

CAS REGISTRYSM: Exact and pattern searching of nucleic acid sequences

November 2008

Table of Contents

Preface	3
Searching exact sequences	5
Using SEQLINK.....	5
Searching partial sequences	6
Pattern searching	8
Gaps	8
Repetition.....	8
Other variability options	9
Order of execution for symbols.....	9
Pattern searching example	10
Searching length.....	13
Searching chemical annotation	14
Appendix: Annotation for chemically modified nucleic acids	15

Preface

This guide provides an overview and examples of exact and pattern searching of nucleic acid sequences in the CAS REGISTRY database on STN.

CAS REGISTRY BLAST[®] similarity searching is available using STN Express[®] or STN[®] on the WebSM. For information, refer to the *CAS REGISTRYSM: BLAST[®] similarity searching via STN Express[®]* guide available at www.cas.org.

For information on searching in REGISTRY on STN, please refer to the REGISTRY Database Summary Sheet available at www.cas.org.

Searching exact sequences

To find an exact sequence of a nucleotide in REGISTRY, enter the sequence in the Exact Sequence Search (/SQEN) field.

The following codes may be used in exact nucleic acid sequence searches:

Code	Name or Definition
A	adenosine
C	cytidine
G	guanosine
T	thymidine (2'-deoxythymidine)
U	uridine (Note: ribothymidine = 5-methyluridine)
I	inosine

Using SEQLINK

The SEQLINK EXACT command is used to locate additional nucleic acid sequences that match a sequence that has already been retrieved from REGISTRY.

Find literature or patents on a diagnostic probe with the sequence CGCCCCTGCGTTACCCTCCCCGCCG.

- 1 Enter REGISTRY.
- 2 Use the SEARCH (or S) command to search the exact sequence in the /SQEN field.
- 3 Display the sequence (SEQ), annotation (NTE), and the Locator (LC) field listing the databases containing references to the CAS Registry Number®.
- 4 Use the SEQLINK command (free of charge) to find related sequences, if any.

```
=> FILE REGISTRY
=> S CGCCCCTGCGTTACCCTCCCCGCCG/SQEN
L1      3 CGCCCCTGCGTTACCCTCCCCGCCG/SQEN

=> D SEQ NTE LC 3
L1     ANSWER 3 OF 3  REGISTRY  COPYRIGHT 2008 ACS on STN

SEQ      1 cgcccctgcg ttaccctccc cgccg
          =====
HITS AT:  1-25

**RELATED SEQUENCES AVAILABLE WITH SEQLINK**
LC   STN Files:  CA, CAPLUS, TOXCENTER, USPATFULL

=> SEQLINK
ENTER TYPE OF LINK (EXACT) OR ?:EXACT
ENTER (L1), L# OR ?:L1
L2      3 SEQLINK EXACT L1
```

5 Enter one or more of the databases containing the CAS Registry Number.

6 Search the REGISTRY L-number (L2).

7 Display the bibliographic information (BIB), abstract (AB), and index entry for the hit sequence (HITSEQ).

```
=> FILE CAPLUS
=> S L2
L3          1 L2

=> D BIB AB HITSEQ
L3 ANSWER 1 OF 1 CAPLUS COPYRIGHT 2008 ACS on STN
AN 1995:884205 CAPLUS Full-text
DN 123:278057
TI Early diagnosis of breast cancer by analysis of
   patterns of gene expression and treatment using the
   BRCA1 gene
IN Holt, Jeffrey T.; Jensen, Roy A.; Page, David L.;
   Obermiller, Patrice S.; Robinson-Benion, Cheryl L.;
   Thompson, Marilyn E.
PA Vanderbilt University, USA
SO PCT Int. Appl., 97 pp.
   CODEN: PIXXD2
DT Patent
LA English
FAN.CNT 1
      PATENT NO.      KIND DATE      APPLICATION NO.  DATE
      -----
PI WO 9519369      A1 19950720 WO 1995-US608 19950117
      .
      .
      .
PRAI US 1994-182961 A 19940114
     US 1995-373799 A 19950117
     WO 1995-US608 W 19950117
AB A method of detecting and diagnosing pre-invasive breast
   cancer by identifying differentially expressed genes in
   early, pre-invasive breast cancer tissue is described.
   Differentially expressed genes can be used as genetic
   markers to indicate the presence of pre-invasive cancerous
   tissues. Microscopically directed tissue sampling
   techniques combined with differential display or
   differential screening of cDNA libraries are used to
   determine differential expression of genes in the early
   stages of breast cancer. Differential expression of genes
   in pre-invasive breast cancer tissue is confirmed by RT-
   PCR, nuclease protection assays and in-situ hybridization
   of ductal carcinoma in situ tissue RNA and control tissue
   RNA. The present invention also provides a method of
   screening for compds. that induce expression of the BRCA1
   gene, whose product neg. regulates cell growth in both
   normal and malignant mammary epithelial cells. The use of
   the BRCA1 gene in gene therapy is also discussed.
IT 169596-15-0
   RL: PRP (Properties); THU (Therapeutic use); BIOL
      (Biological study); USES (Uses)
      (PCR primer, in differential display diagnosis of
      breast cancer; early diagnosis of breast cancer by
      anal. of patterns of gene expression and treatment
      using BRCA1 gene)
RN 169596-15-0 CAPLUS
CN DNA, d(C-G-C-C-C-C-T-G-C-G-T-T-A-C-C-C-T-C-C-C-C-G-
   C-C-G) (9CI) (CA INDEX NAME)
SEQ 1 cgcccctgcg ttaccctccc cgccg
```

Searching partial sequences

To find partial sequences or sequences with gaps, repeating units, or alternate units, search the partial sequence in the Subsequence Search (/SQSN) field in REGISTRY. You can use the codes for specific nucleotides or ambiguity codes.

