RUN BLAST or RUN GETSIM the following search codes have to be specified:

- /SQP for searching peptide sequences
- /SQN for nucleotide sequences
- /TSQN for searching peptide sequences translated from GENESEQ nucleotide sequences.

For the BLAST package four additional search codes are available:

- /SQM (megaBLAST) for searching highly similar nucleotide sequences
- /SQDM (discontiguous megaBLAST) for searching similar nucleotide sequences allowing more mismatches
- /TSQP for searching nucleotide sequences translated from GENESEQ protein sequences
- /TSQNX for searching translated nucleotides form GENESEQ protein sequences

It is recommended to use the search codes /SQM or /SQDM rather than /SQN when searching longer sequences as the response time is much faster. The commands /TSQN, /TSQP and /TSQNX are more time consuming compared to the other commands.

When using the /SQN, /SQM, /SQDM, or /TSQNX option, it is possible to specify whether single (SIN), complementary (COM), or BOTH strands should be searched. The options can be specified with the search code, e.g., /SQN -S COM. If no search option is given, BOTH (both) will be used by BLAST and GETSIM. Note that for the /TSQN option generally both strands will be searched.

GETSIM / BLAST: Types of Searches

Description	Search Code	Search Examples (1)
Peptide homology	/SQP	RUN BLAST L1 /SQP RUN GETSIM L1/SQP
Nucleotide homology	/SQN	RUN BLAST L1 /SQN RUN GETSIM L1/SQN
	/SQM **(2)**	RUN BLAST L1 /SQM
	/SQDM **(2)**	RUN BLAST L1 /SQDM
Translated peptide homology	/TSQN	RUN BLAST L1 /TSQN RUN GETSIM L1 /TSQN
Translated peptide homology from translated peptide	/TSQNX **(2)**	RUN BLAST L1/TSQNX
Translated nucleotide homology	/TSQP **(2)**	RUN BLAST L1 /TSQP

(1) Where L1 is a sequence query generated using the “Structure” page.
(2) BLAST only

The maximum number of hits is by default 15,000 records. The parameter "-maxseq" allows to increase the maximum number of hits to 100,000 records, e.g., =>RUN BLAST L1/SQN -F F -MAXSEQ 100000.

The number of additional results and their relevance in terms of high score and/or high identity values depend on the length of the query sequence and the number of subject sequences in the database.

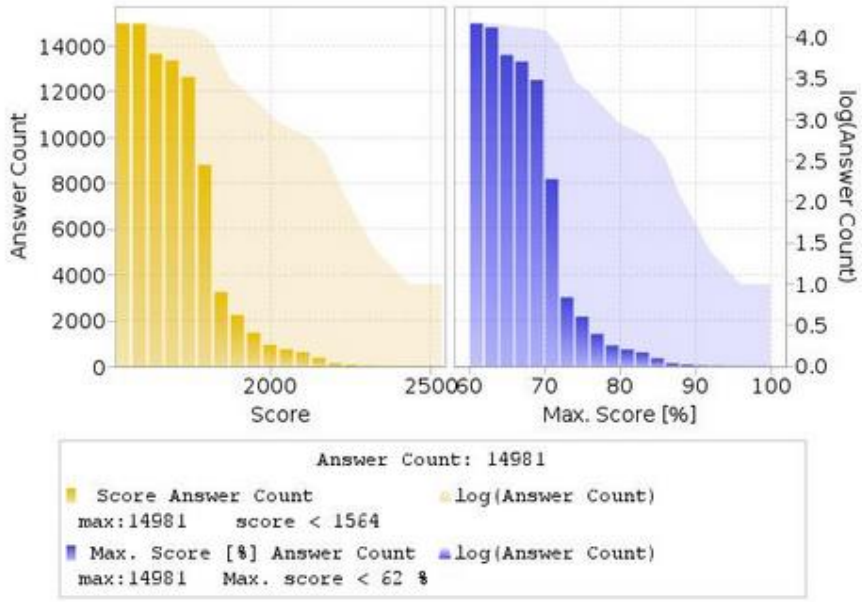
In general, searching a short sequence with -maxseq 100000 may retrieve additional documents with high score and high identity values while searching a longer sequence with -maxseq 100000 may retrieve only additional documents with high identity values.

