Innovation, Science and Economic Development Canada
Canadian Intellectual Property Office

Innovation, Sciences et Développement économique Canada
Office de la propriété intellectuelle du Canada

Canadian Intellectual Property Office

Patent Biological Sequence Listings

Data Dictionary

WIPO Standard ST.25

Version 1.00

2019-01-08

Canada

Ce document est aussi disponible en français sous le titre Listages des séquences biologiques de brevets - Dictionnaire de données. This publication is available upon request in accessible formats.

Contact

Client Service Centre
Canadian Intellectual Property Office
Innovation, Science and Economic Development Canada
Place du Portage I
Room C229, 2nd floor
50 Victoria Street
Gatineau, QC K1A 0C9

Telephone (toll-free): 1-866-997-1936
TTY: 1-866-442-2476
Fax: 819-953-2476
ic.contact-contact.ic@canada.ca

Permission to reproduce

Cat. No. Iu71-4/63-2019E-PDF
ISBN 978-0-660-29225-0

TABLE OF CONTENTS

1.0 Overview of Canadian Biological Sequence Listings Data

Patents apply to newly developed technology as well as to improvements on products or processes. A full description of what constitutes a Patent is described in the Patent Act. A patent is a right, granted by government, to exclude others from making, using, or selling your invention in Canada.

Canadian Biological Sequence Listings data contain information on nucleotide and amino acid sequences. This section will give an overview of what information is supplied and a description of the file structures.

1.1 Biological Sequence Listings Data

A patent's biological sequence listing is presented as a separate part of the description of a patent application/granted patent and as a separate collection of data. The file structure is governed by World Intellectual Property Organization (WIPO) standard ST.25 for the presentation of nucleotide and amino acid sequence listings in patent applications.

Biological Sequence Listings (BSL) data consists of the applicant submitted file (often a text file) and two additional file types (.PEP and .SEQ) generated by the Office. The official documentation for a Biological Sequence Listing patent application/granted patent is the file supplied by the applicant. The Office generated files are created as working files and may be incomplete if the file submitted by the client is incomplete. The Office will request a new sequence listing from the client when it is deemed incomplete.

It should also be noted that applications entering the national phase import the sequence listing from the international application at the World Intellectual Property Organization (WIPO) and at times these downloaded sequence listings are incomplete.

1.2 Weekly Data File

BSL in patent applications and granted patents are included in a weekly update file. Patent data is open to public inspection after a confidentiality period of up to 18 months after the filing date of an application, or the earliest filing date of any previously regularly filed application on which a request for priority has been made.

A weekly collection contains TXT, PEP, and SEQ files. This collection contains updated and new files for the current week. The weekly data file ranges in size from 1 to 350 Mb depending on the volume of activity. The weekly collection includes a report that lists all the files included in that week's extract.

Mandatory data elements in the applicant-provided file include:

- **Bibliographical Information** (applicant name, title of invention)
- **Sequence Listing Information** (number of sequence ID numbers, SEQ ID number, Length, Type, Organism, Sequence).

1.3 Description of ST.25 Structure

The file structure is governed by the World Intellectual Property Organization (WIPO) standard ST.25 for the presentation of nucleotide and amino acid sequence listings in patent applications. All information regarding this standard as it relates to the file structure and technical specification can be found on the WIPO website at:
https://www.wipo.int/export/sites/www/standards/en/pdf/03-25-01.pdf

Weekly update files are structured as follows:

BSL Weekly Zip File:
BSLYYYYWW.zip (ie. BSL201825.zip) contains a folder for the week of the extract and a report file where YYYY represents the year and WW represents the week number (01 to 52 weeks)

- Weekly Extract Folder:
  - YYYY-MM-DD (i.e. 2018-06-23) contains folders named XXXXXXXX (i.e. 02836299) where X represents the patent application number.

    - Patent numbered folders contain PEP, SEQ and or TXT files:
      - These filenames contain a prefix (CA), the patent application number, the production date, the file type and version CAXXXXXXXXYYYYMMDD-DNAvXX.PEP (or .SEQ or .TXT)
      - Examples:
        - CA0278203320180618-DNAv03.PEP;
        - CA0278203320180618-DNAv04.SEQ;
        - CA0278203320180618-DNAv03.TXT)

-Report file:

- Report.txt contains the extract range for the week, the number of patent applications processed, the number of files output and a complete list of all patents along with their PEP, SEQ and TXT filenames included in the weekly extract. Some basic information regarding laid open dates, national entry and PCT are also included in the report.txt file for each patent application in the weekly extract.

