Sequence Motif Searches in CAS REGISTRY℠

STN® THE CHOICE OF PATENT EXPERTS™

FIZ Karlsruhe Leibniz Institute for Information Infrastructure

CAS A division of the American Chemical Society

Agenda

- Overview
- Sequence Code Match (SCM)
- Exact, subsequence and family sequence searches
- Sequences containing uncommon amino acids

STN 3

Search for related sequences

- To find variations of nucleic acids and proteins, including mutations, for diagnosis and treatment of diseases
- To classify evolutionary relationships
- To identify sequence motifs that indicate important structures or functions
- To assess the novelty or non-obviousness of a claimed nucleic acid or protein

STN 4

STN Sequence Databases

- CAS REGISTRY℠
  - Produced by Chemical Abstracts Service
- DGENE (GENESEQ™)
  - Produced by Thomson Reuters
- PCTGEN
  - Produced by FIZ Karlsruhe
- USGENE®
  - Produced by SequenceBase Corporation

STN 5

REGISTRY has extensive sequence coverage

- REGISTRY contains >66M sequences
  - > 57.8M nucleic acid sequences
  - > 8.8M amino acid sequences
- Sequences from non-patent literature, patents, GenBank, registrations via CAS Client Services and USAN Council
- Records include sequences, annotation tables, patent information tables, sequence length, nucleic acid residue counts, links to exact match peptide sequences

STN 6

Strategies for finding sequence information

- Chemical name search
- Structure search
- Similarity search
- Sequence Code Match search

STN 7

Search chemical names in REGISTRY to find sequence information

- Search generic or trade names in the Chemical Name (CN)
- Search name fragments in the Chemical Name Segment (CNS) field
- Display Sequence Identification information

  => S OXYTOCIN/CN; D SQIDE

  => S ?INTERLEUKIN?/CNS; D SQIDE
- Best for finding substances that have unambiguous names

STN 8

CNS searches retrieve amino acid and nucleic acid sequence records

```
=> FILE REGISTRY
=> S ?INTERLEUKIN?/CNS; D SQIDE
L1        41857 ?INTERLEUKIN?/CNS

RN   1421082-51-0  REGISTRY
CN   DNA (human interleukin-12 subunit P40-specifying SEQ ID 22)
     (CA INDEX NAME)
OTHER NAMES:
CN   20: PN: CN102876684 SEQID: 22 claimed DNA
FS   NUCLEIC ACID SEQUENCE
SQL  978
NA   225 a   250 c   298 g   205 t
PATENT ANNOTATIONS (PNTE):
Sequence |Patent
Source   |Reference
=========+=============
Not Given|CN102876684
         |claimed SEQID
         |22
SEQ      1 atggagaccg ataccctgct gctgtgggtg ctgctgctgt gggtgcccgg
        51 atctaccgga atctgggagc tgaagaagga tgtgtatgtg gtggagctgg
       101 attggtaccc tgatgctcca ggcgagatgg tggtgctgac ctgtgacaca
• • •
```

Chemical Names can include patent numbers.
FS = File Segment
NA = Nucleic Acid Counts

STN 9

Chemical Names can also include sequence information

```
=> S (LYS? (XW) PRO? (XW) ALA?)/CNS
L1   166087 (LYS? (XW) PRO? (XW) ALA?)/CNS

=> S L1 AND SQL<=7
L2     3008  L1 AND SQL<=7

=> D SCAN
IN   L-Proline, L-asparaginyl-L-lysyl-L-alanyl-L-leucyl-L-prolyl-L-
     alanyl-
SQL  7
• • •

IN   L-.alpha.-Glutamine, N2,N6-bis(L-seryl-L-asparaginyl-L-threonyl-
     L-seryl-L-.alpha.-glutamyl-L-seryl-L-phenylalanyl)-L-lysyl-L-
     phenylalanyl-L-arginyl-L-valyl-L-threonyl-L-glutaminyl-L-leucyl-
     L-alanyl-L-prolyl-L-lysyl-L-alanyl-L-glutaminyl-L-isoleucyl-L-
     lysyl-
SQL  29,22,7

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

(XW) finds terms in the specified order, with any number of intervening terms. Sequence Length (SQL) is numerically searchable.

STN 10

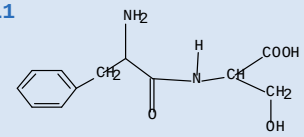
Search structures in REGISTRY to find complete sequence information

- Structure search retrieves chemical modifications to the primary structure
- Small peptides and nucleic acid oligomers are captured in REGISTRY as chemical structures, sequences, or both
- Structures can be drawn with up to 252 non-hydrogen atoms
  - Maximum of ~30 amino acids
  - Maximum of 8 nucleotides

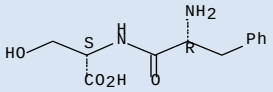
STN 11

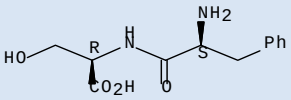
Structure search for Phe-Ser dipeptides