GENESEQ

After a search with BLAST or GETSIM the number of retrieved sequences for the different score values are displayed in two diagrams. The y-axis of these diagrams represents the number of answers (absolute values are displayed as bars, logarithmic values are shaded) and the x-axis the score as the specific degree of similarity for this search. In the left diagram the score values are displayed, in the right diagram the percentage values of the maximum score.

In addition, two score values are given, the highest possible score value defining the maximum score when the query is aligned to itself, and the score of the best answer of the retrieved answer set. Both values are the same, if the query and at least one retrieved sequence are identical.

```
Highest possible score value: 2533.2
Best answer score value: 2533.2
```



Multiple answer sets (L-numbers) can be created with different cut off values for the score and the percentage identity. Five options are available:

1) Select a part of the answer set using the score value from the left histogram. The generated L-number contains all records with a score above the entered value.

```
ENTER EITHER "ALL" TO KEEP ALL ANSWERS
OR ENTER THE MINIMUM SCORE VALUE YOU WISH TO KEEP
OR ENTER THE MINIMUM PERCENT OF SCORE FOLLOWED BY "% SCORE"
OR ENTER THE MINIMUM PERCENT OF IDENTITY FOLLOWED BY "% IDENT"
OR COMBINE MINIMUM PERCENT OF SCORE AND IDENTITY AS "X% SCORE Y% IDENT"
OR ENTER "END". "END" MUST BE ENTERED TO COMPLETE THE RUN COMMAND.
ENTER (ALL) OR ? :2300

L10   RUN STATEMENT CREATED
L10         22 ATGGGATGGAGCTGTATCATCCTCTTCTTGGTAGCAACAGCTACAGGTGT
```

2) Select a part of the answer set using the percentage score value from the right histogram, e.g., "85%" or "85% SCORE". The generated L-number contains all records with a percentage score above the entered value.

```
ENTER EITHER "ALL" TO KEEP ALL ANSWERS
OR ENTER THE MINIMUM SCORE VALUE YOU WISH TO KEEP
OR ENTER THE MINIMUM PERCENT OF SCORE FOLLOWED BY "% SCORE"
OR ENTER THE MINIMUM PERCENT OF IDENTITY FOLLOWED BY "% IDENT"
OR COMBINE MINIMUM PERCENT OF SCORE AND IDENTITY AS "X% SCORE Y% IDENT"
OR ENTER "END". "END" MUST BE ENTERED TO COMPLETE THE RUN COMMAND.
ENTER (ALL) OR ? :85% SCORE

L11   RUN STATEMENT CREATED
L11         343 ATGGGATGGAGCTGTATCATCCTCTTCTTGGTAGCAACAGCTACAGGTGT
```

3) Select a part of the answer set using the percentage identity value, e.g., "100% IDENT".
The generated L-number contains all records with a percentage identity above the entered value.

```
ENTER EITHER "ALL" TO KEEP ALL ANSWERS
OR ENTER THE MINIMUM SCORE VALUE YOU WISH TO KEEP
OR ENTER THE MINIMUM PERCENT OF SCORE FOLLOWED BY "% SCORE"
OR ENTER THE MINIMUM PERCENT OF IDENTITY FOLLOWED BY "% IDENT"
OR COMBINE MINIMUM PERCENT OF SCORE AND IDENTITY AS "X% SCORE Y% IDENT"
OR ENTER "END". "END" MUST BE ENTERED TO COMPLETE THE RUN COMMAND.
ENTER (ALL) OR ? :100% IDENT

L13   RUN STATEMENT CREATED
L13          41 ATGGGATGGAGCTGTATCATCCTCTTCTTGGTAGCAACAGCTACAGGTGT
```

