USGENE® (USPTO Genetic Sequence Database)

Subject Coverage

Nucleotide and amino acid sequence data as submitted by patent applicants to the United States Patent and Trademark Office (USPTO).

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

- All available peptide and nucleic acid sequences from the published applications and issued patents of the United States Patent and Trademark Office (USPTO).
- Extensive bibliographic and text search options, including publication title, abstract, patent assignees at issue, full inventor names, plus the complete set of publication, application, priority, and parent case WIPO/PCT numbers and dates.
- Each record includes the actual sequence and additional information on the sequence, e.g., molecule type and organism, and sequence length.

File Size

- More than 118 million records (12/2023)
- More than 79.6 million nucleic acid sequences (12/2023)
- More than 33.4 million protein sequences (12/2023)

Coverage

1980-present

Updates

Weekly

Language

English

Database Producer

SequenceBase Corporation
3 Dellview Drive
Edison, NJ 08820-2545
USA

Copyright Holder

Sources

Published applications and issued patents of the USPTO.

User Aids

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

USGENE

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), description (DESC), organism species (ORGN), molecule type (MTY), and feature table (FEAT) fields)	None or /BI	S ANAPHYLATOXIN S PLANT GENE# AND RNA	TI, DESC, ORGN, MTY, FEAT
Abstract	/AB	S GLUCOSE/AB	AB
Accession Number	/AN	S 11203790.3/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 US 2001-997425/AP	AI
Application Number, Original	/APO	S US2005-100212 /APO	APO
Application Year **(1)**	/AY	S 2002/AY	AI
Calculated Expiration Date	/XPY	S 20240102/XPD	XPD
Calculated Expiration Year	/XPY	S 2024/XPY	XPD
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 GHRH/DESC	DESC
Document Type (code and text)	/DT (or /TC)	S PATENT/DT	DT
Entry Date **(1)**	/ED	S 20211224/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
Inventor, Address	/INA	S LONDON/INA	IN
Main Claim	/MCLM	S GLUCOSE/MCLM	MCLM, CLM
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
Patent Assignee **(3)**	/PA (or /CS)	S MOLECULAR DYNAMICS/PA	PA
Patent Assignee, Address	/PAA	S NEW YORK/PAA	PA
Patent Country (code and text)	/PC	S US/PC	PI
Patent Information Type	/PIT	S “USA9 CORRECTED PATENT APPLICATION (FROM 2001 ONWARDS)”/PIT	PI
Patent Number **(2)**	/PN	S US11202830/PN	PI
Patent Number Kind Code (**2)**	/PNK	S US11202830B2/PNK	PI
Patent Number, Original	/PNO	S US11202830/PNO	PNO
Patent Number Group **(2)**	/PATS	S US11202830/PATS	PI
Patent Sequence Location	/PSL	S 10/PSL	PSL
Patent Term Adjustment (number of days) **(1)**	/PTA	S 100-150/PTA	XPD
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

General Search Fields (cont'd)

Search Field Name	Search Code	Search Examples	Display Codes
Priority Number **(2)**	/PRN	S EP2001-102050/PRN	PRAI
Priority Number, Original	/PRNO	S DE1980-3023813/PRNO	PRNO
Priority Year **(1)**	/PRY	S 2000-2001/PRY	PRAI
Priority Year, First	/PRYF	S 2015/PRYF	PRAI
Related Application Country	/RLC	S US/RLC	RLI
Related Application Date	/RLD	S 20100106/RLD	RLI
Related Application Number	/RLN	S US1978-910559/RLN	RLI
Related Application Type	/RLT	S EARLIER APPLICATION/RLT	RLI
Related Application Year	/RLY	S 2012/RLY	RLI
Related Publication Country	/RLPC	S WO/RLPC	RLPI
Related Publication Date	/RLPD	S 20140116/RLPD	RLPI
Related Publication Number	/RLPN	S WO2014001422/RLPN	RLPI
Related Publication Year	/RLPY	S 2015/RLPY	RLPI
Sequence Count **(1)**	/SEQC	S 7/SEQC	SEQC
Sequence Key	/SEQK	S A00000ED1BC0D49FACA2F472D1551B121561C12A2A43231981626FB510C442F4/SEQK	SEQK
Sequence Identity Number **(1)**	/SEQN	S 337/SEQN	SEQN
Sequence Source	/SSO	S NCBI/SSO	SSO
Sequence Length **(1)**	/SQL	S 150-175/SQL	SQL
Title (**4)**	/TI	S HYBRIDIZATION ASSAY#/TI	TI
Update Date **(1)**	/UP	S 20211224/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 table	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
CLM	Claims	D CLM
CR	Cross Reference	D CR
DED	Data Entry Date	D DED
DESC	Description	D DESC
DT (TC)	Document Type	D DT
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
MCLM	Main Claim	D MCLM
MTY	Molecule Type	D MTY
NA	Nucleic Acid Table	D NA
ORGN	Organism Name	D ORGN
PA	Patent Assignee	D PA
PI	Patent Information	D PI
PNO	Patent Number, Original	D PNO
PRAI	Priority Information	D PRAI
PRNO	Priority Number, Original	D PRNP
PSL	Patent Sequence Location	D PSL
RLI	Related Application Information	D RLI
RLPI	Related Publication Information	D RLPI
SCORE **(2,3)**	Similarity Score	D SCORE
SEQ **(4)**	Sequence (one-letter codes)	D SEQ
SEQ3 **(4)**	Sequence (three-letter codes)	D SEQ3
SEQC	Sequence Count	D SEQC
SEQK	Sequence Key	D SEQK
SEQN	Sequence Identify Number	D SEQN
SQL	Sequence Length	D 1-20 SQL
SSO	Sequence Source	D SSO
TI	Title	D L7 1-25 TI
UP	Update Date	D AN TI UP
XNTE	Patent Expiration Note	D XNTE
XPD	Calculated Expiration Date	D XPD

