



STN®: Sequence Motif Searching



Today's presenters are...



Gin-Yun
Eggerichs



Alice
Humel-Denton

Why STN®?

- **Comprehensive Sequence Information**
 - Access 4 databases which compile nucleotide and protein sequence data from a wide range of sources including NCBI and patent authorities around the world
- **Most Current Sequence Information**
 - Frequent updates make sequence data available within days of publication
- **Convenience**
 - Use a single search tool to find sequence information from multiple sources
- **Versatility**
 - Choose the search method best suited to answer your sequence questions

STN

COS
SEARCH

3

Agenda

- Overview of the STN sequence databases
- Discuss the difference between a similarity search and a Sequence Code Match (SCM)
- Discuss the types of SCM searches
- Introduce special characters and symbols available for SCM search
- Retrieve sequences with uncommon amino acids
- Tips on how to search for RNA on STN

STN

COS
SEARCH

4

STN Sequence Databases

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

STN

CAS
FIZ Karlsruhe

5

CAS REGISTRY

- Produced by Chemical Abstracts Service
- Contains more than 60.1 million sequences from 1957 to present (Sept 08)
- Sequences are extracted from
 - Journal publications
 - Patents from 51 active patent authorities
 - GenBank[®]
- Intellectually analyzed content
 - Analyzed by CAS scientists who are experts in a variety of scientific disciplines
- Updated daily

Learn more at: www.cas.org

STN

CAS
FIZ Karlsruhe

6



- Produced by the SequenceBase Corporation
- Contain sequences from USPTO *published patent applications* and *granted patents*
- Patent information in the database is as given by patent applicant
- Bibliography, original publication title, abstract, and claims provided for each sequence
- Provides USPTO sequence data within 3 days of publication
- Covers from 1982 to present
- Updated weekly

Learn more at: www.fiz-k.com/usgene

STN

COS
FOR LIFE

7

DGENE

- Produced by Thomson Reuters
- Intellectually annotated database of nucleic acid and polypeptide sequences from Derwent World Patents Index® (DWPISM)
 - Extracted from the basic patent published by 41 patent offices worldwide
- Contains more than 8.1 million nucleic acid sequences (July 08)
- Contains more than 3.4 million protein sequences (July 08)
- Updated every two weeks

STN

COS
FOR LIFE

8

PCTGEN

- Produced by FIZ Karlsruhe
- Covers sequences from patent applications from 2001 to present
- Information as given by patent applicant
- Bibliography and original publication title provided for each sequence
- Contains more than 5.8 million DNA sequence records (July 08)
- Contains more than 795,100 protein sequences (July 08)
- Sequences are typically made available within one day of publication by WIPO
- Updated weekly

STN

CCS
FIZ Karlsruhe

9

BLAST® searching is available on all STN sequence databases

- BLAST (**B**asic **L**ocal **A**lignment **S**earch **T**ool) is the most common type of sequence searching
- Designed to search sequences with speed and minimal sacrifice of sensitivity
- A suite of programs that use algorithms to compare query sequences with the database sequences
 - Local alignment (short sequence) comparison
 - Statistical score helps identify the similarity between the query and the database sequence

STN

CCS
FIZ Karlsruhe

10

Why use BLAST?

- Find similarities between sequences
 - Predict domain locations
 - Infer domain functions
- Explore evolutionary relationships
 - How similar is human gene X vs. mouse gene X?
 - Where do the sequences diverge between organisms?
 - Functionality change?
- Identify DNA from unknown species/source

Searching for short sequences is not optimized in BLAST

- Retrieved sequences with exact pattern match (often referred to as motifs) are often placed at the end of the sequence result list because the rest of the alignment is low in similarity
 - Motifs: short stretch of amino acid or nucleic acid sequences that have a common pattern
 - Low score = less similarity
 - Loss of important information

Sequence Code Match (SCM) retrieves motifs of interest

- STN has the ability to search for a specific motif of interest
- SCM does not assign scores for sequences
 - Important sequences containing the motif are not lost because the rest of the sequence is dissimilar
- SCM allows variability within the sequence query using symbols and special characters

STN

CCS
Purdue

13

Benefits of using SCM

- Allows matching of a specific portion of the sequence
- Allows sequence variations or gaps within a motif
- Ability to specify the motif to be either at the N- or C-terminus
- Ability to specify or exclude residue