```
=> FILE REGISTRY
=>
Uploading . . .
=> D L1
```



```
=> S L1 FAM SAM; S L1 FAM FULL; D SCAN
L3      8  SEA FAM FUL L1

IN    L-Serine, D-phenylalanyl-
MF    C12 H16 N2 O4

Absolute stereochemistry.


IN    D-Serine, N-L-phenylalanyl- (9CI)
MF    C12 H16 N2 O4

Absolute stereochemistry.
```



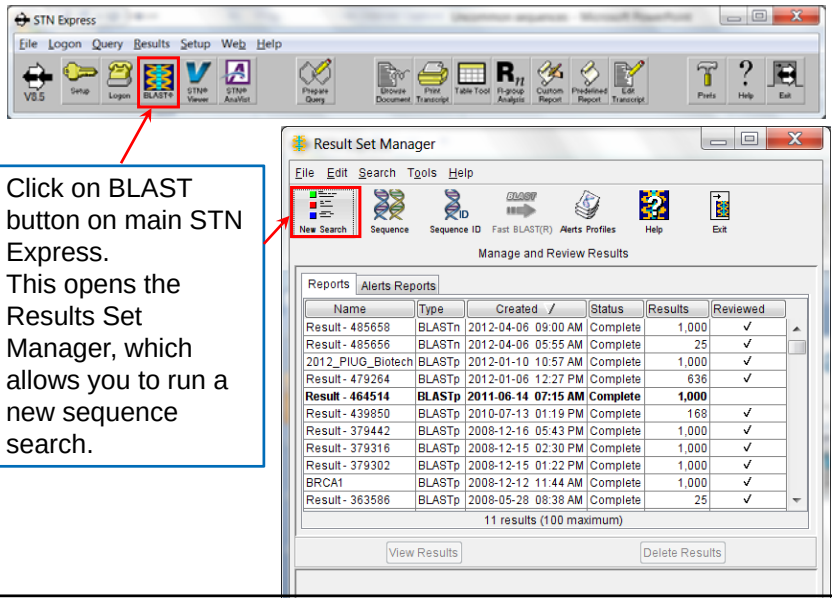


STN 12

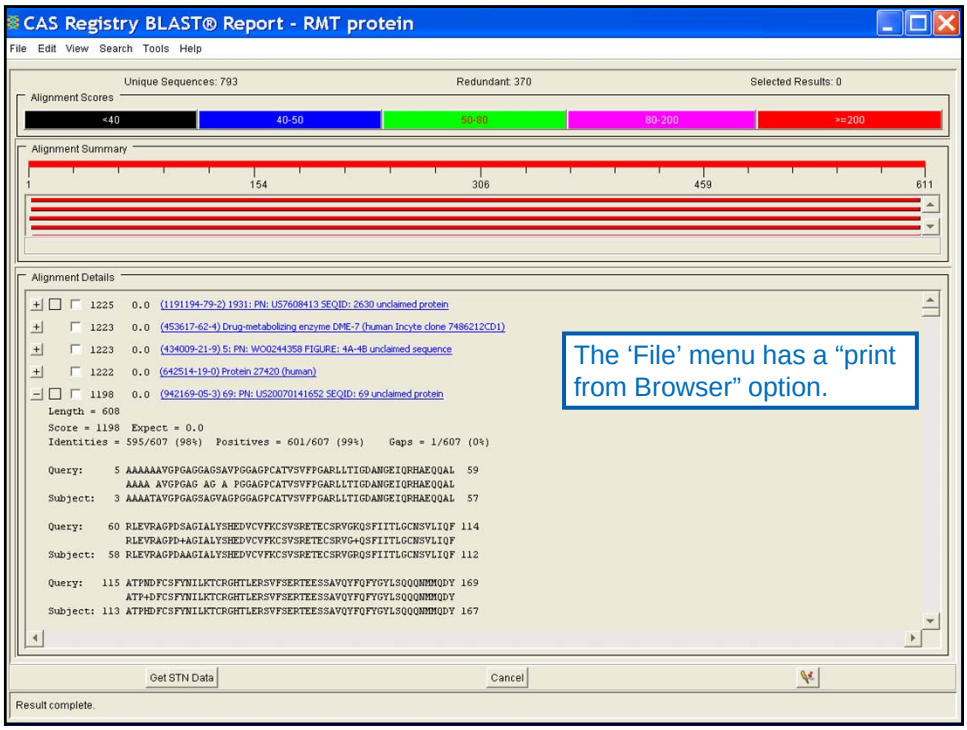
Conduct a BLAST search in REGISTRY

- Basic Local Alignment Search Tool (BLAST) finds regions of local similarity between sequences
- Used to infer functional and evolutionary relationships between sequences and to identify members of gene families
- Compares query algorithmically with a database of sequences using a similarity scoring matrix, code identity measurement, gap penalty and e-value

BLAST search in REGISTRY



1. Click on BLAST button on main STN Express.
2. This opens the Results Set Manager, which allows you to run a new sequence search.



The 'File' menu has a "print from Browser" option.

STN 15

Majority of the proteins contain common amino acids

- 22 common amino acids can be searched with 1-letter or 3-letter codes

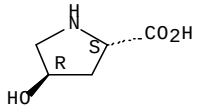
Amino acid	3 letter designation	1 letter designation
Pyrrolysine	Pyl	O
Selenocysteine	Scy	U

Pyrrolysine was added to REGISTRY in 2006.
Selenocysteine is covered through all time periods.

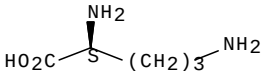
STN 16

Some proteins contain uncommon amino acids

- Formed or modified after translation
  - Example: 4-hydroxyproline (Hyp)



- Formed as intermediates in metabolic pathways
  - Example: Ornithine (Orn)



STN 17

CAS registers uncommon residues with three-letter codes

For a full list of the uncommon residue three-letter codes, see the Quick Reference Guide.

Codes	Uncommon residue
Aaa	Alpha-amino acid
Aad	2-aminoadipic acid (2-aminohexanedioic acid)
Aan	alpha-asparagine
Abu	2-aminobutanoic acid
Aca	2-aminocapric acid (2-aminodecanoic acid)
Agn	alpha-glutamine
Aib	alpha-aminoisobutyric acid
Apm	2-aminopimelic acid (2-aminoheptanedioic acid)
App	gamma-amino-beta-hydroxybenzenepentanoic acid2-aminosuberic acid (2-aminooctanedioic acid)
App	gamma-amino-beta-hydroxybenzenepentanoic acid2-aminosuberic acid (2-aminooctanedioic acid)

STN 18

Amino acid sequence displays can include ambiguity codes

Represents	3-letter designation	1-letter designation
Aspartic acid or Asparagine	Asx	B
