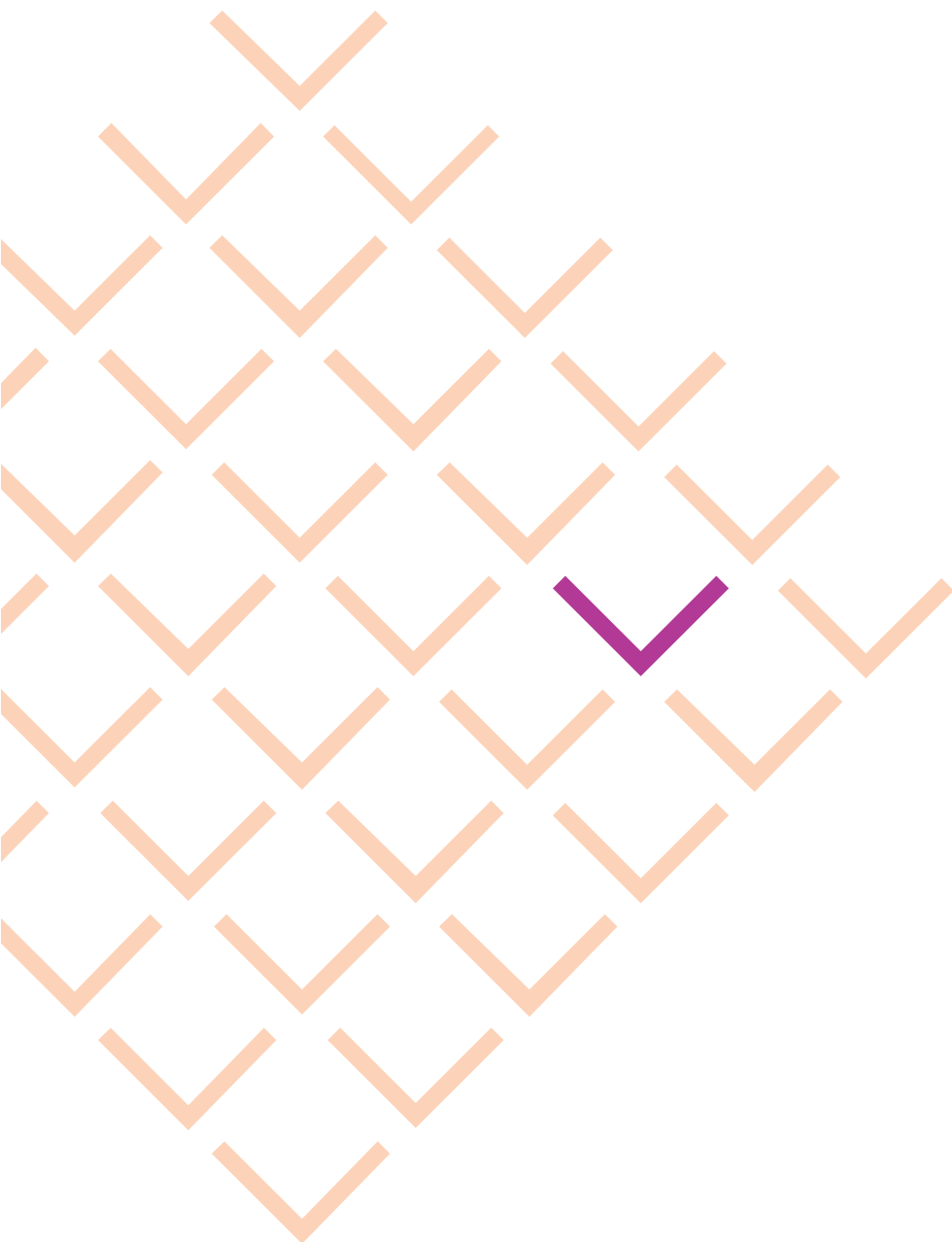




Protein Sequences in the CAS RegistrySM File on STN[®]– Exact and Pattern Searching

A Quick Reference Guide



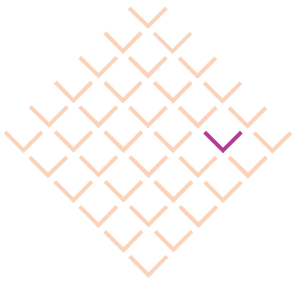
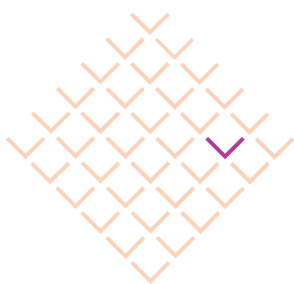


Table of Contents

Preface	2
Protein sequences in REGISTRY	3
Search options	4
Amino acid codes	5
Display options	7
Sample record	8
Searching for exact sequence strings	9
Searching subsequences	10
Searching for functionally similar sequences	11
Searching motifs and patterns	13
Gaps	14
Order of execution of symbols	15
Using SEQLINK	16
Searching length.....	17
Searching chemical annotation	18
Searching for references	19



Preface

This booklet is designed as a quick reference guide to exact and pattern searching of protein sequences in the CAS Registry File on STN.