(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 USGENE are given according to WST.25 of the WIPO.

Predefined Display and Print Formats

Format	Content	Examples
ABS	AN, ED, UP, DED, 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, TI, IN, PA, PI, PIT, AI, RLPI, RLI, PRAI, XPD, XNTE, FS, MTY, PSL, DESC, SSO, ORGN, AB, CLM, SEQC, SEQN, SQL, SEQK, SEQ, AA or NA, FEAT	D ALL
IALL	ALL, indented with text labels	D L2 1-5 IALL
APPS	AI, RLI, PRAI	D APPS
BIB	AN, ED, UP, DED, TI, IN, PA, DT, PI, PIT, AI, RLPI, RLI, PRAI, FS, MTY, PSL, DESC (BIB is the default)	D BIB
IBIB	BIB, indented with text labels	D IBIB
BRIEF	ALL, but with MCLM only	D BRIEF
IBRIEF	BRIEF, indented with text labels	D IBRIEF
FASTA	FASTA format	D FASTA
SCAN	ED, UP, DED, TI, MTY, DESC (random display without answer numbers)	D SCAN
SQIDE	AN, ED, UP, DED, MTY, ORGN, SEQC, SEQN, SQL, SEQK, SEQ, AA or NA, FEAT	D SQIDE
SQ3IDE	AN, ED, UP, DED, MTY, ORGN, SEQC, SEQN, SQL, SEQK, SEQ3, AA or NA, FEAT	D SQ3IDE
TRIAL (TRI, SAM, SAMPLE, FREE)	AN, TI, MTY, DESC, 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
Calculated Expiration Date	XPD	Y	Y
Calculated Expiration Year	XPY	Y	Y
Data Entry Date	DED	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
Inventor Address	INA	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
Organism Name	ORGN	Y	Y
Patent Assignee	PA	Y	Y
Patent Assignee Address	PAA	Y	Y
Patent Country	PC	Y	Y
Patent Information Type	PIT	Y	Y
Patent Number	PN (PI)	Y	Y
Patent Number/Kind Code	PNK	Y	Y
Patent Number, Original	PNO	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
Related Application Country	RLC	Y	Y
Related Application Date	RLD	Y	Y
Related Application Number	RLN	Y	Y
Related Application Type	RLT	Y	Y
Related Application Year	RLY	Y	Y
Related Publication Country	RLPC	Y	Y
Related Publication Date	RLPD	Y	Y
Related Publication Number	RLPN	Y	Y

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

Field Name	Field Code	ANALYZE/ SELECT (1)	SORT
Related Publication Year	RLPY	Y	Y
Sequence Count	SEQC	Y	Y
Sequence Identity Number	SEQN	Y	Y
Sequence Key	SEQK	Y	Y
Sequence Length	SQL	Y	Y
Sequence Source	SSO	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 USGENE database for protein and nucleotide sequence data by similarity (homology). BLAST is provided in USGENE with the permission of the National Center for Biotechnology Information (NCBI) of the National Library of Medicine (NLM). GETSIM uses 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 must be specified:

- /SQP for searching peptide sequences
- /SQN for nucleotide sequences
- /TSQN for searching peptide sequences translated from USGENE 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 USGENE protein sequences
- /TSQNX for searching translated nucleotides form USGENE 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 L /TSQP