4) Select a part of the answer set combining the percentage score and the percentage identity value, e.g., "85% SCORE 100% IDENT". The generated L-number contains all records which have a percentage score and percentage identity above the entered value.

```
ENTER EITHER "ALL" TO KEEP ALL ANSWERS
OR ENTER THE MINIMUM SCORE VALUE YOU WISH TO KEEP
OR ENTER THE MINIMUM PERCENT OF SCORE FOLLOWED BY "% SCORE"
OR ENTER THE MINIMUM PERCENT OF IDENTITY FOLLOWED BY "% IDENT"
OR COMBINE MINIMUM PERCENT OF SCORE AND IDENTITY AS "X% SCORE Y% IDENT"
OR ENTER "END". "END" MUST BE ENTERED TO COMPLETE THE RUN COMMAND.
ENTER (ALL) OR ? :85% SCORE 100% IDENT

L14   RUN STATEMENT CREATED
L14          10 ATGGGATGGAGCTGTATCATCCTCTTCTTGGTAGCAACAGCTACAGGTGT
```

5) Keep the complete answer set with ALL.

```
ENTER EITHER "ALL" TO KEEP ALL ANSWERS
OR ENTER THE MINIMUM SCORE VALUE YOU WISH TO KEEP
OR ENTER THE MINIMUM PERCENT OF SCORE FOLLOWED BY "% SCORE"
OR ENTER THE MINIMUM PERCENT OF IDENTITY FOLLOWED BY "% IDENT"
OR COMBINE MINIMUM PERCENT OF SCORE AND IDENTITY AS "X% SCORE Y% IDENT"
OR ENTER "END". "END" MUST BE ENTERED TO COMPLETE THE RUN COMMAND.
ENTER (ALL) OR ? :ALL

L15   RUN STATEMENT CREATED
L15       14981 ATGGGATGGAGCTGTATCATCCTCTTCTTGGTAGCAACAGCTACAGGTGT
```

In order to complete the RUN BLAST or the RUN GETSIM command, END must be entered.

```
ENTER EITHER "ALL" TO KEEP ALL ANSWERS
OR ENTER THE MINIMUM SCORE VALUE YOU WISH TO KEEP
OR ENTER THE MINIMUM PERCENT OF SCORE FOLLOWED BY "% SCORE"
OR ENTER THE MINIMUM PERCENT OF IDENTITY FOLLOWED BY "% IDENT"
OR COMBINE MINIMUM PERCENT OF SCORE AND IDENTITY AS "X% SCORE Y% IDENT"
OR ENTER "END". "END" MUST BE ENTERED TO COMPLETE THE RUN COMMAND.
ENTER (ALL) OR ? :END
```

An L-number is generated for each selection, which contains all answers of the specified subset. Each L-number can be used for further processing. As the initial L-number is sorted by descending accession number, the selected L-number may be re-arranged by descending similarity score (SORT SCORE D L1) or descending percent identity (SORT IDENT D L1).

The alignment between the retrieved sequence and the query sequence can be displayed as text with the display format ALIGN or as an image with ALIGNG. The top line is the query sequence and the bottom line the hit sequence. Above each alignment the percentage of the BLAST and GETSIM score compared to the query self-score value and the percentage of identity is given. Both values can also be displayed as well with D SCORE and D IDENT. Both BLAST and GETSIM ALIGN format follows the standard convention for NCBI alignment displays. See further details in HELP ALIGNMENT.

```
ALIGNG
```