Specific Code

Specific Code	Name
A	adenosine
C	cytosine
G	guanosine
T	thymidine
U	uridine
I	inosine

Ambiguity Codes

Ambiguity Codes	Definition
M	A or C
R	A or G
W	A or T
S	C or G
Y	C or T
K	G or T
V	A or C or G
H	A or C or T
D	A or G or T
B	C or G or T
X	Uncommon nucleotide (abasic or non-purine/non-pyrimidine base-substituted)
N	Unknown nucleotide: A or C or G or T
Z	Nonspecific nucleotide: matches on any of the ambiguity codes

Pattern searching

Complex pattern searching of nucleic acid subsequences is possible using special notations for gaps, repeating residues, and other types of variability.

Gaps

Use this symbol...	To specify a...	Example
.	Gap of one base	=> S TACGGGG.TG/SQSN
.{m}	Gap of m bases	=> S CTCGTGATTA.{5}GG/SQSN
.{m,u}	Gap of m to u bases	=> S ATGGC.{1,50}ATGGC/SQSN
.?	Gap of zero or one base	=> S GATTA.?TTG/SQSN
.*	Gap of zero or more bases	=> S ATCTTCCTGT.*CCCTC/SQSN
.+	Gap of one or more bases	=> S TACGG.+GAGAGCTT/SQSN

Repetition

Use this symbol...	To...	Example
{ } with a number or range	Repeat the preceding unit	=> S GAAT(TAA){2}/SQSN
?	Repeat the preceding unit zero or one time	=> S CAT(CGA)?GGAC/SQSN
*	Repeat the preceding unit zero or more times	=> S CAT(CTG)*TATT/SQSN
+	Repeat the preceding unit one or more times	=> S CAT(CTG)+TATT/SQSN

Other variability options

Use this symbol...	To...	Example
^	Require the base occur at the beginning or the end of the sequence	=> S ^GGAAGGG/SQSN => S CCTC^/SQSN
[]	Specify alternate bases	=> S CATCTG [CG] C/SQSN
[-]	Exclude a base	=> S TTTGGG [-G] TTT/SQSN
 	Specify alternate sequences	=> S TTA TTG/SQSN
&	Join together sequence queries	=> S L1&L2/SQSN (L1 and L2 are sequence queries)

Order of execution for symbols

More than one symbol may be used to create complex sequence queries. If you do not use parentheses in sequence queries, the operations will be executed 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 |

Pattern searching example

Find patents and literature on the following partial sequence: **AGGGTATAAAAA...(CCA|ATG)**, where is a gap of four nucleotides followed by either CCA or ATG.

- 1 Enter REGISTRY.
- 2 Search the partial sequence in the /SQSN field.
- 3 Display the sequence (SEQ).
- 4 Enter the reference databases containing CAS Registry Numbers for the sequences.
- 5 Enter SET MSTEPS ON to create an L-number for a search in each database.
- 6 Search the REGISTRY L-number (L1). Each database is searched, and an L-number answer set is created in each database. A composite L-number (L6) with all references is created.
- 7 Set the arrangement of answers in database order in the process of duplicate identification or elimination.
- 8 Remove duplicates. Answers are arranged in database order.

```
=> FILE REGISTRY
=> S AGGGTATAAAAA...(CCA|ATG)/SQSN
L1          605 AGGGTATAAAAA...(CCA|ATG)/SQSN
=> D 7 SEQ
L1  ANSWER 7 OF 606  REGISTRY  COPYRIGHT 2008 ACS on STN
SEQ   1  gcagggagag  agaactggcc  agggtataaa  aagggccac  aagagaccgg
          =====
          51  ctctaggatc  ccaaggcca  actcccogaa  ccactcaggg  tcctgtggac
          101 agctcaccta  gtggcaatgg  ctccaggctc  ccggacgtcc  ctgctcctgg
          151 cttttgccct  gctctgctg  ccttggtctc  aagaggtctg  tgccgtccaa
          201 accgttccgt  tatccaggct  ttttgaccac  gctatgctcc  aagcccatcg
          251 cgcgcaccag  ctggccattg  acacctacca  ggagtttagg  ctggaagacg
          301 gcagccgccg  gactgggcag  atcctcaagc  agacctacag  caagtttgac
          351 acaaactcgc  acaaccatga  cgactgctc  aagaactacg  ggctgctcta
          401 ctgcttcagg  aaggacatgg  acaaggtcga  gacattcctg  cgcattggtc
          451 agtgccgctc  tgtggagggc  agctgtggct  tctaggtgcc  cgagtagcat
          501 cctgtgacct  ctcccagtg  cctctcctgg  ccctgaaggt  gccactccag
          551 tgcccaccag  ccttgccta  ataaaattaa  gttgtatcat  ttca
HITS AT:   21-39
=> FILE USPATFULL CAPLUS BIOSIS GENBANK
=> SET MSTEPS ON
SET COMMAND COMPLETED
=> S L1
L2          67 FILE USPATFULL
L3          201 FILE CAPLUS
L4           11 FILE BIOSIS
L5          453 FILE GENBANK
TOTAL FOR ALL FILES
L6          732 L1
=> SET DUPORDER FILE
SET COMMAND COMPLETED
=> DUP REM L6
DUPLICATE IS NOT AVAILABLE IN 'GENBANK'.
ANSWERS FROM THESE FILES WILL BE CONSIDERED UNIQUE
PROCESSING COMPLETED FOR L6
L7          668 DUP REM L6 (64 DUPLICATES REMOVED)
          ANSWERS '1-67' FROM FILE USPATFULL
          ANSWERS '68-206' FROM FILE CAPLUS
          ANSWERS '207-215' FROM FILE BIOSIS
          ANSWERS '216-668' FROM FILE GENBANK
```