Glutamic acid or Glutamine	Glx	Z
Isoleucine or Leucine	Xle	J
CAS uncommon or unspecified	Xxx	X

Records that include an ambiguity code indicate that the author or inventor did not delineate exactly which residue was in that position.

STN 19

Sequence Code Match (SCM)

- Search sequences using 1-letter or 3-letter codes
  - Exact amino acids, e.g., A or 'Ala'
  - Variable amino acids, e.g., B or 'Asx'
  - Any uncommon amino acid, e.g., X or 'Xxx'
  - Specific uncommon amino acids, e.g., 'Hse'
  - Exact nucleic acid, e.g., C
  - Ambiguous nucleic acids, e.g., U (for T or U)
  - Any nucleic acid, e.g., N
  - Unavailable nucleic acid, e.g., X

STN 20

HELP messages in REGISTRY

- HELP AAC - 1- and 3-letter codes for common amino acids
- HELP AAU - 3-letter codes for uncommon amino acids
- HELP BLAST - BLAST similarity searching
- HELP NUC - Codes for nucleic acids
- HELP SSQ - Sequence search methods and fields
- HELP SEQLINK - Sequence linking with SEQLINK
- HELP SQQ - Variability symbols in subsequence queries

STN 21

Sequence Code Match options

Search Type	Proteins	Nucleic Acids
Exact	/SQEP	/SQEN
Exact Family	/SQEFP	Not Applicable
Subsequence	/SQSP	/SQSN
Subsequence Family	/SQSFP	Not Applicable

STN 22

Functionally similar amino acids are searched in a family search

GROUP	AMINO ACIDS
Neutral-Weak Hydrophobic	P, A, G, S, T
Acid Amines-Hydrophilic	Q, N, E, D, B, Z
Basic-Hydrophilic	H, K, R
Hydrophobic	I, M, L, V
Aromatic	F, W, Y
Cross-Linking	C

STN 23

SCM searching symbols allow flexibility

- Specify motif patterns that consist of different amino acids at one or more locations in the sequence
- Specify amino acid gaps
- Search for sequence patterns at either the beginning or the end of a sequence
- Specify the number or range of repeats for amino acids or gaps

Gaps, wildcards, and brackets are invalid for "exact" sequence field codes

STN 24

Motif searching symbols

Symbols	Functions	Examples	Possible answers
^	Search at the beginning or the end of a sequence	=> S ^MCGIL/SQSP => S VCDS^/SQSFP	"MCGIL……“ "……VCDS"
[ ]	Specify alternate residues	=> S LGP[VL]/SQSP	LGPV LGPL
[-] or [~]	Exclude one or more residues	=> S PTGK[-H]/SQSP => S PTGK[~H]/SQSP	PTGKACCD
{#,#} {# - #} {#}	Repeat preceding residue(s)	=> S GG(FL){1,3}/SQSP => S GG(FL){1-3}/SQSP => S GG(FL){3}/SQSP	GGFL GGFLFL GGFLFLFL
.	Specify gap(s) in the sequence	=> S SY.RPG/SQSP => S SY…RPG/SQSP	SYARPG SYAAARPG