Query Length: 303; **Sequence Length:** 591;
Score: 277.2 bits (306), 50.6% of highest possible score 547.7;
Expect value: 1.877e-71;
Identities: 158 / 160 (98.8%);
Query Identity: 52.1%; **Query Coverage:** 52.8%;
Subject Identity: 26.7%; **Subject Coverage:** 27.1%;
Strand: Plus / Plus; **Alignment Length:** 160;

```
Q: 144 TCTGGGCTTCTTGCATTCTGGGACAGCCAAGTCTGTGACTTGCACGTACTCCCCTGCCCT 203
       ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
S: 1   TCTGGGCTTCTTGCATTCTGGGACAGCCAAGTCTGTGACTTGCACGTACTCCCCTGCCCT 60
Q: 204 CAACAAGATGTTTTGCCAACTGGCCAAGACCTGCCCTGTGCAGCTGTGGGTTGATTCCAC 263
       |||||||||||||||||||||||||||||||||||| ||||||||||||||||||||
S: 61  CAACAAGATGTTTTGCCAACTGGCCAAGACCTGCCCTGCGCAGCTGTGGGTTGATTCCA- 119
Q: 264 ACCCCCGCCCGGCACCCGCGTCCGCGCCATGGCCATCTAC                     303
       |||||||||||||||||||||||||||||||||||||||
S: 120 ACCCCCGCCCGGCACCCGCGTCCGCGCCATGGCCATCTAC                     159
```

Advanced User Options for BLAST and GETSIM

For the experienced user of BLAST® and GETSIM a variety of options are available via the STN command line. Altering these parameters will have a profound effect on the outcome of the search. It is strongly recommended that users are completely familiar with NCBI documentation before embarking on customizing any of these settings. For further information see the information on the NCBI website.

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 -F F or RUN BLAST L1/SQP -E 0.1 -M PAM30.

Advanced User Options

Option	Switch	Values
1. Filter	-f	T (True), F (False), Default value is T. If T is set, for peptides the SEG, and for nucleotides the DUST filter is employed.
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 for nucleotides only	-s	1 (SIN), 2 (COM) or 3 (BOTH) default value is 3
5. Matrix for peptides only	-m	*BLAST* BLOSUM62 (default), BLOSUM80, BLOSUM45, PAM30, PAM70 *GETSIM* BL50 (default), BL62, BL80, MD10, MD20, MD40, OPT5, P120, P250, VT160
6. Gap Penalty	-g	Peptides (default): BLAST 11; GETSIM 12 Nucleotides (default): BLAST 5; GETSIM 12
7. Gap Extension	-x	Peptides: BLAST 1; GETSIM 2 Nucleotides (default): BLAST 2; GETSIM 4
8. Penalty for nucleotide mismatch	-q	BLAST: -3 (default); GETIM: -2 (default)
9. Reward for nucleotide match	-r	BLAST: 1 (default); GETSIM: 3 (default)

BLAST Matrix settings (for option 5. Matrix)

Please note that for a certain matrix only a restricted set of possible gap and gap extension values are 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
BLOSUM50	32767	32767
	13	3
	12	3
	11	3
	10	3
	9	3
	16	2
	15	2
	14	2
	13	2 (default)
	12	2
	19	1
	18	1
	17	1
	16	1
	15	1

Matrix	Gap	Gap Extension
BLOSUM90	32767	32767
	9	2
	8	2
	7	2
	6	2
	11	1
	10	1 (default)
PAM30	9	1
	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
PAM250	32767	32767
	15	3
	14	3
	13	3
	12	3
	11	3
	17	2
	16	2
	15	2
	14	2 (default)
	13	2
	21	1
	20	1
	19	1
	18	1
	17	1

Searching Sequence Data with the GETSEQ RUN Package

The GETSEQ run package is a tool to search the GENESEQ 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. The maximum number of hits is 250,000 records.