STN

CCS
Purdue

14

Four types of SCM are available on STN

- SCM
 - EXACT search
 - EXACT FAMILY search
 - SUBSEQUENCE search
 - SUBSEQUENCE FAMILY search
- REGISTRY
 - Use [Search](#) or [S](#) to initiate SCM search
- USGENE, DGENE, PCTGEN
 - Use [RUN GETSEQ](#) to initiate SCM search

STN

CCS
Purdue

15

Sequence Code Match: EXACT search

- EXACT search
 - Matches the sequence query as entered
 - Identical sequences and exact length
- Field codes identify the type of sequence to search, e.g.
 - /SQEN = SeQ uence Exa ct Nu cleotide**
 - /SQEP = SeQ uence Exa ct Peptide**

STN

CCS
Purdue

16

EXACT amino acid search (/SQEP) in REGISTRY

```

=> S NMSTYVDYK/SQEP
      1 NMSTYVDYK/SQEP
143623 SQL=9
L1      1 NMSTYVDYK/SQEP
      (NMSTYVDYK/SQEP AND SQL=9)

=> D SEQ SEQ3
L1 ANSWER 1 OF 1 REGISTRY COPYRIGHT 2008 ACS on STN
SEQ      1 NMSTYVDYK ←
=====
HITS AT: 1-9

SEQ3     1 Asn-Met-Ser-Thr-Tyr-Val-Asp-Tyr-Lys ←
===  ===  ===  ===  ===  ===  ===  ===  ===
HITS AT: 1-9

```

In REGISTRY, amino acids can be displayed as single letter codes (**SEQ**) or three letter codes (**SEQ3**).

TIP: Entering both **SEQ** and **SEQ3** in the same command prompt results in 1 sequence display fee.

STN

CS
Purdue

17

EXACT amino acid search (/SQEP) in USGENE

```

=> RUN GETSEQ NMSTYVDYK/SQEP
L1      1 NMSTYVDYK/SQEP

=> D SSO ORGN SEQ SEQ3
L1 ANSWER 1 OF 1 USGENE COPYRIGHT 2008 SEQUENCEBASE CORP
on STN
SSO PROTEIN; USPTO; APPLICATION
ORGN Yersinia pestis
SEQ      1 nmstyvdyk
=====
HITS AT: 1-9
      1 Asn-Met-Ser-Thr-Tyr-Val-Asp-Tyr-Lys
===  ===  ===  ===  ===  ===  ===  ===  ===
HITS AT: 1-9

```

In DGENE, USGENE, and PCTGEN, **RUN GETSEQ** is required to initiate SCM search.

Sequence source field (**SSO**) provides additional information such as the type of sequence (DNA, RNA, protein), and application or granted patent.

STN

CS
Purdue

18

Sequence Code Match: EXACT FAMILY search

- EXACT FAMILY search
 - Matches the sequence query as entered and allows family substitution to occur
 - Retrieves identical sequences and family sequences with exact length
 - Family substitutions only occur for proteins and not nucleic acids

/SQEFP = SeQUENCE EXACT FAMILY PEPTIDE

STN

COS
PUBLISHER

19

Amino acid family substitutions

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

STN

COS
PUBLISHER

20

EXACT FAMILY amino acid search (/SQEFP) in REGISTRY

```
=> S NPSKVTAYL/SQEFP
      2 NPSKVTAYL/SQEFP
      143623 SQL=9
L1    2 NPSKVTAYL/SQEFP
      (NPSKVTAYL/SQEFP AND SQL=9)

=> D L1 SEQ
L1    ANSWER 1 OF 2  REGISTRY  COPYRIGHT 2008 ACS on STN
SEQ    1 NAPRLSSFL
      =====
HITS AT: 1-9
```

Possible family substitutions for NPSKVTAYL:

	N	P	S	K	V	T	A	Y	L
Q	A	A	H	I	A	P	F	I	
E	G	G	R	M	G	G	W	M	
D	S	P	L	S	S	V			
B	T	T			P	T			
Z									

STN

CCS
Chemical Compound Search

21

Sequence Code Match: SUBSEQUENCE search

- SUBSEQUENCE search
 - Retrieves exact answers plus sequences that are embedded in longer sequence

/SQSN = Sequence **S**ubsequence **N**ucleotides

/SQSP = Sequence **S**ubsequence **P**eptides

STN

CCS
Chemical Compound Search

22

SUBSEQUENCE amino acid search (/SQSP) in REGISTRY

```

=> S NMSTYVDYK/SQSP
L1          138 NMSTYVDYK/SQSP

=> D L1 4 SQIDE
L1  ANSWER 4 OF 138  REGISTRY  COPYRIGHT 2008 ACS on STN
RN  1008818-88-9  REGISTRY
CN  Antigen RECP02221 (Escherichia coli strain UPEC-536)
    (CA INDEX NAME)
OTHER NAMES:
CN  115: PN: WO2008020330 SEQID: 113 claimed protein
FS  PROTEIN SEQUENCE
SQL  375
    
```

In REGISTRY, patent information for some sequences can sometimes be found in the Chemical Name (/CN) field.