No.	Data	ST.25 Numeric Identifier	Mandatory (M), Conditional, Mandatory (C), or Optional (O)	Description
2.0 Canadian Patent Biological Sequence Listings				
1.	**Applicant name**	**<110>** *MOUNT SINAI HOSPITAL*	M	This numeric identifier is followed by the applicant name. This may include an individual and/or a company. Note: If the applicant name is written in characters other than those of the Latin alphabet, the value will be a translation or transliteration.
2.	**Title of invention**	**<120>** *METHODS AND COMPOSITIONS FOR MODULATING A STEROID RECEPTOR*	M	This numeric identifier is followed by the title of the invention.
3.	**File reference**	**<130>** *064016-379280*	C	This numeric identifier is followed by the physical file reference to the application. Note: This numeric identifier appears if the sequence listing was furnished at any time prior to the assignment of an application number.
4.	**Current patent application**	**<140>** *PCT/CA2005/000042*	C	This numeric identifier is followed by the patent application's number. Note: This numeric identifier appears if the sequence listing was furnished following the assignment of an application number.
5.	**Current filing date**	**<141>** *2005-01-14*	C	This numeric identifier is followed by the filing date of the application. Note: This numeric identifier appears if the sequence listing was furnished following the assignment of an application number.
6.	**Earlier patent application**	**<150>** *60/536,598*	C	This numeric identifier is followed by the application number of an earlier application when a sequence listing is filed relating to an application which claims the priority of an earlier application.
7.	**Earlier application filing date**	**<151>** *2004-01-15*	C	This numeric identifier is followed by the date of filing of an earlier application when a sequence listing is filed relating to an application which claims the priority of an earlier application.
8.	**Number of sequence identification numbers**	**<160>** *22*	M	This numeric identifier is followed by the number of sequences (<400>) found in the document.
9.	**Software**	**<170>** *PatentIn version 3.3*	O	This numeric identifier is followed by the software used in the creation of sequence listings for inclusions in patent applications.

No.	Data	ST.25 Numeric Identifier	Mandatory (M), Conditional, Mandatory (C), or Optional (O)	Description
10.	**Sequence identification number**	**<210>** *1*	M	This numeric identifier is followed by the sequence identification number related to the following sequence (<400>).
11.	**Length**	**<211>** *707*	M	This numeric identifier is followed by the length of the sequence (i.e. number of base pairs or amino acids).
12.	**Type**	**<212>** *PRT*	M	This numeric identifier is followed by the type of molecule sequenced. The type can be DNA, RNA or PRT. Note: If a nucleotide sequence contains both DNA and RNA fragments, the value following this numeric identifier will be DNA.
13.	**Organism**	**<213>** *Homo sapiens*	M	This numeric identifier is followed by the Genus Species/scientific name of the molecule sequenced. This value may also be “Artificial Sequence” or “Unknown”.
14.	**Feature**	**<220>**	C	This numeric identifier is always blank, but is followed by other numeric identifiers having a description of points of biological significance in the sequence. Note: This numeric identifier is used when “n” or “Xaa” or a modified base or modified/unusual L-amino acid is used in the sequence or if the organism (numeric identifier <213>) is “Artificial Sequence” or “Unknown”.
15.	**Name/key**	**<221>**	C	This numeric identifier is followed by the feature key representing points of biological significance in the sequence. These feature keys and their definition can be found in Appendix E and Appendix F. Note: This numeric identifier is used when “n” or “Xaa” or a modified base or modified/unusual L-amino acid is used in the sequence.
16.	**Location**	**<222>**	C	This numeric identifier is followed by the location of the points of biological significance in the sequence as per the length of the sequence. Note: This numeric identifier is used when “n” or “Xaa” or a modified base or modified/unusual L-amino acid is used in the sequence.