9 Display references from selected databases.

Answer 10 is from *USPATFULL*.

Answer 75 is from *CAPLUS*.

```
=> D TI PA AB HITRN 10
L7 ANSWER 10 OF 668 USPATFULL on STN      DUPLICATE 17
TI Staphylococcus aureus polynucleotides and sequences
PA Human Genome Sciences, Inc., Rockville, MD, United States
(U.S. corporation)
AB The present invention provides polynucleotide sequences of
the genome of Staphylococcus aureus, polypeptide sequences
encoded by the polynucleotide sequences, corresponding
polynucleotides and polypeptides, vectors and hosts
comprising the polynucleotides, and assays and other uses
thereof. The present invention further provides
polynucleotide and polypeptide sequence information stored
on computer readable media, and computer-based systems and
methods which facilitate its use.
IT 552379-34-7
(nucleotide sequence; Staphylococcus aureus genome
fragment and polypeptide sequences)

=> D L7 BIB AB 75
L7 ANSWER 75 OF 668 CAPLUS COPYRIGHT 2008 ACS on STN
DUPLICATE 18
AN 2003:942764 CAPLUS Full-text
DN 140:3792
TI Genes expressed in atherosclerotic tissue and their
use in diagnosis and pharmacogenetics
IN Nevins, Joseph; West, Mike; Goldschmidt, Pascal
PA Duke University, USA
SO PCT Int. Appl., 408 pp.
CODEN: PIXXD2
DT Patent
LA English
FAN.CNT 5
PATENT NO.      KIND DATE      APPLICATION NO.  DATE
-----
PI WO 2003091391 A2      20031106 WO 2002-XA38221 20021112
      .
      .
      .
AB Genes whose expression is correlated with an determinant of
an atherosclerotic phenotype are provided. Also provided
are methods of using the subject atherosclerotic
determinant genes in diagnosis and treatment methods, as
well as drug screening methods. In addition, reagents and
kits thereof that find use in practicing the subject
methods are provided. Also provided are methods of
determining whether a gene is correlated with a disease
phenotype, where correlation is determined using a Bayesian
anal.
```

Answer 207 is from BIOSIS.

```
=> D L7 207 ALL
L7 ANSWER 207 OF 668 BIOSIS COPYRIGHT (c) 2008 The
  Thomson Corporation on STN
AN 2003:408767 BIOSIS Full-text
DN PREV200300408767
TI Cloning of novel pituitary growth hormone gene from
  Rhinopithecus roxellanae.
AU Ye Chun [Reprint Author]; Zhang Ya-Ping [Reprint
  Author]
CS Laboratory of Molecular Evolution and Genome
  Diversity, Kunming Institute of Zoology, Chinese
  Academy of Sciences, Kunming, 650223, China
  zhangyp@public.km.yn.cn
SO Yichuan, (May 2003) Volume 25, Number 3, pp. 291-
  294. print.
  ISSN: 0253-9772 (ISSN print).
DT Article
LA Chinese
OS DDBJ-AF374232; EMBL-AF374232; GenBank-AF374232;
  DDBJ-AF374234; EMBL-AF374234; GenBank-AF374234;
  DDBJ-AJ297562; EMBL-AJ297562; GenBank-AJ297562;
  DDBJ-AJ297563; EMBL-AJ297563; GenBank-AJ297563;
  DDBJ-J03071; EMBL-J03071; GenBank-J03071; DDBJ-
  L16556; EMBL-L16556; GenBank-L16556
ED Entered STN: 3 Sep 2003
  Last Updated on STN: 3 Sep 2003
AB Putative pituitary growth hormone gene of Rhinopithecus
  roxellanae was cloned and sequenced. All exons sequences
  and deduced amino acid sequence (containing 26 residues
  signal peptide and 191 residues mature protein) were
  obtained. We constructed a phylogenetic tree, which well
  reflected the true evolutionary relationship of pituitary
  growth hormone genes from 7 primates species. From the
  results of amino acids sequence comparison and analysis of
  functionally important sites of growth hormone, pituitary
  growth hormone of macaque from Cercopithecidae and snub-
  nosed golden monkey from Colobidae show little difference.
  We indicated that pituitary growth hormone from
  Cercopithecoidae species have no apparently functional
  difference.
  .
  .
  .
IT Sequence Data
  AF374232: DDBJ, EMBL, GenBank, amino acid sequence,
  nucleotide sequence; AF374234: DDBJ, EMBL, GenBank, amino
  acid sequence, nucleotide sequence; AJ297562: DDBJ, EMBL,
  GenBank, amino acid sequence, nucleotide sequence;
  AJ297563: DDBJ, EMBL, GenBank, amino acid sequence,
  nucleotide sequence; J03071: DDBJ, EMBL, GenBank, amino
  acid sequence, nucleotide sequence; L16556: DDBJ, EMBL,
  GenBank, amino acid sequence, nucleotide sequence
  .
  .
  .
RN 9002-72-6 (growth hormone)
  392361-04-5 (DDBJ, EMBL, GenBank-AF374232)
  392361-06-7 (DDBJ, EMBL, GenBank-AF374234)
  286485-24-3 (DDBJ, EMBL, GenBank-AJ297562)
  286484-88-6 (DDBJ, EMBL, GenBank-AJ297563)
  141145-46-2 (DDBJ, EMBL, GenBank-J03071)
  149765-82-2 (DDBJ, EMBL, GenBank-L16556)
```