Continued on next slide



23

SUBSEQUENCE amino acid search (/SQSP) in REGISTRY

PATENT ANNOTATIONS (PNTE) :

Sequence Source	Patent Reference
Not Given	WO2008020330
	claimed SEQID
	113

Since 1999, sequence patent information can also be found in the Patent Annotations Table (/PNTE).

```

SEQ      1 MKVKVLSLLV PALLVAGAAN AAEVYNKDGK KLDLYGKVDG LHYFSDDKSV
      51 DGDQTYMRLG FKGETQVTDQ LTGYGQWEYQ IQGNAPESN NSWTRVAFAG
     101 LKFQDIGSFD YGRNYGVVYD VTSWTDVLPE FGGDTYGSND FMQQRGNGFA
     151 TYRNTDFFGL VDGLNFAVQY QGQNGSVSGE NDPDFTGHGI TNNGRKALRQ
     201 NGDGVGGSIT YDIEGFGVGA AVSSSKRTDA QNTAAYIGNG DRAETYTGGL
     251 KYDANNIYLA AQYTQTYNAT RVGSLGWANK AQNFEAVAQY QFDGFLRPSV
     301 AYLQSKGKNL GTIGTRNYDD EDILKYVDVG ATYYFNKNMS TYVDYKINLL
                                     === =====
      351 DDNQFTRDAG INTDNIVALG LVYQF
HITS AT:  338-346
    
```



24

Sequence Code Match: SUBSEQUENCE FAMILY search

- SUBSEQUENCE FAMILY search
 - Retrieves EXACT sequence match, SUBSEQUENCE match, and sequences that contain family substitution of amino acid

=> **S DSDGP/SQSFP**

/SQSFP = Sequence **S**ubsequence **F**amily **P**eptides



25

SUBSEQUENCE FAMILY amino acid search (/SQSFP) in REGISTRY

=> **S NMSTYVDYK/SQSFP**

L1 683 NMSTYVDYK/SQSFP

=> **D L1 SEQ NTE**

L1 ANSWER 10 OF 683 REGISTRY COPYRIGHT 2008 ACS on STN

SEQ 1 MVKXXRQALP LXIDGXXYDV SAWVNFHPGG AEIENYQGR DATDAFMVMH

== =====

51 SQEXXDKLKR MPKINXXSEL PPQAAVNEAQ EDFRKLREEL IATGMFDASP

Possible family substitutions for NMSTYVDYK:

N	M	S	T	Y	V	D	Y	K
Q	L	A	A	F	L	Q	F	H
E	I	G	G	W	I	E	W	R
D	V	P	P	M	N			
B	T	S			B			
Z					Z			

YXXYFIG AXXLGMHYQQ MGWLSHDICH

PSVTWWK DRHNAHSAT NVQGHDPDID

QFQYYF LVICILLRFI WCFQSVLTVR

HWTLKXX FHLFFMPSIL TSXLVFFVSE

SVWDGHG FSVGQIHETM NIRRGXXXDW

WVSYQVE QLCQKHNLPY RNPLPHEGLV

Continued on next slide



26

SUBSEQUENCE FAMILY amino acid search (/SQSFP) in REGISTRY

NTE

type	location	description
uncommon	Aaa-4	-
uncommon	Aaa-5	-
uncommon	Aaa-12	-
• • •		
uncommon	Aaa-171	-
uncommon	Aaa-279	-
uncommon	Aaa-280	-
uncommon	Aaa-293	-
uncommon	Aaa-346	-
uncommon	Aaa-347	-
uncommon	Aaa-348	-
uncommon	Aaa-407	-
uncommon	Aaa-408	-
uncommon	Aaa-418	-
uncommon	Aaa-419	-
uncommon	Aaa-422	-

In REGISTRY, chemically modified sequences are listed with their position in the Sequence Annotation Field (**NTE**).