|	Specify alternate residues	=> S ACD|KLM/SQSP	ACD KLM

STN 25

Motif searching symbols (cont.)

Symbols	Functions	Examples	Possible answers
?	Repeat residue(s) zero or one time	=> S FLRR(RP)?K/SQSP	FLRRK FLRRRPK
*	Repeat residue(s) zero or more times	=> S KLK(WD)*N/SQSP	KLKN KLKWDN KLKWDWDN KLKWDWDWDN
+	Repeat residue(s) one or more times	=> S AQP+/SQSP	AQPP AQPPP AQPPPP AQPPPPP
		=> S (AQP)+/SQSP	AQPAQP AQPAQPAQP AQPAQPAQPAQP AQPAQPAQPAQPAQP
&	Join multiple sequence fragments together as one	=> S ACDKLM & KLKWDN/SQSP	ACDKLMKLKWDN

STN 26

Tips

- Use 1-letter codes for common residues
- Use 3-letter codes for uncommon residues
  - Enclose three-letter codes in single quotes
- 1-letter and 3-letter codes can be mixed

  => S 'AIB'A'ABU''PIP'/SQSP
- Search shortcuts for blocking groups in the Notes (NTE) field

  => S BOC/NTE

For a full list of shortcuts for blocking groups, see the Quick Reference Guide.

STN 27

SCM is useful for finding sequence motifs

- Sequence motifs are short, recurring sequence patterns that can have a biological function
- Search motif sequences specific to a function to identify proteins with the same sequence that may have similar function, e.g., the pattern for N-Glycosylation
  - Asparagine followed by anything but proline, followed by either serine or threonine, followed by anything but proline D[-P][ST][-P]

STN 28

Claim 1: A peptide having the sequence

$X_1 – X_2 – PRO – THR – ALA – X_3 – (X_4)_n$

wherein

X_1 is an amino acid other than proline,

X_2 is Lys, His, Trp, Asp, Leu, Pro, or Tyr,

X_3 is an amino acid other than proline,

X_4 is any amino acid, and n repeats 0-6 times.

STN 29

Select the sequence symbols to use in the query

- Use single letter codes to represent the residues
- Use square brackets to represent possible residues at a specific location
- Use square brackets and minus sign to exclude residues
- Use a period (.) to represent any amino acid
- Use curly braces { } to represent repeats

STN 30

Construct the sequence query

Query

[-P][KHWDLPY]PTA[-P].{0-6}/SQSP

Recall the sequence claim:
$X_1 – X_2 – PRO – THR – ALA – X_3 – (X_4)n$
wherein
X_1 = an amino acid other than proline
X_2 = Lys, His, Trp, Asp, Leu, Pro, or Tyr
X_3 = any amino acid except proline
X_4 = any amino acid
n = repeat between 0-6 times

Repeat symbols can not be combined with excluding symbols.

STN 31

Run SQSP sequence query in REGISTRY