Searching length

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

Use this operator...	To indicate...	Example
>	Greater than	=> S SQL>100
<	Less than	=> S SQL<25
=	Equal to	=> S SQL=15 or 15/SQL
<=	Less than or equal to	=> S SQL<=100
>=	Greater than or equal to	=> S SQL=>120
m-n	Range beginning with m and ending with n	=> S 35-100/SQL

Find GCGCTACTGA containing sequences with 20 or fewer nucleotides.

1 Enter *REGISTRY* and search the sequence.

2 Search *SQL<=20* to retrieve only sequences with 20 or fewer residues.

3 Display some answers in the *HIT* format.

```
=> FILE REGISTRY
=> S GCGCTACTGA/SQSN
L3      10910 GCGCTACTGA/SQSN
=> S L3 AND SQL=<20
      4389764 SQL=<20
L4      13 L3 AND SQL=<20
=> D HIT 5-7
L4 ANSWER 5 OF 13 REGISTRY COPYRIGHT 2008 ACS on STN
SQL 19
SEQ      1 aagcauggcg cuacugaaa
              === =====
HITS AT:  8-17
L4 ANSWER 6 OF 13 REGISTRY COPYRIGHT 2008 ACS on STN
SQL 19
SEQ      1 gcaagcaugg cgcuacuga
              = =====
HITS AT:  10-19
L4 ANSWER 7 OF 13 REGISTRY COPYRIGHT 2008 ACS on STN
SQL 19
SEQ      1 gcauggcgcu acugaaagu
              =====
HITS AT:  6-15
```

Searching chemical annotation

In the Annotation (/NTE) field, you can search terms for chemically modified nucleic acids such as:

- Global terms for broad classification of the entire nucleic acid sequence, e.g., metal salt or complex
- Strand-specific terms, e.g., linear or cyclic
- Terms describing the type of chemical modification, e.g., stereoisomer or metal complex

Refer to the Appendix for information on annotation for chemically modified nucleic acids.

In the /NTE field, you can search phrases or single words and combine them using the Boolean operators (AND, OR, NOT). You can use both right and left truncation.

Find sequences with stereochemically modified nucleotides.

1 Enter *REGISTRY*.

2 Search *STEREOISOMER* in the /NTE field.

```
=> FILE REGISTRY
=> S STEREOISOMER/NTE
L1          4532 STEREOISOMER/NTE

=> D SQD

L1  ANSWER 1 OF 4532  REGISTRY  COPYRIGHT 2008 ACS on STN
RN  1042744-59-1  REGISTRY
FS  NUCLEIC ACID SEQUENCE
SQL 18,9,9
NA  5 a  4 c  4 g  3 t  1 u  1 x
NTE multistranded (2)
    modified

-----
type          ----- location -----          description
-----
modified base  u-5[2]                m5u
modified base  u-5[2]                modified uridine
uncommon base  x-4                    unavailable
stereoisomer   u-5[2]                 $\alpha$ -D-ribo
-----

SEQ          1 gcaxatcac
SEQ          1 gtgauatgc
```

Appendix: Annotation for chemically modified nucleic acids

Introduction

Chemically modified nucleic acids with searchable annotation include RNA and DNA sequences containing at least nine nucleotides and peptide nucleic acid sequences composed of at least four base units. The annotation terms and symbols used for describing the chemical structure of modified nucleic acids are defined in Tables 1-11 of this Appendix. The chemical annotation information for modified nucleic acids is located in the Annotation (NTE) field in REGISTRY and is searchable in the /NTE field. The nucleic acid strand, nucleotide residue, and locant positions of the individual modifications are provided as appropriate.

Table 1 lists “Global” terms that are intended to provide a broad description of the modified nucleic acid. All chemically annotated nucleic acids receive the term “modified”. Additional global terms define the number of strands associated with the nucleic acid (i.e., singlestranded, doublestranded, or multistranded) and whether it occurs as a complex with a metal salt or nonmetallic, non-nucleic acid component or exists as an amino acid/peptide conjugate.