27

Summary of the Sequence Code Match options

Search Type	Polypeptides	Nucleic Acids
EXACT	/SQEP	/SQEN
EXACT FAMILY	/SQEFP	—
SUBSEQUENCE	/SQSP	/SQSN
SUBSEQUENCE FAMILY	/SQSFP	—



28

Special motif symbols allow more user control in sequence searching

- Allow users to specify motif patterns that consist of different amino acid(s) at one location of the sequence
- Ability to specify gaps of any size between specific amino acid sequences
- Ability to search for sequence patterns at either beginning or the end of the sequence
- Users can specify the number or range of repeats for amino acid(s) or gaps

STN

CCS
ProteinLife

29

Motif searching symbols and characters

Symbols	Function	Example	Possible Answers
^	Search at the beginning or the end of a sequence	^MCGIL/SQSP VCDS^/SQSFP	"MCGIL....." ".....VCDS"
[]	Specify alternate residues	LGP[VL]/SQSP	LGPV LGPL
[-] or [~]	Exclude one or more residues	PTGK[-H]/SQSP PTGK[~H]/SQSP	PTGKACCD
{#,#} {# - #} {#}	Repeat preceding residue(s)	GG(FL){1,3}/SQSP GG(FL){1-3}/SQSP GG(FL){3}/SQSP	GGFL GGFLFL GGFLFLFL
.	Specify gap(s) in the sequence	SY.RPG/SQSP SY...RPG/SQSP	SYARPG SYAAARPG
	Specify alternate residues	ACD KLM/SQSP A(CD KL)M/SQSP	ACD KLM ACDM AKLM

STN

CCS
ProteinLife

30

Motif searching symbols and characters

Symbols	Function	Examples	Possible Answer
?	Repeat residue(s) zero or one time	FLRR(RP)?K/SQSP	FLRRK FLRR RP K
*	Repeat residue(s) zero or more times	KLK(WD)*N/SQSP	KLKN KLK WD N KLK WDWD N KLK WDWDWD N
+	Repeat residue(s) one or more times	AQP+/SQSP	AQ PP AQ PPP AQ PPPP AQ PPPPP
		(AQP+)/SQSP	AQP AQP AQP AQPAQP AQP AQPAQPAQP AQP AQPAQPAQPAQP
&	Join multiple sequence fragments together as one	ACDKLM & KLKWDN /SQSP	ACDKLMKLKWDN

STN

CS
PROMOTIONS

31

Parentheses determine the order of operation

- Without parentheses, the order of operations are
 - Repetition symbols: ?, *, +
 - Repetition operator: { }
 - Concatenation symbol: &
 - Alternation symbol: |

STN

CS
PROMOTIONS

32

Search Question 1: Duchenne Muscular Dystrophy (DMD) is a recessive disease that results in rapid progression of muscle degeneration. The DMD gene codes for protein, dystrophin, that has a WW domain with a consensus pattern. Find other proteins have the same motif.

W X₁ X₂ X₃ X₄ X₅ X₆ X₃ X₇ X₈ P

X₁ = Any 9-11 amino acids

X₂ = V, F, or L

X₃ = F, Y, or W

X₄ = Any 6-7 amino acids

X₅ = G, S, T, N, or E

X₆ = G, S, T, Q, C, or R

X₇ = Any amino acid except R

X₈ = Any amino acid except S or A

STN

CCS
FINDERS

33

Assemble the motif sequence with STN symbols/characters

- Period (.) represents gaps
- { } represents repeats
- [] represents "OR"
- Minus sign (-) represents "NOT"

W.{9-11}[VFY][FYW].{6-7}[GSTNE][GSTQCR][FYW][-R][-SA]P

STN

CCS
FINDERS

34

Use SUBSEQUENCE (/SQSP) search when using sequence variability symbols

```
=> S W.{9-11}[VFY][FYW].[6-7][GSTNE][GSTQCR][FYW]
[-R][-SA]P/SQSP
```

```
L1 2728 W.{9-11}[VFY][FYW].[6-7][GSTNE][GSTQCR][FYW]
[-R][-SA]P/SQSP
```

SQSP (SUBSEQUENCE) is used because the WW domain is most likely embedded within a larger sequence.

```
=> D SQIDE 10
```

```
L1 ANSWER 10 OF 2728 REGISTRY COPYRIGHT 2008 ACS on STN
RN 1029162-28-4 REGISTRY
CN Crohn's disease-associated protein (human clone
US2008067195-SEQID-482) (CA INDEX NAME)
```

OTHER NAMES:

```
CN 482: PN: WO2008067195 SEQID: 482 claimed protein
FS PROTEIN SEQUENCE
SQL 1455
```

The WW motif search retrieves proteins with similar motifs that are associated with Crohn's disease.