```
=> FILE REGISTRY

=> S [-P][KHWDLPY]PTA[-P].{0-6}/SQSP
L1      101893 [-P][KHWDLPY]PTA[-P].{0-6}/SQSP

=> S L1 AND 6-12/SQL
L2          156 L1 AND 6-12/SQL


=> D L2 SQIDE
```

Gaps, wildcards, and brackets are invalid for a SQEP search.

Refine the answers with sequence length (SQL).

Type D SQIDE to see sequence identification information, including the hit sequence.

STN 32

Example answer

```
RN   141785-31-1  REGISTRY
CN   L-Norleucinamide, L-lysyl-L-alanyl-L-arginyl-L-valyl-L-tyrosyl-
     L-prolyl-L- threonyl-L-alanyl- (9CI)  (CA INDEX NAME)
FS   PROTEIN SEQUENCE; STEREOSEARCH
SQL  9
NTE  modified
----------------------------------------------------------------------
 type           ------ location ------         description
----------------------------------------------------------------------
terminal mod.    Nle-9           -         C-terminal amide
uncommon         Nle-9           -         -
----------------------------------------------------------------------

SEQ      1 KARVYPTAX
                 ======
                  X2
                X1   X3
HITS AT:   4-9
```

This sequence does not have an X4 residue (zero repeats) .

Dependent Claim 2 requires a modification on the alanine residue in the sequence of Claim 1:

$X_1 – X_2 – PRO – THR – ALA – X_3 – (X_4)_n$

Modifications are searched in the NTE field

```
=> S ALA/NTE
L3        60197  ALA/NTE

=> S (ALA (P) MOD?)/NTE
L4        38099 (ALA (P) MOD?)/NTE

=> S (ALA (P) STEREO?)/NTE
L5        19233 (ALA (P) STEREO?)/NTE

=> S (ALA (P) BRIDG?)/NTE
L6         3437 (ALA (P) BRIDG?)/NTE

=> S (ALA (P) REPLAC?)/NTE
L7          391 (ALA (P) REPLAC?)/NTE

=> S L2 AND (L7 OR L8 OR L9 OR L10)
L8            3 L2 AND (L4 OR L5 OR L6 OR L7)
```

Use the (P) operator to ensure that alanine and the modification are listed on the same row within the Notes table.

Information is listed in the NTE table

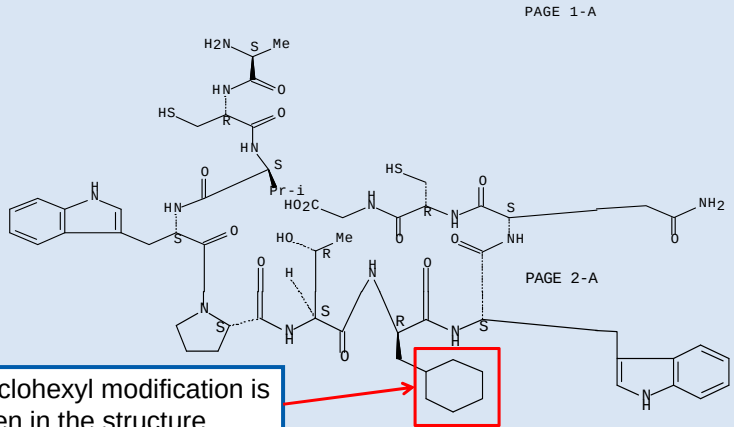
```
=> D SQIDE

