

Knowing whether to use U or T when searching RNA sequences on STN

Introduction

When you are designing a search query for sequence information on STN, it is important to consider how specific databases index RNA sequences.

Extensive sequence information is available in:

- DGENE (GENESEQ™)
- CAS REGISTRYSM
- USGENE® (USPTO Genetic Sequence Database)
- PCTGEN (World Patent Application Biosequences)

These databases derive their sequence content from multiple sources, including journals, patents, and GenBank®. RNA sequences in these databases are indexed with U or T, depending on the source of the data.

For example, the National Center for Biotechnology Information (NCBI) has a policy of converting uracil (U) residues to thymidine (T) for RNA sequences in GenBank. Other sources of sequence data, however, index RNA sequences with U.

This article provides descriptions and examples of how you can effectively search for RNA sequences in STN databases.

DGENE (GENESEQ)

DGENE contains nucleotide and peptide sequences from basic patent documents of 41 patent-issuing authorities. RNA sequences in DGENE are reported with U residues. Therefore, you should include U's in your RNA search query sequences.

```
=> FIL DGENE
=> RUN GETSEQ GGGAAUACCA/SQSN

L6          22 GGGAAUACCA/SQSN

=> D SQIDE

L6  ANSWER 1 OF 22  DGENE  COPYRIGHT 2008  THOMSON REUTERS on STN
AN  ASQ24900  RNA      DGENE
NA  4 A; 2 C; 3 G; 0 T; 1 U; 0 Other
SQL  10
SEQ  1  gggaauacca
      =====
HITS AT:  1-10
```

CAS REGISTRY

RNA sequence records in REGISTRY can include U or T, depending on their original source. RNA sequences from GenBank contain T, but those from non-GenBank sources contain U.

But, REGISTRY automatically helps with this. A subsequence query containing U allows for ambiguous matches on either T or U.

In the following example, two records are retrieved using U in the Subsequence search. The first contains an RNA sequence indexed with U that is found in a PCT application.

```
=> FIL REGISTRY
=> S UGAAGCGGAGCUGGAA/SQSN AND SQL=16

      3 UGAAGCGGAGCUGGAA/SQSN
      101308 SQL=16
L1      3 UGAAGCGGAGCUGGAA/SQSN AND SQL=16

=> S L1 AND RNA/CNS; D SQIDE 1-2

L1      ANSWER 1 OF 2 REGISTRY COPYRIGHT 2008 ACS on STN
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
```

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

```
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
```

USGENE (USPTO Genetic Sequence Database)

USGENE includes sequence records from the USPTO and GenBank. RNA sequences from the USPTO are indexed with U. RNA sequences from GenBank are indexed with T.

Unlike REGISTRY, searches in USGENE for sequences containing U do not automatically result in ambiguous matches on T or U. Searching with U only retrieves sequences containing U. Searching with T only retrieves sequences containing T.

For comprehensive results, it is therefore necessary to conduct two searches: one with T and a second with U.

The following example shows a search in USGENE for an RNA sequence containing U. The retrieved record is from the USPTO and is indexed with U.

```
=> FIL USGENE
=> 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 gggaauacca
        =====
```

The next example shows a search for the same RNA sequence with T instead of U. The retrieved record (indexed with T) is from GenBank (NCBI) with an earlier application date than the USPTO record shown above.

```
=> RUN GETSEQ GGGAATACCA/SQSN
L1          6437 GGGAATACCA/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      B1      20070501
AI  US 2003-730261  20031209
DT  Patent
MTY RNA
SSO NUCLEIC; NCBI; GRANTED
SEQ
      1 gggaatacca
        =====
```


Conclusion

As more patents for DNA-related inventions are issued, comprehensive searches for sequence patentability become increasingly important. Freedom-to-operate, prior art, validity, and infringement patent sequence searches can be conducted in STN databases that contain extensive sequence and patent information. Keeping in mind the differences between these databases can help you improve your RNA sequence search strategies and retrieve more comprehensive results.

Overview of RNA sequence indexing

	Thymidine (T)	Uracil (U)	Simultaneous Search
DGENE		X	n/a
REGISTRY	X	X	yes
USGENE	X	X	no
PCTGEN	X	X	no

Additional Resources

For additional information about the databases mentioned above, refer to the STN Database Summary Sheets at www.stn-international.de.

For more information about sequence searching on STN, visit the Training Center section of www.stn-international.de. Click Materials for Searching STN, and then click Biosequence Searching.