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

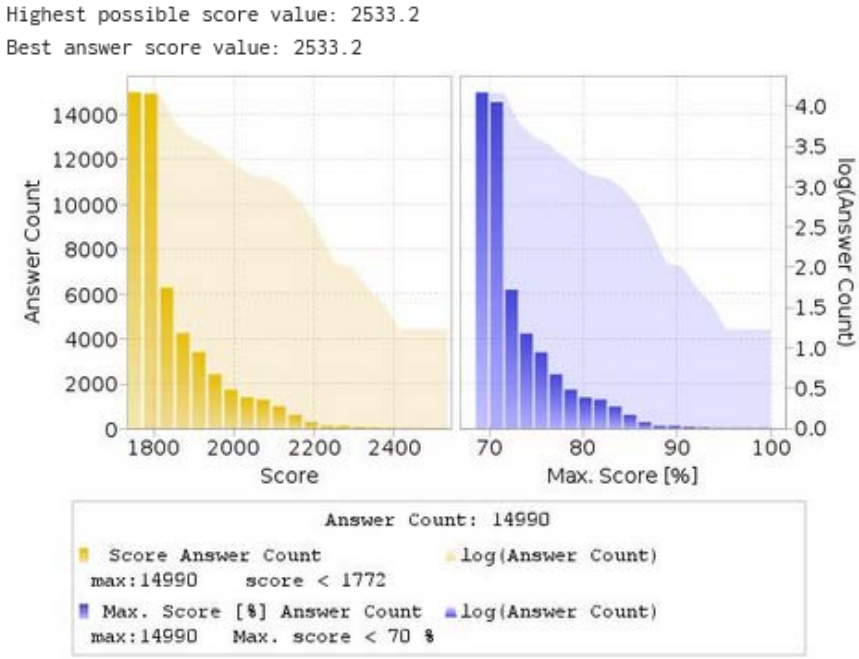
The maximum number of hits is by default 15,000 records. The parameter "-maxseq" allows to increase the maximum number of hits to 100,000 records, e.g., =>RUN BLAST L1/SQN -F F -MAXSEQ 100000.

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.

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.



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 ? :2400

L3    RUN STATEMENT CREATED
L3          17 ATGGGATGGAGCTGTATCATCCTCTTCTTGGTAGCAACAGCTACAGGTGT
```

2) Select a part of the answer set using the percentage score value from the right histogram, e.g., "90%" or "90% 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 ? :90% SCORE

L4    RUN STATEMENT CREATED
L4          101 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

L6    RUN STATEMENT CREATED
L6           78 ATGGGATGGAGCTGTATCATCCTCTTCTTGGTAGCAACAGCTACAGGTGT
```

4) Select a part of the answer set combining the percentage score and the percentage identity value, e.g., "90% 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 ? :90% SCORE 100% IDENT

L7    RUN STATEMENT CREATED
L7           17 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

L8    RUN STATEMENT CREATED
L8       14990 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 USGENE 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^{SM} file.

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

L8     ANSWER 1 OF 28 USGENE COPYRIGHT 2022 SEQUENCEBASE CORP on STN.

ALIGN

  Sequence Length: 43;


  Hits at: 1-11

     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 1 OF 28 USGENE COPYRIGHT 2022 SEQUENCEBASE CORP on STN.
SEQ
           MFTIRSRMCLH
           ===========

  Hits at: 1-11
```

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)+/SQSP CATA(CTG){1,}TATT/SQSN
^ (Caret)	search at the beginning or end of a sequence	^MCGIL/SQS VCDS^/SQSP
|	specifies alternate residues	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 IALL

L1 ANSWER 1 OF 1 USGENE COPYRIGHT 2022 SEQUENCEBASE CORP on STN.

ACCESSION NUMBER: **20210272652.4325** USGENE Full-text

ENTRY DATE: 20211224

UPDATE DATE: 20211224

DATA ENTRY DATE: 20210904

TITLE: Method of finding structural variants for identifying and differentiating species, strains and cells in normal and pathological conditions

INVENTOR(S): Huang Xiaoqiu, Ames, IA

PATENT APPLICANT(S): NO ASSIGNEE AT PUBLICATION

DOCUMENT TYPE: Patent

PATENT INFORMATION: **US 20210272652** A1

PATENT INFO. TYPE: USA1 FIRST PUBLISHED PATENT APPLICATION [FROM 2001 ONWARDS]

APPLICATION INFO.: **US 2020-16805783** 20200301

PRIORITY INFO.: **US 2020-16805783** 20200301

FILE SEGMENT: NUCLEIC; NS

MOLECULE TYPE: DNA

PAT. SEQ. LOC: SEQ ID NO 4325

DESCRIPTION: Aspergillus thermomutatus DNA; 236; sequence 4325 of 84193

SEQUENCE SOURCE: NUCLEIC; PSIPS; APPLICATION

ORGANISM: Aspergillus thermomutatus

ABSTRACT:

Large whole-genome datasets of short reads from species and strains in normal and pathological conditions are processed to find species-, strain- and condition-specific structural variants along with their estimated genome-wide copy numbers. These structural variants provide huge pools of genetic targets with molecular approaches to accurate & fast detection and identification of eukaryotic pathogens such as fungal pathogens and to precise diagnosis and accurate assessment of clinical conditions such as cancer, dementia, Parkinson's disease, Asperger's syndrome.

CLAIMS:

1. A method of finding structural variants for identifying and differentiating species, strains and cells in normal and pathological conditions, comprising:(a) a data storage element storing two or more whole-genome datasets of sequence reads with no genomic location information, where the datasets come from different species or strains, or from cells in normal and pathological conditions; and(b) a processing element associated with the storage element and configured to:i.

...

8. The method of claim 1, wherein the datasets come from human cells in normal and pathological conditions.

9. The method of claim 1, wherein the datasets come from animal cells in normal and pathological conditions.

SEQUENCE COUNT: 84193
SEQUENCE NUMBER: 73765
SEQUENCE LENGTH: 120
SEQUENCE KEY: cca62a890bb34eb456b746ba4090fd0e0ec0ba31ba4e463d849a8b780ba1bc0b

SEQUENCE:

```
   1 aacccaaaaa ggcataatta aactttactt cctctctttc ttcttcccac
  51 tcatcctaac cctactccta atcacataac ctattccccc gagcaatctc
 101 aattacaata tatacaccaa
```

NA

Code	Count	Percent
A	40	33.3
C	41	34.2
G	4	3.3
U	0	0.0
T	35	29.2
B	0	0.0
D	0	0.0
H	0	0.0
I	0	0.0
K	0	0.0
M	0	0.0
R	0	0.0
S	0	0.0
V	0	0.0
W	0	0.0
X	0	0.0
Others	0	0.0

FEATURE TABLE:

Key	Location
USGENE	|
PSIPS	|
misc_feature	|

DISPLAY TRIAL

L1 ANSWER 1 OF 380 USGENE COPYRIGHT 2022 SEQUENCEBASE CORP on STN.
TI STRAIN OF SERRATIA LIQUEFACIENS AND A METHOD OF PRODUCING HELIOTROPIN WITH THE SAME STRAIN
MTY DNA
DESC Artificial DNA; Synthesized; sequence 2 of 2

USGENE

DISPLAY SQIDE

```
L2      ANSWER 1 OF 177 USGENE COPYRIGHT 2022 SEQUENCEBASE CORP on STN.
AN      10920202.4   USGENE
        ED 20211224 UP 20211224 DED 20210218
MTY     protein
ORGN    Artificial Sequence
SEQC    12
SEQN    4
SEQK    d94a209eb0c6c088ad4d4a722bb5f9b6ea4d982dea1564a57cd1c153a2327395

SEQ
               1 mrfnnkmlal aallfaaqas adtlesidnc avgcptggss nvsivrhayt
              51 lnnnsttkfa nwvayhitkd tpasgktrnw ktdpalnpad tlapadytga
             101 naalkvdrgh qaplaslagv sdweslnyls nitpqksdln qgawadledq
             151 erklidradi ssvytvtgpl yerdmgklpg tqkahtipsa ywkvifinns
             201 pavnhyaafl fdqntpkgad fcqfrvtvde iekrtgliiw aglpddvqas
             251 lkskpgvlpe lmgckn

AA
        Code    Count  Percent
           A       33     12.4
           B        0      0.0
           C        4      1.5
           D       19      7.1
           E        8      3.0
           F        8      3.0
           G       16      6.0
           H        5      1.9
           I       12      4.5
           J        0      0.0
           K       16      6.0
           L       24      9.0
           M        4      1.5
           N       19      7.1
           O        0      0.0
           P       15      5.6
           Q        9      3.4
           R        9      3.4
           S       18      6.8
           T       19      7.1
           U        0      0.0
           V       14      5.3
           W        6      2.3
           Y        8      3.0
           Z        0      0.0
      Others        0      0.0
FEATURE TABLE:
Key        |Location|
===========+========+=======================================
USGENE     |1..266  |http://www.sequencebase.com/usgene.php?d
           |        |=10920202.4
other_info|1..266  |Description of Artificial Sequence Synth
           |        |etic polypeptide
           |        |[\\4501-00\3] 266
```

DISPLAY FASTA

```
L7      ANSWER 1 OF 627 USGENE COPYRIGHT 2022 SEQUENCEBASE CORP on STN.

FASTA

        >USGENE|20210371872.9413|protein|sequence 9413 from US20210371872
        mkmslvrplltssekmvasvlferlpvvipkidpivyafqefsfrwrqqyqrrypdefldrsdargkgdyqi
        eyvpapriteadktmieghckelstedstffsmvmlmglqvgslcgifqrnfmnqrkpcvsalslh
```

DISPLAY SEQ3

```
L7      ANSWER 1 OF 627 USGENE COPYRIGHT 2022 SEQUENCEBASE CORP on STN.

SEQ3

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