No.	Data	ST.25 Numeric Identifier	Mandatory (M), Conditional, Mandatory (C), or Optional (O)	Description
17.	**Other information**	**<223>**	C	This numeric identifier is followed by the description of points of biological significance in the sequence. Note: This numeric identifier is used when "n" or "Xaa" or a modified base or modified/unusual L-amino acid is used in the sequence or if the organism (numeric identifier <213>) is "Artificial Sequence" or "Unknown".
18.	**Publication information**	**<300>**	O	This numeric identifier is always blank but is followed by publication information.
19.	**Authors**	**<301>**	O	This numeric identifier is followed by the name of the author of the publication in which the sequence listing can be found.
20.	**Title**	**<302>**	O	This numeric identifier is followed by the title of publication in the journal in which the sequence listing can be found.
21.	**Journal**	**<303>**	O	This numeric identifier is followed by the title of the journal in which the sequence listing can be found.
22.	**Volume**	**<304>**	O	This numeric identifier is followed by the volume of the journal in which the sequence listing can be found.
23.	**Issue**	**<305>**	O	This numeric identifier is followed by the issue of the journal in which the sequence listing can be found.
24.	**Pages**	**<306>**	O	This numeric identifier is followed by the range of pages of the journal in which the sequence listing can be found.
25.	**Date**	**<307>**	O	This numeric identifier is followed by the date of publication of the journal in which the sequence listing can be found.
26.	**Database accession number**	**<308>** *P23246*	O	Accession number assigned by database including database name.
27.	**Database entry date**	**<309>** *2004-06-15*	O	Date of entry in database
28.	**Document number**	**<310>**	O	Document number, for patent-type citations only.
29.	**Filing date**	**<311>**	O	Document filing date, for patent-type citations only.

No.	Data	ST.25 Numeric Identifier	Mandatory (M), Conditional, Mandatory (C), or Optional (O)	Description
30.	**Publication date**	**<312>**	O	Document publication date; for patent-type citations only.
31.	**Relevant residues in SEQ ID NO: x**	**<313>** *(1)..(707)*	O	Sequence ID numbers (SEQ ID NOS) referred to in the cited publication.
32.	**Sequence**	**<400>** *Met Ser Arg Asp Arg Phe Arg Ser Arg Gly Gly Gly Gly Gly Gly Phe* *1 5 10 15* *His Arg Arg Gly Gly Gly Gly Gly Arg Gly Gly Leu His Asp Phe Arg* *20 25 30* *Ser Pro Pro Pro Gly Met Gly Leu Asn Gln Asn Arg Gly Pro Met Gly* *35 40 45*	M	

Appendix A - List of Nucleotides

Nucleotides Symbol	Meaning	Origin of designation
a	a	adenine
g	g	guanine
c	c	cytosine
t	t	thymine
u	u	uracil
r	g or a	purine
y	t/u or c	pyrimidine
m	a or c	amino
k	g or t/u	keto
s	g or c	strong interactions 3H-bonds
w	a or t/u	weak interactions 2H-bonds
b	g or c or t/u	not a
d	a or g or t/u	not c
h	a or c or t/u	not g
v	a or g or c	not t, not u
n	a or g or c or t/u, unknown, or other	any

Appendix B - List of Modified Nucleotides

Symbol	Meaning
ac4c	4-acetylcytidine
chm5u	5-(carboxyhydroxymethyl)uridine
cm	2'-O-methylcytidine
cmnm5s2u	5-carboxymethylaminomethyl-2-thiouridine
cmnm5u	5-carboxymethylaminomethyluridine
d	dihydrouridine
fm	2'-O-methylpseudouridine
gal q	beta, D-galactosylqueuosine
gm	2'-O-methylguanosine
i	inosine
i6a	N6-isopentenyladenosine
m1a	1-methyladenosine
m1f	1-methylpseudouridine
m1g	1-methylguanosine
m1i	1-methylinosine
m22g	2,2-dimethylguanosine
m2a	2-methyladenosine
m2g	2-methylguanosine
m3c	3-methylcytidine
m5c	5-methylcytidine
m6a	N6-methyladenosine
m7g	7-methylguanosine
mam5u	5-methylaminomethyluridine
mam5s2u	5-methoxyaminomethyl-2-thiouridine
man q	beta, D-mannosylqueuosine
mcm5s2u	5-methoxycarbonylmethyl-2-thiouridine
mcm5u	5-methoxycarbonylmethyluridine
mo5u	5-methoxyuridine
ms2i6a	2-methylthio-N6-isopentenyladenosine
ms2t6a	N-((9-beta-D-ribofuranosyl-2-methylthiopurine-6-yl)carbamoyl)threonine
mt6a	N-((9-beta-D-ribofuranosylpurine-6-yl)N-methylcarbamoyl)threonine
mv	uridine-5-oxyacetic acid-methylester
o5u	uridine-5-oxyacetic acid