CAS Registry BLAST® similarity searching is also available using STN® on the WebSM or STN Express® with *Discover!*TM For information, refer to the *CAS Registry BLAST Similarity Searching via STN on the Web* Quick Reference Guide available at:

www.cas.org/ONLINE/QRGUIDES/blast.pdf

The STN Express with *Discover!* User Guide is available from the STN Express Resources web page:

www.cas.org/ONLINE/STN/expresources.html

Protein sequences in REGISTRY

Protein sequence data may be searched and displayed in REGISTRY on STN. References to protein sequences may be searched and displayed in bibliographic files on STN, e.g., CAPlusSM.

Protein sequence information in REGISTRY is compiled by CAS from sequences reported in research articles and patents.

Sequences may be found in REGISTRY for the following classes of proteins and peptides from both journal and patent literature:

- Naturally occurring proteins and peptides
- Sequences deduced from gene translation and reported by the author
- Sequences deduced by gene translation from the GenBank[®] database (registered trademark of the U.S. Department of Health and Human Services)
- Chemically modified peptides and proteins
- Genetically engineered and synthetic proteins
- Multichain proteins
- Cyclic peptides
- Fusion proteins
- Peptide metal complexes
- Sequences containing uncommon amino acids, i.e., not genetically encoded
- Partial protein sequences

Search options

To search for sequence information in REGISTRY, enter the SEARCH (or S) command followed by the search string and a field code.

Search	Field Code	Example	Retrieves
Exact Sequence	/SQEP	S FCFWKTCT/SQEP	Exact match; same length
Subsequence	/SQSP	S LAGLL/SQSP	Sequences in which the query sequence may or may not be embedded
Exact Family	/SQEFP	S YGGFL/SQEFP	Functionally similar amino acids; same length
Subsequence Family	/SQSFP	S ATCXAWV/SQSFP	Functionally similar amino acids; may or may not be embedded
Sequence Length	/SQL	S SQL<=10	Sequences of a certain length
Annotation	/NTE	S MULTICHAIN/NTE	Sequences with the search term in the NTE field

Amino acid codes

These codes are used for displaying or searching protein sequences with chain lengths of four or more. Dipeptides and tripeptides are also included in REGISTRY, but may be searched only by name or structure and not by sequence representation.

For common amino acids, either one-letter or three-letter codes may be used. Three-letter codes are used for uncommon amino acids. When searching, enclose three-letter codes or strings of codes in single quotes. Use dashes to separate three-letter codes in strings.

Common amino acids

1-Letter Code	3-Letter Code	Name
A	Ala	Alanine
B	Asx	Aspartic acid or Asparagine
C	Cys	Cysteine
D	Asp	Aspartic acid
E	Glu	Glutamic acid
F	Phe	Phenylalanine
G	Gly	Glycine
H	His	Histidine
I	Ile	Isoleucine
K	Lys	Lysine
L	Leu	Leucine
M	Met	Methionine
N	Asn	Asparagine
P	Pro	Proline
Q	Gln	Glutamine
R	Arg	Arginine
S	Ser	Serine
T	Thr	Threonine
U	Scy	Selenocysteine
V	Val	Valine
W	Trp	Tryptophan
X	Xxx	Uncommon or Unspecified
Y	Tyr	Tyrosine
Z	Glx	Glutamic acid or Glutamine