Table 2 lists “Strand-specific” terms that define whether the individual strands associated with the nucleic acid are “linear”, “cyclic”, or “homopolymeric”. If two or more strands are present, each strand receives a numerical ranking based first on the total number of individual nucleotide residues beginning at the 5'- (or left) end and proceeding from left to right. The longest strand in a complex is ranked as number “1”, and additional strands are ranked consecutively based on length. In the event that two or more strands are equal in length, the ranking is based on lowest alphabetic of the nucleotide residues beginning at the 5'- (or left) end of each strand. If length and alphabetic result in a tie, further ranking of strands is based on their molecule type with DNA > RNA > peptide nucleic acid (PNA).

Table 3 lists general “Modification” terms that categorize the nucleotide residue modifications in regard to their base and/or linkage and/or sugar structural chemistry (e.g., “modified base”, “modified link”, “stereoisomer”). The location of these modifications is defined by their strand and residue numbers (moving from left to right along the strand).

The remainder of the Tables (4-11) provide lists of descriptors and symbols that more precisely define the general modification terms from Table 3.

When locants are defined for specific chemical descriptors, they follow the accepted nucleotide conventions. Unprimed locants are assigned to the base, and primed locants are assigned to the sugar/sugar substitute in DNA and RNA. The Greek letter “ α ” is the locant assigned to the methyl group at the “5” position of the 2'-deoxythymidine base, and “N” is the locant designation for a substituted amino group on the adenosine, cytidine, or guanosine bases. The locant “P” is used for modifications to the phosphodiester linkage. In peptide nucleic acid (PNA), the purine/pyrimidine bases are assigned the same unprimed locants as in DNA and RNA. However, the 1'-position in PNA is the acetyl carbon to which the base is attached. The amino acid α -nitrogen has the designated locant “N”, and the amino group of the -(2-aminoethyl) attached to “N” is designated as the 5'-position. The 3'-position in PNA is the glycol (or other amino acid) acid carbonyl.

The location of a modified or uncommon nucleic acid linkage is defined as that of the residue which it follows (3'-attachment moving from left to right).

Table 1. Global terms

Term in NTE	Definition
modified	Used for nucleic acid sequences that contain one or more chemical modifications. Also used to define nucleic acids consisting of two or more strands, each of which is a different molecule type (e.g., DNA-RNA, RNA-PNA)
metal salt	Used for nucleic acids that exist as a salt with a metal. The descriptor for these is the metal (e.g., "sodium", "potassium").
complex	Used for nucleic acids that exist as compounds with nonmetallic, non-nucleic acid substances (e.g., drugs, antibiotics). The descriptor for these is "unavailable".
conjugated	Used for nucleic acids that are chemically bonded with at least two amino acids, including peptides and/or proteins. A "conjugated" peptide nucleic acid would contain amino acids in addition to those of the defined backbone.
singlestranded	Used for nucleic acids that consist of one chemically modified strand.
doublestranded	Used for nucleic acids that consist of two strands, at least one of which is chemically modified.
multistranded	Used for nucleic acids that consist of more than two strands, at least one of which is chemically modified.

Table 2. Strand specific terms

Term in NTE	Definition
linear	Used for nucleic acid strands that have free terminal 5'- and 3'-ends that are not chemically esterified and/or bonded to each other.
cyclic	Used for cyclized nucleic acids with no free termini. Cyclic strands are keyed starting at the lowest alphabetic residue position and proceeding leftward to complete the cycle. The initial residue is annotated as a "modified base" - "5'-phosphate" for purposes of cyclization. A strand number is associated with "cyclic" for multistranded nucleic acid complexes.
homopolymer	Used for nucleic acid strands that exist as an indeterminate number of replicates. The 5'-terminal residue is annotated as a "modified base" - "5-phosphate" for purposes of polymerization. The descriptor associated with "homopolymer" is "unavailable". A strand number is associated with "homopolymer" for multistranded nucleic acid complexes. The homopolymer strand in a complex may consist of a single residue type (e.g., 5'-adenylic acid homopolymer).