osyw	wybutoxosine
p	pseudouridine
q	queuosine
s2c	2-thiocytidine
s2t	5-methyl-2-thiouridine
s2u	2-thiouridine
s4u	4-thiouridine
t	5-methyluridine
t6a	N-((9-beta-D-ribofuranosylpurine-6-yl)-carbamoyl)threonine
tm	2’-O-methyl-5-methyluridine
um	2’-O-methyluridine
yw	wybutosine
x	3-(3-amino-3-carboxy-propyl)uridine, (acp3)u

Appendix C - List of Amino Acids

Symbol	**Meaning**
Ala	Alanine
Cys	Cysteine
Asp	Aspartic Acid
Glu	Glutamic Acid
Phe	Phenylalanine
Gly	Glycine
His	Histidine
Ile	Isoleucine
Lys	Lysine
Leu	Leucine
Met	Methionine
Asn	Asparagine
Pro	Proline
Gln	Glutamine
Arg	Arginine
Ser	Serine
Thr	Threonine
Val	Valine
Trp	Tryptophan
Tyr	Tyrosine
Asx	Asp or Asn
Glx	Glu or Gln
Xaa	unknown or other

Appendix D - List of Modified and Unusual Amino Acids

Symbol	Meaning
Aad	2-Aminoadipic acid
bAad	3-Aminoadipic acid
bAla	beta-Alanine, beta-Aminopropionic acid
Abu	2-Aminobutyric acid
4Abu	4-Aminobutyric acid, piperidinic acid
Acp	6-Aminocaproic acid
Ahe	2-Aminoheptanoic acid
Aib	2-Aminoisobutyric acid
bAib	3-Aminoisobutyric acid
Apm	2-Aminopimelic acid
Dbu	2,4 Diaminobutyric acid
Des	Desmosine
Dpm	2,2’-Diaminopimelic acid
Dpr	2,3-Diaminopropionic acid
EtGly	N-Ethylglycine
EtAsn	N-Ethylasparagine
Hyl	Hydroxylysine
aHyl	allo-Hydroxylysine
3Hyp	3-Hydroxyproline
4Hyp	4-Hydroxyproline
Ide	Isodesmosine
aIle	allo-Isoleucine
MeGly	N-Methylglycine, sarcosine
MeIle	N-Methylisoleucine
MeLys	6-N-Methyllysine
MeVal	N-Methylvaline
Nva	Norvaline
Nle	Norleucine
Orn	Ornithine