Uncommon amino acids

3-Letter Code	Name	3-Letter Code	Name
Aaa	α -amino acid	Har	homoarginine
Aad	2-aminoadipic acid (2-aminohexanedioic acid)	Hcy	homocysteine
Aan	α -asparagine	Hhs	homohistidine
Abu	2-aminobutanoic acid	Hiv	2-hydroxyisovaleric acid
Aca	2-aminocaproic acid (2-aminodecanoic acid)	Hse	homoserine
Agn	α -glutamine	Hva	2-hydroxypentanoic acid
Aib	α -aminoisobutyric acid (α -methylalanine)	Hyl	5-hydroxylysine
Apm	2-aminopimelic acid (2-aminoheptanedioic acid)	Hyp	4-hydroxyproline
App	γ -amino- β -hydroxybenzenepentanoic acid	Inc	2-carboxyoctahydroindole
Asu	2-aminosuberic acid (2-aminooctanedioic acid)	Iqc	3-carboxyisoquinoline
Aze	2-carboxyazetidine	Iva	isovaline
Bal	β -alanine	Lac	2-hydroxypropanoic acid (lactic acid)
Bas	β -aspartic acid	Maa	mercaptoacetic acid
Bly	3,6-diaminohexanoic acid (β -lysine)	Mba	mercaptobutanoic acid
Bua	butanoic acid	Mhp	4-methyl-3-hydroxyproline
Bux	4-amino-3-hydroxybutanoic acid	Mpa	mercaptopropanoic acid
Cap	γ -amino- β -hydroxycyclohexanepentanoic acid	Nle	norleucine
Cit	N ⁵ -aminocarbonylornithine	Nty	nortyrosine
Cya	3-sulfoalanine	Nva	norvaline
Dab	2,4-diaminobutanoic acid	Oaa	ω -amino acid
Dpm	diaminopimelic acid	Orn	ornithine
Dpr	2,3-diaminopropanoic acid	Pen	penicillamine (3-mercaptoproline)
Dsu	2,7-diaminosuberic acid (2,7-diaminooctanedioic acid)	Phg	2-phenylglycine
Edc	S-ethylthiocysteine	Pip	2-carboxypiperidine
Ggu	γ -glutamic acid	Sar	sarcosine (N-methylglycine)
Gla	γ -carboxyglutamic acid	Spg	1-amino-1-carboxycyclopentane
Glc	hydroxyacetic acid (glycolic acid)	Sta	statin (4-amino-3-hydroxy-6-methylheptanoic acid)
Glp	pyroglutamic acid	Thi	3-thienylalanine
		Tml	ϵ -N-trimethyllysine
		Tza	3-thiazolylalanine
		Und	undefined
		Wil	α -amino-2,4-dioxypyrimidine-propanoic acid

Display options

To display answers in REGISTRY, enter the DISPLAY (or D) command followed by the L-number resulting from a search, answer numbers or a range of numbers, and the display fields or formats.

Display Fields

Code	Content
RN	CAS Registry Number
CN	Chemical Name
PNTE	Patent Annotation
FS	File Segment
SQL	Sequence Length
NTE	Sequence Annotation
SEQ	Sequence (1-letter codes)
SEQ3	Sequence (3-letter codes)
MF	Molecular Formula
CI	Substance Class Identifier
SR	Source of Registration
LC	CAS Registry Number Locator
DT.CA	CAplus Document Type
RL	CAplus Super Roles
RL.NP	CAplus Super Roles from Non-patents
RL.P	CAplus Super Roles from Patents

Some Display Formats

Format	Content
ALL	All available fields, including sequence data and the 10 most recent CA references
SQD	Sequence data, 1-letter codes
SQD3	Sequence data, 3-letter codes
SQIDE	Sequence data, CN, MF, SR, LC, DT.CA, RL, REF
HIT	All fields containing hit terms
KWIC	All hit terms plus 20 words on either side

Sample record

This record is displayed in the SQIDE format.

```
RN 653601-37-7 REGISTRY
CN L-Tyrosinamide, L-seryl-L-leucyl-L-arginyl-L-arginyl-
L-seryl-L-seryl-L-cysteinyl-N-methyl-L-
phenylalanylglycylglycyl-L-arginyl-L-methionyl-L-
a-aspartyl-L-arginyl-L-isoileucylglycyl-L-alanyl-L-
glutamyl-L-serylglycyl-L-leucylglycyl-L-cysteinyl-L-
asparagyl-L-seryl-L-phenylalanyl-L-arginyl-, cyclic
(7 23)-disulfide (9CI) (CA INDEX NAME)
FS PROTEIN SEQUENCE; STEREOSEARCH
SQL 28
NTE modified
```

```
-----
type          --- location ---      description
-----
terminal mod. Tyr-28          -      C-terminal amide
bridge        Cys-7          - Cys-23    disulfide bridge
modification  Phe-8          -      methyl<Me>
-----
```

```
SEQ      1 SLRRSSCFGG RMDRIGAQSG LGCNSFRY
```

```
**RELATED SEQUENCES AVAILABLE WITH SEQLINK**
```

```
MF C128 H206 N46 O38 S3
```

```
SR CA
```

```
LC STN Files: CA, CAPLUS
```

```
DT.CA CAplus document type: Patent
```

```
RL.P Roles from patents: BIOL (Biological study); PREP
(Preparation); USES (Uses)
```

```
1 REFERENCES IN FILE CA (1907 TO DATE)
```

```
1 REFERENCES IN FILE CAPLUS (1907 TO DATE)
```

Searching for exact sequence strings

To find substances that match the search query exactly and that are of the same length, search the sequence in the Exact Sequence Search (/SQEP) field. This option is most useful when you need to find analogs differing only in chemical modifications.