Nucleotide and protein sequences can be subjected to a GETSEQ search as a query entered directly on the command line using RUN GETSEQ or the query may be created with the QUERY command, and subsequently searched through the GETSEQ run package specifying the query L-number (e.g., RUN GETSEQ L1, if L1 represents the sequence query).

```
=> RUN GETSEQ MCLHFLVLVICIL/SQSP

RUN GETSEQ AT 08:57:25 ON 2021-10-11
COPYRIGHT (C) 2021 FIZ KARLSRUHE on STN

GetSeq motif search by FIZ Karlsruhe; Version: 1.0.0

Query time:          115
L13   RUN STATEMENT CREATED
L13          30 MCLHFLVLVICIL/SQSP
```

Long sequences may be uploaded via the “Structures” page; see details here. The L-number may also derive from a previous sequence search in another STN database with bio sequence search capabilities, e.g., the CAS REGISTRY℠ file.

Any L-numbered sequence answer set from RUN GETSEQ may be combined with any search field in the GENESEQ file, for example => S L1 AND ARTIFICIAL SEQUENCE/ORGN where L1 represents the answer set from a RUN GETSEQ operation.

Hits of the retrieved sequence can be displayed in context of the whole sequences as text with the display format ALIGN or as an image with ALIGNG.

```
=> D ALIGN
L3     ANSWER 1 OF 30 GENESEQ COPYRIGHT 2021 CLARIVATE ANALYTICS on STN.
ALIGN
  Sequence Length: 43;

  Hits at: 8-20
     1 MFTIRSRMCL HFLVLVICIL RECESVCVCV CVCVCLWHLG RVV
     = ==========
```

The HIT display format contains only the part of the hit sequence with the matching residues which are highlighted with double underlining. In addition, the information HITS AT: gives the residue number of the start and end point of the matching part of the hit sequence.

```
 => D HIT
L5     ANSWER 50 OF 147 GENESEQ COPYRIGHT 2021 CLARIVATE ANALYTICS on STN.
SEQ
           SGTTGKPKG
           =========

  Hits at: 413-420 3426-3433 4466-4473
```

Sequence Search Terms

Amino acid and nucleic acid sequences may be searched with the one-letter code, amino acids also with the three-letter codes for common amino acids. Enter HELP AAC for a table of the one- and three-letter codes of the common amino acids and HELP NUC for a table of the codes for nucleic acids.

Uncommon amino acids are represented in the sequence by an 'X' (or 'Xaa'). ‘X’ is used also as an unspecified amino acid since July 2022 with standard ST.26. If you want to search specifically for an 'X' in the sequence, it has to be placed in square brackets, e.g., =>RUN GETSEQ TF[X]C[X]T/SQSP

Terms	Search Examples
One-letter codes for common amino acids	LAGLL/SQSP
Three-letter codes for common amino acids Enclose strings of codes in single quotes and use dashes to separate codes in strings.	'HIS-LEU-TYR-LEU-GLN-TYR-ILE-ARG-LYS-LEU'/SQSFP 'HIS-LEU-TYR-LEU-GLN-TYR-ILE-ARG-LYS-LEU' /SQEP
One-letter codes for nucleic acids	ATGAAN/SQEN CATCTGTATT/SQSN

Types of Sequence Searches

In the GETSEQ run package four options are available for searching polypeptide sequences using amino acid codes and two options for searching nucleic acid sequences.

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**.**	/SQEP	GAPGEK/SQEP 'ASP-HIS-ALA-ILE-HIS' /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.	/SQEFP	YGGFL/SQEFP 'TYR-GLY-GLY-PHE-LEU'/SQEFP
Subsequence, Protein	Search for exact answers plus sequences in which the query sequence is embedded.	/SQSP	LAGLL/SQSP 'ASP-HIS-ALA'/SQSP
Subsequence Family, Protein	Search for exact sequences, subsequences, and answers in which family-equivalent substitution of the query amino acids occurs.	/SQSFP	ATCXAWV/SQSFP 'THR-ASP-SER-GLU-SER-SER-HIS' /SQSFP
Sequence Exact, Nucleic Acid	Search for sequences that match the query. Ambiguity codes for nucleic acids are allowed.	/SQEN	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.	/SQSN	TGGAGAAGGC/SQSN

The families of amino acid equivalents retrieved in the polypeptide family searches SQEFP and SQSFP 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)

Variability Symbols for Sequence Code Match Searches

Variability symbols are allowed in all GETSEQ search options. For more information on specifying variability in sequence code match queries, enter HELP SQQ.