STN

CCS
PattentLife

35

Use SUBSEQUENCE (/SQSP) search when using sequence variability symbols

PATENT ANNOTATIONS (PNTE):

Sequence Source	Patent Reference
Not Given	WO2008067195 claimed SEQID 482

This protein is a claimed protein with a SEQID 482 that contains two WW domains.

```
SEQ 1 MSKSLKKKSH WTSKVHESVI GRNPEGQLGF ELKGAENGQ FPYLGEVKPG
51 KVAYESGSKL VSEELLLEVN ETPVAGLTIR DVLAVIKHCK DPLRLKCVKQ
101 GGIVDKDLRH YLNLRFOKGS VDHELQIIR DNLYLRTVPC TTRPHKEGEV
151 PGVDYIFITV EDFMELEKSG ALLESGTYED NYYGTPKPPA EPAPLLLNVT
201 DQILPGATPS AEGKRKRNKS VSNMEKASIE PPEEEEEERP VVNGNGVVVT
251 PESSEHEDKS AGASGEMPSQ PYPAPVYSQP EELKEQMDDT KPTKPEDNEE
301 PDPLPDNWEM AYTERGEVYF IDHNTKTTSW LDPRLAKKAK PPEECKENEL
====
351 PYGWEKIDDP IYGTYYVDHI NRRTQFENPV LEAKRKLQQH NMPHTELGTK
====
401 PLQAPGFREK PLFTRDASQL KGTFLLSTTLK KSNMGFGFTI IGGDEPDEFL
...

```

STN

CCS
PattentLife

36

Sequence Annotation (/NTE) field contains chemically modified sequence information

- The /NTE field contains additional information
 - Classification of the sequence (i.e. double/single stranded, multichain, linear, or cyclic)
 - Type of chemical modification (i.e. uncommon amino acid/base or bridge)
 - Sequence position where the chemical modification has occurred
 - Terms describing the type of modification (i.e. blocking group, metal complex)
- The /NTE field contains single words that may be searched with Boolean or proximity operators

STN

COS
PentaLife

37

Blocking groups can be searched using full name or the shortcut

```
=> E CYCLOPENTYL/NTE 5
E1      58      CYCLOHEXYLCARBONYL/NTE
E2      33      CYCLOHEXYLOXY/NTE
E3      50 -->  CYCLOPENTYL/NTE
E4      51      CYCLOPENTYLCARBONYL/NTE
E5      25      CYCLOPENTYLOXY/NTE
```

```
=> E CPE/NTE 5
E6      14981   COVALENT/NTE
E7      51      CPC/NTE
E8      50 -->  CPE/NTE
E9      59      CPM/NTE
E10     84      CU/NTE
```

EXPAND on the blocking group name or shortcut before searching.

STN

COS
PentaLife

38

Multiple terms can be used to search in the /NTE field

```
=> S (CYCLOPENTYL AND CYCLIC)/NTE
L1          3 (CYCLOPENTYL AND CYCLIC)/NTE

=> D NTE SEQ

L1  ANSWER 1 OF 3  REGISTRY  COPYRIGHT 2008 ACS on STN
NTE  cyclic
```

type	location	description
bridge	Cys-5 - Cys-10	disulfide bridge
uncommon	Nle-9	-
modification	Ala-1	cyclopentyl<Cpe>

```
SEQ      1 APPQCYPPXC TAS
```

The shortcuts for the blocking groups are found next to the full blocking group name.