Find analogs of the drug Sandostatin with the sequence FCFWKTCCT.

```
=> FILE REGISTRY

=> S FCFWKTCCT/SQEP
L1          401 FCFWKTCCT/SQEP
           (FCFWKTCCT/SQEP AND SQL=8)

=> D L1 SQD 1-2

L1  ANSWER 1 OF 401  REGISTRY  COPYRIGHT 2004 ACS on STN
RN  708972-45-6  REGISTRY
FS  PROTEIN SEQUENCE; STEREOSEARCH
SQL 8
NTE modified (modifications unspecified)
-----
type          location          description
-----
bridge        Cys-2          - Cys-7        disulfide bridge
modification  Phe-1          -              undetermined
                                           modification
-----

SEQ          1 FCFWKTCCT
           =====
HITS AT:    1-8

**RELATED SEQUENCES AVAILABLE WITH SEQLINK**

L1  ANSWER 2 OF 401  REGISTRY  COPYRIGHT 2004 ACS on STN
RN  706790-70-7  REGISTRY
FS  PROTEIN SEQUENCE
SQL 8
NTE modified (modifications unspecified)
-----
type          location          description
-----
bridge        Cys-2          - Cys-7        disulfide bridge
modification  Phe-1          -              undetermined
                                           modification
modification  Thr-8          -              undetermined
                                           modification
-----

SEQ          1 FCFWKTCCT
           =====
HITS AT:    1-8

**RELATED SEQUENCES AVAILABLE WITH SEQLINK**
```

Enter REGISTRY.

Enter S (SEARCH) and the exact sequence in the /SQEP search field. You can use 1-letter codes for common amino acids.

An L-number answer set (L1) is created. The number of sequences retrieved (401) is displayed.

To display sequence data, enter D (or DISPLAY), the L-number, the format, and the answer numbers. The SQD format includes the CAS Registry Number and sequence data using 1-letter codes.

The answers have the same sequence and length, but they differ in chemical annotation in the NTE field.

Searching subsequences

To find answers matching the search sequence plus sequences in which the query sequence may be embedded, search in the Subsequence Search (/SQSP) field.

Find proteins containing the sequence string GLFGRKTGQAP from the human cytochrome c.

```
=> FILE REGISTRY

=> S GLFGRKTGQAP/SQSP
L1          131 GLFGRKTGQAP/SQSP

=> D CN SQL SEQ 1, 14

L1 ANSWER 1 OF 131 REGISTRY COPYRIGHT 2004 ACS on STN
CN GenBank CAE99232 (9CI) (CA INDEX NAME)
OTHER NAMES:
CN GenBank CAE99232 (Translated from: GenBank AX885118)
SQL 98

SEQ 1 MGDVEKGKKI FIMKCSQCHT VEKGGKHKTG PNLHGLFGRK TGQAPGYSYT
          =====
          51 AANKNKGIIW GEDTLMEYLE NPKKYIPGTK MIFVGIKKKE ERADLIAY
HITS AT: 35-45

**RELATED SEQUENCES AVAILABLE WITH SEQLINK**

L1 ANSWER 14 OF 131 REGISTRY COPYRIGHT 2004 ACS on STN
CN GenBank AAA41016 (9CI) (CA INDEX NAME)
OTHER NAMES:
CN GenBank AAA41016 (Translated from: GenBank M20623)
SQL 105

SEQ 1 MGDAEAGKKI FIQKCAQCHT VEKGGKHKTG PNLWGLFGRK TGQAPGFSYT
          =====
          51 DANKNKGVIW TEETLMEYLE NPKKYIPGTK MIFAGIKKKS EREDLIQYLK
          101 EATSS
HITS AT: 35-45

**RELATED SEQUENCES AVAILABLE WITH SEQLINK**
```

Enter REGISTRY.

Search the subsequence in the /SQSP field. You can use 1-letter codes for common amino acids.