Symbol(s)	Function	Query Examples
[ ]	to specify alternate residues	NGSLLAGAYAIST[LV]I/SQSP LGP['VAL-LEU-LYS']/SQSP
[-]	to exclude a specific residue or alternate residues	LGP[-H]/SQSP LGP[-'HIS']/SQSFP LGP[-HL]/SQSP
{m}	to repeat the preceding sequence m times	(FL){2}/SQSP (CTGA){3}/SQSN TAA(TAAA){2}/SQSN
{m,u} or {m-u}	to repeat the preceding sequence m to u times	GG(FL){1,2}/SQSP (CTGA){2,4}/SQSN
? or {0,1} or {0-1}	to repeat the preceding sequence zero or one time	FLRRI(RP)?K/SQSP FLRRI(RP){0,1}K/SQSP CATG(CGTA){0,1}GGAC/SQSN
* or {0,} or {0-}	to repeat the preceding sequence zero or more times	KLK(WD){0,}N/SQSP KLK(WD)*N/SQSP CATAA(CTG){0,}TATT/SQSN
+ or {1,} or {1-}	to repeat the preceding sequence one or more times	KLK(DLE){1,}/SQSP KLK(DLE)+/SQS CATA(CTG){1,}TATT/SQSN
^ (Caret)|	search at the beginning or end of a sequence specifies alternate residues	^MCGIL/SQS VCDS^/SQSP ACDS|KLMP/SQSP
&	to join together sequence expressions or queries (L#s)	

SPECIFYING GAPS IN GETSEQ SEQUENCE QUERIES

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

Sample Records

DISPLAY TRIAL

L1 ANSWER 1 OF 1 GENESEQ COPYRIGHT 2021 CLARIVATE on STN.
AN **BHN82727** GENESEQ
TI Serratia species used in vaccine, composition and kit for preventing or treating disease in fish, preferably in salmonids, has average nucleotide identity of ninety five percent or less with Serratia proteomaculans.
MTY DNA
DESC Serratia sp. gbpA gene PCR primer (237F), SEQ ID 18.
KW GlnNAc binding protein; PCR; antibacterial; bacterial infection; gbpA gene; microorganism; microorganism detection; primer; prophylactic to disease; ss; therapeutic; vaccine antibacterial; veterinary
SQL 25

DISPLAY SQIDE

L3 ANSWER 1 OF 1 GENESEQ COPYRIGHT 2021 CLARIVATE on STN.
AN **BCL98591** GENESEQ ED 20211030 UP 20211030
DED 20160324
MTY DNA
ORGN Streptomyces sp SirexAA-E
SEQN 50
SQL 591
SEQK 5ab00e7851602f4d762ca4ffdce0e930c2ea237105e4009dfb506d92e9a0396a

SEQ

```
      1 atgcggaaaa gggcaagcgc ggccgtcata ggcctggcga tcgccggcgt
     51 ctcgatgttc gccaccagca gtgccagcag ccacggctac accgattccc
    101 ccatcagcag acagaagctg tgtgccaacg gcaccgtcac cggctgcggc
    151 aacatccagt gggagccgca gagcgtcgag ggcccgaagg gcttcccggc
    201 ggcaggtccg gcggacggca agatctgcgc cggcggaaac agctccttcg
    251 ccgcgctcga cgacccgcgc gggggcaact ggcccgccac ccaggtcacc
    301 ggcggccagg gctacaactt ccgctggcag ttcaccgccc gccacgccac
    351 gaccgacttc cggtactaca tcaccaagga cggctgggac tccaccaagc
    401 cgctcaccag ggccgccctg gagtcgcagc ccttcatgac ggtgccgtac
    451 gggaaccagc agcccccggc gaccctgacc caccagggca ccatccccac
    501 ccagaagtcc ggcaagcaca tcatcctggc cgtctggaac gtggctgaca
    551 ccgccaacgc gttctacgcg tgctcggacg tgaagttctg a
```