Appendix E - List of Feature Keys Related to Nucleotide Sequences

Key	Description
allele	a related individual or strain contains stable, alternative forms of the same gene which differs from the presented sequence at this location (and perhaps others)
attenuator	(1) region of DNA at which regulation of termination of transcription occurs, which controls the expression of some bacterial operons; (2) sequence segment located between the promoter and the first structural gene that causes partial termination of transcription
C_region	constant region of immunoglobulin light and heavy chains, and T-cell receptor alpha, beta, and gamma chains; includes one or more exons depending on the particular chain
CAAT_signal	CAAT box; part of a conserved sequence located about 75 bp up-stream of the start point of eukaryotic transcription units which may be involved in RNA polymerase binding; consensus=GG (C or T) CAATCT
CDS	coding sequence; sequence of nucleotides that corresponds with the sequence of amino acids in a protein (location includes stop codon); feature includes amino acid conceptual translation
conflict	independent determinations of the “same” sequence differ at this site or region
D-loop	displacement loop; a region within mitochondrial DNA in which a short stretch of RNA is paired with one strand of DNA, displacing the original partner DNA strand in this region; also used to describe the displacement of a region of one strand of duplex DNA by a single stranded invader in the reaction catalyzed by RecA protein
D-segment	diversity segment of immunoglobulin heavy chain, and T-cell receptor beta chain
enhancer	a cis-acting sequence that increases the utilization of (some) eukaryotic promoters, and can function in either orientation and in any location (upstream or downstream) relative to the promoter
exon	region of genome that codes for portion of spliced mRNA; may contain 5’UTR, all CDSs, and 3’UTR
GC_signal	GC box; a conserved GC-rich region located upstream of the start point of eukaryotic transcription units which may occur in multiple copies or in either orientation; consensus=GGGCGG
gene	region of biological interest identified as a gene and for which a name has been assigned
iDNA	intervening DNA; DNA which is eliminated through any of several kinds of recombination
intron	a segment of DNA that is transcribed, but removed from within the transcript by splicing together the sequences (exons) on either side of it
J_segment	joining segment of immunoglobulin light and heavy chains, and T-cell receptor alpha, beta, and gamma chains
LTR	long terminal repeat, a sequence directly repeated at both ends of a defined sequence, of the sort typically found in retroviruses
mat_peptide	mature peptide or protein coding sequence; coding sequence for the mature or final peptide or protein product following post-translational modification; the location does not include the stop codon (unlike the corresponding CDS)
misc_binding	site in nucleic acid which covalently or non-covalently binds another moiety that cannot be described by any other Binding key (primer_bind or protein_bind)
misc_difference	feature sequence is different from that presented in the entry and cannot be described by any other Difference key (conflict, unsure, old_sequence, mutation, variation, allele, or modified_base)
misc_feature	region of biological interest which cannot be described by any other feature key; a new or rare feature
misc_recomb	site of any generalized, site-specific or replicative recombination event where there is a breakage and reunion of duplex DNA that cannot be described by other recombination keys (iDNA and virion) or qualifiers of source key (/insertion_seq, /transposon, /proviral)
misc_RNA	any transcript or RNA product that cannot be defined by other RNA keys (prim_transcript, precursor_RNA, mRNA, 5’clip, 3’clip, 5’UTR, 3’UTR, exon, CDS, sig_peptide, transit_peptide, mat_peptide, intron, polyA_site,

	rRNA, tRNA, scRNA, and snRNA)
misc_signal	any region containing a signal controlling or altering gene function or expression that cannot be described by other Signal keys (promoter, CAAT_signal, TATA_signal, -35_signal, -10_signal, GC_signal, RBS, polyA_signal, enhancer, attenuator, terminator, and rep_origin)
misc_structure	any secondary or tertiary structure or conformation that cannot be described by other Structure keys (stem_loop and D-loop)
modified_base	the indicated nucleotide is a modified nucleotide and should be substituted for by the indicated molecule (given in the mod_base qualifier value)
mRNA	messenger RNA; includes 5’ untranslated region (5’UTR), coding sequences (CDS, exon) and 3’ untranslated region (3’UTR)
mutation	a related strain has an abrupt, inheritable change in the sequence at this location
N_region	extra nucleotides inserted between rearranged immunoglobulin segments
old_sequence	the presented sequence revises a previous version of the sequence at this location
polyA_signal	recognition region necessary for endonuclease cleavage of an RNA transcript that is followed by polyadenylation; consensus=AATAAA
polyA_site	site on an RNA transcript to which will be added adenine residues by post-transcriptional polyadenylation
precursor_RNA	any RNA species that is not yet the mature RNA product; may include 5’ clipped region (5’clip), 5’ untranslated region (5’UTR), coding sequences (CDS, exon), intervening sequences (intron), 3’ untranslated region (3’UTR), and 3’ clipped region (3’clip)
prim_transcript	primary (initial, unprocessed) transcript; includes 5’ clipped region (5’clip), 5’ untranslated region (5’UTR), coding sequences (CDS, exon), intervening sequences (intron), 3’ untranslated region (3’UTR), and 3’ clipped region (3’clip)
primer_bind	non-covalent primer binding site for initiation of replication, transcription, or reverse transcription; includes site(s) for synthetic, for example, PCR primer elements
promoter	region on a DNA molecule involved in RNA polymerase binding to initiate transcription
protein_bind	non-covalent protein binding site on nucleic acid
RBS	ribosome binding site
repeat_region	region of genome containing repeating units
repeat_unit	single repeat element
rep_origin	origin of replication; starting site for duplication of nucleic acid to give two identical copies
rRNA	mature ribosomal RNA; the RNA component of the ribonucleoprotein particle (ribosome) which assembles amino acids into proteins
S_region	switch region of immunoglobulin heavy chains; involved in the rearrangement of heavy chain DNA leading to the expression of a different immunoglobulin class from the same B-cell
satellite	many tandem repeats (identical or related) of a short basic repeating unit; many have a base composition or other property different from the genome average that allows them to be separated from the bulk (main band) genomic DNA
scRNA	small cytoplasmic RNA; any one of several small cytoplasmic RNA molecules present in the cytoplasm and (sometimes) nucleus of a eukaryote
sig_peptide	signal peptide coding sequence; coding sequence for an N-terminal domain of a secreted protein; this domain is involved in attaching nascent polypeptide to the membrane; leader sequence
snRNA	small nuclear RNA; any one of many small RNA species confined to the nucleus; several of the snRNAs are involved in splicing or other RNA processing reactions