Display chemical names (CN), sequence length (SQL), and sequences using 1-letter codes (SEQ).

Notice the different chemical names and variable sequence length. The query subsequence is highlighted.

Searching for functionally similar sequences

To search for functionally similar sequences, use the “family” search options: Family Exact Sequence Search (/SQEFP) and Family Subsequence Search (/SQSFP). In family searches, each common amino acid in the query has to match either the exact amino acid or a functionally similar “equivalent,” as shown in the following table.

Property	Functionally Similar Amino Acids
Neutral-Weakly Hydrophobic	Ala,Gly,Pro,Ser,Thr (A, G, P, S, T)
Hydrophilic-Acid Amine	Asn,Asp,Gln,Glu (N, D, Q, E)
Hydrophilic-Basic	Arg,His,Lys (R, H, K)
Hydrophobic	Ile,Met,Leu,Val (I, M, L, V)
Hydrophobic-Aromatic	Phe,Trp,Tyr (F, W, Y)
Cross-linking	Cys (C)

Find sequences that are functionally similar to the sequence of synthetic somatostatin (AGCKNFFWKFTTSC).

=> **FILE REGISTRY**

=> **S AGCKNFFWKFTTSC/SQEFP**

L1 300 AGCKNFFWKFTTSC/SQEFP

=> **D CN HIT NTE 1,4**

L1 ANSWER 1 OF 300 REGISTRY COPYRIGHT 2004 ACS on STN
 CN L-Cysteine, glycyglycyl-L-cysteiny-L-lysyl-L-asparaginy-L-phenylalanyl-L-phenylalanyl-L-tryptophyl-L-lysyl-L-threonyl-L-phenylalanyl-L-threonyl-L-seryl- (9CI) (CA INDEX NAME)

OTHER NAMES:

CN 1733: PN: W003060071 SEQID: 1757 unclaimed protein

FS PROTEIN SEQUENCE; STEREOSEARCH

SQL 14

SEQ 1 GGCKNFFWK FTSC
 =====

HITS AT: 1-14

L1 ANSWER 4 OF 300 REGISTRY COPYRIGHT 2004 ACS on STN
 CN L-Cysteinamide, L-alanylglycyl-L-cysteiny-L-lysyl-L-asparaginy-L-phenylalanyl-L-phenylalanyl-L-tryptophyl-L-lysyl-L-alanyl-L-phenylalanyl-L-threonyl-L-seryl-, cyclic (3→14)-disulfide (9CI) (CA INDEX NAME)

RN 556796-24-8 REGISTRY

FS PROTEIN SEQUENCE; STEREOSEARCH

SQL 14

SEQ 1 AGCKNFFWKA FTSC
 =====

HITS AT: 1-14

****RELATED SEQUENCES AVAILABLE WITH SEQLINK****

NTE modified

type	location	description
terminal mod.	Cys-14	- C-terminal amide
bridge	Cys-3	- Cys-14 disulfide bridge

Enter REGISTRY.

Search the sequence of somatostatin in the /SQEFP field.

Display the names (CN), fields in which hit terms occur (HIT), and chemical annotation (NTE).

The sequence length of answers is the same as the length of the query sequence.

Searching motifs and patterns

Complex pattern searching of protein sequences is possible in the /SQSP and /SQSFP subsequence search fields using the Boolean operators (AND, OR, NOT) as well as special characters and symbols.

Symbol(s)	Function	Example	Retrieves
^	Require the string at the beginning or the end of the sequence	^MCGIL/SQSP VCDS~/SQSFP	MCGIL at the beginning VCDS at the end
[]	Specify alternate residues	LGP[VL]/SQSP	LGP followed by either V or L
[-] [~]	Exclude a residue or alternate residues	PTGK[-H]EA/SQSP	PTGKDEA, PTGKNEA, etc.
{ } with a number or range	Repeat the preceding string or residue	GG(FL){1-3}/SQSP	GGFL, GGFLFL, or GGFLFLFL
?	Repeat the preceding string or residue zero or one time	FLRRI(RP)?K/SQSP	FLRRIK or FLRRIRPK
*	Repeat the preceding string or residue zero or more times	KLK(WD)*N/SQSP	KLKN, KLKWDN, KLKWDWDN, KLKWDWDWDN, etc.
+	Repeat the preceding string or residue one or more times	AQP+/SQSP (AQP)+/SQSP	AQP, AQPP, AQPPP, etc. AQP, AQPAQP, AQAQPAQP, etc.
	Specify alternate sequences	S ACD KLM/SQSP	ACD or KLM
&	Join together sequence queries	S L1&L3/SQSFP	Sequence L1 joined to sequence L3

Gaps

To specify a gap, use a period (.) for one residue, a colon (:) for zero or one residue, or a period (.) followed by an appropriate repeat expression.