NA

Code	Count	Percent
A	115	19.5
C	223	37.7
G	179	30.3
T	74	12.5
U	0	0.0
Other	0	0.0

FEATURE TABLE:

Key	Location	Qualifier	
CDS	1..591	*tag= a	
		product	"chitin-binding domain 3 protein"

DISPLAY IALL

L3 ANSWER 1 OF 1 GENESEQ COPYRIGHT 2021 CLARIVATE on STN.
AN **BCL98591** GENESEQ ED 20211030 UP 20211030
DED 20160324 Full-text
TI Digesting a lignocellulosic material, comprises exposing the lignocellulosic material to cultured Streptomyces sp. ActE secretome preparation where at least partial lignocellulosic digestion occurs.
IN Fox BG; Takasuka T; Book AJ; Currie CR
PA FOX B G (FOXB-I)
TAKASUKA T (TAKA-I)
BOOK A J (BOOK-I)
CURRIE C R (CURR-I)
LA English
DT Patent
PI **US 20160032340** A1 20160204
PIT USA1 FIRST PUBLISHED PATENT APPLICATION [FROM 2001 ONWARDS]
AI **US 2015-851812** 20150911
PRAI **US 2011-61579301** 20111222 (61)
US 2011-61579897 20111223 (61)
US 2012-709971 20121210
FS NUCLEIC; NS
CR BCL98576
OS 2016-09721M [15]
MTY DNA
PSL Disclosure; SEQ ID NO 50; 173pp
DESC Streptomyces chitin-binding domain 3 protein CBM33 coding DNA, SEQ:50.
KW chitin-binding domain 3 gene; degradation; ds; feedstuff; gene
ORGN Streptomyces sp SirexAA-E

AB The present invention relates to a method for digesting a lignocellulosic material by exposing the lignocellulosic material to cultured Streptomyces sp. ActE secretome. The invention further claims: (1) a purified preparation comprising the Streptomyces sp. ActE secretome; and (2) a composition useful for digesting lignocellulosic material. The method of the invention is useful for digesting a lignocellulosic material and degradation of biomass which is used as animal feed. The present sequence is a Streptomyces sp. ACTE chitin-binding domain 3 protein CBM33 coding DNA SACTE_2313, which is useful in the method for digesting a lignocellulosic material.

SEQN 50

SQL 591

SEQK 5ab00e7851602f4d762ca4ffdce0e930c2ea237105e4009dfb506d92e9a0396a

SEQ

```
   1 atgcggaaaa gggcaagcgc ggccgtcata ggcctggcga tcgccggcgt
  51 ctcgatgttc gccaccagca gtgccagcag ccacggctac accgattccc
 101 ccatcagcag acagaagctg tgtgccaacg gcaccgtcac cggctgcggc
 151 aacatccagt gggagccgca gagcgtcgag ggcccgaagg gcttcccggc
 201 ggcaggtccg gcggacggca agatctgcgc cggcggaaac agctccttcg
 251 ccgcgctcga cgacccgcgc gggggcaact ggcccgccac ccaggtcacc
 301 ggcggccagg gctacaactt ccgctggcag ttcaccgccc gccacgccac
 351 gaccgacttc cggtactaca tcaccaagga cggctgggac tccaccaagc
 401 cgctcaccag ggccgccctg gagtcgcagc ccttcatgac ggtgccgtac
 451 gggaaccagc agccccggc gaccctgacc caccagggca ccatccccac
 501 ccagaagtcc ggcaagcaca tcatcctggc cgtctggaac gtggctgaca
 551 ccgccaacgc gttctacgcg tgctcggacg tgaagttctg a
```