STN

CCS
Chemical Computing System

39

Uncommon amino acids can be searched on STN using the 3-letter codes

Uncommon amino acids		3-Letter Code Name		3-Letter Code Name	
Aaa	α -amino acid	Har	homocysteine	Mpa	mercaptopropanoic acid
Aad	2-aminoadipic acid (2-aminohexanedioic acid)	Hcy	homocysteine	Nle	norleucine
Aan	α -asparagine	Hha	homohistidine	Nty	nortyrosine
Abu	2-aminobutanoic acid	Hiv	2-hydroxyisovaleric acid	Nva	norvaline
Aca	2-aminocaproic acid (2-aminododecanoic acid)	Hse	homoserine	Oaa	α -amino acid
Agm	α -glutamine	Hva	2-hydroxypentanoic acid	Orn	ornithine
Alb	α -aminobutyric acid (α -methylalanine)	Hyl	5-hydroxylysine	Pen	penicillamine (3-mercaptoproline)
Apim	2-aminopimelic acid (2-aminoheptanedioic acid)	Hyp	4-hydroxyproline	Pbg	2-phenylglycine
App	γ -amino- β -hydroxybenzenepentanoic acid	Inc	2-carboxyoctahydroindole	Pip	2-carboxypiperidine
Aeu	2-aminosuberlic acid (2-aminooctanedioic acid)	Iqc	3-carboxyquinoline	Sar	sarcosine (N-methylglycine)
		Iva	isovaline	Spg	1-amino-1-carboxycyclopentane
		Lac	2-hydroxypropanoic acid (lactic acid)	Sta	statin (4-amino-3-hydroxy-6-methylheptanoic acid)
		Maa	mercaptacetic acid	Thi	3-thiethylalanine
		Mba	mercaptobutanoic acid	Tml	\pm -N-trimethyllysine
		Mhp	4-methyl-5-hydroxyproline	Tza	3-thiazolylalanine
				Und	undefined
				Wll	α -amino-2,4-dioxypyrimidinepropanoic acid
Aze	2-carboxyzetidine				
Bal	β -alanine				
Bas	β -aspartic acid				
Bly	3,6-diaminohexanoic acid (β -lysine)				
Bua	butanoic acid				
Bux	4-amino-3-hydroxybutanoic acid				
Cap	γ -amino- β -hydroxycyclohexanepentanoic acid				
Cit	N ⁶ -aminocarbonylornithine				
Cys	3-sulfalalanine				
Dab	2,4-diaminobutanoic acid				
Dpm	diaminopimelic acid				
Dpr	2,3-diaminopropanoic acid				
Dsu	2,7-diaminooctanedioic acid				
Edc	S-ethylthiocysteine				
Ggu	γ -glutamic acid				
Gla	γ -carboxyglutamic acid				
Glc	hydroxyacetic acid (glycolic acid)				
Glp	pyroglutamic acid				

STN

CCS
Chemical Computing System

40

Uncommon amino acids can be searched on STN using the 3-letter codes

```
=>S MCKKK'NVA'GG'ABU'GGA/SQEP
L1      1 MCKKK'NVA'GG'ABU'GGA/SQEP
      (MCKKK'NVA'GG'ABU'GGA/SQEP AND SQL=12)
```

```
=> D SQIDE
```

```
L1 ANSWER 1 OF 1 REGISTRY COPYRIGHT 2008 ACS on STN
RN 1034249-58-5 REGISTRY
CN INDEX NAME NOT YET ASSIGNED
FS PROTEIN SEQUENCE; STEREOSEARCH
SQL 12
```

ABU = 2-aminobutanoic acid.
NVA = norvaline.

Uncommon amino acids must be searched with single quotation marks (') to the left and right of the 3-letter code.

Continued on next slide

STN

CCS
Chemical Abstracts

41

Uncommon amino acids can be searched on STN using the 3-letter codes

NTE modified (modifications unspecified)

type	location	description
bridge	Cys-2 - Abu-9	covalent bridge
bridge	Nva-6 - Ala-12	covalent bridge
uncommon	→ Nva-6	-
uncommon	→ Abu-9	-

```
SEQ      1 MCKKKXGGXG GA
          =====
```

The sequence annotation field (/NTE) gives the position of the uncommon amino acid.

Within the sequence, the uncommon amino acids are represented by an "X".

TIP: The position of the uncommon amino acid can be searched in the /NTE field (i.e. **S NVA-6/NTE**)

STN

CCS
Chemical Abstracts

42

Searching for RNAs on STN

- The standard policy for all RNA sequences at NCBI is “U” are converted to “T”, including those sourced from patents
 - In USGENE, RNAs from NCBI are represented with “T” instead of “U”
- REGISTRY and DGENE indexes RNA with “U”
 - In REGISTRY, RNAs from NCBI contain “T” instead of “U” whereas non-NCBI RNAs are indexed as “U”