Symbol(s)	Function	Example	Retrieves
.	A gap of one residue	SY.RPG/SQSP	SY followed by one residue followed by RPG
{m} or [m.]	A gap of m residues	SY.{2}RPG/SQSP	SY followed by any two residues followed by RPG
{m,u} or {m-u}	A gap of m to u residues	GFF.{2,10}LSS/SQSP	GFF followed by a gap of 2-10 residues followed by LSS
.? or : or {0,1} or {0-1}	A gap of zero or one residue	AGA.?SRI/SQSFP	AGA followed by zero or one residue followed by SRI
* or {0,} or {0-}	A gap of zero or more residues	HLC.*TYG/SQSP	HLC followed by a gap of zero or more residues followed by TYG
+ or {1,} or {1-}	A gap of one or more residues	SY.+TH/SQSP	SY followed by any number of residues followed by TH

Order of execution of symbols

More than one symbol may be used to create complex sequence queries. If you do not use parentheses in sequence queries, the operations are performed in the following order:

1. Repeat symbols ? or * or +
2. Repeat expressions using curly braces, e.g., {3,6}
3. Concatenation symbol &
4. The vertical bar |

Find atriopeptin analogs containing RSSCF and QSGLG, separated by a gap of zero or any number of amino acids.

```
=> FILE REGISTRY
```

```
=> S RSSCF.*QSGLG/SQSP
```

```
L3          473 RSSCF.*QSGLG/SQSP
```

```
=> D KWIC 2, 5
```

```
L3  ANSWER 2 OF 473  REGISTRY  COPYRIGHT 2004 ACS on STN
```

```
SEQ  101 RRSSCFGGRM DRIGAQSGLG CNSFRY
```

```
=====
```

```
HITS AT: 102-120
```

```
L3  ANSWER 5 OF 473  REGISTRY  COPYRIGHT 2004 ACS on STN
```

```
RN  613263-35-7  REGISTRY
```

```
SEQ   1 SLRRSSCFGG RMDRIGKQSG LGCNSFRY
```

```
=====
```

```
HITS AT: 4-22
```

```
**RELATED SEQUENCES AVAILABLE WITH SEQLINK**
```

Enter REGISTRY.

Search the sequence pattern in the /SQSP field. The symbol .* indicates a gap of any number of amino acids, including zero.

Use the KWIC format to display the hit subsequence in context.

Using SEQLINK

The SEQLINK EXACT command is used to locate additional protein or nucleic acid sequences that have the same sequence structure as that of a protein or nucleic acid sequence that has already been retrieved from REGISTRY. SEQLINK is especially useful after searching with a name, name segments, or CAS Registry Numbers®.

```
=> FILE REGISTRY

=> S 487486-61-3
L1      1 487486-61-3
          (487486-61-3/RN)

=> D SQIDE
L1 ANSWER 1 OF 1 REGISTRY COPYRIGHT 2004 ACS on STN
RN 487486-61-3 REGISTRY
CN GenBank CAA00061 (9CI) (CA INDEX NAME)
OTHER NAMES:
CN GenBank CAA00061 (Translated from: GenBank A00380)
FS PROTEIN SEQUENCE
SQL 28

SEQ      1 SLRRSSCFGG RMDRIGAQSG LGCNSFRY

**RELATED SEQUENCES AVAILABLE WITH SEQLINK**
MF Unspecified
CI MAN
SR GenBank

=> SEQLINK EXACT L1
L2      56 SEQLINK EXACT L1

=> D SQIDE 1
L2 ANSWER 1 OF 56 REGISTRY COPYRIGHT 2004 ACS on STN
RN 653601-38-8 REGISTRY
CN L-Tyrosinamide, N-[[[2-[[[3-(2,5-dihydro-2,5-dioxo-1H-
      :
      :
FS PROTEIN SEQUENCE; STEREOSEARCH
SQL 28
NTE modified (modifications unspecified)
-----
type          --- location ---          description
-----
bridge        Cys-7          - Cys-23          disulfide bridge
-----

SEQ      1 SLRRSSCFGG RMDRIGAQSG LGCNSFRY

**RELATED SEQUENCES AVAILABLE WITH SEQLINK**
MF C141 H222 N48 O44 S3
SR CA
LC STN Files: CA, CAPLUS
DT.CA Cplus document type: Patent
RL.P Roles from patents: BIOL (Biological study); PREP
      (Preparation); USES (Uses)
          1 REFERENCES IN FILE CA (1907 TO DATE)
          1 REFERENCES IN FILE CAPLUS (1907 TO DATE)
```

Enter REGISTRY.