NA

Code	Count	Percent
A	115	19.5
C	223	37.7
G	179	30.3
T	74	12.5
U	0	0.0
Other	0	0.0

FEATURE TABLE:

Key	Location	Qualifier	
CDS	1..591	*tag= a	
		product	"chitin-binding domain 3 protein"

DISPLAY FASTA

```
L4      ANSWER 1 OF 1 GENESEQ COPYRIGHT 2021 CLARIVATE on STN.

FASTA

        >GENESEQ|BFE50692|DNA|sequence 1 from WO2018060498
        atgaacaaaacttcccgtaccctgctctctctgggcctgctgagcgcggccatgttcggcgtttcgcaacag
        gcgaatgcccacggttatgtcgaatcgccggccagccgcgcctatcagtgcaaactgcagctcaacacgcag
        tgcggcagcgtgcagtacgaaccgcagagcgtcgagggcctgaaaggcttcccgcaggccggcccggctgac
        ggccatatcgccagcgccgacaagtccaccttcttcgaactggatcagcaaacgccgacgcgctggaacaag
        ctcaacctgaaaaccggtccgaactcctttacctggaagctgaccgcgcgtcacagcaccaccagctggcgc
        tatttcatcaccaagccgaactgggacgcttcgcagccgctgacccgcgcttcctttgacctgacgccgttc
        tgccagttcaacgacggcggcgccatccctgccgcacaggtcacccaccagtgcaacataccggcagatcgc
        agcggttcgcacgtgatccttgccgtgtgggacatagccgacaccgctaacgccttctatcaggcgatcgac
        gtcaacctgagcaaataa
```

DISPLAY SEQ3

```
L6      ANSWER 1 OF 134465 GENESEQ COPYRIGHT 2021 CLARIVATE on STN.

SEQ3

              1 Met-Ala-Pro-Ala-Ala-Ala-Phe-Leu-Ser-Ala-
             11 Cys-Ala-Ala-Gly-Ser-Ile-Pro-Arg-Ala-Pro-
             21 Phe-Leu-Ile-Pro-Arg-Pro-Leu-Leu-Leu-Pro-
             31 Ile-Pro-Leu-Ser-Pro-Ala-Arg-Trp-Asp-Arg-
             41 Ser-Arg-Ser-Cys-Ser-Leu-Phe-Gly-Val-Gly-
             51 Ala-Asn-Thr-Arg-Arg-Ala-Pro-Thr-Leu-Arg-
             61 Arg-Asn-Ala-Ser-Thr-Glu-Thr-Val-Val-Pro-
             71 Tyr-Val-Pro-Gly-Ser-Gly-Lys-Tyr-Ile-Ala-
             81 Pro-Asp-Tyr-Leu-Val-Lys-Lys-Val-Ser-Ala-
             91 Glu-Glu-Val-Gln-Glu-Leu-Val-Arg-Gly-Gln-
            101 Arg-Lys-Val-Pro-Leu-Ile-Val-Asp-Phe-Tyr-
            111 Ala-Thr-Trp-Cys-Gly-Pro-Cys-Val-Gln-Met-
            121 Ala-Gln-Asp-Ile-Glu-Met-Leu-Ala-Val-Glu-
            131 Tyr-Glu-Asp-Asn-Ala-Leu-Phe-Val-Lys-Val-
            141 Asp-Thr-Asp-Asp-Glu-Tyr-Glu-Phe-Ala-Lys-
            151 Asp-Met-Gln-Val-Arg-Gly-Leu-Pro-Thr-Leu-
            161 Tyr-Phe-Phe-Ser-Pro-Asp-Gln-Asn-Lys-Asp-
            171 Ala-Ile-Arg-Thr-Glu-Gly-Leu-Ile-Pro-Met-
            181 Asp-Met-Ile-Arg-Asn-Ile-Ile-Asp-Asn-Glu-
            191 Leu
```