Table 3. Type of modification terms

Term in NTE	Definition
modified base	Used for nucleotides whose "base" and/or "sugar" have been chemically modified by substitution, esterification, etc. A "modified base" by definition can contain no more than two heteroatom replacements in the purine or pyrimidine ring and can be represented by the normal sequence symbols of "a, c, g, i, u, or t". The "modified base" term is also used to describe nucleotides with modified or substituted ribose sugars, non-ribose sugars, and "nonsugar" (alkyl) replacements. (See Tables 4, 6, and 7 for descriptors.)
uncommon base	Used for nucleotides in which the purine or pyrimidine base has been replaced by some other chemical entity (including purine and pyrimidine ring systems with more than two heteroatom replacements). The sequence symbol used to represent a nucleic acid residue that contains an uncommon base is "x". An "uncommon base" may have a modified ribose or non-ribose sugar associated with it. (Descriptor list is derived from Tables 6 and 7.)
DNA-containing	Used to designate a "2'-deoxyribose"-containing nucleotide (DNA) residue in a RNA or peptide nucleic acid (PNA) sequence. (Table 5 descriptors are da, dc, dg, di, dt, and du.) If a sequence contains an equal number of DNA and RNA residues, it is defined overall as "DNA" for annotation purposes. A sequence must contain more RNA or PNA residues than DNA residues to be defined as "RNA" or "PNA".
RNA-containing	Used to designate a "ribose"-containing nucleotide (RNA) residue in a DNA or PNA sequence. (Table 5 descriptors are ra, rc, rg, ri, and ru.)
PNA-containing	Used to designate a "PNA" residue in a DNA or RNA sequence. (Table 5 descriptors are pa, pc, pg, pi, pt, and pu.) A PNA residue consists of a "N"-substituted amino acid (typically glycine) with an "N"-acetyl-linked nucleic acid "base".
modified link	Used to indicate phosphodiester linkages that have been modified by "P"-substitution or "P"-esterification. The position of a modified link is that of the residue to which it is 3'-attached.
uncommon link	Used when the sugar locants associated with the phosphodiester linkage are modified and/or when the phosphodiester linkage is replaced or lengthened in DNA or RNA. The position of an uncommon link is that of the residue to which it is 3'-attached. (See Table 8 for descriptors.) The term "uncommon link" is also used to annotate chemically modified linkages in PNA sequences, where the normal linkage between the 5'-(2-aminoethyl) and the 3'-carbonyl of the adjoining amino acid residue has been altered.
stereoisomer	Used to indicate modified stereochemistry of the nucleotide sugar and to indicate "R" or "S" stereochemistry associated with the "P" of a substituted phosphodiester linkage. (See Table 10 descriptors.)
metal complex	Used to indicate that the nucleic acid is coordinately complexed with a metal. (See Table 11 for metal descriptors.)
labeled	Used to designate radioisotopic labeling of nucleic acid chemical substituents or atom replacements. (See Table 9 for descriptors.)
covalent bridge	Used to indicate the covalent cross-linking of two or more nucleic acid residues within or between strands (in addition to the normal phosphodiester linkages). The descriptor used is "unavailable" and the annotation indicates the strand(s) and positions of the residues involved.
radical ion	Used when the nucleic acid contains a charged radical ion, generally through esterification or substitution.

Table 4. Shortcut descriptors for “modified base”

Symbol in NTE	Modified Nucleotide	Sequence Symbol
ac4c	<i>N</i> -acetylcytidine	c
ac2g	<i>N</i> -acetylguanosine	g
an4c	<i>N</i> -(4-methoxybenzoyl)cytidine	c
am	2'- <i>O</i> -methyladenosine	a
bz6a	<i>N</i> -benzoyladenosine	a
bz4c	<i>N</i> -benzoylcytidine	c
c7a	7-deazaadenosine	a
c7g	7-deazaguanosine	g
c7i	7-deazainosine	i
cm	2'- <i>O</i> -methylcytidine	c
gm	2'- <i>O</i> -methylguanosine	g
hu	dihydrouridine	u
ib2g	<i>N</i> -(2-methyl-1-oxopropyl)guanosine (<i>N</i> -isobutyrylguanosine)	g
im	2'- <i>O</i> -methylinosine	i
m1a	1-methyladenosine	a
m2a	2-methyladenosine	a
m26a	<i>N,N</i> -dimethyladenosine	a
m6a	<i>N</i> -methyladenosine	a
m3c	3-methylcytidine	c
m5c	5-methylcytidine	c
m1g	1-methylguanosine	g
m2g	<i>N</i> -methylguanosine	g
m22g	<i>N,N</i> -dimethylguanosine	g
m227g	2,2,7-trimethylguanosine	g
m7g	7-methylguanosine	g
m1i	1-methylinosine	i
m7i	7-methylinosine	i
m1p	1-methylpseudouridine	u
m5u	5-methyluridine	u
p	pseudouridine	u
pm	2'- <i>O</i> -methylpseudouridine	u
s2c	2-thiocytidine	c
s6g	6-thioguanosine	g
s6i	6-thioinosine	i
s2t	2-thiothymidine	t
s4t	4-thiothymidine	t
s2u	2-thiouridine	u
s4u	4-thiouridine	u
um	2'- <i>O</i> -methyluridine	u
xan	xanthosine	g

Table 5. Description symbols for hybrid nucleic acids

Hybrid nucleic acid strands contain a mix of DNA, RNA, and/or PNA residues. A “molecule type” (i.e., DNA, RNA, or PNA) is assigned for the entire strand based on the major residue type. Those residues that constitute the minority are annotated using the symbols in this table along with the appropriate broader modification term “DNA-containing”, “RNA-containing”, or “PNA-containing”. In the event that a strand contains an equal number of different residue types, DNA outranks both RNA and PNA and RNA outranks PNA (i.e., DNA > RNA > PNA).

Symbol in NTE	Sequence Symbol	Definition
da	a	Indicates a DNA (2'-deoxyribose-containing) residue in a RNA or a PNA strand
dc	c	
dg	g	
di	i	
dt	t	
du	u	
ra	a	Indicates a RNA (ribose-containing) residue in a DNA or a PNA strand
rc	c	
rg	g	
ri	i	
ru	u	
pa	a	Indicates a PNA residue in a DNA or RNA strand
pc	c	
pg	g	
pi	i	
pt	t	
pu	u	

Table 6 – Description terms for chemical groups and modifications

A locant is assigned to indicate the position of these modifications on the nucleotide base, sugar, or phosphodiester linkage.