source	identifies the biological source of the specified span of the sequence; this key is mandatory; every entry will have, as a minimum, a single source key spanning the entire sequence; more than one source key per sequence is permissable
stem_loop	hairpin; a double-helical region formed by base-pairing between adjacent (inverted) complementary sequences in a single strand of RNA or DNA
STS	Sequence Tagged Site; short, single-copy DNA sequence that characterizes a mapping landmark on the genome and can be detected by PCR; a region of the genome can be mapped by determining the order of a series of STSs
TATA_signal	TATA box; Goldberg-Hogness box; a conserved AT-rich septamer found about 25 bp before the start point of each eukaryotic RNA polymerase II transcript unit which may be involved in positioning the enzyme for correct initiation; consensus=TATA(A or T)A(A or T)
terminator	sequence of DNA located either at the end of the transcript or adjacent to a promoter region that causes RNA polymerase to terminate transcription; may also be site of binding of repressor protein
transit_peptide	transit peptide coding sequence; coding sequence for an N-terminal domain of a nuclear-encoded organellar protein; this domain is involved in post-translational import of the protein into the organelle
tRNA	mature transfer RNA, a small RNA molecule (75-85 bases long) that mediates the translation of a nucleic acid sequence into an amino acid sequence
unsure	author is unsure of exact sequence in this region
V_region	variable region of immunoglobulin light and heavy chains, and T-cell receptor alpha, beta, and gamma chains; codes for the variable amino terminal portion; can be made up from V_segments, D_segments, N_regions, and J_segments
V_segment	variable segment of immunoglobulin light and heavy chains, and T-cell receptor alpha, beta, and gamma chains; codes for most of the variable region (V_region) and the last few amino acids of the leader peptide
variation	a related strain contains stable mutations from the same gene (for example, RFLPs, polymorphisms, etc.) which differ from the presented sequence at this location (and possibly others)
3’clip	3’-most region of a precursor transcript that is clipped off during processing
3’UTR	region at the 3’ end of a mature transcript (following the stop codon) that is not translated into a protein
5’clip	5’-most region of a precursor transcript that is clipped off during processing
5’UTR	region at the 5’ end of a mature transcript (preceding the initiation codon) that is not translated into a protein
-10_signal	pribnow box; a conserved region about 10 bp upstream of the start point of bacterial transcription units which may be involved in binding RNA polymerase; consensus=TAtAaT
-35_signal	a conserved hexamer about 35 bp upstream of the start point of bacterial transcription units; consensus=TTGACa [ ] or TGTTGACA [ ]