RN   365254-57-5  REGISTRY
CN   Glycine, L-alanyl-L-cysteinyl-L-valyl-L-tryptophyl-L-prolyl-L-
     threonyl-3-cyclohexyl-D-alanyl-L-tryptophyl-L-asparaginyl-
     L-cysteinyl- (9CI)  (CA INDEX NAME)
OTHER NAMES:
CN   42: PN: WO0172771 FIGURE: 8-2 claimed protein
FS   PROTEIN SEQUENCE; STEREOSEARCH
SQL  11
NTE  modified (modifications unspecified)
-------------------------------------------------------------------
 type           ------ location ------         description
-------------------------------------------------------------------
modification     Ala-7          -         cyclohexyl<Chx>
-------------------------------------------------------------------
```

Specific modifications can be searched in REGISTRY. See Quick Reference Guide.

Alanine is modified by the addition of cyclohexyl.

CAS analysts include sequence modifications in the structure

```
• • •
SEQ      1 ACVWPTAWNC G
           ======== =
```



Cyclohexyl modification is seen in the structure.

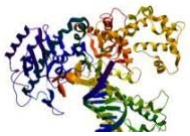
STN 37

Claim 1. Peptides or proteins having the following sequence:

$X_1ITIVFX_2V$

wherein X_1 is an uncommon amino acid, and X_2 is 2-aminobutanoic acid.

STN



Single letter codes and three letter codes can be combined in a query

- Use the code X to search uncommon amino acids in REGISTRY
- Place single quotation marks around the three-letter code
  - Search 2-aminobutanoic acid as 'ABU'
- Claimed sequence: $X_1ITIVFX_2V$
- Query sequence: XITIVF'ABU'V/SQSP

Search sequence in REGISTRY

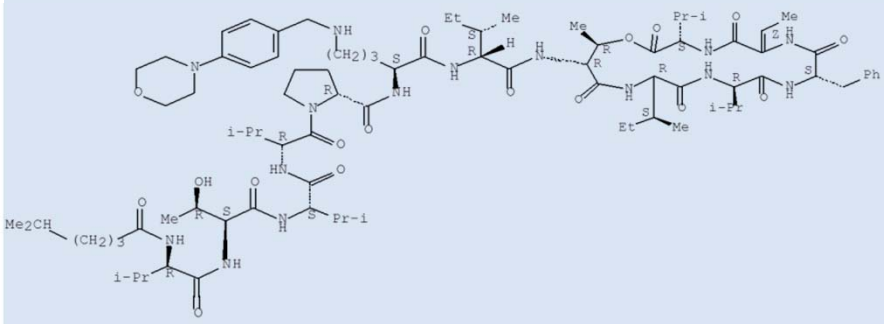
```
=> S XITIVF'ABU'V/SQSP
L1          185 XITIVF'ABU'V/SQSP

=> D SQIDE

RN   1350829-44-5  REGISTRY
CN   INDEX NAME NOT YET ASSIGNED
FS   PROTEIN SEQUENCE; STEREOSEARCH
SQL  13
NTE  modified (modifications unspecified)
----------------------------------------------------------------------
 type            ------ location ------         description
----------------------------------------------------------------------
bridge           Thr-8       - Val-13      lactone
uncommon         Orn-6           -         -
uncommon         Abu-12          -         -
modification     Val-1           -         undetermined modification
modification     Orn-6           -         undetermined modification
----------------------------------
SEQ      1 VTVVPXITIV FXV
                   ===== ===
```

Modifications that are not clearly stated within a document will be indexed as "undetermined modification."

Uncommon amino acids and modified will display as X in 1-letter codes.

Sequence structure includes both common and uncommon residues

```
Absolute stereochemistry.  Rotation (-).
Double bond geometry as shown.
```



STN 41

Claim 2. The peptide of Claim 1, wherein X_1 is epsilon-N-trimethyllysine.

STN 42

Refine the answers with a specific uncommon amino acid residue

TML is epsilon-N-trimethyllysine