Term in NTE	Definition
ac	acetyl
an	anisoyl (4-methoxybenzoyl)
aza	substitution of "N" for "C" in nucleotide base
boc	[(1,1-dimethylethoxy) carbonyl]
br	bromo
bu	butyl
bz	benzoyl
bzl	benzyl (phenylmethyl)
cbz	[(phenylmethoxy) carbonyl]
cho	formyl
cl	chloro
(2-clph)	(2-chlorophenyl)
(3-clph)	(3-chlorophenyl)
(4-clph)	(4-chlorophenyl)
deamino	loss of the -NH ₂ group from the adenosine, cytidine, or guanosine base
deaza	substitution of "C" for "N" in nucleotide base
deoxo	loss of the "=O" (oxo) group from a nucleotide base or a phosphodiester linkage
deoxy	loss of the "-OH" (hydroxy) group from a sugar or a phosphodiester linkage
dmt	dimethoxytrityl or [bis(4-methoxyphenyl)phenylmethyl]
dnp	2,4-dinitrophenyl
dns	dansyl or [[5-(dimethylamino)-1-naphthalenyl]sulfonyl]
et	ethyl
ethenyl	-CH=CH ₂
fl	fluoro
fmoc	[(9H-fluoren-9-ylmethoxy) carbonyl]
ib	isobutyryl (2-methyl-1-oxopropyl)
ibu	isobutyl
io	iodo
ipr	isopropyl
me	methyl
mmt	monomethoxytrityl or [(4-methoxyphenyl)diphenylmethyl]
mo	methoxy
moe	(2-methoxyethyl)
nh ₂	amino
oh	hydroxyl
ph	phenyl
phosphate	phosphate ester
diphosphate	diphosphate ester
triphosphate	triphosphate ester
tetraphosphate	tetraphosphate ester
pentaphosphate	pentaphosphate ester
hexaphosphate	hexaphosphate ester
heptaphosphate	heptaphosphate ester
octaphosphate	octaphosphate ester
nonaphosphate	nonaphosphate ester
decaphosphate	decaphosphate ester
pr	propyl
1-propynyl	-C≡C-CH ₃
2-propenyl	-CH ₂ -CH=CH ₂
sbu	sec-butyl or -CH-CH ₂ -CH ₃ CH ₃

sh	mercapto
tbu	tert-butyl or $\begin{array}{c} \text{CH}_3 \\ \\ - \text{C}-\text{CH}_3 \\ \\ \text{CH}_3 \end{array}$
thio	Replacement of an oxygen (=O or -OH) with a sulfur (=S or -SH). Includes replacement of the ribose or deoxyribose ring "O" with "S" (annotated as "modified base" - "4'-thio").
dithio	Replacement of both oxygens (=O and -OH) on the "P" of a phosphodiester linkage with sulfur
thp	(tetrahydro-2H-pyran-2-yl)
tms	trimethylsilyl
tos	tosyl or [(4-methylphenyl)sulfonyl]
tr	trityl or (triphenylmethyl)

Table 7. Generic Description Terms

The generic terms from this table are used to describe modifications involving common large molecules and when other more specific description terms are too narrow. These terms will include a locant when appropriate.

Term in NTE	Definition
biotin-linked	Used to indicate conjugation with a biotin molecule by means of substitution, esterification, etc.
cyanine dye-linked	Used to indicate conjugation with a cyanine dye molecule (e.g., Cy3 Cy5) by means of substitution, esterification, etc.
digoxigenin-linked	Used to indicate conjugation with a digoxigenin molecule by means of substitution, esterification, etc.
ester	Used to indicate modification via esterification when no other more specific descriptors are available. This includes 5'- and 3'-terminal nucleic acid esters and "P"-esters involving the phosphodiester linkage.
fluorescein-linked	Used to indicate conjugation with a fluorescein molecule or derivative by means of substitution, esterification, etc.
glycosylated	Used to indicate conjugation with a carbohydrate molecule by means of substitution or esterification, in addition to the normal nucleotide backbone sugars.
modified adenosine modified cytidine modified guanosine modified inosine modified thymidine modified uridine	Used when the purine/pyrimidine base has been modified with up to two heteroatom ring system replacements, including additional substitutions on the base. This is also used for "locked" nucleotides containing modified sugars with locking groups like 2'-O,4'-C-methylene. These modified nucleotides are represented by the normal symbols: a, c, g, i, u, or t.
phosphonate	Used to indicate esterification with phosphonic acid or substituted phosphonic acid.
phosphoramidate	Used to indicate esterification with phosphoramidic acid or N-substituted phosphoramidic acid.
phosphorothioate	Used to indicate esterification with phosphorothioic acid or substituted phosphorothioic acid.
photoadduct	Used to describe radiation-induced cyclo or dicyclo adducts of nucleotide bases other than "thymidine dimers". If the adduct bridges two different strands, the corresponding strand and residue positions are defined.
porphyrin-linked	Used to indicate conjugation with a porphyrin molecule or derivative by means of substitution, esterification, etc.
psoralen-linked	Used to indicate conjugation with a psoralen molecule by means of substitution, esterification, etc.
rhodamine-linked	Used to indicate conjugation with a rhodamine dye molecule or derivative by means of substitution, esterification, etc.
substituted	Used to indicate a chemical substitution on a nucleic acid when no more specific descriptors apply. Substitutions occur on the bases, sugars, and phosphodiester linkages of DNA and RNA and on the amino acids, bases, and strand termini of PNAs.
steroid-linked	Used to indicate conjugation with a molecule containing a steroid ring system by means of substitution, esterification, etc.

thymidine dimer	Used to describe cyclo or dicyclo adducts formed from two thymidine bases usually in response to ionizing radiation.
unavailable	This term is used when none of the other descriptors or terms apply. It is used with the strand-specific term "homopolymer" and with the modification term "covalent bridge" and commonly also with "uncommon base" and "uncommon link".