Appendix F - List of Feature Keys Related to Protein Sequences

Key	Description
CONFLICT	different papers report differing sequences
VARIANT	authors report that sequence variants exist
VARSPLIC	description of sequence variants produced by alternative splicing
MUTAGEN	site which has been experimentally altered
MOD_RES	post-translational modification of a residue
ACETYLATION	N-terminal or other
AMIDATION	generally at the C-terminal of a mature active peptide
BLOCKED	undetermined N- or C-terminal blocking group
FORMYLATION	of the N-terminal methionine
GAMMA-CARBOXYGLUTAMIC ACID HYDROXYLATION	of asparagine, aspartic acid, proline or lysine
METHYLATION	generally of lysine or arginine
PHOSPHORYLATION	of serine, threonine, tyrosine, aspartic acid or histidine
PYRROLIDONE CARBOXYLIC ACID	N-terminal glutamate which has formed an internal cyclic lactam
SULFATATION	generally of tyrosine
LIPID	covalent binding of a lipidic moiety
MYRISTATE	myristate group attached through an amide bond to the N-terminal glycine residue of the mature form of a protein or to an internal lysine residue
PALMITATE	palmitate group attached through a thioether bond to a cysteine residue or through an ester bond to a serine or threonine residue
FARNESYL	farnesyl group attached through a thioether bond to a cysteine residue
GERANYL-GERANYL	geranyl-geranyl group attached through a thioether bond to a cysteine residue
GPI-ANCHOR	glycosyl-phosphatidylinositol (GPI) group linked to the alpha-carboxyl group of the C-terminal residue of the mature form of a protein
N-ACYL DIGLYCERIDE	N-terminal cysteine of the mature form of a prokaryotic lipoprotein with an amide-linked fatty acid and a glyceryl group to which two fatty acids are linked by ester linkages
DISULFID	disulfide bond; the ‘FROM’ and ‘TO’ endpoints represent the two residues which are linked by an intra-chain disulfide bond; if the ‘FROM’ and ‘TO’ endpoints are identical, the disulfide bond is an interchain one and the description field indicates the nature of the cross-link
THIOLEST	thiolester bond; the ‘FROM’ and ‘TO’ endpoints represent the two residues which are linked by the thiolester bond
THIOETH	thioether bond; the ‘FROM’ and ‘TO’ endpoints represent the two residues which are linked by the thioether bond
CARBOHYD	glycosylation site; the nature of the carbohydrate (if known) is given in the description field
METAL	binding site for a metal ion; the description field indicates the nature of the metal
BINDING	binding site for any chemical group (co-enzyme, prosthetic group, etc.); the chemical nature of

	the group is given in the description field
SIGNAL	extent of a signal sequence (prepeptide)
TRANSIT	extent of a transit peptide (mitochondrial, chloroplastic, or for a microbody)
PROPEP	extent of a propeptide
CHAIN	extent of a polypeptide chain in the mature protein
PEPTIDE	extent of a released active peptide
DOMAIN	extent of a domain of interest on the sequence; the nature of that domain is given in the description field
CA_BIND	extent of a calcium-binding region
DNA_BIND	extent of a DNA-binding region
NP_BIND	extent of a nucleotide phosphate binding region; the nature of the nucleotide phosphate is indicated in the description field
TRANSMEM	extent of a transmembrane region
ZN_FING	extent of a zinc finger region
SIMILAR	extent of a similarity with another protein sequence; precise information, relative to that sequence is given in the description field
REPEAT	extent of an internal sequence repetition
HELIX	secondary structure: Helices, for example, Alpha-helix, 3(10) helix, or Pi-helix
STRAND	secondary structure: Beta-strand, for example, Hydrogen bonded beta-strand, or Residue in an isolated beta-bridge
TURN	secondary structure Turns, for example, H-bonded turn (3-turn, 4-turn or 5-turn)
ACT_SITE	amino acid(s) involved in the activity of an enzyme
SITE	any other interesting site on the sequence
INIT_MET	the sequence is known to start with an initiator methionine
NON_TER	the residue at an extremity of the sequence is not the terminal residue; if applied to position 1, this signifies that the first position is not the N-terminus of the complete molecule; if applied to the last position, it signifies that this position is not the C-terminus of the complete molecule; there is no description field for this key
NON_CONS	non consecutive residues; indicates that two residues in a sequence are not consecutive and that there are a number of unsequenced residues between them
UNSURE	uncertainties in the sequence; used to describe region(s) of a sequence for which the authors are unsure about the sequence assignment