Conduct a search.

Enter SEQLINK EXACT L1. L2 contains the CAS Registry Number from L1, plus 55 additional CAS Registry Numbers that have the same sequence as the sequence in L1.

Searching length

You can refine a sequence search by combining it with a search of sequence length in the /SQL field. You can use the following operators to search sequence lengths.

Operator	Definition	Example
>	Greater than	SQL>100
<	Less than	SQL<25
=	Equal to	SQL=15 or 15/SQL
<=	Less than or equal to	SQL<=100
>=	Greater than or equal to	SQL=>120
m-n	Range beginning with m and ending with n	35-100/SQL

Find RGDF containing peptides with 10 or fewer amino acids.

```
=> FILE REGISTRY
=> S RGDF/SQSP
L1          6872 RGDF/SQSP
=> S L1 AND SQL<=10
L2          733 L1 AND SQL<=10
=> D KWIC 1, 11
L2 ANSWER 1 OF 733 REGISTRY COPYRIGHT 2004 ACS on STN
SQL 4
SEQ        1 RGDF
          ====
HITS AT:   1-4
**RELATED SEQUENCES AVAILABLE WITH SEQLINK**
L2 ANSWER 11 OF 733 REGISTRY COPYRIGHT 2004 ACS on
STN
SQL 10,5,5
SEQ        1 RGDFQ
          ====
HITS AT:   1-4
SEQ        1 RGDFQ
          ====
HITS AT:   1-4
**RELATED SEQUENCES AVAILABLE WITH SEQLINK**
```

Enter REGISTRY and search the sequence.

Search SQL<=10 to retrieve only sequences with 10 or fewer residues.

The KWIC format displays the search terms in context of the fields in which they occur.

Searching chemical annotation

In the Annotation (/NTE) field, you can search the following types of information for chemically modified sequences:

- Terms for broad classification of the entire protein sequence, e.g., multichain, linear, or cyclic
- Terms for the type of chemical modification, e.g., uncommon amino acid or bridge
- Location of the amino acid where the chemical modification has occurred
- Terms describing the chemical modification made, e.g., the name of a blocking group, metal complex, or a bridge

In the /NTE field, you can search phrases or single words and combine them by the Boolean operators (AND, OR, NOT). When you enter terms with punctuation, the phrase is searched. When you enter terms separated by spaces, terms are searched in the same modification, in any order, and any number of words apart. You can use both right and left truncation. A term with left truncation must contain at least four characters, e.g., => S ?CHLOR?/NTE.

Find multichain sequences.

```
=> FILE REGISTRY

=> S MULTICHAIN/NTE
L3      14446 MULTICHAIN/NTE

=> D KWIC

L3      ANSWER 1 OF 14446 REGISTRY COPYRIGHT 2004 ACS on
STN
NTE     multichain
        modified (modifications unspecified)
```

Find sequences with the blocking group ethoxycarbonyl, also known as Eoc.

```
=> FILE REGISTRY

=> S EOC/NTE
L4      192 EOC/NTE

=> D KWIC

L4      ANSWER 1 OF 192 REGISTRY COPYRIGHT 2004 ACS on STN
NTE     modified (modifications unspecified)

-----
type           ----- location -----      description
-----
modification   Thr-1           -           ethoxycarbonyl
                                     <Eoc>
-----
```

Searching for references

To find references to protein sequences obtained in REGISTRY, use the resulting L-number as a search term in STN databases containing CAS Registry Numbers, e.g., CAPLUS, USPATFULL.

Find patents on peptides containing RGDF.

=> FILE REGISTRY

=> S RGDF/SQSP

L1 7950 RGDF/SQSP

=> FILE CAPLUS

=> S L1 AND PATENT/DT

L2 1170 L2 AND PATENT/DT

=> D 7 BIB AB

L2 ANSWER 7 OF 1170 CAPLUS COPYRIGHT 2004 ACS on STN

AN 2004:60714 CAPLUS Full-text

TI Gene expression profiles for marker genes of intestinal-type gastric tumors and uses in cancer diagnosis

IN Nakamura, Yusuke; Furukawa, Yoichi

PA Oncotherapy Science, Inc., Japan; Japan as Represented by the President of the University of Tokyo

SO PCT Int. Appl., 63 pp.