STN

COS
PSEARCH

43

In REGISTRY, “U” will search for both “T” and “U”

```
=> S UGAAGCGGAGCUGGAA/SQSN AND SQL=16 AND RNA/CNS; D
SQSIDE 1-2
```

```
L1 ANSWER 1 OF 2 REGISTRY
RN 866168-32-3 REGISTRY
CN RNA (U-G-A-A-G-C-G-G-A-G-C-U-G-G-A-A) (9CI) (CA INDEX NAME)
OTHER NAMES:
CN 4: PN: WO2005092393 FIGURE: 1 claimed RNA
FS NUCLEIC ACID SEQUENCE
SQL 16
PATENT ANNOTATIONS (PNTE):
Sequence | Patent
Source | Reference
=====+=====
Not Given|WO2005092393
          |claimed FIGURE
          |1
SEQ      1 ugaagcggag cuggaa
```

Two records are retrieved using “U” in the SUBSEQUENCE search.

The first record is a claimed RNA found in a PCT application. The RNA is indexed as “U”.

STN

COS
PSEARCH

44

In REGISTRY, "U" will search for both "T" and "U"

```

L1 ANSWER 2 OF 2 REGISTRY COPYRIGHT 2008 ACS on STN
RN 422620-34-6 REGISTRY
CN RNA (human microRNA miR-115 gene,) (9CI) (CA INDEX
NAME)
OTHER NAMES:
CN GenBank AF480513
FS NUCLEIC ACID SEQUENCE
SQL 16
NA 5 a 2 c 7 g 2 t

SEQ 1 tgaagcggag ctggaa
=====

```

The second record is a RNA with "T" indexed instead of "U". The original source of this sequence is GenBank.

In REGISTRY using "U" in a SUBSEQUENCE (/SQSN) search retrieves RNAs that are indexed as either "T" or "U".



In USGENE, use both T and U for a comprehensive RNA search

```

=> RUN GETSEQ GGGAAUACCA/SQSN
L1 20 GGGAAUACCA/SQSN

=> S L1 AND SQL=10; D BIB MTY SSO SEQ

L2 ANSWER 1 OF 1 USGENE COPYRIGHT 2008 SEQUENCEBASE CORP on STN
AN 20080167258.17 RNA USGENE
TI Trans-excision-splicing ribozyme and methods of use
(Published Application)
IN Testa Stephen M. (Lexington, KY); Bell Michael A.
(Lexington, KY)
PA UNIVERSITY OF KENTUCKY RESEARCH FOUNDATION (Lexington KY)
PI US 20080167258 A1 20080710
AI US 2007-723456 20070320
DT Patent
MTY RNA
SSO NUCLEIC; USPTO; APPLICATION
SEQ 1 gggaaauacca
=====

```

This RNA record (indexed with "U") is from USPTO.

In USGENE, "U" retrieve only sequences containing "U".



In USGENE, use both T and U for a comprehensive RNA search

```
=> RUN GETSEQ GGAATACCA/SQSN
L1          6437 GGGAAUACCA/SQSN

=> S L1 AND SQL=10;D BIB MTY SSO SEQ

L2  ANSWER 1 OF 1 USGENE COPYRIGHT 2008 SEQUENCEBASE CORP on STN
AN   7211661.17 RNA          USGENE
TI   Trans-excision-splicing ribozyme and methods of use (Patent)
IN   Testa Stephen M. (Lexington, KY); Bell Michael A.
      (Lexington, KY)
PA   University of Kentucky Research Foundation (Lexington KY)
PI   US 7211661          BI   20070501
AI   US 2003-730261     BI   20031209
DI   Patent
MTY  RNA
SSO  NUCLEIC; NCB1; GRANTED
SEQ  1 gggaatacca
      =====
```

This RNA record (indexed with "T") is from NCBI with an earlier application date than the USPTO record.

In USGENE, "T" retrieve only sequences containing "T".

STN

CCS
SequenceBase

47

Summary

- STN has 4 sequence databases that can be accessed to achieve comprehensive search results
- SCM search allows more user control in sequence variations than BLAST
- Modification sequences can be searched using text in the sequence annotation table (/NTE)
- Uncommon amino acids are searched as three-letter codes with single quotations before and after the code

STN

CCS
SequenceBase

48

Summary

- Uncommon amino acid position can be specified in the /NTE field
- In REGISTRY, “U” can retrieve RNAs that are indexed as “U” or “T”
- In USGENE, both “U” and “T” must be searched separately for the search results to be comprehensive

STN

COS
Karlsruhe

49

STN

FIZ Karlsruhe Home | Terms and Conditions | Contact
About us | Guided Tour | STN Easy | STN on the Web | Site Map | Search Site

INPADOCDB - the new enhanced INPADOC file on STN

Your Connection to Science and Technology

Get Connected! [info](#)
STN International connects scientists, engineers and anyone who needs technical information to the world's most complete and authoritative databases.