```
=> S L1 AND TML/NTE; D SQIDE
           279 TML/NTE
L2           2 L1 AND TML/NTE

L2   ANSWER 1 OF 2  REGISTRY
RN   847968-30-3  REGISTRY
CN   Kahalalide F, 1-[N-[(4S)-4-methyl-1-oxohexyl]-D-valine]-6-[6-
     (trimethylammonio)-L-norleucine]- (9CI)  (CA INDEX NAME)
FS   PROTEIN SEQUENCE; STEREOSEARCH
SQL  13
NTE  modified (modifications unspecified)
-------------------------------------------------------------------
 type            ------ location ------         description
-------------------------------------------------------------------
bridge           Thr-8       - Val-13      lactone
uncommon         Tml-6           -         -
uncommon         Abu-12          -         -
modification     Val-1           -         undetermined modification
-------------------------------------------------------------------
```

STN 43

Claim 1. A composition comprising a peptide having the following sequence

X_1MYX_2KT

wherein X_1 is D-phenylalanine, and

X_2 is D-tryptophan.

Query

FMYWKT/SQSP

STN 44

D-amino acids are included in the Notes table

```
=> FILE REGISTRY
=> S FMYWKT/SQSP
L1            5 FMYWKT/SQSP
=> S L1 AND ((PHE (P) D) AND (TRP (P) D))/NTE
L2            1 L1 AND ((PHE (P) D) AND (TRP (P) D))/NTE
=> D NTE SEQ
L2   ANSWER 1 OF 1  REGISTRY
NTE  modified (modifications unspecified)
-----------------------------------------------------------------
 type            ------ location ------         description
-----------------------------------------------------------------
bridge           Phe-1       - Lys-9       covalent bridge
stereo           Phe-1           -         D
stereo           Trp-4           -         D
-----------------------------------------------------------------
SEQ      1 FMYWKTMTK
           ======
HITS AT:   1-6
```

Search in the NTE field with the (P) operator.

D-amino acids are listed as *STEREO*.

STN 45

Claim 1. A peptide comprising a carboxy terminal sequence

$$KTDSX_1VCDS$$

wherein X_1 is a gap of zero to ten residues, or functional equivalents thereof.

Query

KTDS.{0-10}VCDS^/SQSFP

Possible family substitutions for KTDS_VCDS:

K	T	D	S	V	C	D	S
H	A	B	A	I	C	B	A
R	G	E	G	M		E	G
	P	N	P	L		N	P
	S	Q	T			Q	T

STN 46

Use family searches to find functional equivalents

```
=> FILE REGISTRY

=> S KTDS.{0-10}VCDS^/SQSFP
L1        38 S KTDS.{0-10}VCDS^/SQSFP

=> D IDE SEQ SEQ3
RN   1282622-50-7  REGISTRY
CN   L-Alanine, L-seryl-L-isoleucyl-L-seryl-L-arginyl-L-threonyl-L-
     .alpha.-glutamyl-L-alanyl-L-alanyl-L-.alpha.-aspartyl-L-leucyl-
     L-cysteinyl-L-glutaminyl-  (CA INDEX NAME)
SEQ       1 SISRTEAADL CQA
                ======= ===
HITS AT:    4-13
SEQ3      1 Ser-Ile-Ser-Arg-Thr-Glu-Ala-Ala-Asp-Leu-
                        === === === === === === ===
         11 Cys-Gln-Ala
            === === ===
HITS AT:    4-13
```

Alternate display formats can be used. SEQ3 displays the sequence in 3-letter codes.

Summary

- Use SCM symbols to find sequences with variations
- Search sequence modifications in the NTE field
- Search uncommon amino acids with single quotations before and after the 3-letter code
- D-amino acids are easily retrieved using the NTE field

48

Resources

- Download PDF files with information on blocking groups, annotations, BLAST searching, and exact and pattern match searching of Sequences http://www.cas.org/training/stn/database-specific
- Send sequence-related comments or suggestions to the CAS Customer Center
  - help@cas.org
  - 1-800-753-4227
- View additional e-seminars on sequence searching

http://www.stn-international.com/recorded_events.html

For more information …

CAS
help@cas.org
Support and Training:
www.cas.org

FIZ Karlsruhe
helpdesk@fiz-karlsruhe.de
Support and Training:
www.stn-international.de