CODEN: PIXXD2

DT Patent

LA English

FAN.CNT 1

	PATENT NO.	KIND	DATE	APPLICATION NO.	DATE
PI	WO 2004007770	A2	20040122	WO 2003-JP8651	20030708

AB The present invention provides sensitive, specific and convenient diagnostic methods that correlate the expression of marker genes for distinguishing between benign and malignant lesions of intestinal-type gastric cancers and for identifying the presence or absence of lymph-node metastasis (i.e., identifying the metastatic phenotype). In one embodiment, the diagnostic method involves the scoring of gene expression profiles that discriminate between lymph node pos. tumors and lymph node neg. tumors. The predictive score calculated acts as diagnostic indicator that can objectively indicate whether a sample tissue has the metastatic phenotype. The present invention further provides methods of diagnosing intestinal-type gastric cancer in a subject, methods of screening for therapeutic agents useful in the treatment of intestinal-type gastric cancer, methods of treating intestinal-type gastric cancer and method of vaccinating a subject against intestinal-type gastric cancer.

Enter REGISTRY.

Search the subsequence in the /SQSP field.

An L-number (L1) is assigned to the REGISTRY answer set.

Enter CAPLUS.

Combine the search of L1 with PATENT/DT to restrict answers to patent documents citing the subsequences retrieved in REGISTRY.

To display bibliographic information and the abstract, enter BIB AB as the display format.

Find U.S. patents on the peptides containing the residues RGDF.

=> FILE REGISTRY

=> S RGDF/SQSP

L1 7950 RGDF/SQSP

=> FILE USPATFULL

=> S L1

L2 1076 L1

=> D BIB AB 1

L2 ANSWER 1 OF 1076 USPATFULL on STN
AN 2004:217835 USPATFULL Full-text
TI Gene disruption methodologies for drug target
discovery
IN Roemer, Terry, Montreal, CANADA
Jiang, BO, Montreal, CANADA
Boone, Charles, Toronto, CANADA
Bussey, Howard, Westmount, CANADA
PA Elitra Pharmaceuticals Inc., San Diego, CA, United
States (U.S. corporation)
PI US 6783985 B1 20040831
AI US 2001-792024 20010220 (9)
PRAI US 2000-183534P 20000218 (60)
DT Utility
FS GRANTED
EXNAM Primary Examiner: Ketter, James; Assistant Examiner:
Lambertson, David
LREP Day, Jones
CLMN Number of Claims: 29
ECL Exemplary Claim: 1
DRWN 10 Drawing Figure(s); 7 Drawing Page(s)
LN.CNT 12866

CAS INDEXING IS AVAILABLE FOR THIS PATENT.

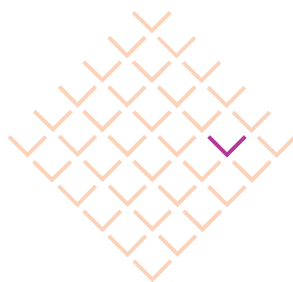
AB The present invention provides methods and compositions that enable the experimental determination as to whether any gene in the genome of a diploid pathogenic organism is essential, and whether it is required for virulence or pathogenicity. The methods involve the construction of genetic mutants in which one allele of a specific gene is inactivated while the other allele of the gene is placed under conditional expression. The identification of essential genes and those genes critical to the development of virulent infections, provides a basis for the development of screens for new drugs against such pathogenic organisms. The present invention further provides *Candida albicans* genes that are demonstrated to be essential and are potential targets for drug screening. The nucleotide sequence of the target genes can be used for various drug discovery purposes, such as expression of the recombinant protein, hybridization assay and construction of nucleic acid arrays. The uses of proteins encoded by the essential genes, and genetically engineered cells comprising modified alleles of essential genes in various screening methods are also encompassed by the invention.

Enter REGISTRY and search the sequence.

Enter USPATFULL for access to full text of U.S. patents. CAS Registry Numbers are available for chemical patents.

Search the REGISTRY sequence search L-number (L1) in USPATFULL to find U.S. patents citing the retrieved sequences.

Enter the DISPLAY command to view the entire text or portions of the patents. To display only the bibliographic information and the abstract, enter BIB AB as the format.



*CAS is a division of the
American Chemical Society®*

CAS2052-1104
November 2004

CAS Customer Care
Phone: 800-753-4227 (North America)
614-447-3700 (worldwide)
Fax: 614-447-3751
E-mail: help@cas.org
Internet: www.cas.org/supp.html