Table 8. Description terms for uncommon linkages

Uncommon linkages in DNA and RNA can be lengthened derivatives of the normal phosphodiester linkage or involve phosphodiester or non-phosphodiester linkages with locant sets other than 3'→5'. When the phosphodiester linkage is replaced with a non-phosphate linkage, it is annotated as “uncommon link” – “unavailable”. Uncommon linkages in PNA involve alterations of the normal amide linkage.

Each of the multiphosphate terms including diphosphate, triphosphate, tetraphosphate, pentaphosphate, hexaphosphate, heptaphosphate, octaphosphate, nonaphosphate, or decaphosphate may be paired with any of the locant sets.

<u>Multiphosphate</u>	<u>Uncommon Locant Sets</u>
diphosphate	2'→2'
triphosphate	2'→3'
tetraphosphate	2'→5'
pentaphosphate	3'→2'
hexaphosphate	3'→3'
heptaphosphate	5'→2'
octaphosphate	5'→3'
nonaphosphate	5'→5'
decaphosphate	unavailable

Table 9. Description terms for isotopically “labeled” nucleotides

Nucleic acids with isotopic modifications are annotated with the term “labeled” and the proper isotope descriptor from the following list. When appropriate, the position of the label is accompanied with a locant designation. For example, a DNA strand with a ³²P-phosphate ester on its 5'-end would be annotated as “labeled” – 5'-P32 with the location being that of the first nucleotide residue.

This isotope list continues to be updated as needed.

Bromine	Br76
Carbon	C10, C11, C12, C13, C14, C15
Fluorine	F18
Hydrogen	H2, H3
Iodine	I125, I131
Nitrogen	N12, N13, N15, N16
Oxygen	O15, O17, O18
Phosphorus	P29, P30, P32, P33
Sulfur	S32, S33, S34, S35, S36, S37, S38

Table 10. Description Terms for Stereoisomers

Stereochemical modifications of the normal ribose sugar in RNA residues and the normal 2'-deoxyribose sugar in DNA residues are annotated with the term "stereoisomer" and the following list of specific description terms. Nucleotides containing more highly modified sugars or sugar stereochemistry not on this list would be annotated with the term "modified base" along with the descriptors "modified adenosine", "modified cytidine", "modified inosine", "modified thymidine", or "modified uridine". The "R" and "S" stereochemical descriptors are used with locant "P" to describe the stereochemistry of substituted phosphodiester linkages.

α -D-arabino
 α -L-arabino
 β -D-arabino
 β -L-arabino
 α -D-erythro
 α -L-erythro
 β -D-erythro
 β -L-erythro
 α -D-lyxo
 α -L-lyxo
 β -D-lyxo
 β -L-lyxo
 α -D-ribo

α -L-ribo
 β -D-ribo
 β -L-ribo
 α -D-threo
 α -L-threo
 β -D-threo
 β -L-threo
 α -D-xylo
 α -L-xylo
 β -D-xylo
 β -L-xylo
R
S

Table 11. Description terms for “Metal Complex”

When nucleic acids are modified with coordinated metals by esterification or substitution, they are annotated with the term “metal complex” and the following list of specific metal descriptors. The location of the metal complex is annotated with a locant position when appropriate.

Symbol in NTE	Metal
Ac	Actinium
Ag	Silver
Al	Aluminum
Am	Americium
Au	Gold
Ba	Barium
Be	Beryllium
Bi	Bismuth
Bk	Berkelium
Ca	Calcium
Cd	Cadmium
Ce	Cerium
Cf	Californium
Cm	Curium
Co	Cobalt
Cr	Chromium
Cs	Cesium
Cu	Copper
Dy	Dysprosium
Er	Erbium
Es	Einsteinium
Eu	Europium
Fe	Iron
Fm	Fermium
Fr	Francium
Ga	Gallium
Gd	Gadolinium
Ge	Germanium
Hf	Hafnium
Hg	Mercury
Ho	Holmium
In	Indium
Ir	Iridium
K	Potassium
La	Lanthanum
Li	Lithium
Lr	Lawrencium
Lu	Lutetium
Md	Mendelevium
Mg	Magnesium
Mn	Manganese
Mo	Molybdenum
Na	Sodium
Nb	Niobium
Nd	Neodymium
Ni	Nickel
No	Nobelium
Np	Neptunium
Os	Osmium

Pa	Protactinium
Pb	Lead
Pd	Palladium
Pm	Promethium
Po	Polonium
Pr	Praseodymium
Pt	Platinum
Pu	Plutonium
Ra	Radium
Rb	Rubidium
Re	Rhenium
Rh	Rhodium
Ru	Ruthenium
Sb	Antimony
Sc	Scandium
Sm	Tin
Sr	Strontium
Ta	Tantalum
Tb	Terbium
Tc	Technetium
Th	Thorium
Ti	Titanium
Tl	Thallium
Tm	Thulium
U	Uranium
V	Vanadium
W	Tungsten
Y	Yttrium
Yb	Ytterbium
Zn	Zinc
Zr	Zirconium



A division of the
American Chemical Society

November 2008
CAS2536-1108

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