From Your Desktop [info](#)
Select your preferred STN interface and:

- ask questions simply or by using sophisticated search commands
- identify published research and patents in all scientific fields
- retrieve original full-text articles and patents on the Web
- search chemical substance information by structure, name, or CAS Registry Numbers (CAS Number)

Be Confident [info](#)
You can use STN with confidence because the system and the more than 220 databases it brings you are operated by some of the most respected scientific organizations in the world.

[USGENE is now available](#)

STN Service Centers [info](#)

- FIZ Karlsruhe in Europe
- CAS in North America
- Japan Association for International Chemical Information (JAIC)

© FIZ Karlsruhe 2007 - Last Update: 09/26/2007 01:45:08 - Inprint

Training Center

- Workshops
- e-Seminars
- Getting Started with STN
- Materials for Searching STN
- STN Express
- Interactive Training
- STN Easy Demo

STN Archive

- STN News
- STN Brochures
- Presentations
- e-Seminar Archive

Press Room

- Contact
- Press Releases

Please visit www.stn-international.com for more FIZ Karlsruhe training information.

STN

COS
Karlsruhe

50

ACS | Journals | CMB | CAS

Advanced Search

Home | About CAS | Our Expertise | Solutions | Products & Services | Support & Training | News & Events

Home • Support • STN • e-Seminars and Training

STN e-Seminars and Training

STN provides a variety of free training resources for new users as well as for experienced STN searchers:

- e-Seminars
- Instructor-Led Training
- Self-Directed Learning
- Additional Training Resources

e-Seminars

e-Seminars are interactive, web-based seminars that bring professional training to your desktop. View the sessions live, or access a variety of previously recorded sessions to view at your convenience:

- CAS e-Seminars
- FIZ Karlsruhe e-Seminars

-TOP-

Instructor-Led Training

Instructor-led training provides opportunities to develop searching skills or to simply learn about new features and enhancements on STN:

- STN Workshops, User Updates, and Patent Forums
- STN Virtual Workshops
- STN Patents & Pizza sessions
- FIZ Karlsruhe Workshops in the USA

Please visit www.cas.org for more CAS training information.

STN 51

2008 STN Virtual Workshops

- Virtual workshops allow for hands-on training in our virtual lab along with an Applications Specialist
- Workshops offered
 - STN Basics
 - REGISTRY
 - Structure Searching
 - Patent Basics
 - STN Intermediate
- Private workshops available by request

Click on the Training Center tab at <http://casevents.webex.com>

STN 52

Questions and answers...

All e-Seminars				
STN				
		Product:	Category:	Language:
		STN	All	All <input type="button" value="Display"/>
<small>All event times in: Eastern DT</small>				
Date & Time	Event	Category	Subcategory	
August 2008				
August 26, 2008 13:00 - 14:00 Eastern DT	STN: Finding Licensing Information on STN®	Search Techniques	Intermediate	Enroll
September 2008				
September 11, 2008 9:00 - 10:00 Eastern DT	STN: Finding Licensing Information on STN®			Enroll
September 30, 2008 13:00 - 14:00 Eastern DT	STN: Sequence Motif Searching	Search Techniques	Intermediate	Enroll
October 2008				
October 9, 2008 9:00 - 10:00 Eastern DT	STN: Sequence Motif Searching	Sequences	Intermediate	Enroll
October 20, 2008 13:00 - 14:00 Eastern DT	STN: What's New in STN Express®	STN Express®	Intermediate	Enroll
November 2008				
November 13, 2008 9:00 - 10:00 Eastern ST	STN: What's New in STN Express®			Enroll
November 25, 2008 13:00 - 14:00 Eastern ST	STN: Polymers in Scientific Literature	Polymers	Intermediate	Enroll
December 2008				
December 11, 2008 9:00 - 10:00 Eastern ST	STN: Polymers in Scientific Literature	Polymers	Intermediate	Enroll
December 16, 2008 13:00 - 14:00 Eastern ST	STN: Dealing with Large Answer Sets	Search Techniques	Intermediate	Enroll
January 2009				
January 8, 2009 9:00 - 10:00 Eastern ST	STN: Dealing with Large Answer Sets			Enroll

<http://casevents.webex.com>



53